

CHEMOINFORMATICS APPROACHES TO STRUCTURE- AND LIGAND-BASED DRUG DESIGN, VOLUME II

EDITED BY: Adriano D. Andricopulo and Leonardo L. G. Ferreira
PUBLISHED IN: Frontiers in Pharmacology





frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88976-631-4

DOI 10.3389/978-2-88976-631-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

CHEMOINFORMATICS APPROACHES TO STRUCTURE- AND LIGAND-BASED DRUG DESIGN, VOLUME II

Topic Editors:

Adriano D. Andricopulo, University of Sao Paulo, Brazil

Leonardo L. G. Ferreira, University of São Paulo, Brazil

Citation: Andricopulo, A. D., Ferreira, L. L. G., eds. (2022). Chemoinformatics Approaches to Structure- and Ligand-Based Drug Design, Volume II. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-631-4

Table of Contents

- 05 Editorial: Chemoinformatics Approaches to Structure- and Ligand-Based Drug Design, Volume II**
Leonardo L. G. Ferreira and Adriano D. Andricopulo
- 08 In silico Analyses of Immune System Protein Interactome Network, Single-Cell RNA Sequencing of Human Tissues, and Artificial Neural Networks Reveal Potential Therapeutic Targets for Drug Repurposing Against COVID-19**
Andrés López-Cortés, Patricia Guevara-Ramírez, Nikolaos C. Kyriakidis, Carlos Barba-Ostria, Ángela León Cáceres, Santiago Guerrero, Esteban Ortiz-Prado, Cristian R. Munteanu, Eduardo Tejera, Doménica Cevallos-Robalino, Ana María Gómez-Jaramillo, Katherine Simbaña-Rivera, Adriana Granizo-Martínez, Gabriela Pérez-M, Silvana Moreno, Jennyfer M. García-Cárdenas, Ana Karina Zambrano, Yunierkis Pérez-Castillo, Alejandro Cabrera-Andrade, Lourdes Puig San Andrés, Carolina Proaño-Castro, Jhommara Bautista, Andreina Quevedo, Nelson Varela, Luis Abel Quiñones and César Paz-y-Miño
- 32 Antioxidant Activity, Molecular Docking, Quantum Studies and In Vivo Antinociceptive Activity of Sulfonamides Derived From Carvacrol**
Aldo S. de Oliveira, Luana C. Llanes, Ricardo J. Nunes, Catharina Nucci-Martins, Anacleto S. de Souza, David L. Palomino-Salcedo, María J. Dávila-Rodríguez, Leonardo L. G. Ferreira, Adair R. S. Santos and Adriano D. Andricopulo
- 48 3D-QSAR, Molecular Docking, and MD Simulations of Anthraquinone Derivatives as PGAM1 Inhibitors**
Yuwei Wang, Yifan Guo, Shaojia Qiang, Ruyi Jin, Zhi Li, Yuping Tang, Elaine Lai Han Leung, Hui Guo and Xiaojun Yao
- 66 Machine Learning Enables Accurate and Rapid Prediction of Active Molecules Against Breast Cancer Cells**
Shuyun He, Duancheng Zhao, Yanle Ling, Hanxuan Cai, Yike Cai, Jiquan Zhang and Ling Wang
- 85 Multiparameter Optimization of Trypanocidal Cruzain Inhibitors With In Vivo Activity and Favorable Pharmacokinetics**
Ivani Pauli, Celso de O. Rezende Jr., Brian W. Slafer, Marco A. Dessoy, Mariana L. de Souza, Leonardo L. G. Ferreira, Abraham L. M. Adjanohun, Rafaela S. Ferreira, Luma G. Magalhães, Renata Krogh, Simone Michelin-Duarte, Ricardo Vaz Del Pintor, Fernando B. R. da Silva, Fabio C. Cruz, Luiz C. Dias and Adriano D. Andricopulo
- 106 ER/AR Multi-Conformational Docking Server: A Tool for Discovering and Studying Estrogen and Androgen Receptor Modulators**
Feng Wang, Shuai Hu, De-Qing Ma, Qiuye Li, Hong-Cheng Li, Jia-Yi Liang, Shan Chang and Ren Kong
- 113 Introduction to the BioChemical Library (BCL): An Application-Based Open-Source Toolkit for Integrated Cheminformatics and Machine Learning in Computer-Aided Drug Discovery**
Benjamin P. Brown, Oanh Vu, Alexander R. Geanes, Sandeepkumar Kothiwale, Mariusz Butkiewicz, Edward W. Lowe Jr, Ralf Mueller, Richard Pape, Jeffrey Mendenhall and Jens Meiler

- 143 Combined Machine Learning and GRID-Independent Molecular Descriptor (GRIND) Models to Probe the Activity Profiles of 5-Lipoxygenase Activating Protein Inhibitors**
Hafiza Aliza Khan and Ishrat Jabeen
- 158 The Impact of the Secondary Binding Pocket on the Pharmacology of Class A GPCRs**
Attila Egyed, Dóra Judit Kiss and György M. Keserű
- 185 Limits of Prediction for Machine Learning in Drug Discovery**
Modest von Korff and Thomas Sander
- 195 From Data to Knowledge: Systematic Review of Tools for Automatic Analysis of Molecular Dynamics Output**
Hanna Baltrukevich and Sabina Podlowska
- 211 Pocket2Drug: An Encoder-Decoder Deep Neural Network for the Target-Based Drug Design**
Wentao Shi, Manali Singha, Gopal Srivastava, Limeng Pu, J. Ramanujam and Michal Brylinski
- 223 A Set of Experimentally Validated Decoys for the Human CC Chemokine Receptor 7 (CCR7) Obtained by Virtual Screening**
Matic Proj, Steven De Jonghe, Tom Van Loy, Marko Jukič, Anže Meden, Luka Ciber, Črtomir Podlipnik, Uroš Grošelj, Janez Konc, Dominique Schols and Stanislav Gobec
- 236 A New Strategy for Multitarget Drug Discovery/Repositioning Through the Identification of Similar 3D Amino Acid Patterns Among Proteins Structures: The Case of Tafluprost and its Effects on Cardiac Ion Channels**
Alejandro Valdés-Jiménez, Daniel Jiménez-González, Aytug K. Kiper, Susanne Rinné, Niels Decher, Wendy González, Miguel Reyes-Parada and Gabriel Núñez-Vivanco
- 247 Identification of Potent and Selective JAK1 Lead Compounds Through Ligand-Based Drug Design Approaches**
Sathya Babu, Santhosh Kumar Nagarajan, Sruthy Sathish, Vir Singh Negi, Honglae Sohn and Thirumurthy Madhavan
- 271 Drugsniffer: An Open Source Workflow for Virtually Screening Billions of Molecules for Binding Affinity to Protein Targets**
Vishwesh Venkatraman, Thomas H. Colligan, George T. Lesica, Daniel R. Olson, Jeremiah Gaiser, Conner J. Copeland, Travis J. Wheeler and Amitava Roy
- 282 Machine Learning in Antibacterial Drug Design**
Marko Jukič and Urban Bren
- 293 Structure-Based Design of 2-Aminopurine Derivatives as CDK2 Inhibitors for Triple-Negative Breast Cancer**
Hanzhi Liang, Yue Zhu, Zhiyuan Zhao, Jintong Du, Xinying Yang, Hao Fang and Xuben Hou
- 308 Probabilistic Pocket Druggability Prediction via One-Class Learning**
Riccardo Aguti, Erika Gardini, Martina Bertazzo, Sergio Decherchi and Andrea Cavalli



Editorial: Chemoinformatics Approaches to Structure- and Ligand-Based Drug Design, Volume II

Leonardo L. G. Ferreira* and Adriano D. Andricopulo*

Laboratory of Medicinal and Computational Chemistry, Center for Research and Innovation in Biodiversity and Drug Discovery, Physics Institute of Sao Carlos, University of Sao Paulo, Sao Carlos, Brazil

Keywords: drug design, machine learning, drug discovery, molecular docking, virtual screening, medicinal chemistry, QSAR

Editorial on the Research Topic

Chemoinformatics Approaches to Structure- and Ligand-Based Drug Design, Volume II

"Chemoinformatics Approaches to Structure- and Ligand-Based Drug Design, Volume II" follows the success of the first volume of this Research Topic (RT) (Ferreira and Andricopulo, 2018). The field has been more relevant than ever, especially in pandemic times, when Covid-19 hit the world and the scientific community was urged to come up with fast and cost-effective solutions (Robinson et al., 2022). Apart from the pandemic, chemoinformatics has been a core component in outstanding developments across different therapeutic areas and will continue to be a strategic innovation driver in the drug research and development (R&D) process (Chen et al., 2018; Ferreira and Andricopulo, 2019; Jiménez-Luna et al., 2021).

The second volume of this RT contains reviews and original research articles covering up-to-date research on machine learning (ML), multiparameter optimization (MPO), quantitative structure-activity relationships (QSAR), chemoinformatics servers, virtual screening, pharmacokinetics, among other equally relevant topics. More than 140 authors from all over the world contributed to the 20 articles that are part of this volume. Chemoinformatics investigations applied to different conditions such as Covid-19, cancer, Chagas disease, inflammation, pain, and immunological diseases are included. Additionally, novel approaches to pocket druggability analysis, multi-target drug discovery, artificial neural networks, multi-conformation molecular docking, molecular dynamics, and quantum studies are provided. Regarding target-based efforts, key aspects of intermolecular recognition are reported for a variety of proteins, including cruzain, G-protein coupled receptors (GPCR), phosphoglycerate mutase 1 (PGAM1), glutamate receptor, 5-lipoxygenase-activating protein (FLAP), Janus kinase 1 (JAK1), CC chemokine receptor 7 (CCR7), and cyclin-dependent kinase 2 (CDK2).

An MPO campaign combining computational and experimental approaches yielded a series of novel cruzain inhibitors (Pauli et al.). These compounds showed *in vitro* and *in vivo* trypanocidal activity along with low toxicity and suitable pharmacokinetics, contributing to the advance of Chagas disease drug discovery. Another study that integrated organic synthesis, biological evaluation, and molecular modeling (Oliveira et al.) resulted in the discovery of a series of carvacrol-derived sulfonamides with potent antioxidant, antinociceptive, and anti-edematogenic activities. Moreover, 3D-QSAR models (Wang et al.) were integrated with molecular docking and molecular dynamics to investigate anthraquinone-based PGAM1 inhibitors. Molecular modeling was also applied in combination with virtual screening and molecular docking to investigate novel inhibitors of JAK1 (Babu et al.), a critical enzyme for intracellular signal transduction and the development of numerous types of cancer. Given the importance of GPCRs in drug design, a review article covers

OPEN ACCESS

Edited and reviewed by:

Heike Wulff,
University of California, Davis,
United States

*Correspondence:

Leonardo L. G. Ferreira
leonardo@ifsc.usp.br
Adriano D. Andricopulo
aandrico@ifsc.usp.br

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 16 May 2022

Accepted: 14 June 2022

Published: 29 June 2022

Citation:

Ferreira LLG and Andricopulo AD
(2022) Editorial: Chemoinformatics
Approaches to Structure- and Ligand-
Based Drug Design, Volume II.
Front. Pharmacol. 13:945747.
doi: 10.3389/fphar.2022.945747

recent discoveries on allosteric GPCR ligands and bitopic modulators (Egyed et al.). The role of the GPCR secondary site was examined in terms of its effects on properties such as binding affinity, selectivity, and kinetics. A further review article examines the currently available tools used to analyze molecular dynamics results (Baltrukevich and Podlowska).

Chemokines play a critical role in immunological signaling and, therefore, can be explored as drug targets in different diseases such as immunological and inflammatory conditions and cancer (Salem et al., 2021). A set of experimentally validated decoys were identified for CCR7 using a structure-based virtual screening approach (Proj et al.). In addition to the traditional single-target drug design paradigm, a new multitarget strategy is reported in this RT (Valdés-Jiménez et al.). A computational tool was developed to explore and identify druggable 3D arrangements across different proteins. This algorithm allows the comparison of quaternary structures and the evaluation of druggability from the 3D structural pattern. However, defining druggable and non-druggable protein cavities is neither a trivial nor an obvious task (Ehrt et al., 2019). Departing from the commonly used two-class classification models, a one-class approach to assess druggability (Aguti et al.) using a probabilistic kernel is communicated in this RT. The workflow proved to be feasible in removing or reducing biases in the classification of druggable pockets. Virtual screening has become an important tool in drug discovery as it allows a preliminary evaluation of large compound collections in short timelines and costs (Ferreira and Andricopulo, 2021). A novel virtual screening workflow (Venkatraman et al.) that can sample billions of compounds and supports parallel and cloud computing is reported. A collection of approximately 3.7 billion compounds against three Sars-CoV-2 proteins were used to evaluate the effectiveness of the new virtual screening pipeline. Another target-based study focuses on CDK2, which participates in the regulation of the cell cycle and is a critical player in cancer emergence. A series of aminopurine derivatives was designed as novel CDK2 inhibitors (Liang et al.) with high selectivity concerning other CDK isoforms. Anti-proliferative activity against triple-negative breast cancer cells (TNBC) was shown, which makes this series suitable starting points for optimization.

ML has been a hot topic in drug discovery, which is reflected in the number of articles on this theme published in this RT. Novel molecular targets for Covid-19 drug repositioning were identified (López-Cortés et al.) by a combination of artificial neural

networks, single-cell RNA sequencing, and interactome analyses of the immunological system proteins. After a screen of more than 1,500 proteins, 25 putative molecular targets were identified. Interestingly, datasets containing more than 50,000 structurally diverse compounds with reported activity against several breast cancer cell lines were used to generate predictive models (He et al.). As a result, a web server was created to predict the activity of query compounds against breast cancer cell lines. Another online tool reported in this RT performs multi-conformational molecular docking (Wang et al.) on estrogen (ER α and ER β) and androgen (AR) receptors. In addition, this interface runs 2D similarity searches against a database of known ER α , ER β , and AR ligands. ML was also applied to identify the 2D features associated with the anti-inflammatory properties of FLAP inhibitors (Aliza Khan and Jabeen), which can assist the design of optimized anti-inflammatory agents. Furthermore, this RT brings to the readers an interesting analysis of the extrapolation limits of different regression methods (von Korff and Sander) applied to drug discovery along with an ML-based QSAR model (Brown et al.) for the estimation of molecular properties in drug design. In the field of deep learning, this article Research Topic features a report of a deep graph neural network (Shi et al.) to predict the interaction of small-molecule compounds with protein binding cavities. An additional important topic is drug resistance to antibiotics, which has emerged as a major health concern all over the world. ML has been applied to the field to identify novel chemical matter able to circumvent the main resistance mechanisms found in bacteria (Chowdhury et al., 2020). A review article examines recent machine learning studies (Jukič and Bren) applied to the identification of novel non-peptidic and peptidic antibacterial compounds and drug targets.

This RT encloses articles that cover a broad range of chemoinformatics applications to drug discovery and its many interfaces with the chemical and biological sciences. The knowledge shared through this RT could not be more relevant and timely. We hope that the findings, insights, and analyses reported herein contribute to the advance of drug discovery and, ultimately, to the promotion of human health.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

- Chen, H., Kogej, T., and Engkvist, O. (2018). Cheminformatics in Drug Discovery, an Industrial Perspective. *Mol. Inf.* 37, e1800041. doi:10.1002/minf.201800041
- Chowdhury, A. S., Call, D. R., and Broschat, S. L. (2020). PARGT: a Software Tool for Predicting Antimicrobial Resistance in Bacteria. *Sci. Rep.* 10, 11033. doi:10.1038/s41598-020-67949-9
- Ehrt, C., Brinkjost, T., and Koch, O. (2019). Binding Site Characterization - Similarity, Promiscuity, and Druggability. *Medchemcomm* 10, 1145–1159. doi:10.1039/c9md00102f
- Ferreira, L. L. G., and Andricopulo, A. D. (2019). ADMET Modeling Approaches in Drug Discovery. *Drug Discov. Today* 24, 1157–1165. doi:10.1016/j.drudis.2019.03.015
- Ferreira, L. L. G., and Andricopulo, A. D. (2018). Editorial: Chemoinformatics Approaches to Structure- and Ligand-Based Drug Design. *Front. Pharmacol.* 9, 1416. doi:10.3389/fphar.2018.01416
- Ferreira, L. L. G., and Andricopulo, A. D. (2021). "Structure-Based Drug Design," in *Burger's Medicinal Chemistry, Drug Discovery and Development*. Editors D. J. Abraham and M. Myers. 8th edition (John Wiley & Sons), 1–54. doi:10.1002/0471266949.bmc141.pub2

- Jiménez-Luna, J., Grisoni, F., Weskamp, N., and Schneider, G. (2021). Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opin. Drug Discov.* 16, 949–959. doi:10.1080/17460441.2021.1909567
- Robinson, P. C., Liew, D. F. L., Tanner, H. L., Grainger, J. R., Dwek, R. A., Reisler, R. B., et al. (2022). COVID-19 Therapeutics: Challenges and Directions for the Future. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2119893119. doi:10.1073/pnas.2119893119
- Salem, A., Alotaibi, M., Mroueh, R., Basheer, H. A., and Afarinkia, K. (2021). CCR7 as a Therapeutic Target in Cancer. *Biochim. Biophys. Acta Rev. Cancer.* 1875, 188499. doi:10.1016/j.bbcan.2020.188499

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ferreira and Andricopulo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

Edited by:

Roberto Paganelli,
University of Studies G. d'Annunzio
Chieti and Pescara, Italy

Reviewed by:

Pascal Falter-Braun,
Helmholtz-Gemeinschaft Deutscher
Forschungszentren (HZ), Germany
Chandra C. Ghosh,
Harvard Medical School,
United States

*Correspondence:

Andrés López-Cortés
aalc84@gmail.com
Luis Abel Quiñones
lquinone@med.uchile.cl

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Translational Pharmacology,
a section of the journal
Frontiers in Pharmacology

Received: 27 August 2020

Accepted: 11 January 2021

Published: 26 February 2021

Citation:

López-Cortés A, Guevara-Ramírez P,
Kyriakidis NC, Barba-Ostria C,
León Cáceres Á, Guerrero S,
Ortiz-Prado E, Munteanu CR, Tejera E,
Cevallos-Robalino D,
Gómez-Jaramillo AM,
Simbaña-Rivera K,
Granizo-Martínez A, Pérez-M G,
Moreno S, García-Cárdenas JM,
Zambrano AK, Pérez-Castillo Y,
Cabrera-Andrade A,
Puig San Andrés L, Proaño-Castro C,
Bautista J, Quevedo A, Varela N,
Quiñones LA and Paz-y-Miño C (2021)
In silico Analyses of Immune System
Protein Interactome Network, Single-
Cell RNA Sequencing of Human
Tissues, and Artificial Neural Networks
Reveal Potential Therapeutic Targets
for Drug Repurposing
Against COVID-19.
Front. Pharmacol. 12:598925.
doi: 10.3389/fphar.2021.598925

In silico Analyses of Immune System Protein Interactome Network, Single-Cell RNA Sequencing of Human Tissues, and Artificial Neural Networks Reveal Potential Therapeutic Targets for Drug Repurposing Against COVID-19

Andrés López-Cortés^{1,2,3†*}, Patricia Guevara-Ramírez^{1†}, Nikolaos C. Kyriakidis^{4†}, Carlos Barba-Ostria^{4†}, Ángela León Cáceres^{5,6,7}, Santiago Guerrero¹, Esteban Ortiz-Prado⁴, Cristian R. Munteanu^{2,8,9}, Eduardo Tejera¹⁰, Doménica Cevallos-Robalino¹¹, Ana María Gómez-Jaramillo¹², Katherine Simbaña-Rivera⁴, Adriana Granizo-Martínez¹³, Gabriela Pérez-M¹⁴, Silvana Moreno¹⁵, Jennyfer M. García-Cárdenas¹, Ana Karina Zambrano^{1,8}, Yunierkis Pérez-Castillo¹⁰, Alejandro Cabrera-Andrade^{2,10}, Lourdes Puig San Andrés¹, Carolina Proaño-Castro¹⁶, Jhommara Bautista¹⁷, Andreina Quevedo¹, Nelson Varela^{3,18}, Luis Abel Quiñones^{3,18*} and César Paz-y-Miño¹

¹Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Quito, Ecuador, ²RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, A Coruña, Spain, ³Latin American Network for the Implementation and Validation of Clinical Pharmacogenomics Guidelines (RELIVAF-CYTED), Madrid, Spain, ⁴One Health Research Group, Faculty of Medicine, Universidad de Las Américas (UDLA), Quito, Ecuador, ⁵Heidelberg Institute of Global Health, Faculty of Medicine, Heidelberg University, Heidelberg, Germany, ⁶Instituto de Salud Pública, Facultad de Medicina, Pontificia Universidad Católica del Ecuador, Quito, Ecuador, ⁷Tropical Herping, Quito, Ecuador, ⁸Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), A Coruña, Spain, ⁹Centro de Información en Tecnologías de la Información y las Comunicaciones (CITIC), A Coruña, Spain, ¹⁰Grupo de Bio-Quimiinformática, Universidad de Las Américas (UDLA), Quito, Ecuador, ¹¹Hospital General del Sur de Quito, Instituto Ecuatoriano de Seguridad Social, Quito, Ecuador, ¹²Faculty of Medicine, Pontifical Catholic University of Ecuador, Quito, Ecuador, ¹³Carrera de Medicina, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Quito, Ecuador, ¹⁴Centro Clínico Quirúrgico Ambulatorio Hospital del Día El Batán, Instituto Ecuatoriano de Seguridad Social, Quito, Ecuador, ¹⁵Department of Plant Biology, Faculty of Natural Resources and Agricultural Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden, ¹⁶Fundación Futuro, Quito, Ecuador, ¹⁷Facultad de Ingeniería y Ciencias Aplicadas-Biotecnología, Universidad de Las Américas, Quito, Ecuador, ¹⁸Laboratory of Chemical Carcinogenesis and Pharmacogenetics, Department of Basic-Clinical Oncology, Faculty of Medicine, University of Chile, Santiago, Chile

Background: There is pressing urgency to identify therapeutic targets and drugs that allow treating COVID-19 patients effectively.

Methods: We performed *in silico* analyses of immune system protein interactome network, single-cell RNA sequencing of human tissues, and artificial neural networks to reveal potential therapeutic targets for drug repurposing against COVID-19.

Results: We screened 1,584 high-confidence immune system proteins in ACE2 and TMPRSS2 co-expressing cells, finding 25 potential therapeutic targets significantly overexpressed in nasal goblet secretory cells, lung type II pneumocytes, and ileal absorptive enterocytes of patients with several immunopathologies. Then, we

performed fully connected deep neural networks to find the best multitask classification model to predict the activity of 10,672 drugs, obtaining several approved drugs, compounds under investigation, and experimental compounds with the highest area under the receiver operating characteristics.

Conclusion: After being effectively analyzed in clinical trials, these drugs can be considered for treatment of severe COVID-19 patients. Scripts can be downloaded at <https://github.com/muntisa/immuno-drug-repurposing-COVID-19>.

Keywords: COVID-19, immune system, single-cell RNA sequencing, artificial neural networks, drug repurposing

INTRODUCTION

The first zoonotic transmission of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was located in China in December 2019 (Tay et al., 2020), and it is the causative agent of the coronavirus disease 2019 (COVID-19) (Sanders et al., 2020). The World Health Organization (WHO) declared the outbreak of COVID-19 as a Public Health Emergency of International Concern on January 30, 2020, and a pandemic on March 11, 2020 (Gao Q et al., 2020). Classified in the *Coronaviridae* family and *Betacoronavirus* genus, SARS-CoV-2 is the seventh CoV known to infect humans, along with 229E, NL63, OC43, HKU1, SARS-CoV, and Middle East respiratory syndrome (MERS) (Oberfeld et al., 2020). Coronaviruses cause mild to severe respiratory diseases and have high mutation rates that result in high genetic diversity, plasticity, and adaptability to invade a wide range of hosts (Peiris et al., 2004).

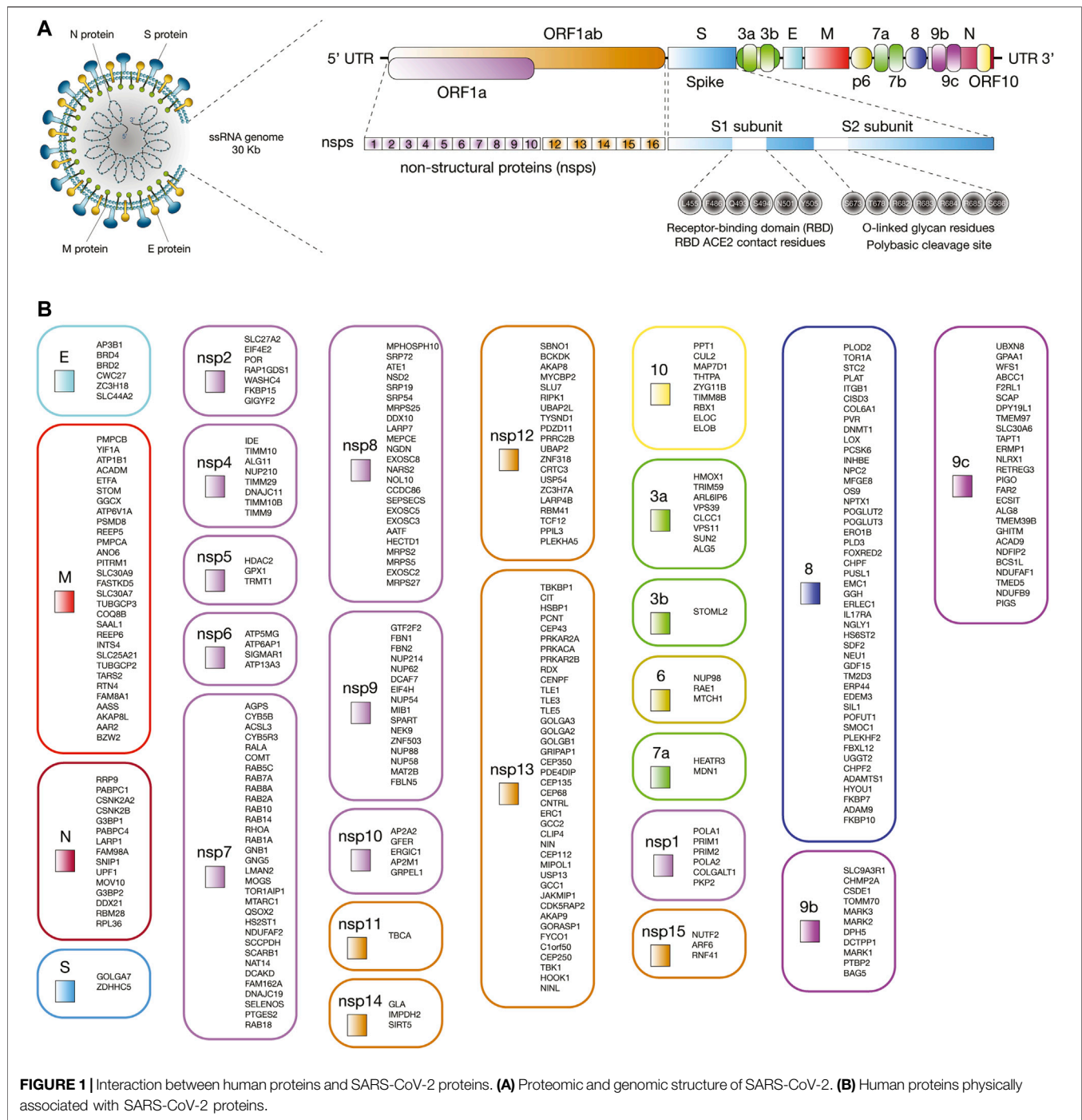
The first genome of SARS-CoV-2 named Wuhan-Hu-1 (NCBI reference sequence NC_045512) was isolated and sequenced in China in January 2020 (Zhou P et al., 2020; Zhu et al., 2020). SARS-CoV-2 is a single-stranded positive-sense RNA virus of about 30 kb in length (Zhou P et al., 2020; Ziegler et al., 2020). The genomic structure is comprised of a 5' terminal cap structure, 14 open reading frames (ORFs) encoding 29 proteins, and a 3' poly A tail (Wu A et al., 2020). ORF1a and ORF1ab are the largest genes and codify 16 non-structural proteins (nsp1 to nsp16). According to Gordon et al. (2020), nsps are involved in antiviral response (nsp1), viral replication (the nsp3-nsp4-nsp6 complex), the protease 3C^{pro} (nsp5) (Zhang L et al., 2020), the RNA polymerase (the nsp7-nsp8 complex), the single-strand RNA binding (nsp9), the methyltransferase activity (nsp10 and nsp16), the RNA-dependent RNA polymerase (nsp12) (Gao Y et al., 2020), the helicase/triphosphatase (nsp13), the 3'-5' exonuclease (nsp14), the uridine-specific endoribonuclease (nsp15), and the RNA-cap methyltransferase (nsp16) (Gordon et al., 2020). Lastly, the 3' terminus contains genes that codify the spike (S) glycoprotein, the envelope (E) protein, the membrane (M) glycoprotein, the nucleocapsid (N) protein, and several accessory proteins (3a, 3b, p6, 7a, 7b, 8, 9b, 9c, and 10) (Figure 1A) (Wu A et al., 2020; Wu C et al., 2020).

COVID-19 is caused when SARS-CoV-2 exploits the host cell machinery for its own replication and spread (Ortiz-Prado et al., 2020). SARS-CoV-2 entry into human cells is mediated by the S glycoprotein that forms homotrimers protruding from the viral surface (Walls et al., 2020). S1 and S2 are two functional subunits

of the S glycoprotein. Six receptor-binding domain (RBD) amino acids (L455, F486, Q493, S494, N501, and Y505) of the S1 subunit directly bind to the peptide domain of angiotensin-converting enzyme 2 (ACE2) human receptor protein (Andersen et al., 2020; Cao et al., 2020; Wang Q et al., 2020; Yan et al., 2020). The affinity constant for RBD of SARS-CoV-2 to ACE2 is greater than that of SARS-CoV by as much as a factor of 10–15 (Wang Q et al., 2020; Wang Y et al., 2020; Wrapp et al., 2020). S glycoprotein is cleaved by the cathepsin L (CTSL) protease (Muus et al., 2020), and the transmembrane serine protease (TMPRSS2) in a functional polybasic (furin) cleavage site at the S1-S2 boundary flanked for O-linked glycans (Hoffmann et al., 2020; Walls et al., 2020). S2 subunit mediates subsequent fusion between the human and viral membranes (Kirchdoerfer et al., 2016; Yuan et al., 2017).

ACE2 is a type I membrane protein widely expressed in nasal goblet secretory cells, lung type II pneumocytes, ileal absorptive enterocytes, kidney proximal tubule cells, gallbladder basal cells, among other human cells (Deng et al., 2020; Lamers et al., 2020; Singh et al., 2020; Sungnak et al., 2020; Ziegler et al., 2020), and participates in the maturation of angiotensin, a peptide hormone that controls blood pressure and vasoconstriction (Donoghue et al., 2000). After virus entry, many severe ill COVID-19 patients developed clinical manifestations such as cough, mild fever, dyspnea, lung edema, severe hypoxemia, acute respiratory distress syndrome (ARDS) (Montenegro et al., 2020), acute lung injury (Blanco-Melo et al., 2020), interstitial pneumonia, increased concentrations of fibrinogen and D-dimer plasma levels (Spiezia et al., 2020; Tang et al., 2020), elevated levels of pro-inflammatory chemokines and cytokines such as interleukin (IL) 6 (Herold et al., 2020; Sarzi-Puttini et al., 2020), low levels of type I and III interferons (IFNs) (Blanco-Melo et al., 2020), high levels of lactate dehydrogenase, hyperferritinemia, idiopathic thrombocytopenic purpura caused by spleen atrophy (Zulfiqar et al., 2020), formation of hyaline membrane (Yao et al., 2020), hilar lymph node necrosis, lymphopenia (Terpos et al., 2020), intravascular coagulopathy (Fogarty et al., 2020), pulmonary thromboembolism (Rotzinger et al., 2020), hypotension (Rentsch et al., 2020), cerebrovascular events (Mao et al., 2020), severe metabolic acidosis, kidney and hepatic dysfunctions (Zhang C et al., 2020), secondary infections, septic shock (Li H et al., 2020), and multi-organ failure (Wang Q et al., 2020; Gupta et al., 2020; Wadman et al., 2020).

Additionally, SARS-CoV-2 interacts with the immune system triggering dysfunctional immune responses to



COVID-19 progression (Tay et al., 2020). Given that an excessive inflammatory response to the novel coronavirus is thought to be a major cause of disease severity and death (Blanco-Melo et al., 2020; Mehta et al., 2020), a better understanding of the immunological underpinnings is required to identify potential therapeutic targets. To fill in this gap, we performed *in silico* analyses of immune system protein-protein interactome (PPI) network, single-cell RNA sequencing (scRNA-seq) of human tissues, and artificial neural

networks to reveal potential therapeutic targets for drug repurposing against COVID-19.

METHODS

Protein Sets

We have retrieved the 332 human proteins physically associated with 26 of the 29 SARS-CoV-2 proteins proposed by Gordon *et al*

(Figure 1B; Supplementary Table S1; Gordon et al., 2020). We have also retrieved a total of 3,885 immune system proteins from several databases such as the International ImmunoGeneTics information system (<http://www.imgt.org>) (Giudicelli et al., 2005; Lefranc et al., 2009; Lefranc et al., 2015), the InnateDB database (<https://www.innatedb.com/>) (Breuer et al., 2013), and the David Bioinformatics Resource (<https://david.ncifcrf.gov/>) (Huang et al., 2009b; Huang et al., 2009a) using the gene ontology (GO) terms: 0002376 immune system process, 0045087 innate immune response, and 0002250 adaptive immune response. Lastly, both protein sets were integrated to identify the highest confidence interactions and to design the immune system PPI network.

Protein-Protein Interactome Network

The immune system PPI network with a highest confidence cutoff of 0.9 and zero node addition was created between the human proteins physically associated with SARS-CoV-2 and their first neighboring proteins of the immune system. This network was generated using the human proteome of the Cytoscape StringApp (Szkarczyk et al., 2015; Doncheva et al., 2019), which imports protein-protein interaction data from the STRING database (Szkarczyk et al., 2015). The degree centrality represents the number of edges the node has in a network (López-Cortés et al., 2018; López-Cortés et al., 2020b), and it was calculated using the CytoNCA app (Tang et al., 2015). All nodes and edges were organized through the organic layout, which produces clear representations of complex networks, and lastly, the immune system PPI network was visualized through the Cytoscape software v.3.7.1 (Shannon et al., 2003).

Interestingly, Overmyer *et al.* published a large-scale multi-omic analysis identifying 146 significantly expressed proteins in patients with severe COVID-19 (Overmyer et al., 2020). We located these proteins in our immune system PPI network and generated the immune system PPI subnetwork encompassing the significantly expressed proteins in severe COVID-19 and their first neighbor nodes (cutoff = 0.9). Subsequently, we ranked the overexpressed and underexpressed proteins according to the highest degree centrality.

Additionally, Bouhaddou *et al.* published the global phosphorylation landscape of SARS-CoV-2 infection identified 97 significantly expressed proteins in Vero E6 cells (Bouhaddou et al., 2020). We located these proteins in both networks and ranked the phosphorylated proteins according to the highest degree centrality. Lastly, human proteins physically associated with the SARS-CoV-2 proteins, immune system proteins, significantly expressed proteins in severe COVID-19, and significantly expressed phosphorylated proteins in SARS-CoV-2 infection in Vero E6 cells were differentiated by colors in both the immune system PPI network and subnetwork.

Functional Enrichment Analysis

The functional enrichment analysis gives curated signatures of protein sets generated from omics-scale experiments (Reimand et al., 2019). We performed the enrichment analysis to validate the correlation between the immune system PPI subnetwork and biological annotations related to severe COVID-19, using the protein set of the immune system PPI network as background set.

The enrichment was calculated using g:Profiler version e101_eg48_p14_baf17f0 (<https://biit.cs.ut.ee/gprofiler/gost>) to obtain significant annotations (Benjamini-Hochberg false discovery rate - FDR < 0.001) related to GO: biological processes, the Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathways, and Reactome signaling pathways (Wang et al., 2016; Slenter et al., 2018; Raudvere et al., 2019; Jassal et al., 2020). Lastly, the enrichment analysis was visualized in a Manhattan plot, and the significant terms related to the immunopathology of severe COVID-19 were manually curated.

Single-Cell RNA Sequencing Data

Ziegler *et al.* analyzed human scRNA-seq data to uncover potential targets of SARS-CoV-2 amongst tissue-resident cell subsets. They discovered *ACE2* and *TMPRSS2* co-expressing in goblet secretory cells from nasal passages, type II pneumocytes from lung epithelial cells, and absorptive enterocytes from ileal epithelial cells (Ziegler et al., 2020).

After constructing the immune system PPI network between the human proteins physically associated with the SARS-CoV-2 proteins, immune system proteins, and significantly expressed proteins in severe COVID-19, we compared the transcriptomics data of the network nodes between 10 nasal passage cells (goblet cell, basal cell of olfactory epithelium, ciliated cell, endothelial cell, fibroblast cell, glandular epithelial cell, mast cell, myeloid cell, plasma cell, and T cell), 15 lung epithelial cells (ciliated cell, lymphatic cell, fibroblast 1, fibroblast 2, macrophage 1, macrophage 2, macrophage 3, mast cell, monocytes 1, monocytes 2, neutrophil cell, proliferating cell, T cell, type I pneumocytes, and type II pneumocytes), and 9 ileal epithelial cells (cycling stem cell, early enterocyte 1, early enterocyte 2, absorptive enterocyte, enteroendocrine cell, goblet cell, quiescent stem cell, TA G1S cell, and TA G2M cell) to identify significantly expressed genes in goblet secretory cells, type II pneumocytes, and absorptive enterocytes.

The transcriptomics data was taken from the 'COVID-19 Studies' section of the Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/covid19), and the Alexandria Project (<https://alexandria-scrna-data-library.readthedocs.io/en/latest/introduction.html>). The three single-cell databases analyzed were: 1) nasal passage cells (Ordovas-Montanes et al., 2018) (https://singlecell.broadinstitute.org/single_cell/study/SCP253/allergic-inflammatory-memory-in-human-respiratory-epithelial-progenitor-cells#study-visualize), 2) lung epithelial cells (Ziegler et al., 2020) (https://singlecell.broadinstitute.org/single_cell/study/SCP814/human-lung-hiv-tb-co-infection-ace2-cells#study-visualize), and 3) ileal epithelial cells (Fujii et al., 2018) (https://singlecell.broadinstitute.org/single_cell/study/SCP817/comparison-of-ace2-and-tmprss2-expression-in-human-duodenal-and-ileal-tissue-and-organoid-derived-epithelial-cells#study-visualize). Lastly, it is important to clarify that the scRNA-seq analyses were done in cells non exposed to the novel coronavirus.

The criteria of analysis of transcriptomics data of nasal passage cells, lung epithelial cells, and ileal epithelial cells was the following: 't-distributed stochastic neighbor embedding (t-SNE) cell types' as load cluster, 'cell type ontology label' as

selected annotation, and ‘all cells’ as subsampling threshold. Additionally, we adjust the mRNA expression taking into account the Z-scores, that is, overexpressed mRNAs with Z-scores > 2 and underexpressed mRNAs with Z-scores < -2. Regarding visualization of transcriptomics data, we designed heatmaps to compare the expression between cell types, dot plots to visualize the percentage of cells expressing, box plots to compare the expression scores of multiple genes for each cell type taking into account the mean log normalized expression, and 2D t-SNE to visualize the expression score of significantly expressed multiple genes per subpopulation cell.

Drug Repurposing

After identifying the significantly expressed biological molecules present in the scRNA-seq analyses of *ACE2* and *TMPRSS2* co-expressing human cells, we evaluated the druggability of these molecules, and subsequently perform the drug repurposing analysis.

From all the 75 previously identified and significantly expressed biological molecules, only 31 had identification number in the ChEMBL database (<https://www.ebi.ac.uk/chembl>) (Gaulton et al., 2017), and from these 31 proteins, all compounds were extracted from ChEMBL as follow: 1) all reported interactions with (IC50, Ki, EC50, and GI50) where extracted from ChEMBL version 26; 2) all extracted interactions were labeled as active (1) or inactive (0) if values are less than 10 μ M; and 3) if more than one report (active or inactive) is available for the same compound-target interaction, the final criteria (active or inactive) was assigned considering the 75% of the information or rejected otherwise. From the 31 proteins, only 25 had identified molecules with active/inactive interactions after considering the previous filters. Hence, we identified 25 potential therapeutic targets for drug repurposing against COVID-19.

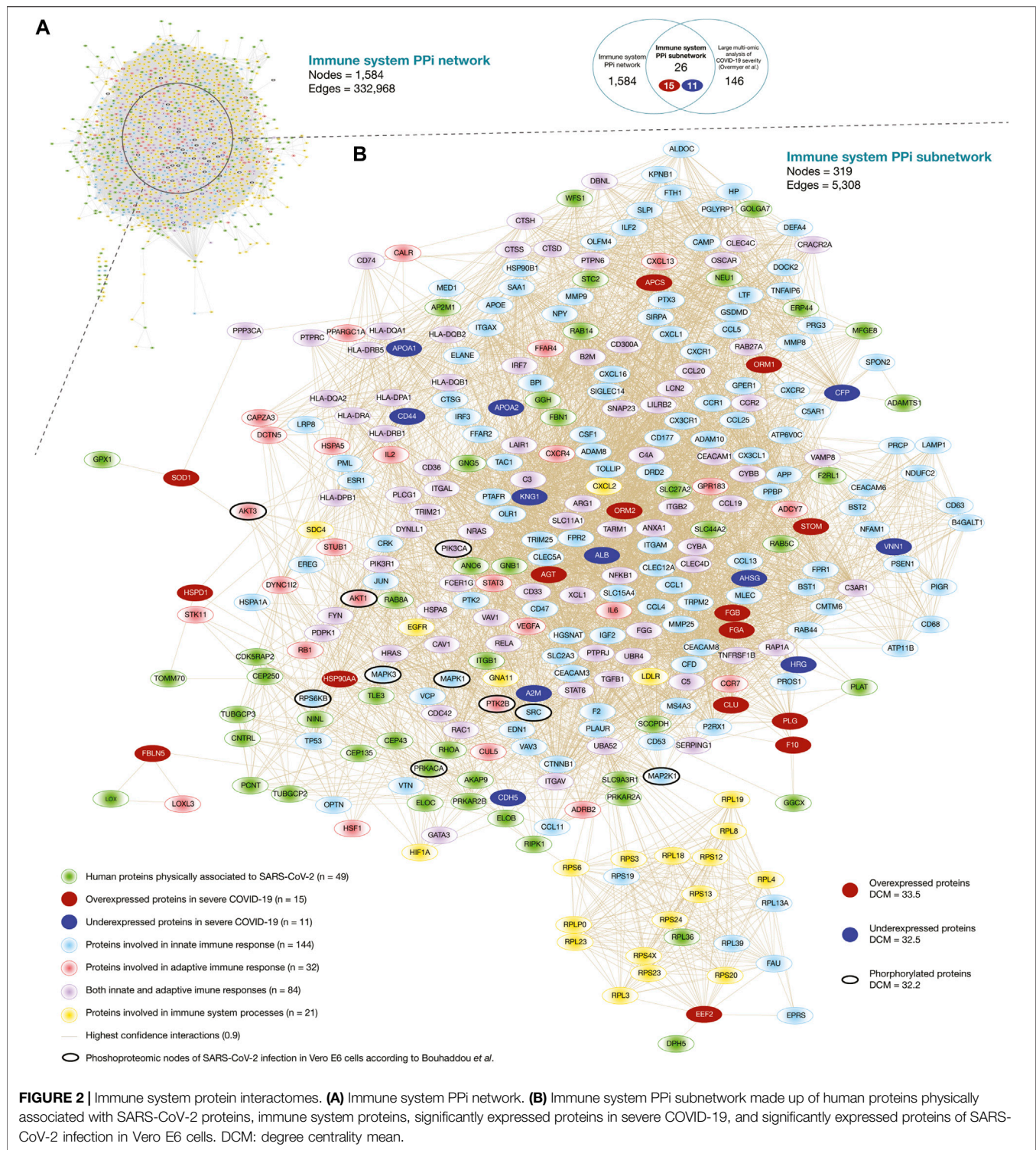
DeepChem package and Python Jupyter Notebooks (Oliver, 2013) were used to predict if drugs (DrugBank compounds) could be active for multiple protein targets (Oliver, 2013) (<https://github.com/deepchem/deepchem>). DrugBank (<https://www.drugbank.ca/>) contains comprehensive information about drugs, their mechanism of action, and their targets (Wishart et al., 2006; Wishart et al., 2018). The calculations used the GPU of Google Colab and the correspondent scripts could be found at GitHub repository: <https://github.com/muntisa/immuno-drug-repurposing-COVID-19>. The fully-connected deep neuronal networks (FCNNs) have been used to find the best multitask classification model using 1,024 molecular circular fingerprints (CFPs) as input descriptors for 15,377 ChEMBL compounds and activity (1/0) for the 25 therapeutic targets as outputs/tasks (Wu et al., 2018). The best model resulted from a grid search for the best parameters have been used to predict the activity of 10,672 drugs for the 25 targets. The performance of the classifiers used during the training, grid search and test evaluation of the best model was the area under curve (AUC) of the receiver operating characteristic (ROC) curve (AUROC) (Hastie et al., 2009), the default metric in DeepChem package. The ROC curve is defined by the True Positive Rate (TPR) (or Sensitivity) vs. the False Positive Rate (FPR) (or 1-Specificity) for each of the class of the multi-task classifier for different class probability thresholds.

$TPR = TP/(TP + FN)$, $FPR = FP/(FP + TN)$, where TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative (from the confusion matrix that summarizes the results of testing the classifier). AUROC represents the area under the ROC curve, with values between 0 and 1 (1 = perfect model; 0.5 = no skill/random model).

The main script of the repository (Immuno-Drug-Repurposing-DeepChem-MultitaskClassification.ipynb) is presenting all the methodology with python code and results. The repository folder “datasets” contains the dataset with the ChEMBL ID, SMILES formula, and the class of protein target (multiclass_origDS_noDB.csv). The dataset that will be used by the classifier contains the SMILES formulas of 15,377 ChEMBL compounds that interacts with 25 different protein targets with the following UniProt IDs: O00571, P00533, P01024, P01130, P04233, P07339, P08962, P09668, P11021, P15291, P16070, P17301, P21741, P25774, P25963, P26006, P27361, P35222, P40763, P50591, P55085, Q15904, Q16665, Q99519, and Q99814. This means that the dataset was composed by 15,377 examples with 25 classes. The multi-task classification model will be able to predict if a compound with a SMILE formula could have one or more protein targets simultaneously, using separated tasks/outputs for each of the 25 proteins. It is not a simple classification with only an output (class) that can predict only a protein value from the 25 possible targets). The prediction molecules that will be evaluated with the best classifier can be found in DB_toPredict.csv (DataBank ID, SMILES formulas, and the classes to predict). The input SMILES formulas will be used to calculate molecular descriptors for all molecules (as model inputs).

In the first step, CFPs molecular descriptor have been calculated for both ChEMBL dataset and DrugBank prediction set (Gaulton et al., 2017; Wishart et al., 2018) as a vector of 1,024 values for each compound. Thus, the dataset to build the future classifier has 1,024 input features in 15,377 examples with 25 output classes (protein target).

In order to build a classifier (model), the training of the model should be done with a training subset and the final model should be tested for performance with a test subset that was not used during the training process. In addition, if different classifiers with different parameters are used during the training, there is a need of an extra validation subset to decide the best classifier (model) using a specific metrics (in our scripts: AUROC). Thus, the dataset was splitted into 80%-10%-10% training-validation-test subsets using RandomStratifiedSplitter (to maintain the same ratio between the examples in all 25 classes as in the initial dataset). The training and validation subsets were used to find the best hyperparameters for the FCNN with 1,000 neurons (MultitaskClassifier from DeepChem package). The constant parameters are activation functions as relu, momentum of 0.9, weights initialization using Glorot uniform method (Xavier uniform initializer), learning rate of 1e-3, decay of 1e-6, 10 epochs, a single hidden layer (additional parameters could be found in the main notebook of the repository). During the grid search for the best model, 64 classifiers have been optimized with different combination of the following parameters: batch size = (128, 515), dropouts = (0.0, 0.1, 0.2, 0.3), batch normalization =



(False, True), and hidden layer sizes (number of neurons) = (100, 500, 1,000, 1,024). Thus, the training subset was used for training of each model/classifier and the validation subset was used to decide the best model.

The test set was used to verify the performance of the best model for each task/protein target (see **Supplementary**

Table S2). AUROC for the test subset was between 0.935 and 1.000 (mean AUROC = 0.989; standard deviation (SD) = 0.019). Additional results such as the AUROC values for training, validation and test subset for each protein target (task/class) are presented into the folder “results” as multitasks_metrics_best.csv.

The best model has 1,000 neurons in a hidden layer (dropout of 0.5) with all parameters as 'activation': 'relu', 'momentum': 0.9, 'batch_size': 124, 'init': 'glorot_uniform', 'data_shape': (1024), 'learning_rate': 0.001, 'decay': 1e-06, 'nb_epoch': 1, 'nesterov': False, 'dropouts': (0.5), 'nb_layers': 1, 'batchnorm': False, 'layer_sizes': (1000), 'weight_init_stddevs': (0.1), 'bias_init_consts': (1.0), 'penalty': 0.0. This classifier was used to predict the activity of 10,672 drugs from DataBank for the 25 immune system targets: DDX3X, EGFR, C3, LDLR, CD74, CTSD, CD63, CTSH, HSPA5, B4GALT1, CD44, ITGA2, MDK, CTSS, NFKBIA, ITGA3, MAPK3, CTNNB1, STAT3, TNFSF10, F2RL1, ATP6AP1, HIF1A, NEU1, and EPAS1 (see **Supplementary Table S7** and multitasks_predictions_best.csv in repository folder "results"). Lastly, the best predicted drug-target associations were evaluated according to its first ATC level (https://www.whocc.no/atc_ddd_index/), drug category, mechanism of action, approval status by the US Food & Drug Administration (FDA) or the European Medicines Agency (EMA), the pharmacological indications, and the current involvement in COVID-19 clinical trials (<https://www.clinicaltrials.gov/ct2/results?cond=COVID-19>).

RESULTS

Immune System Protein-Protein Interactome Network

In biological systems, specialized pathogens (i.e., SARS-CoV-2) employ a suite of virulent proteins, which interact with key targets in host interactomes to extensively rewire the flow of information and cause diseases, such as COVID-19 (Vidal et al., 2011; Pan et al., 2016; Kumar et al., 2020). The human proteins physically associated with SARS-CoV-2 are the first line of host proteins, which also interacts with molecular components involved in a wide spectrum of biological processes and signaling pathways within the cell. Therefore, analyzing the interactome of immune system proteins may reveal novel components in SARS-CoV-2 immunopathogenesis.

Here, we generated the immune system PPI network encompassing 1,584 nodes and 332,968 edges (**Figure 2A**). Of them, 256 human proteins physically associated with SARS-CoV-2 proteins had high-confidence interactions (cutoff = 0.9) with 1,390 immune system proteins belonging to the first neighbor nodes (**Supplementary Table S3**). The degree centrality mean of the human proteins physically associated with SARS-CoV-2 proteins was 23.6, and proteins with the highest degree centrality were GNB1, GNG5, RBX1, RHOA, and TCEB1. On the other hand, the degree centrality mean of the immune system protein was 44.5, and proteins with the highest degree centrality were UBA52, APP, FPR2, NCBP1, and NCBP2. Additionally, we have identified 40 significantly expressed phosphorylated proteins of SARS-CoV-2 infection according to the global phosphorylation landscape in Vero E6 cells published by Bouhaddou et al. (2020). The degree centrality mean of the phosphorylated proteins was 59.8, and proteins with the highest degree centrality were PIK3CA, MAPK1, MAPK3, SRC, and AKT1 (**Supplementary Table S4**). Lastly,

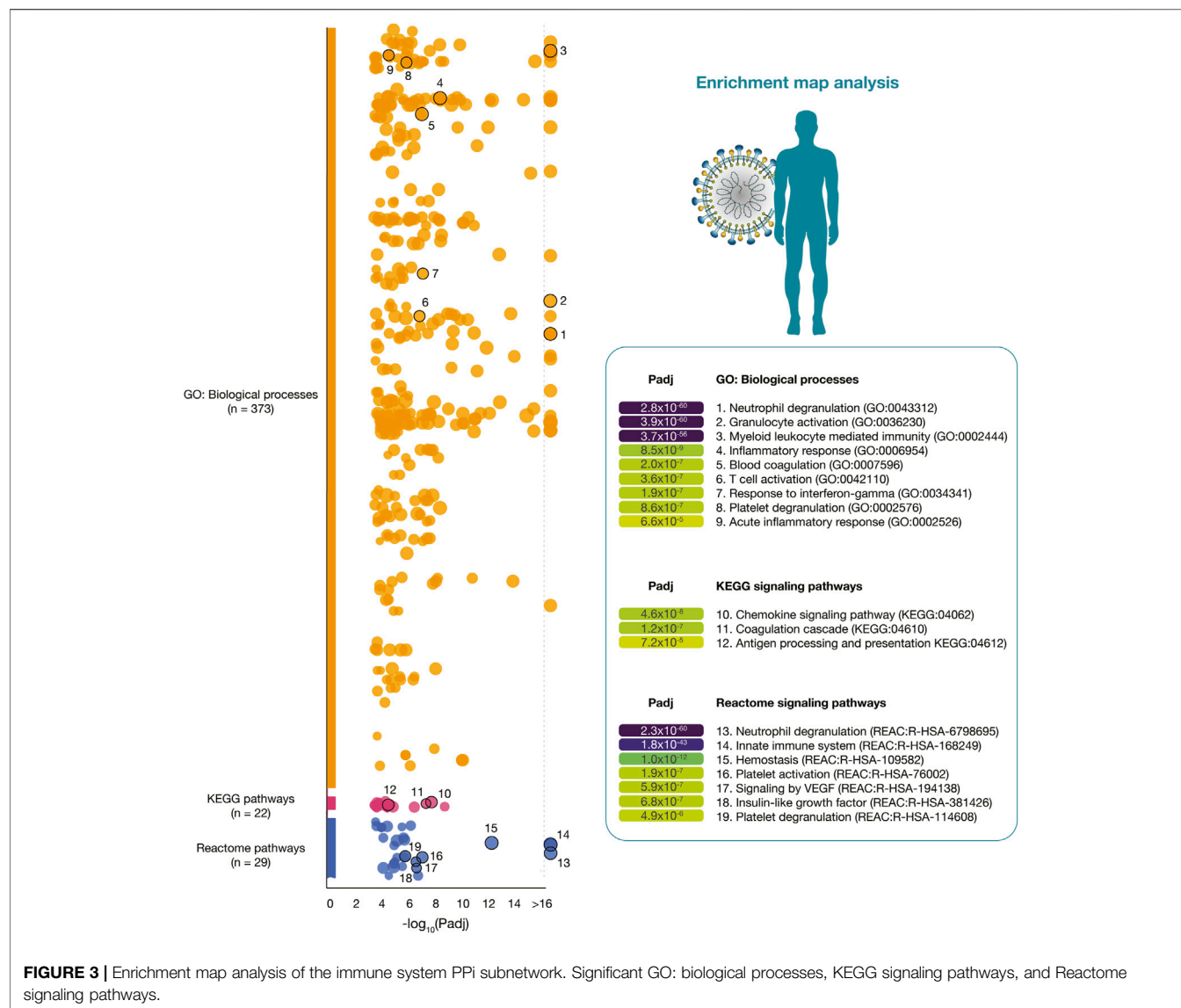
Supplementary Figure S1 details an expanded visualization of the immune system PPI network.

Figure 2B shows the immune system PPI subnetwork encompassing 319 nodes and 5,308 edges. Of them, 26 significantly expressed proteins in severe COVID-19 (15 overexpressed and 11 underexpressed) (Overmyer et al., 2020) had high-confidence interactions (cutoff = 0.9) with 49 human proteins physically associated with SARS-CoV-2 proteins, and with 281 immune system proteins belonging to the first neighbor nodes. The degree centrality mean of the overexpressed proteins was 33.5, and proteins with the highest degree centrality were STOM, HSP90AA1, AGT, ORM1, and ORM2. On the other hand, the degree centrality mean of the underexpressed protein was 32.5, and proteins with the highest degree centrality were KNG1, CFP, ALB, AHSR, and APOA1. Additionally, we have identified 10 significantly expressed phosphorylated proteins of SARS-CoV-2 infection in Vero E6 cells in our subnetwork. The degree centrality mean of the phosphorylated proteins was 32.2, and proteins with the highest degree centrality were PIK3CA, MAPK1, SRC, MAPK3, and AKT1 (**Supplementary Table S4**). Although it has been shown that hubs of high-degree nodes are targets of numerous human viral (Calderwood et al., 2007; De Chassey et al., 2008; Gulbahce et al., 2012; Pan et al., 2016; Huttlin et al., 2017), and are highly correlated with pathogenicity in cancer (López-Cortés et al., 2018; López-Cortés et al., 2020b; Cabrera-andrade, 2020), COVID-19 is a novel disease and requires more in-depth studies.

Functional Enrichment Analysis

The functional enrichment analysis was performed to validate the correlation between the immune system PPI subnetwork and biological annotations related to severe COVID-19. Therefore, after generating the subnetwork encompassing 319 immune system proteins, we performed a functional enrichment analysis using g:Profiler to obtain significant annotations (Benjamini-Hochberg FDR < 0.001) related to GO: biological processes, KEGG signaling pathways, and Reactome signaling pathways (Wang et al., 2016; Slenter et al., 2018; Raudvere et al., 2019; Jassal et al., 2020).

Figure 3 details a Manhattan plot of 373 GO: biological processes, 22 KEGG signaling pathways, and 29 Reactome signaling pathways significantly associated with the 319 immune system proteins. However, after a manual curation of GO terms related to the immunopathology of severe COVID-19, the most significant GO: biological processes were neutrophil degranulation (2.8×10^{-60}), granulocyte activation (3.9×10^{-60}), myeloid leukocyte mediated immunity (3.7×10^{-56}), inflammatory response (8.5×10^{-9}), blood coagulation (2.0×10^{-7}), T-cell activation (3.6×10^{-7}), response to interferon-gamma (1.9×10^{-7}), platelet degranulation (8.6×10^{-7}), and acute inflammatory response (6.6×10^{-5}). The most significant KEGG signaling pathways related to severe COVID-19 were chemokine signaling pathway (4.6×10^{-8}), coagulation cascade (1.2×10^{-7}), and antigen presentation (7.2×10^{-5}). Lastly, the most significant Reactome signaling pathways related to severe COVID-19 were neutrophil degranulation ($2.3 \times$



10^{-60}), innate immune system (1.8×10^{-43}), hemostasis (1.0×10^{-12}), signaling by VEGF (5.9×10^{-7}), insulin-like growth factor (6.8×10^{-7}), and platelet degranulation (4.9×10^{-6}) (Supplementary Table S5).

Single-Cell RNA Sequencing Data Analysis

Omics medicine has evolved the way for identifying therapeutically actionable targets for complex diseases. However, one of the major limitations is the gene expression variability due to the cellular heterogeneity of organs (Gawel et al., 2019). Single-cell biology is a powerful approach that provides unprecedented resolution to the cellular and molecular underpinnings of biological processes and signaling pathways of diseases in order to find therapeutic targets (Ballestar et al., 2020). For instance, the significant overexpression of programmed death 1 (PD-1) in innate lymphoid cells as therapeutic target for cancer immunotherapy (Yu et al., 2016).

Regarding COVID-19, there are several single-cell studies focused on understanding the transcriptional and proteomics insights into the host response for drug discovery (Ballestar et al., 2020; Yang X et al., 2020; Di Giorgio et al., 2020; Wu M et al., 2020; Park and Lee, 2020; Prokop et al., 2020). Ziegler et al. discovered *ACE2* and *TMPRSS2* co-expressing cells in nasal goblet secretory cells, lung type II pneumocytes, and ileal absorptive enterocytes through scRNA-seq data analyses (Ziegler et al., 2020). Once we delimited the interactions between human proteins physically associated with SARS-CoV-2, and immune system proteins (immune system PPI network), we analyzed the transcriptomics data of the 1,584 nodes using three single-cell databases incorporated into the 'COVID-19 Studies' section of the Alexandria Project (see Methods), in order to reveal potential therapeutic targets for drug repurposing against COVID-19.

Chronic rhinosinusitis samples (18,036 cells) developed by allergic inflammation, and nasal scraping samples (18,704 cells)

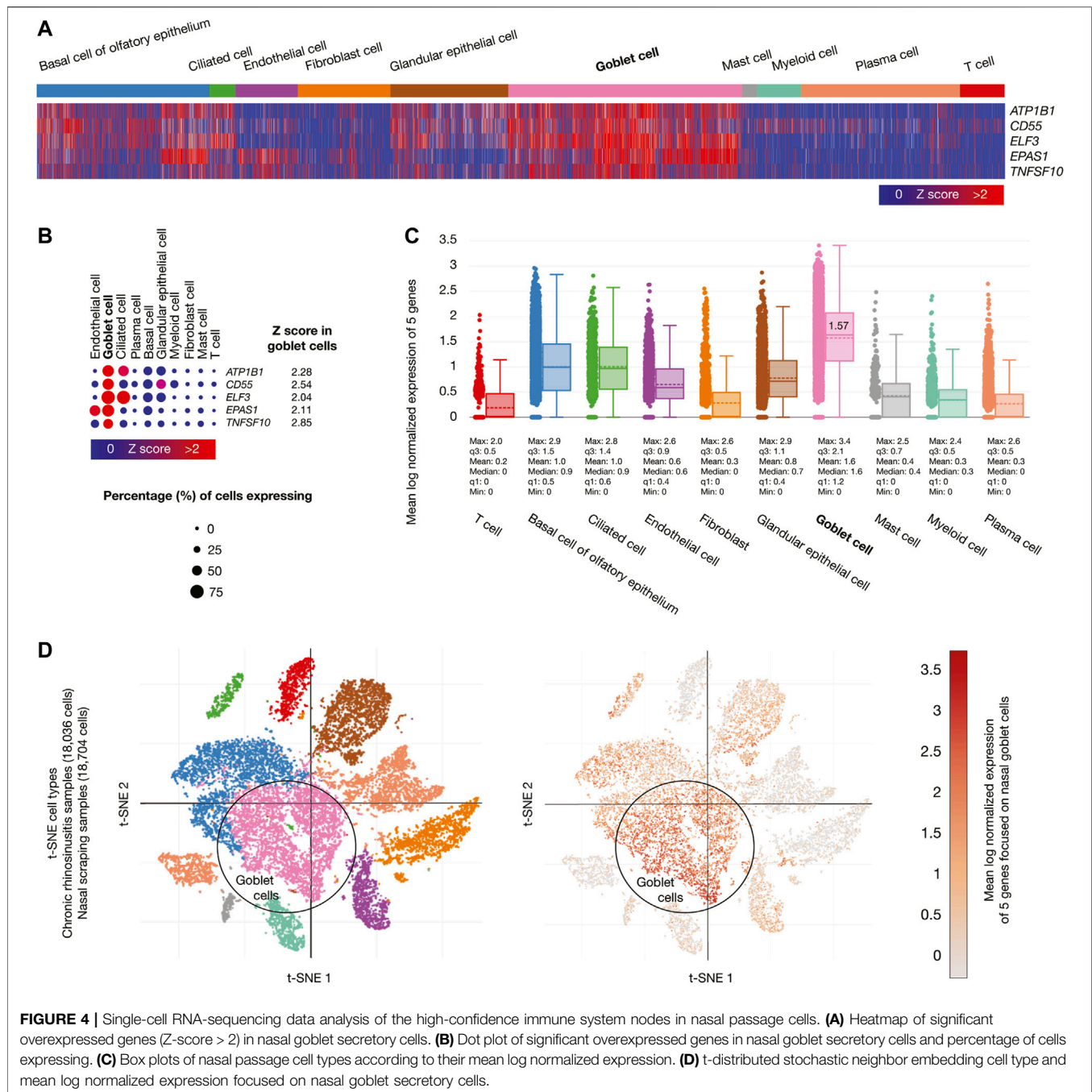


FIGURE 4 | Single-cell RNA-sequencing data analysis of the high-confidence immune system nodes in nasal passage cells. **(A)** Heatmap of significant overexpressed genes (Z-score > 2) in nasal goblet secretory cells. **(B)** Dot plot of significant overexpressed genes in nasal goblet secretory cells and percentage of cells expressing. **(C)** Box plots of nasal passage cell types according to their mean log normalized expression. **(D)** t-distributed stochastic neighbor embedding cell type and mean log normalized expression focused on nasal goblet secretory cells.

conform the nasal passage cells. **Figure 4A** shows a heatmap of the five genes whose mRNAs were significantly overexpressed (Z-score > 2) in goblet cells. **Figure 4B** shows a dot plot detailing the five overexpressed genes, its Z-scores between 2.04 and 2.85, and the percentage of goblet cells expressing the overexpressed genes (>50%). **Figure 4C** shows box plots comparing the mean log normalized expression of the five overexpressed genes in nasal passage cells. Goblet cells had the highest mean log normalized expression (1.57) compared to the other cells. **Figure 4D** projected the expression scores of the significantly

expressed multiple genes ($n = 5$) onto 2D t-SNEs per subpopulation cell (total = 10 subpopulation cells). In summary, five immune system genes were overexpressed in the goblet cells from nasal passages.

Epithelial cells of lung tissue (18,915 cells) were the second single-cell database analyzed. **Figure 5A** shows a heatmap of the 46 genes whose mRNAs were significantly overexpressed in lung type II pneumocytes. **Figure 5B** shows a dot plot detailing the 46 overexpressed genes, its Z-scores between 2.05 and 3.61, and the percentage of type II pneumocytes expressing the

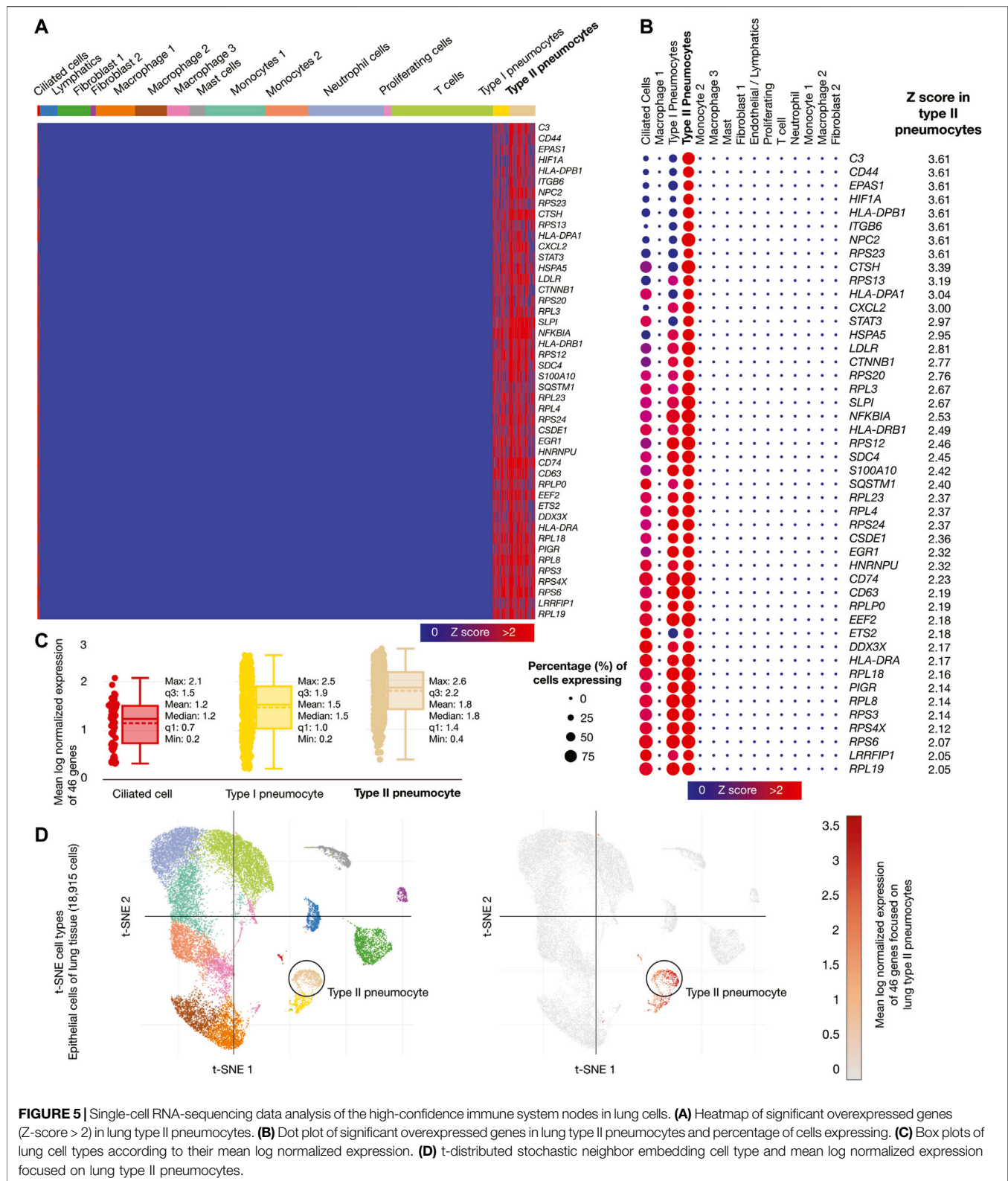


FIGURE 5 | Single-cell RNA-sequencing data analysis of the high-confidence immune system nodes in lung cells. **(A)** Heatmap of significant overexpressed genes (Z-score > 2) in lung type II pneumocytes. **(B)** Dot plot of significant overexpressed genes in lung type II pneumocytes and percentage of cells expressing. **(C)** Box plots of lung cell types according to their mean log normalized expression. **(D)** t-distributed stochastic neighbor embedding cell type and mean log normalized expression focused on lung type II pneumocytes.

overexpressed genes (>50%). **Figure 5C** shows box plots comparing the mean log normalized expression of the 46 overexpressed genes in lung cells. Type II pneumocytes had

the highest mean log normalized expression (1.78) compared to other cells. **Figure 5D** projected the expression scores of the significantly expressed multiple genes ($n = 46$) onto 2D t-SNEs

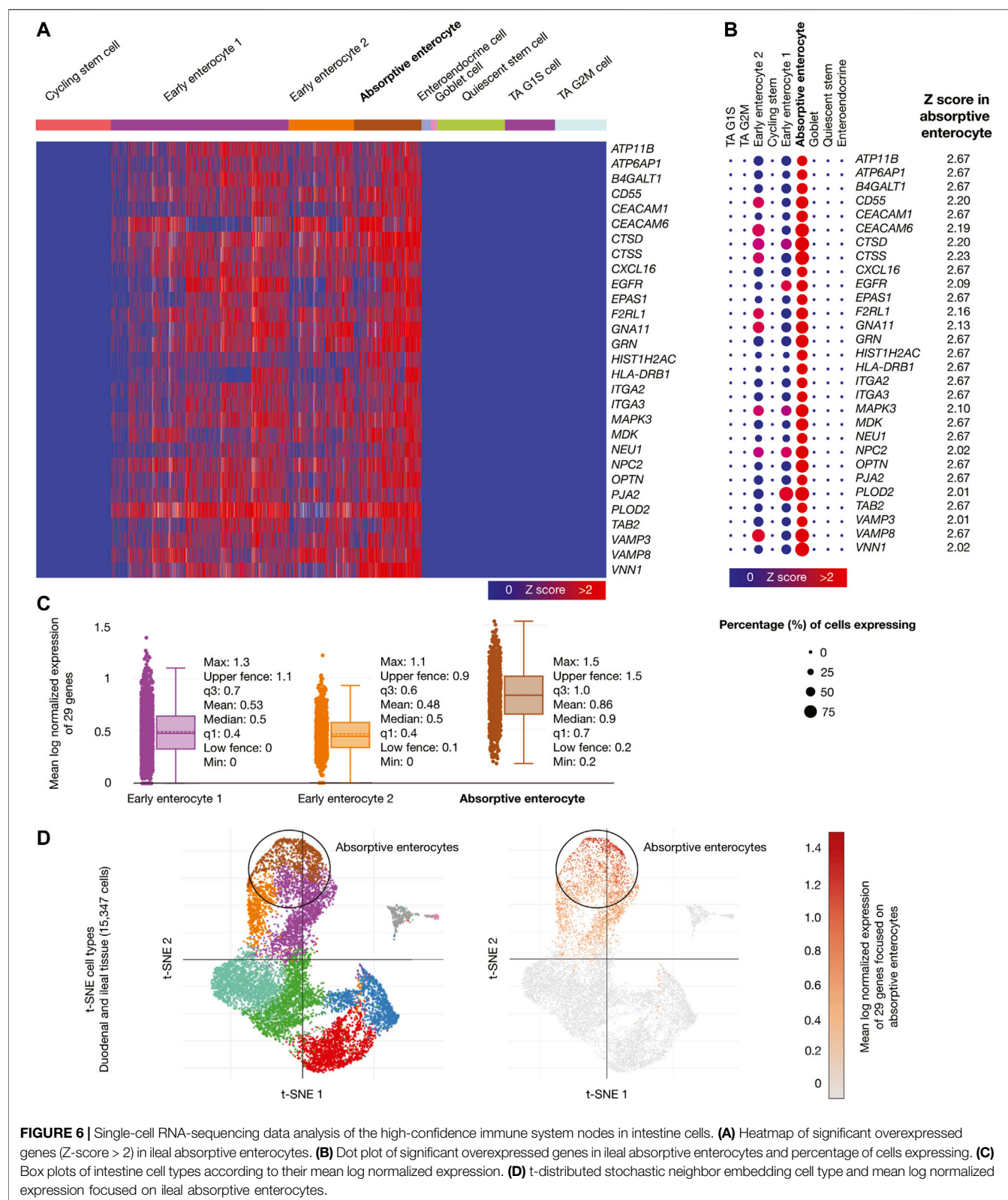
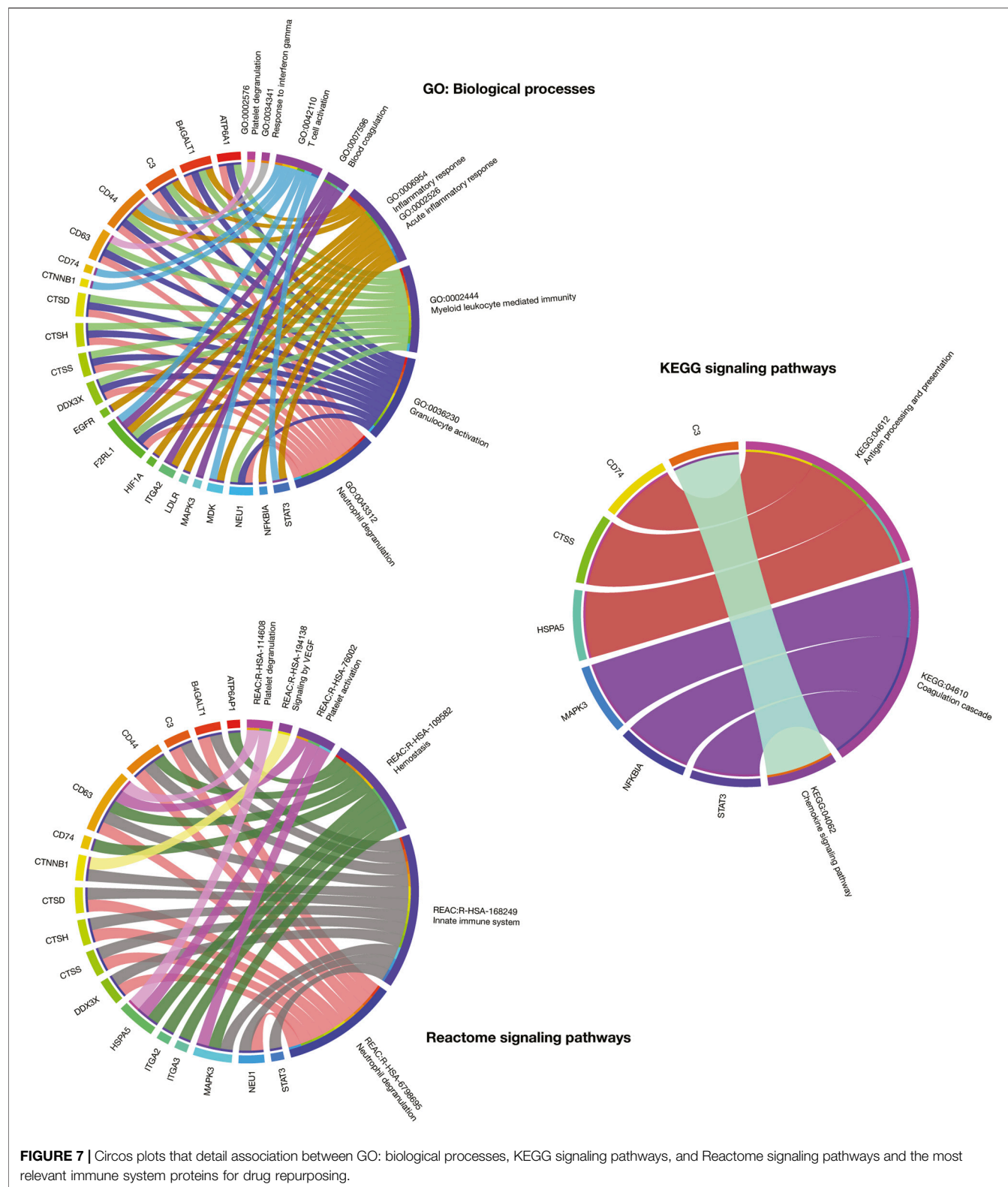


FIGURE 6 | Single-cell RNA-sequencing data analysis of the high-confidence immune system nodes in intestine cells. **(A)** Heatmap of significant overexpressed genes (Z -score > 2) in ileal absorptive enterocytes. **(B)** Dot plot of significant overexpressed genes in ileal absorptive enterocytes and percentage of cells expressing. **(C)** Box plots of intestine cell types according to their mean log normalized expression. **(D)** t-distributed stochastic neighbor embedding cell type and mean log normalized expression focused on ileal absorptive enterocytes.

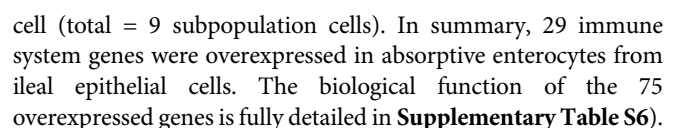
per subpopulation cell (total = 15 subpopulation cells). In summary, 46 immune system genes were overexpressed in type II pneumocytes from lung cells.

Samples from adult human duodenum and ileum (15,347 cells) were the third single-cell database analyzed. **Figure 6A** shows a heatmap of genes whose mRNAs were significantly



overexpressed in ileal absorptive enterocytes. **Figure 6B** shows a dot plot detailing the 29 overexpressed genes, its Z-scores between 2.02 and 2.67, and the percentage of ileal absorptive

enterocytes expressing the overexpressed genes (>50%). **Figure 6C** shows box plots comparing the mean log normalized expression of the 29 overexpressed genes in ileal



Drug Repurposing

The current work proposes an innovative virtual high-throughput screening to predict the activity of 10,672 compounds for 25 immune system targets fully detailed in the **Supplementary Table S7**. The other 50 targets had not identified molecules with active/inactive interactions in the ChEMBL database as previously explained in **Methods** section. Interestingly, the 25 potential therapeutic targets analyzed not only were relevant in the immune system PPi subnetwork and the scRNA-seq analyses, but also had significant associations with biological processes and signaling pathways relevant to severe COVID-19 (Overmyer et al., 2020). For instance, ATP6A1, B4GALT1, C3, CD44, CD63, CTSD, CTSH, CTSS, DDX3X, F2RL1, and NEU1 were involved in neutrophil degranulation; F2RL1, ITGA2, MAPK3, NFKBIA, and STAT3 in blood coagulation or coagulation cascade; ATP6A1, CD44, CD63, CD74, HSPA5, ITGA2, ITGA3, and MAPK3 in hemostasis; lastly, CD63, HSPA5, and MAPK3 in platelet degranulation (**Figure 7**).

The classification model was based on the molecular Circular Fingerprints descriptors (calculated using SMILES formulas) of 15,377 ChEMBL compounds and its 25 therapeutic targets as outputs/tasks. The best model was obtained after a hyperparameter grid search (64 topologies) as a fully connected deep neuronal networks with 1,000 neurons in one hidden layer, with the mean AUROC of 0.989 ± 0.019 (AUROC between 0.935 and 1.000 for 25 classes). Our free GitHub repository contains the Jupyter notebook as python script using DeepChem methodology, datasets, calculated descriptors, best model, metrics of the model, and predictions. After applying the best classification model, we evaluated drugs taking into account the first ATC levels associated to COVID-19 symptoms, drug category, mechanism of action, pharmacological indications, and the best ranked AUROC values (threshold > 0.8). Consequently, on one hand, we obtained 44 approved drugs, 16 compounds under investigation, and 35 experimental compounds with the highest affinities for 15 immune system proteins (**Supplementary Table S8**). On the other hand, we obtained four approved drugs, nine compounds under investigation, and 16 experimental compounds with the highest multi-target affinities for nine immune system proteins (**Supplementary Table S9**).

Figure 8 details the AUROC affinity score of the best-predicted experimental compounds, compounds under investigation, and approved drugs per immune system protein target and multi-targets. We found eleven different categories of approved drugs, the anti-neoplastic and immunomodulating agents were lanreotide, enzalutamide, topotecan, erlotinib, methotrexate, imatinib, pemetrexel, lapatinib, sunitinib, vandetanib, midostaurin, bosutinib, axitinib, ruxolitinib, afatinib, ibrutinib, duvelisib, and gilterintinib; the anti-hemorrhagic agent was fostamatinib; the anti-inflammatory agents were clobetasol propionate, nedocromil, oxaprozin, and beclomethasone dipropionate; the anti-malarial agent was halofantrine; the anti-parathyroid agent was etelcalcetide; the anti-viral agents were amprenavir, atazanavir, saquinavir,

darunavir, fosamprenavir, lopinavir, paritrapevir, nelfinavir, pibrentasvir, zanamivir, peramivir, and rilpivirine; the antioxidant agent was allopurinol; the cardiovascular agents were aliskiren, zofenopril, digitoxin, torasemide, and triamterene; the central nervous system agents were citicoline and cabergoline; the growth hormone-releasing hormone was tesamorelin; and the only antibiotic was rosoxacin.

Interestingly, 13 (27%) of the 48 best-predicted approved drugs are currently involved in approximately 54 COVID-19 clinical trials as detailed in **Figure 9**. The cardiovascular agents with clinical trials are aliskiren, torasemide, and triamterene. Aliskiren had an AUROC affinity of 0.993 on CTSD, and it is a renin inhibitor used to treat hypertension; torasemide had an AUROC affinity of 1.0 on EGFR, and it is used to treat edema associated with heart, renal, and hepatic failures; and triamterene had an AUROC affinity of 1.0 on EGFR, and it is used to treat hypertension. The anti-viral agents with clinical trials are atazanavir, darunavir, and lopinavir. Atazanavir had an AUROC affinity of 0.997 on CTSD; darunavir had an AUROC affinity of 0.999 on CTSD, and lopinavir had an AUROC affinity of 1.0 on CTSD. All of them are protease inhibitors used to treat HIV infection. The anti-neoplastic and immunomodulating agents with clinical trials are enzalutamide, methotrexate, imatinib, ruxolitinib, ibrutinib, and duvelisib. Enzalutamide had an AUROC affinity of 0.983 on CTSS, and it is an androgen receptor inhibitor to treat prostate cancer; methotrexate had an AUROC affinity of 1.0 on EGFR, and it is an antimetabolite used to treat breast cancer, lung cancer, head and neck cancer, and non-Hodgkin's lymphoma; imatinib had an AUROC affinity of 1.0 on EGFR, and it is a BCR/ABL kinase inhibitor used to treat chronic myeloid leukemia, acute lymphoblastic leukemia, and gastrointestinal stromal tumors; ruxolitinib had an AUROC affinity of 1.0 on EGFR, and it is an inhibitor of JAK1/2 to reduce the hyperinflammation during cytokine storm in thrombocytopenia myelofibrosis; ibrutinib had an AUROC affinity of 1.0 on EGFR, and it is an inhibitor of the Bruton tyrosine kinase causing protection against immune-induced lung injury; and duvelisib had an AUROC affinity of 1.0 on EGFR, and it is a PI3K inhibitor involved in the immune homeostasis restoration and viral replication inhibition. Finally, the anti-hemorrhagic agent with clinical trial was fostamatinib, which had an AUROC affinity of 1.0 on EGFR, and it is an inhibitor of spleen tyrosine kinase used to treat chronic immune thrombocytopenia (**Supplementary Table S10**; Wishart et al., 2018).

DISCUSSION

Since the finding of patient zero in China, a wide spectrum of clinical manifestations has been discovered, as we have understood the COVID-19 disease. The most common initial symptoms are cough, fever, anorexia, and dyspnea (Wang D et al., 2020; Berlin et al., 2020). The most common clinical features in severe COVID-19 patients are dyspnea, severe hypoxemia, lung edema, respiratory failure, ARDS (Montenegro et al., 2020), lymphopenia (Terpos et al., 2020),

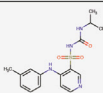
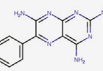
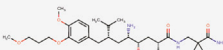
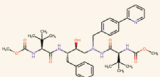

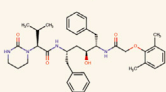
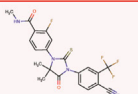
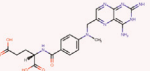
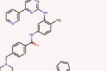

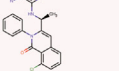

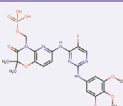
| Drug Category | Drug | Structure | Pharmacological indication | COVID-19 clinical trial identifier |
|---|-------------------------|---|--|--|
| Cardiovascular agents | Torsemide DB00214 |  | Renin-Angiotensin System (RAS). Edema associated with heart, renal or hepatic failures. | NCT04467931 |
| | Triamterene DB00384 |  | Renin-Angiotensin System (RAS). Hypertension. | NCT04467931 |
| | Aliskiren DB09026 |  | Renin-Angiotensin System (RAS). Hypertension, hypotension. | ChiCTR2000032314 |
| Anti-viral agents | Atazanavir DB01072 |  | Protease inhibitor. HIV infection | NCT04459286, NCT04452565, NCT04468087 |
| | Darunavir DB01264 |  | Protease inhibitor. HIV infection | NCT04425382, NCT04252274 |
| | Lopinavir BD01601 |  | Protease inhibitor. HIV infection | NCT04372628, NCT04455958, NCT04499677, NCT04425382, NCT04466241, NCT04364022, NCT04328012, NCT04255017, NCT04321174, NCT04295551, NCT04521400, NCT04261907, NCT04315948, NCT04275388, NCT04276688, NCT04386876 |
| Anti-neoplastic and immunomodulating agents | Enzalutamide DB08899 |  | Androgen receptor inhibitor. Metastatic prostate cancer. | NCT04456049, NCT04475601 |
| | Methotrexate DB00563 |  | Antimetabolite. Breast cancer, lung cancer, T cell lymphoma, head and neck cancer. | NCT04352465, NCT04610567 |
| | Imatinib DB00619 |  | Kinase inhibitor. Leukemia, gastrointestinal stromal tumors. | NCT04357613, NCT04346147, NCT04394416, NCT04356495 |
| | Ibrutinib DB09053 |  | Inhibition of Bruton tyrosine kinase. Mantle cell lymphoma, chronic lymphocytic leukemia. | NCT04375397, NCT04439006 |
| | Duvelisib DB11952 |  | PI3K inhibitor. Leukemia and lymphoma. | NCT04372602, NCT04487886 |
| | Ruxolitinib DB08877 |  | Kinase inhibitor. Thrombocythemia myelofibrosis. | NCT04348071, NCT04338958, NCT04362137, NCT04359290, NCT04355793, NCT04377620, NCT04366232, NCT04334044, NCT04374149, NCT04338958, NCT04337359, NCT04331665, NCT04361903, NCT04348695, NCT04424056 |
| Anti-hemorrhagic agent | Fostamatinib DB12010 |  | Kinase inhibitor. Chronic immune thrombocytopenia. | NCT04629703, NCT04581954, NCT04579393 |

FIGURE 9 | Best-predicted approved drugs involved in COVID-19 clinical trials. Cardiovascular agents, anti-viral agents, anti-neoplastic and immunomodulating agents, and anti-hemorrhagic agent with their respective clinical trial identifier number, pharmacological indication, and chemical structure according to DrugBank.

cardiac arrhythmias, rhabdomyolysis, hyperferritinemia, intravascular coagulopathy (Fogarty et al., 2020), and pulmonary thromboembolism (Rotzinger et al., 2020). Also, it has been observed that 15% of patients required supplemental oxygen (Young et al., 2020), and 5% of patients required mechanical ventilation. In addition, the smaller percentage of patients who required mechanical ventilation suffered comorbidities that lead to sepsis and septic shock (Rhee et al., 2020). Nowadays, it is known that SARS-CoV-2 is

capable of reaching other organs depending on the host (Yang W et al., 2020). Different studies worldwide refer that clinical presentation vary between individuals, presenting manifestations not only respiratory tract infection, but also blood, skin, kidney, liver, ocular symptoms, neurologic signs, among others (Adhikari et al., 2020; Wang Q et al., 2020). Therefore, it is necessary to continuously review the reports on clinical manifestations in order to get to know the behavior of this disease as well as to think over the physiopathological

mechanisms that allows us to better understand the related complications (Gupta et al., 2020; Wadman et al., 2020).

The effective immune response of the host, including the innate and adaptive ones, against SARS-CoV-2 seems to be essential to control and solve the infection. However, the clinical seriousness of COVID-19 could be associated to the excessive production of pro-inflammatory cytokines, known as 'cytokine storm' (Fajgenbaum and June, 2020; Hussman, 2020), or to the excessive production of bradykinin peptides, known as 'bradykin storm' (Garvin et al., 2020). This clinical paradigm is still to be figured out, and that is why the effective treatment is still uncertain. It is indispensable to understand the immunological responses that are triggered off since the beginning of the infection with SARS-CoV-2, so as to make progress in search of effective therapeutic strategies.

Innate immune response executes the first line of antiviral defense and is essential to obtain immunity against viruses (Zhong et al., 2020). Pattern recognition receptors (PRRs), codified by germline DNA, are responsible for recognizing widely common molecular patterns shared by pathogens of a certain group. Single-stranded and double-stranded viral RNAs produced during the replication phase of SARS-CoV-2 are recognized by endosomal TLRs (TLR7 and TLR8 or TLR3, respectively) and cytosolic RIG-I like receptors (RLRs), mainly RIG-I and MDA-5. After PRR engagement, downstream signaling pathways trigger the activation and nuclear translocation of key transcription factors, such as NF- κ B, AP-1 and interferon regulatory factors (IRFs), and the ensuing expression of inflammasome activation and anti-viral cytokines (Lee et al., 2020). Among the most relevant cytokines we can find interleukins (IL-1, IL-6, and IL-18), pro-inflammatory TNF- α and TNF- β , and type I and III IFNs (Blanco-Melo et al., 2020; Herold et al., 2020; McKechnie and Blish, 2020). Consequently, cytokines induce antiviral processes potentiating the innate and adaptive immune responses, limiting CoVs replication capacity and inducing the elimination of the virus cell reservoirs (Channappanavar et al., 2019; Blanco-Melo et al., 2020). However, CoVs have developed mechanisms of immune evasion where viral factors inhibit viral recognition by PRR sensing, and cytokine expression and secretion. Individuals with severe COVID-19 have demonstrated remarkably impaired type I IFN values as compared to mild patients (Hadjadj et al., 2020), and the interferon-induced overexpression of ACE2 may be involved (Ziegler et al., 2020).

Mucosal immune responses against viruses are orchestrated by myeloid cells such as macrophages, conventional DCs, plasmacytoid DCs, and monocyte-derived DCs (Guilliams et al., 2013). Accumulating evidence suggests that deregulation of myeloid cell-mediated responses potentially triggers lymphopenia, cytokine release syndrome, acute respiratory distress syndrome (Mehta et al., 2020), and pathogenic inflammation with high level secretion of IL-6, IL-2, IL-7, IFN- γ , IFN-I, and type III IFNs (Shi et al., 2019) in COVID-19 patients with severe clinical manifestations.

Innate lymphoid cells (ILCs) are lymphoid-like immune cells that lack the expression of rearranged antigen receptors. The non-cytotoxic group I, II, and III ILCs and the cytotoxic natural killer

(NK) cells form the ILC family (Vivier et al., 2018). Several clinical data have reported that NK cells decrease in peripheral blood of severe patients (Song et al., 2020; Yu et al., 2020). An *in vitro* study has identified that the CXCL9-11 chemokines are overexpressed in lung cells infected with SARS-CoV-2, suggesting that the CXCR3 signaling pathway drives NK cells from peripheral blood to lungs in COVID-19 patients (Liao M et al., 2020). In addition, NK cells have the quality to induce lysis of infected cells causing severe hypoxemia and contributing to the cytokine storm resulting in ARDS.

T cells are involved in fundamental processes in viral infections. CD8 T cells eliminate infected cells and CD4 T cells help B cells for antibody production. Nevertheless, immunopathology is generated when T cells are dysregulated. Several reports have shown that moderate to severe COVID-19 patients with lymphopenia drastically reduce CD8 T cell and CD4 T cells in peripheral blood (Nie et al., 2020; Wen et al., 2020; Zeng et al., 2020). T cells reduction in the blood is also a contribution of mechanisms such as inflammatory cytokine milieu, which is why lymphopenia has a correlation with TNF- α , IL-6, and IL-10 (Diao et al., 2020; Wan et al., 2020). Conversely, clinical reports have shown that convalescent patients have low pro-inflammatory cytokine levels paired with restored bulk T cell frequencies (Diao et al., 2020).

The humoral immune response plays a main role in the clearance of cytopathic viruses and its memory response prevents reinfection. According to Huang et al. and Wu et al., IgM, IgA, and neutralizing IgG antibodies can be detected in 12, 14 and 10–14 days, respectively, after symptom onset on average, suggesting that SARS-CoV-2 causes a robust B cell response in the majority of COVID-19 patients (Wu F et al., 2020; Huang et al., 2020). Indeed, antibodies binding the RBD of the S glycoprotein can have neutralizing properties, blocking virus interactions with the human protein receptor ACE2 (Ju et al., 2020), thereby inhibiting/preventing target cell infection. The B cell response to SARS-CoV-2 protects from the primary infection and extends immunity against reinfection due to memory B cells that can respond quickly by producing high affinity neutralizing antibodies. However, it is yet impossible to predict the duration of memory responses due to the timing of the COVID-19 pandemic.

There is currently a limited number of known risk factors that confer susceptibility to COVID-19. Several routine blood tests and immunological biomarkers have been suggested to classify patients with mild and severe symptoms. The routine blood test biomarkers currently suggested are lymphocyte count (Tan et al., 2020), neutrophil to lymphocyte ratio (Liu et al., 2020b), C-reactive protein (Ji et al., 2020), lactate dehydrogenase (Xiang et al., 2020), ferritin (Bataille et al., 2020), D-dimer and coagulation parameters (Zhou et al., 2020b), serum amyloid protein (Ji et al., 2020), N terminal pro B type natriuretic peptide (Gao L et al., 2020), platelet count (Qu et al., 2020), ultrasensitive troponin, and creatine kinase MB (Akhmerov and Marbán, 2020). On the other hand, immunological biomarkers associated with different COVID-19 outcomes are CD4⁺, CD8⁺, and NK cell count (Nie et al., 2020); PD-1 and Tim-3 expression on T cells (Diao et al., 2020); phenotypic changes in peripheral

blood monocytes (Zhang D et al., 2020); expression levels of IP-10, MCP-3, IL-1ra (Yang Y et al., 2020); IL-6 (Chen et al., 2020), IL-8, IL-10, IL-2R, IL-1 β (Gong et al., 2020), IL-4 (Fu et al., 2020), IL-18, granulocyte macrophage colony stimulating factor (GM-CSF) (Zhou et al., 2020a), IL-2, IFN- γ (Liu et al., 2020a), and anti-SARS-CoV-2 antibodies (Zhang B et al., 2020; Fu et al., 2020).

In this study, we performed proteomics, transcriptomics, and artificial neural network analyses to reveal potential therapeutic targets for drug repurposing to treat severe COVID-19. Firstly, we generated an immune system PPI network encompassing 1,584 nodes and 332,968 edges. Of them, 256 human proteins physically associated with SARS-CoV-2 proteins (Gordon et al., 2020) had high-confidence interactions with 1,390 immune system proteins. The degree centrality mean of the human proteins physically associated with SARS-CoV-2 was 23.6. GNB1, with the highest degree centrality, acts as a modulator in transmembrane signaling systems, including the GTPase activity (Gordon et al., 2020). The degree centrality mean of the immune system proteins was 44.5. UBA52, with the highest degree centrality, acts as a fusion protein that regulates ubiquitination of ribosome (Kobayashi et al., 2016). Lastly, the degree centrality mean of the phosphorylated proteins was 59.8. PIK3CA had the highest degree centrality and significant underexpression in SARS-CoV-2 infection in Vero E6 cells (Bouhaddou et al., 2020; **Figure 2A**; **Supplementary Figure S1**).

Overmyer *et al.* published a large-scale multi-omic analysis and identified 146 significantly expressed proteins in severe COVID-19 (Overmyer et al., 2020). We located these proteins and their high-confidence interactions in the immune system PPI network and subsequently generated the immune system PPI subnetwork encompassing 319 nodes and 5,308 edges. Of them, 26 significantly expressed proteins in severe COVID-19 (Overmyer et al., 2020) had high-confidence interactions with 49 human proteins physically associated with SARS-CoV-2 proteins, and with 281 immune system proteins. The degree centrality mean of the overexpressed proteins was 33.5. STOM, with the highest degree centrality, is located in cell membranes regulating ion channels and transporters. Loss of localization of the encoded protein is associated with hemolytic anemia shown in COVID-19 patients (Algassim et al., 2020). The degree centrality mean of the underexpressed proteins was 32.5. KNG1, with the highest degree centrality, is the precursor for bradykin synthesis, and is involved in the coagulation system dysfunction of severe COVID-19 (Sidarta-Oliveira et al., 2020). Lastly, the degree centrality mean of the phosphorylated proteins was 32.2, and PIK3CA had the highest degree centrality (**Figure 2B**).

SARS-CoV-2 employs a suite of virulent proteins that interact with key targets in host interactomes to extensively rewire the flow of information and cause COVID-19 (Vidal et al., 2011; Pan et al., 2016; Kumar et al., 2020). Although it has been shown that hubs of high-degree nodes are targets of numerous human viral (Calderwood et al., 2007; De Chasseay et al., 2008; Gulbahce et al., 2012; Pan et al., 2016; Huttlin et al., 2017), COVID-19 is a novel disease and requires more in-depth studies. Therefore, we performed a functional enrichment analysis to validate the correlation between the subnetwork proteins and COVID-19

signatures published in studies worldwide (**Figure 3**). After a manual curation of gene ontology terms, the most significant biological processes were neutrophil degranulation (Shen et al., 2020), granulocyte activation (Yang L et al., 2020), myeloid leukocyte mediated immunity (Chen and John Wherry, 2020), inflammatory response (Jose and Manuel, 2020; Merad and Martin, 2020), blood coagulation (Vinayagam and Sattu, 2020), T-cell activation (Chen and John Wherry, 2020), response to interferon-gamma (Hu et al., 2020), platelet degranulation (Kuchi Bhotla et al., 2020), and acute inflammatory response (Manjili et al., 2020). The most significant KEGG pathways were chemokine signaling pathway (Chua et al., 2020), coagulation cascade (Overmyer et al., 2020), and antigen presentation (Li X et al., 2020). Lastly, the most significant Reactome signaling pathways were neutrophil degranulation (Wang J et al., 2020), innate immune system (Ahmed-Hassan et al., 2020), hemostasis (Liao D et al., 2020), signaling by VEGF (Kong et al., 2020), insulin-like growth factor (Winn, 2020), and platelet degranulation (Overmyer et al., 2020).

According to Buccitelli & Selbach (Buccitelli and Selbach, 2020), proteomics and transcriptomics typically show reasonable correlation, and integrating both types of data can reveal exciting biology and gene expression patterns. In light of this approach, the 'COVID-19 Studies' section of the Alexandria Project represents a large effort to characterize this immunopathology from a transcriptomics view. Ziegler *et al.* analyzed human scRNA-seq data to uncover potential targets of SARS-CoV-2 amongst tissue-resident cell subsets. They discovered ACE2 and TMPRSS2 co-expressing in goblet cells from nasal passage cells, type II pneumocytes from lung epithelial cells, and absorptive enterocytes from ileal epithelial cells (Ziegler et al., 2020). Therefore, after generating our immune system PPI network, we screened the 1,584 nodes into 10 nasal passage cells, 15 lung epithelial cells, and nine ileal epithelial cells to identify potential therapeutic targets for drug repurposing against COVID-19.

We found 75 significantly overexpressed molecules (Z-score > 2) in nasal goblet secretory cells (n = 5) (**Figure 4**), lung type II pneumocytes (n = 46) (**Figure 5**), and ileal absorptive enterocytes (n = 29) (**Figure 6**; Reimand et al., 2019). Subsequently, we analyzed the druggability of these 75 molecules (**Methods** section), and identified 25 potential therapeutic targets with ChEMBL ID and identified molecules with active/inactive interactions.

Meaningfully, these potential therapeutic targets not only were relevant in both the immune system PPI subnetwork and the scRNA-seq data, but also were involved in biological processes and signaling pathways related to severe COVID-19, such as neutrophil degranulation, blood coagulation or coagulation cascade, hemostasis, and platelet degranulation (**Figure 7**; Overmyer et al., 2020). Several studies worldwide have correlated these potential therapeutic targets with COVID-19. For instance, MAPK3 and EGFR showed kinase activity in the global phosphorylation landscape of SARS-CoV-2 infection according to Bouhaddou *et al.* (Bouhaddou et al., 2020). CTSD, CD63, MKD, NFKBIA, MAPK3, STAT3, TNFSF10, F2RL1, HIF1A, NEU1, and EPAS1 were identified as

significantly expressed targets in patients with severe COVID-19 according to Aschenbrenner *et al* (Aschenbrenner *et al.*, 2020). C3, LDLR, CTSB, B4GALT1 and NFKB1A were significantly expressed targets in COVID-19 according to Alsamman & Zayed (Alsamman and Zayed, 2020). HSPA5 was associated with the viral entry, the endoplasmic reticulum stress, and anti-clotting agents according to Law *et al* (Law *et al.*, 2020). CD44 was involved in the extravasation cascade with significant expression in severe COVID-19 according to Chua *et al* (Chua *et al.*, 2020). Basu *et al* found significant expression of the ITGA2 and ITGA3 integrins in COVID-19 patients (Basu *et al.*, 2020). DDX3X was involved in the coronavirus-host protein-protein interactions according to Perrin-Cocon *et al* (Perrin-Cocon *et al.*, 2020). Daniloski *et al* showed that ATP6AP1 induces shared transcriptional changes in cholesterol biosynthesis in human cells with SARS-CoV-2 infection (Daniloski *et al.*, 2020). Lastly, CD74, CTSS and CTNBN1 were identified as potential targets for SARS-CoV-2 diagnosis and treatment according to Vastrad *et al* (Vastrad *et al.*, 2020).

There is currently an urgent need for effective COVID-19 drugs. High-throughput screening for drug discovery has been important in finding antiviral drugs focused on the SARS-CoV-2 spike protein (Nicholas and Jeremy, 2020) and the main protease (M^{pro}), as detailed in our previous study (Tejera *et al.*, 2020). However, computational structure-based drug discovery focused on immune system proteins is imperative to select potential drugs that, after being effectively analyzed in cell lines (i.e., African green monkey cells) and clinical trials, these can be considered for treatment of complex symptoms of COVID-19 patients. Drug repurposing offers a potentially rapid mechanism to deployment, since the safety profiles are known (Cabrera-Andrade *et al.*, 2020; Phimister *et al.*, 2020).

We performed fully connected deep neuronal networks to predict drugs with the highest affinities per target and multi-targets. We identified 47 approved drugs, 25 compounds under investigation, and 50 experimental compounds with the highest AUROCs for 15 (60%) of the 25 potential therapeutic targets. The best-predicted approved drugs were enrolled in ten different categories: anti-neoplastic and immunomodulating agents, anti-hemorrhagic agents, anti-inflammatory agents, anti-parathyroid agents, anti-viral agents, anti-oxidant agents, cardiovascular agents, central nervous system agents, growth hormone-releasing hormone, and antibiotics (see **Results** section and **Figure 8**).

There are around 4,000 clinical trials on COVID-19 using small molecules as single or combination agents with other anti-viral agents worldwide. Interestingly, 54 clinical trials currently correspond to 13 (27%) of the 48 best-predicted approved drugs found in our study (**Figure 9**). The cardiovascular agents implicated in the renin-angiotensin system are aliskiren, triamterene, and torasemide. Aliskiren and triamterene are renin inhibitors used to treat hypertension; and torasemide is used to treat edema associated with heart, renal, and hepatic failures. According to Garvin *et al.*, the renin-angiotensin system is an important pathway linked to hypertension and hypotension in COVID-19 patients because it maintains a balance of blood pressure (Garvin *et al.*, 2020). The anti-viral agents are atazanavir,

darunavir, and lopinavir. All of them are protease inhibitors used to treat HIV infection. According to Mahdi *et al.*, targeting of SARS-CoV-2 M^{pro} by HIV protease inhibitors might be of limited clinical potential due to the high concentration of drug required to achieve this inhibition. However, any potential beneficial effect in COVID-19 context might be attributed to acting on other molecular targets (Mahdi *et al.*, 2020). The anti-neoplastic and immunomodulating agents are enzalutamide, methotrexate, imatinib, ruxolitinib, ibrutinib, and duvelisib. Enzalutamide is an androgen receptor inhibitor to treat prostate cancer; methotrexate is an antimetabolite that inhibits the dihydrofolate reductase and is used to treat breast cancer, lung cancer, head and neck cancer, and non-Hodgkin's lymphoma; imatinib is a BCR/ABL kinase inhibitor used to treat chronic myeloid leukemia, acute lymphoblastic leukemia, and gastrointestinal stromal tumors; ruxolitinib is a Janus kinase 1 and 2 inhibitor that reduces the hyperinflammation during cytokine storm in thrombocytopenia myelofibrosis; ibrutinib is an inhibitor of the Bruton tyrosine kinase causing protection against immune-induced lung injury; and duvelisib is a PI3K inhibitor involved in the immune homeostasis restoration and viral replication inhibition. According to Saini *et al.*, three hallmarks of cancer, namely immune dysfunction, inflammation, and coagulopathy are also seen in patients with SARS-CoV-2 infection, providing a biological rationale for testing anti-neoplastic agents for their ability to control the severe COVID-19 symptoms. However, these anti-neoplastic drugs should be evaluated carefully through well-designed and often novel trial platforms to avoid detrimental effects in future treatments (Saini *et al.*, 2020). Finally, the anti-hemorrhagic agent, fostamatinib, is an inhibitor of spleen tyrosine kinase used to treat chronic immune thrombocytopenia. According to Kost-Alimova *et al.*, elevated mucin-1 (MUC1) protein levels predict acute lung injury and ARDS with poor clinical outcomes, and fostamatinib has been shown to reduce MUC1 abundance in a relevant pre-clinical model and has demonstrated safety profile in patients (Kost-Alimova *et al.*, 2020; Tabassum *et al.*, 2020).

Despite enormous scientific effort in drug repurposing studies to inhibit SARS-CoV-2 proteins or control severe COVID-19 symptoms, significant limitations exist. The main concern associated with drug repurposing studies involves the implementation of well-designed validation assays through clinical trials. Other main concerns are related to obtaining the correct therapeutic doses, safety results to avoid detrimental effects of repurposed drugs after treatments, and delivery capabilities worldwide (Parvathaneni and Gupta, 2020). All of this carried out counter clock due to the health emergency triggered by the pandemic. However, the positive side of this enormous scientific effort is to put forward recommendations for transforming today's tools into solutions for future pandemics according to The National Symposium on Drug Repurposing for Future Pandemics, on behalf of the National Science Foundation.

The current COVID-19 pandemic offers a unique opportunity to strengthen mechanisms that promote the use of drug repurposing processes—considering the drug safety profile and the possibility of originate different adverse reactions in patients with distinct concomitant diseases—; inclusively, in the ongoing or

future clinical trials, having the potential to reduce the time and costs for finding potential solutions to the current pandemic. Additionally, contributing to future analysis for high threat pathogens and rare diseases. This idea is welcomed by some other authors who conveyed on the potential of drug repurposing for common national and global health benefits (Yan, 2017). Between the several advantages of this process, the one which leads efforts to the use of the current information -on human pharmacology and toxicology-of safe and affordable generic drugs, is worth to remark. As also stated by Guy et al. (2020), along with this statement, there is the urge to motivate the transparency and compliance of the highest ethical principles for the conduction of studies, including as a key potential for drug repurposing, the visualization and sharing of negative results. Mainly, promoting and assuring that well-designed randomized clinical trials are timely implemented, especially during health emergencies and crises. In this sense, drug repurposing will be fulfilling its main objective: proposing potential, prompt, cost-effective, and safe solutions for the public and global health problems, with a human-centered approach.

The COVID-19 pandemic has evidenced that there is a strong urge to strengthen health systems with a major emphasis on health prevention and the major need, especially of low and middle income countries, to publicly invest on research and development. Consequently, the benefits of innovation and the results of research should be always available and affordable to anyone in need, to comply with the goal of public health (Røttingen et al., 2012). This is of particular importance during the current pandemic situation and on its aftermath.

From a global health perspective, initiatives directed to the improvement of rapid data sharing are critical during health emergency. This rapid sharing includes undoubtedly a transboundary collaboration founded on the principles of reliability and accuracy of the data (The Lancet, 2020). Meaningfully, for preventing potential new or existing pathogens to become high threats to human health and global security, non-commercial basic research on microorganisms should be assured. Additionally, introducing and promoting genomic epidemiology and strengthening global laboratory alliances would contribute to the national and global rapid detection and containment of outbreaks, as also promoted by the WHO. Accordingly, every country is sovereign and should guarantee the protection and regulation of the use of its biological resources, specifically working toward the Fair and Equitable Sharing of Benefits. Nevertheless, international conventions on the topic and national legislations should include fast track options for research on pathogens (Knauf et al., 2019). Relevantly, the links between human, environmental, and

animal health - the One Health approach-are widely recognized to be effective toward the prevention and reduction of the emergence and re-emergence of potential pandemic agents (El Zowalaty and Järhult, 2020). This, not only pursuing to diminish the impact of epidemics or pandemics in the health systems, but also to underpin and reinforce economic, development, and social benefits.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the article and in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

ALC conceived the subject, the conceptualization of the study, and wrote the manuscript. PG, NK, and CB edited the manuscript and gave valuable scientific input. CM and ET performed the artificial neural networks. AL, SG, EO, DC, AG, KS, AG, GP, SM, JG, AZ, YP, AC, LP, CP, JB, AQ, NV, LQ, and CP gave valuable scientific input and did data curation. Lastly, all authors reviewed and approved the manuscript.

FUNDING

Publication of this article was funded by Universidad UTE-Ecuador, and ANID grant COVID0789-Chile. This manuscript has been released as a pre-print at ChemRxiv, (López-Cortés et al., 2020a). Additionally, this work was supported by a) the Latin American Society of Pharmacogenomics and Personalized Medicine (SOLFAGEM), and b) the Consolidation and Structuring of Competitive Research Units - Competitive Reference Groups (ED431C 2018/49), funded by the Ministry of Education, University and Vocational Training of the Xunta de Galicia endowed with EU FEDER funds.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2021.598925/full#supplementary-material>.

REFERENCES

- Adhikari, S. P., Meng, S., Wu, Y. J., Mao, Y. P., Ye, R. X., Wang, Q. Z., et al. (2020). Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. *Infect. Dis. Poverty* 9, 29. doi:10.1186/s40249-020-00646-x
- Ahmed-Hassan, H., Sisson, B., Shukla, R. K., Wijewantha, Y., Funderburg, N. T., Li, Z., et al. (2020). Innate immune responses to highly pathogenic coronaviruses and other significant respiratory viral infections. *Front. Immunol.* 11, 1979. doi:10.3389/fimmu.2020.01979
- Akhmerov, A., and Marbán, E. (2020). COVID-19 and the heart. *Circ. Res.* 126, 1443–1455. doi:10.1161/CIRCRESAHA.120.317055
- Algassim, A. A., Elghazaly, A. A., Alnahdi, A. S., Mohammed-Rahim, O. M., Alanazi, A. G., Aldhuwayhi, N. A., et al. (2020). Prognostic significance of

- hemoglobin level and autoimmune hemolytic anemia in SARS-CoV-2 infection. *Ann. Hematol.* [Epub ahead of print]. doi:10.1007/s00277-020-04256-3
- Alsamman, A., and Zayed, H. (2020). The transcriptomic profiling of COVID-19 compared to SARS MERS, Ebola, and H1N1. *bioRxiv*. doi:10.1101/2020.05.06.080960
- Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi:10.1038/s41591-020-0820-9
- Aschenbrenner, A., Mouktaroudi, M., Kraemer, B., Antonakos, N., Oestreich, M., Gkizeli, K., et al. (2020). Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. *Genome Med.* 13, 7. doi:10.1186/s13073-020-00823-5
- Ballestar, E., Farber, D. L., Glover, S., Horwitz, B., Meyer, K., Nikolić, M., et al. (2020). Single cell profiling of COVID-19 patients: an international data resource from multiple tissues. *medRxiv*. doi:10.1101/2020.11.20.20227355
- Basu, A., Sarkar, A., and Maulik, U. (2020). Study of cell to cell transmission of SARS CoV 2 virus particle using gene network from microarray data. *bioRxiv*. doi:10.1101/2020.05.26.116780
- Bataille, S., Pedinielli, N., and Bergougnoux, J.-P. (2020). Could ferritin help the screening for COVID-19 in hemodialysis patients? *Kidney Int.* 98, 235–236. doi:10.1016/j.kint.2020.04.017
- Berlin, D. A., Gulick, R. M., and Martinez, F. J. (2020). Severe Covid-19. *N. Engl. J. Med.* 383, 2451–2460. doi:10.1056/NEJMc2009575
- Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Möller, R., et al. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* 181, 1036–1045.e9. doi:10.1016/j.cell.2020.04.026
- Bouhaddou, M., Memon, D., Meyer, B., White, K. M., Rezeli, V. V., Correa Marrero, M., et al. (2020). The global phosphorylation landscape of SARS-CoV-2 infection. *Cell* 182, 685–712.e19. doi:10.1016/j.cell.2020.06.034
- Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., et al. (2013). InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation. *Nucleic Acids Res.* 41, D1228–D1233. doi:10.1093/nar/gks1147
- Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* 21, 630–644. doi:10.1038/s41576-020-0258-4
- Cabrera-andrade, A. (2020). Gene prioritization through consensus strategy, enrichment methodologies analysis, and networking for osteosarcoma pathogenesis. *Int. J. Mol. Sci.* 21, 1053. doi:10.3390/ijms21031053
- Cabrera-Andrade, A., López-Cortés, A., Jaramillo-Koupermann, G., González-Díaz, H., Pazos, A., Munteanu, C. R., et al. (2020). A multi-objective approach for anti-osteosarcoma cancer agents discovery through drug repurposing. *Pharmaceuticals* 13, 409. doi:10.3390/ph13110409
- Calderwood, M. A., Venkatesan, K., Xing, L., Chase, M. R., Vazquez, A., Holthaus, A. M., et al. (2007). Epstein-Barr virus and virus human protein interaction maps. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7606–7611. doi:10.1073/pnas.070232104
- Cao, Y., Li, L., Feng, Z., Wan, S., Huang, P., Sun, X., et al. (2020). Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* 6, 11. doi:10.1038/s41421-020-0147-1
- Channappanavar, R., Fehr, A. R., Zheng, J., Wohlford-Lenane, C., Abrahante, J. E., Mack, M., et al. (2019). IFN-I response timing relative to virus replication determines MERS coronavirus infection outcomes. *J. Clin. Invest.* 129, 3625–3639. doi:10.1172/JCI126363
- Chen, Z., and John Wherry, E. (2020). T cell responses in patients with COVID-19. *Nat. Rev. Immunol.* 20, 529–536. doi:10.1038/s41577-020-0402-6
- Chen, X., Zhao, B., Qu, Y., Chen, Y., Xiong, J., Feng, Y., et al. (2020). Detectable serum SARS-CoV-2 viral load (RNAemia) is closely correlated with drastically elevated interleukin 6 (IL-6) level in critically ill COVID-19 patients. *Clin. Infect. Dis.* 71, 1937–1942. doi:10.1093/cid/ciaa449
- Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., et al. (2020). COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* 38, 970–979. doi:10.1038/s41587-020-0602-4
- Daniloski, Z., Jordan, T. X., Wessels, H.-H., Hoagland, D. A., Kasela, S., Legut, M., et al. (2020). Identification of required host factors for SARS-CoV-2 infection in human cells. *Cell* 184, 92–105.e16. doi:10.1016/j.cell.2020.10.030
- De Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugué, S., et al. (2008). Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 4, 230. doi:10.1038/msb.2008.66
- Deng, Y. Y., Zheng, Y., Cai, G. Y., Chen, X. M., and Hong, Q. (2020). Single-cell RNA sequencing data suggest a role for angiotensin-converting enzyme 2 in kidney impairment in patients infected with 2019-nCoV. *Chin. Med. J.* 133, 1129–1131. doi:10.1097/CM9.0000000000000783
- Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G., and Conticello, S. G. (2020). Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* 6, eabb5813. doi:10.1126/sciadv.abb5813
- Diao, B., Wang, C., Tan, Y., Chen, X., Liu, Y., Ning, L., et al. (2020). Reduction and functional exhaustion of T cells in patients with Coronavirus Disease 2019 (COVID-19). *medRxiv*. doi:10.1101/2020.02.18.20024364
- Doncheva, N. T., Morris, J. H., Gorodkin, J., and Jensen, L. J. (2019). Cytoscape StringApp: network analysis and visualization of proteomics data. *J. Proteome Res.* 18, 623–632. doi:10.1021/acs.jproteome.8b00702
- Donoghue, M., Hsieh, F., Baronas, E., Godbout, K., Gosselin, M., Stagliano, N., et al. (2000). A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1-9. *Circ. Res.* 87, E1–E9. doi:10.1161/01.res.87.5.e1
- El Zowalaty, M. E., and Järhult, J. D. (2020). From SARS to COVID-19: a previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans – Call for a One Health approach. *One Health* 9, 100124. doi:10.1016/j.onehlt.2020.100124
- Fajgenbaum, D. C., and June, C. H. (2020). Cytokine storm. *N. Engl. J. Med.* 383, 2255–2273. doi:10.1056/NEJMra2026131
- Fogarty, H., Townsend, L., Ni Cheallaigh, C., Bergin, C., Martin-Loeches, I., Browne, P., et al. (2020). COVID-19 coagulopathy in caucasian patients. *Br. J. Haematol.* 189, 1044–1049. doi:10.1111/bjh.16749
- Fu, S., Fu, X., Song, Y., Li, M., Pan, P., Tang, T., et al. (2020). Virologic and clinical characteristics for prognosis of severe COVID-19: a retrospective observational study in Wuhan, China. *medRxiv*. doi:10.1101/2020.04.03.20051763
- Fujii, M., Matano, M., Toshimitsu, K., Takano, A., Mikami, Y., Nishikori, S., et al. (2018). Human intestinal organoids maintain self-renewal capacity and cellular diversity in Niche-Inspired culture condition. *Cell Stem Cell* 23, 787–793.e6. doi:10.1016/j.stem.2018.11.016
- Gao, L., Jiang, D., Wen, X. S., Cheng, X. C., Sun, M., He, B., et al. (2020). Prognostic value of NT-proBNP in patients with severe COVID-19. *Respir. Res.* 21, 83. doi:10.1186/s12931-020-01352-w
- Gao, Q., Bao, L., Mao, H., Wang, L., Xu, K., Yang, M., et al. (2020). Rapid development of an inactivated vaccine for SARS-CoV-2. *bioRxiv*. doi:10.1101/2020.04.17.046375
- Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., et al. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 80, eabb7498. doi:10.1126/science.abb7498
- Garvin, M. R., Alvarez, C., Miller, J. I., Prates, E. T., Walker, A. M., Amos, B. K., et al. (2020). A mechanistic model and therapeutic interventions for covid-19 involving a ras-mediated bradykinin storm. *Elife* 9, e59177. doi:10.7554/eLife.59177
- Gaulton, A., Hersey, A., Nowotka, M. L., Patricia Bento, A., Chambers, J., Mendez, D., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.* 45, D945–D954. doi:10.1093/nar/gkw1074
- Gawel, D. R., Serra-Musach, J., Lilja, S., Aagesen, J., Arenas, A., Asking, B., et al. (2019). A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. *Genome Med.* 11, 47. doi:10.1186/s13073-019-0657-3
- Giudicelli, V., Chaume, D., and Lefranc, M. P. (2005). IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33, D256–D261. doi:10.1093/nar/gki010
- Gong, J., Dong, H., Xia, S. Q., Huang, Y. Z., Wang, D., Zhao, Y., et al. (2020). Correlation analysis between Disease severity and inflammation-related parameters in patients with COVID-19 pneumonia. *medRxiv*. doi:10.1101/2020.02.25.20025643

- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468. doi:10.1038/s41586-020-2286-9
- Guilliams, M., Lambrecht, B. N., and Hammad, H. (2013). Division of labor between lung dendritic cells and macrophages in the defense against pulmonary infections. *Mucosal Immunol.* 6, 464–473. doi:10.1038/mi.2013.14
- Gulbahce, N., Yan, H., Dricot, A., Padi, M., Byrdson, D., Franchi, R., et al. (2012). Viral perturbations of host networks reflect disease etiology. *PLoS Comput. Biol.* 8, e1002531. doi:10.1371/journal.pcbi.1002531
- Gupta, A., Madhavan, M. V., Sehgal, K., Nair, N., Mahajan, S., Sehrawat, T. S., et al. (2020). Extrapulmonary manifestations of COVID-19. *Nat. Med.* 26, 1017–1032. doi:10.1038/s41591-020-0968-3
- Guy, R. K., DiPaola, R. S., Romanelli, F., and Dutch, R. E. (2020). Rapid repurposing of drugs for COVID-19. *Science* 368, 829–830. doi:10.1126/science.abb9332
- Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Pere, H., et al. (2020). Impaired type I interferon activity and exacerbated inflammatory responses in severe Covid-19 patients. *medRxiv*. doi:10.1101/2020.04.19.20068015
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning the elements of statistical learning*Data mining, inference, and prediction. 2nd Edn New York: Springer-Verlag.
- Herold, T., Jurinovic, V., Arnreich, C., Hellmuth, J. C., Bergwelt-Baildon, M., Klein, M., et al. (2020). Level of IL-6 predicts respiratory failure in hospitalized symptomatic COVID-19 patients. *medRxiv*. doi:10.1101/2020.04.01.20047381
- Hoffmann, M., Kleine-Weber, H., Schroeder, S., Mü, M. A., Drosten, C., Pö, S., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor article SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 181, 271–280.e8. doi:10.1016/j.cell.2020.02.052
- Hu, Z.-J., Xu, J., Yin, J.-M., Li, L., Hou, W., Zhang, L.-L., et al. (2020). Lower circulating interferon-gamma is a risk factor for lung fibrosis in COVID-19 patients. *Front. Immunol.* doi:10.3389/fimmu.2020.585647
- Huang, A. T., Garcia-Carreras, B., Hitchings, M. D. T., Yang, B., Katzelnick, L., Rattigan, S. M., et al. (2020). A systematic review of antibody mediated immunity to coronaviruses: antibody kinetics, correlates of protection, and association of antibody responses with severity of disease. *medRxiv*. doi:10.1101/2020.04.14.20065771
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi:10.1093/nar/gkn923
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211
- Hussman, J. P. (2020). Cellular and molecular pathways of COVID-19 and potential points of therapeutic intervention. *Front. Pharmacol.* 11, 1169. doi:10.3389/fphar.2020.01169
- Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* 545, 505–509. doi:10.1038/nature22366
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi:10.1093/nar/gkz1031
- Ji, W., Bishnu, G., Cai, Z., and Shen, X. (2020). Analysis clinical features of COVID-19 infection in secondary epidemic area and report potential biomarkers in evaluation. *medRxiv*. doi:10.1101/2020.03.10.20033613
- Jose, R. J., and Manuel, A. (2020). COVID-19 cytokine storm: the interplay between inflammation and coagulation. *Lancet Respir. Med.* 8, e46–e47. doi:10.1016/S2213-2600(20)30216-2
- Ju, B., Zhang, Q., Ge, X., Wang, R., Yu, J., Shan, S., et al. (2020). Potent human neutralizing antibodies elicited by SARS-CoV-2 infection. *bioRxiv*. doi:10.1101/2020.03.21.990770
- Kirchdoerfer, R. N., Cottrell, C. A., Wang, N., Pallesen, J., Yassine, H. M., Turner, H. L., et al. (2016). Pre-fusion structure of a human coronavirus spike protein. *Nature* 531, 118–121. doi:10.1038/nature17200
- Knauf, S., Abel, L., and Hallmaier-Wacker, L. K. (2019). The nagoya protocol and research on emerging infectious diseases. *Bull. World Health Organ.* 97, 379. doi:10.2471/BLT.19.232173
- Kobayashi, M., Oshima, S., Maeyashiki, C., Nibe, Y., Otsubo, K., Matsuzawa, Y., et al. (2016). The ubiquitin hybrid gene UBA52 regulates ubiquitination of ribosome and sustains embryonic development. *Sci. Rep.* 6, 36780. doi:10.1038/srep36780
- Kong, Y., Han, J., Wu, X., Zeng, H., Liu, J., and Zhang, H. (2020). VEGF-D: a novel biomarker for detection of COVID-19 progression. *Crit. Care* 24, 373. doi:10.1186/s13054-020-03079-y
- Kost-Alimova, M., Sidhom, E.-H., Satyam, A., Chamberlain, B. T., Dvela-Levitt, M., Melanson, M., et al. (2020). A high-content screen for mucin-1-reducing compounds identifies fostamatinib as a candidate for rapid repurposing for acute lung injury. *Cell Rep. Med.* 1, 100137. doi:10.1016/j.xcrm.2020.100137
- Kuchi Bhotla, H., Kaul, T., Balasubramanian, B., Easwaran, M., Arumugam, V. A., Pappusamy, M., et al. (2020). Platelets to surrogate lung inflammation in COVID-19 patients. *Med. Hypotheses* 143, 110098. doi:10.1016/j.mehy.2020.110098
- Kumar, N., Mishra, B., Mehmood, A., Athar, M., and Mukhtar, M. S. (2020). Integrative network biology framework elucidates molecular mechanisms of SARS-CoV-2 pathogenesis. *iScience* 23, 101526. doi:10.1016/j.isci.2020.101526
- Lamers, M. M., Beumer, J., van der Vaart, J., Knoops, K., Puschhof, J., Breugem, T. I., et al. (2020). SARS-CoV-2 productively infects human gut enterocytes. *Science* 369, eabc1669. doi:10.1126/science.abc1669
- Law, J. N., Akers, K., Tasnina, N., Della Santina, C. M., Kshirsagar, M., Klein-Seetharaman, J., et al. (2020). Identifying human interactors of SARS-CoV-2 proteins and drug targets for COVID-19 using network-based label propagation. *arXiv*. arXiv:2006.01968v2.
- Lee, S., Channappanavar, R., and Kanneganti, T.-D. (2020). Coronaviruses: innate immunity, inflammasome activation, inflammatory Cell Death, and Cytokines. *Trends Immunol.* 41, 1083–1099. doi:10.1016/j.it.2020.10.005
- Lefranc, M. P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., et al. (2015). IMGT R, the international ImMunoGeneTics information system R 25 years on. *Nucleic Acids Res.* 43, D413–D422. doi:10.1093/nar/gku1056
- Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., et al. (2009). IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Res.* 37, D1006–D1012. doi:10.1093/nar/gkn838
- Li, H., Liu, L., Zhang, D., Xu, J., Dai, H., Tang, N., et al. (2020). SARS-CoV-2 and viral sepsis: observations and hypotheses. *Lancet* 395, P1517–P1520. doi:10.1016/S0140-6736(20)30920-X
- Li, X., Geng, M., Peng, Y., Meng, L., and Lu, S. (2020). Molecular immune pathogenesis and diagnosis of COVID-19. *J. Pharm. Anal.* 10, 102–108. doi:10.1016/j.jpha.2020.03.001
- Liao, D., Zhou, F., Luo, L., Xu, M., Wang, H., Xia, J., et al. (2020). Haematological characteristics and risk factors in the classification and prognosis evaluation of COVID-19: a retrospective cohort study. *Lancet Haematol.* 7, E671–E678. doi:10.1016/S2352-3026(20)30217-9
- Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). The landscape of lung bronchoalveolar immune cells in COVID-19 revealed by single-cell RNA sequencing. *medRxiv*. doi:10.1101/2020.02.23.20026690
- Liu, J., Li, S., Liu, J., Liang, B., Wang, X., Wang, H., et al. (2020). Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine* 55, 102763. doi:10.1016/j.ebiom.2020.102763
- Liu, J., Liu, Y., Xiang, P., Pu, L., Xiong, H., Li, C., et al. (2020). Neutrophil-to-lymphocyte ratio predicts severe illness patients with 2019 novel Coronavirus in the early stage. *medRxiv*. doi:10.1101/2020.02.10.20021584
- López-Cortés, A., Guevara-Ramírez, P., Kyriakidis, N. C., Barba-Ostria, C., León Cáceres, Á., Guerrero, S., et al. (2020a). Silico analyses of immune system protein interactome network, single-cell RNA sequencing of human tissues, and artificial neural networks reveal potential therapeutic targets for drug repurposing against COVID-19. *chemRxiv*. doi:10.26434/chemrxiv.12408074.v1
- López-Cortés, A., Paz-y-Miño, C., Cabrera-Andrade, A., Barigye, S. J., Munteanu, C. R., González-Díaz, H., et al. (2018). Gene prioritization, communality analysis, networking and metabolic integrated pathway to better understand breast cancer pathogenesis. *Sci. Rep.* 8, 16679. doi:10.1038/s41598-018-35149-1
- López-Cortés, A., Paz-y-Miño, C., Guerrero, S., Cabrera-Andrade, A., Barigye, S. J., Munteanu, C. R., et al. (2020b). OncoOmics approaches to reveal essential

- genes in breast cancer: a panoramic view from pathogenesis to precision medicine. *Sci. Rep.* 10, 5285. doi:10.1038/s41598-020-62279-2
- Mahdi, M., Mótýán, J. A., Szojka, Z. I., Golda, M., Miczi, M., and Tózsér, J. (2020). Analysis of the efficacy of HIV protease inhibitors against SARS-CoV-2's main protease. *Virology* 17, 190. doi:10.21203/rs.3.rs-40776/v1
- Manjili, R. H., Zarei, M., Habibi, M., and Manjili, M. H. (2020). COVID-19 as an acute inflammatory disease. *J. Immunol.* 205, 12–19. doi:10.4049/jimmunol.2000413
- Mao, L., Jin, H., Wang, M., Hu, Y., Chen, S., He, Q., et al. (2020). Neurologic manifestations of hospitalized patients with Coronavirus Disease 2019 in Wuhan, China. *JAMA Neurol.* 77, 683–690. doi:10.1001/jamaneurol.2020.1127
- McKechnie, J. L., and Blish, C. A. (2020). The innate immune system: fighting on the front lines or fanning the flames of COVID-19? *Cell Host Microbe* 27, 863–869. doi:10.1016/j.chom.2020.05.009
- Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S., and Manson, J. J. (2020). COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 395, P1033–P1034. doi:10.1016/S0140-6736(20)30628-0
- Merad, M., and Martin, J. C. (2020). Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. *Nat. Rev. Immunol.* 20, 355–362. doi:10.1038/s41577-020-0331-4
- Micholas, S., and Jeremy, C. S. (2020). Repurposing therapeutics for COVID-19: supercomputer-based docking to the SARS-CoV-2 viral spike protein and viral spike protein-human ACE2 interface. *chemRxiv*. doi:10.26434/chemrxiv.11871402.v4
- Montenegro, F., Unigarro, L., Paredes, G., Moya, T., Romero, A., Torres, L., et al. (2020). Acute respiratory distress syndrome (ARDS) caused by the novel coronavirus disease (COVID-19): a practical comprehensive literature review. *Expert Rev. Respir. Med.* 15, 183–195. doi:10.1080/17476348.2020.1820329
- Muus, C., Luecken, M. D., Eraslan, G., Waghay, A., Heimberg, G., Sikkema, L., et al. (2020). Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of SARS-CoV-2 viral entry and highlights inflammatory programs in putative target cells. *bioRxiv*. doi:10.1101/2020.04.19.049254
- Nie, S., Zhao, X., Zhao, K., Zhang, Z., Zhang, Z., and Zhang, Z. (2020). Metabolic disturbances and inflammatory dysfunction predict severity of coronavirus disease 2019 (COVID-19): a retrospective study. *medRxiv*. doi:10.1101/2020.03.24.20042283
- Oberfeld, B., Achanta, A., Carpenter, K., Chen, P., Gilette, N. M., Langat, P., et al. (2020). SnapShot: COVID-19. *Cell* 181, 954–954.e1. doi:10.1016/j.cell.2020.04.013
- Oliver, J. (2013). *Deep learning for the life sciences applying deep learning to genomics, microscopy, drug discovery, and more*. Sebastopol, CA: O'Reilly Media
- Ordovas-Montanes, J., Dwyer, D. F., Nyquist, S. K., Buchheit, K. M., Vukovic, M., Deb, C., et al. (2018). Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* 560, 649–654. doi:10.1038/s41586-018-0449-8
- Ortiz-Prado, E., Simbaña-Rivera, K., Gómez-Barreno, L., Rubio-Neira, M., Guaman, L. P., Kyriakidis, N. C., et al. (2020). Clinical, molecular and epidemiological characterization of the SARS-CoV2 virus and the Coronavirus disease 2019 (COVID-19), a comprehensive literature review. *Diagn. Microbiol. Infect. Dis.* 98, 115094. doi:10.1016/j.diagmicrobio.2020.115094
- Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., et al. (2020). Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst.* 12, 23–40.e7. doi:10.1016/j.cels.2020.10.003
- Pan, A., Lahiri, C., Rajendiran, A., and Shanmugham, B. (2016). Computational analysis of protein interaction networks for infectious diseases. *Brief. Bioinform.* 17, 517–526. doi:10.1093/bib/bbv059
- Park, J. H., and Lee, H. K. (2020). Re-analysis of single cell transcriptome reveals that the NR3C1-CXCL8-neutrophil axis determines the severity of COVID-19. *Front. Immunol.* 11, 2145. doi:10.3389/fimmu.2020.02145
- Parvathaneni, V., and Gupta, V. (2020). Utilizing drug repurposing against COVID-19 – efficacy, limitations, and challenges. *Life Sci.* 259, 118275. doi:10.1016/j.lfs.2020.118275
- Peiris, J. S. M., Guan, Y., and Yuen, K. Y. (2004). Severe acute respiratory syndrome. *Nat. Med.* 349, 2431–2441. doi:10.1038/nm1143
- Perrin-Cocon, L., Diaz, O., Jacquemin, C., Barthel, V., Ogire, E., Ramière, C., et al. (2020). The current landscape of coronavirus-host protein-protein interactions. *J. Transl. Med.* 18, 319. doi:10.1186/s12967-020-02480-z
- Phimister, E. G., Parks, J. M., and Smith, J. C. (2020). How to discover antiviral drugs quickly. 382, 2261–2264. doi:10.1056/NEJMcibr2007042
- Prokop, J. W., Shankar, R., Gupta, R., Leimanis, M. L., Nedveck, D., Uhl, K., et al. (2020). Virus-induced genetics revealed by multidimensional precision medicine transcriptional workflow applicable to COVID-19. *Physiol. Genomics* 52, 255–268. doi:10.1152/physiolgenomics.00045.2020
- Qu, R., Ling, Y., Zhang, Y. hui. zhi., Wei, L. ya., Chen, X., Li, X. mian., et al. (2020). Platelet-to-lymphocyte ratio is associated with prognosis in patients with coronavirus disease-19. *J. Med. Virol.* 92, 1533–1541. doi:10.1002/jmv.25767
- Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198. doi:10.1093/nar/gkz369
- Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* 14, 482–517. doi:10.1038/s41596-018-0103-9
- Rentsch, C., Kidwai-Khan, F., Tate, J., Park, L., King, J., Skanderson, M., et al. (2020). Covid-19 testing, Hospital Admission, and intensive care among 2,026,227 United States Veterans aged 54–75 Years. *medRxiv*. doi:10.1101/2020.04.09.20059964
- Rhee, C., Chiotos, K., Cosgrove, S. E., Heil, E. L., Kadri, S. S., Kalil, A. C., et al. (2020). Infectious Diseases society of America position paper: recommended revisions to the national severe sepsis and septic shock early management Bundle (SEP-1) sepsis quality measure. *Clin. Infect. Dis. ciao059* doi:10.1093/cid/cia059
- Røttingen, J. A., Chamas, C., Goyal, L. C., Harb, H., Lagrader, L., and Mayosi, B. M. (2012). Securing the public good of health research and development for developing countries. *Bull. World Health Organ.* 90, 398–400. doi:10.2471/BLT.12.105460
- Rotzinger, D. C., Beigelman-Aubry, C., von Garnier, C., and Qanadli, S. D. (2020). Pulmonary embolism in patients with COVID-19: time to change the paradigm of computed tomography. *Thromb. Res.* 190, 58–59. doi:10.1016/j.thromres.2020.04.011
- Saini, K. S., Lanza, C., Romano, M., de Azambuja, E., Cortes, J., de las Heras, B., et al. (2020). Repurposing anticancer drugs for COVID-19-induced inflammation, immune dysfunction, and coagulopathy. *Br. J. Cancer* 123, 694–697. doi:10.1038/s41416-020-0948-x
- Sanders, J. M., Monogue, M. L., Jodlowski, T. Z., and Cutrell, J. B. (2020). Pharmacologic treatments for Coronavirus Disease 2019 (COVID-19): a review. *JAMA – J. Am. Med. Assoc.* 323, 1824–1836. doi:10.1001/jama.2020.6019
- Sarzi-Puttini, P., Giorgi, V., Sirotti, S., Marotto, D., Ardizzone, S., Rizzardini, G., et al. (2020). COVID-19, cytokines and immunosuppression: what can we learn from severe acute respiratory syndrome?. *Clin. Exp. Rheumatol.* 38, 337–342.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303
- Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., et al. (2020). Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell* 182, 59–72.e15. doi:10.1016/j.cell.2020.05.032
- Shi, C. S., Nabar, N. R., Huang, N. N., and Kehrl, J. H. (2019). SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes. *Cell Death Discov.* 5, 101. doi:10.1038/s41420-019-0181-7
- Sidarta-Oliveira, D., Jara, C. P., Ferruzzi, A. J., Skaf, M. S., Velander, W. H., Araujo, E. P., et al. (2020). SARS-CoV-2 receptor is co-expressed with elements of the kinin-kallikrein, renin-angiotensin and coagulation systems in alveolar cells. *Sci. Rep.* 10, 19522. doi:10.1038/s41598-020-76488-2
- Singh, M., Bansal, V., and Feschotte, C. (2020). A single-cell RNA expression map of human coronavirus entry factors. *bioRxiv*. doi:10.1101/2020.05.08.084806
- Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics

- to other omics research. *Nucleic Acids Res.* 46, D661–D667. doi:10.1093/nar/gkx1064
- Song, C.-Y., Xu, J., He, J.-Q., and Lu, Y.-Q. (2020). COVID-19 early warning score: a multi-parameter screening tool to identify highly suspected patients. *medRxiv*. doi:10.1101/2020.03.05.20031906
- Spiezia, L., Boscolo, A., Poletto, F., Cerruti, L., Tiberio, I., Campello, E., et al. (2020). COVID-19-related severe hypercoagulability in patients admitted to intensive care unit for acute respiratory failure. *Thromb. Haemost.* 120, 998–1000. doi:10.1055/s-0040-1710018
- Sungnak, W., Huang, N., Bécavin, C., Berg, M., Queen, R., Litvinukova, M., et al. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* 26, 681–687. doi:10.1038/s41591-020-0868-6
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi:10.1093/nar/gku1003
- Tabassum, N., Zhang, H., and Stebbing, J. (2020). Repurposing fostamatinib to combat SARS-CoV-2-induced acute lung injury. *Cell Rep. Med.* 1, 100145. doi:10.1016/j.xcrm.2020.100145
- Tan, L., Wang, Q., Zhang, D., Ding, J., Huang, Q., Tang, Y. Q., et al. (2020). Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal Transduct. Target. Ther.* 5, 33. doi:10.1038/s41392-020-0148-4
- Tang, N., Li, D., Wang, X., and Sun, Z. (2020). Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. *J. Thromb. Haemost.* 18, 844–847. doi:10.1111/jth.14768
- Tang, Y., Li, M., Wang, J., Pan, Y., and Wu, F. X. (2015). CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *BioSystems* 127, 67–72. doi:10.1016/j.biosystems.2014.11.005
- Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A., and Ng, L. F. P. (2020). The trinity of COVID-19: immunity, inflammation and intervention. *Nat. Rev. Immunol.* 20, 363–374. doi:10.1038/s41577-020-0311-8
- Tejera, E., Munteanu, C. R., López-Cortés, A., Cabrera-Andrade, A., and Pérez-Castillo, Y. (2020). Drugs repurposing using QSAR, docking and molecular dynamics for possible inhibitors of the SARS-CoV-2 Mpro protease. *Molecules* 25, 5172. doi:10.3390/molecules25215172
- Terpos, E., Ntanasis-Stathopoulos, I., Elalamy, I., Kastritis, E., Sergentanis, T. N., Politou, M., et al. (2020). Hematological findings and complications of COVID-19. *Am. J. Hematol.* 95, 834–847. doi:10.1002/ajh.25829
- The Lancet (2020). Emerging understandings of 2019-nCoV. *Lancet* 395, 311. doi:10.1016/S0140-6736(20)30186-0
- Vastrad, B., Vastrad, C., and Tengli, A. (2020). Identification of potential mRNA panels for severe acute respiratory syndrome coronavirus 2 (COVID-19) diagnosis and treatment using microarray dataset and bioinformatics methods. *Biotech* 10, 422. doi:10.1007/s13205-020-02406-y
- Vidal, M., Cusick, M. E., and Barabási, A. L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi:10.1016/j.cell.2011.02.016
- Vinayagam, S., and Sattu, K. (2020). SARS-CoV-2 and coagulation disorders in different organs. *Life Sci.* 260, 118431. doi:10.1016/j.lfs.2020.118431
- Vivier, E., Artis, D., Colonna, M., Diefenbach, A., Di Santo, J. P., Eberl, G., et al. (2018). Innate lymphoid cells: 10 years on. *Cell* 174, 1054–1066. doi:10.1016/j.cell.2018.07.017
- Wadman, M., Couzin-Frankel, J., Kaiser, J., and Maticic, C. (2020). A rampage through the body. *Science* 368, 356–360. doi:10.1126/science.368.6489.356
- Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181, 281–292.e6. doi:10.1016/j.cell.2020.02.058
- Wan, S., Yi, Q., Fan, S., Lv, J., Zhang, X., Guo, L., et al. (2020). Characteristics of lymphocyte subsets and cytokines in peripheral blood of 123 hospitalized patients with 2019 novel coronavirus pneumonia (NCP). *medRxiv*. doi:10.1101/2020.02.10.20021832
- Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA - J. Am. Med. Assoc.* 323, 1061–1069. doi:10.1001/jama.2020.1585
- Wang, F., Lu, J., Peng, X., Wang, J., Liu, X., Chen, X., et al. (2016). Integrated analysis of microRNA regulatory network in nasopharyngeal carcinoma with deep sequencing. *J. Exp. Clin. Cancer Res.* 35, 17. doi:10.1186/s13046-016-0292-4
- Wang, J., Li, Q., Yin, Y., Zhang, Y., Cao, Y., Lin, X., et al. (2020). Excessive neutrophils and neutrophil extracellular traps in COVID-19. *Front. Immunol.* 11, 2063. doi:10.3389/fimmu.2020.02063
- Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., et al. (2020). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 181, 894–904.e9. doi:10.1016/j.cell.2020.03.045
- Wang, Y., Liu, M., and Gao, J. (2020). Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proc. Natl. Acad. Sci. U.S.A.* 117, 13967–13974. doi:10.1073/pnas.2008209117
- Wen, W., Su, W., Tang, H., Le, W., Zhang, X., Zheng, Y., et al. (2020). Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discov.* 6, 431. doi:10.1038/s41421-020-0168-9
- Winn, B. J. (2020). Is there a role for insulin-like growth factor inhibition in the treatment of COVID-19-related adult respiratory distress syndrome?. *Med. Hypotheses* 144, 110167. doi:10.1016/j.mehy.2020.110167
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi:10.1093/nar/gkx1037
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672. doi:10.1093/nar/gkj067
- Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C. L., Abiona, O., et al. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260–1263. doi:10.1126/science.aax0902
- Wu, A. Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* 27, 325–328. doi:10.1016/j.chom.2020.02.001
- Wu, C. Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., et al. (2020). Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B* 10, 766–788. doi:10.1016/j.apsb.2020.02.008
- Wu, F. Wang, A., Liu, M., Wang, Q., Chen, J., Xia, S., et al. (2020). Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and their implications. *medRxiv* doi:10.2139/ssrn.3566211
- Wu, M. Chen, Y., Xia, H., Wang, C., Tan, C. Y., Cai, X., et al. (2020). Transcriptional and proteomic insights into the host response in fatal COVID-19 cases. *Proc. Natl. Acad. Sci. U.S.A.* 117, 28336–28343. doi:10.1073/pnas.2018030117
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530. doi:10.1039/c7sc02664a
- Xiang, J., Wen, J., Yuan, X., Xiong, S., Zhou, X., Liu, C., et al. (2020). Potential biochemical markers to identify severe cases among COVID-19 patients. *medRxiv*. doi:10.1101/2020.03.19.20034447
- Yan, Q. (2017). *Translational bioinformatics and systems biology methods for personalized medicine* Cambridge, United States: Academic Press
- Yan, R., Zhang, Y., Guo, Y., Xia, L., and Zhou, Q. (2020). Structural basis for the recognition of the 2019-nCoV by human ACE2. *bioRxiv*. doi:10.1101/2020.02.19.956946
- Yang, L., Liu, S., Liu, J., Zhang, Z., Wan, X., Huang, B., et al. (2020). COVID-19: immunopathogenesis and Immunotherapeutics. *Signal Transduct. Target. Ther.* 5, 128. doi:10.1038/s41392-020-00243-2
- Yang, W., Cao, Q., Qin, L., Wang, X., Cheng, Z., Pan, A., et al. (2020). Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19): A multi-center study in Wenzhou city, Zhejiang, China. *J. Infect.* 80, 388–393. doi:10.1016/j.jinf.2020.02.016
- Yang, X., Kui, L., Tang, M., Li, D., Wei, K., Chen, W., et al. (2020). High-throughput transcriptome profiling in Drug and biomarker Discovery. *Front. Genet.* 11, 19. doi:10.3389/fgene.2020.00019
- Yang, Y., Shen, C., Li, J., Yuan, J., Yang, M., Wang, F., et al. (2020). Exuberant elevation of IP-10, MCP-3 and IL-1ra during SARS-CoV-2 infection is associated with disease severity and fatal outcome. *medRxiv*. doi:10.1101/2020.03.02.20029975
- Yao, X.-H., He, Z.-C., Li, T.-Y., Zhang, H.-R., Wang, Y., Mou, H., et al. (2020). Pathological evidence for residual SARS-CoV-2 in pulmonary tissues of a ready-for-discharge patient. *Cell Res.* 30, 541–543. doi:10.1038/s41422-020-0318-5

- Young, B. E., Ong, S. W. X., Kalimuddin, S., Low, J. G., Tan, S. Y., Loh, J., et al. (2020). Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. *JAMA - J. Am. Med. Assoc.* 323, 1488–1494. doi:10.1001/jama.2020.3204
- Yu, L., Tong, Y., Shen, G., Fu, A., Lai, Y., Zhou, X., et al. (2020). Immunodepletion with hypoxemia: a potential high risk subtype of coronavirus Disease 2019. *medRxiv*. doi:10.1101/2020.03.03.20030650
- Yu, Y., Tsang, J. C. H., Wang, C., Clare, S., Wang, J., Chen, X., et al. (2016). Single-cell RNA-seq identifies a PD-1hi ILC progenitor and defines its development pathway. *Nature* 539, 102–106. doi:10.1038/nature20105
- Yuan, Y., Cao, D., Zhang, Y., Ma, J., Qi, J., Wang, Q., et al. (2017). Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat. Commun.* 8, 15092. doi:10.1038/ncomms15092
- Zeng, Q., Li, Y., Huang, G., Wu, W., Dong, S., and Xu, Y. (2020). Mortality of COVID-19 is associated with cellular immune function compared to immune function in Chinese han population. *medRxiv*. doi:10.1101/2020.03.08.20031229
- Zhang, B., Zhou, X., Zhu, C., Feng, F., Qiu, Y., Feng, J., et al. (2020). Immune phenotyping based on neutrophil-to-lymphocyte ratio and IgG predicts disease severity and outcome for patients with COVID-19. *medRxiv*. doi:10.1101/2020.03.12.20035048
- Zhang, C., Shi, L., and Wang, F. S. (2020). Liver injury in COVID-19: management and challenges. *Lancet Gastroenterol. Hepatol.* 5, P428–P430. doi:10.1016/S2468-1253(20)30057-1
- Zhang, D., Guo, R., Lei, L., Liu, H., Wang, Y., Wang, Y., et al. (2020). COVID-19 infection induces readily detectable morphological and inflammation-related phenotypic changes in peripheral blood monocytes, the severity of which correlate with patient outcome. *medRxiv*. doi:10.1101/2020.03.24.20042655
- Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., et al. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. *Science* 368, 409–412. doi:10.1126/science.abb3405
- Zhong, J., Tang, J., Ye, C., and Dong, L. (2020). The immunology of COVID-19: is immune modulation an option for treatment?. *Lancet Rheumatol.* 2, E428–E436. doi:10.1016/S2665-9913(20)30120-X
- Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi:10.1038/s41586-020-2012-7
- Zhou, Y., Fu, B., Zheng, X., Wang, D., Zhao, C., qi, Y., et al. (2020a). Aberrant pathogenic GM-CSF+ T cells and inflammatory CD14+CD16+ monocytes in severe pulmonary syndrome patients of a new coronavirus. *bioRxiv*. doi:10.1101/2020.02.12.945576
- Zhou, Y., Yang, Z., Guo, Y., Geng, S., Gao, S., Ye, S., et al. (2020b). A new predictor of Disease severity in patients with COVID-19 in Wuhan, China. *medRxiv*. doi:10.1101/2020.03.24.20042119
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi:10.1056/nejmoa2001017
- Ziegler, C. G. K., Allon, S. J., Nyquist, S. K., Mbano, I. M., Miao, V. N., Tzouanas, C. N., et al. (2020). SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. *Cell* 181, 1016–1035.e19. doi:10.1016/j.cell.2020.04.035
- Zulfiqar, A.-A., Lorenzo-Villalba, N., Hassler, P., and Andr  s, E. (2020). Immune thrombocytopenic purpura in a patient with Covid-19. *N. Engl. J. Med.* 382, e43. doi:10.1056/nejmc2010472

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright   2021 L  pez-Cort  s, Guevara-Ram  rez, Kyriakidis, Barba-Ostria, Le  n C  ceres, Guerrero, Ortiz-Prado, Munteanu, Tejera, Cevallos-Robalino, G  mez-Jaramillo, Simba  na-Rivera, Granizo-Mart  nez, P  rez-M, Moreno, Garc  a-C  rdenas, Zambrano, P  rez-Castillo, Cabrera-Andrade, Puig San Andr  s, Proa  o-Castro, Bautista, Quevedo, Varela, Qui  nes and Paz-y-Mi  o. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Antioxidant Activity, Molecular Docking, Quantum Studies and *In Vivo* Antinociceptive Activity of Sulfonamides Derived From Carvacrol

Aldo S. de Oliveira^{1,2*}, Luana C. Llanes³, Ricardo J. Nunes⁴, Catharina Nucci-Martins^{5,6}, Anacleto S. de Souza², David L. Palomino-Salcedo², María J. Dávila-Rodríguez⁷, Leonardo L. G. Ferreira^{2*}, Adair R. S. Santos⁵ and Adriano D. Andricopulo²

¹Department of Exact Sciences and Education, Federal University of Santa Catarina-UFSC, Blumenau, Brazil, ²Laboratory of Medicinal and Computational Chemistry, Institute of Physics of São Carlos, University of São Paulo-USP, São Carlos, Brazil, ³Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, CA, United States, ⁴Department of Chemistry, Federal University of Santa Catarina-UFSC, Florianópolis, Brazil, ⁵Department of Physiological Sciences, Center of Biological Sciences, Federal University of Santa Catarina-UFSC, Florianópolis, Brazil, ⁶Department of Structural and Functional Biology, Institute of Biology, University of Campinas-UNICAMP, Campinas, Brazil, ⁷Department of Chemistry, Federal University of São Carlos-UFSCar, São Carlos, Brazil

OPEN ACCESS

Edited by:

Marco Ragusa,
University of Catania, Italy

Reviewed by:

Liu Shudong,
Agricultural University of Hebei, China
Mohaddeseh Abouhosseini Tabari,
Amol University of Special Modern
Technologies, Iran
Maria Grazia Morgese,
University of Foggia, Italy

*Correspondence:

Aldo S. de Oliveira
aldo.sena@ufsc.br
Leonardo L. G. Ferreira
leonardo@ifsc.usp.br

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 03 October 2021

Accepted: 05 November 2021

Published: 23 November 2021

Citation:

de Oliveira AS, Llanes LC, Nunes RJ,
Nucci-Martins C, de Souza AS,
Palomino-Salcedo DL,
Dávila-Rodríguez MJ, Ferreira LLG,
Santos ARS and Andricopulo AD
(2021) Antioxidant Activity, Molecular
Docking, Quantum Studies and *In Vivo*
Antinociceptive Activity of
Sulfonamides Derived From Carvacrol.
Front. Pharmacol. 12:788850.
doi: 10.3389/fphar.2021.788850

The synthesis and antioxidant, antinociceptive and antiedematogenic activities of sulfonamides derived from carvacrol—a druglike natural product—are reported. The compounds showed promising antioxidant activity, and sulfonamide derived from morpholine (**S1**) demonstrated excellent antinociceptive and antiedematogenic activities, with no sedation or motor impairment. The mechanism that underlies the carvacrol and derived sulfonamides' relieving effects on pain has not yet been fully elucidated, however, this study shows that the antinociceptive activity can be partially mediated by the antagonism of glutamatergic signaling. Compound **S1** presented promising efficacy and was predicted to have an appropriate medicinal chemistry profile. Thus, derivative **S1** is an interesting starting point for the design of new leads for the treatment of pain and associated inflammation and prooxidative conditions.

Keywords: sulfonamides, pain, carvacrol, molecular modeling, antioxidant

INTRODUCTION

Pain is a major sorrowful condition that affects children, adolescents (Guindon et al., 2007; Schmidt et al., 2010) and adults (Loeser and Treede, 2008) in several pathologies, including cancer (Ling et al., 2012). Pain can impair daily activities, diminish life quality, and cause significant psychological conditions (Rowlingson, 2000).

Pain is a clinically meaningful sign for the detection and evaluation of many diseases. Its perception is complex, involving two distinct components, an emotional and a physiological or sensorial component, called nociception (Tominaga et al., 2003). Animal models used for the evaluation of antinociceptive activity involve several nociceptive responses generated by chemical, mechanical or thermal stimuli (Silva et al., 2013).

Despite advances in the pharmacokinetics and pharmacodynamics of analgesic agents, their high toxicity is a determinant of conflicting clinical results due to the need for drug associations and

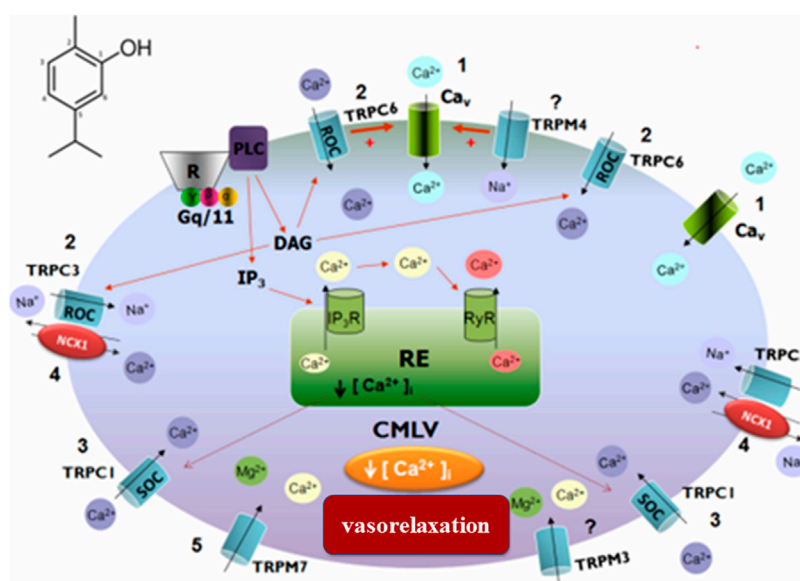


FIGURE 1 | Schematic representation of the probable signaling pathway of the vasorelaxant effect induced by carvacrol. 1) Blockage of calcium influx through the Cav; 2) Blockade of calcium influx through ROC and/or TRPC6; 3) Blockade of calcium influx through SOC and/or TRPC1; 4) Action on NCX1 by activation of TRPC3; 5) Inhibition of TRPM7.

interactions, especially in chronic pain due to its bioplasticity, and association with clinical conditions of anxiety and depression that reduce the quality of life of patient.

Sound evidence indicates that amino acids, mainly glutamate, found in C and Aδ fibers, play a fundamental role in the transmission of pain, as they provoke post-synaptic depolarization and the propagation of nociceptive information (Verri et al., 2006). Besides, abnormal excitability caused by inflammation or injury usually results from increased expression or activation of receptors, which may be stimulated by glutamate, favoring the maintenance of the painful stimulus (Rao, 2009; Salvemini et al., 2011). Therefore, substances capable of causing selective changes in glutamatergic signaling may give rise to new analgesic and anti-inflammatory agents.

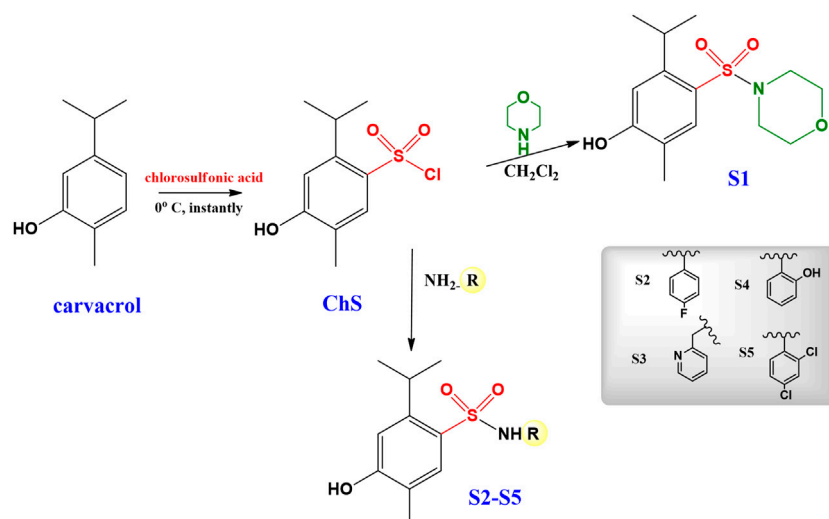
Upon inflammatory reactions, pro-inflammatory chemical messengers stimulate resident cells, recruit nociceptors and cells, and drive pain conduction (Manchope et al., 2016). Furthermore, augmented oxidative stress upon inflammation promotes nociception. For example, Reactive Nitrogen Species (RNS) and Reactive Oxygen Species (ROS) in a direct and indirect manner promote sensitization and activation of nociceptors (Maioli et al., 2015). The unbalance between oxidative and antioxidative agents in inflammatory reactions promotes oxidative stress (Biswas, 2016). Even though many analgesic agents can be used for the therapy of pain, research on novel drug candidates is needed considering that the current analgesics cause a broad diversity of adverse effects (Burgess and Williams, 2010).

Natural product structural motifs have been an invaluable source of new chemical matter for drug design and medicinal chemistry (Rodrigues et al., 2016). Recently, natural product research in the industry has decreased because of compatibility

problems between natural-product extract collections and high-throughput screening platforms (Koehn and Carter, 2005). In this scenario, the monoterpene phenol 2-methyl-5-isopropyl-phenol, known as carvacrol, is a simple molecule with no stereogenic centers, with druglike properties and whose derivatives can be used for structure-activity relationship (SAR) studies. Along with the anti-inflammatory activity of carvacrol (Arigesavan and Sudhandiran, 2015), researchers have been interested in studying the analgesic action of this monoterpene.

Calcium and potassium channels are also directly related to the transmission of painful impulses since they are central for the release of neurotransmitters from nociceptor terminals. In this sense, studies demonstrate that carvacrol promotes a vasorelaxant response in upper mesenteric artery rings in rats, potentially because it inhibits the influx of calcium ions mediated by voltage-sensitive calcium channels (Cav), as well as the receptor-operated channel (ROC) (Pires et al., 2015). Stock-actuated calcium channels (SOC) seem to be associated with classical TRP receptors (C6, C1, and TRPC) and also with melastatin TRP receptor channel inhibition (TRPM7) (Figure 1). The observed vasorelaxant activity may be involved in the hypotensive response detected in *in vivo* studies (Dantas et al., 2015).

Melo et al. (2010) demonstrated that doses of 12.5, 25, and 50 mg/kg of carvacrol, administered orally, have an anxiolytic effect and do not alter the locomotor activity of the animals. In a previous study, we demonstrated that some synthetic sulfonamides derived from carvacrol at a dose of 30 mg/kg, intraperitoneal (ip), are able to reduce streptozotocin-induced Alzheimer's disease deficits, in addition to producing anxiolytic and antioxidant effects, without affecting locomotor activity of animals (de Souza et al., 2020). Also, it was confirmed that carvacrol, administered orally, at single doses of 50 and



SCHEME 1 | Synthesis of the carvacrol-derived sulfonamides.

100 mg/kg, produces significant inhibition of nociception caused by chemical (formalin and acetic acid) and thermal stimulations (hot-plate test) (Cavalcante Melo et al., 2012). Furthermore, part of the mechanism by which carvacrol exerts its effects was demonstrated by Zotti et al. (2013). The authors found that carvacrol administered orally for seven consecutive days (12.5 mg/kg) was able to increase dopamine and serotonin levels in the prefrontal cortex and hippocampus. Following these findings, it has been demonstrated that carvacrol promotes antinociceptive effects by a mechanism that is independent on the activation of the opioid machinery and the L-arginine-nitric oxide (NO) pathway (Cavalcante Melo et al., 2012).

Sulfonamides derived from carvacrol have been investigated recently, for which antibacterial properties (Oliveira et al., 2020) and potential candidates for the development of drugs for the treatment of Alzheimer's disease have been reported (De Souza et al., 2020). As mentioned above and due to the analgesic and anti-inflammatory potential of carvacrol, in this research, the antinociceptive potential of these sulfonamides was investigated. Thus, this investigation is the first report to demonstrate the potential antioxidant activity of sulfonamides derived from carvacrol. Furthermore, this is the first report of sulfonamides derived from carvacrol, rationally designed to the effective control of pain via inhibition of the glutamatergic system. Additionally, molecular docking and quantum investigations were carried out to rationalize the *in vitro* and *in vivo* data.

Despite advances in the pharmacokinetics and pharmacodynamics of analgesic agents, their high toxicity is a determinant of conflicting clinical results due to the need for drug associations and interactions, especially in chronic pain due to its bioplasticity, and association with clinical conditions of anxiety and depression that reduce the quality of life of patient (Berman and Bausell, 2000; Jensen et al., 2001). Therefore, the development of new

chemotherapeutic agents for pain treatment, which is the objective of this research, is extremely relevant in the context of public health worldwide.

MATERIALS AND METHODS

Synthesis of Sulfonamides

All the solvents used were analytically pure. The reagents 5-isopropyl-2-methylphenol (carvacrol), chlorosulfonic acid, morpholine, 4-fluoroaniline, pyridin-2-yl methanamine, 2-hydroxyaniline, 2,4-dichloroaniline were obtained from Sigma Aldrich.

The synthesis sulfonamides **S1–S5**, as already described in the literature (de Oliveira et al., 2016) was performed in two steps: firstly, the synthesis of 4-hydroxy-2-isopropyl-5-methylbenzene-1-sulfonyl chloride (ChS) was performed, subsequently, the ChS was used in reactions with different amines (**Scheme 1**). ChS was obtained from the reaction of carvacrol to six equivalents of chlorosulfonic acid. The sulfonamides obtained in this study were prepared from ChS with two equivalents of amine added slowly. Reactions were followed by thin layer chromatography (TLC). All sulfonamides were purified by acid-base extraction and the compounds were duly characterized by spectroscopic and spectrometric techniques.

Behavioral Tests

Animal Models

Animal care and *in vivo* procedures were carried out according to the ethical guides for the study in conscious animals of experimental pain (Zimmermann, 1983). The experiments were carried out after protocol approval from the Ethics Committee of the Federal University of Santa Catarina—UFSC (protocol PP00745). Male Swiss mice (25–35 g) were obtained

from UFSC. Animals were maintained in a 12 h light/12 h dark cycle (lights on at 6:00 a.m.) under a temperature of $22 \pm 2^\circ\text{C}$ with water and food *ad libitum*. At least 1 h before the tests, the animals were acclimatized to the laboratory conditions. The tests were executed from 8:00 a.m. to 12:00 a.m. The number of animals and noxious stimulation intensity were kept at the minimum needed to obtain consistent results.

Drugs and Reagents

The following substance was used: L-glutamic acid hydrochloride (Sigma–Aldrich, St. Louis, MO, United States). This formulation has a glutamate content of $\geq 99\%$ measured by HPLC, according to the manufacturer's technical sheet. The carvacrol, used in this work, was obtained commercially in liquid form by Sigma–Aldrich, whose density is 0.976 g/ml at 20°C (lit.), melting point $3\text{--}4^\circ\text{C}$ (lit.) with a concentration of 98 %. Glutamate was solubilized in isotonic saline solution (0.9% NaCl), and carvacrol and sulfonamides derived from carvacrol (S1–S5, Scheme 1) were dissolved in saline plus Tween 80. Tween 80 did not exceed a 5% final concentration and did not show any activity by itself. Control groups for each delivery route were given isotonic saline with Tween 80 at 5%.

Glutamate-Induced Nociception

To demonstrate the possible interplay between the carvacrol derivatives and the glutamatergic system, we evaluated whether the compounds would antagonize the glutamate-induced pain behavior of paw licking and biting. This glutamate-induced model of nociception was reported previously (Beirith et al., 2002; Meotti et al., 2010). A $20\ \mu\text{l}$ glutamate solution ($20\ \mu\text{mol/paw}$, in saline, with pH adjusted to 7.4) was administered intraplantarly (i.pl.) in the ventral face of the right hind paw. After the administration of glutamate, the mice were monitored for 15 min. Nociception was monitored by measuring with a chronometer the amount of time that mice spent licking and biting the injected paw. The mice were given vehicle intragastrically (i.g.) (10 ml/kg) or carvacrol derivatives (0.0003, 0.003, and 0.03 mg/kg) 1 h before glutamate administration.

Additionally, the thickness of the animal paw was measured with a digital micrometer (0–25 mm) before and after the nociceptive response induced by glutamate (i.pl.) to evaluate the paw edema. The difference in thickness (mm) of the hind paw, immediately before and after the test of glutamate, was considered as an index of edema.

Evaluation of Locomotor Activity

The open-field test is widely used to assess spontaneous locomotor activity in animals to exclude possible nonspecific effects of a drug on the central nervous system (CNS), causing sedation or motor dysfunction. This is an important measure to check for possible false positives in pain studies, as these parameters can be easily confused with an analgesic effect of the evaluated drug and cause research bias. Thus, to examine the activity of the carvacrol derivatives on spontaneous locomotion, the open-field test was performed as described above (Nucci-Martins et al., 2016; de Souza et al., 2020). The open-field test device was a wooden box

($40 \times 60 \times 50\text{ cm}$). The floor was split into 12 equal squares, and the number of squares that the animal covered with all paws in a 6 min session was registered. Mice were given the compounds (i.g., 0.0003, 0.003, and 0.03 mg/kg) or vehicle (i.g., 10 ml/kg) 1 h before the test. Healthy mice that were not submitted to painful stimuli were used for the assessment of locomotor activity in the open-field experiment.

Statistical Analyses

Results are reported as average values \pm standard deviation (SD) with the exception of ID_{50} and EC_{50} values, which were calculated from single experiments using nonlinear regression implemented in GraphPad 7.0 (GraphPad software, San Diego, CA, United States). The glutamate test with paw edema measurement and the open-field test showed a normal data distribution in line with the Shapiro–Wilk threshold ($p = 0.05$) and, thus, were submitted to one-way ANOVA analysis and to Dunnett test for multiple analyses. Only p -values below 0.05 were taken as significant ($p < 0.05$).

Antioxidant Assays

Scavenging Assay—Nitric Oxide

NO scavenging assay was performed using the method reported by Sens et al. (2018). In this assay, sodium nitroprusside generates NO radicals ($\text{NO}\bullet$) which react with oxygen to generate nitrite ions. The production of the nitrite ions is then determined with the Griess reagent (1% sulfanilamide, 2% H_3PO_4 and 0.1% naphthylethylenediamine dihydrochloride). NO scavenging activity was measured by adding 1.5 ml phosphate buffer saline (0.2 M, pH 7.4) and 1 ml sodium nitroprusside (10 mM) to several concentrations of the test compounds (25, 50, 75, and 100 mg ml^{-1}) and incubating the reaction mixture for 150 min (25°C). Next, 1 ml of Griess reagent was added to 1 ml of the reaction solution. A wavelength of 546 nm was set to measure absorbance (A), and the results of antioxidant assays were expressed as EC_{50} .

Scavenging Assay—Hydrogen Peroxide

The H_2O_2 scavenging activity showed by the compounds was measured spectrophotometrically using a method reported previously (Sens et al., 2018). A 40 mM H_2O_2 solution was made in phosphate buffer (pH 7.4). 25, 50, 75, and 100 mg ml^{-1} test compound solutions in phosphate buffer (3.4 ml) were added to the H_2O_2 solution (0.6 ml). Absorbance was monitored at a wavelength of 230 nm. The percentage of H_2O_2 scavenging was calculated, and the results were expressed as EC_{50} .

Computational Studies

Small-Molecule Modeling and Preparation

All compounds were built in the Avogadro program (Hanwell et al., 2012). The structures of the compounds were optimized at pH 7.4 to simulate the conditions found experimentally. Next, the compounds were minimized with the MMFF94s force field (Halgren, 1996) and the conjugate gradient method.

Density Functional Theory

All energy values of the lowest unoccupied molecular orbitals (LUMO) and highest occupied molecular orbitals (HOMO) were computed by the GAMESS (General Atomic and Molecular

Electronic Structure System) software (Schmidt et al., 1993). In the calculation of simple energy, the Becke's three-parameter hybrid functional, the Lee–Yang–Parr correlation (B3LYP) functional (Nageswari et al., 2018) and the 6–31G(d, p) basis set were used in these molecular systems in gas phase, considering the neutral and singlet structures. The computation was run considering the Slater exchange potential correlation and the grid methodology. The Hückel method (Hückel, 1931) generated an initial estimate of molecular orbitals and electronic density. Consequently, the self-consistent field (SCF) convergence was attributed by the restricted Hartree-Fock (RHF) method (Schmidt et al., 1993), which was limited to 30 iteration cycles. LUMO and HOMO potentials were compared with the experimental results of NO (EC_{50}^{NO}) and peroxide ($EC_{50}^{H_2O_2}$) elimination activities. Finally, HOMO-biological activity (EC_{50}^{NO} and $EC_{50}^{H_2O_2}$) linear regression models were developed.

Molecular Docking

The PDB (Berman et al., 2002) was searched for structures of *Rattus norvegicus* bound to antagonist corresponding to the UniProt Gene Names Grin1 and Grin2A-D (NMDA receptors; 23 structures found); Gria1-4 (AMPA receptors; 16 structures found); Grik1-5 (Kainate receptors; 20 structures found); Grm1 and Grm5 (mGluR Group I receptors; 1 structure found); Grm2–3 (mGluR Group II receptors; no structures found) and Grm4–8 (mGluR Group III receptors; no structures found). When more than one structure was available, a direct comparison of the binding sites was performed to evaluate their plasticity and select the smallest subset of structures capable of representing it. For each subset, ensemble docking calculations were performed. After identifying the structure of each receptor with a higher affinity for the compounds, docking simulations were performed individually. The structural data of the heme domain of rat neuronal NO synthase bound to 6-(3-fluoro-5-(3-(methylamino)prop-1-yn-1-yl)phenethyl)-4-methylpyridin-2-amine (PDB 6NGJ) was additionally used.

In all docking calculations, performed with GOLD v.5.6.1 and the ChemPLP (Korb et al., 2009) scoring function, the receptors were kept rigid, and the ligands were treated with full flexibility. The receptors were prepared using GOLD, and structural water molecules were not considered. The atoms up to a distance of 8 Å from the crystallographic ligands in both the ensemble and individual docking simulations were considered to define the binding sites. PyMOL v.1.8 (Schrodinger, New York, NY) was used to create the receptor-ligand figures.

Molecular Properties and Pharmacokinetics

Molinspiration Chemoinformatics was used for calculating Octanol-Water Partition Coefficient (milogP), number of atoms (natoms), Topological Polar Surface Area (TPSA), molecular weight (MW), hydrogen bond donors (HBD) and hydrogen bond acceptors (HBA), rotatable bonds (NRB), Molecular Volume, and Lipinski RO5 violations.

The SwissADME tool (<http://www.swissadme.ch>) was employed for the generation of the Bioavailability Radar, and assess lipophilicity, druglikeness, medicinal chemistry and pharmacokinetics parameters.

TABLE 1 | Antioxidant activity of sulfonamides derived from carvacrol.

| Compound | NO scavenging activity EC_{50} (μ M) | H ₂ O ₂ scavenging activity EC_{50} (μ M) |
|---------------|--|---|
| S1 | 12.25 \pm 0.12 | 13.13 \pm 0.11 |
| S2 | 18.11 \pm 0.14 | 20.16 \pm 0.17 |
| S3 | 12.14 \pm 0.28 | 13.85 \pm 0.33 |
| S4 | 18.76 \pm 0.22 | 20.28 \pm 0.14 |
| S5 | 12.04 \pm 0.11 | 13.12 \pm 0.18 |
| Ascorbic acid | 14.72 \pm 0.23 | 16.3 \pm 0.26 |

RESULTS

The synthetic procedures for the sulfonamides **S1–S5** (Scheme 1), following a recently reported methodology (de Oliveira et al., 2016), were performed in good yields (85–95%).

Antioxidant Activity

The antioxidant activity of the sulfonamides derived from carvacrol (Table 1) was analyzed by the NO and H₂O₂ scavenging activity assays.

Quantum Studies

The electronic properties were directly correlated with the antioxidant activity of the molecules. The E_{HOMO} and E_{LUMO} indicate the molecule's ability to donate and receive electron density, respectively. The difference between the two energy levels is termed the band gap and gives an estimate of the reactivity of a molecule. The distance between the HOMO and LUMO energy levels is inversely proportional to the reactivity the compound. The HOMO and LUMO potentials and band gap of the carvacrol derivatives are shown in Figure 2.

Figure 3 shows the correlation between HOMO energy and experimental EC_{50}^{NO} and $EC_{50}^{H_2O_2}$. The correlation coefficients r^2 and Person's coefficient (r) of the EC_{50}^{NO} versus E^{HOMO} were 0.87 and 0.93, respectively. For $EC_{50}^{H_2O_2}$ versus E^{HOMO} , r^2 and r were 0.88 and 0.94, respectively. The angular coefficient values of the equations $EC_{50}^{NO} = EC_{50}^{NO}(E^{HOMO})$ and $EC_{50}^{H_2O_2} = EC_{50}^{H_2O_2}(E^{HOMO})$ were, respectively, 10.28 ± 2.03 and $11.30 \pm 2.10 \mu\text{mol. (LeV)}^{-1}$. In addition, the linear coefficients were 87.90 ± 14.49 and $96.63 \pm 14.96 \mu\text{mol. (LeV)}^{-1}$, respectively. From these equations, the minimal values of E^{HOMO} (i.e., $EC_{50}^{NO} = EC_{50}^{H_2O_2} = 0$) can find the maximal activity. Thus, with the HOMO energy tending to -8.55 eV for both equations, the maximal elimination of NO and H₂O₂ is reached for both experiments.

Antinociceptive Activity

For a better understanding of the antinociceptive effect of sulfonamides derived from carvacrol (**S1–S5**), we used the model of glutamate-induced (i.pl.) nociception. This method allowed us to investigate the possible interaction of peripheral antinociceptive action of the analyzed compounds with the glutamatergic system. The results are shown in Figure 4.

Figure 5 shows the results of treatment with carvacrol and its derivatives on paw edema induced by glutamate (i.pl.). Our results show that only **S1** and **S5** were able to significantly reduce edema.

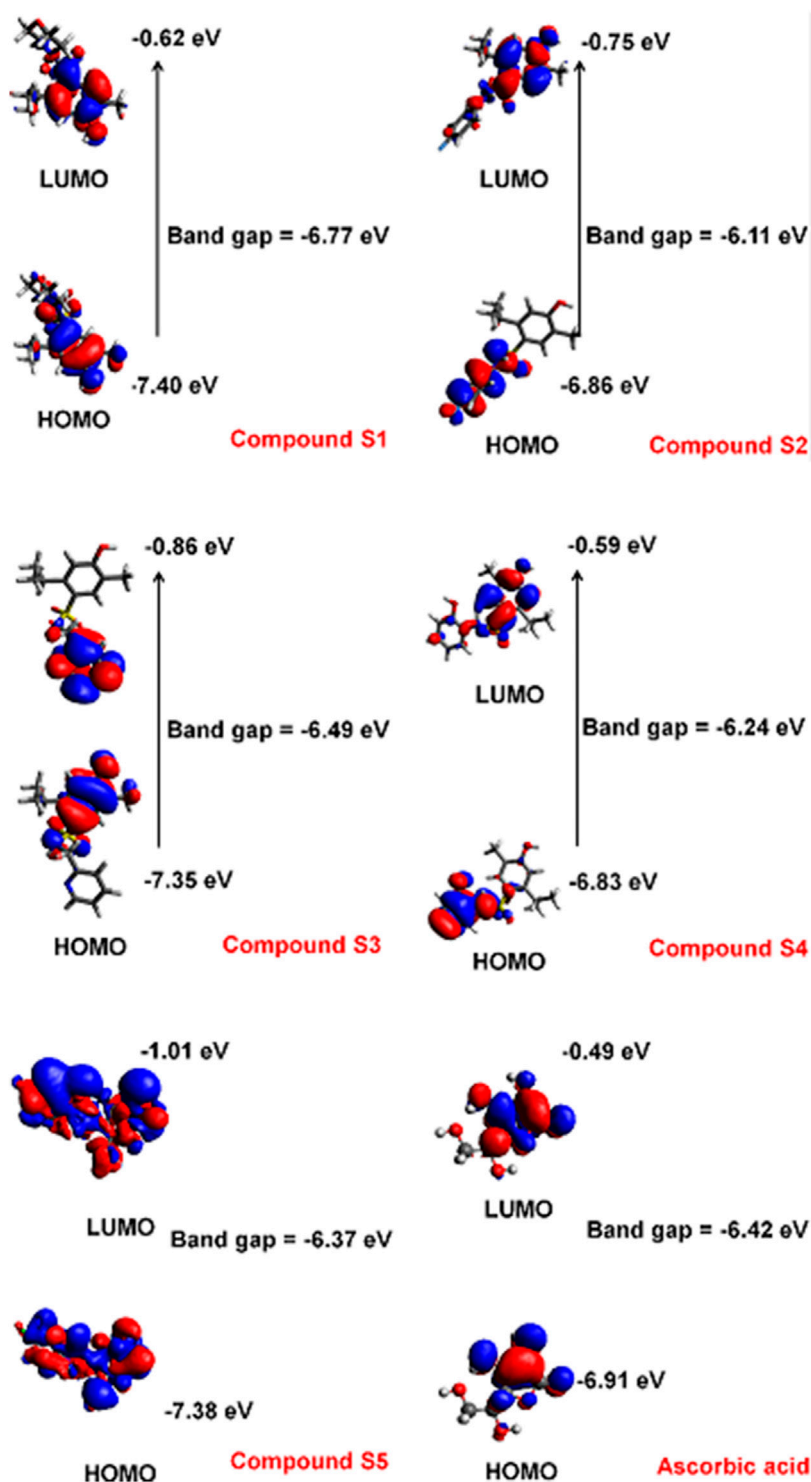
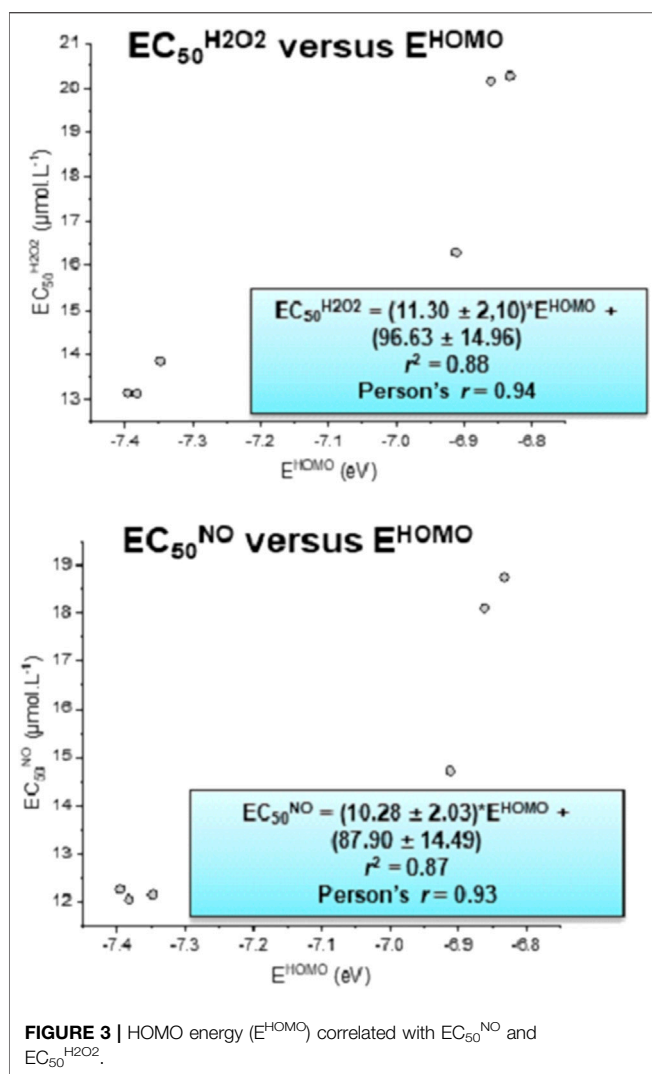


FIGURE 2 | HOMO and LUMO potentials of the carvacrol derivatives estimated by the B3LYP method and 6-31G(d,p) basis set.

However, **S1** inhibited edema more effectively and dose-dependently. The percent inhibition values were: $36 \pm 10\%$, $47 \pm 6\%$, and $73 \pm 12\%$ for **S1** at 0.0003, 0.003 and 0.03 mg/kg i.g., respectively; $19 \pm 9\%$, $33 \pm 6\%$, and $28 \pm 7\%$ for **S5** at 0.0003, 0.003, and 0.03 mg/kg i.g.,

respectively. The value of ID_{50} for compound **S1** was 0.002 (0.0009–0.005) mg/kg. Furthermore, the calculated values for the ID_{50} antiedematogenic effect of **S1** (0.002 mg/kg) agree with the dose found in the glutamate test, showing homogeneity of the data in this



group. Thus, we suggest that **S1** may be an interesting target for the reduction of edema in inflammatory conditions.

Figure 6 shows that intragastric administration of carvacrol and compounds **S1**, **S2**, **S3**, **S4**, and **S5** at doses ranging from 0.0003 to 0.03 mg/kg had no effect on the locomotion of animals in comparison with the animals in the control group, suggesting that the compounds do not induce impairment of motor function in the animals. These results exclude the possibility that the antinociceptive action of carvacrol and its derivatives is nonspecifically associated with activity on the peripheral or central levels of locomotion control, such as sedation or motor dysfunction.

Molecular Docking

As previously shown (Fundytus, 2001), the administration of glutamate receptor (GluR) antagonists has an analgesic effect on peripheral pain. To assess whether the mechanism of action of **S1–S5** is likely to involve these receptors, molecular docking simulations were performed over different GluR structures of *Rattus norvegicus* bound to antagonists (**Supplementary Table S1**).

For the predicted binding modes of **S1–S5**, the main interactions involving the common scaffold are hydrogen bonds with Gln405, Arg523, Thr518 and Ser572 and a displaced π -stacking interaction with Phe484. Of these, the interactions with Arg523, Thr518, and Phe484 are also observed for the crystallographic antagonist TK40 (Ravn et al., 2013). The main interactions observed for carvacrol are only hydrogen bonds with Pro516 and Thr518 and the displaced π -stacking interaction with Phe484 (**Figure 7**). For the different *R* groups, mainly van der Waals interactions were established. Only for the *R* groups of **S3** and **S5**, $-\text{CH}\cdots\pi$ interactions with Leu538 and Ser572, respectively, were observed. Among all five molecules, **S4** established the lowest number of contacts. The scores of each analyzed pose are presented in **Supplementary Table S2**.

The three levels of perception of pain—the cerebral (Dickenson, 1995), spinal and peripheral (Gordh et al., 1995)—appear to be affected by NO. This compound is an essential regulator of various immune and inflammatory functions (Moncada et al., 1991). In this work, we investigated, besides the NO scavenging activity, the possible intermolecular interactions between the sulfonamides and NO synthase. First, to validate the molecular docking protocol, redocking analysis (**Figure 8**) of 5,6,7,8-tetrahydrobiopterin (the crystallographic ligand, PDB ID 6NGJ) (Do et al., 2019) was carried out with GOLD. The ligand occupied the same interaction site in molecular docking when compared to the crystallographic structure, with emphasis on hydrogen bond interactions with Ser334, Val677, and Arg 596 and a π interaction with Trp678.

The molecular docking results agree with the results obtained in the *in vitro* (NO scavenging activity) and *in vivo* tests. All compounds showed an inhibitory profile against NO synthase, except **S4**, which was not effective in all performed assays. The two most active compounds, **S1** and **S5** presented lower ID_{50} values and higher values for the scoring function, which demonstrate the high correlation between the *in vivo* and *in silico* results. The higher activity of these compounds was probably due to π stacking interactions and a hydrogen bond between compounds **S1** and **S5** and Trp 678 (**Figure 9**), which were also observed for the co-crystallized ligand, but was not found for the other sulfonamides. The scores of each analyzed pose are presented in **Supplementary Table S3**.

Molecular Properties

Physicochemical and topological parameters of compounds **S1–S5** were estimated to evaluate their pharmacokinetics profile. The octanol–water partition coefficient (miLogP), topological polar surface area (TPSA), molecular weight (MW), number of atoms, hydrogen-bond acceptors (HBA) and hydrogen-bond donors (HBD), number of rotatable bonds (NRB), Lipinski RO5 violations, and molecular volume are presented in **Table 2**. The *silico*-derived descriptor values were compared with the solubility and permeability filters for drug candidates reported by Lipinski (Barret, 2018), Oprea and Veber (Veber et al., 2002).

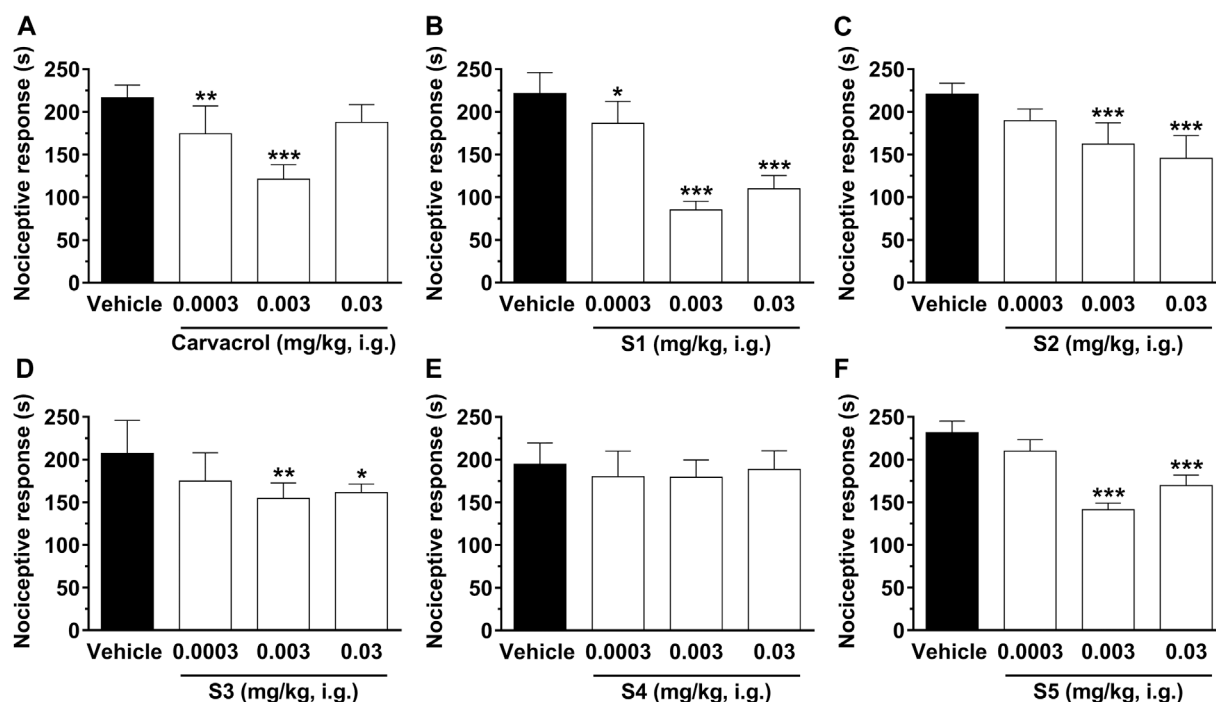


FIGURE 4 | Effect of compounds on nociception induced by glutamate (i.pl.) in mice. The pain behavior, translated by the nociceptive response of licking/biting hind paws induced by glutamate (i.pl.), was evaluated 1 hour after treatment with carvacrol (A), S1 (B), S2 (C), S3 (D), S4 (E) and S5 (F) at doses ranging from 0.0003, 0.003, and 0.03 mg/kg, i.g., (open bars) or vehicle/control (closed bar). Each bar denotes the average response for 6–8 animals, and the vertical lines represent the SEM (standard error of mean). Asterisks (*) indicate the significance in comparison with the control group animals (* $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$). One-way ANOVA and Dunnett test for multiple comparisons were used to determine the statistical significance.

The SwissADME web tool used to calculate the parameters is available at <http://www.swissadme.ch> and allows straightforward submission and analysis. It allows different input methods, multi-molecule computation, and offers the possibility to view and save results for each molecule, in addition to an interactive and intuitive visualization tool. To study the ADME parameters of the most active sulfonamide in the *in vitro* and phenotypic tests (S1), the Bioavailability Radar (Figure 10), lipophilicity, drug likeness (Figure 11), medicinal chemistry and pharmacokinetics (Figure 12) parameters were analyzed.

The Bioavailability Radar (Figure 10) provides a graphical output for the drug-likeness of a compound. The central shaded surface is the optimal domain for lipophilicity (XLOGP3 from -0.7 to $+5.0$), size (MW from 150 to 500 g/mol), polarity (TPSA from 20 to 130 Å²), aqueous solubility ($\log S \leq 6$), saturation (fraction of sp³ carbons ≥ 0.25), and flexibility (rotatable bonds ≤ 9). Compound S1 falls within the optimal range for all parameters.

In addition, S1 has a good medicinal chemistry and synthetic accessibility profile, which is very important in obtaining a drug that can be commercially distributed at a more affordable price. Moreover, S1 has high gastrointestinal absorption (GI) and blood-brain barrier permeability (Figure 12).

DISCUSSION

Antioxidant Activity

The evaluation of the antioxidant activity of a series of compounds should be performed in more than one experiment, allowing for the reliability of the results (Sens et al., 2018). Diverse *in vitro* antioxidant assays have been published. Herein, the antioxidant ability of derivatives S1–S5 was determined in two *in vitro* tests. Subsequently, the results of these tests were correlated with the findings from the HOMO and LUMO studies.

Compounds S1, S3, and S5 were more active than ascorbic acid (AA), which was used as the reference compound. Compound S5 showed the highest activity, and S4 demonstrated to be the least active. A linear correlation was found between both experimental results ($EC_{50}^{H_2O_2} = 1.085EC_{50}^{NO} + 0.2250$; $r^2 = 0.99$).

NO plays a critical part in the control of multiple physiological responses. Also, the NO cascade is associated with many conditions, including Alzheimer's disease (Di Meo et al., 2016). H₂O₂ readily decomposes into water and oxygen, resulting in the production of hydroxyl radicals (OH•), lipid peroxidation and DNA injury, which makes it a target for research of new compounds with antioxidant properties (Phaniendra et al., 2015).

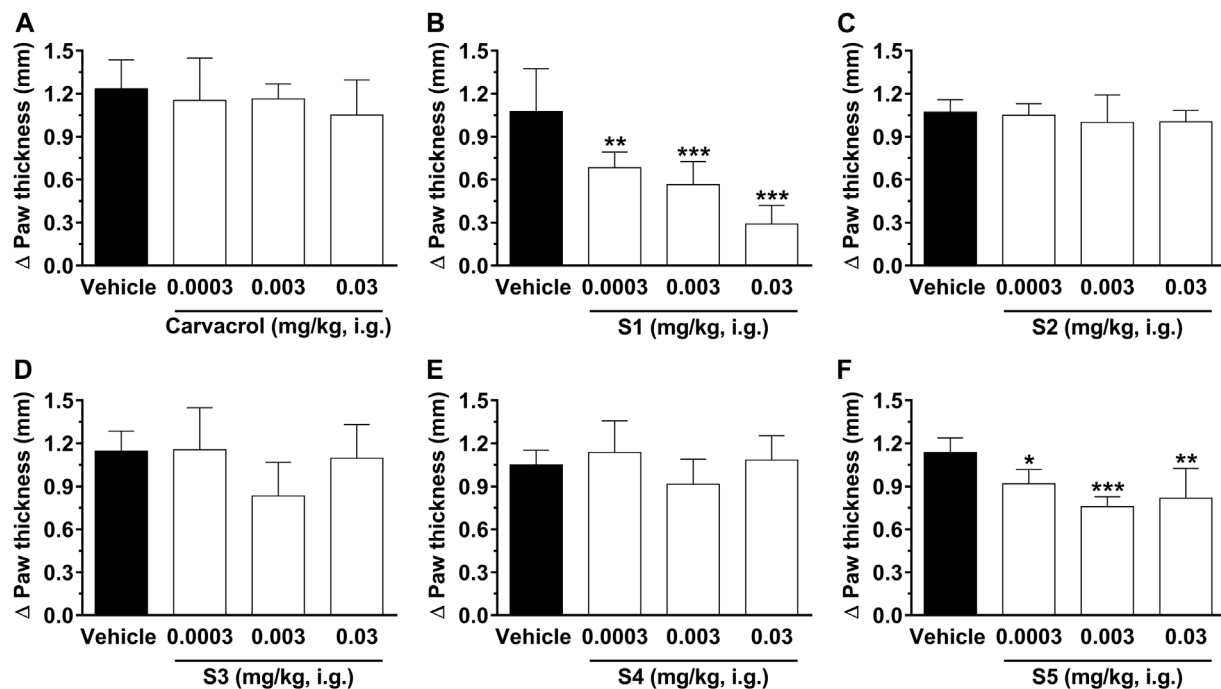


FIGURE 5 | Effect of compounds on paw edema in rats induced by glutamate (i.p.). The edema was evaluated 1 hour after treatment with carvacrol (A), S1 (B), S2 (C), S3 (D), S4 (E) and S5 (F) at doses ranging from 0.0003, 0.003, and 0.03 mg/kg, i.g., (open bars) or vehicle/control (closed bar). The animal paw thickness difference was observed before and after the glutamate test. Each bar denotes the average response for 6–8 animals, and the vertical lines represent SD. Asterisks (*) indicate the significance in comparison with the control group animals (* $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$). One-way ANOVA and Dunnett test for multiple comparisons were used to determine the statistical significance.

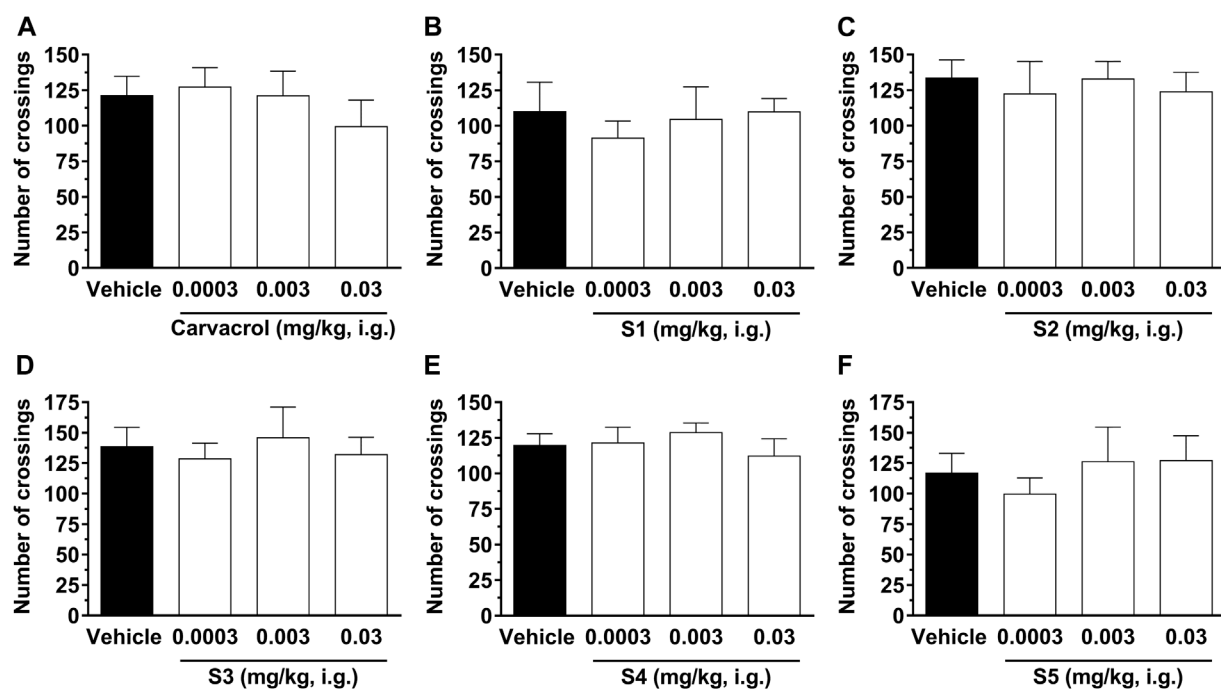


FIGURE 6 | Effect of compounds on the spontaneous locomotion of animals. The crossings were evaluated 1 h after treatment with carvacrol (A), S1 (B), S2 (C), S3 (D), S4 (E) and S5 (F) at doses ranging from 0.0003, 0.003, and 0.03 mg/kg, i.g., (open bars) or vehicle/control (closed bar). Each bar denotes the average values for 6–8 animals, and the vertical lines represent SD. One-way ANOVA and Dunnett test for multiple comparisons were used to determine the statistical significance.

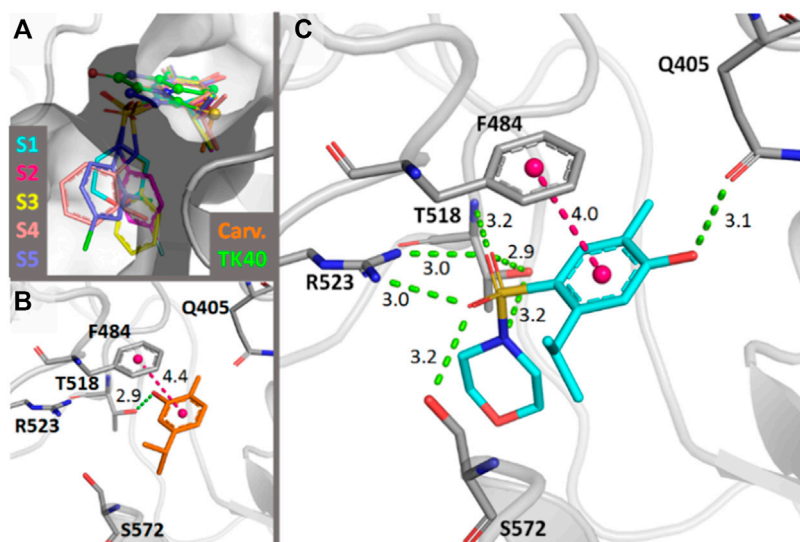


FIGURE 7 | (A) Docking-predicted binding modes in the *rattus norvegicus* NMDA-glycine binding site (PDB ID 4KFQ). The carbon atoms of each molecule are represented in a different color. The carbon atoms of the crystallographic antagonist TK40 are shown in green. **(B)** Main interactions established by carvacrol in the predicted binding mode. **(C)** Main interactions found by S1 in the predicted binding mode. Hydrogen bonds are represented in green and π -interactions in magenta. Distances are in Å.

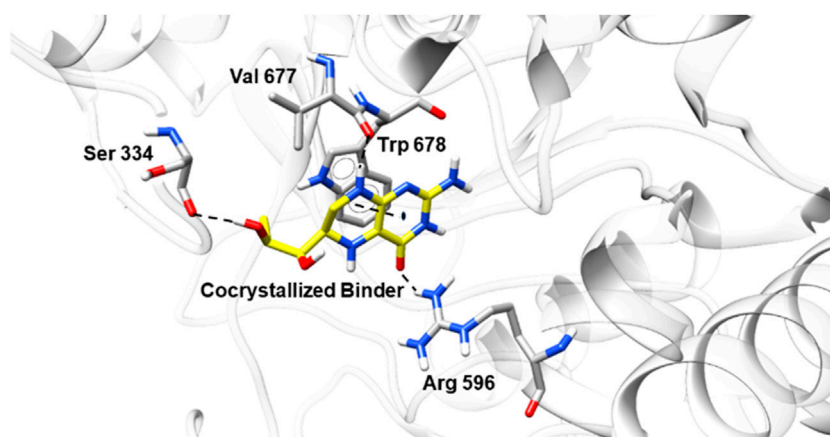


FIGURE 8 | Conformation of the crystallographic ligand in the binding site of NO synthase (PDB ID 6NGJ) after the redocking studies.

Extensive research has revealed that NO plays an essential role in several biological processes, such as neurotransmission, immune defense, and regulation of cell death (Snider and McMahon, 1998). The early 20th century witnessed the discovery of the role played by NO in nociception in both the central and peripheral levels (Zhuo and Gebhart, 1997). One of the physiological functions of NO was initially found in the vasculature; it was shown that the role of endothelium-derived relaxation factor (EDRF) could be quantitatively explained by the formation of NO by endothelial cells (Moncada and Higgs, 2006).

Treatment of pain with NO donors began with the use of nitroglycerin (NTG), which figures among the oldest treatments for ischemic heart disease (Boden et al., 2015). Discovered in

1847, NTG was used for the therapy of pain in angina pectoris for 100 years. However, its mechanism of action was not revealed until EDRF was identified as NO (Marsh and Marsh, 2000). Independently, NO was found to be an endogenous activator of soluble guanylate cyclase, resulting in the formation of cyclic GMP (cGMP), which acts as a second messenger in many cells, including the sensory neurons (Pereira et al., 2011).

NO is a highly reactive chemical messenger diffusible through the cytoplasmic membranes that is critical for the control of neuronal transmission, inflammation, cytotoxicity, and neural plasticity (Pacher et al., 2007). NO modulates the excitability of spinal sensory neurons and contributes to pain in different ways. The control of NO biosynthesis is regulated by NO synthase

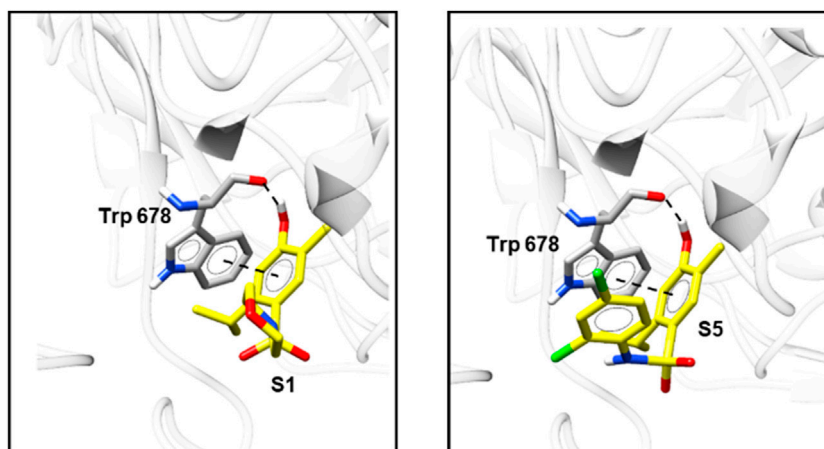


FIGURE 9 | Top-scoring docking poses for **S1** and **S5** in the binding site of NO synthase (PDB ID 6NGJ).

TABLE 2 | Molecular properties of sulfonamides **S1–S5**.

| Property | S1 | S2 | S3 | S4 | S5 |
|-------------------------------------|-----------|-----------|-----------|-----------|-----------|
| miLogP | 2.43 | 4.20 | 2.57 | 3.77 | 5.32 |
| TPSA (\AA^2) | 66.84 | 66.40 | 79.29 | 86.62 | 66.40 |
| Natoms | 20 | 22 | 22 | 22 | 23 |
| MW | 299.39 | 323.39 | 320.41 | 321.40 | 374.29 |
| HBA | 5 | 4 | 5 | 5 | 4 |
| HBD | 1 | 2 | 2 | 3 | 2 |
| nviolations | 0 | 0 | 0 | 0 | 1 |
| NRB | 3 | 4 | 5 | 4 | 4 |
| Molecular volume (\AA^3) | 268.14 | 278.75 | 286.46 | 281.83 | 300.89 |

(NOS) enzymes. Three NO synthase isoforms (NOS; EC 1.14.13.39) catalyze the production of NO (Förstermann and Sessa, 2012). They use O_2 and L-arginine as substrates and flavin mononucleotide (FMN), flavin adenine dinucleotide (FAD), reduced nicotinamide-adenine-dinucleotide phosphate

(NADPH), and tetrahydrobiopterin (BH₄) as cofactors (Förstermann and Sessa, 2012). In this work, molecular docking was used to investigate NOS inhibition by the carvacrol derivatives.

Quantum Studies

The HOMO profile showed a variation of the charge density among the carvacrol derivatives. The HOMO and LUMO energies of compound **S1** is -7.40 and -0.62 eV, respectively. The electronic density is concentrated in the phenol group for HOMO and LUMO. Compound **S2**, however, differs regarding the position of the charge density for these orbitals. In HOMO (-6.86 eV), the orbitals are concentrated on the fluoro-phenyl group. This is because fluorine tends to attract electron density (electronegative atom). In LUMO (-0.75 eV), the electronic density tends to be favorable in the phenol group. The band gap in this compound is -6.11 eV. Compound **S3** has HOMO and LUMO energies of -7.35 and -0.86 eV, respectively. The electronic density of HOMO tends to be located at the phenol. In

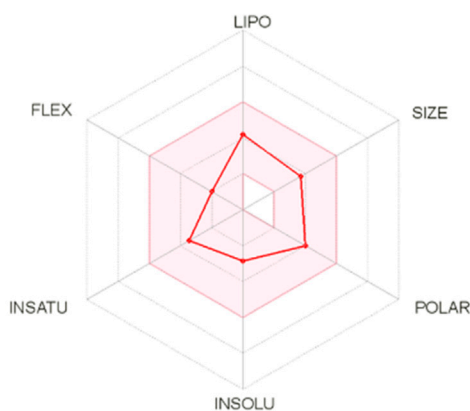
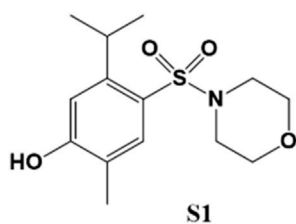


FIGURE 10 | The Bioavailability Radar for **S1**. The figure was generated online using SwissADME. Compound **S1** combines good hydrophobicity and solubility, which is vital for membrane transport and permeability. Also, it does not violate any of the filters proposed by Lipinski, Ghose, Veber, Egan, and Muegge (Figure 11).

| Lipophilicity | | Druglikeness | |
|-----------------------------------|------|-----------------------|------------------|
| Log P _{o/w} (iLOGP) | 2.40 | Lipinski | Yes; 0 violation |
| Log P _{o/w} (XLOGP3) | 1.83 | Ghose | Yes |
| Log P _{o/w} (WLOGP) | 2.54 | Veber | Yes |
| Log P _{o/w} (MLOGP) | 0.89 | Egan | Yes |
| Log P _{o/w} (SILICOS-IT) | 1.75 | Muegge | Yes |
| Consensus Log P _{o/w} | 1.88 | Bioavailability Score | 0.55 |

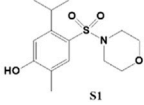


FIGURE 11 | Lipophilicity and drug likeness for **S1**. These parameters were generated online using SwissADME.

LUMO, however, the electronic density concentrates in the region of the pyridinic group. In compound **S4**, the HOMO charge density surrounds the phenolic substituent (−6.83 eV). In LUMO, however, the charge density concentrates in the carvacrol fragment (−0.59 eV). Differently from the other compounds, the charge distribution in **S5** distributes throughout the structure in HOMO (−7.38 eV) and LUMO (−1.01 eV). In HOMO, the positive density concentrates on the sulfonamide group and *p*-fluorine atom. In LUMO, however, the same region is predominantly negative throughout the structure. In ascorbic acid, the dihydroxyfuran has the HOMO electron density (−6.91 eV) close to the hydroxyl groups in the resonant region. In LUMO (−0.49 eV), the signal of electronic density changes and concentrates close to the oxygen atom of the furan group.

Antinociceptive Activity

Injection of glutamate (i.pl.) in the mouse paw causes significant paw edema and nociception (Beirith et al., 2002; Meotti et al., 2010). **Figure 4** shows that systemic administration of carvacrol, **S1**, **S2**, **S3**, and **S5** significantly inhibits nociception induced by injection of 20 μmol/paw glutamate, suggesting that these compounds have an important therapeutic effect for the treatment of acute pain of inflammatory origin, probably due to a decrease in peripheral glutamatergic signaling. Treatment with the compounds significantly reduced pain behavior induced by glutamate (i.pl.), characterized by spontaneous licking/biting of the injected hind paw. Carvacrol was able to reduce nociceptive behavior by 19 ± 6 and 44 ± 8% at 0.0003 and 0.003 mg/kg, respectively. Moreover, the sulfonamides derived from carvacrol showed the following percent inhibitions: 16 ± 5%, 62 ± 5%, and 50 ± 7% for **S1** at 0.0003, 0.003, and 0.03 mg/kg, respectively; 26 ± 5% and 34 ± 6% for **S2** at 0.003 and 0.03 mg/kg, respectively; 25 ± 9% and 22 ± 5% for **S3** at 0.003 and 0.03 mg/kg, respectively; 39 ± 8% and 27 ± 13% for **S5** at 0.003 and 0.03 mg/kg, respectively.

The calculated mean ID₅₀ value for sulfonamides derived from carvacrol was 0.002 (0.001–0.002) mg/kg for **S1**, 0.442 (0.063–0.387) mg/kg. Thus, the results of the present study demonstrate that carvacrol and **S1**, **S2**, **S3**, and **S5** reduce nociception induced by glutamate (i.pl.), suggesting that inhibition of the stimulatory mechanism via peripheral

| Medicinal Chemistry | | Pharmacokinetics | |
|-------------------------|---------|--------------------------------------|------------|
| PAINS | 0 alert | GI absorption | High |
| Brenk | 0 alert | BBB permeant | Yes |
| Leadlikeness | Yes | P-gp substrate | No |
| Synthetic accessibility | 2.77 | CYP1A2 inhibitor | No |
| | | CYP2C19 inhibitor | Yes |
| | | CYP2C9 inhibitor | No |
| | | CYP2D6 inhibitor | No |
| | | CYP3A4 inhibitor | No |
| | | Log K _p (skin permeation) | −6.83 cm/s |

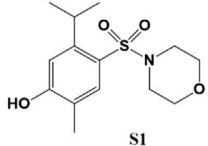


FIGURE 12 | Medicinal Chemistry and pharmacokinetics for **S1**. These parameters were generated online using SwissADME.

glutamatergic neurotransmission may contribute, at least in part, to the antinociceptive effect of these compounds. In addition, we would like to highlight that carvacrol and compounds **S1**, **S2**, **S3**, and **S5** may be interesting lead compounds for acute pain, especially **S1** (0.003 mg/kg) since it presented the highest efficacy among the analyzed compounds.

Importantly, the compounds derived from carvacrol, selected to carry out the *in vivo* experiments, were chosen from the results presented in the molecular docking, quantum studies, and the *in vitro* antioxidant activity. Our results corroborate previous results (Arigesavan and Sudhandiran, 2015) which also found antioxidant and anti-inflammatory effects after treatment with carvacrol, using a carcinogenicity model in the colon of rats. Moreover, previous studies demonstrated that carvacrol attenuates mechanical hypernociception induced by carrageenan (Guimarães et al., 2012) and the acute pain acetic acid-induced abdominal constriction and formalin (Cavalcante Melo et al., 2012). Also, it was shown (Barnwal et al., 2018) that carvacrol increased the activities of antioxidant enzymes and downregulated expression by reducing the inflammation marker in positively dyed cells (iNOS, NF-κB, and COX-2) in a pulmonary toxicity model. These data from the literature reinforce the antinociceptive, anti-inflammatory, and antioxidant potential of carvacrol observed in our study.

Findings from the literature (Pacher et al., 2007; Förstermann and Sessa, 2012) indicate that superoxide (SO, O^{2•(−)}) and peroxynitrite (PN, ONOO[−]), the product of its reaction) are essential for the emergence of pain caused by different etiologies. These findings reinforce the concept that ROS play an essential part in NMDA activation, which is a critical ionotropic glutamatergic receptor, which contributes to central and peripheral pain. Therefore, this study supports previous results (Wang et al., 2004) that stated that superoxide mediates hyperalgesia (increased sensitivity to painful stimulation) through M40403, a manganese(II) complex with a bis(cyclohexylpyridine-substituted) macrocyclic ligand, which is a superoxide dismutase mimetic. These findings disclosed the central role played by superoxide in the peripheral signaling of nociception. In addition, it was shown that the M40403 antihyperalgesic activity could not be reverted by naloxone, which excludes the participation of opioid signaling cascades.

Moreover, so far, few studies have investigated the effect of carvacrol on neurotransmitter modulation. The studies by Zotti et al. (2013) demonstrated that carvacrol, when ingested regularly in low concentrations, influences brain activity by increasing the levels of neurotransmitters such as serotonin and dopamine, which can determine feelings of well-being and reinforcing positive effects. Thus, our interest in investigating the glutamatergic system has arisen, considering that glutamate is a major mediator in the CNS, mediating excitatory neurotransmission in mammals, including in sensory neurons that convey pain, being strongly involved in the stimulation of peripheral and central pain. Therefore, our findings are unprecedented and relevant as they demonstrate the inhibitory capacity of carvacrol on the peripheral glutamatergic pathway.

It was shown (Kuo et al., 2017) that carvacrol mitigated injury in tissues and inflammation derived from periodontitis induced by ligation. Besides that, carvacrol proved to attenuate inflammatory response induced by carrageenan, decreasing mouse paw edema (Guimarães et al., 2012). These data from the literature support the anti-inflammatory, antinociceptive and antiedematogenic effects of carvacrol observed in our study. Importantly, paw edema and pain induced by glutamate are essentially associated with non-NMDA ionotropic glutamate receptors and NO production, a vasodilator, and an important neurotransmitter (Beirith et al., 2002). When in excess, it may be involved in the production of oxidative lesions in proteins. These findings reinforce the importance of studying glutamate-induced paw edema and nociception, as well as the beneficial effects of carvacrol and its derivatives found in this study.

Our results agree with literature data which demonstrated that carvacrol had no effect on the spontaneous locomotion in mice (Cavalcante Melo et al., 2012; Guimarães et al., 2012). However, these studies used a curve of carvacrol doses ranging from 25 to 100 mg/kg in the open-field test and we are the first group to test a much lower dose curve for carvacrol (0.0003, 0.003, and 0.03 mg/kg) in pain, edema, and spontaneous locomotion. In addition, Guimarães et al. (2012) demonstrated that carvacrol at a dose of 100 mg/kg reduced the animals' ambulation in the open-field test, 30 min after intraperitoneal administration, showing that this dose is not safe as it causes nonspecific effects on locomotor activity and should be excluded in future pain studies. It is already well described that some drugs can cause motor slowness (bradykinesia) or even act as a muscle relaxant, causing non-specific changes in the locomotor activity of animals (Cartmell et al., 1991). In addition, drugs like benzodiazepines and other anxiolytics decrease the exploratory behavior of animals (Hazim et al., 2014). In this regard, it was demonstrated (Coderre and van Empel, 1994) that many glutamate antagonists, primarily via ionotropic NMDA receptor, such as the receptor channel block MK-801, produce significant antinociceptive effects, but decrease exploratory behavior of animals. In contrast, our results demonstrate that the intragastric treatment with the tested compounds can induce a significant antinociceptive effect via inhibition of peripheral glutamate, without causing any detectable motor dysfunction. Thus, carvacrol and its derivatives **S1**, **S2**, **S3**, and **S4** at doses up

to 0.03 mg/kg have an attractive analgesic potential to treat acute pain without causing CNS sedation.

Molecular Docking

In general, no significant binding modes were obtained concerning poses matching the available structural criteria of known antagonists (Ramírez and Caballero, 2018). Only the docking simulations in the NMDA-GluN₁ glycine binding site (LBD-GluN₁) excelled, which agrees with previous observations for selective ligands of this site, such as HA-966, "which barely interacts with other ionotropic glutamate receptors" (Planells-Cases et al., 2005).

Considering the docking results and the non-ataxic effects of the compounds at the administered doses, the compounds are likely to be partial agonists, instead of agonists of the NMDA-GluN₁ glycine binding site, such as rapastinel (Wood et al., 2008) (GLYX-13 or BV-102), (+)-HA-966 (Millan and Seguin, 1993) and the recently reported 1-amino-1-cyclobutanecarboxylic acid (Fung et al., 2019).

Molecular Properties

The Lipinski RO5 applies to compounds that are active after oral administration. The RO5 includes four physicochemical property ranges ($\log P \leq 5$, $MW \leq 500$, $HBD \leq 5$ and $HBA \leq 10$) that are present in 90% of the drugs that are active after oral administration and have reached phase II clinical development (Barret, 2018). The sulfonamides investigated in this work are within the RO5 desirable range, except for the $miLogP$ of sulfonamide **S5** ($miLogP = 5.32$), which is slightly higher than expected.

TPSA correlates with a compound's ability to permeate biological membranes through passive transport. Medicinal chemists use TPSA as an important parameter to optimize drug permeation through membranes. Molecules having TPSA values higher than 140 \AA^2 are likely to permeate poorly into cell membranes (Pajouhesh and Lenz, 2005). For molecules that are required to act in the CNS, penetration into the blood-brain barrier is needed, which requires a TPSA lower than 90 \AA^2 (Hitchcock and Pennington, 2006). All investigated sulfonamides are in accordance with these parameters. A molecule that has a higher number of rotatable bonds becomes more flexible and have a good binding affinity with the binding pocket. For a potential drug candidate, Veber proposed that NRB should be ≤ 10 . All investigated sulfonamides are following this parameter.

The molecular volume assesses the transport properties of molecules such as blood-brain barrier penetration. The calculated values for this property are in line with the values expected for drug candidates.

During the discovery of novel drugs, molecules with useful therapeutic properties and low levels of toxicity are highly desirable. In this process, knowledge of the absorption, distribution, metabolism, and excretion profiles (ADME) is essential. It is well-known that the early evaluation of ADME during the drug discovery process reduces the attrition rates during clinical development.

CONCLUSION

In this study, we report the SAR for a series of carvacrol-derived sulfonamides. The antioxidant and antinociceptive activities of compounds **S1–S5** were investigated using *in vitro* and *in vivo* assays. All the sulfonamides showed antioxidant activity in the *in vitro* tests comparable to that of the control compound (ascorbic acid). The results gathered in the *in vitro* antioxidant tests were linearly compared to the binding energies of the HOMO frontier orbital ($r^2 = 0.87$ and 0.88) calculated by DFT. The results of this study demonstrate that carvacrol and its derivatives **S1**, **S2**, **S3**, and **S5** were able to reduce nociception induced by glutamate (i.pl.). Moreover, these findings show that the intragastric treatment with the tested compounds can induce a significant antinociceptive effect via inhibition of glutamatergic peripheral system without causing any detectable motor dysfunction, and not affecting the locomotor activity of mice. Thus, carvacrol and compounds **S1**, **S2**, **S3**, and **S5** at doses up to 0.03 mg/kg have an attractive analgesic potential to treat acute pain with no CNS sedation. Docking simulations highlighted the interactions between the compounds and the NMDA-GluN₁ glycine binding site, which suggested that these molecules act as selective partial agonists. Besides, compounds **S1–S5** exhibit physicochemical parameters and pharmacokinetics compatible with drug candidates. Overall, sulfonamides **S1–S5** are suitable starting points for further molecular optimization.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

REFERENCES

- Arigesavan, K., and Sudhandiran, G. (2015). Carvacrol Exhibits Anti-Oxidant and Anti-Inflammatory Effects against 1, 2-Dimethyl Hydrazine Plus Dextran Sodium Sulfate Induced Inflammation Associated Carcinogenicity in the Colon of Fischer 344 Rats. *Biochem. Biophys. Res. Commun.* 461, 314–320. doi:10.1016/j.bbrc.2015.04.030
- Barnwal, P., Vafa, A., Afzal, S. M., Shahid, A., Hasan, S. K., Alpashree, et al. (2018). Benzo(a)pyrene Induces Lung Toxicity and Inflammation in Mice: Prevention by Carvacrol. *Hum. Exp. Toxicol.* 37, 752–761. doi:10.1177/0960327117735572
- Barret, R. (2018). “Lipinski’s Rule of Five,” in *Medicinal Chemistry: Fundamentals*. Editor R. Barret (Elsevier), 97–100. doi:10.1016/b978-1-78548-288-5.50006-8
- Beirith, A., Santos, A. R. S., and Calixto, J. B. (2002). Mechanisms Underlying the Nociception and Paw Oedema Caused by Injection of Glutamate into the Mouse Paw. *Brain Res.* 924, 219–228. doi:10.1016/s0006-8993(01)03240-1
- Berman, B. M., and Bausell, B. R. (2000). The Use of Non-Pharmacological Therapies by Pain Specialists. *Pain* 85, 313–315. doi:10.1016/S0304-3959(00)00258-X
- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., et al. (2002). The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 58, 899–907. doi:10.1107/S0907444902003451
- Biswas, S. K. (2016). Does the Interdependence Between Oxidative Stress and Inflammation Explain the Antioxidant Paradox? *Oxid. Med. Cell. Longev.* 2016, 5698931. doi:10.1155/2016/5698931

ETHICS STATEMENT

The animal study was reviewed and approved by the Ethics Committee of the Federal University of Santa Catarina—UFSC (protocol number PP00745).

AUTHOR CONTRIBUTIONS

AO: writing—original draft, review & editing, synthesis, *in vitro* antioxidant tests, quantum studies, molecular docking, molecular properties, supervision; LL: synthesis; RN: supervision; RAY: supervision; CN-M: writing—original draft, *in vivo* tests; AS: quantum studies; DP-S: writing—original draft, molecular docking; MD-R: writing—original draft, molecular docking; LF: writing—review & editing, molecular modeling; AA: writing—review & editing, supervision; AS: writing—review & editing, supervision.

FUNDING

The National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES, funding code 001, and access to GOLD suite CSD-System Software through Dot.Lib Brazil), and the Sao Paulo Research Foundation (FAPESP, CIBFar grant 2013/07600-3), Brazil.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2021.788850/full#supplementary-material>

- Boden, W. E., Padala, S. K., Cabral, K. P., Buschmann, I. R., and Sidhu, M. S. (2015). Role of Short-Acting Nitroglycerin in the Management of Ischemic Heart Disease. *Drug Des. Devel. Ther.* 9, 4793–4805. doi:10.2147/DDDT.S79116
- Burgess, G., and Williams, D. (2010). The Discovery and Development of Analgesics: New Mechanisms, New Modalities. *J. Clin. Invest.* 120, 3753–3759. doi:10.1172/JCI43195
- Cartmell, S. M., Gelgor, L., and Mitchell, D. (1991). A Revised Rotarod Procedure for Measuring the Effect of Antinociceptive Drugs on Motor Function in the Rat. *J. Pharmacol. Methods* 26, 149–159. doi:10.1016/0160-5402(91)90063-b
- Cavalcante Melo, F. H., Rios, E. R. V., Rocha, N. F. M., Cito, M. D. C. D. O., Fernandes, M. L., and De Sousa, D. P. (2012). Antinociceptive Activity of Carvacrol (5-Isopropyl-2-Methylphenol) in Mice. *J. Pharm. Pharmacol.* 64, 1722–1729. doi:10.1111/j.2042-7158.2012.01552.x
- Coderre, T. J., and Van Empel, I. (1994). The Utility of Excitatory Amino Acid (EAA) Antagonists as Analgesic Agents. II. Assessment of the Antinociceptive Activity of Combinations of Competitive and Non-Competitive NMDA Antagonists with Agents Acting at Allosteric-Glycine and Polyamine Receptor. *Pain* 59, 353–359. doi:10.1016/0304-3959(94)90021-3
- Dantas, B. P. V., Alves, Q. L., de Assis, K. S., Ribeiro, T. P., and de Almeida, M. M. (2015). Participation of the TRP Channel in the Cardiovascular Effects Induced by Carvacrol in Normotensive Rat. *Vascul. Pharmacol.* 67–69, 48–58. doi:10.1016/j.vph.2015.02.016
- de Oliveira, A. S., Llanes, L. C., Brighente, I. M. C., Nunes, R. J., Yunes, R. A., Máximo Junior, N., et al. (2016). New Sulfonamides Derived from Carvacrol:

- Compounds with High Antibacterial Activity against Resistant *Staphylococcus aureus* Strains. *J. Biosci. Med.* 4, 105–114. doi:10.4236/jbm.2016.47011
- de Souza, M. M., Andreolla, M. C., Ribeiro, T. C., Gonçalves, A. E., Medeiros, A. R., de Souza, A. S., et al. (2020). Structure–Activity Relationships of Sulfonamides Derived from Carvacrol and Their Potential for the Treatment of Alzheimer's Disease. *RSC Med. Chem.* 11, 307–316. doi:10.1039/d0md00009d
- Di Meo, S., Reed, T. T., Venditti, P., and Victor, V. M. (2016). Role of ROS and RNS Sources in Physiological and Pathological Conditions. *Oxid. Med. Cell. Longev.* 2016, 1245049. doi:10.1155/2016/1245049
- Dickenson, A. H. (1995). Spinal Cord Pharmacology of Pain. *Br. J. Anaesth.* 75, 193–200. doi:10.1093/bja/75.2.193
- Do, H. T., Li, H., Chreifi, G., Poulos, T. L., and Silverman, R. B. (2019). Optimization of Blood-Brain Barrier Permeability with Potent and Selective Human Neuronal Nitric Oxide Synthase Inhibitors Having a 2-Aminopyridine Scaffold. *J. Med. Chem.* 62, 2690–2707. doi:10.1021/acs.jmedchem.8b02032
- Förstermann, U., and Sessa, W. C. (2012). Nitric Oxide Synthases: Regulation and Function. *Eur. Heart J.* 33 (7), 829–837. doi:10.1093/eurheartj/ehr304
- Fundytus, M. E. (2001). Glutamate Receptors and Nociception: Implications for the Drug Treatment of Pain. *CNS Drugs* 15, 29–58. doi:10.2165/00023210-200115010-00004
- Fung, T., Asiri, Y. I., Taheri, K., Wall, R., Schwarz, S. K. W., Puil, E., et al. (2019). Antinociception by Intrathecal Delivery of the Novel Non-Opioid 1-Amino-1-Cyclobutanecarboxylic Acid. *Eur. J. Pain* 23, 260–271. doi:10.1002/ejp.1301
- Gordh, T., Karlsten, R., and Kristensen, J. (1995). Intervention with Spinal NMDA, Adenosine, and NO Systems for Pain Modulation. *Ann. Med.* 27, 229–234. doi:10.3109/07853899509031964
- Guimarães, A. G., Xavier, M. A., de Santana, M. T., Camargo, E. A., Santos, C. A., and Brito, F. A. (2012). Carvacrol Attenuates Mechanical Hypernociception and Inflammatory Response. *Naunyn Schmiedeberg's Arch. Pharmacol.* 385, 253–263. doi:10.1007/s00210-011-0715-x
- Guindon, J., Walczak, J. S., and Beaulieu, P. (2007). Recent Advances in the Pharmacological Management of Pain. *Drugs* 67, 2121–2133. doi:10.2165/00003495-200767150-00002
- Halgren, T. A. (1996). Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput.* 17, 490–519. doi:10.1002/(sici)1096-987x(199604)17:5/6<490::aid-jcc1>3.0.co;2-p
- Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., and Hutchison, G. R. (2012). Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *J. Cheminform.* 4, 17. doi:10.1186/1758-2946-4-17
- Hazim, A. I., Ramanathan, S., Parthasarathy, S., Muzaimi, M., and Mansor, S. M. (2014). Anxiolytic-like Effects of Mitragnyne in the Open-Field and Elevated Plus-Maze Tests in Rats. *J. Physiol. Sci.* 64, 161–169. doi:10.1007/s12576-014-0304-0
- Hitchcock, S. A., and Pennington, L. D. (2006). Structure-Brain Exposure Relationships. *J. Med. Chem.* 49, 7559–7583. doi:10.1021/jm060642i
- Hückel, E. (1931). Quantentheoretische Beiträge Zum Benzolproblem - I. Die Elektronenkonfiguration Des Benzols Und Verwandter Verbindungen. *Z. für Phys.* 70, 204–286. doi:10.1007/BF01339530
- Jensen, T. S., Gottrup, H., Kasch, H., Nikolajsen, L., Terkelsen, A. J., and Witting, N. (2001). Has Basic Research Contributed to Chronic Pain Treatment? *Acta Anaesthesiol. Scand.* 45, 1128–1135. doi:10.1034/j.1399-6576.2001.450913.x
- Koehn, F. E., and Carter, G. T. (2005). The Evolving Role of Natural Products in Drug Discovery. *Nat. Rev. Drug Discov.* 4, 206–220. doi:10.1038/nrd1657
- Korb, O., Stützel, T., and Exner, T. E. (2009). Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* 49, 84–96. doi:10.1021/ci800298z
- Kuo, P. J., Hung, T. F., Lin, C. Y., Hsiao, H. Y., Fu, M. W., and Hong, P. (2017). Carvacrol Ameliorates Ligation-Induced Periodontitis in Rats. *J. Periodontol.* 88, e120–e128. doi:10.1902/jop.2017.160618
- Ling, C. C., Lui, L. Y., and So, W. K. (2012). Do Educational Interventions Improve Cancer Patients' Quality of Life and Reduce Pain Intensity? Quantitative Systematic Review. *J. Adv. Nurs.* 68, 511–520. doi:10.1111/j.1365-2648.2011.05841.x
- Loeser, J. D., and Treede, R. D. (2008). The Kyoto Protocol of IASP Basic Pain Terminology. *Pain* 137, 473–477. doi:10.1016/j.pain.2008.04.025
- Maioli, N. A., Zarpelon, A. C., Mizokami, S. S., Calixto-Campos, C., Guazelli, C. F. S., and Hohmann, M. S. N. (2015). The Superoxide Anion Donor, Potassium Superoxide, Induces Pain and Inflammation in Mice through Production of Reactive Oxygen Species and Cyclooxygenase-2. *Braz. J. Med. Biol. Res.* 48, 321–331. doi:10.1590/1414-431X20144187
- Manchope, M. F., Calixto-Campos, C., Coelho-Silva, L., Zarpelon, A. C., Pinho-Ribeiro, F. A., and Georgetti, S. R. (2016). Naringenin Inhibits Superoxide Anion-Induced Inflammatory Pain: Role of Oxidative Stress, Cytokines, Nrf-2 and the No-CGMP-PKG-KATP Channel Signaling Pathway. *PLoS One* 11, e0153015. doi:10.1371/journal.pone.0153015
- Marsh, N., and Marsh, A. (2000). The Short History of Nitroglycerine and Nitric Oxide in Pharmacology and Physiology. *Clin. Exp. Pharmacol. Physiol.* 27, 313–319. doi:10.1046/j.1440-1681.2000.03240.x
- Melo, F. H. C., Venâncio, E. T., De Sousa, D. P., de Franca Fonteles, M. M., De Vasconcelos, S. M. M., Viana, G. S. B., et al. (2010). Anxiolytic-like Effect of Carvacrol (5-Isopropyl-2-Methylphenol) in Mice: Involvement with GABAergic Transmission. *Fundam. Clin. Pharmacol.* 24, 437–443. doi:10.1111/j.1472-8206.2009.00788.x
- Meotti, F. C., Coelho Idos, S., and Santos, A. R. (2010). The Nociception Induced by Glutamate in Mice is Potentiated by Protons Released into the Solution. *J. Pain* 11, 570–578. doi:10.1016/j.jpain.2009.09.012
- Millan, M. J., and Seguin, L. (1993). (+)-HA 966, a Partial Agonist at the Glycine Site Coupled to NMDA Receptors, Blocks Formalin-Induced Pain in Mice. *Eur. J. Pharmacol.* 238, 445–447. doi:10.1016/0014-2999(93)90884-k
- Moncada, S., and Higgs, E. A. (2006). The Discovery of Nitric Oxide and Its Role in Vascular Biology. *Br. J. Pharmacol.* 147 (Suppl. 1), S193–S201. doi:10.1038/sj.bjp.0706458
- Moncada, S., Rees, D. D., Schulz, R., and Palmer, R. M. J. (1991). Development and Mechanism of a Specific Supersensitivity to Nitrovasodilators after Inhibition of Vascular Nitric Oxide Synthesis *In Vivo*. *Proc. Natl. Acad. Sci. U. S. A.* 88, 2166–2170. doi:10.1073/pnas.88.6.2166
- Nageswari, G., George, G., Ramalingam, S., and Govindarajan, M. (2018). Electronic and Vibrational Spectroscopic (FT-IR and FT-Raman) Investigation Using Ab Initio (HF) and DFT (B3LYP and B3PW91) and HOMO/LUMO/MEP Analysis on the Structure of L-Serine Methyl Ester Hydrogen Chloride. *J. Mol. Struct.* 1166, 422–441. doi:10.1016/j.molstruc.2018.04.014
- Nucci-Martins, C., Nascimento, L. F., Venzke, D., Brethanha, L. C., Sako, A. V. F., Oliveira, A. S., et al. (2016). Antinociceptive Effect of Hydroalcoholic Extract and Isoflavone Isolated from Polygala Molluginifolia in Mice: Evidence for the Involvement of Opioid Receptors and TRPV1 and TRPA1 Channels. *Phytomedicine* 23, 429–440. doi:10.1016/j.phymed.2016.02.002
- Oliveira, A. S. de., de Souza, L. F. S., Nunes, R. J., Johann, S., Palomino-Salcedo, D. L., Ferreira, L. L. G., et al. (2020). Antioxidant and Antibacterial Activity of Sulfonamides Derived from Carvacrol: A Structure-Activity Relationship Study. *Curr. Top. Med. Chem.* 20, 173–181. doi:10.2174/1568026619666191127144336
- Pacher, P., Beckman, J. S., and Liaudet, L. (2007). Nitric Oxide and Peroxynitrite in Health and Disease. *Physiol. Rev.* 87, 315–424. doi:10.1152/physrev.00029.2006
- Pajouhesh, H., and Lenz, G. R. (2005). Medicinal Chemical Properties of Successful Central Nervous System Drugs. *NeuroRx* 2, 541–553. doi:10.1602/neurorx.2.4.541
- Pereira, A. C., Paulo, M., Araújo, A. V., Rodrigues, G. J., and Bendhack, L. M. (2011). Nitric Oxide Synthesis and Biological Functions of Nitric Oxide Released from Ruthenium Compounds. *Braz. J. Med. Biol. Res.* 44, 947–957. doi:10.1590/s0100-879x2011007500084
- Phaniendra, A., Jestadi, D. B., and Periyasamy, L. (2015). Free Radicals: Properties, Sources, Targets, and Their Implication in Various Diseases. *Indian J. Clin. Biochem.* 30, 11–26. doi:10.1007/s12291-014-0446-0
- Pires, P. W., Sullivan, M. N., Pritchard, H. A. T., Robinson, J. J., and Earley, S. (2015). Unitary TRPV3 Channel Ca²⁺ Influx Events Elicit Endothelium-Dependent Dilation of Cerebral Parenchymal Arterioles. *Am. J. Physiol. Circ. Physiol.* 309, H2031–H2041. doi:10.1152/ajpheart.00140.2015
- Planells-Cases, R., Perez-Paya, E., Messeguer, A., Carreno, C., and Ferrer-Montiel, A. (2003). Small Molecules Targeting the NMDA Receptor Complex as Drugs for Neuropathic Pain. *Mini Rev. Med. Chem.* 3, 749–756. doi:10.2174/1389557033487782
- Ramirez, D., and Caballero, J. (2018). Is it Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? *Molecules* 23, 1038. doi:10.3390/molecules23051038
- Rao, K. S. (2009). Free Radical Induced Oxidative Damage to DNA: Relation to Brain Aging and Neurological Disorders. *Indian J. Biochem. Biophys.* 46, 9–15.

- Ravn, P., Madhurantakam, C., Kunze, S., Matthews, E., Priest, C., O'Brien, S., et al. (2013). Structural and Pharmacological Characterization of Novel Potent and Selective Monoclonal Antibody Antagonists of Glucose-Dependent Insulinotropic Polypeptide Receptor. *J. Biol. Chem.* 288, 19760–19772. doi:10.1074/jbc.M112.426288
- Rodrigues, T., Reker, D., Schneider, P., and Schneider, G. (2016). Counting on Natural Products for Drug Design. *Nat. Chem.* 8, 531–541. doi:10.1038/nchem.2479
- Rowlingson, J. C. (2000). Textbook of Pain. *Anesth. Analg.* 91, 1315. doi:10.1213/0000539-200011000-00066
- Salvemini, D., Little, J. W., Doyle, T., and Neumann, W. L. (2011). Roles of Reactive Oxygen and Nitrogen Species in Pain. *Free Radic. Biol. Med.* 51, 951–966. doi:10.1016/j.freeradbiomed.2011.01.026
- Schmidt, B. L., Hamamoto, D. T., Simone, D. A., and Wilcox, G. L. (2010). Mechanism of Cancer Pain. *Mol. Interv.* 10, 164–178. doi:10.1124/mi.10.3.7
- Schmidt, M. W., Baldridge, K. K., Boatz, J. A., Elbert, S. T., Gordon, M. S., Jensen, J., et al. (1993). General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* 14, 1347–1363. doi:10.1002/jcc.540141112
- Sens, L., De Oliveira, A. S., Mascarello, A., Brighente, I. M. C., Yunes, R. A., and Nunes, R. J. (2018). Synthesis, Antioxidant Activity, Acetylcholinesterase Inhibition and Quantum Studies of Thiosemicarbazones. *J. Braz. Chem. Soc.* 29, 343–352. doi:10.21577/0103-5053.20170146
- Silva, J. C., Raquel, S., Lima, G. De., Gonçalves, R., Júnior, D. O., and Guedes, R. (2013). Modelos Experimentais Para Avaliação da Atividade Antinociceptiva de Produtos Naturais : Uma Revisão. *Rev. Bras. Farm.* 94, 18–23.
- Snider, W. D., and McMahon, S. B. (1998). Tackling Pain at the Source: New Ideas about Nociceptors. *Neuron* 20, 629–632. doi:10.1016/s0896-6273(00)81003-x
- Tominaga, M., Numazaki, M., Iida, T., and Tominaga, T. (2003). Molecular Mechanisms of Nociception. *Nihon Shinkei Seishin Yakurigaku Zasshi* 23, 139–147.
- Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* 45, 2615–2623. doi:10.1021/jm020017n
- Verri, W. A., Cunha, T. M., Parada, C. A., Poole, S., Cunha, F. Q., and Ferreira, S. H. (2006). Hypernociceptive Role of Cytokines and Chemokines: Targets for Analgesic Drug Development? *Pharmacol. Ther.* 112, 116–138. doi:10.1016/j.pharmthera.2006.04.001
- Wang, Z. Q., Porreca, F., Cuzzocrea, S., Galen, K., Lightfoot, R., Masini, E., et al. (2004). A Newly Identified Role for Superoxide in Inflammatory Pain. *J. Pharmacol. Exp. Ther.* 309, 869–878. doi:10.1124/jpet.103.064154
- Wood, P. L., Mahmood, S. A., and Moskal, J. R. (2008). Antinociceptive Action of GLYX-13: An N-Methyl-D-Aspartate Receptor Glycine Site Partial Agonist. *Neuroreport* 19, 1059–1061. doi:10.1097/WNR.0b013e32830435c9
- Zhuo, M., and Gebhart, G. F. (1997). Biphasic Modulation of Spinal Nociceptive Transmission From the Medullary Raphe Nuclei in the Rat. *J. Neurophysiol.* 78, 746–758. doi:10.1152/jn.1997.78.2.746
- Zimmermann, M. (1983). Ethical Guidelines for Investigations of Experimental Pain in Conscious Animals. *Pain* 16, 109–110. doi:10.1016/0304-3959(83)90201-4
- Zotti, M., Colaianna, M., Morgese, M. G., Tucci, P., Schiavone, S., Avato, P., et al. (2013). Carvacrol: From Ancient Flavoring to Neuromodulatory Agent. *Molecules* 18, 6161–6172. doi:10.3390/molecules18066161

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 de Oliveira, Llanes, Nunes, Nucci-Martins, de Souza, Palomino-Salcedo, Dávila-Rodríguez, Ferreira, Santos and Andricopulo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



3D-QSAR, Molecular Docking, and MD Simulations of Anthraquinone Derivatives as PGAM1 Inhibitors

Yuwei Wang^{1*}, Yifan Guo¹, Shaojia Qiang², Ruyi Jin¹, Zhi Li¹, Yuping Tang¹, Elaine Lai Han Leung^{3,4}, Hui Guo^{1*} and Xiaojun Yao^{3,4*}

¹College of Pharmacy, Shaanxi University of Chinese Medicine, Xianyang, China, ²School of Pharmacy, Lanzhou University, Lanzhou, China, ³Dr. Neher's Biophysics Laboratory for Innovative Drug Discovery, Macau University of Science and Technology, Macau, China, ⁴State Key Laboratory of Quality Research in Chinese Medicine, Macau University of Science and Technology, Macau, China

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Teodorico Castro Ramalho,
Universidade Federal de Lavras, Brazil
Dharmendra Kumar Yadav,
Gachon University, South Korea

*Correspondence:

Yuwei Wang
wangyw@sntcm.edu.cn
Hui Guo
guohui@sntcm.edu.cn
Xiaojun Yao
xjyao@must.edu.mo

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 25 August 2021

Accepted: 01 November 2021

Published: 25 November 2021

Citation:

Wang Y, Guo Y, Qiang S, Jin R, Li Z, Tang Y, Leung ELH, Guo H and Yao X (2021) 3D-QSAR, Molecular Docking, and MD Simulations of Anthraquinone Derivatives as PGAM1 Inhibitors. *Front. Pharmacol.* 12:764351. doi: 10.3389/fphar.2021.764351

PGAM1 is overexpressed in a wide range of cancers, thereby promoting cancer cell proliferation and tumor growth, so it is gradually becoming an attractive target. Recently, a series of inhibitors with various structures targeting PGAM1 have been reported, particularly anthraquinone derivatives. In present study, the structure–activity relationships and binding mode of a series of anthraquinone derivatives were probed using three-dimensional quantitative structure–activity relationships (3D-QSAR), molecular docking, and molecular dynamics (MD) simulations. Comparative molecular field analysis (CoMFA, $r^2 = 0.97$, $q^2 = 0.81$) and comparative molecular similarity indices analysis (CoMSIA, $r^2 = 0.96$, $q^2 = 0.82$) techniques were performed to produce 3D-QSAR models, which demonstrated satisfactory results, especially for the good predictive abilities. In addition, molecular dynamics (MD) simulations technology was employed to understand the key residues and the dominated interaction between PGAM1 and inhibitors. The decomposition of binding free energy indicated that the residues of F22, K100, V112, W115, and R116 play a vital role during the ligand binding process. The hydrogen bond analysis showed that R90, W115, and R116 form stable hydrogen bonds with PGAM1 inhibitors. Based on the above results, 7 anthraquinone compounds were designed and exhibited the expected predictive activity. The study explored the structure–activity relationships of anthraquinone compounds through 3D-QSAR and molecular dynamics simulations and provided theoretical guidance for the rational design of new anthraquinone derivatives as PGAM1 inhibitors.

Keywords: PGAM1, molecular docking, molecular dynamics simulation, CoMFA, CoMSIA

INTRODUCTION

Reprogramming energy metabolism has been regarded as one of the 10 essential hallmarks of cancer cells (Hanahan and Weinberg, 2011), which was called the “Warburg effect.” In 1924, Warburg found that cancer cells are more likely to metabolize glucose by means of aerobic glycolysis instead of oxidative phosphorylation as in normal cells (Wang et al., 2018a; Huang et al., 2019b). Cancer metabolic reprogramming is the performance of adapting to the environment during tumor formation or metastasis. More and more scientists are focusing on the pivotal enzymes in the metabolic reprogramming of cancer cells in order to find new cancer treatment targets (Wang et al., 2018b).

Phosphoglycerate mutase 1 (PGAM1) is a key enzyme that catalyzes the invertible conversion of 3-phosphoglycerate (3-PG) and 2-phosphoglycerate (2-PG) during the process of glycolysis (Fothergill-Gilmore and Watson, 1989). Recent studies have proven that once the expression of PGAM1 is upregulated, it will promote tumor cell proliferation and tumor growth in coordination with glycolysis and biosynthesis (Hitosugi et al., 2012). PGAM1 regulates the proliferation of cancer cells in term of biosynthesis regulation, partly by regulating intracellular levels of its product 2-PG and 3-PG (Hitosugi et al., 2012). In the oxidative pentose phosphate pathway (PPP), 3-PG inhibits 6-phosphogluconate dehydrogenase after binding, while 2-PG feedback control of the levels of through activates 3-phosphoglycerate dehydrogenase. In addition, PGAM1 is overexpressed in multiple cancers (Li and Liu, 2020), including ovarian cancer (Zhang et al., 2020), non-small-cell lung cancer (NSCLC) (Li et al., 2020), colorectal cancer (Liu et al., 2008; Lei et al., 2011), pancreatic ductal adenocarcinoma (PDAC) (Liu et al., 2018), prostate cancer (PCa) (Wen et al., 2018), and glioma (Xu et al., 2016). Particularly, high expression of PGAM1 was associated with poor prognosis in NSCLC patients (Sun et al., 2018; Li et al., 2020). Downregulation of the expression of PGAM1 or suppression of its metabolic activity will lead to weakened cell proliferation and tumor growth (Hitosugi et al., 2012; Peng et al., 2016; Liu et al., 2018). Thus, PGAM1 is considered to be an emerging target for cancer treatment.

Due to the important role of PGAM1 in the occurrence and development of tumors, many researchers have focused on the discovery and characterization of small molecules that can target and modulate the metabolic activity of PGAM1 (Huang et al., 2019a). MJE3 was first revealed as a covalent PGAM1 inhibitor on Lys 100 by the Cravatt group in 2005 (Evans et al., 2005). (-)-Epigallocatechin-3-gallate (EGCG) is a natural product extracted from green tea, which was first discovered as a non-substrate competitive PGAM1 inhibitor with potent inhibition activity against PGAM1 (Li et al., 2017). Anthraquinone derivatives PGMI-004A (Hitosugi et al., 2012) and xanthone derivatives (Wang et al., 2018b) were identified as allosteric PGAM1 inhibitors by the Zhou group, which exhibited moderate inhibition activity on PGAM1. As another anthraquinone derivative, HKB99 was identified to allosterically obstruct the activation of PGAM1, thereby affecting its catalytic activity and the intermolecular interaction of ACTA2 (Huang et al., 2019c; Liang et al., 2021). Based on the excellent anticancer activity of PGMI-004A and HKB99, new small molecules with the anthraquinone core have been synthesized, which may have similar mechanisms of action and therapeutic potential. Therefore, the design and development of novel small molecules with an anthraquinone core targeting PGAM1 may prove to be an effective strategy for the treatment of cancer cells.

Computer-aided drug design is an effective tool in the drug discovery and design process. It can not only be used to predict the activity of small molecules, explain the action mechanism, and provide guidance for the design of more effective drug molecules but also reduce the consumption of manpower and material resources (Jorgensen, 2004). To elucidate the

structure–activity relationships and provide optimization guidance for anthraquinone derivatives, 62 collected compounds were employed to construct 3D-QSAR models using CoMFA and CoMSIA methods. According to the contour maps by 3D-QSAR and the crucial residues by MD simulations, 7 compounds with high predictive activity were designed. This study will provide a valuable theoretical basis for the activity prediction and structural modification of targeted PGAM1 inhibitors containing anthraquinone structures.

MATERIALS AND METHODS

Data Sets and Preparation

In order to ensure the reliability of activity values and reduce accidental errors, a set of 78 PGAM1 inhibitors were retrieved from different literature sources in terms of the same group (Wang et al., 2018a; Wang et al., 2018b; Huang et al., 2019a; Huang et al., 2019b). The molecular structure and experimental bioactivity of all chemicals are listed in **Table 1**. First, corresponding IC_{50} values of experimental bioactivity expressed in nM were converted into negative logarithm ($-\lg IC_{50}$) and acted as the dependent variable for the QSAR modeling. According to the diversity of the molecular structure and activities, all compounds were split into a training set and a test set at a ratio of approximately 4:1. Finally, 62 compounds were selected randomly as the training set and the remaining 16 compounds as the test set. The molecular structure of each compound was determined using ChemDraw 18.0 and then imported to SYBYL 6.9 (SYBYL, XX) to minimize the energy based on the Tripos force field with a convergence criterion of 0.01 kcal/mol. The Gasteiger–Hückel method was employed to calculate the partial atomic charges. Then, the multisearch strategy was performed to obtain the lowest energy conformation, and the lowest energy geometry after being filled with energy was reserved for alignment.






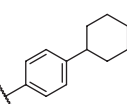
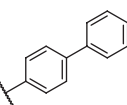

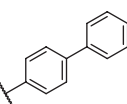
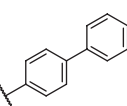
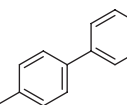
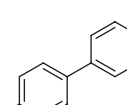
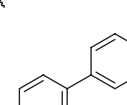
Molecular Alignment

Molecular alignment in terms of the same structure is considered to be one of the most significant elements in the process of built 3D-QSAR modeling. Hence, molecular alignment based on the most active molecule, 35, was employed by atom-by-atom fits. After a common substructure is set, the dominant conformations of the remaining 77 compounds are selected for superimposition.

Construction of CoMFA and CoMSIA Models

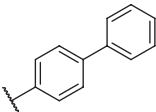
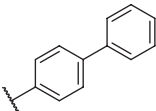
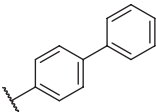
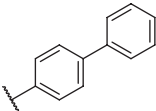
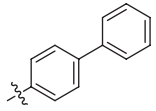





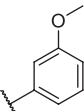
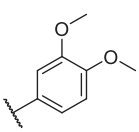
The 3D-QSAR model for the training set compound was built after alignment by using SYBYL 6.9 software. The CoMFA (Cramer et al., 1988b) and CoMSIA (Cramer et al., 1988b) are the most widely used methods for constructing 3D-QSAR. The CoMFA and CoMSIA descriptors were obtained by placing the superposed compound in a 3D cubic lattice with a grid spacing of 2 Å. Using the SP^3 hybrid carbon as the probe atom, the Lennard–Jones and the coulomb potential were applied to obtain the steric field energy and electrostatic field energy of

TABLE 1 | Structure and corresponding activity data of reported PGAM1 inhibitors.

| Number | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ | X | IC ₅₀ (μM) | pIC ₅₀ |
|-----------------|---|----------------|----------------|-------------------|----------------|-------------------|---|--------------------------|-------------------|
| 1 |  | H | OH | H | H | H | O | 10.10 | 5.00 |
| 2 |  | H | OH | H | H | H | O | 13.20 | 4.88 |
| 3 ^a |  | H | OH | H | H | H | O | 6.40 | 5.19 |
| 4 |  | H | OH | H | H | H | O | 10.2 | 4.99 |
| 5 |  | H | OH | H | H | H | O | 8.40 | 5.08 |
| 6 |  | H | OH | H | H | H | O | 5.90 | 5.23 |
| 7 |  | H | OH | H | H | H | O | 5.50 | 5.26 |
| 8 |  | H | OH | H | H | H | O | 6.00 | 5.22 |
| 9 ^a |  | H | H | H | H | H | O | 14.3 | 4.84 |
| 10 |  | H | H | H | H | -OCH ₃ | O | 6.50 | 5.19 |
| 11 ^a |  | H | H | H | H | -CH ₃ | O | 8.60 | 5.07 |
| 12 |  | H | H | -OCH ₃ | H | H | O | 4.60 | 5.34 |
| 13 |  | H | H | -CH ₃ | H | H | O | 8.00 | 5.10 |

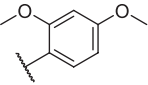
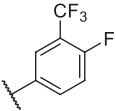
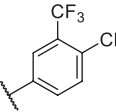
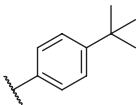
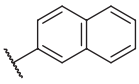
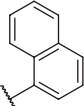
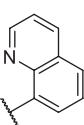
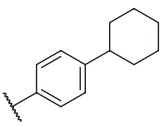
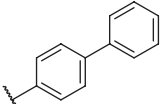
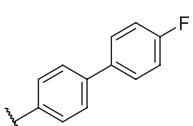
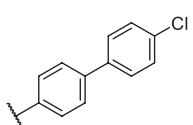
(Continued on following page)

TABLE 1 | (Continued) Structure and corresponding activity data of reported PGAM1 inhibitors.

| Number | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ | X | IC ₅₀ (μ M) | pIC ₅₀ |
|-----------------|---|------------------|----------------|---------------------|----------------|----------------|------|--------------------------------|-------------------|
| 14 |  | H | H | Cl | H | H | O | 3.50 | 5.46 |
| 15 |  | H | H | F | H | H | O | 13.7 | 4.86 |
| 16 |  | H | H | -NO ₂ | H | H | O | 2.10 | 5.68 |
| 17 |  | H | H | OH | H | H | O | 6.40 | 5.19 |
| 18 |  | H | H | -COOCH ₃ | H | H | O | 2.70 | 5.57 |
| 19 |  | -CH ₃ | OH | H | H | H | -C=O | 5.37 | 5.27 |
| 20 | | OH | H | H | H | H | -C=O | 2.05 | 5.69 |
| 21 |  | OH | H | H | H | H | -C=O | 1.75 | 5.76 |
| 22 |  | OH | H | H | H | H | -C=O | 1.50 | 5.82 |
| 23 ^a |  | OH | H | H | H | H | -C=O | 0.36 | 6.44 |
| 24 |  | OH | H | H | H | H | -C=O | 0.84 | 6.08 |
| 25 |  | OH | H | H | H | H | -C=O | 0.55 | 6.26 |
| 26 |  | OH | H | H | H | H | -C=O | 0.48 | 6.32 |

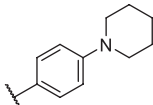
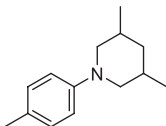
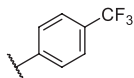
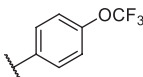
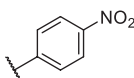
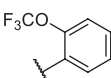
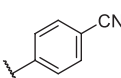
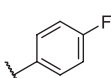

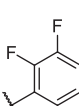
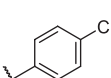
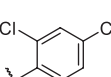

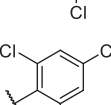
(Continued on following page)

TABLE 1 | (Continued) Structure and corresponding activity data of reported PGAM1 inhibitors.

| Number | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ | X | IC ₅₀ (μ M) | pIC ₅₀ |
|-----------------|---|----------------|----------------|----------------|----------------|----------------|------|--------------------------------|-------------------|
| 27 |  | OH | H | H | H | H | -C=O | 2.81 | 5.55 |
| 28 |  | OH | H | H | H | H | -C=O | 2.86 | 5.54 |
| 29 |  | OH | H | H | H | H | -C=O | 0.63 | 6.20 |
| 30 |  | OH | H | H | H | H | -C=O | 0.55 | 6.26 |
| 31 |  | OH | H | H | H | H | -C=O | 0.49 | 6.31 |
| 32 ^a |  | OH | H | H | H | H | -C=O | 0.19 | 6.72 |
| 33 ^a |  | OH | H | H | H | H | -C=O | 1.29 | 5.89 |
| 34 |  | OH | H | H | H | H | -C=O | 2.05 | 5.69 |
| 35 |  | OH | H | H | H | H | -C=O | 0.097 | 7.01 |
| 36 |  | OH | H | H | H | H | -C=O | 0.25 | 6.60 |
| 37 ^a |  | OH | H | H | H | H | -C=O | 0.26 | 6.59 |

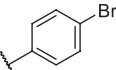


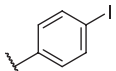

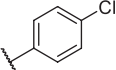
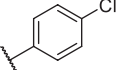
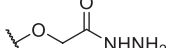

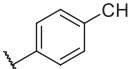
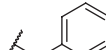
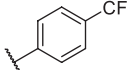
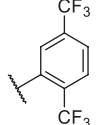

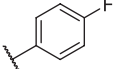
(Continued on following page)

TABLE 1 | (Continued) Structure and corresponding activity data of reported PGAM1 inhibitors.

| Number | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ | X | IC ₅₀ (μM) | pIC ₅₀ |
|-----------------|---|----------------|----------------|----------------|----------------|----------------|------|--------------------------|-------------------|
| 38 |  | OH | H | H | H | H | -C=O | 0.14 | 6.85 |
| 39 |  | OH | H | H | H | H | -C=O | 0.33 | 6.48 |
| 40 |  | OH | H | H | H | H | -C=O | 2.60 | 5.59 |
| 41 |  | OH | H | H | H | H | -C=O | 0.54 | 6.27 |
| 42 |  | OH | H | H | H | H | -C=O | 0.90 | 6.05 |
| 43 |  | OH | H | H | H | H | -C=O | 0.47 | 6.33 |
| 44 |  | OH | H | H | H | H | -C=O | 2.20 | 5.66 |
| 45 ^a |  | OH | H | H | H | H | -C=O | 0.61 | 6.21 |
| 46 |  | OH | H | H | H | H | -C=O | 0.54 | 6.27 |
| 47 |  | OH | H | H | H | H | -C=O | 0.79 | 6.10 |
| 48 ^a |  | OH | H | H | H | H | -C=O | 0.89 | 6.05 |
| 49 |  | OH | H | H | H | H | -C=O | 0.27 | 6.57 |
| 50 |  | OH | H | H | H | H | -C=O | 0.28 | 6.55 |
| 51 |  | OH | H | H | H | H | -C=O | 0.89 | 6.05 |

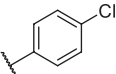
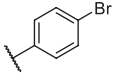
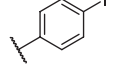
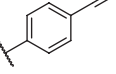
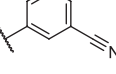
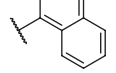
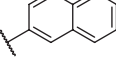
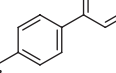
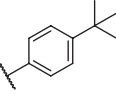
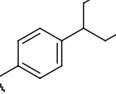
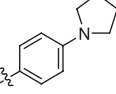
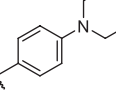
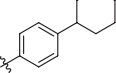
(Continued on following page)

TABLE 1 | (Continued) Structure and corresponding activity data of reported PGAM1 inhibitors.

| Number | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ | X | IC ₅₀ (μ M) | pIC ₅₀ |
|-----------------|---|---|----------------|----------------|----------------|----------------|------|--------------------------------|-------------------|
| 52 |  | OH | H | H | H | H | -C=O | 0.90 | 6.05 |
| 53 |  | OH | H | H | H | H | -C=O | 0.26 | 6.59 |
| 54 |  | OH | H | H | H | H | -C=O | 0.20 | 6.70 |
| 55 |  | OH | H | H | H | H | -C=O | 0.35 | 6.46 |
| 56 |  | OH | H | H | H | H | -C=O | 0.47 | 6.33 |
| 57 |  | OH | H | H | H | H | -C=O | 0.26 | 6.59 |
| 58 |  | -OCH ₃ | H | H | H | H | -C=O | 2.92 | 5.53 |
| 59 |  |  | H | H | H | H | -C=O | 2.00 | 5.70 |
| 60 |  | OH | OH | H | H | H | O | 2.80 | 5.55 |
| 61 |  | OH | OH | H | H | H | O | 7.20 | 5.14 |
| 62 |  | OH | OH | H | H | H | O | 1.90 | 5.72 |
| 63 |  | OH | OH | H | H | H | O | 3.50 | 5.46 |
| 64 |  | OH | OH | H | H | H | O | 6.30 | 5.20 |
| 65 ^a |  | OH | OH | H | H | H | O | 5.80 | 5.24 |

(Continued on following page)

TABLE 1 | (Continued) Structure and corresponding activity data of reported PGAM1 inhibitors.

| Number | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ | X | IC ₅₀ (μM) | pIC ₅₀ |
|-----------------|---|----------------|----------------|----------------|----------------|----------------|---|--------------------------|-------------------|
| 66 |  | OH | OH | H | H | H | O | 5.50 | 5.26 |
| 67 |  | OH | OH | H | H | H | O | 3.60 | 5.44 |
| 68 ^a |  | OH | OH | H | H | H | O | 2.90 | 5.54 |
| 69 ^a |  | OH | OH | H | H | H | O | 1.90 | 5.72 |
| 70 ^a |  | OH | OH | H | H | H | O | 4.20 | 5.38 |
| 71 ^a |  | OH | OH | H | H | H | O | 2.10 | 5.68 |
| 72 |  | OH | OH | H | H | H | O | 1.70 | 5.77 |
| 73 |  | OH | OH | H | H | H | O | 1.60 | 5.80 |
| 74 |  | OH | OH | H | H | H | O | 1.20 | 5.92 |
| 75 |  | OH | OH | H | H | H | O | 2.60 | 5.59 |
| 76 ^a |  | OH | OH | H | H | H | O | 0.50 | 6.30 |
| 77 ^a |  | OH | OH | H | H | H | O | 2.70 | 5.57 |
| 78 |  | OH | OH | H | H | H | O | 1.00 | 6.00 |

^aTest set for the validation of the 3D-QSAR model.

each lattice point. The contributions of the hydrogen bond acceptor field, hydrogen bond donor field, and hydrophobic field were calculated by the probe atom. The partial least squares method (Cramer et al., 1988a) was employed to deal with the linear correlation between the CoMFA and CoMSIA fields and biological activity. The cross-validation correlation coefficient (q^2) and optimum number of components (N) were obtained using the leave-one-out method for cross-validation analysis. In addition, the r_m^2 (Roy et al., 2013; Cardoso et al., 2016), r_{pred}^2 , external standard deviation error of prediction (SDEPext), and applicability domain (Roy et al., 2015; de Assis et al., 2016) were also calculated to evaluate the performance of built models.

Evaluation of the 3D-QSAR Models

The predictive capabilities of built 3D-QSAR models were evaluated *via* the test set of 16 compounds. After all compounds were superimposed upon compound 35, the pIC_{50} values of all compounds were estimated through the built CoMFA and CoMSIA models.

Molecular Docking

To obtain more accurate docking results, the resolution of all crystal structures of PGAM1 in complex with small molecules obtained from the RSCB Protein Data Bank (PDB) was compared, and 5Y35, with the best resolution of 1.99 Å, was preserved as the docking template. Subsequently, the Protein Preparation Wizard module within (Schrödinger, 2015) was utilized to preprocess the crystal structure, including adding hydrogens and side chains, deleting water molecules, and calculating partial charges and protonation states by using the OPLS2005 force field (Jorgensen et al., 1996). Then, a grid box centered at the native ligand with a similar size was produced to determine the binding pocket of PGAM1 by using the Grid Generation module of the Schrödinger package. All molecules were preprocessed using the LigPrep module implemented in the Schrödinger package, and the ionization states were calculated using Epik (Shelley et al., 2007) at pH = 7.0 ± 2.0. Finally, all chemicals were docked into the binding pocket of PGAM1 and evaluated using the standard precision (SP) mode of Glide. The scale factor was set at 0.8, and the partial charge intercept was set at 0.15. The 10,000 poses of each ligand during the initial docking phase were preserved for evaluation.

Molecular Dynamics Simulations

To obtain the structural basis and significant residues involved in the process of ligand binding, molecular dynamics simulations were employed in terms of the crystal structure of compounds 23 and 49 using Amber16 (Case et al., 2005). The general AMBER force field (GAFF) (Wang et al., 2004) was employed to parameterize the compounds, while the AMBER ff14SB force field (Maier et al., 2015) was employed for the PGAM1 structure. The partial charges of compounds were calculated by using the restrained electrostatic potential fitting procedure (Bayly et al., 1993; Cieplak et al., 1995; Fox and Kollman, 1998) based on the electrostatic potentials calculated using the Hartree–Fock (HF) method with the 6-31G* basis set in the Gaussian 09 package

(Frisch et al., 2009). Then, the complex was solvated in a cubic box of TIP3P waters, with the solute 10 Å away from the water box boundary. After adding sodium ions to neutralize each system, the steepest descent method followed by the conjugate-gradient method were employed to minimize the system every 2,500 steps. Subsequently, each system was heated in the NVT ensemble from 0 to 300 K in 50 ps restraint on backbone atoms. The restraint force was gradually decreased from 5 to 0.1 kcal/(mol Å²) within 0.9 ns. Under a periodic boundary condition, 50 ns MD simulations were performed at 300 K and 1 atm without any restraint. The particle mesh Ewald method (Linse and Linse, 2014) was used to calculate the long-range electrostatic interactions, and the SHAKE method (Ryckaert et al., 1977) was employed to constrain all covalent bonds containing hydrogen atoms.

Trajectory Analysis

After the MD simulation finished, trajectories were dissected *via* the Cpptraj module (Roe and Cheatham, 2013) in AmberTools 16. First, the root mean square deviations (RMSDs) value was calculated in terms of the last 10 ns of each MD trajectory. Second, the molecular mechanics/generalized born surface area (MM/GBSA) approach (Massova and Kollman, 2000) was applied to calculate the binding free energy. After withdrawing a total of 2,500 snapshots, the MM/GBSA calculation was executed on each snapshot. The binding free energy (ΔG_{bind}) was calculated as follows (Hou et al., 2011; Sun et al., 2014):

$$\Delta G_{bind} = G_{complex} - (G_{protein} + G_{ligand})$$

where the energy term (G) is estimated as follows:

$$G = E_{vdw} + E_{ele} + G_{GB} + G_{GBSUR}$$

In the equations above, the E_{vdw} , E_{ele} , G_{GB} , and G_{GBSUR} represent van der Waals, electrostatic energy, the electrostatic contribution to the solvation free energy, and non-polar contribution to the solvation free energy, respectively. The changes of conformational entropy were ignored. Moreover, the total free energy was decomposed to each residue in PGAM1 to obtain the crucial residues contributed to the ligand binding process.

RESULTS AND DISCUSSION

CoMFA and CoMSIA Models

In the present study, a series of 78 PGAM1 inhibitors were obtained. The molecular structures and pIC_{50} values of all molecules are listed in **Table 1**. The quality of molecular superposition is considered to be one of the important factors affecting 3D-QSAR prediction accuracy (Cho et al., 1996). On the basis of the structure and bioactivity of PGAM1 inhibitors, the compounds in the training set were aligned to compound 35, which had the highest activity based on the common substructure. It can be seen from **Figure 1** that the common skeleton of all molecules is overlapped. However, the side chains of several compounds surround the common skeleton due to the

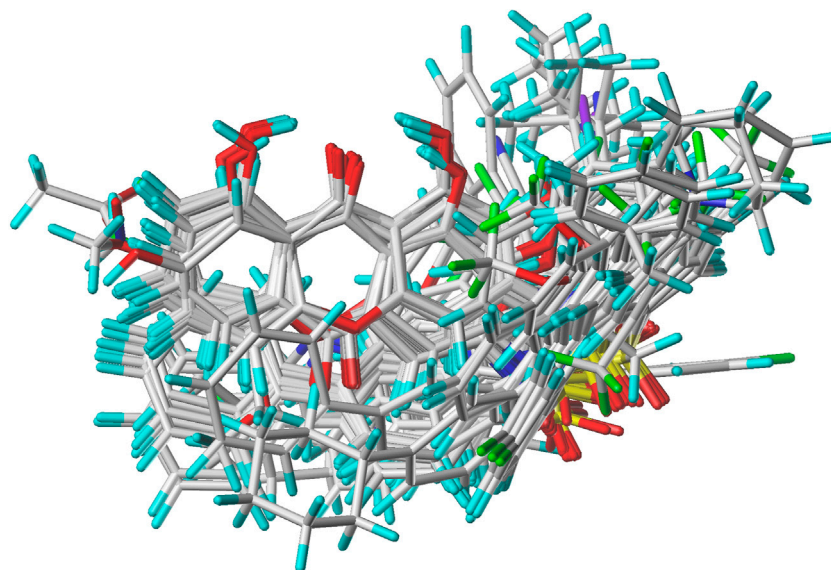


FIGURE 1 | Structural alignment of all the molecules in the training set based on the common substructure of compound 35.

large difference. Then, the 3D-QSAR models of CoMFA and CoSIA were successfully developed.

To examine the predictive ability and reliability of the built model, q^2 and r^2 were applied to evaluate the predictive power of the built 3D-QSAR model, r^2 , F , and SEE values were employed to assess the reliability of the model, and r_m^2 , r_{pred}^2 , and $SDEP_{ext}$ values were utilized for external validation of the model. **Table 2** lists the classical parameter statistics of CoMFA and CoMSIA models. In general, $r^2 > 0.7$ and $q^2, r_m^2, r_{pred}^2 > 0.5$ are necessary for a good model (Pratim Roy et al., 2009). As shown in **Table 2**, the values of q^2 , N , SEE , r^2 , r_m^2 , r_{pred}^2 , $SDEP_{ext}$, and F are 0.81, 6, 0.106, 0.97, 0.78, 0.89, 0.22, and 258.06, respectively. The results show that the built CoMFA model exhibits a good stability and predictive ability. The contribution of the steric field and the electrostatic field is 81 and 19%, respectively, indicating that the biological activity of compounds is more affected by the steric field. In addition, the predicted activity of the new chemical is only valid when the predicted compound falls within the applicability domain of the developed model (Roy et al., 2015). The calculated results show that all compounds are within the application domain of the built CoMFA model, so this prediction result is reliable.

Different field combinations of CoMSIA models were constructed, and it had been proved that CoMSIA-SEHA is the best model. Based on this model, the values of q^2 , N , SEE , r^2 , r_m^2 , r_{pred}^2 , $SDEP_{ext}$, and F are 0.82, 6, 0.11, 0.96, 0.79, 0.89, 0.23, and 228.71, respectively. In this model, the contribution of the steric field is 20%, that of the electrostatic field is 22%, that of the hydrophobic field is 40%, and that of the hydrogen bond acceptor field is 18%, respectively. The results show that the hydrophobic field has a greater effect on the bioactivity of the PGAM1 inhibitors. The calculation results of the application domain show that almost all the compounds are within the application domain of the CoSIA model, except for compound 24 with an

TABLE 2 | Summary of CoMFA and CoMSIA models.

| PLS statistics | CoMFA | CoMSIA |
|------------------------|--------|--------|
| q^2 | 0.81 | 0.82 |
| N | 6 | 6 |
| r^2 | 0.97 | 0.96 |
| F | 258.06 | 228.71 |
| r_m^2 | 0.78 | 0.79 |
| r_{pred}^2 | 0.89 | 0.89 |
| $SDEP_{ext}$ | 0.22 | 0.23 |
| SEE | 0.11 | 0.11 |
| Steric | 0.81 | 0.20 |
| Electrostatic | 0.19 | 0.22 |
| Hydrophobic | - | 0.40 |
| Hydrogen bond acceptor | - | 0.18 |

S_{new} of 3.87 and compound 25 with an S_{new} of 4.06. By analyzing the descriptors in CoMSIA, we found that compounds 24 and 25 have the largest electrostatic field contribution. The experimental and predicted values of the biological activity of the training set and the test set in the established CoMFA and CoMSIA models are shown in **Table 3**.

The scatter plot of the experimental and predicted values of the studied PGAM1 inhibitor is shown in **Figure 2**. It can be seen from **Figure 2** that the experimental and predicted bioactivity values of all molecules are distributed around the $Y = X$ equation, indicating that the predicted values are in good accord with the experimental values, which further demonstrates that the model has good predictive ability.

Contour Maps Analysis of CoMFA and CoMSIA

The structure–activity relationships between PGAM1 inhibitors and activity can be well demonstrated by using 3D contour maps

TABLE 3 | Experimental pIC₅₀ (Exp.), predicted pIC₅₀ (Pred.), and corresponding residuals (Res.) of the anthraquinone derivatives.

| Number | pIC ₅₀ | CoMFA | | CoMSIA | |
|--------|-------------------|-------|-------|--------|-------|
| | Exp | Pred | Res | Pred | Res |
| 1 | 5.00 | 5.00 | 0.00 | 5.02 | 0.03 |
| 2 | 4.88 | 5.00 | 0.12 | 5.05 | 0.17 |
| 3 | 5.19 | 4.98 | -0.21 | 5.14 | -0.05 |
| 4 | 4.99 | 4.94 | -0.05 | 5.11 | 0.12 |
| 5 | 5.08 | 5.11 | 0.04 | 5.11 | 0.03 |
| 6 | 5.23 | 5.34 | 0.11 | 5.15 | -0.08 |
| 7 | 5.26 | 5.28 | 0.02 | 5.31 | 0.05 |
| 8 | 5.22 | 5.16 | -0.06 | 5.18 | -0.05 |
| 9 | 4.84 | 5.33 | 0.49 | 5.29 | 0.44 |
| 10 | 5.19 | 5.18 | -0.01 | 5.29 | 0.10 |
| 11 | 5.07 | 5.26 | 0.19 | 5.21 | 0.15 |
| 12 | 5.34 | 5.35 | 0.01 | 5.24 | -0.10 |
| 13 | 5.10 | 5.27 | 0.17 | 5.31 | 0.21 |
| 14 | 5.46 | 5.34 | -0.12 | 5.29 | -0.17 |
| 15 | 4.86 | 4.94 | 0.08 | 4.84 | -0.03 |
| 16 | 5.68 | 5.78 | 0.10 | 5.67 | -0.01 |
| 17 | 5.19 | 5.23 | 0.03 | 5.17 | -0.02 |
| 18 | 5.57 | 5.38 | -0.19 | 5.51 | -0.05 |
| 19 | 5.27 | 5.46 | 0.19 | 5.31 | 0.04 |
| 20 | 5.69 | 5.61 | -0.08 | 5.74 | 0.05 |
| 21 | 5.76 | 5.70 | -0.06 | 5.74 | -0.02 |
| 22 | 5.82 | 5.84 | 0.02 | 5.76 | -0.07 |
| 23 | 6.44 | 6.33 | -0.12 | 6.32 | -0.13 |
| 24 | 6.08 | 6.09 | 0.01 | 5.88 | -0.20 |
| 25 | 6.26 | 6.44 | 0.18 | 6.26 | 0.00 |
| 26 | 6.32 | 6.38 | 0.06 | 6.24 | -0.08 |
| 27 | 5.55 | 5.59 | 0.03 | 5.48 | -0.07 |
| 28 | 5.54 | 5.54 | -0.01 | 5.46 | -0.08 |
| 29 | 6.20 | 6.21 | 0.01 | 6.25 | 0.04 |
| 30 | 6.26 | 6.27 | 0.01 | 6.33 | 0.07 |
| 31 | 6.31 | 6.35 | 0.04 | 6.42 | 0.11 |
| 32 | 6.72 | 6.61 | -0.11 | 6.51 | -0.21 |
| 33 | 5.89 | 5.96 | 0.07 | 6.45 | 0.56 |
| 34 | 5.69 | 5.65 | -0.04 | 5.69 | 0.00 |
| 35 | 7.01 | 7.08 | 0.07 | 6.97 | -0.04 |
| 36 | 6.60 | 6.69 | 0.09 | 6.53 | -0.07 |
| 37 | 6.59 | 6.84 | 0.26 | 6.80 | 0.21 |
| 38 | 6.85 | 6.84 | -0.01 | 6.82 | -0.03 |
| 39 | 6.48 | 6.48 | 0.00 | 6.66 | 0.18 |
| 40 | 5.59 | 5.61 | 0.03 | 5.53 | -0.06 |
| 41 | 6.27 | 6.19 | -0.07 | 6.33 | 0.06 |
| 42 | 6.05 | 6.08 | 0.03 | 6.13 | 0.08 |
| 43 | 6.33 | 6.31 | -0.02 | 6.39 | 0.06 |
| 44 | 5.66 | 5.77 | 0.12 | 5.53 | -0.12 |
| 45 | 6.21 | 6.21 | 0.00 | 6.38 | 0.17 |
| 46 | 6.27 | 6.43 | 0.16 | 6.37 | 0.10 |
| 47 | 6.10 | 6.09 | -0.01 | 6.14 | 0.04 |
| 48 | 6.05 | 6.05 | 0.00 | 6.11 | 0.06 |
| 49 | 6.57 | 6.53 | -0.04 | 6.49 | -0.08 |
| 50 | 6.55 | 6.36 | -0.19 | 6.40 | -0.16 |
| 51 | 6.05 | 6.13 | 0.08 | 6.24 | 0.19 |
| 52 | 6.05 | 6.04 | 0.00 | 6.32 | 0.27 |
| 53 | 6.59 | 6.54 | -0.05 | 6.51 | -0.08 |
| 54 | 6.70 | 6.50 | -0.20 | 6.49 | -0.21 |
| 55 | 6.46 | 6.11 | -0.34 | 6.39 | -0.07 |
| 56 | 6.33 | 6.49 | 0.16 | 6.41 | 0.08 |
| 57 | 6.59 | 6.55 | -0.03 | 6.44 | -0.15 |
| 58 | 5.53 | 5.39 | -0.15 | 5.56 | 0.02 |
| 59 | 5.70 | 5.69 | -0.01 | 5.72 | 0.02 |
| 60 | 5.55 | 5.40 | -0.15 | 5.52 | -0.03 |
| 61 | 5.14 | 5.33 | 0.19 | 5.47 | 0.33 |
| 62 | 5.72 | 5.72 | -0.01 | 5.72 | 0.00 |

(Continued in next column)

TABLE 3 | (Continued) Experimental pIC₅₀ (Exp.), predicted pIC₅₀ (Pred.), and corresponding residuals (Res.) of the anthraquinone derivatives.

| Number | pIC ₅₀ | CoMFA | | CoMSIA | |
|--------|-------------------|-------|-------|--------|-------|
| | Exp | Pred | Res | Pred | Res |
| 63 | 5.46 | 5.43 | -0.03 | 5.46 | 0.00 |
| 64 | 5.20 | 5.17 | -0.03 | 5.21 | 0.01 |
| 65 | 5.24 | 5.52 | 0.29 | 5.48 | 0.24 |
| 66 | 5.26 | 5.34 | 0.08 | 5.10 | -0.16 |
| 67 | 5.44 | 5.33 | -0.12 | 5.48 | 0.04 |
| 68 | 5.54 | 5.34 | -0.20 | 5.47 | -0.07 |
| 69 | 5.72 | 5.39 | -0.33 | 5.69 | -0.04 |
| 70 | 5.38 | 5.41 | 0.03 | 5.11 | -0.27 |
| 71 | 5.68 | 5.31 | -0.37 | 5.63 | -0.04 |
| 72 | 5.77 | 5.78 | 0.01 | 5.70 | -0.07 |
| 73 | 5.80 | 5.78 | -0.02 | 5.75 | -0.05 |
| 74 | 5.92 | 5.92 | 0.00 | 5.88 | -0.04 |
| 75 | 5.59 | 5.51 | -0.08 | 5.52 | -0.07 |
| 76 | 6.30 | 6.30 | 0.00 | 6.27 | -0.03 |
| 77 | 5.57 | 5.39 | -0.18 | 5.57 | 0.00 |
| 78 | 6.00 | 5.96 | -0.04 | 6.01 | 0.01 |

to display the QSAR equation. The field type Stdev* Coeff was used to generate 3D contour maps. As shown in **Figures 3, 4**, compound 35 with the best anti-PGAM1 activity was selected as the template compound to dissect the results of CoMFA and CoMSIA models.

The contour map of the steric field of CoMFA is shown in **Figure 3A**, and the effect of the steric field on the activity is shown in green and yellow. The presence of green regions around the molecule indicates that the group with a large connecting space contributes to increasing the activity of the compound, while the presence of yellow regions indicates that the group with a large connecting space may decrease the activity of the compound. As can be seen from **Figure 3A**, there is a green region distributed on the R₁ substituent, so the introduction of a slightly larger volume of groups at the R₁ substituent site is conducive to the improvement of the activity of the compound. For example, compound 22 (pIC₅₀ = 5.82) with a benzene ring was significantly higher than compound 19 (pIC₅₀ = 5.27) in bioactivity. The contour map of the electrostatic field of CoMFA is shown in **Figure 3B**, and the effect of the electrostatic field on the activity is shown in blue and red. The blue regions around the molecule indicate that the connection of the electron-donating group is beneficial to the improvement of the activity of the compound, while the red regions indicate that the connection of the electron-withdrawing group is beneficial to the improvement of the activity of the compound. From **Figure 3B**, we can see that the connection of electron-withdrawing groups near the R₁ substituent is conducive to improving the activity of the compound, so it can explain how the activity of compound 22 (pIC₅₀ = 5.82) is higher than that of compound 19 (pIC₅₀ = 5.27). There is a blue region around the R₂ substituents of anthraquinone, where the introduction of electron groups is beneficial. For example, the bioactivity of compound 72 (pIC₅₀ = 5.77) with a hydroxyl group was significantly higher than that of compound 8 (pIC₅₀ = 5.22).

The contour map of the steric field (**Figure 4A**) and the electrostatic field (**Figure 4B**) of the CoMSIA is very similar

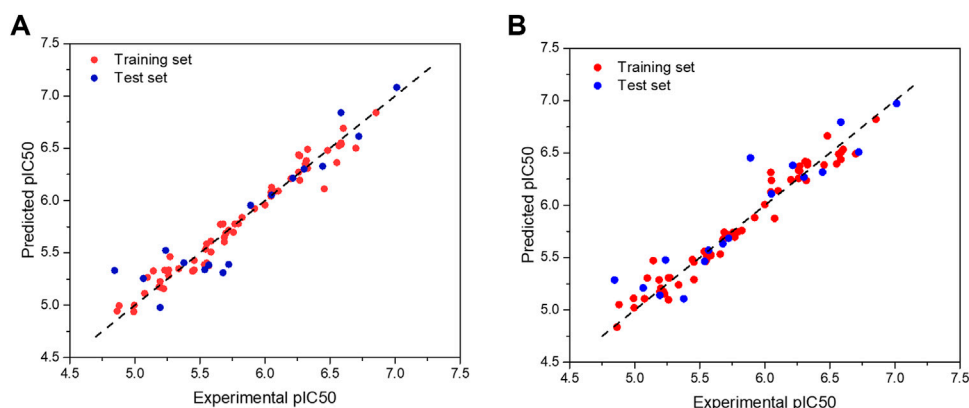


FIGURE 2 | Scatter plot of experimental and predicted bioactivity values (pIC_{50}) of the CoMFA (A) and CoMSIA models (B), respectively.

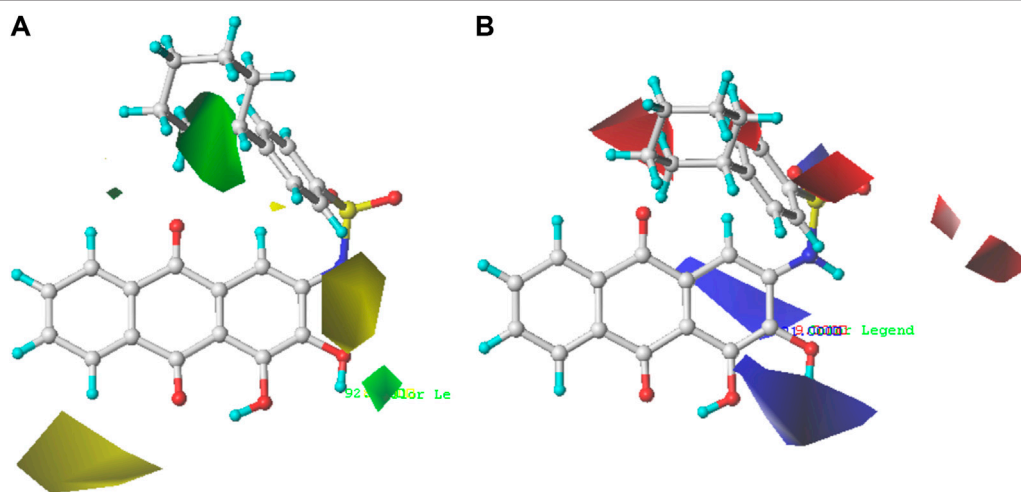


FIGURE 3 | Steric contour map (A) and electrostatic contour map (B) of the CoMFA model based on molecule 35. Green regions represent bulky groups that increase anti-PGAM1 activity, while yellow regions represent sterically disfavored regions. Blue regions show where positive groups are beneficial for increasing anti-PGAM1 activity, and red regions show where negative groups are favored.

to the CoMFA model, so they will not be explained here. The contour map of the hydrophobic field of the CoMSIA model is shown in **Figure 4C**. The cyan regions represent how the introduction of the hydrophobic group is favorable to the activity, while the white regions represent how the introduction of the hydrophilic group is favorable to the activity. There is a cyan region near the R_1 substituent, indicating that the introduction of the hydrophobic group is very helpful to the improvement of the activity. Therefore, the biological activity of compound 22 ($pIC_{50} = 5.82$) is higher than that of compound 19 ($pIC_{50} = 5.27$). The contour map of the hydrogen bond receptor field of CoMSIA is shown in **Figure 4D**. The orange area is where the hydrogen bond acceptor group is conducive to the activity of the compound, and the purple area is where the hydrogen bond donor group is conducive to the activity of the compound. As shown in **Figure 4D**, there are purple

regions with substituents of R_6 and R_2 , where hydrogen bond donors can be imported to improve the anti-PGAM1 activity of the chemical. Moreover, a large purple region is near the nitrogen atom on the amino group, suggesting that the group may be a hydrogen bond donor.

Based on the outcome of CoMFA and CoMSIA analysis, we obtained the structure–activity relationship diagram of anthraquinone compounds (see **Figure 5**). The introduction of hydrogen bond donors in Region A is beneficial to improving the activity of the compounds, such as the carbonyl group. The group with a large space in Region B is conducive to the activity of the compounds, such as biphenyl or p-cyclohexylbenzene (Huang et al., 2019b). The introduction of the hydrophilic group in Region C is conducive to the activity, such as hydroxyl groups (Wang et al., 2018a). The group with a small space in Region D can improve the activity of the compound, such as hydrogen.

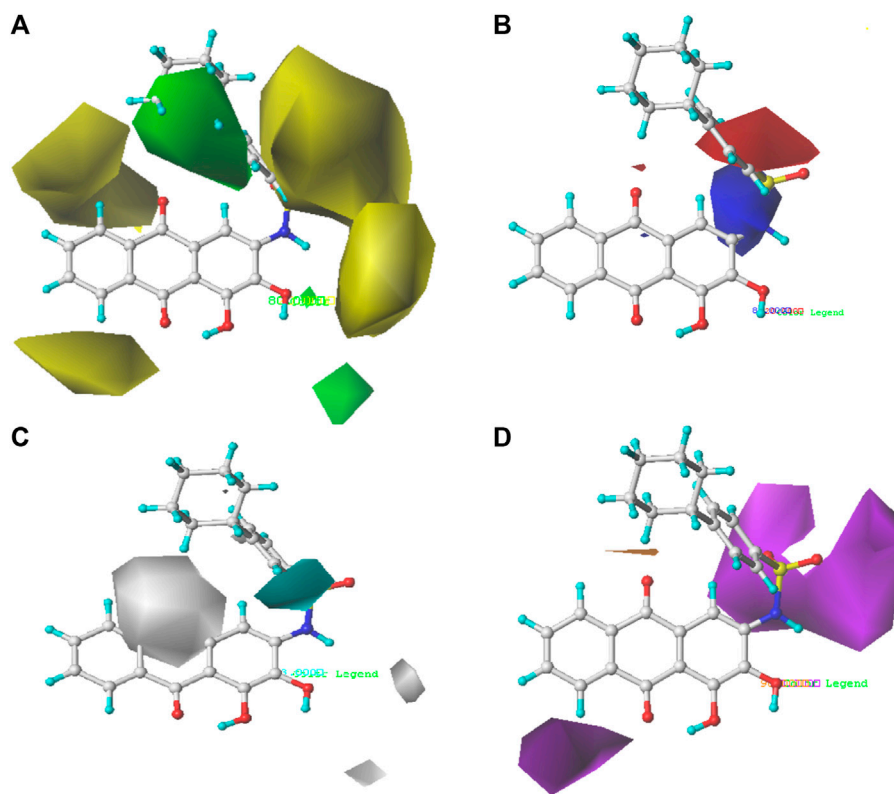


FIGURE 4 | Steric contour map (A), electrostatic contour map (B), hydrophobic contour map (C), and hydrogen bond acceptor contour map (D) of the CoMSIA model based on molecule 35. Green regions are sterically favored regions, while yellow regions are sterically unfavored regions. Blue regions are where electron-donating groups are favored, and red regions are where electron-withdrawing groups are favored. The cyan regions are where the hydrophobic group is favorable to the activity, while the white regions are where the hydrophilic group is favorable to the activity.

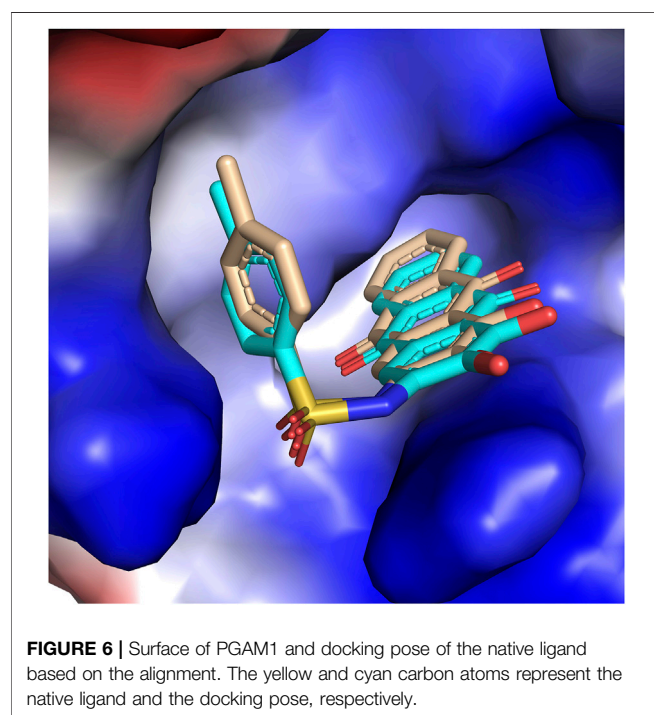
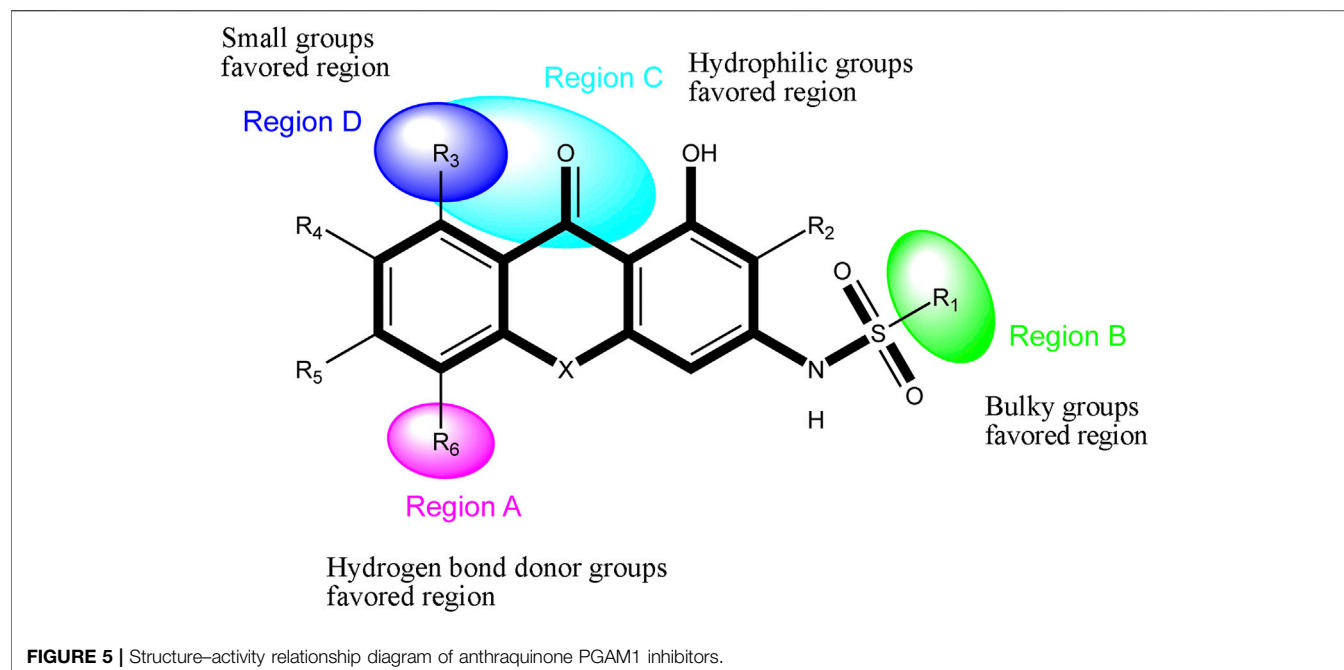
Molecular Docking Analysis

The molecular docking method was employed to interpret the 3D-QSAR result and study the structural basis between PGAM1 and inhibitors. First, the reliability of the glide docking algorithm with the SP mode was evaluated by redocking analysis. It can be seen from **Figure 6** that the redocking conformations of the molecule are well superimposed with the initial structure in PGAM1 protein. The RMSD value between docking conformation and native conformation is 0.005Å. The results suggest that the glide algorithm exhibits a good performance for the PGAM1 protein, which can reproduce the binding pose of the native ligand. Subsequently, all chemicals were docked into the binding site of PGAM1. However, we discover that the docking scores of these compounds are not correlated with the inhibitory activity, and the r^2 of pIC₅₀ vs. the docking score is 0.051, which demonstrates the fact that glide docking is not appropriate for all compounds. We speculate that one of the most important reasons is that 3-PG plays an important role in the process of compounds binding to PGAM1, and the glide scoring function currently used is not suitable for this system. In addition, because PGAM1 catalyzes the conversion of 3-PG to 2-PG in the physiological process, the current docking simulation methods cannot completely simulate this process. Therefore, the docking score and activity do not show a correlation.

Molecular Dynamics Simulations

In order to further analyze the atomic details of the interaction between small molecules and PGAM1, molecular dynamics simulations were employed based on the co-crystal complex of compounds 23 (PDB ID: 5Y35) and 49 (PDB ID: 6ISN) using Amber 16, respectively. 50 ns simulation was performed for each complex. The RMSD plots of Cα, residues within the range of ligand 5Å, ligand, and 3-PG for complexes were shown in **Figure 7**. By monitoring the fluctuation of RMSDs, it can be found that the RMSD fluctuation of each system after 20 ns are all within the range of 2Å. Moreover, the fluctuation of binding free energy over time was also monitored. As shown in **Supplementary Figure S1**, binding free energy of each system fluctuates around 30 kcal/mol after 35 ns. In summary, these results indicate that the two systems finally reached a stable state.

During the process of small molecules binding to PGAM1, the hydrogen bond plays an important role as one of the most important non-bonding interactions. In order to explore the interaction between small molecules and PGAM1, the changes of the hydrogen bond between each residue of PGAM1 and the inhibitor were also monitored. The fraction of the hydrogen bond is greater than 10% as listed in **Table 4**. The results show that two hydrogen bonds formed between compounds 23 and 49 and Arg116, and the total occupancies are 180.12% and 38.48%,



respectively. The results indicate that the hydrogen bonds formed between Arg116 of PGAM1 and inhibitors play a remarkable role in the binding of molecules. Besides, another hydrogen bond is also formed between compound 23 and Arg90 with the occupancy of 12.14%. It is precisely because the small molecules form hydrogen bonds with Arg116 and Arg90 to fix the anthraquinone skeleton of the compounds that compounds 23 and 49 are stably binding with PGAM1.

Binding Free Energy Calculation

The binding free energy is used as a reference standard for evaluating the activity of molecules. It is generally believed that the lower the binding value, the more stable the complex formed by the protein and the small molecule. To evaluate the binding affinity of each complex, the MM/GBSA method was performed to calculate the binding free energy of inhibitors. It can be seen from **Table 5** that the binding free energy of compounds 23 and 49 are -27.40 kcal/mol and -27.85 kcal/mol, respectively, which are consistent with their biological activities. Among them, van der Waals energies (ΔE_{vdw}) are -38.68 kcal/mol and -41.63 kcal/mol, respectively, and their values are much lower than other energy terms, indicating that hydrophobic interaction is the major contributor to the ligand binding process. In addition, electrostatic energy (ΔE_{ele}) also contributes significantly to the binding free energy, which indicates that electrostatic interaction also plays a vital role in ligand binding. It is worth noting that the polar contribution (ΔG_{GB}) is not conducive to ligand binding, which may be attributed to the large size of the binding pocket and the exposure of the hydrophobic ligand to the solvent.

To further confirm the key residues referred to in the ligand binding process, MM/GBSA calculation was performed to decompose the binding free energy into inhibitor-residue pairs. It can be seen from **Figure 8** that the primary residues with binding free energy less than -1 kcal/mol contributing to the ligand binding are F22, K100, V112, W115, and R116. In order to further observe the orientation of compounds and the position of the key residues, we extracted the average structure (see **Figure 9**). It can be seen from **Figure 9** that compounds 23 and 49 adopt a similar binding pose, which is surrounded by those critical residues. Compound 23 forms three hydrogen bonds with R90, W115, and R116. Among the three of them, R90 and

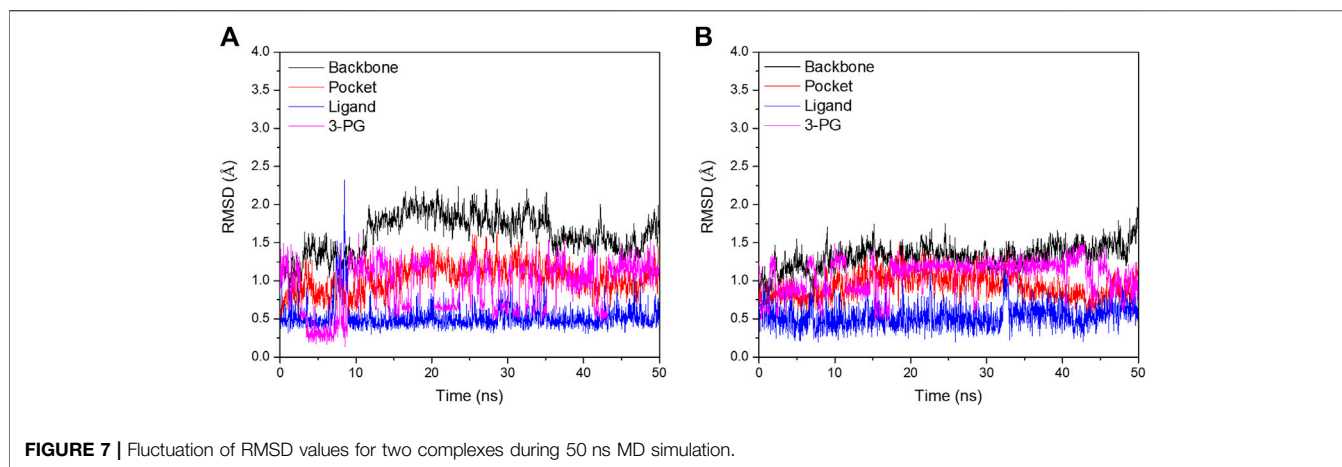


FIGURE 7 | Fluctuation of RMSD values for two complexes during 50 ns MD simulation.

TABLE 4 | Changes of the hydrogen bond over the MD simulations.

| Complex | Donor | Acceptor | Occupancy (%) | Distance (Å) | Angle (°) |
|-------------------|-------------|-----------|---------------|--------------|-----------|
| PGAM1-Compound 23 | Arg116@N-H | Ligand@O5 | 75.08 | 2.93 | 152.74 |
| | Arg116@NE-H | Ligand@O5 | 59.12 | 3.11 | 144.82 |
| | Arg116@NE-H | Ligand@N1 | 45.92 | 3.24 | 152.12 |
| | Arg90@N-H | Ligand@O1 | 12.24 | 3.12 | 130.43 |
| PGAM1-Compound 49 | Arg116@N-H | Ligand@O1 | 20.20 | 2.96 | 148.26 |
| | Arg116@NE-H | Ligand@O1 | 18.28 | 3.05 | 147.35 |

TABLE 5 | Calculated binding energy (kcal/mol) of inhibitor binding to PGAM1.

| Terms | PGAM1-Compound 23 | PGAM1-Compound 49 |
|--------------------|-------------------|-------------------|
| ΔE_{ele} | -26.51 ± 7.59 | -20.76 ± 6.23 |
| ΔE_{vdw} | -38.68 ± 2.92 | -41.63 ± 3.31 |
| ΔG_{gas} | -65.19 ± 8.35 | -62.39 ± 8.01 |
| ΔG_{GB} | 41.62 ± 5.97 | 38.35 ± 5.21 |
| ΔG_{GBSUR} | -3.82 ± 0.16 | -3.81 ± 0.15 |
| ΔG_{sol} | 37.79 ± 5.91 | 34.55 ± 5.14 |
| ΔG_{bind} | -27.40 ± 4.21 | -27.85 ± 3.68 |

$$\Delta G_{gas} = \Delta E_{ele} + \Delta E_{vdw}$$

$$\Delta G_{sol} = \Delta G_{GB} + \Delta G_{GBSUR}$$

$$\Delta G_{bind} = \Delta G_{gas} + \Delta G_{sol}$$

R116 show higher fraction in hydrogen bond analysis, while the bond length of W115 is 3.4 Å due to weak potency. For compound 49, there is no hydrogen bond formed between compound 49 and key residues, which may be due to the low occupancy.

Design New PGAM1 Inhibitors

According to the structure–activity relationships obtained from CoMFA and CoMSIA models, seven molecules with the anthraquinone skeleton were designed as potential PGAM1 inhibitors by introducing new substituents at different positions of compound 35 (see **Table 6**). Compounds 79 and 80 were designed by adding the hydrogen bond donor in the R₆ position to form the key hydrogen bond. Compounds 81, 82, and 83 were designed by introducing the substituent in the R₁ position

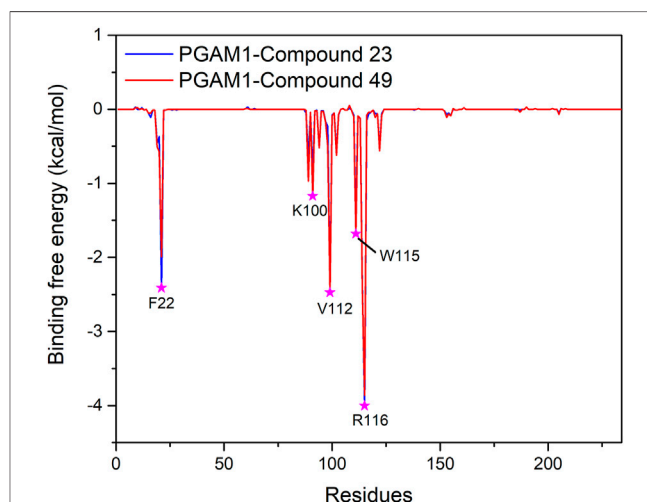


FIGURE 8 | Binding free energy decomposition plots for the two systems.

to increase volume. Based on the contribution of the steric and hydrogen bond donor, compounds 84 and 85 were designed. The pIC₅₀ values of designed compounds were predicted by built CoMFA and CoMSIA models. As shown in **Table 6**, all of the designed compounds exhibit better inhibitory activity targeting PGAM1 than compound 35, and the predictive values are in accordance with the summarized structure–activity relationships.

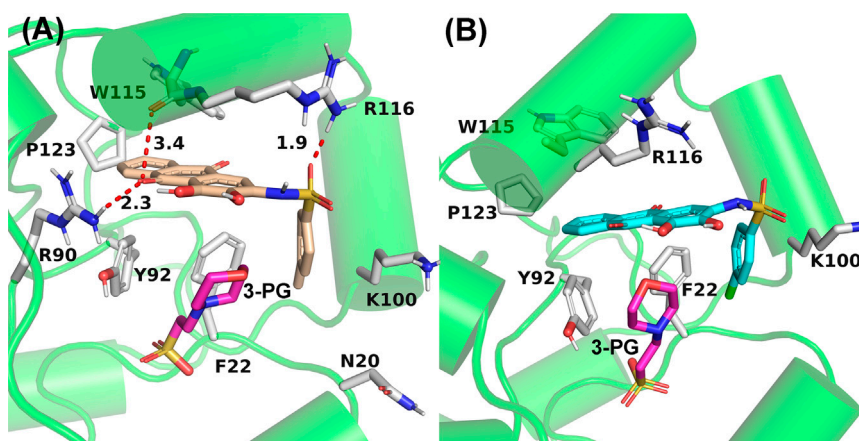


FIGURE 9 | Average structures of PGAM1 with compounds 23 **(A)** and 49 **(B)**. The bonds of residues and ligands are represented in stick, and the carbon atoms of compound 23, compound 49, and residues are represented in yellow, cyan, and white, respectively. The red dotted line represents the hydrogen bond.

TABLE 6 | Newly designed PGAM1 inhibitors and the corresponding predicted activity value.

| Number | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ | X | CoMFA | CoMSIA |
|--------|----------------|----------------|----------------|----------------|----------------|-----------------|------|-------|--------|
| 79 | | OH | H | H | H | OH | -C=O | 7.07 | 7.03 |
| 80 | | OH | H | H | H | NH ₂ | -C=O | 7.05 | 6.99 |
| 81 | | OH | H | H | H | H | -C=O | 7.16 | 6.83 |
| 82 | | OH | H | H | H | H | -C=O | 7.14 | 7.07 |
| 83 | | OH | H | H | H | H | -C=O | 7.10 | 7.03 |
| 84 | | OH | H | H | H | OH | -C=O | 7.14 | 6.85 |
| 85 | | OH | H | H | H | OH | -C=O | 7.13 | 7.00 |

CONCLUSION

In the present study, a combined strategy of 3D-QSAR, molecular docking, and molecular dynamics simulations was applied to explore the structure–activity relationships of anthraquinone analogs. The built CoMFA ($q^2 = 0.81$, $r^2 = 0.97$, $r_m^2 = 0.78$, $r_{pred}^2 = 0.89$) and CoMSIA ($q^2 = 0.82$, $r^2 = 0.96$, $r_m^2 = 0.79$, $r_{pred}^2 = 0.89$)

models have achieved satisfactory results in terms of the statistical results. The results show that the built models have good internal and external predictive power. The acquired contour maps elaborate the structure–activity relationships of anthraquinone derivatives and successfully predict the activity of the test set. According to the results of contour maps, the introduction of hydrogen bond donors in Region A, the group with a large

space in Region B, the hydrophilic group in Region C, and the group with a small space in Region D could improve the activity of the compounds. The calculated results of binding free energy suggest that van der Waals interaction is the major contributor to the ligand binding process. The decomposition binding free energy and hydrogen bond show that small molecules with the anthraquinone core mainly interact with F22, R90, K100, V112, W115, and R116 of PGAM1. Based on these findings, 7 new compounds with the anthraquinone core were designed, and the predicted results show that all of the designed compounds exhibit great inhibitory activity against PGAM1. The constructed 3D-QSAR model will provide theoretical guidance for improving the activity of anthraquinone derivatives and help to develop inhibitors with potent anti-PGAM1 activity.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

YW, HG, and XY designed this study. YG, SQ, and ZL carried out computational modeling and data analysis and wrote the

manuscript. RJ, YT, and EL revised this manuscript. All authors have read and approved the final manuscript.

FUNDING

This study was supported by the Natural Science Foundation of Shaanxi Province (Project No. 2021JQ-734), the National Natural Science Foundation of China (Project No. 82003653), Shaanxi University of Chinese Medicine (Project No. 2020XG01), and the Subject Innovation Team of Shaanxi University of Chinese Medicine (Project No. 2019-PY02).

ACKNOWLEDGMENTS

I wish to thank “Xternal Validation Plus” (<https://dtclab.webs.com/software-tools>) and “application domain” (<https://dtclab.webs.com/software-tools>) tools for calculating the external validation index.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2021.764351/full#supplementary-material>

REFERENCES

- Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: the RESP Model. *J. Phys. Chem.* 97 (40), 10269–10280. doi:10.1021/j100142a004
- Cardoso, G. G., Maria, D. A. T., Assis Leticia, C., Castro, R. T., and Ferreira, D. C. E. F. (2016). Quantitative Structure-Activity Relationship Studies for Potential Rho-Associated Protein Kinase Inhibitors. *J. Chem.* 2016, 9198582. doi:10.1155/2016/9198582
- Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., et al. (2005). The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* 26 (16), 1668–1688. doi:10.1002/jcc.20290
- Cho, S. J., Garsia, M. L., Bier, J., and Tropsha, A. (1996). Structure-Based Alignment and Comparative Molecular Field Analysis of Acetylcholinesterase Inhibitors. *J. Med. Chem.* 39 (26), 5064–5071. doi:10.1021/jm950771r
- Cieplak, P., Cornell, W. D., Bayly, C., and Kollman, P. A. (1995). Application of the Multimolecule and Multiconformational RESP Methodology to Biopolymers: Charge Derivation for DNA, RNA, and Proteins. *J. Comput. Chem.* 16 (11), 1357–1377. doi:10.1002/jcc.540161106
- Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988b). Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* 110 (18), 5959–5967. doi:10.1021/ja00226a005
- Cramer, R. D., III, Bunce, J. D., Patterson, D. E., and Frank, I. E. (1988a). Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* 7 (1), 18–25. doi:10.1002/qsar.19880070105
- de Assis, T. M., Gajo, G. C., de Assis, L. C., Garcia, L. S., Silva, D. R., Ramalho, T. C., et al. (2016). QSAR Models Guided by Molecular Dynamics Applied to Human Glucokinase Activators. *Chem. Biol. Drug Des.* 87 (3), 455–466. doi:10.1111/cbdd.12683
- Evans, M. J., Saghatelian, A., Sorensen, E. J., and Cravatt, B. F. (2005). Target Discovery in Small-Molecule Cell-Based Screens by *In Situ* Proteome Reactivity Profiling. *Nat. Biotechnol.* 23 (10), 1303–1307. doi:10.1038/nbt1149
- Fothergill-Gilmore, L. A., and Watson, H. C. (1989). The Phosphoglycerate Mutases. *Adv. Enzymol. Relat. Areas Mol. Biol.* 62, 227–313. doi:10.1002/9780470123089.ch6
- Fox, T., and Kollman, P. A. (1998). Application of the RESP Methodology in the Parametrization of Organic Solvents. *J. Phys. Chem. B* 102 (41), 8070–8079. doi:10.1021/jp9717655
- Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., and Fox, D. J. (2009). *Gaussian 09*. Wallingford: Gaussian, Inc.
- Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013
- Hitosugi, T., Zhou, L., Elf, S., Fan, J., Kang, H. B., Seo, J. H., et al. (2012). Phosphoglycerate Mutase 1 Coordinates Glycolysis and Biosynthesis to Promote Tumor Growth. *Cancer Cell* 22 (5), 585–600. doi:10.1016/j.ccr.2012.09.020
- Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model.* 51 (1), 69–82. doi:10.1021/ci100275a
- Huang, K., Jiang, L., Li, H., Ye, D., and Zhou, L. (2019a). Development of Anthraquinone Analogues as Phosphoglycerate Mutase 1 Inhibitors. *Molecules* 24 (5), 845. doi:10.3390/molecules24050845
- Huang, K., Jiang, L., Liang, R., Li, H., Ruan, X., Shan, C., et al. (2019b). Synthesis and Biological Evaluation of Anthraquinone Derivatives as Allosteric Phosphoglycerate Mutase 1 Inhibitors for Cancer Treatment. *Eur. J. Med. Chem.* 168, 45–57. doi:10.1016/j.ejmech.2019.01.085
- Huang, K., Liang, Q., Zhou, Y., Jiang, L. L., Gu, W. M., Luo, M. Y., et al. (2019c). A Novel Allosteric Inhibitor of Phosphoglycerate Mutase 1 Suppresses Growth and Metastasis of Non-small-cell Lung Cancer. *Cell Metab* 30 (6), 1107–e1108. doi:10.1016/j.cmet.2019.09.014
- Jorgensen, W. L. (2004). The many Roles of Computation in Drug Discovery. *Science* 303 (5665), 1813–1818. doi:10.1126/science.1096361

- Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* 118 (45), 11225–11236. doi:10.1021/ja9621760
- Lei, Y., Huang, K., Gao, C., Lau, Q. C., Pan, H., Xie, K., et al. (2011). Proteomics Identification of ITGB3 as a Key Regulator in Reactive Oxygen Species-Induced Migration and Invasion of Colorectal Cancer Cells. *Mol. Cell Proteomics* 10 (10), M110–M005397. doi:10.1074/mcp.M110.005397
- Li, F., Yang, H., Kong, T., Chen, S., Li, P., Chen, L., et al. (2020). PGAM1, Regulated by miR-3614-5p, Functions as an Oncogene by Activating Transforming Growth Factor- β (TGF- β) Signaling in the Progression of Non-small Cell Lung Carcinoma. *Cell Death Dis* 11 (8), 710. doi:10.1038/s41419-020-02900-4
- Li, N., and Liu, X. (2020). Phosphoglycerate Mutase 1: Its Glycolytic and Non-glycolytic Roles in Tumor Malignant Behaviors and Potential Therapeutic Significance. *Oncotargets Ther.* 13, 1787–1795. doi:10.2147/OTT.S238920
- Li, X., Tang, S., Wang, Q. Q., Leung, E. L., Jin, H., Huang, Y., et al. (2017). Identification of Epigallocatechin-3- Gallate as an Inhibitor of Phosphoglycerate Mutase 1. *Front. Pharmacol.* 8, 325. doi:10.1073/pnas.191455711610.3389/fphar.2017.00325
- Liang, Q., Gu, W. M., Huang, K., Luo, M. Y., Zou, J. H., Zhuang, G. L., et al. (2021). HKB99, an Allosteric Inhibitor of Phosphoglycerate Mutase 1, Suppresses Invasive Pseudopodia Formation and Upregulates Plasminogen Activator Inhibitor-2 in Erlotinib-Resistant Non-small Cell Lung Cancer Cells. *Acta Pharmacol. Sin* 42 (1), 115–119. doi:10.1038/s41401-020-0399-1
- Linse, B., and Linse, P. (2014). Tuning the Smooth Particle Mesh Ewald Sum: Application on Ionic Solutions and Dipolar Fluids. *J. Chem. Phys.* 141 (18), 184114. doi:10.1063/1.4901119
- Liu, L., Wang, S., Zhang, Q., and Ding, Y. (2008). Identification of Potential Genes/proteins Regulated by Tiam1 in Colorectal Cancer by Microarray Analysis and Proteome Analysis. *Cell Biol Int* 32 (10), 1215–1222. doi:10.1016/j.cellbi.2008.07.004
- Liu, X., Weng, Y., Liu, P., Sui, Z., Zhou, L., Huang, Y., et al. (2018). Identification of PGAM1 as a Putative Therapeutic Target for Pancreatic Ductal Adenocarcinoma Metastasis Using Quantitative Proteomics. *Oncotargets Ther.* 11, 3345–3357. doi:10.2147/OTT.S162470
- Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theor. Comput* 11 (8), 3696–3713. doi:10.1021/acs.jctc.5b00255
- Massova, I., and Kollman, P. A. (2000). Combined Molecular Mechanical and Continuum Solvent Approach (MM-PBSA/GBSA) to Predict Ligand Binding. *Perspect. Drug Discov. Des.* 18 (1), 113–135. doi:10.1023/A:1008763014207
- Peng, X. C., Gong, F. M., Chen, Y., Qiu, M., Cheng, K., Tang, J., et al. (2016). Proteomics Identification of PGAM1 as a Potential Therapeutic Target for Urothelial Bladder Cancer. *J. Proteomics* 132, 85–92. doi:10.1016/j.jprot.2015.11.027
- Pratim Roy, P., Paul, S., Mitra, I., and Roy, K. (2009). On Two Novel Parameters for Validation of Predictive QSAR Models. *Molecules* 14 (5), 1660–1701. doi:10.3390/molecules14051660
- Roe, D. R., and Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theor. Comput* 9 (7), 3084–3095. doi:10.1021/ct400341p
- Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., and Das, R. N. (2013). Some Case Studies on Application of "r(m)2" Metrics for Judging Quality of Quantitative Structure-Activity Relationship Predictions: Emphasis on Scaling of Response Data. *J. Comput. Chem.* 34 (12), 1071–1082. doi:10.1002/jcc.23231
- Roy, K., Kar, S., and Ambure, P. (2015). On a Simple Approach for Determining Applicability Domain of QSAR Models. *Chemometrics Intell. Lab. Syst.* 145, 22–29. doi:10.1016/j.chemolab.2015.04.013
- Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 23 (3), 327–341. doi:10.1016/0021-9991(77)90098-5
- Schrödinger, L. (2015). New York, NY: Schrödinger, LLC.
- Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007). Epik: a Software Program for pK(a) Prediction and Protonation State Generation for Drug-like Molecules. *J. Comput. Aided Mol. Des.* 21 (12), 681–691. doi:10.1007/s10822-007-9133-z
- Sun, H., Li, Y., Tian, S., Xu, L., and Hou, T. (2014). Assessing the Performance of MM/PBSA and MM/GBSA Methods. 4. Accuracies of MM/PBSA and MM/GBSA Methodologies Evaluated by Various Simulation Protocols Using PDBbind Data Set. *Phys. Chem. Chem. Phys.* 16 (31), 16719–16729. doi:10.1039/C4CP01388C
- Sun, Q., Li, S., Wang, Y., Peng, H., Zhang, X., Zheng, Y., et al. (2018). Phosphoglyceric Acid Mutase-1 Contributes to Oncogenic mTOR-Mediated Tumor Growth and Confers Non-small Cell Lung Cancer Patients with Poor Prognosis. *Cell Death Differ* 25 (6), 1160–1173. doi:10.1038/s41418-017-0034-y
- SYBYL, V. (XX). SYBYL. St. Louis, MO: Tripos Inc..
- Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and Testing of a General Amber Force Field. *J. Comput. Chem.* 25 (9), 1157–1174. doi:10.1002/jcc.20035
- Wang, P., Jiang, L., Cao, Y., Ye, D., and Zhou, L. (2018a). The Design and Synthesis of N-Xanthone Benzenesulfonamides as Novel Phosphoglycerate Mutase 1 (PGAM1) Inhibitors. *Molecules* 23 (6), 1396. doi:10.3390/molecules23061396
- Wang, P., Jiang, L., Cao, Y., Zhang, X., Chen, B., Zhang, S., et al. (2018b). Xanthone Derivatives as Phosphoglycerate Mutase 1 Inhibitors: Design, Synthesis, and Biological Evaluation. *Bioorg. Med. Chem.* 26 (8), 1961–1970. doi:10.1016/j.bmc.2018.02.044
- Wen, Y. A., Zhou, B. W., Lv, D. J., Shu, F. P., Song, X. L., Huang, B., et al. (2018). Phosphoglycerate Mutase 1 Knockdown Inhibits Prostate Cancer Cell Growth, Migration, and Invasion. *Asian J. Androl.* 20 (2), 178–183. doi:10.4103/aja.aja_57_17
- Xu, Z., Gong, J., Wang, C., Wang, Y., Song, Y., Xu, W., et al. (2016). The Diagnostic Value and Functional Roles of Phosphoglycerate Mutase 1 in Glioma. *Oncol. Rep.* 36 (4), 2236–2244. doi:10.3892/or.2016.5046
- Zhang, C., Li, Y., Zhao, W., Liu, G., and Yang, Q. (2020). Circ-PGAM1 Promotes Malignant Progression of Epithelial Ovarian Cancer through Regulation of the miR-542-3p/CDC5L/PEAK1 Pathway. *Cancer Med.* 9 (10), 3500–3521. doi:10.1002/cam4.2929

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Wang, Guo, Qiang, Jin, Li, Tang, Leung, Guo and Yao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Machine Learning Enables Accurate and Rapid Prediction of Active Molecules Against Breast Cancer Cells

Shuyun He^{1,2†}, Duancheng Zhao^{1,2†}, Yanle Ling^{1,2}, Hanxuan Cai^{1,2}, Yike Cai³, Jiquan Zhang^{4*} and Ling Wang^{1,2*}

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Cristian Axenie,
Technische Hochschule Ingolstadt,
Germany
Ran Su,
Tianjin University, China

*Correspondence:

Jiquan Zhang
zjqgmc@163.com
Ling Wang
lingwang@scut.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 17 October 2021

Accepted: 02 December 2021

Published: 17 December 2021

Citation:

He S, Zhao D, Ling Y, Cai H, Cai Y,
Zhang J and Wang L (2021) Machine
Learning Enables Accurate and Rapid
Prediction of Active Molecules Against
Breast Cancer Cells.
Front. Pharmacol. 12:796534.
doi: 10.3389/fphar.2021.796534

¹Guangdong Provincial Key Laboratory of Fermentation and Enzyme Engineering, Guangdong Provincial Engineering and Technology Research Center of Biopharmaceuticals, School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China, ²Joint International Research Laboratory of Synthetic Biology and Medicine, Guangdong Provincial Engineering and Technology Research Center of Biopharmaceuticals, School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China, ³Center for Certification and Evaluation, Guangdong Drug Administration, Guangzhou, China, ⁴State Key Laboratory of Functions and Applications of Medicinal Plants, College of Pharmacy, Guizhou Provincial Engineering Technology Research Center for Chemical Drug R&D, Guizhou Medical University, Guiyang, China

Breast cancer (BC) has surpassed lung cancer as the most frequently occurring cancer, and it is the leading cause of cancer-related death in women. Therefore, there is an urgent need to discover or design new drug candidates for BC treatment. In this study, we first collected a series of structurally diverse datasets consisting of 33,757 active and 21,152 inactive compounds for 13 breast cancer cell lines and one normal breast cell line commonly used in in vitro antiproliferative assays. Predictive models were then developed using five conventional machine learning algorithms, including naïve Bayesian, support vector machine, k-Nearest Neighbors, random forest, and extreme gradient boosting, as well as five deep learning algorithms, including deep neural networks, graph convolutional networks, graph attention network, message passing neural networks, and Attentive FP. A total of 476 single models and 112 fusion models were constructed based on three types of molecular representations including molecular descriptors, fingerprints, and graphs. The evaluation results demonstrate that the best model for each BC cell subtype can achieve high predictive accuracy for the test sets with AUC values of 0.689–0.993. Moreover, important structural fragments related to BC cell inhibition were identified and interpreted. To facilitate the use of the model, an online webserver called ChemBC (<http://chembc.idruglab.cn/>) and its local version software (<https://github.com/idruglab/ChemBC>) were developed to predict whether compounds have potential inhibitory activity against BC cells.

Keywords: breast cancer, machine learning, graph neural networks, molecular fingerprints, structural fragments

1 INTRODUCTION

According to the latest data on the global cancer burden for 2020 released by the International Agency for Research on Cancer of the World Health Organization, breast cancer (BC) surpassed lung cancer in 2020 to become the most common cancer worldwide. BC is the leading cause of cancer-related death among women worldwide (Sung et al., 2021). BC consists of the uncontrolled proliferation of mammary epithelial cells under the action of many carcinogenic factors (Escala-Garcia et al., 2020), including alcohol consumption, smoking, overweight, and mammographic density. BC is classified according to the expression of the estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki-67 into five subtypes: Luminal A, Luminal B (HER2-positive or HER2-negative), HER2-positive, and triple-negative breast cancer (TNBC) (Harbeck et al., 2013). Among these BC subtypes, TNBC is associated with poor survival mediated by treatment resistance, and it is the most difficult to treat with curative intent (Liao et al., 2021). Several drugs (e.g., anthracyclines and trastuzumab) have been approved by the U.S. Food and Drug Administration (FDA) for the treatment of BC; however, issues such as poor efficacy, toxicity, adverse drug reactions, and the emergence of drug resistance have limited their clinical use (Brower, 2013; Cameron et al., 2017; Shah and Gradishar, 2018; Daniyal et al., 2021; Li and Li, 2021). Therefore, there is an urgent need to discover and develop new drugs for the treatment of BC, particularly for TNBC.

Innovative drugs (or active molecules) can be identified through two mainstream screening methods: phenotypic-based screening and target-based screening. Target-based screening has been widely used to discover new drugs for the treatment of human diseases in both the pharmaceutical industry and academia for more than 30 years (Chen et al., 2014; Zhang et al., 2014; Wang et al., 2017a; Luo and Wang, 2017; Moffat et al., 2017; Shang et al., 2017). Target-based screening has several advantages, including simplicity, lower cost, and easy to achieve efficient structure-activity relationship (SAR) for lead optimization (Croston, 2017). However, there are two major concerns associated with target-based approaches: 1) the identification and validation of druggable targets is difficult, and if a selected target is undruggable, it may lead practitioners to pursue projects and compounds that fail to translate into clinical results (Croston, 2017) and 2) the conventional “one drug, one target” paradigm has shown unsatisfactory clinical results in human complex diseases (e.g., cancer (Wermuth, 2004), Alzheimer’s disease (Wang et al., 2017b; Albertini et al., 2021), and infectious diseases (Morphy et al., 2004; Li et al., 2019). Phenotypic-based screening (e.g., whole-cell activity), an original but indispensable drug screening method, has gained attention in recent years because of the number of discovered and approved drugs (Liu et al., 2019; Childers et al., 2020; Berg, 2021; Quancard et al., 2021). Two influential analyses by Swinney and Anthony in 2011 and Swinney in 2013 highlighted that the majority of first-in-class drugs (new chemical entities, NME) approved between 1999 and 2008 were identified through phenotypic screening approaches

compared with target-based screening methods. In reality, most FDA approvals of first-in-class drugs originated from phenotypic screening before their precise mechanisms of action or molecular targets were elucidated.

Although phenotype-based screening has advantages over target-based screening for drug discovery, it is unscalable, costly, and does not contribute to the understanding of the mechanism of action of drugs. Several important technologies including affinity-based approaches, functional genetic approaches, cellular profiling approaches, and knowledge-based (computational) approaches are currently available and can be used to characterize the direct and indirect target space of bioactive compounds from phenotypic screening (Schirle and Jenkins, 2016; Sydow et al., 2019; Hughes et al., 2021).

Increased amounts of phenotypic pharmacological data on cancer, Alzheimer’s disease, and infectious diseases have been accumulated in the past 3 decades. Inspired by the available phenotypic screening data, several efficient and cost-saving computational models have been developed to accelerate the drug design and discovery process (Zoffmann et al., 2019; Buckner et al., 2020; Chandrasekaran et al., 2021; Malandraki-Miller and Riley, 2021). For example, in 2020, Stokes et al. first reported directed message passing neural network models using a collection of 2,335 compounds for those that inhibited the growth of *Escherichia coli* (phenotype screening data) and then identified the lead compound halicin with broad-spectrum antibacterial activity (Stokes et al., 2020). Other machine learning-based models have been established to identify new agents against Methicillin-Resistant *Staphylococcus aureus* (Wang et al., 2016b), *Mycobacterium tuberculosis* (Ye et al., 2021), *Pseudomonas aeruginosa* (Fields et al., 2020), *Plasmodium falciparum* (Ashdown et al., 2020), and *Schistosoma* (Zheng et al., 2021). In the field of anticancer drug design and discovery, phenotypic whole cell-based screening methods have substantially advanced our ability to identify new anticancer drugs. In previous studies, we reported the development of computational models using integrated NCI-60 cell-based phenotype screening data to identify new anticancer agents (e.g., **G03** and **I2**) with significant inhibitory activity against various cancer cell lines (Guo et al., 2019; Luo et al., 2019). Although the reported integrated computational anticancer models provided valuable data for discovering anticancer agents, these models cannot distinguish or selectively predict specific cancer cell subtypes (such as BC and its subtypes). In addition, these prediction models have not been developed into easy-to-use tools (e.g., local software packages or online prediction platforms), which limits the use of these models by practitioners in the field.

In the present study, we expanded our earlier efforts aimed at developing reliable computational cell-based models to predict cell inhibitory activity in BC and subtypes and provided a free platform to share our models. A total of 588 cell-based models for BC and subtypes were developed using five conventional machine learning (ML) and five deep learning (DL) algorithms based on three major types of molecular descriptors, fingerprints, and graphs. We used the local outlier factor (LOF) (Breunig et al., 2000) algorithm to evaluate the applicability domain of the best

model for each BC cell line and applied the SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017; Lundberg et al., 2020) algorithm to highlight significant structural fragments. Finally, an online platform (<http://chembc.idruglab.cn/>) and local software (<https://github.com/idruglab/ChemBC>) were constructed based on reliable models to contribute to future research.

2 METHODS

2.1 Dataset Collection and Preparation

All quantitative compound-cell associations (cell-based assays, assay type: F) for available BC cell lines and normal BC cell lines were collected from ChEMBL (Mendez et al., 2019) (downloaded in March 2021) after the exclusion of metastatic cell lines. Each BC cell dataset was then processed using the following steps: 1) compounds with biological activity reported as IC_{50} , EC_{50} , or GI_{50} were kept, whereas molecules that had no bioactivity record were removed; 2) the units of bioactivity (i.e., g/mL, M, nM) were converted into the standard unit in μM ; 3) for a molecule with multiple bioactivity values, the final bioactivity value was obtained by averaging the available bioactivity records; 4) according to previous studies (Fields et al., 2020; Ye et al., 2021), compounds with bioactivity values (e.g., IC_{50} , EC_{50} , GI_{50}) $\leq 10 \mu M$ were considered as active and vice versa; molecules whose labels could not be unequivocally assigned (e.g., activity $< 100 \mu M$ or activity $> 1 \mu M$) were excluded from the dataset; 5) all molecules were processed by removing salt and optimized based on the MMFF94X force field using MOE software (version 2018) with the default parameters. Finally, 14 cell lines with the number of active molecules (actives) and inactive molecules (inactives) > 50 were retained. Each cell-compound dataset was randomly split into three sub-datasets: training (80%), validation (10%), and test (10%). All datasets used for the models described in the present study are freely available at <https://github.com/idruglab/ChemBC>.

2.2 Molecular Representations Calculation

Choosing suitable molecular representations is essential for developing acceptable and robust QSAR models. To a certain extent, the molecular representation determines the upper limit of the accuracy of the model. To fully characterize the chemical information of these molecules, three distinct types of features were calculated and used, including molecular descriptors-, fingerprints-, and graph-based representations. RDKit descriptors (RDKitDes), a set of 208 descriptors, were used. Four fingerprint-based features including Morgan fingerprints (ECFP-like, 1024-bits) (Rogers and Hahn, 2010), MACCS keys (166-bits) (Durant et al., 2002), AtomParis fingerprints (2048-bits) (Carhart et al., 1985), and 2D Pharmacophore Fingerprints (PharmacPFP, 38-bits) (Gobbi and Poppinger, 1998) were implemented. The molecular descriptor- and fingerprint-based representations were calculated using RDKit (Landrum, 2016) (version: 2020.03.1).

The molecular graph (G) representative consisted of two matrices for a given molecule: the $N \times N$ adjacency matrix A ,

representing a graph structure; and the $N \times F$ node-feature matrix X , where N is the number of nodes and F is the number of node features. The node-feature matrix contained the following atom features: atom type, formal charge, hybridization, number of bound hydrogens, aromaticity, number of degrees, number of hydrogens, chirality, and partial charge. The edge representation contained bond type, whether the atoms in the pair are in the same ring, whether the bond is conjugated or not, and stereo configuration of a bond (Kearnes et al., 2016). Most of them were encoded in a one-hot manner into a molecular graph. In this study, molecular graph-based representations were generated using Deepchem (version: 2.5.0). For example, the MolGraphConvFeaturizer module was used to calculate the molecular graphs of Attentive FP, GAT, and MPNN models, and the ConvMolFeaturizer (Duvenaud et al., 2015) module was used to calculate the molecular graph of the GCN model.

2.3 Machine Learning Algorithms and Model Construction

Five conventional ML algorithms (i.e., RF, SVM, XGBoost, KNN, and NB) and five DL algorithms (i.e., DNN, GCN, GAT, MPNN, and Attentive FP) were used to develop classification models for discriminating actives from inactives against breast cell lines. The RF, SVM, KNN, and NB models were constructed using the Scikit-learn (Pedregosa et al., 2011) python package (<https://github.com/scikit-learn/scikit-learn>, version: 0.24.1); the XGBoost (Chen and Guestrin, 2016) models were developed using the XGBoost python package (<https://github.com/dmlc/xgboost>, version: 1.3.3); and other graph-based models were established using the DeepChem python package (<https://deepchem.io/>). All descriptor- and fingerprint-based models and graph-based DL models were trained on CPU [Intel(R) Xeon(R) Silver 4216 CPU @ 2.10 GHz] and GPU [NVIDIA Corporation GV100GL (Tesla V100 PCIe 32 GB)], respectively. In addition, we used grid search to optimize hyperparameters for each model. Detailed these modeling methods and their hyperparameters are briefly described as follows.

2.3.1 Random Forest

RF is a representative ensemble learning approach. It establishes a classifier or regressor by an ensemble of individual decision trees and makes predictions as final output by vote or by averaging multiple decision trees (Svetnik et al., 2003). Compared with a decision tree, RF has high prediction accuracy, good tolerance to outliers and noise, and is not easy to overfit. To obtain the best RF model, the following five hyperparameters were optimized: `n_estimators` (10–500), `criterion` (“gini” and “entropy”), `max_depth` (0–15), `min_samples_leaf` (1–10), and `max_features` (“log2”, “auto” and “sqrt”).

2.3.2 Support Vector Machine

SVM is a supervised ML algorithm that can be used for both classification and regression tasks (Zernov et al., 2003). The basic idea underlying SVM is to find the optimal hyperplane in the feature space that can be obtained by maximizing the boundary between classes in N -dimensional space, which distinguishes objects with different class labels. SVM has been widely used

in drug discovery-relevant applications such as compound activity and property prediction (Heikamp and Bajorath, 2014). In the training of SVM models, two hyperparameters, Kernel coefficient (gamma, “auto”, 0.1–0.2) and penalty parameter C of the error term (C, from 1 to 100), were optimized.

2.3.3 Extreme Gradient Boosting

XGBoost is one of the so-called ensemble learning algorithms under the Gradient Boosting framework and has achieved state-of-the-art ranking results in many ML competitions. It has been widely used in molecular property/activity prediction tasks (Jiang Z. et al., 2021; Li et al., 2021; Ye et al., 2021). Seven hyperparameters were optimized in the training of XGBoost models: learning_rate (0.01–0.1), gamma (0–0.1), min_child_weight (1–3), max_depth (3–5), n_estimators (50–100), subsample (0.8–1.0), and colsample_bytree (0.8–1.0).

2.3.4 K-Nearest Neighbor

The basic idea of the KNN ML algorithm (Cover and Hart, 1967) is to identify the k training samples closest to the test samples in the training set based on distance measures (e.g., Euclidean, Manhattan, and Jaccard distance), and to make a prediction based on the information of the k samples. The default distance measure Euclidean was used in this study. The following three hyperparameters were optimized: n_neighbors (1–5), p (1–2), and weight function (“uniform”, “distance”).

2.3.5 Naïve Bayes

NB is a classic classification ML method based on Bayes’ theorem (Duda and Hart, 1973) and independent assumption of characteristic conditions. For a given dataset, the joint probability distribution of input and output is first learned based on the independent hypothesis of characteristic conditions. NB is also widely used in drug discovery practices (Wang et al., 2016b; Wang et al., 2016a; Wang et al., 2016b; Guo et al., 2020). Two hyperparameters were optimized: alpha (0.01–1) and binarize (0, 0.5, 0.8).

2.3.6 Deep Neural Networks

DNN is a typical DL algorithm and is essentially an artificial neural network (McCulloch and Pitts, 1943) with multiple hidden layers. It consists of many independent neurons, each of which collects information from its connected neurons, and the aggregated information is then activated through a nonlinear activation function. The following key hyperparameters were optimized: dropouts (0.1, 0.2, 0.5), layer_sizes (64, 128, 256, 512) and weight_decay_penalty (0.01, 0.001, 0.0001).

2.3.7 Graph Convolutional Network

GCN is a classic neural network that can use graph-structured data as input (Kipf and Welling, 2016). It is composed of graph convolution layers, a readout layer, fully connected layers, and an output layer. The core idea of graph convolution is to use edge information for aggregating node information, thereby generating a new node representation. Various GCN frameworks have been proposed. Duvenaud et al. (2015) introduced a convolutional neural network that allows end-to-end learning of prediction pipelines. In this study, we used

Duvenaud’s GCN method, and the following hyperparameters were optimized: weight_decay (0, 10e-8, 10e-6, 10e-4), graph_conv_layers [(64, 64), (128, 128), (256, 256)], learning rate (0.01, 0.001, 0.0001) and dense_layer_size (64, 128, 256).

2.3.8 Graph Attention Network

Attention mechanism (AM) is one component of a neural network architecture, which can be embedded in the DL models to automatically learn and calculate the contribution of input data to output data. GCN cannot complete the inductive task, namely, dynamic graph problems, and it is not easy for GCN to assign different learning weights to different neighbors. GAT (Veličković et al., 2017) introduces an AM to address the disadvantages of previous approaches based on GCN or its approximation. The weight of the features of adjacent nodes depends entirely on the features of the nodes and is independent of the graph structure. In the training of the GAT model, the following hyperparameters were optimized: weight_decay (0, 10e-8, 10e-6, 10e-4), learning rate (0.01, 0.001, 0.0001), n_attention_heads (8, 16, 32), and dropouts (0, 0.1, 0.3, 0.5).

2.3.9 Message Passing Neural Network

MPNN, proposed by Gilmer et al. (2017), is a common graph neural network (GNN) framework for chemical prediction tasks. It can directly learn the molecular characteristics from the molecular diagram and is not affected by the graph isomorphism. In the training of the MPNN model, six hyperparameters were optimized: weight_decay (10e-8, 10e-6, 10e-4), learning rate (0.01, 0.001, 0.0001), graph_conv_layers [(64, 64), (128, 128), (256, 256)], num_layer_set2set (2, 3, 4), node_out_feats (16, 32, 64), and edge_hidden_feats (16, 32, 64).

2.3.10 Attentive FP

Attentive FP, which was proposed by Xiong et al. (Xiong et al., 2020), is currently a state-of-the-art GNN model for molecular property prediction, and what is learned from the established model is interpretable. It allows the model to focus on the most relevant parts of the input by applying a graph AM. Herein, the main hyperparameters were optimized as follows: dropout (0, 0.1, 0.5), graph_feat_size (50, 100, 200), num_timesteps (1, 2, 3), num_layers (2, 3, 4), learning rate (0.0001, 0.001, 0.01), and weight_decay (0, 0.01, 0.0001).

2.4 Performance Evaluation of Models

The following classification evaluation metrics were used to evaluate the performance of the classification models: specificity (SP/TNR), sensitivity (SE/TPR/Recall), accuracy (ACC), F1-measure (F1 score), Matthews correlation coefficient (MCC), the area under the receiver operating characteristic (AUC), and Balanced accuracy (BA). These evaluation metrics are defined as follows:

$$SP = \frac{TN}{TN + FP} \quad (1)$$

$$SE = \frac{TP}{TP + FN} \quad (2)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (4)$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \quad (5)$$

$$BA = \frac{TPR + TNR}{2} = \frac{SE + SP}{2} \quad (6)$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

2.5 Model Interpretation

The interpretation of complex ML models remains a challenge because ML algorithms are often a “black box”. Accordingly, we used a recently-developed model-agnostic interpretation framework termed SHapley Additive exPlanation (SHAP) to interpret the established ML models presented in this study. Inspired by the idea of cooperative game theory, the SHAP method constructs an additive explanatory model. In this model, all features are considered contributors. For each prediction sample, the model generates a predicted value, and the SHAP value is the value assigned to each feature in the sample. The greater the SHAP value, the greater the contribution of the corresponding feature to the ML model. The SHAP value is calculated as follows:

$$y_i = y_{base} + f(X_{i1}) + f(X_{i2}) + \cdots + f(X_{ik}) \quad (7)$$

where X_i represents the sample, X_{ij} represents the j feature of this sample, y_i represents the predicted value of the model for this sample, y_{base} represents the baseline of the entire model (usually the mean of the target variable for all samples), $f(X_{ij})$ is the SHAP value of X_{ij} . Intuitively, $f(X_{i1})$ is the contribution value of the first feature in sample i to the final predicted value y_i . When $f(X_{i1}) > 0$, it indicates that this feature improves the predicted value and has a positive effect. On the contrary, it shows that this feature reduces the predicted value and has a reverse effect. Collectively, SHAP value can reflect the influence of the feature in each sample and show the positive and negative influence of the feature.

2.6 Model Applicability Domain

According to the principles of the Organization for Economic Co-operation and Development (OECD), it is necessary to determine the applicability domain (AD) of the QSAR model because of the limited structural diversity of the molecules used in the training dataset. From the perspective of ML, a suitable AD can prevent the prediction deviation from being too large because the feature range of the samples to be tested is too different from the training dataset samples. Therefore, effective identification of Out-of-Domain compounds is the basis for ensuring the reliability of the established model. We used the LOF algorithm (Breunig et al., 2000) to detect super-applicability domain compounds for the best model for each BC or normal breast cell line. LOF is based on the concept of local density, where the local area is given by k -nearest neighbors, whose distance is used to estimate the density. Regions of similar density can be identified by

comparing the local density of an object with that of its neighbors, and points that are much lower in density than their neighbors are considered outliers.

3 RESULTS

3.1 Dataset Analysis and Model Construction

According to the above-predefined criteria, 14 breast-associated cell lines were obtained and distributed as follows: 1) two Luminal A subtypes including MCF-7 and T-47D; 2) two Luminal B subtypes including BT-474 and MDA-MB-361; 3) three HER-2+ subtypes including MDA-MB-435, MDA-MB-453, and SK-BR-3; 4) six TNBC subtypes including Bcap37, BT-20, BT-549, HS-578T, MDA-MB-231, and MDA-MB-468; and 5) one normal breast cell line, HBL-100. Accordingly, we selected these cell-based phenotypical datasets for subsequent modeling. The model construction pipeline is shown in **Figure 1**. Details on the 14 cell lines and their corresponding cell-associated compound datasets are summarized in **Table 1**. The compiled cell-based phenotype datasets included 34,801 unique compounds and 54,909 cell-compound associations. Among them, in 14 cell line datasets, 33,757 compounds were labeled as actives and 21,152 compounds were labeled as inactive (Supplementary Figure S1A). Supplementary Figure S1B shows the proportions of actives and inactive in the 14 cell datasets (due to the natural, although it may not be the best, we did not add theoretical decoys to deliberately balance the data), with active compounds accounting for approximately 40–78%.

The structural diversity and chemical space of compounds in datasets play a key role in the predictive ability of the ML models. Bemis–Murcko scaffold analysis (Bemis and Murcko, 1996) showed that the proportion of the scaffolds for each BC cell line dataset was between 19.70 and 53.41% (**Table 1**), suggesting that the anti-BC compounds of each cell line were structurally more diverse. In addition, the chemical space of the compounds in each dataset can be depicted in a two-dimensional space using molecular weight (MW) and AlogP. As shown in **Supplementary Figure S2**, the training, validation, and test set compounds were distributed over a wide range of MW (108.10–5,714.45) and AlogP (−55.54–42.62), demonstrating that the compounds in the modeling datasets have a broad chemical space. Based on the three different types of molecular features (i.e., molecular descriptors-, fingerprints-, and graph-based features) and the selected ten types of ML algorithms, 476 single models and 112 fusion models were developed. All models were optimized based on the validation sets and selected based on the F1 score (Kc et al., 2021). The best models were selected for the evaluation of external test datasets. The performance of the established models is discussed in the following sections.

3.2 Performance of Descriptor-Based Prediction Models for Breast-Associated Cells

Firstly, 84 predictive models were constructed based on the RDKit-descriptors using five traditional types of ML

TABLE 1 | Breast cell line datasets used in this study.

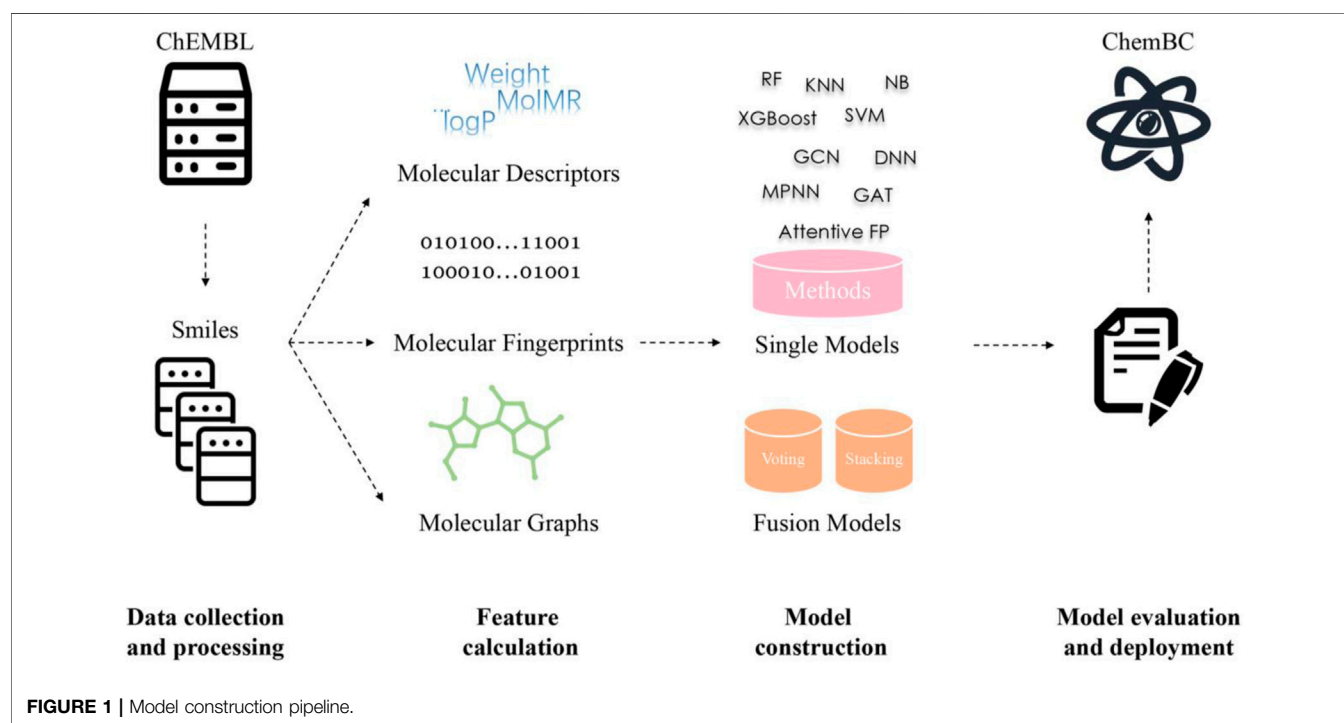
| Cell lines | Classification | No. of compounds | No. of scaffolds | Scaffolds/compounds (%) |
|------------|------------------------|------------------|------------------|-------------------------|
| MDA-MB-435 | HER-2+ ^a | 3,030 | 870 | 28.71 |
| MDA-MB-453 | HER-2+ | 440 | 215 | 48.86 |
| SK-BR-3 | HER-2+ | 2026 | 571 | 28.18 |
| MCF-7 | Luminal A ^b | 29,378 | 5,787 | 19.70 |
| T-47D | Luminal A | 3,135 | 926 | 29.54 |
| BT-474 | Luminal B ^c | 811 | 308 | 37.98 |
| MDA-MB-361 | Luminal B | 367 | 196 | 53.41 |
| HL-100 | Normal cell line | 316 | 110 | 34.81 |
| Bcap37 | TNBC ^d | 275 | 73 | 26.55 |
| BT-20 | TNBC | 292 | 146 | 50.00 |
| BT-549 | TNBC | 1,182 | 497 | 42.05 |
| HS-578T | TNBC | 469 | 215 | 45.84 |
| MDA-MB-231 | TNBC | 11,202 | 2,672 | 23.85 |
| MDA-MB-468 | TNBC | 1986 | 685 | 34.49 |

^aHER-2+: HER2-positive breast cancers.

^bLuminal A: Luminal A breast cancer is hormone-receptor positive (estrogen-receptor and/or progesterone-receptor positive), HER2-negative, and has low levels of the protein Ki-67, which helps control how fast cancer cells grow.

^cLuminal B: Luminal B breast cancer is hormone-receptor positive (estrogen-receptor and/or progesterone-receptor positive), and either HER2 positive or HER2 negative with high levels of Ki-67.

^dTNBC: triple-negative breast cancer.



algorithms (KNN, NB, RF, SVM, and XGBoost) and one deep learning DNN method. For these traditional ML methods, the optimized RDKit-descriptors were obtained using the SelectPercentile module (Percentile = 30) implemented in the scikit-learn package and then used as input features to construct models. Each model is denoted as a combination of a given molecular representation and ML algorithm (e.g., RF:RDKitDes). For each cell dataset and the corresponding ML methods, hyperparameters were optimized based on the validation sets (detailed in the Methods section), and the best set of

hyperparameters are shown in **Supplementary Table S1**. The detailed performance results for descriptor-based models are listed in **Supplementary Table S2**. The performance of the models (F1 score, BA, and AUC) for the test sets is summarized in **Figure 2**. Overall, most descriptor-based models performed well in BC cell inhibitory prediction tasks, achieving a mean F1 score and BA value > 0.5. The RF model performed the best in all cell lines, with higher average F1 scores (0.840 ± 0.073), BA (0.725 ± 0.073), and AUC (0.835 ± 0.067). Meanwhile, the XGBoost model also achieved good and/or

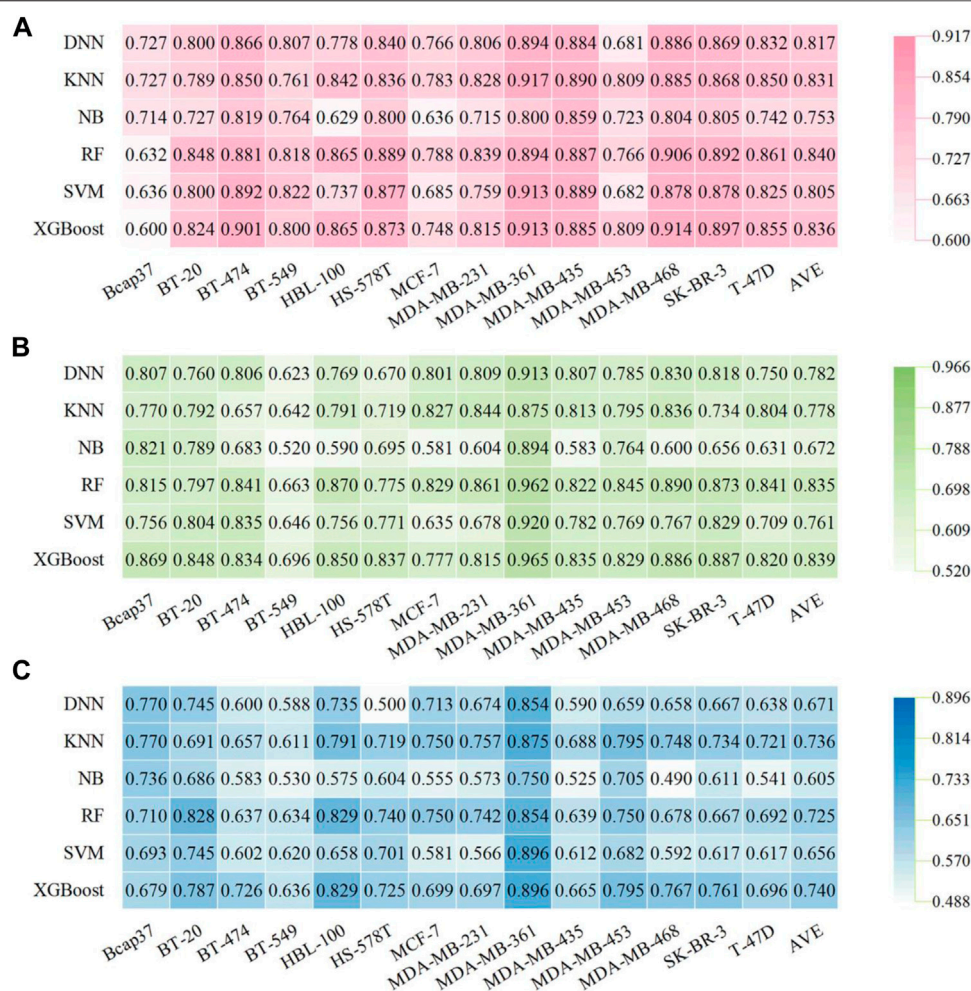


FIGURE 2 | Performance of descriptor-based BC prediction models. **(A)** F1 scores of descriptor-based models. **(B)** AUC results of descriptor-based models. **(C)** BC results of descriptor-based models.

comparable performance results (Figure 2). The detailed best-performing RF:RdkitDes models results were achieved in five breast cancer cell lines (BT-20, HS-578T, MCF-7, MDA-MB-231, and T-47D), while the XGBoost:RDKitDes models also showed superior performance in five breast-associated cell lines (BT-474, HBL-100, MDA-MB-453, MDA-MB-468, and SK-BR-3). The KNN:RDKitDes models exhibited the best performance in the Bcap37, MDA-MB-361, and MDA-MB-435 cell lines. The SVM:RDKitDes models performed well in BT-549.

3.3 Performance of Fingerprint-Based Prediction Models for Breast-Associated Cells

There were 336 models developed based on four types of fingerprints (Morgan, MACCS, Atompairs, and PharmacPFP) using six types of ML algorithms (KNN, NB, RF, SVM, XGBoost, and DNN). The detailed performance results for fingerprint-based models are listed in Supplementary Tables S3–S6. The F1,

AUC, and BA values of the test sets are shown in Figures 3, 4 and Supplementary Figure S3. Taking the average F1 score as a point metric into consideration, the numbers of cell lines for which each model was identified as the best-performing are shown in Figure 5. No model, fingerprint, or ML algorithm could be identified as the best-performing for the 14 cell line datasets, demonstrating that it is necessary to screen different fingerprints and different ML algorithms for the current breast cell-associated modeling datasets (Figures 5B–F). Although the characteristics of the four molecular fingerprints are different, the RF models performed better than the other five ML models against most of the 14 cell lines (Figures 3, 4, 5A). Meanwhile, the Morgan fingerprint represents the best molecular feature representation because the ML models based on Morgan fingerprints achieved the best results for these modeling datasets (Table 2). Global analysis of four fingerprint-based models also demonstrated that RF methods can achieve a better performance than other ML methods, with the highest average F1 score (0.848 ± 0.006), BA (0.750 ± 0.013), and AUC (0.853 ± 0.009).

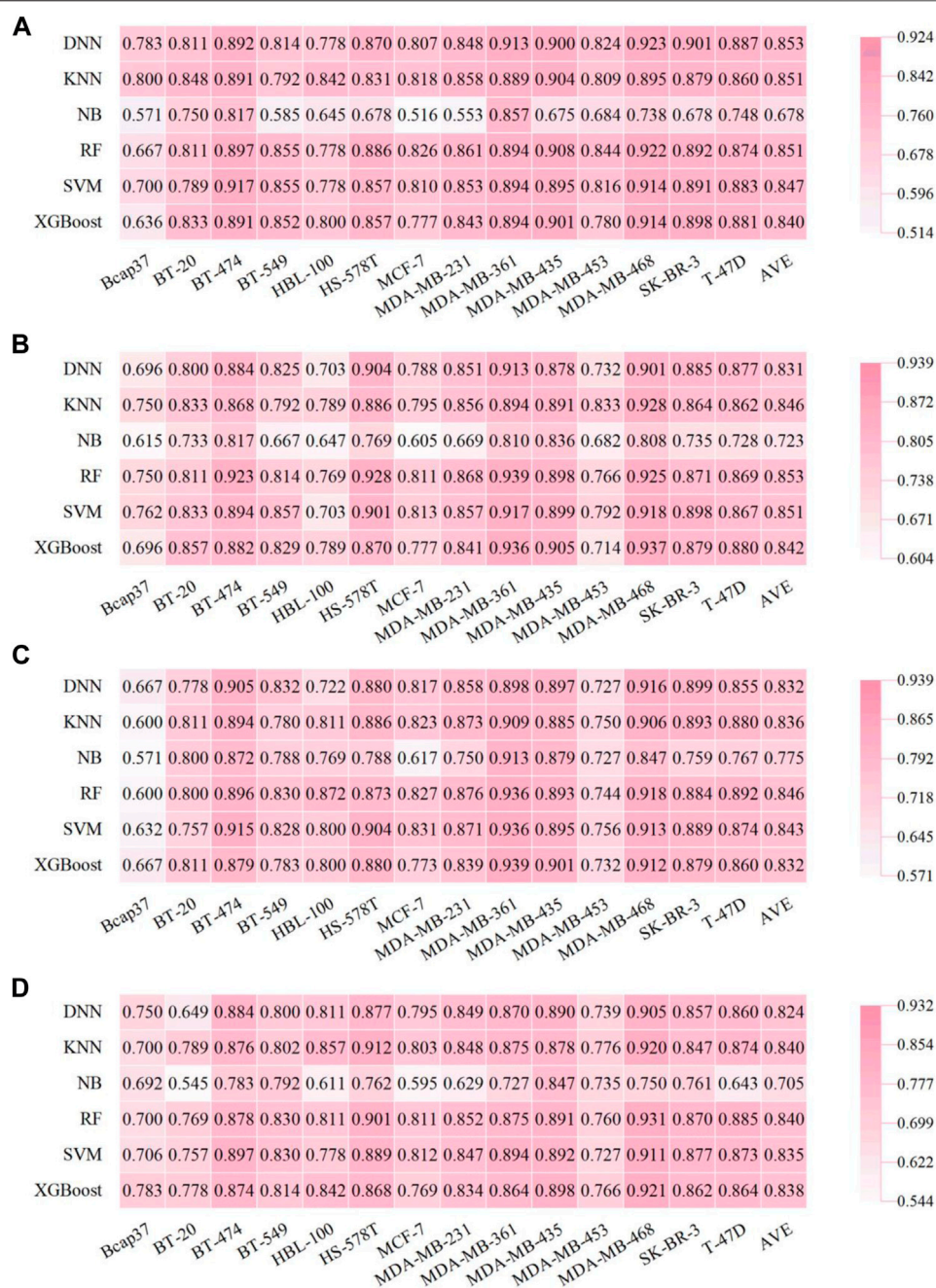


FIGURE 3 | Performance of fingerprint-based BC prediction models. **(A)** F1 scores of the AtomPairs-based models. **(B)** F1 scores of the MACCS-based models. **(C)** F1 scores of the Morgan-based models. **(D)** F1 scores of the Pharmacophore-based models.

3.4 Performance of Graph-Based Prediction Models for Breast-Associated Cells

Compared with the traditional pre-tailored molecular descriptors and/or fingerprints, the key feature of GNN is its capacity to automatically learn task-specific molecular representations using graph convolutions. The SOAT accuracies of GNN models and their variants (e.g., GCN, MPNN, GAT, and Attentive FP) have been reported in various molecular property prediction tasks (Wu

et al., 2018; Yang et al., 2019; Xiong et al., 2020). Therefore, 56 molecular graph-based models were established using four types of DL algorithms, including GCN, MPNN, GAT, and Attentive FP. The detailed performance results of molecular graph-based models are listed in **Supplementary Table S7**. As shown in **Figure 6**, the Attentive FP models exhibited the overall best performance compared with other GNN methods, with a relatively higher average F1 score (0.831 ± 0.070) and AUC (0.809 ± 0.086). The BA results are shown in **Supplementary Figure S4**. **Figure 6C**

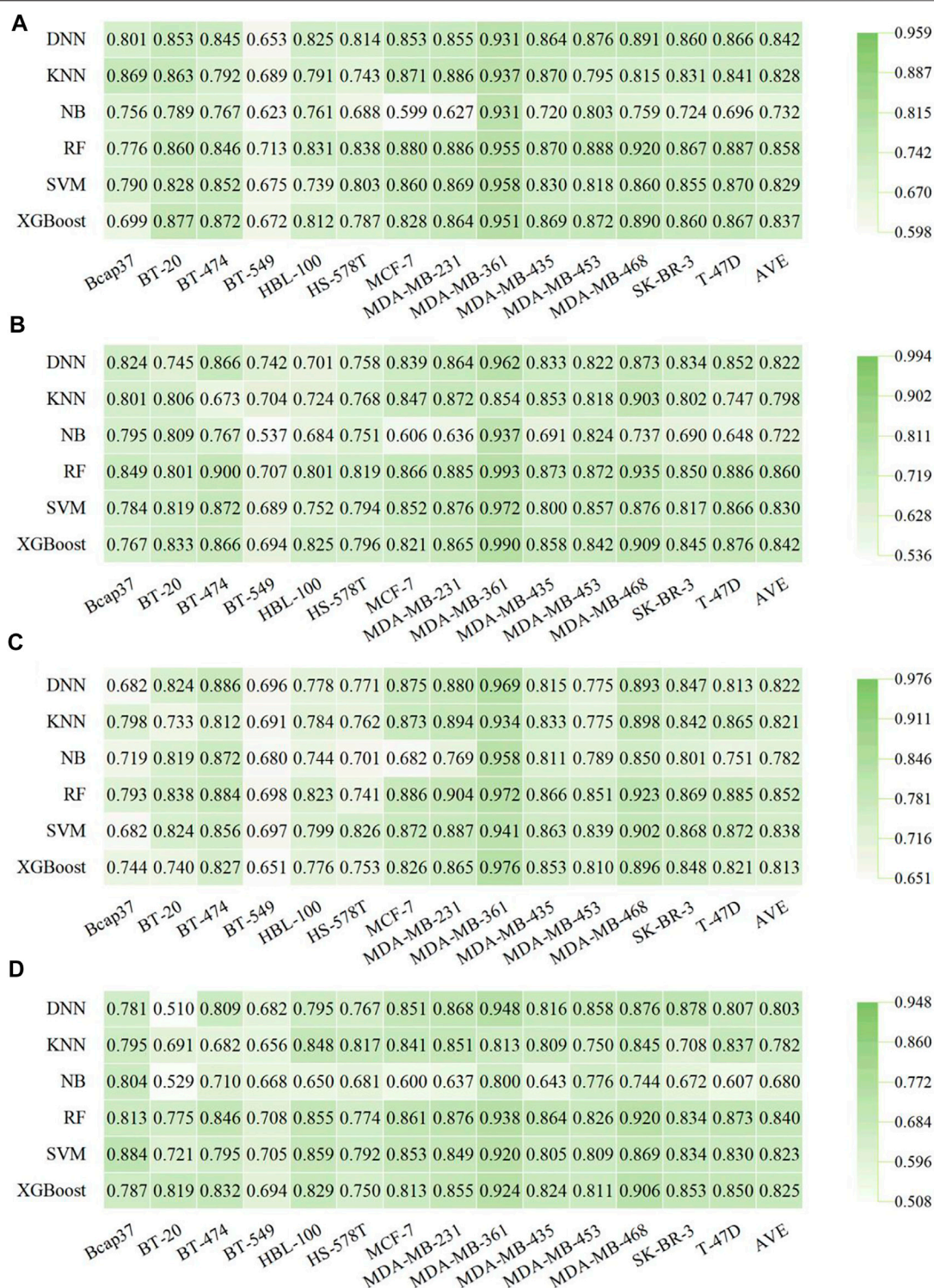


FIGURE 4 | Performance of fingerprint-based BC prediction models. **(A)** AUC results of the AtomPairs-based models. **(B)** AUC results of the MACCS-based models. **(C)** AUC results of the Morgan-based models. **(D)** AUC results of the Pharmacophore-based models.

shows that the Attentive FP models performed the best in six breast cancer cell lines including Bcap37, MCF-7, MDA-MB-453, MDA-MB-468, SK-BR-3, and T-47D, making it the most frequent choice.

The GCN models showed the best performance in four breast cell lines (BT-549, HBL-100, MDA-MB-231, and MDA-MB-361), the MPNN models performed the best in BT-20 and BT-474 cell lines,

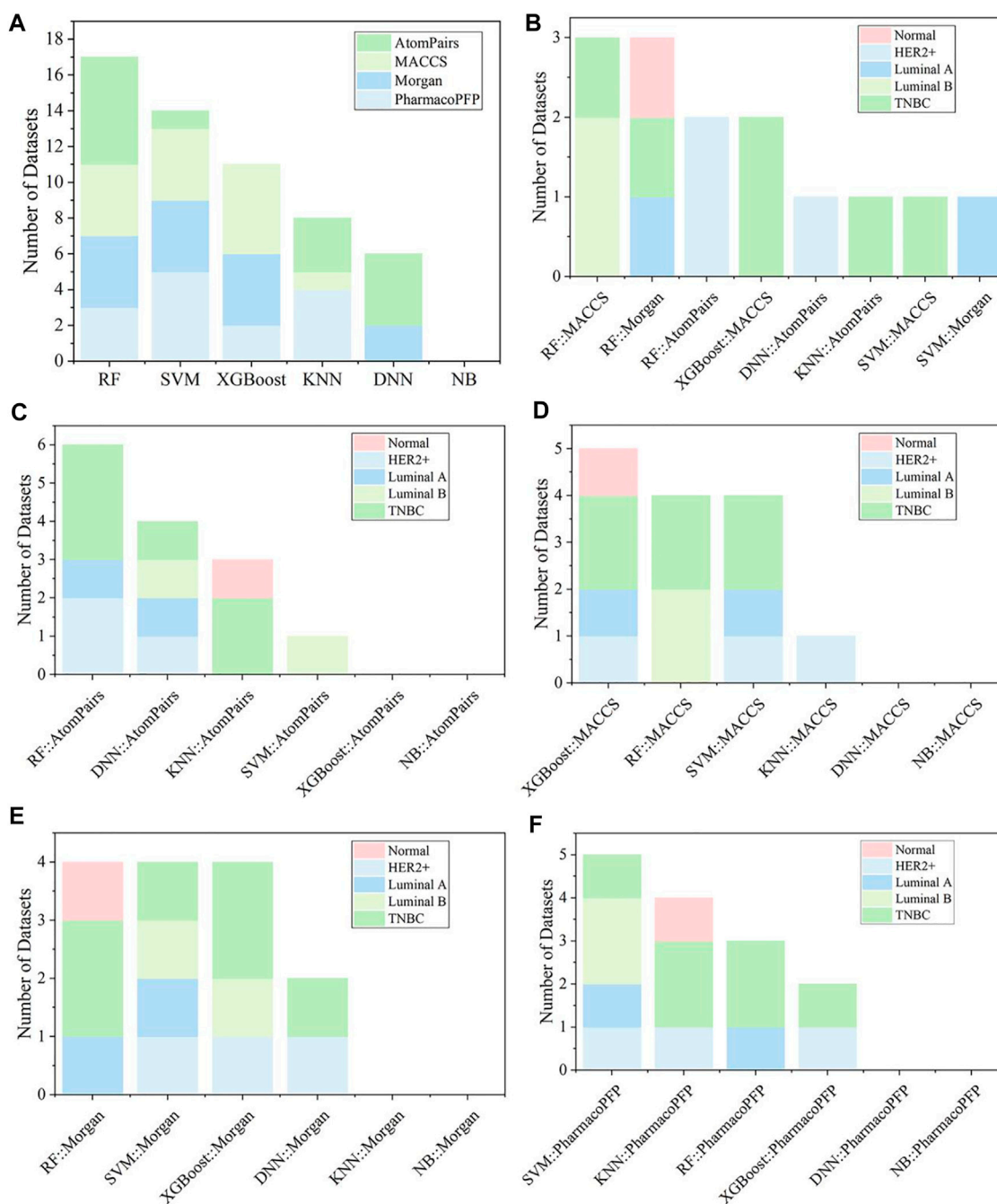


FIGURE 5 | (A) Summary of the optimal models for each fingerprint-based feature. **(B)** The best models among various fingerprint-based models for different kinds of breast cell lines. The optimal models based on **(C)** AtomPairs, **(D)** MACCS, **(E)** Morgan, and **(F)** PharmacolPP for different subtypes of breast cell lines.

and the GAT models performed the best in HS-578T and MDA-MB-435 cell lines.

One advantage of the DL model is its capacity for multi-task model building for attribute-related datasets to improve the accuracy of the single-task model (Li et al., 2018). Therefore, the multi-task models were trained by the entire 13 breast cancer cell-compound datasets based on the features of the Morgan fingerprints using DNN and molecular graphs

using GCN, Attentive FP. **Supplementary Table S8** shows that the AUC of the multi-task models was not better than that of the single-task models. Further data point distribution analysis found that the number of common compounds shared by 13 cell line datasets was small (only 12 molecules, **Supplementary Figure S5**), which explains the poor performance results (**Supplementary Table S8**) of the multi-task models.

TABLE 2 | Optimal models in different datasets and the evaluation of test datasets.

| Molecular features | Algorithms | F1 scores ^j | BA ^k | AUC ^l |
|--------------------|----------------------|------------------------|-----------------|------------------|
| Morgan | DNN ^a | 0.832 ± 0.080 | 0.735 ± 0.058 | 0.822 ± 0.078 |
| | KNN ^b | 0.836 ± 0.084 | 0.771 ± 0.063 | 0.821 ± 0.069 |
| | NB ^c | 0.775 ± 0.094 | 0.720 ± 0.079 | 0.782 ± 0.078 |
| | RF ^d | 0.846 ± 0.087 | 0.754 ± 0.068 | 0.852 ± 0.072 |
| | SVM ^e | 0.843 ± 0.084 | 0.747 ± 0.067 | 0.838 ± 0.072 |
| | XGBoost ^f | 0.832 ± 0.076 | 0.728 ± 0.062 | 0.813 ± 0.079 |
| | Mean | 0.827 ± 0.026 | 0.743 ± 0.019 | 0.821 ± 0.024 |
| MACCS | DNN | 0.831 ± 0.076 | 0.737 ± 0.060 | 0.822 ± 0.067 |
| | KNN | 0.846 ± 0.050 | 0.759 ± 0.056 | 0.798 ± 0.067 |
| | NB | 0.723 ± 0.077 | 0.637 ± 0.073 | 0.722 ± 0.103 |
| | RF | 0.853 ± 0.066 | 0.761 ± 0.064 | 0.860 ± 0.067 |
| | SVM | 0.851 ± 0.064 | 0.755 ± 0.059 | 0.830 ± 0.068 |
| | XGBoost | 0.842 ± 0.074 | 0.760 ± 0.056 | 0.842 ± 0.068 |
| | Mean | 0.824 ± 0.050 | 0.735 ± 0.049 | 0.812 ± 0.049 |
| AtomPairs | DNN | 0.853 ± 0.050 | 0.759 ± 0.057 | 0.842 ± 0.063 |
| | KNN | 0.851 ± 0.037 | 0.781 ± 0.051 | 0.828 ± 0.064 |
| | NB | 0.678 ± 0.099 | 0.668 ± 0.083 | 0.732 ± 0.085 |
| | RF | 0.851 ± 0.066 | 0.753 ± 0.054 | 0.858 ± 0.059 |
| | SVM | 0.847 ± 0.062 | 0.737 ± 0.069 | 0.829 ± 0.066 |
| | XGBoost | 0.840 ± 0.074 | 0.755 ± 0.041 | 0.837 ± 0.075 |
| | Mean | 0.820 ± 0.070 | 0.742 ± 0.039 | 0.821 ± 0.045 |
| Molecular Graph | Attentive FP | 0.831 ± 0.070 | 0.721 ± 0.086 | 0.809 ± 0.087 |
| | GAT ^g | 0.810 ± 0.086 | 0.695 ± 0.088 | 0.774 ± 0.075 |
| | GCN ^h | 0.818 ± 0.076 | 0.710 ± 0.091 | 0.798 ± 0.100 |
| | MPNN ⁱ | 0.821 ± 0.080 | 0.696 ± 0.109 | 0.781 ± 0.090 |
| | Mean | 0.820 ± 0.009 | 0.708 ± 0.011 | 0.793 ± 0.015 |
| PharmacoPFP | DNN | 0.824 ± 0.072 | 0.705 ± 0.091 | 0.803 ± 0.105 |
| | KNN | 0.840 ± 0.060 | 0.755 ± 0.075 | 0.782 ± 0.070 |
| | NB | 0.705 ± 0.088 | 0.619 ± 0.075 | 0.680 ± 0.080 |
| | RF | 0.840 ± 0.064 | 0.731 ± 0.070 | 0.840 ± 0.060 |
| | SVM | 0.835 ± 0.068 | 0.722 ± 0.064 | 0.823 ± 0.059 |
| | XGBoost | 0.838 ± 0.049 | 0.727 ± 0.072 | 0.825 ± 0.058 |
| | Mean | 0.814 ± 0.054 | 0.710 ± 0.047 | 0.792 ± 0.059 |
| RDKit | DNN | 0.817 ± 0.063 | 0.671 ± 0.089 | 0.782 ± 0.070 |
| | KNN | 0.831 ± 0.053 | 0.736 ± 0.065 | 0.778 ± 0.068 |
| | NB | 0.753 ± 0.068 | 0.605 ± 0.083 | 0.672 ± 0.108 |
| | RF | 0.840 ± 0.073 | 0.725 ± 0.073 | 0.835 ± 0.067 |
| | SVM | 0.805 ± 0.091 | 0.656 ± 0.086 | 0.761 ± 0.077 |
| | XGBoost | 0.836 ± 0.084 | 0.740 ± 0.071 | 0.839 ± 0.060 |
| | Mean | 0.814 ± 0.032 | 0.689 ± 0.054 | 0.778 ± 0.061 |

^aDNN: Deep neural networks.^bKNN: K-Nearest Neighbor.^cNB: Naïve Bayesian.^dRF: Random forest.^eSVM: Support vector machine.^fXGBoost: Extreme gradient boosting.^gGCN: Graph convolutional networks.^hGAT: Graph attention network.ⁱMPNN: Message passing neural networks.^jF1 scores: F1-measure.^kBA: Balanced accuracy.^lAUC: Area under the receiver operating characteristics curve.

3.5 The Optimal Model for Each Breast Cell Line and Further Validation

Comparison of the established molecular descriptor-, fingerprint-, and graph-based models showed that Eq. 1 the RF algorithm had a better performance capability than the other five ML methods, with higher average metric values of F1 score, BA, and AUC

(Table 2) in both descriptor- and fingerprint-based models, while XGBoost also achieved comparable results for these 14 modeling datasets (Table 2 and Figure 5A); 2) among the established 56 graph-based models, Attentive FP architecture outperformed the other three deep graph learning approaches (i.e., GCN, MPNN, and GAT) on average across all 14 datasets (Table 2);

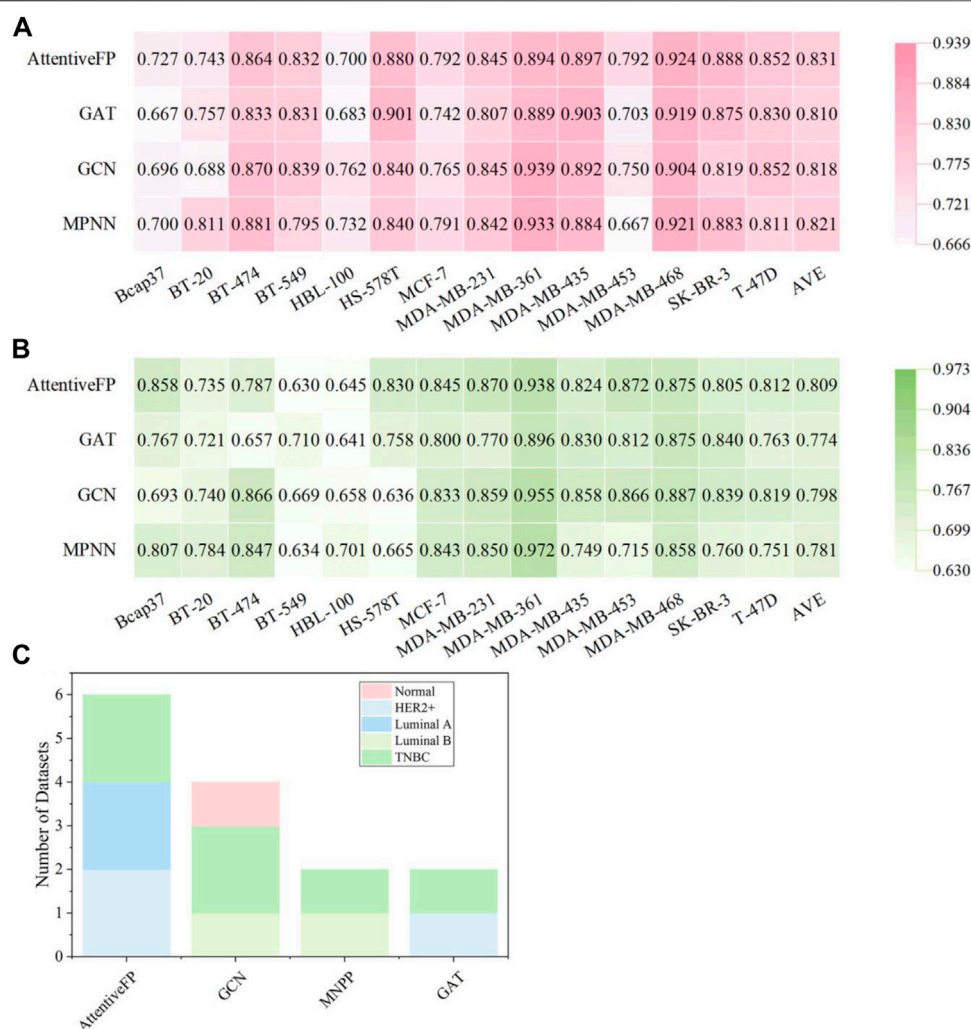


FIGURE 6 | Performance of graph-based BC prediction models. **(A)** F1 scores of graph-based models. **(B)** AUC results of graph-based models. **(C)** The optimal models based on molecular graph for different subtypes of breast cell lines.

and 3) the performance of molecular fingerprint-based models is generally better than that of both descriptor- and graph-based models at least in these 14 datasets (Table 2), implying that graph DL methods do not achieve better results than the traditional ML learning methods (especially for the two most efficient algorithms, XGBoost and RF), which is consistent with a recent systematic comparison study (Jiang D. et al., 2021).

According to the metrics of F1 score, BA, and AUC from the test sets, the optimal in silico predictive model for each breast cell line is listed in **Supplementary Table S9**. Fingerprint-based RF models performed the best because they ranked first in eight of 14 cell lines. Fingerprint-based XGBoost and SVM models are tied for second place and performed best in two of 14 breast cell lines each. For example, the RF:Morgan model achieved higher prediction results against MDA-MB-231 and T-47D breast cancer cell lines, with ACC values of 83.7 and 84.0%, respectively, and AUC values of 0.904 and 0.885, respectively. The lack of

selectivity for cancer cells rather than normal cells is one of the main factors that limit the development of anticancer drugs for clinical use (Dy and Adjei, 2013; Guo et al., 2020). For one normal breast cell line (HBL-100), the RF:Morgan model also showed good prediction results, with ACC and AUC values of 83.9%, and 0.823, respectively, suggesting that this model can be used to detect whether a given molecule selectively inhibits breast cancer cells over normal human breast cells.

Model fusion may improve the classification prediction performance of a single model by combining the classification prediction results from the corresponding multiple models. Both voting and stacking methods were used in this study for model fusion. As shown in Table 2, Morgan fingerprint-based models performed the best in different kinds of fingerprint-based models with an average F1 score of 0.827 ± 0.026 , and RF, XGBoost, and SVM algorithms performed best in most of the datasets (Figures

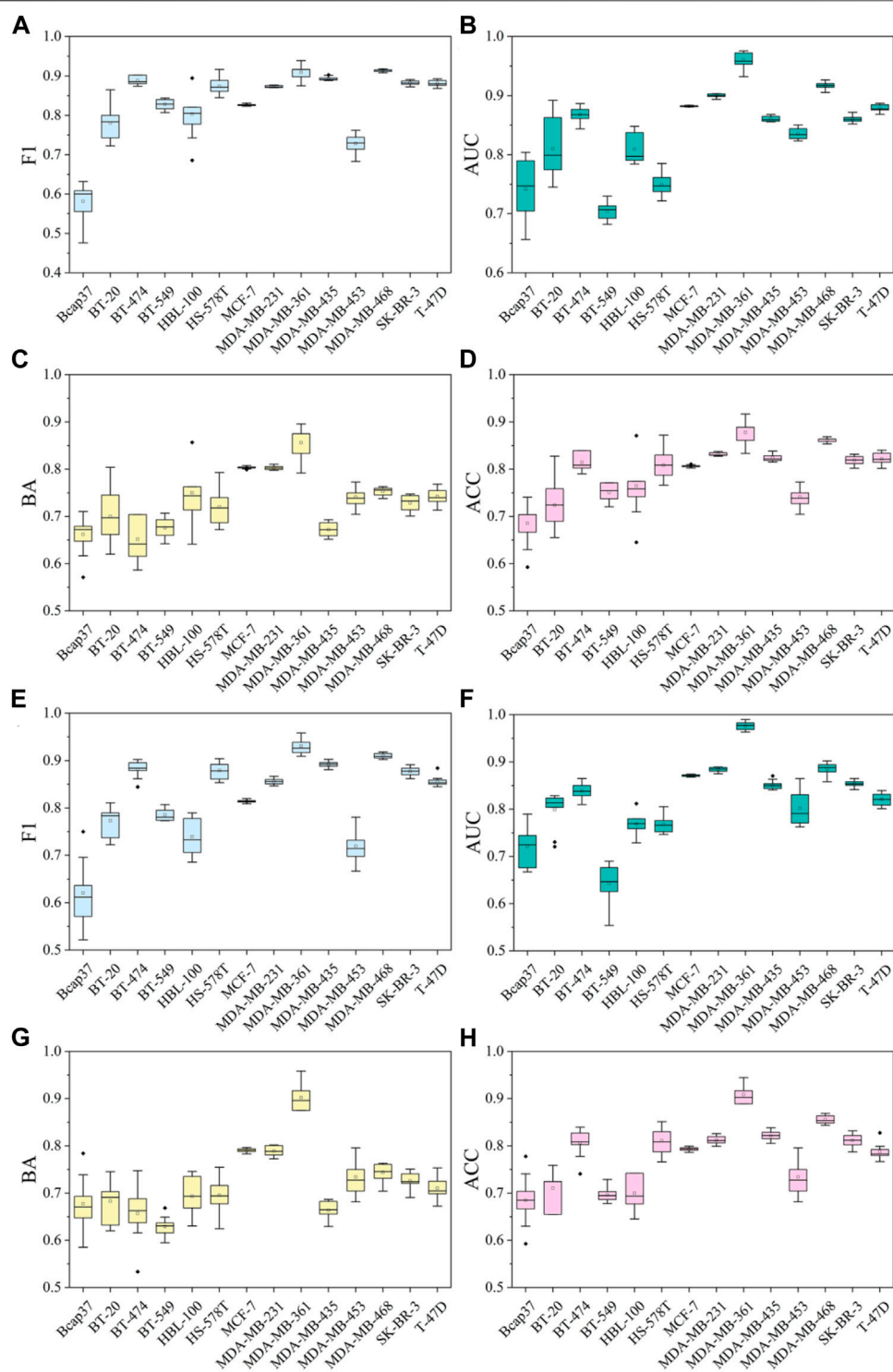


FIGURE 7 | The performance of 10-fold cross-validation results in RF:Morgan and XGBoost:Morgan models. **(A–D)** F1 scores, AUC, BA, and ACC results in RF: Morgan models. **(E–H)** F1 scores, AUC, BA, and ACC results in XGBoost:Morgan models.

5A,E). Therefore, RF, SVM, and XGBoost models for model fusion were applied based on Morgan fingerprints. A total of 112 fusion models were established, and detailed performance

results for these voting and stacking models are listed in **Supplementary Tables S10, S11**. As shown in **Supplementary Figure S6**, the average F1 scores of voting or

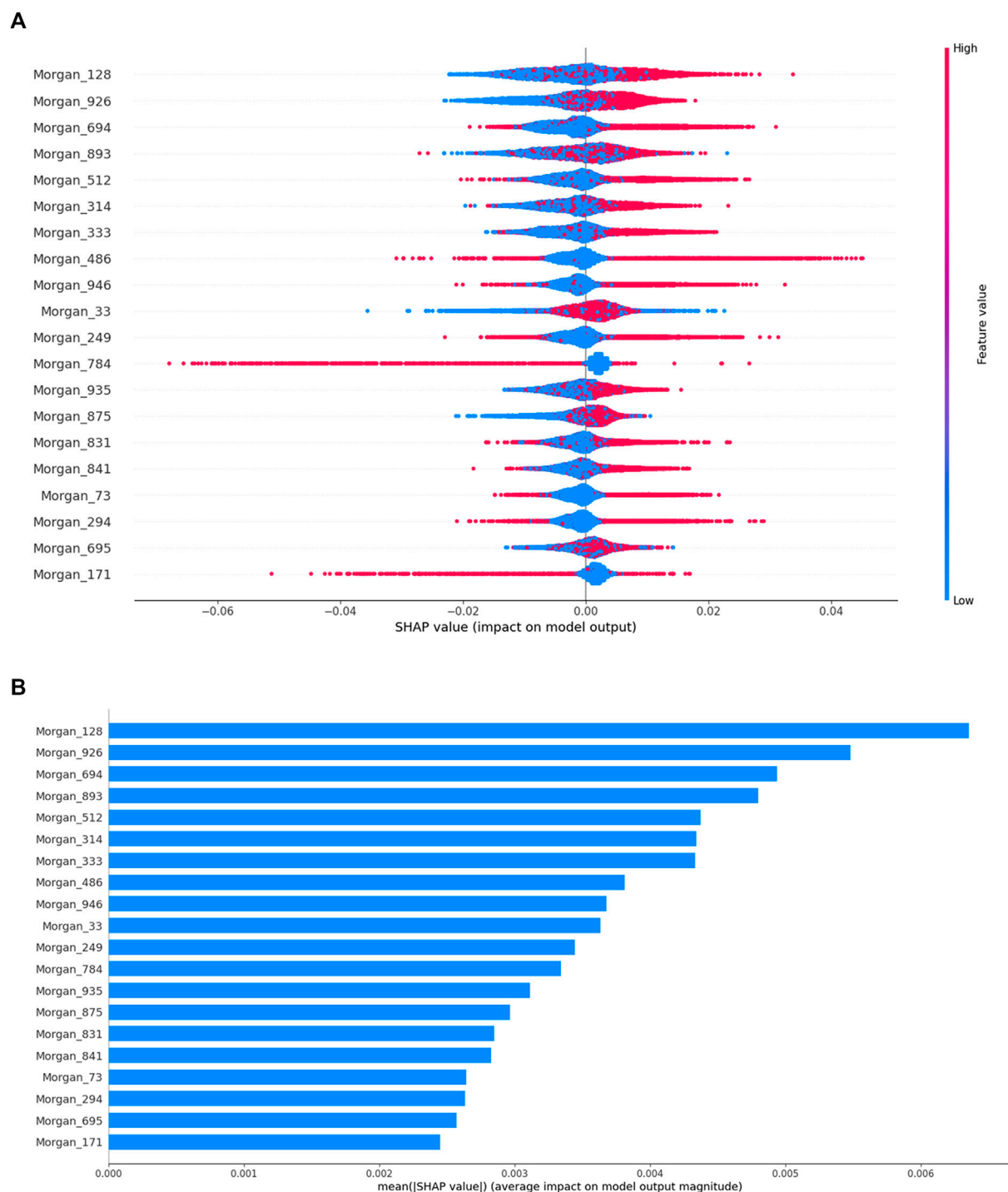


FIGURE 8 | Based on the top 20 most important features of the RF:Morgan model in MDA-MB-231, **(A)** the SHAP values for each molecular substructure, and **(B)** the mean of the absolute value of the SHAP value for each molecular substructure.

stacking models were similar in each dataset. In all the datasets of breast cell lines, the RF + XGBoost voting model showed the best average performance among fusion models, with average F1, BA, and AUC of 0.849 ± 0.066 , 0.749 ± 0.075 , and 0.845 ± 0.075 , respectively. The fusion models based on Morgan

fingerprints were slightly but not significantly better than the single models.

To validate the stability and reliability of the models presented, 10-fold cross-validation and 10 different random seeds of data were used to retrain the models based on the

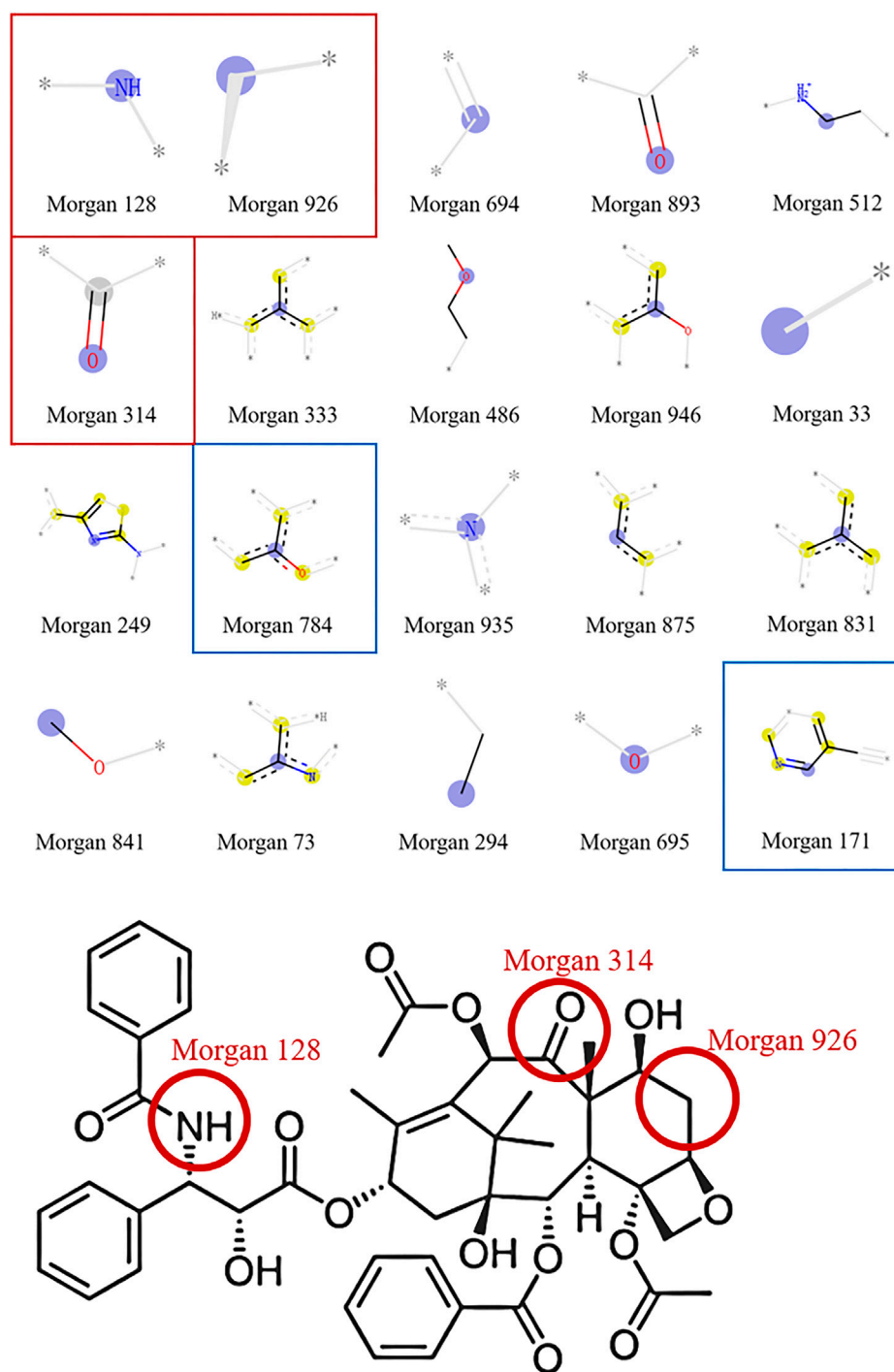


FIGURE 9 | Important molecular substructures of the RF:Morgan model in MDA-MB-231 and the chemical structural of paclitaxel.

combination of Morgan fingerprints and two ML algorithms (RF and XGBoost). The performance of 10-fold cross-validation classification models is summarized in **Supplementary Table S12** and **Figure 7**. Overall, all RF:Morgan models performed well, showing high F1 scores of 0.582–0.914, AUC values of 0.704–0.960, and ACC values of 0.685–0.878. XGBoost:Morgan models showed a similar trend in the 10-fold cross-validation

experiment. In 14 cell line datasets, both RF:Morgan and XGBoost:Morgan models consistently exhibited better performance with different seeds (**Supplementary Figure S7**), and the performance showed comparable or smaller variation compared with the previous models based on a specific random seed. Taken together, these results demonstrate that the models presented in this study show stability and reliability. Y-scrambling testing was

system called ChemBC (<http://chembc.idruglab.cn/>). To expand the AD threshold of the established model, we retained models for each breast cell line according to the combination of Morgan fingerprint and RF using the entire dataset, and then implemented these retained models into ChemBC and its local version. According to the 10-fold cross-validation (AUC = 0.780–0.928, ACC = 0.714–0.880), the retrained models for 14 breast cell line datasets showed excellent predictive performance. ChemBC was developed based on the Django framework using the Python package. The main functional module of ChemBC is prediction (**Figure 10**) in which users can upload and/or online draw a structure to easily and quickly predict the inhibitory activity against 13 breast cancer cell lines and one normal breast cell line. In addition, a local version executable software (<https://github.com/idruglab/ChemBC>) was developed to perform large-scale VS screening.

Taking paclitaxel as an example, it has a predicted score of 1.0 in the MDA-MB-231 model, proving that it has a strong inhibitory effect on the MDA-MB-231 cell line. Meanwhile, it has a predicted score of 0.8 in the normal breast cell line (HBL-100), suggesting that it is also toxic to the normal breast cell. Therefore, the ChemBC webserver can not only predict whether the compound has an inhibitory effect on breast cancer cells but also predict whether the compound is toxic to one normal breast cell.

4 CONCLUSION

In this study, we collected datasets of phenotypic compound-cell association bioactivity toward 13 breast cancer cell lines and one normal breast cell line and constructed 588 models based on three molecular representatives, including molecular descriptors, fingerprints, and graphs using five conventional ML and five DL algorithms. Compared with these established models, the performance of RF:Morgan models was superior to that of the other models based on molecular descriptors and graphs. Based on RF:Morgan models, the important favorable and unfavorable fragments for each breast cell line generated using SHAP algorithms will be helpful for lead optimization or the design of new agents with better anti-BC activity. Although some fusion models based on voting and stacking methods showed better performance than single models, the observed improvement was minor. Finally, the online platform ChemBC and its local version

software were developed based on well-established models, which could contribute to research aimed at designing and discovering new anti-BC agents. With the growth of compound toxicity data for BC and normal breast cell lines, we will add more prediction models in future studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

LW conceived and designed the experiments. SH, DZ, YL, and HC collected and processed the data, implemented the algorithm and created the web-server. SH performed the analysis and wrote the manuscript. LW offered support and critically revised the manuscript. JZ and YC are cooperators. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (Nos 81973241 and 82060625), the Natural Science Foundation of Guangdong Province (2020A1515010548), the Guizhou Provincial Natural Science Foundation ((2020)1Z073), and the National Science Foundation of Health and Family planning Commission of Guizhou Province (gzwjkj2019-1-178).

ACKNOWLEDGMENTS

We acknowledge the use of computational resources from the SCUT supercomputing platform.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2021.796534/full#supplementary-material>

REFERENCES

- Albertini, C., Salerno, A., de Sena Murteira Pinheiro, P., and Bolognesi, M. L. (2021). From Combinations to Multitarget-Directed Ligands: A Continuum in Alzheimer's Disease Polypharmacology. *Med. Res. Rev.* 41 (5), 2606–2633. doi:10.1002/med.21699
- Ashdown, G. W., Dimon, M., Fan, M., Sánchez-Román Terán, F., Witmer, K., Gaboriau, D. C. A., et al. (2020). A Machine Learning Approach to Define Antimalarial Drug Action from Heterogeneous Cell-Based Screens. *Sci. Adv.* 6 (39), eaba9338. doi:10.1126/sciadv.aba9338
- Bemis, G. W., and Murcko, M. A. (1996). The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* 39 (15), 2887–2893. doi:10.1021/jm9602928
- Berg, E. L. (2021). The Future of Phenotypic Drug Discovery. *Cell Chem. Biol.* 28 (3), 424–430. doi:10.1016/j.chembiol.2021.01.010
- Breunig, M. M., Kriegl, H.-P., Ng, R. T., and Sander, J. (2000). "LoF, SIGMOD Rec.," in proceedings of the 2000 ACM SIGMOD international conference on Management of data, 93–104. doi:10.1145/335191.335388
- Brower, V. (2013). Cardiotoxicity Debated for Anthracyclines and Trastuzumab in Breast Cancer. *J. Natl. Cancer Inst.* 105 (12), 835–836. doi:10.1093/jnci/djt161
- Buckner, F. S., Buchynskyy, A., Nagendar, P., Patrick, D. A., Gillespie, J. R., Herbst, Z., et al. (2020). Phenotypic Drug Discovery for Human African

- Trypanosomiasis: A Powerful Approach. *Trop. Med. Infect. Dis.* 5 (1), 23. doi:10.3390/tropicalmed5010023
- Cameron, D., Piccart-Gebhart, M. J., Gelber, R. D., Procter, M., Goldhirsch, A., de Azambuja, E., et al. (2017). 11 Years' Follow-Up of Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Early Breast Cancer: Final Analysis of the HERceptin Adjuvant (HERA) Trial. *Lancet* 389 (10075), 1195–1205. doi:10.1016/S0140-6736(16)32616-2
- Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985). Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* 25 (2), 64–73. doi:10.1021/ci00046a002
- Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D., and Carpenter, A. E. (2021). Image-based Profiling for Drug Discovery: Due for a Machine-Learning Upgrade. *Nat. Rev. Drug Discov.* 20 (2), 145–159. doi:10.1038/s41573-020-00117-w
- Chen, L., Wang, L., Gu, Q., and Xu, J. (2014). An In Silico Protocol for Identifying mTOR Inhibitors from Natural Products. *Mol. Divers.* 18 (4), 841–852. doi:10.1007/s11030-014-9543-5
- Chen, T., and Guestrin, C. (2016). "Xgboost: A Scalable Tree Boosting System," in proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.
- Childers, W. E., Elokely, K. M., and Abou-Gharbia, M. (2020). The Resurrection of Phenotypic Drug Discovery. *ACS Med. Chem. Lett.* 11 (10), 1820–1828. doi:10.1021/acsmchemlett.0c00006
- Cover, T., and Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Trans. Inform. Theor.* 13 (1), 21–27. doi:10.1109/TIT.1967.1053964
- Croston, G. E. (2017). The Utility of Target-Based Discovery. *Expert Opin. Drug Discov.* 12 (5), 427–429. doi:10.1080/17460441.2017.1308351
- Daniyal, A., Santoso, I., Gunawan, N. H. P., Barliana, M. I., and Abdullah, R. (2021). Genetic Influences in Breast Cancer Drug Resistance, Bctt. *Breast cancer* 13, 59–85. doi:10.2147/BCTT.S284453
- Duda, R. O., and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.
- Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* 42 (6), 1273–1280. doi:10.1021/ci010132r
- Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). "Convolutional Networks on Graphs for Learning Molecular Fingerprints," in 29th Annual Conference on Neural Information Processing Systems. doi:10.1021/ci010132r
- Dy, G. K., and Adjei, A. A. (2013). Understanding, Recognizing, and Managing Toxicities of Targeted Anticancer Therapies. *CA Cancer J. Clin.* 63 (4), 249–279. doi:10.3322/caac.21184
- Escala-Garcia, M., Morra, A., Canisius, S., Chang-Claude, J., Kar, S., Zheng, W., et al. (2020). Breast Cancer Risk Factors and Their Effects on Survival: a Mendelian Randomisation Study. *BMC Med.* 18 (1), 327. doi:10.1186/s12916-020-01797-2
- Fields, F. R., Freed, S. D., Carothers, K. E., Hamid, M. N., Hammers, D. E., Ross, J. N., et al. (2020). Novel Antimicrobial Peptide Discovery Using Machine Learning and Biophysical Selection of Minimal Bacteriocin Domains. *Drug Dev. Res.* 81 (1), 43–51. doi:10.1002/ddr.21601
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). "Neural Message Passing for Quantum Chemistry," in International Conference on Machine Learning: PMLR, 1263–1272.
- Gobbi, A., and Poppinga, D. (1998). Genetic Optimization of Combinatorial Libraries. *Biotechnol. Bioeng.* 61 (1), 47–54. doi:10.1002/(sici)1097-0290(199824)61:1<47:aid-bit9>3.0.co;2-z
- Guo, Q., Luo, Y., Zhai, S., Jiang, Z., Zhao, C., Xu, J., et al. (2019). Discovery, Biological Evaluation, Structure-Activity Relationships and Mechanism of Action of Pyrazolo[3,4-b]pyridin-6-One Derivatives as a New Class of Anticancer Agents. *Org. Biomol. Chem.* 17 (25), 6201–6214. doi:10.1039/c9ob00616h
- Guo, Q., Zhang, H., Deng, Y., Zhai, S., Jiang, Z., Zhu, D., et al. (2020). Ligand- and Structural-Based Discovery of Potential Small Molecules that Target the Colchicine Site of Tubulin for Cancer Treatment. *Eur. J. Med. Chem.* 196, 112328. doi:10.1016/j.ejmech.2020.112328
- Harbeck, N., Thomssen, C., and Gnant, M. (2013). St. Gallen 2013: Brief Preliminary Summary of the Consensus Discussion. *Breast Care (Basel)* 8 (2), 102–109. doi:10.1159/000351193
- Heikamp, K., and Bajorath, J. (2014). Support Vector Machines for Drug Discovery. *Expert Opin. Drug Discov.* 9 (1), 93–104. doi:10.1517/17460441.2014.866943
- Hughes, R. E., Elliott, R. J. R., Dawson, J. C., and Carragher, N. O. (2021). High-content Phenotypic and Pathway Profiling to advance Drug Discovery in Diseases of Unmet Need. *Cel Chem. Biol.* 28 (3), 338–355. doi:10.1016/j.chembiol.2021.02.015
- Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., et al. (2021a). Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminform* 13 (1), 12. doi:10.1186/s13321-020-00479-8
- Jiang, Z., Xu, J., Yan, A., and Wang, L. (2021b). A Comprehensive Comparative Assessment of 3D Molecular Similarity Tools in Ligand-Based Virtual Screening. *Brief. Bioinf* 22 (6), bbab231. doi:10.1093/bib/bbab231
- Kc, G. B., Bocci, G., Verma, S., Hassan, M. M., Holmes, J., Yang, J. J., et al. (2021). A Machine Learning Platform to Estimate Anti-SARS-CoV-2 Activities. *Nat. Mach. Intell.* 3 (6), 527–535. doi:10.1038/s42256-021-00335-w
- Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided Mol. Des.* 30 (8), 595–608. doi:10.1007/s10822-016-9938-8
- Kipf, T. N., and Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks. arXiv:1609.02907.
- Landrum, G. (2016). RDKit: Open-Source Cheminformatics Software, 2016. Available: <http://www.rdkit.org>.
- Li, S., Ding, Y., Chen, M., Chen, Y., Kirchmair, J., Zhu, Z., et al. (2021). HDAC3i-Finder: A Machine Learning-based Computational Tool to Screen for HDAC3 Inhibitors. *Mol. Inf.* 40 (3), 2000105. doi:10.1002/minf.202000105
- Li, X., Xu, Y., Lai, L., and Pei, J. (2018). Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharm.* 15 (10), 4336–4345. doi:10.1021/acs.molpharmaceut.8b00110
- Li, Y., and Li, Z. (2021). Potential Mechanism Underlying the Role of Mitochondria in Breast Cancer Drug Resistance and its Related Treatment Prospects. *Front. Oncol.* 11, 629614. doi:10.3389/fonc.2021.629614
- Li, Y., Zhao, C., Zhang, J., Zhai, S., Wei, B., and Wang, L. (2019). HybridMolDB: A Manually Curated Database Dedicated to Hybrid Molecules for Chemical Biology and Drug Discovery. *J. Chem. Inf. Model.* 59 (10), 4063–4069. doi:10.1021/acs.jcim.9b00314
- Liao, M., Zhang, J., Wang, G., Wang, L., Liu, J., Ouyang, L., et al. (2021). Small-Molecule Drug Discovery in Triple Negative Breast Cancer: Current Situation and Future Directions. *J. Med. Chem.* 64 (5), 2382–2418. doi:10.1021/acs.jmedchem.0c01180
- Liu, P., Li, H., Li, S., and Leung, K. S. (2019). Improving Prediction of Phenotypic Drug Response on Cancer Cell Lines Using Deep Convolutional Network. *BMC Bioinformatics* 20 (1), 408. doi:10.1186/s12859-019-2910-6
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* 2 (1), 56–67. doi:10.1038/s42256-019-0138-9
- Lundberg, S. M., and Lee, S. I. (2017). "A Unified Approach to Interpreting Model Predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).
- Luo, Y., and Wang, L. (2017). Discovery and Development of ATP-Competitive mTOR Inhibitors Using Computational Approaches. *Curr. Pharm. Des.* 23 (29), 4321–4331. doi:10.2174/1381612823666170710150604
- Luo, Y., Zeng, R., Guo, Q., Xu, J., Sun, X., and Wang, L. (2019). Identifying a Novel Anticancer Agent with Microtubule-Stabilizing Effects through Computational Cell-Based Bioactivity Prediction Models and Bioassays. *Org. Biomol. Chem.* 17 (6), 1519–1530. doi:10.1039/c8ob02193g
- Malandraki-Miller, S., and Riley, P. R. (2021). Use of Artificial Intelligence to Enhance Phenotypic Drug Discovery. *Drug Discov. Today* 26 (4), 887–901. doi:10.1016/j.drudis.2021.01.013
- McCulloch, W. S., and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.* 5 (4), 115–133. doi:10.1007/bf02478259
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2019). ChEMBL: towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* 47 (D1), D930–D940. doi:10.1093/nar/gky1075
- Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., and Prunotto, M. (2017). Opportunities and Challenges in Phenotypic Drug Discovery: an Industry Perspective. *Nat. Rev. Drug Discov.* 16 (8), 531–543. doi:10.1038/nrd.2017.111

- Morphy, R., Kay, C., and Rankovic, Z. (2004). From Magic Bullets to Designed Multiple Ligands. *Drug Discov. Today* 9 (15), 641–651. doi:10.1016/S1359-6446(04)03163-0
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Quancard, J., Bach, A., Cox, B., Craft, R., Finsinger, D., Guéret, S. M., et al. (2021). The European Federation for Medicinal Chemistry and Chemical Biology (EFMC) Best Practice Initiative: Phenotypic Drug Discovery. *ChemMedChem* 16 (11), 1736–1739. doi:10.1002/cmdc.202100041
- Rogers, D., and Hahn, M. (2010). Extended-connectivity Fingerprints. *J. Chem. Inf. Model.* 50 (5), 742–754. doi:10.1021/ci100050t
- Schirle, M., and Jenkins, J. L. (2016). Identifying Compound Efficacy Targets in Phenotypic Drug Discovery. *Drug Discov. Today* 21 (1), 82–89. doi:10.1016/j.drudis.2015.08.001
- Shah, A. N., and Gradishar, W. J. (2018). Adjuvant Anthracyclines in Breast Cancer: What Is Their Role. *Oncologist* 23 (10), 1153–1161. doi:10.1634/theoncologist.2017-0672
- Shang, J., Dai, X., Li, Y., Pistozzi, M., and Wang, L. (2017). HybridSim-VS: a Web Server for Large-Scale Ligand-Based Virtual Screening Using Hybrid Similarity Recognition Techniques. *Bioinformatics* 33 (21), 3480–3481. doi:10.1093/bioinformatics/btx418
- Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell* 180 (2), 688–702.e13. doi:10.1016/j.cell.2020.01.021
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660
- Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* 43 (6), 1947–1958. doi:10.1021/ci034160g
- Sydow, D., Burggraaf, L., Szengel, A., van Vlijmen, H. W. T., IJzerman, A. P., van Westen, G. J. P., et al. (2019). Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.* 59 (5), 1728–1742. doi:10.1021/acs.jcim.8b00832
- Velicković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph Attention Networks. arXiv:1710.10903.
- Wang, L., Li, Y. C., Xu, M. Y., Pang, X. Q., Liu, Z. H., and Tan, W. (2016b). Chemical Fragment-Based CDK4/6 Inhibitors Prediction and Web Server. *RSC Adv.* 6 (21), 16972–16981. doi:10.1039/c5ra23289a
- Wang, L., Chen, L., Yu, M., Xu, L. H., Cheng, B., Lin, Y. S., et al. (2016a). Discovering New mTOR Inhibitors for Cancer Treatment through Virtual Screening Methods and *In Vitro* Assays. *Sci. Rep.* 6 (1), 18987–19013. doi:10.1038/srep18987
- Wang, L., Pang, X., Li, Y., Zhang, Z., and Tan, W. (2017a). RADER: a Rapid DEcoy Retriever to Facilitate Decoy Based Assessment of Virtual Screening. *Bioinformatics* 33 (8), 1235–1237. doi:10.1093/bioinformatics/btw783
- Wang, L., Wang, Y., Tian, Y., Shang, J., Sun, X., Chen, H., et al. (2017b). Design, Synthesis, Biological Evaluation, and Molecular Modeling Studies of Chalcone-Rivastigmine Hybrids as Cholinesterase Inhibitors. *Bioorg. Med. Chem.* 25 (1), 360–371. doi:10.1016/j.bmc.2016.11.002
- Wang, L., Li, Y., Xu, M., Pang, X., Liu, Z., Tan, W., et al. (2016b). Chemical Fragment-Based CDK4/6 Inhibitors Prediction and Web Server. *RSC Adv.* 6 (21), 16972–16981. doi:10.1039/c5ra23289a
- Wermuth, C. G. (2004). Multitargeted Drugs: the End of the "One-Target-One-Disease" Philosophy. *Drug Discov. Today* 9 (19), 826–827. doi:10.1016/S1359-6446(04)03213-1
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: a Benchmark for Molecular Machine Learning. *Chem. Sci.* 9 (2), 513–530. doi:10.1039/c7sc02664a
- Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., et al. (2020). Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* 63 (16), 8749–8760. doi:10.1021/acs.jmedchem.9b00959
- Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., et al. (2019). Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* 59 (8), 3370–3388. doi:10.1021/acs.jcim.9b00237
- Ye, Q., Chai, X., Jiang, D., Yang, L., Shen, C., Zhang, X., et al. (2021). Identification of Active Molecules against Mycobacterium tuberculosis through Machine Learning. *Brief. Bioinf.* 22 (5), bbab068. doi:10.1093/bib/bbab068
- Zernov, V. V., Balakin, K. V., Ivaschenko, A. A., Savchuk, N. P., and Pletnev, I. V. (2003). Drug Discovery Using Support Vector Machines. The Case Studies of Drug-Likeness, Agrochemical-Likeness, and Enzyme Inhibition Predictions. *J. Chem. Inf. Comput. Sci.* 43(6), 2048–2056. doi:10.1021/ci0340916
- Zhang, W., Wang, L., Zhang, L., Chen, W., Chen, X., Xie, M., et al. (2014). Synthesis and Biological Evaluation of Steroidal Derivatives as Selective Inhibitors of AKR1B10. *Steroids* 86, 39–44. doi:10.1016/j.steroids.2014.04.010
- Zheng, J. X., Xia, S., Lv, S., Zhang, Y., Bergquist, R., and Zhou, X. N. (2021). Infestation Risk of the Intermediate Snail Host of Schistosoma Japonicum in the Yangtze River Basin: Improved Results by Spatial Reassessment and a Random Forest Approach. *Infect. Dis. Poverty* 10 (1), 74. doi:10.1186/s40249-021-00852-1
- Zoffmann, S., Vercruysse, M., Benmansour, F., Maunz, A., Wolf, L., Blum Marti, R., et al. (2019). Machine Learning-Powered Antibiotics Phenotypic Drug Discovery. *Sci. Rep.* 9 (1), 5013. doi:10.1038/s41598-019-39387-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 He, Zhao, Ling, Cai, Cai, Zhang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Multiparameter Optimization of Trypanocidal Cruzain Inhibitors With *In Vivo* Activity and Favorable Pharmacokinetics

Ivani Pauli^{1†}, Celso de O. Rezende Jr.^{2†}, Brian W. Slafer², Marco A. Desso², Mariana L. de Souza¹, Leonardo L. G. Ferreira¹, Abraham L. M. Adjanohun³, Rafaela S. Ferreira³, Luma G. Magalhães¹, Renata Krogh¹, Simone Michelin-Duarte¹, Ricardo Vaz Del Pintor⁴, Fernando B. R. da Silva⁴, Fabio C. Cruz⁴, Luiz C. Dias^{2*} and Adriano D. Andricopulo^{1*}

OPEN ACCESS

Edited by:

Salvatore Salomone,
University of Catania, Italy

Reviewed by:

Aparecida Donizette Malvezi,
State University of Londrina, Brazil
Galia Andrea Ramirez-Toloza,
University of Chile, Chile

*Correspondence:

Adriano D. Andricopulo
aandrico@ifsc.usp.br
Luiz C. Dias
ldias@unicamp.br

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 10 September 2021

Accepted: 22 November 2021

Published: 05 January 2022

Citation:

Pauli I, Rezende Jr. CO, Slafer BW, Desso MA, de Souza ML, Ferreira LLG, Adjanohun ALM, Ferreira RS, Magalhães LG, Krogh R, Michelin-Duarte S, Del Pintor RV, da Silva FBR, Cruz FC, Dias LC and Andricopulo AD (2022) Multiparameter Optimization of Trypanocidal Cruzain Inhibitors With *In Vivo* Activity and Favorable Pharmacokinetics. *Front. Pharmacol.* 12:774069. doi: 10.3389/fphar.2021.774069

¹Laboratório de Química Medicinal e Computacional, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, Brazil, ²Instituto de Química, Universidade Estadual de Campinas, Campinas, Brazil, ³Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, ⁴Departamento de Farmacologia, Universidade Federal de São Paulo, São Paulo, Brazil

Cruzain, the main cysteine protease of *Trypanosoma cruzi*, plays key roles in all stages of the parasite's life cycle, including nutrition acquisition, differentiation, evasion of the host immune system, and invasion of host cells. Thus, inhibition of this validated target may lead to the development of novel drugs for the treatment of Chagas disease. In this study, a multiparameter optimization (MPO) approach, molecular modeling, and structure-activity relationships (SARs) were employed for the identification of new benzimidazole derivatives as potent competitive inhibitors of cruzain with trypanocidal activity and suitable pharmacokinetics. Extensive pharmacokinetic studies enabled the identification of metabolically stable and permeable compounds with high selectivity indices. CYP3A4 was found to be involved in the main metabolic pathway, and the identification of metabolic soft spots provided insights into molecular optimization. Compound **28**, which showed a promising trade-off between pharmacodynamics and pharmacokinetics, caused no acute toxicity and reduced parasite burden both *in vitro* and *in vivo*.

Keywords: chagas disease, cruzain, medicinal chemistry, drug design, multiparameter optimization, pharmacokinetics, molecular modeling

INTRODUCTION

Endemic in Latin America, Chagas disease affects 6–7 million people worldwide and has become an emerging public health problem in nonendemic countries¹. Among nonendemic nations, the greatest burden occurs in the United States, which is estimated to have approximately 300,000 cases of the disease (Pérez-Molina and Molina, 2018). Chagas disease kills ~12,000 people annually, and 70 million people are at risk of infection in the Americas². Moreover, the disease is an important cause of infectious cardiopathy worldwide, playing a key role in the global prevalence of cardiovascular

¹[https://www.who.int/news-room/fact-sheets/detail/chagas-disease-\(american-trypanosomiasis\)](https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)).

²<https://www.paho.org/en/topics/chagas-disease>.

disease (Bern, 2015; Cucunubá et al., 2016). Chagas disease significantly impacts the productivity of endemic countries, which are estimated to lose more than US \$7.2 billion per year because of the disease (GBD DALYs and HALE Collaborators, 2016; Arnal et al., 2019). According to the World Health Organization (WHO), the development of innovative therapeutic approaches is required for this neglected tropical disease (NTD) because of the lack of efficient control measures and the insufficient research and development (R&D) funding. The need for novel therapeutic approaches has become more evident this year, as the WHO released a new roadmap for NTDs for 2021–2030, whose target is to eliminate the epidemics of these diseases by 2030. Chemotherapy for Chagas disease consists of benznidazole (BZ) and nifurtimox, two nitro compounds that have limited efficacy and produce serious adverse reactions that lead up to 40% of patients to discontinue treatment (Rodrigues Coura and De Castro, 2002). Given these shortcomings, the development of novel, effective and safe drugs for the treatment of Chagas disease is critically needed.

Cruzain (EC 3.4.22.51), the main cysteine protease of *Trypanosoma cruzi*, has been broadly explored as a molecular target in Chagas disease drug discovery (Engel et al., 1998; McKerrow, 1999; Jose Cazzulo et al., 2001; Massarico Serafim et al., 2014). This enzyme plays a key role in all stages of the parasite's life cycle, participating in processes such as nutrition, differentiation, evasion of the host immune system, and invasion of host cells (Ferreira and Andricopulo, 2017). Genetic studies of *T. cruzi* and the efficacy of cruzain inhibitors in reducing parasite load *in vivo* have validated the enzyme as a molecular target for the discovery of novel drugs for Chagas disease (Zanatta et al., 2008; Doyle et al., 2011; Ndao et al., 2014). Following these investigations, various classes of cruzain inhibitors, such as nitroalkenes, vinyl sulfones, thiosemicarbazones, and triazoles, have been described in the literature (Rogers et al., 2012; Avelar et al., 2015; Espíndola et al., 2015; Neitz et al., 2015; Latorre et al., 2016). In this work, we describe the design, synthesis, and *in vitro* and *in vivo* evaluations of novel benzimidazole derivatives. In addition to improving pharmacodynamic properties, such as binding affinity and potency, we evaluated the pharmacokinetic (PK) profile of newly synthesized and previously described benzimidazoles (Ferreira et al., 2014) by applying a multiparameter optimization (MPO) approach. MPO has increasingly been adopted in the early phases of pharma R&D to exclude pipeline compounds that feature poor PK profiles as early as possible (Eddershaw et al., 2000; Andricopulo and Montanari, 2005; Wang et al., 2007; Wang, 2009; Wang and Skolnik, 2009). This study led to the discovery of potent cruzain inhibitors with trypanocidal activity and innovatively contributed to the identification of compounds with improved safety and PK profiles to be explored for Chagas disease drug discovery.

MATERIALS AND METHODS

Expression and Purification

Pro-cruzain truncated at the C-terminus was expressed and purified using a previously described protocol (Ferreira et al.,

2019). *Escherichia coli* (strain M15) cultures were grown overnight at 37°C and 200 rpm in Luria Bertani (LB) medium supplemented with ampicillin (100 µg/ml) and kanamycin (50 µg/ml). Next, the cultures were diluted 10-fold in fresh LB medium supplemented with 0.5 M NaCl, 0.2% glucose, 1 mM betaine, 0.5 M sorbitol, 100 µg/ml ampicillin, and 50 µg/ml kanamycin and incubated at 37°C and 200 rpm. At an optical density (OD₆₀₀) of 0.9, the cultures were incubated at 47°C for 20 min to promote the expression of chaperones. Then, the expression of cruzain was induced by adding isopropyl β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.2 mM, which was followed by overnight incubation of the cultures at 20°C and 200 rpm. Next, the cultures were centrifuged (5,000 rpm, 30 min, 4°C), and the cells were suspended in 50 ml of lysis buffer (300 mM NaCl, 50 mM Tris-HCl, and 1.6 mg/ml lysozyme, pH 8.0) per liter of culture and lysed by sonication (12 cycles of 30 s). This cell lysate was centrifuged (9,000 rpm, 30 min, 4°C), and the supernatant was collected. Cruzain was precipitated by incubation with 35% ammonium sulfate (2 h), and this suspension was centrifuged at 9,000 rpm for 30 min at 4°C. The precipitated cruzain was resuspended in lysis buffer, and the sample was dialyzed to eliminate ammonium sulfate. The soluble fraction of the dialysate was loaded on a Ni-NTA column (Qiagen, Hilden, Germany), and the contaminants were washed using washing buffer (300 mM NaCl, 50 mM Tris-HCl, and 10 mM imidazole, pH 8.0). Cruzain was eluted by applying an increasing imidazole gradient: 25, 50, 75, 100, and 250 mM. The fractions containing cruzain were pooled together and dialyzed against 1.5 L of 0.1 M acetate buffer, pH 5.5, and then concentrated to 0.5 mg/ml. Pro-cruzain was activated by incubation with activation buffer (100 mM sodium acetate, pH 5.5, 10 mM EDTA, 5 mM DTT, and 1 M NaCl) at 37°C. The activation of cruzain was monitored by following the enzymatic activity at 30-min intervals, and the process was observed to stop after approximately 1 h. After activation, the enzyme was diluted 20-fold in binding buffer (20 mM sodium phosphate and 150 mM NaCl, pH 7.2) and added to thiopropyl Sepharose 6B resin (GE Healthcare Life Sciences, Pittsburgh, PA). After overnight incubation at 4°C, the resin was loaded on a column, and cruzain was eluted with binding buffer supplemented with 20 mM DTT. Fractions containing cruzain were pooled together and stored in 0.1 M sodium acetate, pH 5.5, at –80°C.

Enzyme Kinetics Assays

Cruzain activity was followed by monitoring the cleavage of the fluorogenic substrate Z-Phe-Arg-aminomethyl coumarin (Z-FR-AMC), as previously described (Ferreira et al., 2019), using 96-well flat-bottom black plates and wavelengths of 355 nm for excitation and 460 nm for emission. All cruzain assays were performed in 0.1 M sodium acetate buffer with 5 mM dithiothreitol (DTT) and 0.01% Triton X-100, pH 5.5. The final concentration of cruzain was 1.5 nM, and the substrate concentration was 5.0 µM ($K_m = 1.6 \mu\text{M}$), except in the experiments for K_i determination, in which several concentrations of substrate were used. The cleavage of the

substrate was monitored for 5 min, and the activity was calculated based on the initial reaction rates compared with the rate of a DMSO control at 30°C. The IC₅₀ values were independently calculated by considering the rate measurements for at least six inhibitor concentrations, each evaluated in triplicate. To determine the mechanism of cruzain inhibition, eight concentrations of the substrate Z-FR-AMC and four concentrations of the inhibitor were employed, each in triplicate. Kinetic parameters were determined using the SigmaPlot (Systat Software Inc., Erkrath, Germany) enzyme kinetics module. Compounds were tested in two or three independent experiments. All enzyme assays were performed using varying Triton X-100 concentrations (0, 0.01, and 0.1%) (Ferreira et al., 2009). Compound concentrations were 100 µM in the single-dose percentage inhibition assays.

Rhodesain Assays

Rhodesain activity was measured using a fluorescence-based assay as previously described (Fonseca et al., 2015). The cleavage rates of the fluorogenic substrate Z-Phe-Arg-aminomethyl coumarin (Z-FR-AMC) were monitored at wavelengths of 340 nm for excitation and 440 nm for emission. All assays were performed in triplicate in a 0.1 M sodium acetate buffer, pH 5.5, with 1 mM beta-mercaptoethanol and 0.01% Triton X-100. The final concentration of rhodesain was 3 nM, and the substrate concentration was 2.5 µM. The cleavage of the substrate was followed by continuous reading for 5 min, and enzyme activity in the presence of 100 µM of each potential inhibitor was calculated based on initial velocity rates compared to DMSO controls. All compounds were tested in triplicate in three independent experiments.

Molecular Docking

The three-dimensional structures of the cruzain inhibitors were constructed using the standard geometric parameters embedded in SYBYL-X 2.1 (Certara, Princeton, NJ). Each compound was energetically minimized employing the Tripos force field (Clark et al., 1989) and Powell conjugate gradient method (Powell, 1977), with a convergence value of 0.05 kcal/mol.Å, and the Gasteiger-Hückel model was used for charge calculation (Gasteiger and Marsili, 1980). The molecules were docked using GOLD 5.3 (Cambridge Crystallographic Data Centre, Cambridge, United Kingdom) (Jones et al., 1997; Verdonk et al., 2003) against the X-ray structure of cruzain (PDB ID 3KKU, 1.28 Å) (Ferreira et al., 2010). The preparation of the cruzain structure consisted of removing all water molecules and inserting hydrogen atoms. The active site Cys25 was kept negatively charged, and His162 was kept protonated. The binding site was defined as a sphere with a 10 Å radius centered on the Cys25 sulfur atom. The default GOLD parameters were applied for the molecular docking runs, except for the search efficiency, which was changed to its maximum value of 200%. The generated poses were evaluated using the GoldScore scoring function, and the analysis of the binding conformations was visualized using PyMOL 3.1 (Schrödinger, New York, NY) (Lill and Danielson, 2011).

Biological Assays Against *T. cruzi* Intracellular Amastigotes

Biological assays against *T. cruzi* intracellular amastigotes were performed as reported previously using the *T. cruzi* Tulahuen strain, which is genetically engineered to express the *E. coli* β-galactosidase gene *lacZ* (Buckner et al., 1996). β-Galactosidase catalyzes a colorimetric reaction with chlorophenol red β-D-galactopyranoside (CPRG, Sigma Chemical Co., St. Louis, MO) as the substrate. The assays were conducted in 96-well tissue culture plates, and the compounds to be tested were prepared in 100% DMSO. Epimastigotes were maintained in liver infusion tryptone (LIT) enriched with 10% fetal calf serum (FCS), streptomycin, and penicillin at 28°C. Epimastigotes were converted to trypomastigotes by incubation in Grace's insect medium (Sigma-Aldrich, St. Louis, MO) enriched with 10% FCS at 28°C. Human HFF-1 fibroblasts were seeded at 2×10^3 /well in 80 µl of RPMI 1640 without phenol red and incubated overnight at 37°C and 5% CO₂. Trypomastigotes were seeded at 1.0×10^4 /well in 20 µl of RPMI 1640, and the plates were incubated at 37°C and 5% CO₂. The next day, the synthesized compounds were added (50 µl) in 3-fold serial dilutions at concentrations ranging from 0.4 to 300 µM, and the plates were incubated at 37°C and 5% CO₂. Each compound concentration was assayed in triplicate. After 120 h, 50 µl of chlorophenol red β-D-galactopyranoside (CPRG, Sigma-Aldrich) and IGEPAL CA-630 (Sigma-Aldrich) at a final concentration of 0.1% were added. The absorbance was measured at a wavelength of 570 nm in an automated microplate reader. The data were transferred to SigmaPlot 10.0 (Systat Software Inc., Erkrath, Germany) to determine the IC₅₀ values. Benznidazole (BZ, Sigma-Aldrich) was used as a positive control, and untreated wells (100% parasite growth) were used as negative controls in all plates. All compounds were tested in three independent assays.

Cytotoxicity Assays in HFF-1 Fibroblasts

The synthesized compounds were evaluated for their cytotoxicity against HFF-1 cells using the MTS assay (Promega, Madison, WI) (Bartrop et al., 1991) as previously described (Ferreira et al., 2019). HFF-1 fibroblasts were plated at 2×10^3 /well in 96-well culture plates in RPMI 1640 without phenol red enriched with 10% FCS and incubated overnight at 37°C and 5% CO₂. Next, 7 concentrations (0.1–100 µM) of the compounds were added in 3-fold serial dilutions, each concentration in triplicate, and the plates were incubated for 72 h at 37°C and 5% CO₂. Next, 20 µl of MTS was added to each well, and the plates were incubated for an additional 4 h at 37°C and 5% CO₂. The absorbance was measured at 490 nm using a spectrophotometer, and the data were transferred to SigmaPlot 10.0 (Systat Software Inc., Erkrath, Germany) to determine the IC₅₀ values. Doxorubicin (Sigma-Aldrich) was used as a positive control, and untreated wells (100% growth) were used as negative controls in all plates. All compounds were tested in two independent assays.

In Vitro Metabolic Stability in Liver Microsomes

Isolated mouse (BD Gentest, Bedford, MA) and human liver microsomes (XenoTech, Kansas City, KS) (Plant, 2004) were

added at a final concentration of 0.25 mg/ml to a solution containing 40 mM dibasic potassium phosphate and 10 mM monobasic potassium phosphate. Stock solutions of compounds at 5 mM were prepared in 100% DMSO. A 50:50 quenching solution of acetonitrile (ACN) and methanol (MeOH) was prepared. An NADPH solution was prepared at 10 mM. The preparations containing the microsomes were added to each well of the incubation plate (450 μ l), which was then heated to 37°C for 10 min. The compounds were added to the respective wells of the test plate (2 μ l). Then, 300 μ l of the microsome preparation was added to each well of the test plate. The test plate was heated under gentle rotation for 5 min at 37°C. Next, 90 μ l of the mixture contained in the test plate was added to the incubation plate, making a final volume of 540 μ l. Samples were collected at the following incubation times: 0, 5, 10, 15, 20, and 30 min. To the 0 min sample plate, quenching solution (180 μ l), NADPH (6 μ l), and the incubation plate mixture (54 μ l) were added to each well. The sample plate was then sealed, homogenized, and stored at 4°C. Then, 54 μ l of the NADPH solution was added to the incubation plate, which was homogenized. Before the collection of each sample at the established times, 45 μ l of the quenching solution was added to each well of the corresponding sample plates. After reaching the incubation times, 60 μ l of the mixture contained in the incubation plate was added to the corresponding sample plates. The sample plates were then sealed and stored at 4°C. After the collection of the last sample (30 min), all sample plates were centrifuged at 3,800 rpm for 30 min. The supernatant from each well was collected and transferred to clean plates for mass spectrometry. Five to 10 μ l of each sample was injected into an AB Sciex Triple Quad 5500 LC-MS/MS instrument.

***In Vitro* Metabolic Stability Using Recombinant CYP Enzymes**

Stock solutions of compounds at 5 mM were prepared in 100% DMSO. An NADPH solution at 10 mM was prepared in a 50 mM potassium phosphate buffer. A solution at 100 pmol/ml of each of the recombinant CYP450 enzymes (1A2, 2C8, 2C9, 2C19, 2D6, and 3A4) was prepared in 50 mM potassium phosphate buffer (Proctor et al., 2004). A 50:50 quenching solution of ACN and MeOH was prepared. 320 μ l of each CYP solution was added to the incubation plate, and 1 μ l of each compound was added to the compound plate. Then, 100 μ l of each CYP solution was added to the compound plate, and 10 μ l from the compound plate was added to the incubation plate. The incubation plate was heated to 37°C with 600 rpm rotation for 10 min. To the 0 min plate, 30 μ l of quenching solution, 1 μ l of NADPH, and 9 μ l of the incubation plate solution were added to each well, and the plate was sealed and stored at 4°C. Next, 10- μ l samples were collected from the incubation plate at different incubation times (5, 10, 20, 30, and 60 min). For each time point, a separate plate containing 30 μ l of quenching solution and incubation plate solution was used. After collecting the last sample, the plates were centrifuged for 20 min at 3,000 rpm. The supernatant from each well was collected and transferred to new plates for mass spectrometry. Five to 10 μ l of

each sample was injected into an AB Sciex Triple Quad 5500 LC-MS/MS instrument. The compounds were tested at a final concentration of 5 μ M.

***In Vitro* Metabolic Stability in Hepatocytes**

Rat and human hepatocytes were used (McGinnity, et al., 2004). Samples were collected at six different time points during the incubation period and analyzed by LC-MS/MS to determine the $T_{1/2}$ and the intrinsic clearance. Stock solutions of the CYP inhibitors azamulin (8.75 mM) and 1-ABT (350 mM) were prepared. Next, the test compounds were added to 96-well plates and incubated in Williams medium (Invitrogen, A12176-01) supplemented with 2 mM L-glutamine and 15 mM HEPES to reach a final concentration of 2 μ M. Then, compounds (50 μ l) were transferred to other plates, one plate for each time point (0, 15, 30, 60, 120, and 240 min), and incubated at 37°C and 5% CO₂ for 30 min. Cryopreserved hepatocytes were heated in a wet bath at 37°C and dispensed in InVitroGRO HT medium supplemented with 10% FCS, 0.15 μ M hydrocortisone, 0.2 mg/ml BSA, fructose, insulin, and amino acids. The cells were centrifuged at 500 rpm for 5 min. The supernatant was discarded, and the cells were resuspended in 1 ml of incubation medium heated to 37°C. Next, the cells were counted and diluted to 1×10^6 /ml in incubation medium. The cell suspension was divided into 3 samples: without CYP inhibitors; with 25 μ M azamulin, a CYP3A4 inhibitor; and with 1 mM 1-ABT, an inhibitor of all CYPs. Each hepatocyte sample with 12,500 cells/well was incubated with either azamulin or 1-ABT for 30 min (except the group without CYP inhibitor). To the plates with the test compounds, 12,500 cells/well were added. The plates were incubated in a shaker at 37°C, 5% CO₂, and 300 rpm. At the specified time points, the hepatocyte enzymatic activity was interrupted by the addition of 75 μ l of cold ACN, and the samples were read by LC-MS/MS.

Parallel Artificial Membrane Permeability Assay

The permeability of the compounds was assessed using the PAMPA method (Yu et al., 2015). The test compounds were dissolved in DMSO to a concentration of 5 mM. Next, the compounds were diluted in a stock plate in saline phosphate buffer (PBS), pH 6.5, containing 1% DMSO to a concentration of 1 μ M. Then, 300 μ l of each test compound was added to the donor plate. Afterward, 200 μ l of PBS buffer, pH 7.4, was added to the acceptor plate. The donor plate was attached to the acceptor plate. The assembled acceptor-donor plate was then incubated at 37°C for 5 h under gentle agitation. To analyze the concentration of the compounds by mass spectrometry, two analysis plates (one for the donor and another for the acceptor plate) containing 300 μ l of MeOH:ACN 50:50 were prepared. To the donor analysis plate, 90 μ l of PBS, pH 6.5, and 10 μ l of the content of the donor plate were added. For the acceptor analysis plate, 100 μ l of the content of the acceptor plate was added. To monitor any potential decomposition/intrinsic instability of the test compounds in solution, samples from the stock plate were subjected to LC-MS/MS. The compounds that presented $P_e \geq 1.5 \times 10^{-6}$ cm/s were

classified as permeable, while compounds that presented $P_e < 1.5 \times 10^{-6}$ cm/s were considered poorly permeable or nonpermeable.

Experimental Distribution Coefficient

For the eLogD assays (Waring, 2010), each compound (5 μ l) from a 5 mM stock solution was diluted in 245 μ l of a solution containing buffers A (5% MeOH, 10 mM ammonium acetate, pH 7.4) and B (100% MeOH, pH 7.4) (50:50). Nine control compounds with known eLogD and column retention times (acyclovir, atenolol, antipyrine, fluconazole, metoprolol, carbamazepine, ketoconazole, tolinaftate, and amiodarone; eLogD values ranging from -1.86 to 6.1) were subjected to LC-MS/MS in triplicate before and after the analysis of the test compounds. Retention times were recorded for each control and test compound in a C18 column. The retention time of each control compound was plotted against the respective eLogD values described in the literature (Lombardo et al., 2002; Lombardo et al., 2004; Alelyunas et al., 2010). The resulting linear equation ($y = mx + b$) was used to calculate the eLogD values of the test compounds, in which x is the retention time in minutes and y is the eLogD value.

Fraction Unbound

The fraction unbound (f_u) (Masimirembwa et al., 2003) of the test compounds was determined after incubation with different media, namely, plasma, microsomes, and buffer with 10% FCS. Equilibrium dialysis was performed in a 96-well plate (HT-Dialysis, Gales Ferry, CT) in which each well was divided by a semipermeable membrane (12–14 kDa cutoff). The test compounds diluted in either plasma, microsome suspension, or buffer were added to one side of the membrane. Potassium phosphate buffer (50 mM, pH 7.4) was added to the other side of the membrane. A standard curve was used to calculate the compound concentration (f_u) on each side of the membrane. Stock solutions of each compound were prepared in 100% DMSO to obtain a final concentration of 1 mM. The following compound concentrations were used: 1, 2, 20, 200, 2,000 and 5,000 nM. To 500 μ l of medium, 0.5 μ l of the compound stock solution was added, and this was applied to one side of the well. Each compound was evaluated in triplicate. Next, the plate was assembled, sealed, and incubated at 37°C under rotation (150 rpm) for 4 h. After the incubation period, the plates were subjected to LC-MS/MS.

Biotransformation and Analysis of Metabolites

Stock solutions of compounds at 5 mM were prepared in 100% DMSO for the biotransformation analyses (Obach, 1999). Test solutions of compounds at 1 mM were prepared in H₂O/MeOH: 2/1 (v/v). A solution of 10 mM NADPH, a 50:50 quenching solution of ACN and MeOH, and 100 mM phosphate buffer was prepared. Isolated mouse (BD Gentest, Bedford, MA) and human liver microsomes (XenoTech, Kansas City, KS) were dissolved to 2 mg/ml in phosphate buffer. One plate for the 0 min time point and another for the 60 min time point were prepared. To these plates, 178 μ l of the microsome solution and 2 μ l of the

compound stock solution (final test concentration of 5 μ M) were added. The plates were incubated at 37°C for 5 min under gentle agitation. Next, 400 μ l of the quenching solution was added to the 0 min plate, which was followed by the addition of 20 μ l of NADPH. The plate was sealed and kept under refrigeration (4°C). To the 60 min plate, 20 μ l of NADPH was added, and the plate was sealed and incubated at 37°C for 60 min under gentle agitation. Next, 400 μ l of quenching solution was added to the 60 min plate. The two plates were centrifuged (4°C, 3300 rpm, 30 min), and the supernatant was collected for mass spectrometry.

In Vivo Pharmacokinetics

The *in vivo* pharmacokinetic profiles of the compounds were determined using male CD1 mice weighing 50 g (Davies and Morris, 1993; Liu and Jia, 2007). The compounds were administered in a single dose orally (0.5 mg/kg) and intravenously (0.5 mg/kg). Stock solutions of the compounds in DMSO were diluted in Tween 80, PEG-400, and D5W (5% dextrose in water) at a ratio of 2:5:20:73 (v/v). The injection volume was 10 ml/kg. The remaining plasma concentration was monitored over time by LC-MS/MS by collecting blood samples (40 μ l) at 10, 25, and 50 min and 1, 3, 6, 9, 12, and 24 h after administration of the compound.

Pharmacokinetics Analysis by LC-MS/MS

For the biotransformation experiments and analysis of metabolites in microsomes, an Ultra-High-Pressure Liquid Chromatography instrument (UHPLC, Thermo Accela, Waltham, MA) (Spaggiari et al., 2014) connected to an automatic sample injector and a 1250 series pump was used. The UHPLC system was connected to a Thermo Fisher (Waltham, MA) LTQ Orbitrap mass spectrometer. For all other analyses, an AB Sciex Triple Quad 5500 coupled to a UHPLC equipped with a UV 1290 diode detector (Agilent Technologies, Santa Clara, CA) and a CTC PAL self-collecting system (LEAP Technologies, Carrboro, NC) was used. Q1 MS positive ion mode (300–500 Da) was used to detect the ions of the parent compounds. The UV detector was operated in spectral mode (250–280 nm). A Hypersil Gold C18 (2.1 mm \times 100 mm, 1.9 μ m, Thermo Fisher) HPLC column was used. The mobile phases were solvent A (0.1% formic acid in water) and solvent B (0.1% formic acid in ACN). The flow was adjusted to 0.55 ml/min, and the injection volume was adjusted to 20 μ l. The gradient started with 1% solvent B for 0.4 min, reached 40% (solvent B) in 2.3 min and 95% (solvent B) in 0.67 min, was maintained for 0.5 min, and returned to the initial condition of 1%. This condition was maintained for 1 min before injection of the next sample. The peak area ratio (peak area of the test compounds/peak area of the control compounds) was converted to the percentage of remaining compound, with the 0 min time point ratio set to 100%. $T_{1/2}$ and CL_{int} were calculated from the percentage of remaining compound versus the incubation time. From the resulting function, the slope (k) was determined. The equations $T_{1/2}(\text{min}) = \ln(2)/k$ and $CL_{int \text{ in vitro}} (\mu\text{L}/\text{min}/\text{mg}) = k \cdot 1000/0.25$ were used to determine $T_{1/2}$ and CL_{int} .

Animals for the *In Vivo* Assays

Thirty-day-old female Swiss mice weighing 20–25 g and procured from the Center for the Development of Experimental Models for Medicine and Biology (CEDEME/UNIFESP) served as the subjects for these experiments. Animals were housed (5–6 per cage) in polypropylene cages and kept under controlled temperature (22–23°C) and humidity on a 12-h light/dark cycle (12 h light, 12 h dark; lights on at 6:30 am). Rodent chow and water were available *ad libitum* throughout the experiments. The Committee of Ethics in Research of the Universidade Federal de São Paulo approved all the experiments (CEUA n° 5301080816).

Chemistry

Unless stated otherwise, all reactions were performed under an atmosphere of argon with dry solvents and magnetic stirring (detailed organic synthesis methods are in the **Supplementary Material**). Dichloromethane (DCM) and triethylamine (Et₃N) were distilled from CaH₂. Tetrahydrofuran (THF) was distilled from sodium/benzophenone. Dimethyl formamide (DMF) was purchased from Aldrich (anhydrous) and used without further purification. Yields refer to homogeneous materials obtained after purification of reaction products by flash column chromatography using silica gel (200–400 mesh) or recrystallization. Analytical thin-layer chromatography was performed on silica gel 60 and GF (5–40 µm thickness) plates, and the plates were treated with a basic potassium permanganate stain or ninhydrin solution, heated and visualized under UV light. Melting points were measured with a Buchi M-565 instrument and are uncorrected. ¹H and proton-decoupled ¹³C NMR spectra were acquired in CDCl₃, CD₃OD or *d*₆-DMSO at 250 MHz (¹H) and 62.5 MHz (¹³C) (Bruker DPX250), 400 MHz (¹H) and 100 MHz (¹³C) (Bruker AVANCE 400), 500 MHz (¹H) and 125 MHz (¹³C) (Varian Inova 500), or 600 MHz (¹H) and 150 MHz (¹³C) (Bruker AVANCE 600). Chemical shifts (δ) are reported in ppm using residual undeuterated solvent as an internal standard (CDCl₃ at 7.26 ppm, CD₃OD at 3.31 ppm, *d*₆-DMSO at 2.50 ppm, and TMS at 0.00 ppm for ¹H NMR spectra and CDCl₃ at 77.16 ppm, CD₃OD at 49.0 ppm, *d*₆-DMSO at 39.52 ppm for ¹³C NMR spectra). Multiplicity data are reported as follows: s = singlet, d = doublet, t = triplet, q = quartet, br s = broad singlet, dd = doublet of doublets, dt = doublet of triplets, app d = apparent doublet, app t = apparent triplet, m = multiplet, and br m = broad multiplet. The multiplicity is followed by the coupling constant(s) in Hz and integration. High-resolution mass spectrometry (HRMS) was measured using electrospray ionization (ESI) (Waters xevo Q-tof, Thermo LTQ-FT ultra, or Thermo Q Exactive) or using electron ionization (EI) (GCT Premier Waters). The synthesis and characterization of compounds 1, 17, 18, 31–63 were previously reported (Ferreira et al., 2014).

RESULTS AND DISCUSSION

Synthesis of Novel Benzimidazole Derivatives

Phenoxyacetic acids of type I were prepared from the corresponding substituted phenols by nucleophilic substitution

with 2-bromoacetic acid or nucleophilic substitution with alkyl 2-bromoacetic ester, followed by ester hydrolysis (**Scheme 1A**). A subsequent reaction of activated carboxylic acid I with amine II led to the formation of amides 1–4, 6, 8, 10–12, 14, and 16–18. Alcohols 5 and 7 and aniline 9 were prepared by reduction reactions of imides 4 and 6 with sodium borohydride and nitrobenzene derivative 8 using hydrogenation under Pd/C catalysis. Carboxylic acid derivatives 13 and 15 were synthesized by hydrolysis under basic conditions of methyl esters 12 and 14, respectively. *N*-alkylated compounds 19–29 were synthesized by *N*-alkylation of the benzimidazole moiety of compounds 1, 17, and 18 with different electrophiles. *N*-Phenyl derivative 30 was prepared as described in **Scheme 1B** by an amidation reaction followed by cyclization and dehydration.

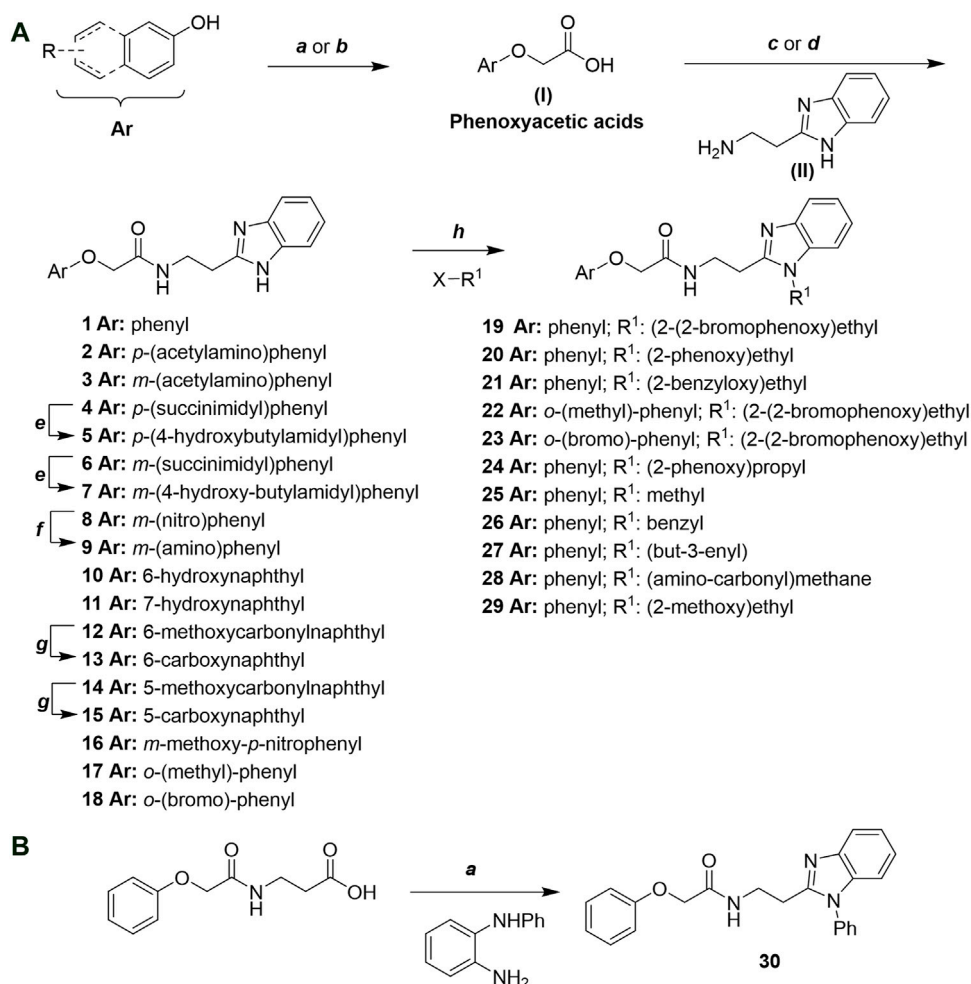
Design of Novel Cruzain Inhibitors

In this work, we designed a series of cruzain inhibitors based on a previously identified benzimidazole derivative (18, **Figure 1A**) (Ferreira et al., 2010; Ferreira et al., 2014). Considering the lead-like profile of compound 18 and its activity against cruzain and *T. cruzi*, we selected this compound for a lead optimization program and, for the first time, the pharmacokinetics and the *in vivo* trypanocidal ability of this molecule and its analogs were investigated. We explored compound 18 by appending diverse substituents at the phenyl and benzimidazole rings to improve both the interaction with cruzain and the PK profile. By adding substituents at the phenyl ring, we aimed to enhance the selectivity for cruzain over other proteases by promoting hydrogen bonding with Glu208, a critical residue located in the S2 subsite of the active site (**Figure 1B**). Glu208 is absent in most other proteases, including human cathepsins. We additionally focused on increasing the affinity and potency of the compounds by exploring *N*-substitutions at the benzimidazole and enabling additional interactions with the S1 and S1' subsites.

Exploring the Benzimidazole and Phenyl Rings

The structure and activity against cruzain of *N*-substituted benzimidazoles are summarized in **Table 1**. Three out of the derivatives that were initially evaluated showed IC₅₀ values below 3 µM. Only compounds lacking the *o*-bromine at the substituent appended to the benzimidazole core were active against cruzain. No significant variation in the percent inhibition values was observed for different Triton X-100 concentrations (0, 0.01, and 0.1%), demonstrating that the inhibitors do not act as aggregators (**Supplementary Table S1**).

The mechanisms of action of compounds 20 and 24 were determined by measuring their remaining enzymatic activity in the presence of distinct concentrations of the substrate and inhibitors. Double reciprocal Lineweaver-Burk plots (**Figure 2**) showed that unlike the benzimidazole analogs previously described (Ferreira et al., 2014), compounds 20 and 24 act as noncompetitive cruzain inhibitors with a higher affinity for the free enzyme than for the corresponding enzyme-substrate complex. The typical behavior of noncompetitive inhibitors



Scheme 1 | (A) Reagents and conditions: (a) i) ethyl 2-bromoacetate, K₂CO₃, DMF, r.t., 4–6 h; ii) NaOH (6 mol. L⁻¹), MeOH, r.t., 30 min; iii) HCl (6 mol. L⁻¹), 0°C, 10 min; (b) i) benzyl 2-bromoacetate, K₂CO₃, DMF, r.t., 4–6 h; ii) Pd/C (20%), H₂(g), EtOAc, MeOH, r.t., 1–2 h; (c) i) oxalyl chloride, DMF, DCM, r.t., 1 h; ii) *N*-Hydroxysuccinimide, DCM, triethylamine, 0°C, 30 min; iii) II, sodium carbonate, EtOAc, r.t., 1 h; (d) II, EDC, HOBt, trimethylamine, DMF, r.t., 8–15 h; (e) sodium borohydride, MeOH, THF, r.t., 5 h; (f) Pd/C (20%), H₂(g), MeOH, r.t., 2 h; (g) i) NaOH (6 mol. L⁻¹), MeOH, r.t., 20 min; ii) HCl (6 mol. L⁻¹), 0°C, 10 min; (h) haloalkyls, 18-crown-6, potassium *tert*-butoxide, THF, r.t. or 45°C, 13–48 h. **(B) Reagents and conditions:** (a) i) oxalyl chloride, DMF, DCM, r.t., 30 min, ii) *N*¹-phenylbenzene-1,2-diamine, *n*-butanol, 110°C, 18 h.

was additionally confirmed in another experiment, in which no significant variation in IC₅₀ values was observed with increasing substrate concentrations at a constant protein concentration (Supplementary Table S2).

Next, novel compounds were synthesized to evaluate the effect of growing the *N*-substituent on the mechanism of action against cruzain. Compound **1** (IC₅₀ = 10.9 μM, Table 1), which, in contrast with lead compound **18**, lacks the *o*-bromine at the phenyl ring, is more than 10-fold less potent than **18** (IC₅₀ = 0.8 μM). Installing a methyl group as the *N*-substituent also resulted in a decrease in activity (**25**, IC₅₀ = 8.6 μM). As shown in Figure 3, compounds **1** and **25** act as competitive inhibitors. Growing the *N*-substituent to a benzyl (**26**, IC₅₀ = 1.1 μM) enhanced the activity; however, expanding to but-3-enyl (**27**, IC₅₀ = 13.7 μM) significantly reduced the activity. Interestingly, in contrast with compounds **1** and **25**, compounds **26** and **27** act as

noncompetitive inhibitors (Figure 3). No significant variation in the IC₅₀ values of **26** and **27** was observed with increasing substrate and constant protein concentrations, further corroborating the noncompetitive inhibition mechanism (Supplementary Table S2).

These results clearly highlight the role played by the *N*-substituents in the mechanism of cruzain inhibition. The importance of the amine was previously demonstrated by replacing the nitrogen with an oxygen atom, and the activity was lost (Ferreira et al., 2014). Additionally, the distinct *N*-substituents allowed us to correlate the substituent volume with the mechanism of inhibition. The lack of a substituent (**1**) or the presence of a methyl (**25**) results in competitive inhibition, while bulkier groups such as benzyl (**26**) and but-3-enyl (**27**) lead to noncompetitive inhibition.

In the next step, we explored substitutions at the phenyl ring. Given that the phenyl ring of lead compound **18** occupies the S2 pocket of the cruzain-binding site, we expanded the

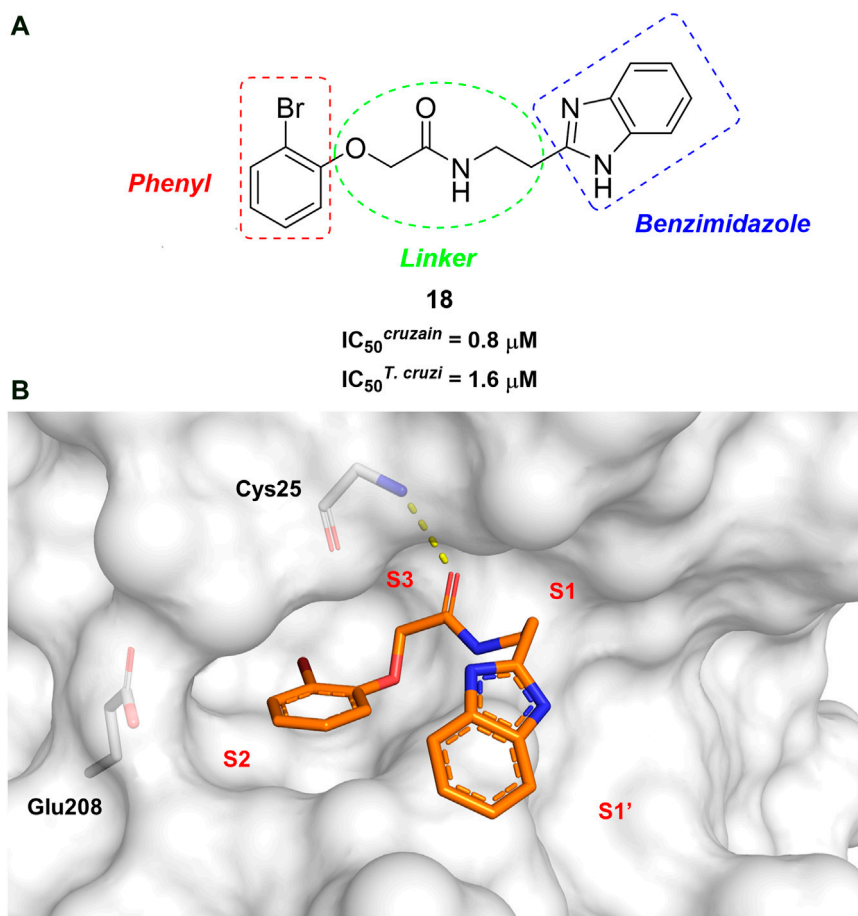


FIGURE 1 | (A) Cruzain inhibitor **18** was used as the lead compound for the design of novel benzimidazole derivatives. **(B)** X-ray structure of compound **18** in complex with cruzain (PDB 3KKU, 1.28 Å). Binding site residues (carbon in gray) and compound **18** (carbon in orange) are shown as sticks. A hydrogen bond is shown as a dashed line. Cruzain subsites are labeled as S1, S1', S2, and S3.

phenyl into a naphthyl system and appended different hydrogen bond donors and acceptors to the phenyl ring. The goal was to explore a potential interaction with Glu208. As shown in **Table 2**, among the 15 synthesized compounds, the three naphthyl analogs with either hydroxyl or ether at the *meta* or *para* positions were the most potent: 6-hydroxynaphthyl **10** ($IC_{50} = 3.4 \mu M$), 7-hydroxynaphthyl **11** ($IC_{50} = 2.7 \mu M$), and 6-methoxycarbonyl **12** ($IC_{50} = 2.3 \mu M$). The design concept was corroborated by molecular docking runs, which predicted the formation of a hydrogen bond between the hydroxyl groups of **10** and **11** and Glu208 (**Figure 4**). To further corroborate the formation of a hydrogen bond with Glu208, we evaluated the activities of six compounds against the enzyme rhodesain, a cysteine protease that has a similar active site to that of cruzain, in which Glu208 is replaced with an alanine residue (Lima et al., 2013). The activities of the compounds against cruzain were significantly more pronounced than their activities against rhodesain, indicating the importance of interactions with Glu208 for inhibition by **10** and **11** (**Supplementary Table S3**).

Trypanocidal Activity, Physicochemical Profile, and Cytotoxicity

After the enzyme inhibition studies, active compounds were evaluated for their activity against *T. cruzi* intracellular amastigotes and PK properties (**Table 3**). Among the *N*-substituted analogs, compounds **20** ($IC_{50} = 2.04 \mu M$) and **24** ($IC_{50} = 1.43 \mu M$) were equipotent to the reference drug BZ ($IC_{50} = 1.45 \mu M$). The only inactive compound in this series was the *N*-methyl analog **25**. In general, these compounds are more lipophilic than BZ, as shown by the $LogP$ and $eLogD$ values. Among the molecules in **Table 3**, six were classified as high-permeability compounds (PAMPA higher than 1.5×10^{-6} cm/s), and nine were classified as low-permeability compounds (PAMPA lower than 1.5×10^{-6} cm/s).

Most compounds with substituents on the phenyl ring were active against *T. cruzi*, with IC_{50} values in the low micromolar range (**Table 3**). The exception was compound **13**, which has a 6-carboxynaphthyl moiety. Compound **13** showed moderate activity against cruzain ($IC_{50} = 24.2 \mu M$) in addition to a $LogP$ value higher than those of the other analogs. The combination of these two properties may be the cause of the lack of trypanocidal activity of this compound.

TABLE 1 | Structure and activity against cruzain of new *N*-substituted benzimidazole derivatives.^a

| Compound | Structure | % Cruzain inhibition (100 μ M) ^a | IC ₅₀ (μ M) ^b |
|----------|-----------|---|--|
| 1 | | 90 | 10.9 \pm 1.0 |
| 19 | | 72 | ND |
| 20 | | 92 | 1.04 \pm 0.7 |
| 21 | | 76 | 1.69 \pm 0.4 |
| 22 | | 36 | ND |
| 23 | | 49 | ND |

(Continued in next column)

TABLE 1 | (Continued) Structure and activity against cruzain of new *N*-substituted benzimidazole derivatives.^a

| Compound | Structure | % Cruzain inhibition (100 μ M) ^a | IC ₅₀ (μ M) ^b |
|----------|-----------|---|--|
| 24 | | 81 | 2.2 \pm 1.2 |
| 25 | | 87 | 8.6 \pm 1.7 |
| 26 | | 77 | 1.1 \pm 0.2 |
| 27 | | 79 | 13.7 \pm 1.4 |
| 28 | | 81 | 12.1 \pm 2.4 |
| 29 | | 79 | 8.8 \pm 1.8 |
| 30 | | 79 | 8.6 \pm 2.6 |

^aThe percentage of inhibition refers to the mean of three experimental measures.^bIC₅₀ values were determined independently in triplicate using at least six distinct inhibitor concentrations, and the values represent the mean \pm SD of 2–3 independent assays.

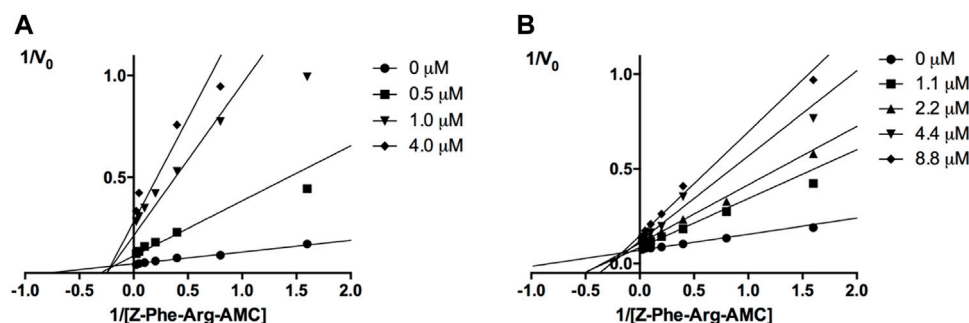


FIGURE 2 | Lineweaver-Burk plots for compounds **20** (A) and **24** (B). Each curve represents a different inhibitor concentration.

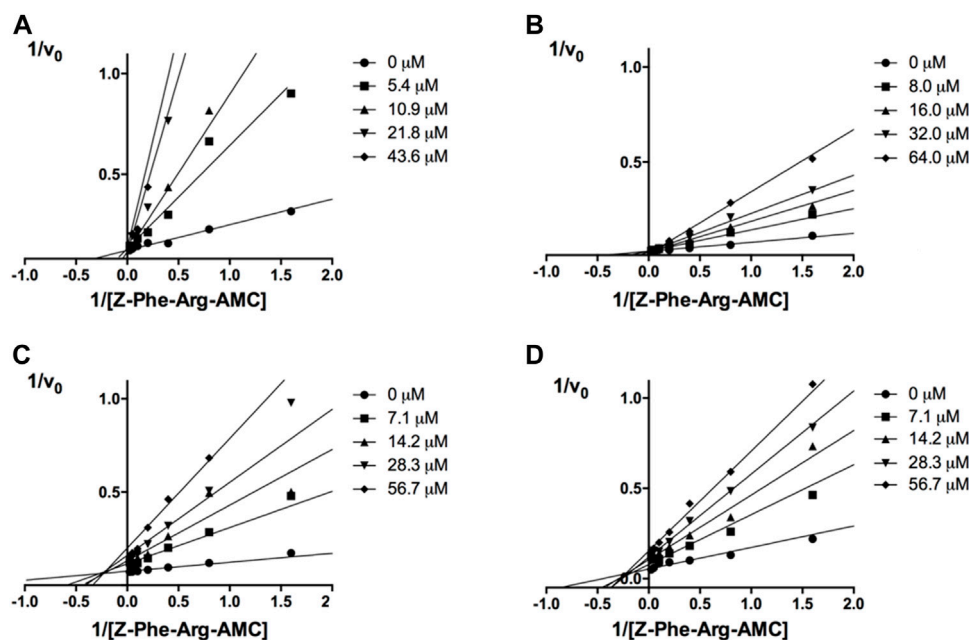


FIGURE 3 | Lineweaver-Burk plots for compounds **1** (A); **25** (B); **27** (C); and **26** (D). Each curve represents a different inhibitor concentration.

The benzimidazole derivatives were further evaluated regarding their cytotoxicity against human HFF-1 fibroblasts, which were used as host cells for *T. cruzi* (Table 4). Selectivity indices (SI), which express the ratio between the IC_{50} values for HFF-1 cells and *T. cruzi*, were calculated. Overall, the evaluated compounds exhibited no significant toxicity against human HFF-1 fibroblasts. Three compounds showed SI values comparable to or greater than that of the reference drug BZ (SI > 33): **18** (SI > 61), **17** (SI > 35), and **37** (SI > 34). It is worth noting that compounds **1** (SI > 26) and **8** (SI > 29) also exhibited suitable SI values.

Determination of *In Vitro* and *In Vivo* Metabolic Stability

A series of 10 benzimidazole derivatives were selected based on their activity against cruzain and *T. cruzi* to undergo PK studies, including

in vitro and *in vivo* metabolism. Table 5 shows the *in vitro* results for CL_{int} after incubation with human and mouse microsomes, *fu*, LogD, and PAMPA. Corrected clearance values (CL_{int_u}) were obtained by calculating the ratio between CL_{int} and *fu*. It is important to note that only unbound drug molecules are available for clearance, interaction with metabolizing enzymes and transporters, equilibration into tissues, and pharmacological activity. Thus, PK, pharmacodynamics, and toxicity are driven by unbound drug concentrations (Zamek-Gliszczyński, et al., 2011). As such, protein binding (PPB) in plasma, microsomes, and target tissues is routinely evaluated in drug discovery to determine the respective *fu* values (Wang, et al., 2014). The drug-like space for unbound clearance lies at approximately 10 L/h/kg. As shown in Table 5, all benzimidazole derivatives have CL_{int_u} values much higher than the drug-like reference and that of the reference drug BZ. Compounds **17**, **18**, **37**, which have $IC_{50}^{T. cruzi}$ values comparable to that of BZ, have CL_{int_u} values ranging from ~10

TABLE 2 | Structure and activity against cruzain of new benzimidazoles with substituents at the phenyl ring.^a

| Compound | Structure | % Cruzain inhibition (100 μ M) ^a | IC ₅₀ (μ M) ^b |
|----------|-----------|---|--|
| 2 | | 79 | 4.5 \pm 0.5 |
| 3 | | 70 | 28.1 \pm 3.1 |
| 4 | | 20 | ND |
| 5 | | 87 | ND |
| 6 | | 38 | ND |
| 7 | | 75 | ND |
| 8 | | 83 | 13.5 \pm 2.6 |
| 9 | | 90 | 18.2 \pm 1.8 |
| 10 | | 90 | 3.4 \pm 0.9 |
| 11 | | 96 | 2.7 \pm 0.7 |
| 12 | | 100 | 2.3 \pm 0.6 |
| 13 | | 89 | 24.2 \pm 4.5 |

(Continued in next column)

TABLE 2 | (Continued) Structure and activity against cruzain of new benzimidazoles with substituents at the phenyl ring.^a

| Compound | Structure | % Cruzain inhibition (100 μ M) ^a | IC ₅₀ (μ M) ^b |
|----------|-----------|---|--|
| 14 | | 92 | 8.3 \pm 2.1 |
| 15 | | 58 | > 100 |
| 16 | | 62 | ND |

^aThe percentage of inhibition refers to the mean of three experimental measures.^bIC₅₀ values were determined independently in triplicate using at least six distinct inhibitor concentrations, and the values represent the mean \pm SD of 2–3 independent assays.

to 21 times higher than that of BZ, which could undermine the achievement of the bioavailability levels required for biological response. Most compounds listed in **Table 5** had PAMPA values higher than 1.5×10^{-6} cm/s and were classified as having good permeability.

Next, the same set of molecules was evaluated for their *in vivo* PK profile (**Table 6**). From this assay, information such as $T_{1/2}$, plasma clearance (CL_p), and bioavailability (F) were obtained. As observed for the *in vitro* assays, all benzimidazoles had high unbound clearance compared to that of BZ. The *in vivo* assays reinforced the concept that the high clearance may be the reason for the very low oral bioavailability (F) observed for the benzimidazoles (0–35%) compared to that of BZ (90%).

In vitro experiments are faster and less expensive than *in vivo* assays. Accessing the *in vitro-in vivo* correlation (IVIVC) for metabolic stability is important to demonstrate whether one can rely on *in vitro* studies and keep the use of animals to a minimum for a series of molecules. The lack of IVIVC is also informative, indicating that other metabolic routes are likely to be responsible for the observed *in vivo* clearance. In our experiments, a positive IVIVC was observed for *fu*-corrected clearance (**Figure 5**), which allowed us to rely on *in vitro* assays for the prediction of *in vivo* metabolic stability and prioritize compounds for further studies.

Determination of Metabolic Stability in Human Hepatocytes and Identification of CYP450 Isoforms

All compounds evaluated showed high clearance values, reaching 60–180% of mouse liver blood flow (5.4 L/h/kg), which is a plausible explanation for their low bioavailability (0–35%). To better understand pathways involved in the elimination of the compounds, metabolic stability studies in human hepatocytes,

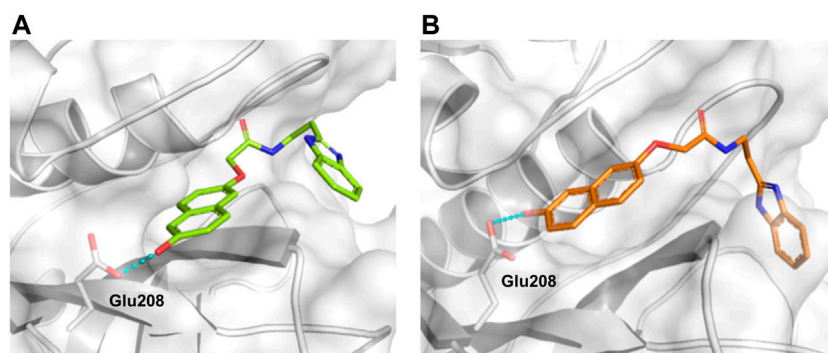


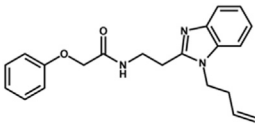
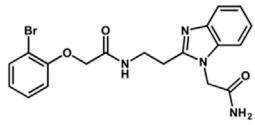
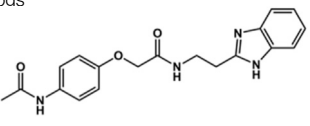
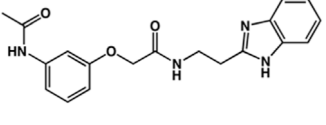
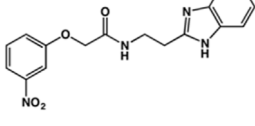
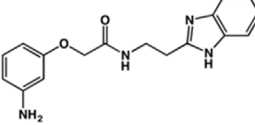
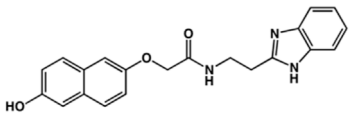
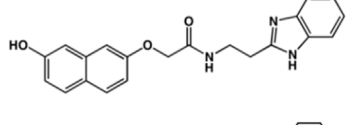
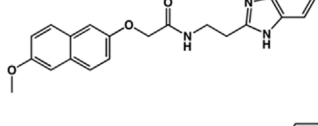
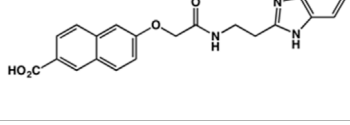
FIGURE 4 | Molecular docking predicted the binding conformations of compounds **10** (A) and **11** (B) in complex with cruzain (PDB 3KKU, 1.28 Å), showing the formation of hydrogen bonds (dashed lines) between the hydroxyl groups and Glu208. Binding site residues (carbon in gray) and compounds **10** and **11** (carbon in green and orange, respectively) are shown as sticks.

TABLE 3 | *In vitro* activity against *T. cruzi* and physicochemical properties of a subset of the benzimidazoles.

| Compound | Structure | IC ₅₀ ^{<i>T. cruzi</i>} (μM) ^a | PAMPA (×10 ⁻⁶ cm/s) | eLogD | LogP | PSA (Å ²) |
|-----------------------|-----------|--|--------------------------------|-------|------|-----------------------|
| BZ ^a | | 1.45 ± 0.4 | 3.17 | 0.84 | 1.00 | 92.70 |
| N-substituted Analogs | | | | | | |
| 1 | | 3.9 ± 0.3 | 4.21 | 2.90 | 2.41 | 67.00 |
| 20 | | 2.04 ± 0.6 | 4.30 | 4.24 | 1.57 | 104.38 |
| 24 | | 1.43 ± 0.4 | 2.29 | 4.45 | 4.31 | 65.38 |
| 25 | | ≅ 100 | 1.46 | 2.93 | 2.62 | 56.15 |
| 26 | | 7.4 ± 2 | 8.72 | 4.06 | 4.20 | 56.15 |

(Continued on following page)

TABLE 3 | (Continued) *In vitro* activity against *T. cruzi* and physicochemical properties of a subset of the benzimidazoles.

| Compound | Structure | IC ₅₀ ^{<i>T. cruzi</i>} (μM) ^a | PAMPA (×10 ⁻⁶ cm/s) | eLogD | LogP | PSA (Å ²) |
|----------------------------|---|--|--------------------------------|-------|------|-----------------------|
| 27 |  | 6.9 ± 2.2 | 9.06 | 3.75 | 3.56 | 56.15 |
| 28 |  | 6.8 ± 0.9 | 0.71 | 2.16 | 1.47 | 99.24 |
| Phenyl-substituted Analogs | | | | | | |
| 2 |  | 4.5 ± 0.6 | 0.21 | 1.92 | 1.53 | 96.10 |
| 3 |  | 5.0 ± 1.0 | 0.17 | 2.26 | 1.53 | 96.10 |
| 8 |  | 3.5 ± 0.7 | 2.17 | 2.98 | 2.31 | 112.83 |
| 9 |  | 5.4 ± 0.9 | 0.38 | 1.88 | 1.67 | 93.03 |
| 10 |  | 14.6 ± 0.7 | 0.21 | 3.17 | 3.08 | 87.24 |
| 11 |  | 2.8 ± 0.6 | 0.17 | 2.26 | 3.08 | 87.24 |
| 12 |  | 16.6 ± 2.4 | 0.34 | 3.89 | 3.18 | 93.31 |
| 13 |  | >50 | 0.83 | ND | 3.31 | 76.24 |

^aIC₅₀ values represent the mean ± SD of three independent assays; BZ, benzimidazole. eLogD and PAMPA were experimentally determined. LogP and PSA were predicted computationally.

TABLE 4 | Biological evaluation of a subset of the benzimidazoles against *T. cruzi* and human HFF-1 fibroblasts.

| Compound | IC ₅₀ ^{T. cruzi} (μM) ^a | IC ₅₀ ^{HFF-1} (μM) ^b | SI ^c |
|-------------|--|---|-----------------|
| BZ | 3.00 ± 0.60 | >100 | >33 |
| Doxorubicin | — | 0.26 | — |
| 1 | 3.9 ± 0.3 | >100 | >26 |
| 2 | 12.1 ± 1.3 | >100 | >8 |
| 3 | 5.0 ± 1.0 | >100 | >20 |
| 5 | ~ 50.00 | >100 | >2 |
| 8 | 3.5 ± 0.7 | >100 | >29 |
| 10 | 14.6 ± 0.7 | >100 | >7 |
| 17 | 2.81 ± 0.75 | >100 | >35 |
| 18 | 1.63 ± 0.57 | >100 | >61 |
| 20 | 2.04 ± 0.60 | >30 | >14 |
| 24 | 1.43 ± 0.40 | >30 | >21 |
| 26 | 7.40 ± 2.00 | >100 | >13 |
| 27 | 6.90 ± 2.20 | >100 | >14 |
| 28 | 6.8 ± 0.9 | >100 | >15 |
| 32 | 7.90 ± 2.13 | >100 | >12 |
| 33 | 6.68 ± 2.35 | >100 | >15 |
| 34 | 16.22 ± 3.51 | >30 | >2 |
| 35 | 46.12 ± 6.21 | >100 | > 2 |
| 37 | 2.90 ± 0.66 | > 100 | >34 |
| 38 | 11.14 ± 3.19 | >100 | >9 |

^aIC₅₀ values represent the mean ± SD of three independent assays.^bIC₅₀ values represent the mean ± SD of two independent assays.^cSelectivity index (SI) = IC₅₀^{HFF-1}/IC₅₀^{T. cruzi}.

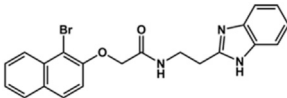
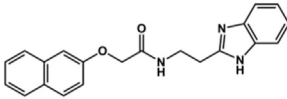
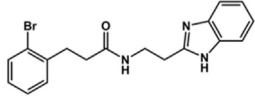
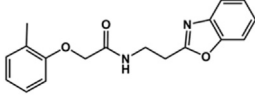
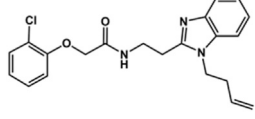
which contain the whole set of human phase I and phase II hepatic metabolizing enzymes, were conducted. This experiment was performed in the absence of CYP inhibitors for the determination of the total clearance (Phase I + Phase II); in the presence of 1-ABT (a CYP450 inhibitor) for the determination of the fraction of the compounds that are metabolized by the remaining Phase I as well as the conjugating Phase II enzymes; and in the presence of azamulin (a CYP3A4 inhibitor) for the identification of the fraction metabolized by this isoform. This last assay provides critical information for prioritizing compounds since CYP3A4 plays a key role in drug-drug interactions (DDIs) and is associated with adverse effects and low efficacy when two or more drugs are taken together. Phase I metabolism, performed mainly by the CYP450 family, was responsible for 56–95% of the metabolism of the compounds (mean = 78.1 ± 12.9%). A much lower contribution to total clearance was observed for all other Phase I enzymes, which were responsible for 5–35% of the metabolism (mean = 21.9 ± 13.6%). The central role played by CYP3A4 became apparent when the CYP3A4 inhibitor azamulin was used in the assay: the resulting clearance values were approximately 40% lower, reaching a minimum value of 31.6% and a maximum value of 59.5% when compared with total clearance values (**Supplementary Figure S1**).

TABLE 5 | *In vitro* PK profile of benzimidazole analogs.

| Compound | Structure | CL _{int} (L/h/kg) human | CL _{int} (L/h/kg) mouse | fu | CL _{int,u} (L/h/kg) human | CL _{int,u} (L/h/kg) mouse | eLogD | PAMPA (×10 ⁻⁶ cm/s) |
|-----------|-----------|-------------------------------------|-------------------------------------|-----|---------------------------------------|---------------------------------------|-------|-----------------------------------|
| BZ | | 1.5 | 4.4 | 1.0 | 1.5 | 4.4 | 0.8 | 3.2 |
| 17 | | 23.9 | 300.0 | 0.8 | 31.2 | 392.2 | 3.6 | 5.9 |
| 18 | | 9.2 | 191.0 | 0.6 | 16.0 | 334.5 | 3.9 | 5.6 |
| 31 | | 5.6 | 74.6 | 0.9 | 6.6 | 87.9 | 2.6 | 2.2 |
| 32 | | 25.5 | 526.0 | 0.7 | 39.2 | 808.0 | 3.6 | 4.9 |
| 33 | | 10.4 | 161.0 | 0.8 | 12.5 | 193.3 | 2.8 | 0.7 |

(Continued on following page)

TABLE 5 | (Continued) *In vitro* PK profile of benzimidazole analogs.

| Compound | Structure | CL _{int} (L/h/kg) human | CL _{int} (L/h/kg) mouse | fu | CL _{int,u} (L/h/kg) human | CL _{int,u} (L/h/kg) mouse | eLogD | PAMPA ($\times 10^{-6}$ cm/s) |
|----------|---|-------------------------------------|-------------------------------------|-----|---------------------------------------|---------------------------------------|-------|-----------------------------------|
| 34 |  | 49.9 | 607.0 | 0.1 | 539.5 | 6562.2 | 4.4 | 0.3 |
| 35 |  | 21.1 | 565.0 | 0.5 | 43.9 | 1174.6 | 3.9 | 0.7 |
| 36 |  | 28.8 | 705.0 | 0.8 | 38.2 | 935.0 | 3.8 | 4.7 |
| 37 |  | 22.9 | 745.0 | 0.9 | 26.9 | 874.4 | 3.9 | 19.7 |
| 38 |  | 230.0 | 934.0 | 0.5 | 501.1 | 2034.9 | 4.1 | 7.7 |

CL_{int}, intrinsic clearance after incubation with human and mouse microsomes; fu, fraction unbound; CL_{int,u}, corrected clearance (CL_{int}/fu); eLogD, experimentally determined distribution coefficient; PAMPA, parallel artificial membrane permeability assay.

After identifying the central role of CYP3A4 in the metabolism of this series, the next step was to identify the involved isoforms using recombinant CYP enzymes. CL_{int} was determined based on the residual amount of the compound over time. Additionally, the contribution of each CYP isoform to metabolism was calculated based on their relative abundance in humans. The information generated by this experiment was essential to assess the risk of potential drug-drug interactions (DDIs) for this series of compounds. Molecules eliminated through multiple pathways have reduced DDI potential and are therefore more suitable for advancing to further steps in a drug discovery pipeline. The benzimidazole derivatives are mainly metabolized by CYP3A4 (23–90%) (**Supplementary Table S4**). Although modest, a contribution from isoforms CYP2D6 and CYP1A2 is observed, featuring an attractive profile from a DDI perspective despite the high clearance values.

Identification of Sites of Metabolism

All studies on benzimidazoles showed that these molecules are metabolically unstable, and biotransformation mediated by CYP3A4 is the major metabolic route. Therefore, studies to identify the molecular sites of metabolism (SOM) were performed. These studies can enable the blockage of these sites by the inclusion of blocking groups to achieve appropriate levels of metabolic stability. Ideally, these molecular changes should not significantly affect the potency toward the molecular target. The test compounds were then incubated with human and mouse liver microsomes. Most of the metabolites were found to be oxidation products mainly of the linker and benzimidazole

moieties (**Supplementary Figures S2–S5**). Strategies to block these SOMs could include switching the amide position and adding halogen atoms to the linker. At the benzimidazole ring, N substitutions and the addition of halogens could be explored (**Supplementary Figure S6**).

Metabolic Stability Studies for an Additional Set of Benzimidazole Derivatives

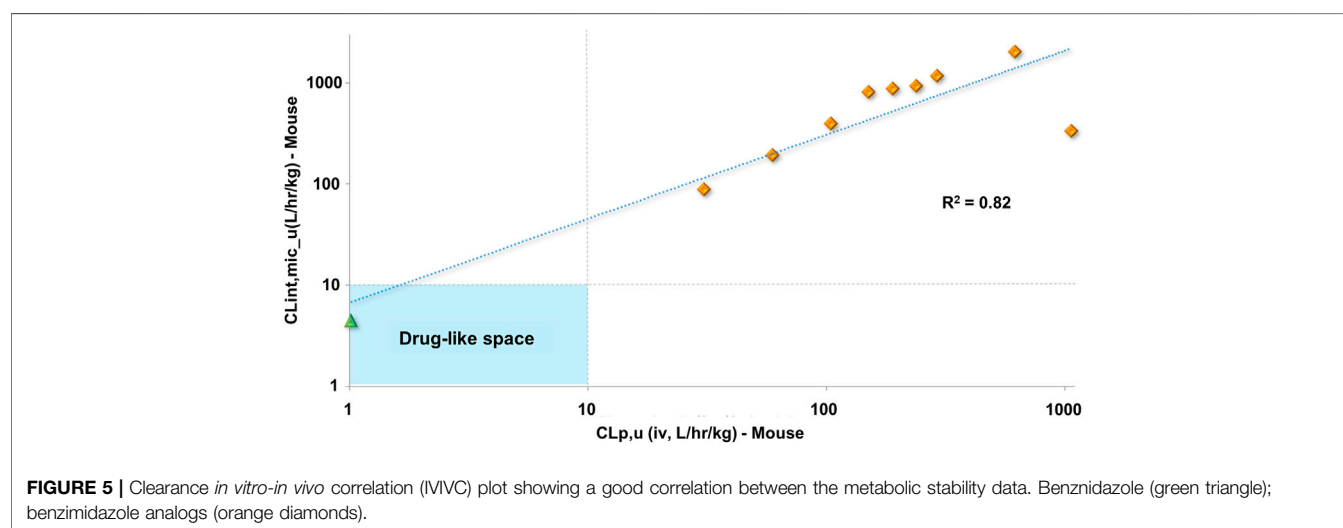
After completing the PK profile for 10 molecules (set 1), an additional set of 55 compounds (set 2) was evaluated to provide additional information for the establishment of a SAR for metabolic stability. The results for clearance after incubation with human and mouse liver microsomes, eLogD, and PAMPA are summarized in **Supplementary Table S5**. Some set 2 compounds with lower clearance values than those of set 1 molecules were identified, some of which exhibited clearance values comparable to that of BZ (**Supplementary Figure S7**).

The clearance values listed in **Supplementary Table S5** show that the presence of substituents on the phenyl ring might influence the metabolic stability of the compounds. Substituents at *para* and *meta*, for example, led to the most stable compounds with the lowest CL_{int,u} values. Among the compounds with substituents at *para*, compounds **2**, **4** and **5** are highlighted, for which clearance values are in the same range of BZ (CL_{int,u} = 1.50 L/h/kg). In addition, compounds **6** (CL_{int,u} = 3.54 L/h/kg), **7** (CL_{int,u} = 1.50 L/h/kg), and **3** (CL_{int,u} = 6.90 L/h/kg), with substituents at *meta*, showed drug-like profiles for metabolic stability. Substituents at *ortho* also led to stable

TABLE 6 | *In vivo* PK profiles of selected benzimidazole derivatives.

| Cpd | $T_{1/2}$ (h) | C_0 (ng/ml) | IV | | | | | PO | | | | |
|-----------|---------------|---------------|-------------|---------------|--------------------------|-----------------------|----------------------------|---------------|-------------------|---------------|---------------|-------|
| | | | V SS (L/kg) | AUC (ng·h/ml) | CL _p (L/h/kg) | <i>f</i> _u | CL _{p,u} (L/h/kg) | $T_{1/2}$ (h) | C_{max} (ng/ml) | T_{max} (h) | AUC (ng·h/ml) | F (%) |
| BZ | 0.8 | 961 | 1.1 | 1.020 | 1.0 | 0.99 | 1.0 | 1.5 | 404 | 0.4 | 1.040 | 90 |
| 17 | 0.5 | 153 | 4.4 | 83 | 6.3 | 0.06 | 104.3 | — | 0.0 | — | 0.0 | 0.0 |
| 18 | 0.2 | 283 | 1.8 | 84 | 3.2 | 0.03 | 103.7 | — | — | 0.3 | 1.0 | 1.2 |
| 31 | 0.2 | 232 | 1.9 | 67 | 4.6 | 0.15 | 29.9 | — | 12.8 | 0.3 | — | — |
| 32 | 0.2 | 220 | 2.5 | 53 | 5.9 | 0.04 | 157.4 | — | — | 0.3 | 1.0 | 1.9 |
| 33 | 0.2 | 658 | 0.6 | 160 | 9.5 | 0.16 | 60.6 | 0.4 | 71.7 | 0.3 | 56.5 | 35.2 |
| 34 | 0.3 | 255 | 1.9 | 116 | 7.6 | 0.00 | 2,874.7 | 0.6 | 7.6 | 0.3 | 6.5 | 5.6 |
| 35 | 0.2 | 228 | 2.0 | 86 | 5.8 | 0.02 | 334.6 | 0.2 | 3.7 | 0.3 | 1.7 | 2.0 |
| 36 | 0.5 | 144 | 7.2 | 55 | 9.5 | 0.04 | 231.5 | — | 0.0 | — | 0.0 | 0.0 |
| 37 | 0.2 | 169 | 2.4 | 53 | 9.5 | 0.05 | 187.1 | — | 14.6 | 0.3 | — | — |
| 38 | 0.4 | 126 | 3.3 | 82 | 6.2 | 0.01 | 430.5 | — | 2.5 | 0.3 | — | — |

IV, intravenous administration; PO, oral administration; $T_{1/2}$, plasma half-life; C_0 , concentration at time = 0; VSS, steady-state volume of distribution; AUC, area under the curve; CL_p, plasma clearance; *f*_u, fraction unbound; CL_{p,u}, plasma clearance corrected for the fraction unbound; C_{max} , peak plasma concentration; T_{max} , time of peak plasma concentration; F, bioavailability.



compounds: **31** ($CL_{int,u} = 6.6$ L/h/kg), **42** ($CL_{int,u} = 4.64$ L/h/kg), and **43** ($CL_{int,u} = 3.94$ L/h/kg).

Set 2 compounds did not show significant structural variability in the linker region. Among the few exceptions are compounds **48**, in which sulfur replaced the linker oxygen ($CL_{int,u} = 11.38$ L/h/kg), and **58** ($CL_{int,u} = 16.99$ L/h/kg) and **49** ($CL_{int,u} = 3.28$ L/h/kg), in which the position of the phenoxy fragment was modified by the introduction of a methylene group at the linker. At the benzimidazole ring, the introduction of a hydrophilic amide led to high metabolic stability (**28**, $CL_{int,u} = 1.53$ L/h/kg). It is important to highlight the influence exerted by the physicochemical nature of the substituents at the phenyl and the benzimidazole on the stability of the compounds (Supplementary Figure S8). The introduction of hydrophobic substituents at the phenyl resulted in high clearance values, such as those observed for compounds **44** ($CL_{int,u} = 70.62$ L/h/kg), **51** ($CL_{int,u} = 65.56$ L/h/kg), **52** ($CL_{int,u} = 2,122.45$ L/h/kg), and **60** ($CL_{int,u} = 194.48$ L/h/kg). Hydrophobic substituents at the benzimidazole (**19**, **20**, **21**, **23**, **26**, **27**, and **30**) followed the

same trend, with $CL_{int,u}$ values ranging from 70.23 to 3,366.34 L/h/kg. Overall, the clearance values for the benzimidazole derivatives increased with increasing hydrophobicity (Supplementary Figure S9). Seven set 2 compounds (**1**, **2**, **3**, **8**, **10**, **11**, and **28**) with suitable trypanocidal activity and *in vitro* clearance underwent *in vivo* PK studies. Overall, the set 2 compounds exhibited lower $CL_{p,u}$ values compared to those of the set 1 analogs (Supplementary Table S6), with benzimidazoles **2** and **28** showing the most promising profiles ($CL_{p,u}$ of 4.16 and 3.98, respectively). Additionally, similar to the profile observed for the set 1 compounds, a good correlation between *in vitro* and *in vivo* clearance was found for the set 2 benzimidazoles (Figure 6).

In Vivo Toxicity and Trypanocidal Activity

Compound **28** ($IC_{50}^{T. cruzi} = 6.8$ μ M) was selected for a proof-of-concept study given its suitable balance between pharmacodynamics and PK properties. Initially, we determined the doses that elicited no acute toxicity. The compound, solubilized in 10% DMSO aqueous solution, was

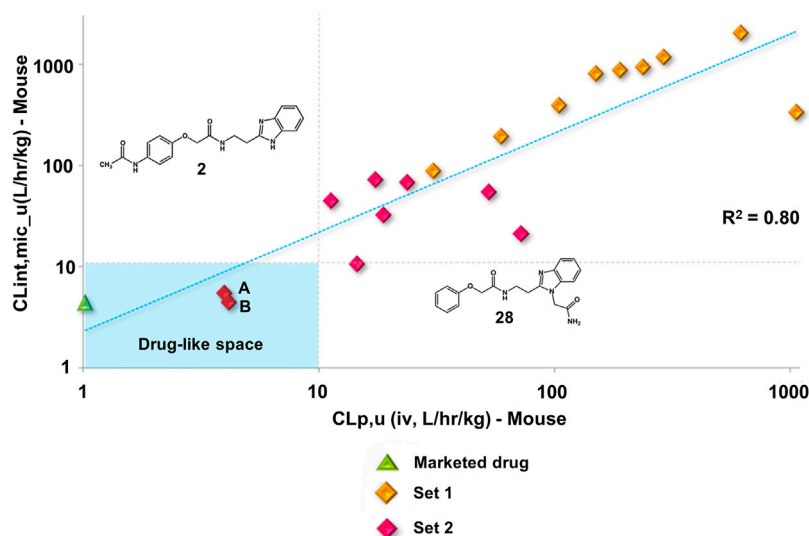


FIGURE 6 | Clearance *in vitro-in vivo* correlation (IVVC) plot showing a good correlation between the metabolic stability data. Benznidazole (green triangle); set 1 compounds (orange diamonds); set 2 compounds (red diamonds). Compounds **2** (A) and **28** (B).

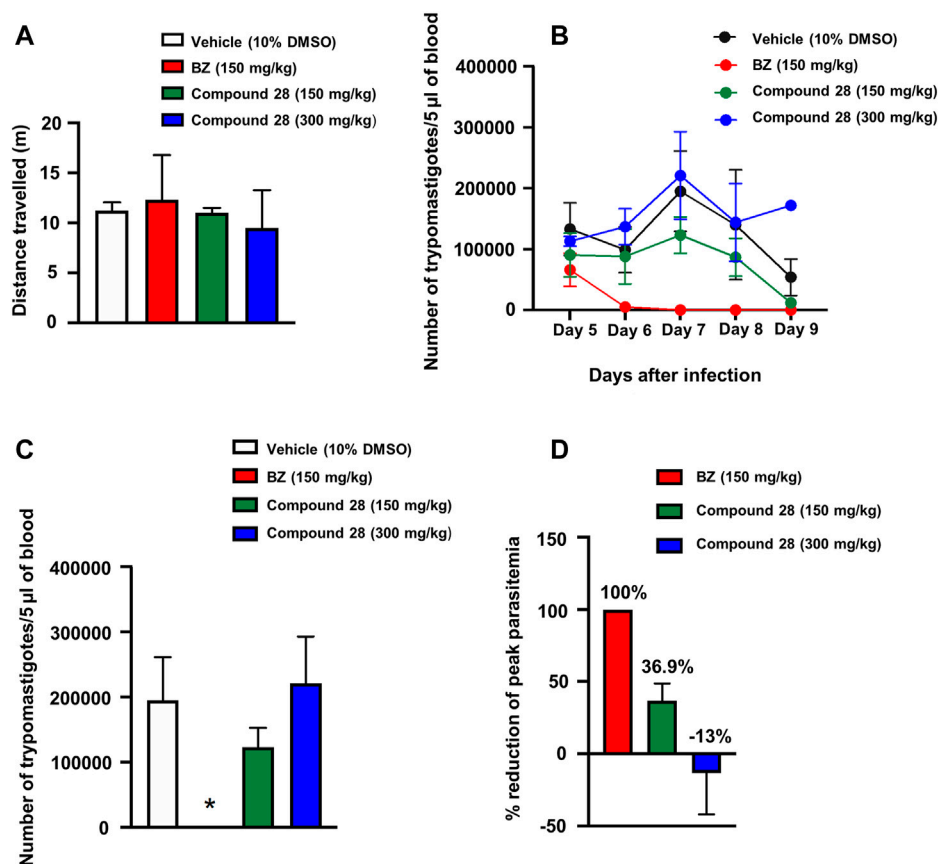


FIGURE 7 | Acute toxicity and trypanocidal activity *in vivo*. **(A)** Open-field test. Mice were orally treated with vehicle (10% DMSO) or benznidazole (BZ) at doses of 150 mg/kg or **28** at doses of 150 and 300 mg/kg. **(B)** Parasitemia during *T. cruzi* infection in mice treated with vehicle, BZ or **28** (150 and 300 mg/kg) expressed as the number of trypanastigotes per 5 µl of blood. The data represent the mean parasitemia \pm SEM (4–8 animals per group) for all assays. **(C)** Peak parasitemia expressed as the number of trypanastigotes per 5 µl of blood in mice treated with vehicle, BZ or **28** (150 and 300 mg/kg) (* $p < 0.05$ when compared to vehicle and other groups). **(D)** Reduction of peak parasitemia (seventh day of infection) in mice treated with vehicle, BZ or **28** (150 and 300 mg/kg). Vehicle solution: 0.9% NaCl + 10% DMSO.

administered orally in a single dose of 150 and 300 mg/kg of body weight in female Swiss mice. Parameters related to behavior, autonomic functions, neurological activity, and mortality were assessed as toxicity signs. Soon after the administration of the compound, mice were placed in a circular open-field arena (40 cm diameter) with 50-cm-high walls to assess motor deficits. Mortality and clinical signs associated with toxicity were recorded 0.5, 2, 4, 8, and 24 h after the single-dose administration. After this period, toxicity signs were assessed once a day for two consecutive weeks. The reference drug BZ at 150 mg/kg and vehicle (10% DMSO) were administered as controls. No toxicity signs were observed within the 2 weeks of observation for any of the tested doses. Additionally, no mortality was observed (**Figure 7A**). One-way ANOVA did not reveal any difference among the groups [$F_{3,4} = 0.15$; $p = 0.9238$], indicating that treatment with **28** at doses of 150 and 300 mg/kg did not cause locomotor deficits.

Considering the favorable acute toxicity results, the *in vivo* trypanocidal activity of **28** was determined at single doses of 150 and 300 mg/kg for 5 days. Female Swiss mice were infected with *T. cruzi* (Y strain) (Rodriguez et al., 2014) and treated via gavage with five daily doses of BZ (150 mg/kg of body weight), **28** (150 and 300 mg/kg of body weight) and vehicle (10% DMSO). Treatment started on day five after infection with *T. cruzi*. The following parameters were evaluated in these experiments: level of parasitemia after the treatment, suppression of peak parasitemia (day seven after the infection), and reduction of parasitemia on the peak day (day seven after infection). Parasitemia was expressed as the number of *T. cruzi* trypomastigotes per 5 μ l of blood and was calculated using the Brener method (Brener, 1962). Repeated ANOVA measures considering the factors treatment and day (repeated measure) showed a main effect of treatment ($F_{3,18} = 12.68$; $p < 0.05$) (**Figure 7B**). On the seventh day of treatment, when parasitemia reached its peak, one-way ANOVA indicated a treatment effect ($F_{3,16} = 7.47$; $p = 0.002$). Post hoc analyses indicated that the treatment with BZ significantly decreased parasite burden compared with the other treatments ($p < 0.05$) (**Figure 7C**). BZ and **28** (150 mg/kg) reduced peak parasitemia by 100 and 36.9%, respectively, compared with the vehicle. At a dose of 300 mg/kg, benzimidazole **28** showed an increase of 13% in the peak parasitemia when compared with the vehicle (**Figure 7D**). This dose-response effect is likely associated with the modulation of physiological systems that increase the susceptibility of the animals to infection with *T. cruzi* at high doses of the compound. Further tests with lower doses (75 mg/kg and 37.5 mg/kg) showed no reduction in parasitemia levels (**Supplementary Figure S10**). The results of the *in vivo* studies indicate a moderate ability of **28** to suppress peak parasitemia at 150 mg/kg.

Considering that few molecular targets are validated in NTDs (De Rycker et al., 2018) and the relatively unsuitable compounds regarding toxicity and drug-likeness that have been historically explored in the area, the findings reported herein address an important gap in Chagas disease drug discovery. Regardless of the mechanism of action, it is noteworthy that in rare cases a compound succeeds in terms of efficacy in Chagas disease *in vivo* infection models. This is a major hurdle in the field that can

be related to the complex life-cycle biology of *T. cruzi*, and the many poorly understood aspects of the interplay between the parasite and the host (Libisch et al., 2021). Among the compounds that reached this milestone, we can highlight vinyl sulfone K777, CYP51-inhibitor azoles (including posaconazole), and cruzain-inhibitor triazoles and carbamoyl imidazoles (Ferraz et al., 2007; McKerrow et al., 2009; Brak et al., 2010; de Souza et al., 2020). K777 was a landmark in the field as it was the first compound to show the possibility to enter clinical trials for Chagas disease. However, tolerability issues in dogs and primates during the preclinical phase hampered the progression of this compound toward clinical development. After the failure of K777, it was discussed whether the toxicity issues would be due to the irreversible mechanism of action and resulting lack of selectivity of K777 over other proteases, which could include human proteases. The case of K777 highlights the importance of designing reversible cruzain inhibitors with improved selectivity as are the benzimidazole derivatives investigated in this work. In this study, we adopted the strategy of diversifying the substitution pattern at the phenyl and benzimidazole regions. This approach led to an enhanced interaction with cruzain and, by enabling the formation of a hydrogen bond with Glu208, it improves the selectivity for cruzain over other proteases. Glu208 is part of the S2 subsite in the cruzain active site and is lacking in most other proteases such as human cathepsins. The role played by the formation of a hydrogen bond with Glu208 was investigated by evaluating a set of compounds against rhodesain, a cysteine protease that has an active site that resemble that of cruzain in which an alanine replaces Glu208. The compounds were far more active against cruzain over rhodesain, which indicates the important part played by Glu208 in selectivity toward cruzain.

Another key finding reported in Chagas disease drug design was the identification of CYP51-inhibitor antifungal azoles (Ferraz et al., 2007). These compounds, particularly Posaconazole and E1224 (the ravuconazole prodrug), showed promising suppressive effects in parasite burden in animal models of Chagas disease. However, their failure in clinical trials raised fruitful discussions regarding the mechanism of action of the compounds. Although these azoles displayed a remarkable suppressive effect, they failed in providing sustained parasite clearance when opposed to benzimidazole. These studies served to establish the landmark that *T. cruzi* CYP51 is not a molecular target to be pursued in Chagas disease drug discovery. These previous findings demonstrate the critical importance of target validation and identification of compounds that act by different modes of action, for example, the modulation of cruzain. The compounds studied herein showed a moderate reduction of parasite burden and, therefore, open novel possibilities for future work on this molecular target. Moreover, the benzimidazoles did not demonstrate toxicity in animal studies, which, as seen in the K777 case, can be an issue of cysteine protease inhibitors. The best compound (**28**) administered orally to mice in a single dose of 150 and 300 mg/kg showed no toxicity signs for any of the doses and, importantly, no mortality was observed.

Another important class of compounds is triazole-based cruzain inhibitors, whose representative analogs showed promising *in vivo* efficacy (Brak et al., 2010; Neitz et al.,

2015). These non-peptidic ketones irreversibly inactivate cruzain by attaching covalently to Cys25, which can raise selectivity issues and, therefore, be a drawback for further development. Regarding the PK profile, optimization of these triazoles resulted in enhanced bioavailability and exposure after oral dosing, although they proved to inhibit CYP3A4, the most important CYP isoform for the elimination of xenobiotics. The best compound identified herein (**28**), showed a suitable tradeoff among pharmacodynamics and PK properties. Regarding its mechanism of action, inhibitor **28** is a reversible inhibitor and interacts with Glu208, which reduces the probability of inhibition of human proteases. Additionally, the extensive PK studies enabled the identification of permeable, metabolically stable, and bioavailable compounds with high selectivity indices, and that are metabolized mainly by CYP3A4. Incubation of the compounds with isolated recombinant CYPs using CYP3A4 and pan-CYP inhibitors as controls showed that the benzimidazoles do not inhibit CYP3A4. These findings are pivotal in the context of drug-drug interactions, particularly in the case of chagasic patients who need to use different drugs to mitigate the complications of the disease.

CONCLUSION

An MPO strategy for the optimization of benzimidazole derivatives as antichagasic agents was developed. This strategy relied on the parallel optimization of activity against cruzain and *T. cruzi*, selectivity, and PK parameters such as metabolic stability and permeability. New compounds were synthesized, and previously synthesized analogs were thoroughly evaluated for PK properties. Newly introduced *N*-substituents at the benzimidazole ring revealed that increasing bulkiness at this site modifies the mechanism of action toward cruzain from competitive to noncompetitive. These results introduce new and interesting aspects regarding the binding mode and mechanism of action of cruzain inhibitors. Newly designed phenyl-substituted analogs showed increased inhibition of cruzain over rhodesain, demonstrating the key role played by Glu208 in the selective inhibition of cruzain over other proteases. Some of the benzimidazole derivatives showed appropriate metabolic stability and clearance values comparable to those of drug-like molecules. Phase I oxidation reactions catalyzed by CYP3A4 were detected as the main elimination pathway, and the identified sites of metabolism provided insights into the improvement of metabolic stability. Moreover, the analysis of the *in vitro* trypanocidal and cytotoxicity data revealed a sound selectivity index for the investigated compounds, indicating a low potential for toxicity.

The applied MPO approach enabled the prioritization of compounds considering an appropriate combination of *in vitro* activity, toxicity, and PK properties. The gathered *in vitro* data supported *in vivo* PK studies for representative compounds. A solid IVIVC was obtained, demonstrating the high predictive ability of the *in vitro* PK models for the corresponding *in vivo* endpoints. Finally, acute toxicity and efficacy studies were conducted for compound **28**, which showed no toxicity signs and a moderate reduction in peak parasitemia at 150 mg/kg.

Importantly, the knowledge gathered in this study opens novel opportunities to understand the molecular aspects of cruzain inhibition, enabling the discovery of compounds with a good trade-off between pharmacodynamics and pharmacokinetics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The animal study was reviewed and approved by The Committee of Ethics in Research of the Universidade Federal de São Paulo (CEUA No. 5301080816) and AbbVie DMPK Department.

AUTHOR CONTRIBUTIONS

IP conceptualization, writing, molecular design and biological evaluation; CORJ: conceptualization, writing, molecular design and organic synthesis; BS: organic synthesis; MD: organic synthesis; MS: molecular design and *in vitro* experiments; LF: writing, molecular design and *in vitro* experiments; ALMA: molecular design and *in vitro* experiments; RF: molecular design and *in vitro* experiments; LM, RK, and SM-D: *in vitro* experiments; FS: *in vivo* experiments; RD: *in vivo* experiments; FC: *in vivo* experiments; LD: conceptualization, supervision, organic synthesis and writing; and ADA: conceptualization, supervision, molecular design, fund acquisition and writing.

FUNDING

The National Council for Scientific and Technological Development (CNPq), Brazil, the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil, the Sao Paulo Research Foundation (FAPESP), Brazil (CEPID-CIBFar grant 13/07600-3, IP grants 11/13789-6 and 14/26324-0) for financial support and post-doctoral fellowship for CORJ, grant 158926/2014-5.

ACKNOWLEDGMENTS

The authors acknowledge the Abbvie DMPK Department for the support in the DMPK experiments and Luiz Severino da Silva for the support in the *in vivo* experiments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2021.774069/full#supplementary-material>

REFERENCES

- Alelyunas, Y. W., Pelosi-Kilby, L., Turcotte, P., Kary, M. B., and Spreen, R. C. (2010). A High Throughput Dried DMSO LogD Lipophilicity Measurement Based on 96-Well Shake-Flask and Atmospheric Pressure Photoionization Mass Spectrometry Detection. *J. Chromatogr. A*. 1217, 1950–1955. doi:10.1016/j.chroma.2010.01.071
- Andricopulo, A. D., and Montanari, C. A. (2005). Structure-Activity Relationships for the Design of Small-Molecule Inhibitors. *Mini Rev. Med. Chem.* 5, 585–593. doi:10.2174/1389557054023224
- Arnal, A., Waleckx, E., Rico-Chávez, O., Herrera, C., and Dumonteil, E. (2019). Estimating the Current burden of Chagas Disease in Mexico: A Systematic Review and Meta-Analysis of Epidemiological Surveys from 2006 to 2017. *Plos Negl. Trop. Dis.* 13, e0006859. doi:10.1371/journal.pntd.0006859
- Avelar, L. A., Camilo, C. D., de Albuquerque, S., Fernandes, W. B., Gonçalves, C., Kenny, P. W., et al. (2015). Molecular Design, Synthesis and Trypanocidal Activity of Dipeptidyl Nitrides as Cruzain Inhibitors. *Plos Negl. Trop. Dis.* 9, e0003916. doi:10.1371/journal.pntd.0003916
- Bartrop, J. A., Owen, T. C., Cory, A. H., and Cory, J. G. (1991). 5-(3-Carboxymethoxyphenyl)-2-(4,5-Dimethylthiazolyl)-3-(4-Sulfonyl) tetrazolium, Inner Salt (MTS) and Related Analogs of 3-(4,5-Dimethylthiazolyl)-2,5-Diphenyltetrazolium Bromide (MTT) Reducing to Purple Water-Soluble Formazans as Cell-Viability Indicators. *Bioorg. Med. Chem. Lett.* 1, 611–614. doi:10.1016/S0960-894X(01)81162-8
- Bern, C. (2015). Chagas' Disease. *N. Engl. J. Med.* 373, 456–466. doi:10.1056/NEJMr1410150
- Brak, K., Kerr, I. D., Barrett, K. T., Fuchi, N., Debnath, M., Ang, K., et al. (2010). Nonpeptidic Tetrafluorophenoxymethyl Ketone Cruzain Inhibitors as Promising New Leads for Chagas Disease Chemotherapy. *J. Med. Chem.* 53, 1763–1773. doi:10.1021/jm901633v
- Brener, Z. (1962). Therapeutic Activity and Criterion of Cure on Mice Experimentally Infected with Trypanosoma Cruzi. *Rev. Inst. Med. Trop. Sao Paulo* 4, 389–396.
- Buckner, F. S., Verlinde, C. L., La Flamme, A. C., and Van Voorhis, W. C. (1996). Efficient Technique for Screening Drugs for Activity against Trypanosoma Cruzi Using Parasites Expressing Beta-Galactosidase. *Antimicrob. Agents Chemother.* 40, 2592–2597. doi:10.1128/AAC.40.11.2592
- Clark, M., Cramer, R. D., III, and Van Opdenbosch, N. (1989). Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* 10, 982–1012. doi:10.1002/jcc.540100804
- Cucunubá, Z. M., Okuwoga, O., Basáñez, M. G., and Nouvellet, P. (2016). Increased Mortality Attributed to Chagas Disease: A Systematic Review and Meta-Analysis. *Parasit. Vectors* 9, 42. doi:10.1186/s13071-016-1315-x
- Davies, B., and Morris, T. (1993). Physiological Parameters in Laboratory Animals and Humans. *Pharm. Res.* 10, 1093–1095. doi:10.1023/a:1018943613122
- De Rycker, M., Baragana, B., Duce, S. L., and Gilbert, I. H. (2018). Challenges and Recent Progress in Drug Discovery for Tropical Diseases. *Nature* 559, 498–506. doi:10.1038/s41586-018-0327-4
- de Souza, M. L., de Oliveira Rezende Junior, C., Ferreira, R. S., Espinoza Chávez, R. M., Ferreira, L. L. G., Slafer, B. W., et al. (2020). Discovery of Potent, Reversible, and Competitive Cruzain Inhibitors with Trypanocidal Activity: A Structure-Based Drug Design Approach. *J. Chem. Inf. Model.* 60, 1028–1041. doi:10.1021/acs.jcim.9b00802
- Doyle, P. S., Zhou, Y. M., Hsieh, I., Greenbaum, D. C., McKerrow, J. H., and Engel, J. C. (2011). The Trypanosoma Cruzi Protease Cruzain Mediates Immune Evasion. *Plos Pathog.* 7, e1002139. doi:10.1371/journal.ppat.1002139
- Eddershaw, P. J., Beresford, A. P., and Bayliss, M. K. (2000). ADME/PK as Part of a Rational Approach to Drug Discovery. *Drug Discov. Todaytoday* 5, 409–414. doi:10.1016/s1359-6446(00)01540-3
- Engel, J. C., Doyle, P. S., Palmer, J., Hsieh, I., Bainton, D. F., and McKerrow, J. H. (1998). Cysteine Protease Inhibitors Alter Golgi Complex Ultrastructure and Function in Trypanosoma Cruzi. *J. Cel Sci.* 111, 597–606. doi:10.1242/jcs.111.5.597
- Espíndola, J. W., Cardoso, M. V., Filho, G. B., Oliveira E Silva, D. A., Moreira, D. R., Bastos, T. M., et al. (2015). Synthesis and Structure-Activity Relationship Study of a New Series of Antiparasitic Aryloxyl Thiosemicarbazones Inhibiting Trypanosoma Cruzi Cruzain. *Eur. J. Med. Chem.* 101, 818–835. doi:10.1016/j.ejmech.2015.06.048
- Ferraz, M. L., Gazzinelli, R. T., Alves, R. O., Urbina, J. A., and Romanha, A. J. (2007). The Anti-Trypanosoma Cruzi Activity of Posaconazole in a Murine Model of Acute Chagas' Disease Is Less Dependent on Gamma Interferon Than that of Benznidazole. *Antimicrob. Agents Chemother.* 51, 1359–1364. doi:10.1128/AAC.01170-06
- Ferreira, L. G., and Andricopulo, A. D. (2017). Targeting Cysteine Proteases in Trypanosomatid Disease Drug Discovery. *Pharmacol. Ther.* 180, 49–61. doi:10.1016/j.pharmthera.2017.06.004
- Ferreira, R. A. A., Pauli, I., Sampaio, T. S., de Souza, M. L., Ferreira, L. L. G., Magalhães, L. G., et al. (2019). Structure-Based and Molecular Modeling Studies for the Discovery of Cyclic Imides as Reversible Cruzain Inhibitors with Potent Anti-Trypanosoma Cruzi Activity. *Front. Chem.* 7, 798. doi:10.3389/fchem.2019.00798
- Ferreira, R. S., Bryant, C., Ang, K. K., McKerrow, J. H., Shoichet, B. K., and Renslo, A. R. (2009). Divergent Modes of Enzyme Inhibition in a Homologous Structure-Activity Series. *J. Med. Chem.* 52, 5005–5008. doi:10.1021/jm9009229
- Ferreira, R. S., Dessoy, M. A., Pauli, I., Souza, M. L., Krogh, R., Sales, A. I., et al. (2014). Synthesis, Biological Evaluation, and Structure-Activity Relationships of Potent Noncovalent and Nonpeptidic Cruzain Inhibitors as Anti-Trypanosoma Cruzi Agents. *J. Med. Chem.* 57, 2380–2392. doi:10.1021/jm401709b
- Ferreira, R. S., Simeonov, A., Jadhav, A., Eidam, O., Mott, B. T., Keiser, M. J., et al. (2010). Complementarity between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* 53, 4891–4905. doi:10.1021/jm100488w
- Fonseca, N. C., da Cruz, L. F., da Silva Villela, F., do Nascimento Pereira, G. A., de Siqueira-Neto, J. L., Kellar, D., et al. (2015). Synthesis of a Sugar-Based Thiosemicarbazone Series and Structure-Activity Relationship versus the Parasite Cysteine Proteases Rhodensin, Cruzain, and Schistosoma Mansoni Cathepsin B1. *Antimicrob. Agents Chemother.* 59, 2666–2677. doi:10.1128/AAC.04601-14
- Gasteiger, J., and Marsili, M. (1980). Iterative Partial Equalization of Orbital Electronegativity-A Rapid Access to Atomic Charges. *Tetrahedron* 36, 3219–3228. doi:10.1016/0040-4020(80)80168-2
- GBD DALYs and HALE Collaborators (2016). Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 315 Diseases and Injuries and Healthy Life Expectancy (HALE), 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015. *Lancet* 388, 1603–1658. doi:10.1016/S0140-6736(16)31460-X
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* 267, 727–748. doi:10.1006/jmbi.1996.0897
- Jose Cazzulo, J., Stoka, V., and Turk, V. (2001). The Major Cysteine Proteinase of Trypanosoma Cruzi: A Valid Target for Chemotherapy of Chagas Disease. *Curr. Pharm. Des.* 7, 1143–1156. doi:10.2174/1381612013397528
- Latorre, A., Schirmeister, T., Kesselring, J., Jung, S., Johé, P., Hellmich, U. A., et al. (2016). Dipeptidyl Nitroalkenes as Potent Reversible Inhibitors of Cysteine Proteases Rhodensin and Cruzain. *ACS Med. Chem. Lett.* 7, 1073–1076. doi:10.1021/acsmedchemlett.6b00276
- Libisch, M. G., Rego, N., and Robello, C. (2021). Transcriptional Studies on Trypanosoma Cruzi - Host Cell Interactions: A Complex Puzzle of Variables. *Front. Cel. Infect. Microbiol.* 11, 692134. doi:10.3389/fcimb.2021.692134
- Lill, M. A., and Danielson, M. L. (2011). Computer-Aided Drug Design Platform Using PyMOL. *J. Comput. Aided Mol. Des.* 25, 13–19. doi:10.1007/s10822-010-9395-8
- Lima, A. P., Reis, F. C., and Costa, T. F. (2013). Cysteine Peptidase Inhibitors in Trypanosomatid Parasites. *Curr. Med. Chem.* 20, 3152–3173. doi:10.2174/0929867311320250009
- Liu, X., and Jia, L. (2007). The Conduct of Drug Metabolism Studies Considered Good Practice (I): Analytical Systems and In Vivo Studies. *Curr. Drug Metab.* 8, 815–821. doi:10.2174/138920007782798153
- Lombardo, F., Obach, R. S., Shalaeva, M. Y., and Gao, F. (2004). Prediction of Human Volume of Distribution Values for Neutral and Basic Drugs. 2. Extended Data Set and Leave-Class-Out Statistics. *J. Med. Chem.* 47, 1242–1250. doi:10.1021/jm030408h

- Lombardo, F., Obach, R. S., Shalaeva, M. Y., and Gao, F. (2002). Prediction of Volume of Distribution Values in Humans for Neutral and Basic Drugs Using Physicochemical Measurements and Plasma Protein Binding Data. *J. Med. Chem.* 45, 2867–2876. doi:10.1021/jm0200409
- Masimirembwa, C. M., Bredberg, U., and Andersson, T. B. (2003). Metabolic Stability for Drug Discovery and Development: Pharmacokinetic and Biochemical Challenges. *Clin. Pharmacokinet.* 42, 515–528. doi:10.2165/00003088-200342060-00002
- Massarico Serafim, R. A., Gonçalves, J. E., de Souza, F. P., de Melo Loureiro, A. P., Storpirtis, S., Krogh, R., et al. (2014). Design, Synthesis and Biological Evaluation of Hybrid Bioisoster Derivatives of N-Acylhydrazones and Furoxan Groups with Potential and Selective Anti-Trypanosoma Cruzi Activity. *Eur. J. Med. Chem.* 82, 418–425. doi:10.1016/j.ejmech.2014.05.077
- McKerrow, J. H. (1999). Development of Cysteine Protease Inhibitors as Chemotherapy for Parasitic Diseases: Insights on Safety, Target Validation, and Mechanism of Action. *Int. J. Parasitol.* 29, 833–837. doi:10.1016/S0020-7519(99)00044-2
- McKerrow, J. H., Doyle, P. S., Engel, J. C., Podust, L. M., Robertson, S. A., Ferreira, R., et al. (2009). Two Approaches to Discovering and Developing New Drugs for Chagas Disease. *Mem. Inst. Oswaldo Cruz.* 104 (Suppl. 1), 263–269. doi:10.1590/s0074-02762009000900034
- Ndao, M., Beaulieu, C., Black, W. C., Isabel, E., Vasquez-Camargo, F., Nath-Chowdhury, M., et al. (2014). Reversible Cysteine Protease Inhibitors Show Promise for a Chagas Disease Cure. *Antimicrob. Agents Chemother.* 58, 1167–1178. doi:10.1128/AAC.01855-13
- Neitz, R. J., Bryant, C., Chen, S., Gut, J., Hugo Caselli, E., Ponce, S., et al. (2015). Tetrafluorophenoxymethyl Ketone Cruzain Inhibitors with Improved Pharmacokinetic Properties as Therapeutic Leads for Chagas' Disease. *Bioorg. Med. Chem. Lett.* 25, 4834–4837. doi:10.1016/j.bmcl.2015.06.066
- Obach, R. S. (1999). Prediction of Human Clearance of Twenty-Nine Drugs from Hepatic Microsomal Intrinsic Clearance Data: An Examination of *In Vitro* Half-Life Approach and Nonspecific Binding to Microsomes. *Drug Metab. Dispos.* 27, 1350–1359.
- Pérez-Molina, J. A., and Molina, I. (2018). Chagas Disease. *The Lancet* 391, 82–94. doi:10.1016/S0140-6736(17)31612-4
- Plant, N. (2004). Strategies for Using *In Vitro* Screens in Drug Metabolism. *Drug Discov. Today* 9, 328–336. doi:10.1016/s1359-6446(03)03019-8
- Powell, M. J. D. (1977). Restart Procedures for the Conjugate Gradient Method. *Math. Programming* 12, 241–254. doi:10.1007/BF01593790
- Proctor, N. J., Tucker, G. T., and Rostami-Hodjegan, A. (2004). Predicting Drug Clearance from Recombinantly Expressed CYPs: Intersystem Extrapolation Factors. *Xenobiotica* 34, 151–178. doi:10.1080/00498250310001646353
- Rodríguez, H. O., Guerrero, N. A., Fortes, A., Santi-Rocca, J., Gironès, N., and Fresno, M. (2014). Trypanosoma Cruzi Strains Cause Different Myocarditis Patterns in Infected Mice. *Acta Trop.* 139, 57–66. doi:10.1016/j.actatropica.2014.07.005
- Rodrigues Coura, J., and De Castro, S. L. (2002). A Critical Review on Chagas Disease Chemotherapy. *Mem. Inst. Oswaldo Cruz.* 97, 3–24. doi:10.1590/s0074-02762002000100001
- Rogers, K. E., Keränen, H., Durrant, J. D., Ratnam, J., Doak, A., Arkin, M. R., et al. (2012). Novel Cruzain Inhibitors for the Treatment of Chagas' Disease. *Chem. Biol. Drug Des.* 80, 398–405. doi:10.1111/j.1747-0285.2012.01416.x
- Spaggiari, D., Geiser, L., and Rudaz, S. (2014). Coupling Ultra-high-pressure Liquid Chromatography with Mass Spectrometry for *In-Vitro* Drug-Metabolism Studies. *Trac Trends Anal. Chem.* 63, 129–139. doi:10.1016/j.trac.2014.06.021
- Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved Protein-Ligand Docking Using GOLD. *Proteins* 52, 609–623. doi:10.1002/prot.10465
- Wang, H., Zrada, M., Anderson, K., Katwaru, R., Harradine, P., Choi, B., et al. (2014). Understanding and Reducing the Experimental Variability of *In Vitro* Plasma Protein Binding Measurements. *J. Pharm. Sci.* 103, 3302–3309. doi:10.1002/jps.24119
- Wang, J. (2009). Comprehensive Assessment of ADMET Risks in Drug Discovery. *Curr. Pharm. Des.* 15, 2195–2219. doi:10.2174/138161209788682514
- Wang, J., and Skolnik, S. (2009). Recent Advances in Physicochemical and ADMET Profiling in Drug Discovery. *Chem. Biodivers.* 6, 1887–1899. doi:10.1002/cbdv.200900117
- Wang, J., Urban, L., and Bojanic, D. (2007). Maximising Use of *In Vitro* ADMET Tools to Predict *In Vivo* Bioavailability and Safety. *Expert Opin. Drug Metab. Toxicol.* 3, 641–665. doi:10.1517/17425255.3.5.641
- Waring, M. J. (2010). Lipophilicity in Drug Discovery. *Expert Opin. Drug Discov.* 5, 235–248. doi:10.1517/17460441003605098
- Yu, H., Wang, Q., Sun, Y., Shen, M., Li, H., and Duan, Y. (2015). A New PAMPA Model Proposed on the Basis of a Synthetic Phospholipid Membrane. *PLoS One* 10, e0116502. doi:10.1371/journal.pone.0116502
- Zamek-Gliszczyński, M. J., Ruterbories, K. J., Ajamie, R. T., Wickremsinhe, E. R., Pothuri, L., Rao, M. V., et al. (2011). Validation of 96-Well Equilibrium Dialysis with Non-Radiolabeled Drug for Definitive Measurement of Protein Binding and Application to Clinical Development of Highly-Bound Drugs. *J. Pharm. Sci.* 100, 2498–2507. doi:10.1002/jps.22452
- Zanatta, N., Amaral, S. S., Dos Santos, J. M., de Mello, D. L., Fernandes, Lda. S., Bonacorso, H. G., et al. (2008). Convergent Synthesis and Cruzain Inhibitory Activity of Novel 2-(N'-Benzylidenehydrazino)-4-Trifluoromethyl-Pyrimidines. *Bioorg. Med. Chem.* 16, 10236–10243. doi:10.1016/j.bmc.2008.10.052

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Pauli, Rezende Jr., Slafer, Dessoy, de Souza, Ferreira, Adjanohun, Ferreira, Magalhães, Krogh, Michelan-Duarte, Del Pintor, da Silva, Cruz, Dias and Andricopulo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



ER/AR Multi-Conformational Docking Server: A Tool for Discovering and Studying Estrogen and Androgen Receptor Modulators

Feng Wang¹, Shuai Hu², De-Qing Ma¹, Qiuye Li², Hong-Cheng Li², Jia-Yi Liang², Shan Chang^{2*} and Ren Kong^{2*}

¹Changzhou University Huaide College, Taizhou, China, ²Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, School of Chemical and Environmental Engineering, Jiangsu University of Technology, Changzhou, China

OPEN ACCESS

Edited by:

Leonardo L. G. Ferreira,
University of São Paulo, Brazil

Reviewed by:

Darren Fayne,
Trinity College Dublin, Ireland
Aldo Oliveira,
Federal University of Santa Catarina,
Brazil

*Correspondence:

Shan Chang
schang@jsut.edu.cn
Ren Kong
rkong@jsut.edu.cn

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 24 October 2021

Accepted: 03 January 2022

Published: 24 January 2022

Citation:

Wang F, Hu S, Ma D-Q, Li Q, Li H-C,
Liang J-Y, Chang S and Kong R (2022)
ER/AR Multi-Conformational Docking
Server: A Tool for Discovering and
Studying Estrogen and Androgen
Receptor Modulators.
Front. Pharmacol. 13:800885.
doi: 10.3389/fphar.2022.800885

The prediction of the estrogen receptor (ER) and androgen receptor (AR) activity of a compound is quite important to avoid the environmental exposures of endocrine-disrupting chemicals. The Estrogen and Androgen Receptor Database (EARDDB, <http://eardb.schanglab.org.cn/>) provides a unique collection of reported ER α , ER β , or AR protein structures and known small molecule modulators. With the user-uploaded query molecules, molecular docking based on multi-conformations of a single target will be performed. Moreover, the 2D similarity search against known modulators is also provided. Molecules predicted with a low binding energy or high similarity to known ER α , ER β , or AR modulators may be potential endocrine-disrupting chemicals or new modulators. The server provides a tool to predict the endocrine activity for compounds of interests, benefiting for the ER and AR drug design and endocrine-disrupting chemical identification.

Keywords: estrogen receptor (ER), androgen receptor (AR), molecular docking, similarity search, web-server

INTRODUCTION

With the development of chemistry technology, numerous natural or non-natural compounds are synthesized and used in the daily life of human beings such as medicines, perfumes, food additives, automobiles, electronics, pesticides, textiles, plastics, and so on (Barr Dana et al., 2005; Judson et al., 2011). It is noted that a number of the compounds act as endocrine-disrupting chemicals (EDCs) with the potential to interfere the hormone systems in human or wild lives (Schug et al., 2011; Dionisio et al., 2015). The occupational and environmental exposures of EDCs are strongly correlated with the adverse health outcomes such as reproductive health, development disorders, oncological, immunological and cardiovascular disease, obesity, and neurobehavior disorders (Elobeid and Allison 2008; Yilmaz et al., 2020; Boudalia et al., 2021; O'Shaughnessy et al., 2021; Priya et al., 2021). Numerous efforts have been taken to identify that if a compound is endocrine-active or not. The Endocrine Disruptor Screening Program (EDSP) and the Toxicology Testing in the 21st Century (Tox21) projects set up various *in vitro* or *in vivo* assays to measure the potential effects of chemicals on the endocrine system in humans or wildlife (Judson et al., 2010; Willett et al., 2011; Judson et al., 2015; Yilmaz et al., 2020). However, high costs and low speed make the experimental methods not fulfill the need of testing the rapid increased number of synthetic chemicals in use. Currently, only a small fraction of compounds have the experimental determined endocrine activity data available (Egeghy et al., 2012; Tickner et al., 2019). It is of great need to develop the predictive models to provide clues of the compounds' endocrine activity.

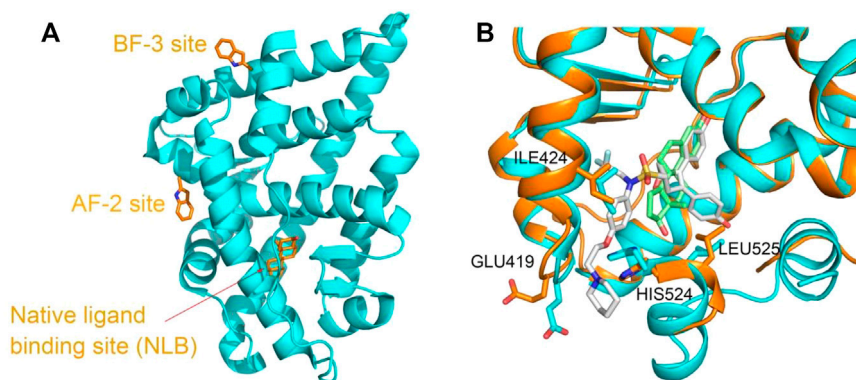


FIGURE 1 | (A) Three binding sites in the ligand-binding domain of the AR depicted based on PDBID 2POI and **(B)** the ligand-binding site flexibility of ER α in complex with different compounds, estradiol ((8R,9S,13S,14S,17S)-13-methyl-6,7,8,9,11,12,14,15,16,17-decahydrocyclopenta[a]phenanthrene-3,17-diol) and 7AI ((1S,2R,4S)-5,6-bis(4-hydroxyphenyl)-N-{4-[2-(piperidin-1-yl)ethoxy]phenyl}-N-(2,2,2-trifluoroethyl)-7-oxabicyclo[2.2.1]hept-5-ene-2-sulfonamide) (structures from PDBID 5GS4 and 7RRX). Proteins are represented in a cartoon model, and compounds or residues, in a stick model. In **(B)**, proteins from 5GS4 and 7RRX are colored in cyan and orange, and the compounds estradiol and 7AI are colored in green and gray, respectively. Residues undergoing considerable conformation changes, such as ILE 424, GLU419, HIS524, and LEU525 are depicted in the stick model.

The *in silico* methods, especially the ligand-based QSAR (quantitative structure–activity relationships) approaches and structural base docking methods, are widely used in the computer-aided drug design field (Mao et al., 2021; Sabe et al., 2021). These methods are applied to predict the compound's activities against endocrine-related proteins, such as the estrogen receptor (ER) and androgen receptor (AR) (Schneider et al., 2019). The CERAPP (Collaborative Estrogen Receptor Activity Prediction Project) and CoMPARA (Collaborative Modeling Project for Androgen Receptor Activity) construct various predictive models trained by different QSAR approaches for estrogen or androgen receptor activity prediction (Mansouri et al., 2016; Mansouri et al., 2020). Both categorical and continuous models are built based on the dataset provided by the U.S. EPA, and a consensus model was obtained by weighting the models on scores based on evaluated accuracies of single models. Shen et al. collected the estrogenic activity data from public sources and developed QSAR models based on the dataset (Shen et al., 2013). Machine learning and deep learning methods are also applied in EDC predictions and achieved a relative high overall predictive accuracy (Zhang et al., 2017). One of the important limitations of the QSAR-based model is the quality of data for model training (Maggiore 2006). The structure-based docking method provides significant complementation for EDC prediction. The “Endocrine disruptome” server provides docking models for 14 nuclear hormone receptors such as ER, AR, glucocorticoid receptor, liver X receptors, etc. (Kolšek et al., 2014). However, only 18 structures are incorporated in the server.

ER and AR are also pivotal therapeutic targets due to the roles in regulation of development, endocrinology, and metabolism, and numerous compounds are developed to modulate the protein functions. The nuclear receptors including the ER and AR are composed of several functional domains, such as the N-terminal domain (NTD), the DNA-binding domain (DBD), and the ligand-binding domain (LBD) (McEwan 2009). The investigation of PDB structures show that except the native hormone-binding pocket, there are activation function 2 (AF-

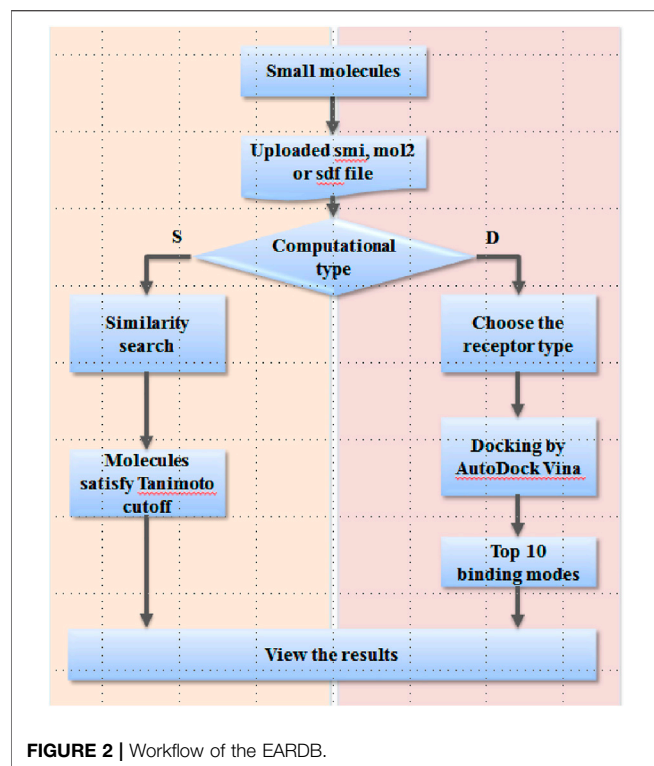
2) pocket and binding function 3 (BF-3) pocket located on the ligand-binding domain (LBD) of the ER and AR (Estébanez-Perpiñá et al., 2007; Axerio-Cilies et al., 2011; Lack et al., 2011; Buzón et al., 2012). Compounds bound to either of the pockets are demonstrated to interfere with the protein functions and the related signal pathway as agonists or antagonists (Moore et al., 2010; Nwachukwu et al., 2017) (Figure 1A). Even for the same pocket, such as the native ligand-binding pocket, considerable conformational changes occurred upon various types of compound binding (Min et al., 2021) (Figure 1B).

Current used small molecule docking programs such as AutoDock Vina can only consider the flexibility of the small molecule while keeping the protein conformation fixed (Trott and Olson 2010). The bias will be introduced only when one of the protein structures is used as the receptor for docking experiments. To integrate the conformation change information from the resolved structures, here we built docking models based on all available complex structures of the ER and AR and constructed a docking server (Estrogen and Androgen Receptor Database, EARDDB, <http://earddb.schanglab.org.cn/>). The user can easily dock the compound of interest to the conformational ensembles of ER α , ER β , and AR by several simple clicks. The top 10 highest docking score poses among all the ensembles are returned. The compound fit to any of the pocket may be potential EDCs of the ER or AR. In addition, the 2D similarity search function for the known ER or AR modulators is also implemented based on data retrieved from the binding database. The aim is to provide a tool to predict any possible ER or AR effectors with multi-conformational docking models and ligand-based similarity.

MATERIAL AND METHODS

Construction of Docking Models

By using the advance search option, the structures of ER α , ER β , and AR are retrieved from the PDB (the Protein Data Bank) with



the uniprot accession numbers of P03372, Q92731, and P10275. The structures with small molecules bound with the protein are used to generate the docking models. Totally, 282, 29, and 81 PDB structures are downloaded for proteins ER α , ER β , and AR, respectively. For a crystal structure with multiple chains, each chain is separated and

treated as a unique docking model. Then, the protein structure and ligand structure are split into different files. AutoDock Vina is chosen as the docking tool due to its excellent performance in systematic docking program evaluations (Trott and Olson 2010; Wang et al., 2016). The protein structures are operated by Receptor_prepare.py to remove waters and add polar hydrogens and Kollman charges to obtain the receptor pdbqt files. The docking box is defined by using the geometric center of the native binding ligand from the original PDB as the box center with $28 \times 28 \times 28$ Å in size to include the entire binding site. Ligand_prepare.py is used to generate the ligand pdbqt file. The flexible bonds are set as default, and the Gasteiger charge is computed for the ligand. Both the receptor and ligand pdbqt files are generated in the neutral pH condition. The top ten docking poses are allowed to output with the docking score. In the re-dock experiment, the native binding ligands are docked into the corresponding receptors to validate the docking protocol. The receptor files which possess docking poses less than 2 Å rmsd with the native binding ligand are incorporated to the web server to evaluate the user-submitted small molecules.

Activity Data Curation

We also collected the activity data of the reported ER α , ER β , and AR agonist or antagonist from the binding database (<http://www.bindingdb.org>) (Chen et al., 2001; Liu et al., 2007; Gilson et al., 2016) to enable the user to evaluate the 2D similarity of his own compound and the known ER/AR modulators. All the records with the uniprot accession numbers of P03372, Q92731, and P10275 are retrieved from the binding database and implemented in the local server.

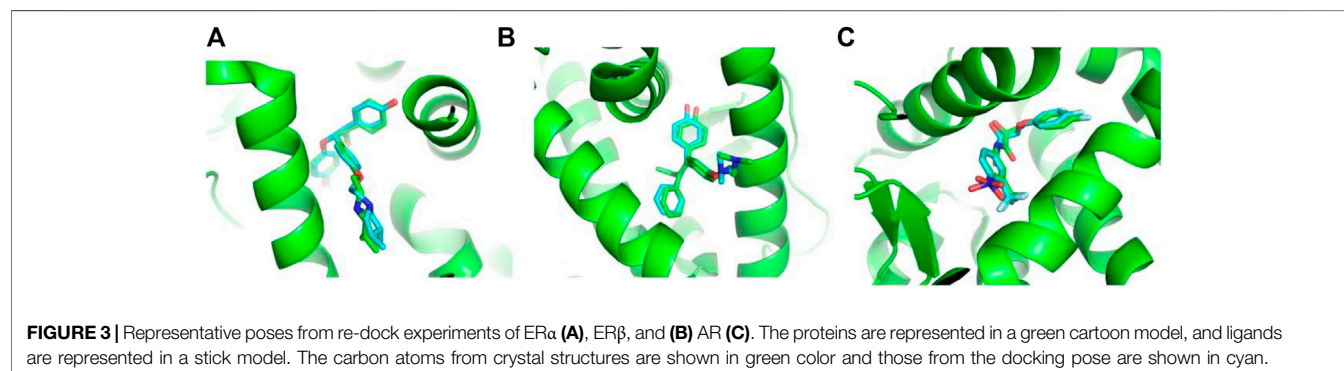
Structure Similarity Search

In the structure similarity search function, the smi, mol2, or sdf file of the interested compound need to be provided by the user.

TABLE 1 | Number of structures and docking models for ER α , ER β , and AR.

| Protein name | Number of crystal structures with the small molecule ligand | Number of docking modes in re-dock experiment | Number of successful docked systems in re-dock experiment ^a |
|--------------|---|---|--|
| ER α | 282 | 609 | 580 |
| ER β | 32 | 66 | 62 |
| AR | 81 | 105 | 91 |

^aDocking models with the lowest RMSD of docking poses less than 3.0 Å are defined as successful docking systems.



The Tanimoto similarity coefficient between the user-uploaded small molecule and the ligands in the EARDDB is computed by Open Babel 2.3.2 based on the linear fingerprint of fragment indices. The Tanimoto coefficient is a value range from 0 to 1, representing the level of similarity between two molecules. The value of 1 is the highest similarity and indicates the same molecules, and the value of 0 is the lowest similarity. The EARDDB will automatically calculate the Tanimoto value for each ligand in the database with the query compound and will only return the ligands with similarity values higher than the user-defined cutoff with a descending rank.

Implementation

The EARDDB is installed on CentOS 7.6 server workstations. The webserver platform is constructed by Apache 2.4.6, and the website was built with PHP 5.6.4. The MySQL 5.7 Database management system was used to organize, manage, and sort data of various types. The open-source Java viewer NGL is embedded on the webpage for 3D molecular visualization (<https://nglviewer.org/>). Open Babel 2.3.2 is used for format transformation, 3D coordinate generation, and 2D similarity search for the uploaded files (O'Boyle et al., 2011). AutoDock Vina 1.1.2 (Trott and Olson 2010) is used to obtain the docking scores and binding modes with default settings.

RESULTS AND DISCUSSION

Overview of the Database

The EARDDB (<http://eardb.schanglab.org.cn/>) currently implements two major functions, the 2D chemical similarity search for known ER/AR modulators and the online docking module to predict the potential ER/AR modulator (**Figure 2**). The ligand database contains about 7,800 unique compounds associated with 13,190 related activity records of the ER and AR from *Homo sapiens*. For each ER/AR modulator entry, the molecular chemical name, the 2D chemical structure and monomer ID from the binding database are provided. The online docking module provides a web-based interface to predict the binding mode and binding affinity for the user-uploaded compounds with the protein of interest. There are three types of protein targets, including ER α , ER β , and AR. For each type of target, structures in complex with differential compounds are retrieved from the RCSB Protein Data Bank (Rose et al., 2011). Totally, 580, 62, and 91 docking models derived from the experimental structures of ER α , ER β , and AR are available on the server.

The workflow of the web server is shown in **Figure 2**. For the user-interested compound, the strict smi, mol2, or sdf format file is needed to upload to the server. Two computational types are provided as following: "S" for similarity search and "D" for docking. For similarity search, the Tanimoto cutoff needs to be defined by the user. By submission of the job, Open Babel 2.3.2 is launched on the server to retrieve any compounds with Tanimoto values greater than the cutoff value. A molecular table is presented on the webpage to display the results, and also, a tab-delimited txt file is provided to download. For the

A Input webpage for multi-conformational docking of ER α , ER β , or AR. The page shows the EARDDB logo and a navigation bar. Below the navigation bar, there are sections for 'Additional Notes' and 'Docking'. The 'Docking' section includes an 'Upload File' button, a text input for the file name (example: FHM.smi, FHM.mol2, FHM.sdf), a 'Choose A Receptor' dropdown menu (options: AR, ER α , ER β), and an 'Enter Mail Address' field. A 'Submit' button is at the bottom.

B Result page of multi-conformational docking (example running for FHM). The page shows a 'View Docking Results' section with a table of docking scores. The table has columns for Rank, Receptor Name, and docking score (kcal/mol). The table lists 10 docking models. Below the table, there is a 'Download Result' button. To the right of the table, there is a 3D visualization of the docking results, showing the protein structure and the docked ligand. The visualization includes a 'Style' dropdown menu (options: Cartoon, Ribbon, Surface) and a 'Color' dropdown menu (options: Random, Spin, Reset).

C Result page of multi-conformational docking (example running for DDT). The page shows a 'View Docking Results' section with a table of docking scores. The table has columns for Rank, Receptor Name, and docking score (kcal/mol). The table lists 10 docking models. Below the table, there is a 'Download Result' button. To the right of the table, there is a 3D visualization of the docking results, showing the protein structure and the docked ligand. The visualization includes a 'Style' dropdown menu (options: Cartoon, Ribbon, Surface) and a 'Color' dropdown menu (options: Random, Spin, Reset).

FIGURE 4 | (A) Input webpage for multi-conformational docking of ER α , ER β , or AR; (B) The result page of multi-conformational docking (example running for FHM); (C) The result page of multi-conformational docking (example running for DDT).

docking module, the target type needs to be selected as the first step. By submission of the job, ligand preparation, a series of molecular docking experiments against all structures of the specific target type will be automatically carried out on the EARDDB server. The top 10 models ranked by the predicted binding affinities are kept and visualized in 3D by NGL. A package of docking results is also provided to download from the results page.

A

Similarity Search

• Upload File:

(Example :FHM.smi)

• Tanimoto:

B

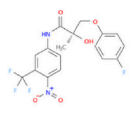
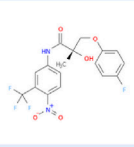
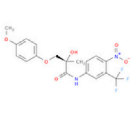
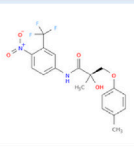
| LigandName | Image | Tanimoto | MonomerID |
|--|--|----------|-----------------------|
| (2R)-3-(4-fluorophenoxy)-2-hydroxy-2-methyl-N-[4-nitro-3-(trifluoromethyl)phenyl]propanamide::Nonsteroidal AR Ligand, R-5 |  | 1 | 18664 |
| (2S)-3-(4-fluorophenoxy)-2-hydroxy-2-methyl-N-[4-nitro-3-(trifluoromethyl)phenyl]propanamide::BMC185567 S-1::ChEMBL124718::Nonsteroidal AR Ligand, S-5 |  | 1 | 18663 |
| (2S)-2-hydroxy-3-(4-methoxyphenoxy)-2-methyl-N-[4-nitro-3-(trifluoromethyl)phenyl]propanamide::Nonsteroidal AR Ligand, S-13 |  | 0.916667 | 18672 |
| (2S)-2-hydroxy-2-methyl-3-(4-methylphenoxy)-N-[4-nitro-3-(trifluoromethyl)phenyl]propanamide::Nonsteroidal AR Ligand, S-12 |  | 0.916667 | 18671 |

FIGURE 5 | Input (A) and output (B) webpages for 2D similarity search.

Validation of Docking Models

To consider the conformational change of proteins upon binding different ligands, we retrieved all the protein–small molecule complex structures of ER α , ER β , and AR from the PDB website. As shown in Table 1, there are 282, 32, and 81 complex structures for ER α , ER β , and AR, respectively. Based on these structures, totally 621, 66, and 105 docking models were obtained by separating chains. Then, the re-dock experiments were performed to dock the native ligand from the experimental structures to the protein active site. It is necessary to validate the parameters set in the docking protocol, as well as if the protein structure is qualified for docking. The root mean square deviation (RMSD) between the docking poses generated by Vina and the native ligand structure from the original experimental structure are used to evaluate the accuracy of re-dock experiments. As shown in Table 1 and Supplementary Table S1, among 780 docking models, 733 models obtained docking poses with RMSD less than 2.0 Å (Supplementary Table S1). The re-dock results of 1XP1 (ER α) (Blizzard et al., 2005), 2FSZ (ER β) (Wang et al., 2006), and 2AX8 (AR) (Bohl et al., 2005) are shown here as examples. The docking poses and the original crystal structures are presented in Figure 3. The RMSD values of 1XP1 (ER α), 2FSZ (ER β), and 2AX8 (AR) are 1.48, 0.87, and 1.12 Å, respectively.

The docking poses superimpose well with the native ligand structure from the original crystal structures, indicating the docking procedure is able to recover the experimental structures. To include more conformational diverse structures, the docking models with RMSD less than 2.0 Å are considered as successful docking systems and kept in the receptor database to provide readily to dock function on the web server.

Multiple Conformation-Based Docking

Multiple structures of ER α , ER β , and AR are collected from the PDB database and prepared as docking models. The user could upload the smi, sdf, or mol2 file of one small molecule and then choose a receptor for docking (Figure 4A). An email address is needed to receive the message of job submission and job status. After the job is completed, a notification email will be sent to the user's email address. An investigation drug of AR, ligand name from the PDB as FHM, was used as an example for multiple conformational docking. FHM is a native ligand in the crystal structure with PDBID as 2AXA. The sdf file of FHM was uploaded, and the receptor type of the AR was selected. With one click on submission button, the job is submitted to the server conveniently. After about 3 h running, ninety-three docking experiments were finished. On the "Job Status" page, it is

shown that the job was completed. By click the result link from the email or by search for the job ID in the “Check Result” page, the result page can be accessed. As shown in **Figure 4B**, a table of the top 10 lowest energy poses is provided in the result webpage. The PDBID of the receptor model and the corresponding docking score are presented. For FHM, its original receptor file 2AXA (chain A) obtained the lowest docking score as -11.0 kcal/mol. The binding poses could also be explored by the NGL molecule viewing window on the page. As the docking score of Vina is fitted to the binding affinity, molecules with a low docking score may be a possible modulator.

DDT (4,4-dichlorodiphenyltrichloroethane) is used as another example. As a pesticide, DDT was banned because it acts as an endocrine-disrupting chemical with ER α agonist activity (Nwachukwu et al., 2017). The sdf file of the compound was uploaded to the server, and ER α was chosen as the receptor to perform the multi-conformation-based docking. The results are displayed in **Figure 4C**. It is showed that the compound bound well with various PDB structures of ER α , and the top ten lowest docking scores ranged from -9.3 to -9.1 kcal/mol, indicating DDT as a strong ER α binder (**Figure 4C**).

3.4 2D Similarity Search for Known ER α , ER β , or AR Modulators

Ligand-based chemical structure similarity search is provided on the web server. As shown in **Figure 5A**, the user can choose a molecule in the smiles format from the local storage and upload it to the server. Here, we also take FHM as an example. The default value of 0.6 is taken as the Tanimoto cutoff. As shown in **Figure 5B**, the monomerID from the binding database, ligand name, 2D chemical structure, and Tanimoto value are presented. A hyperlink is added to the monomer ID, and by clicking the ID, the user will be led to the binding database webpage for detailed information, such as activity values, assay description, and publication. The first hit in the table with the Tanimoto value of 1.0 is the compound itself. Molecules highly similar to the potent modulator may have the same function and could be an endocrine active compound.

CONCLUSION

A web server, EARDDB database, was constructed to predict the potential ER α , ER β , and AR modulators. Both structure-based

methods, the multi-conformation docking, and ligand-based method, and 2D similarity search are provided on the server. The investigation of the available PDB structures of ER α , ER β , and AR showed that there are several ligand binding sites on these targets and considerable binding site plasticity. Thus, over 600 docking models are prepared and allowed to settle on the web server to provide docking against the conformational ensemble of each target type. The results of re-dock experiments suggest the docking procedures could reproduce most of the experiment structures of the protein–small molecule complexes. The server provides either the protein structure-based docking function or ligand-based similarity search function to estimate if the query compound is a potential ER α , ER β , or AR modulator. It will benefit for either the ER or AR drug design or EDC prediction.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

RK and SC designed the system and experiments. FW, SH, DM, QL, HL, and JL performed the experiments, data analysis, and the data acquisition. FW, SH, and DM wrote the webpage and constructed the web server. RK, FW, and SC wrote the main manuscript text. All the authors reviewed and approved the manuscript.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (81603152) and the fund of Changzhou Sci. and Tech. Program (CE20205033).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.800885/full#supplementary-material>

REFERENCES

- Axerio-Cilies, P., Lack, N. A., Nayana, M. R., Chan, K. H., Yeung, A., Leblanc, E., et al. (2011). Inhibitors of Androgen Receptor Activation Function-2 (AF2) Site Identified Through Virtual Screening. *J. Med. Chem.* 54, 6197–6205. doi:10.1021/jm200532b
- Barr, D. B., Wang, R. Y., and Needham, L. L. (2005). Biologic Monitoring of Exposure to Environmental Chemicals Throughout the Life Stages: Requirements and Issues for Consideration for the National Children's Study. *Environ. Health Perspect.* 113, 1083–1091. doi:10.1289/ehp.7617
- Blizzard, T. A., DiNinno, F., Morgan, J. D., Chen, H. Y., Wu, J. Y., Kim, S., et al. (2005). Estrogen Receptor Ligands. Part 9: Dihydrobenzoxathiin SERAMs with Alkyl Substituted Pyrrolidine Side Chains and Linkers. *Bioorg. Med. Chem. Lett.* 15, 107–113. doi:10.1016/j.bmcl.2004.10.036
- Bohl, C. E., Miller, D. D., Chen, J., Bell, C. E., and Dalton, J. T. (2005). Structural Basis for Accommodation of Nonsteroidal Ligands in the Androgen Receptor. *J. Biol. Chem.* 280, 37747–37754. doi:10.1074/jbc.M507464200
- Boudalia, S., Bousbia, A., Boumaaza, B., Oudir, M., and Canivenc Lavier, M. C. (2021). Relationship Between Endocrine Disruptors and Obesity with a Focus on Bisphenol A: A Narrative Review. *Bioimpacts.* 11, 289–300. doi:10.34172/bi.2021.33
- Buzón, V., Carbó, L. R., and Estruch, S. B. (2012). A Conserved Surface on the Ligand Binding Domain of Nuclear Receptors for Allosteric Control. *Mol. Cell Endocrinol.* 348, 394–402. doi:10.1016/j.mce.2011.08.012

- Chen, X., Lin, Y., and Gilson, M. K. (2001). The Binding Database: Overview and User's Guide. *Biopolymers*. 61, 127–141. doi:10.1002/1097-0282(2002)61:2<127AID-BIP10076>3.0.CO;2-N
- Dionisio, K. L., Frame, A. M., Goldsmith, M. R., Wambaugh, J. F., Liddell, A., Cathey, T., et al. (2015). Exploring Consumer Exposure Pathways and Patterns of Use for Chemicals in the Environment. *Toxicol. Rep.* 2, 228–237. doi:10.1016/j.toxrep.2014.12.009
- Egeghy, P. P., Judson, R., Gangwal, S., Mosher, S., Smith, D., Vail, J., et al. (2012). The Exposure Data Landscape for Manufactured Chemicals. *Sci. Total Environ.* 414, 159–166. doi:10.1016/j.scitotenv.2011.10.046
- Elobeid, M. A., and Allison, D. B. (2008). Putative Environmental-Endocrine Disruptors and Obesity: A Review. *Curr. Opin. Endocrinol. Diabetes Obes.* 15, 403–408. doi:10.1097/MED.0b013e32830ce95c
- Estébanez-Perpiñá, E., Arnold, L. A., Arnold, A. A., Nguyen, P., Rodrigues, E. D., Mar, E., et al. (2007). A Surface on the Androgen Receptor that Allosterically Regulates Coactivator Binding. *Proc. Natl. Acad. Sci. U S A*. 104, 16074–16079. doi:10.1073/pnas.0708036104
- Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2016). BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* 44, D1045–D1053. doi:10.1093/nar/gkv1072
- Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., et al. (2010). In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. *Environ. Health Perspect.* 118, 485–492. doi:10.1289/ehp.0901392
- Judson, R. S., Kavlock, R. J., Setzer, R. W., Hubal, E. A., Martin, M. T., Knudsen, T. B., et al. (2011). Estimating Toxicity-Related Biological Pathway Altering Doses for High-Throughput Chemical Risk Assessment. *Chem. Res. Toxicol.* 24, 451–462. doi:10.1021/tx100428e
- Judson, R. S., Maggiantay, F. M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., et al. (2015). Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicol. Sci.* 148, 137–154. doi:10.1093/toxsci/kfv168
- Kolšek, K., Mavri, J., Sollner Dolenc, M., Gobec, S., and Turk, S. (2014). Endocrine Disruptome--An Open Source Prediction Tool for Assessing Endocrine Disruption Potential Through Nuclear Receptor Binding. *J. Chem. Inf. Model.* 54, 1254–1267. doi:10.1021/ci400649p
- Lack, N. A., Axerio-Cilies, P., Tavassoli, P., Han, F. Q., Chan, K. H., Feau, C., et al. (2011). Targeting the Binding Function 3 (BF3) Site of the Human Androgen Receptor Through Virtual Screening. *J. Med. Chem.* 54, 8563–8573. doi:10.1021/jm201098n
- Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* 35, D198–D201. doi:10.1093/nar/gkl999
- Maggiora, G. M. (2006). On Outliers and Activity Cliffs--Why QSAR Often Disappoints. *J. Chem. Inf. Model.* 46, 1535. doi:10.1021/ci060117s
- Mansouri, K., Abdelaziz, A., Rybacka, A., Roncaglioni, A., Tropsha, A., Varnek, A., et al. (2016). CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* 124, 1023–1033. doi:10.1289/ehp.1510267
- Mansouri, K., Kleinsteuer, N., Abdelaziz, A. M., Alberga, D., Alves, V. M., Andersson, P. L., et al. (2020). CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environ. Health Perspect.* 128, 27002. doi:10.1289/EHP5580
- Mao, J., Akhtar, J., Zhang, X., Sun, L., Guan, S., Li, X., et al. (2021). Comprehensive Strategies of Machine-Learning-Based Quantitative Structure-Activity Relationship Models. *iScience* 24, 103052. doi:10.1016/j.isci.2021.103052
- McEwan, I. J. (2009). "Nuclear Receptors: One Big Family," in *The Nuclear Receptor Superfamily: Methods and Protocols*. Editor I. J. McEwan (Totowa, NJ: Humana Press). doi:10.1007/978-1-60327-575-0_1
- Min, J., Nwachukwu, J. C., Min, C. K., Njeri, J. W., Srinivasan, S., Rangarajan, E. S., et al. (2021). Dual-Mechanism Estrogen Receptor Inhibitors. *Proc. Natl. Acad. Sci. U S A* 118, e2101657118. doi:10.1073/pnas.2101657118
- Moore, T. W., Mayne, C. G., and Katzenellenbogen, J. A. (2010). Minireview: Not Picking Pockets: Nuclear Receptor Alternate-Site Modulators (NRAMs). *Mol. Endocrinol.* 24, 683–695. doi:10.1210/me.2009-0362
- Nwachukwu, J. C., Srinivasan, S., Bruno, N. E., Nowak, J., Wright, N. J., Minutolo, F., et al. (2017). Systems Structural Biology Analysis of Ligand Effects on ERα Predicts Cellular Response to Environmental Estrogens and Anti-hormone Therapies. *Cell Chem Biol.* 24, 35–45. doi:10.1016/j.chembiol.2016.11.014
- O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An Open Chemical Toolbox. *J. Cheminform.* 3, 33. doi:10.1186/1758-2946-3-33
- O'Shaughnessy, K. L., Fischer, F., and Zenclussen, A. C. (2021). Perinatal Exposure to Endocrine Disrupting Chemicals and Neurodevelopment: How Articles of Daily Use Influence the Development of Our Children. *Best Pract. Res. Clin. Endocrinol. Metab.* 35, 101568. doi:10.1016/j.beem.2021.101568
- Priya, K., Setty, M., Babu, U. V., Paiand K. S. R. (2021). Implications of Environmental Toxicants on Ovarian Follicles: How it Can Adversely Affect the Female Fertility? *Environ. Sci. Pollut. Res. Int.* 28, 67925–67939. doi:10.1007/s11356-021-16489-4
- Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., et al. (2011). The RCSB Protein Data Bank: Redesigned Web Site and Web Services. *Nucleic Acids Res.* 39, D392–D401. doi:10.1093/nar/gkq1021
- Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E. M., Govender, T., Naicker, T., et al. (2021). Current Trends in Computer Aided Drug Design and a Highlight of Drugs Discovered via Computational Techniques: A Review. *Eur. J. Med. Chem.* 224, 113705. doi:10.1016/j.ejmech.2021.113705
- Schneider, M., Pons, J. L., Labesse, G., and Bourguet, W. (2019). In Silico Predictions of Endocrine Disruptors Properties. *Endocrinology*. 160, 2709–2716. doi:10.1210/en.2019-00382
- Schug, T. T., Janesick, A., Blumberg, B., and Heindel, J. J. (2011). Endocrine Disrupting Chemicals and Disease Susceptibility. *J. Steroid Biochem. Mol. Biol.* 127, 204–215. doi:10.1016/j.jsbmb.2011.08.007
- Shen, J., Xu, L., Fang, H., Richard, A. M., Bray, J. D., Judson, R. S., et al. (2013). EADB: An Estrogenic Activity Database for Assessing Potential Endocrine Activity. *Toxicol. Sci.* 135, 277–291. doi:10.1093/toxsci/kft164
- Tickner, J., Jacobs, M., Malloy, T., Buck, T., Stone, A., Blake, A., et al. (2019). Advancing Alternatives Assessment for Safer Chemical Substitution: A Research and Practice Agenda. *Integr. Environ. Assess. Manag.* 15, 855–866. doi:10.1002/ieam.4094
- Trott, O., and Olson, A. J. (2010). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* 31, 455–461. doi:10.1002/jcc.21334
- Wang, Y., Chirgatz, N. Y., Briggs, S. L., Khan, S., Jensen, E. V., and Burris, T. P. (2006). A Second Binding Site for Hydroxytamoxifen within the Coactivator-Binding Groove of Estrogen Receptor Beta. *Proc. Natl. Acad. Sci. U S A*. 103, 9908–9911. doi:10.1073/pnas.0510596103
- Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., et al. (2016). Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein-Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power. *Phys. Chem. Chem. Phys.* 18, 12964–12975. doi:10.1039/c6cp01555g
- Willett, C. E., Bishop, P. L., and Sullivan, K. M. (2011). Application of an Integrated Testing Strategy to the U.S. EPA Endocrine Disruptor Screening Program. *Toxicol. Sci.* 123, 15–25. doi:10.1093/toxsci/kfr145
- Yilmaz, B., Terekeci, H., Sandal, S., and Kelestimur, F. (2020). Endocrine Disrupting Chemicals: Exposure, Effects on Human Health, Mechanism of Action, Models for Testing and Strategies for Prevention. *Rev. Endocr. Metab. Disord.* 21, 127–147. doi:10.1007/s11154-019-09521-z
- Zhang, Q., Yan, L., Wu, Y., Ji, L., Chen, Y., Zhao, M., et al. (2017). A Ternary Classification Using Machine Learning Methods of Distinct Estrogen Receptor Activities within a Large Collection of Environmental Chemicals. *Sci. Total Environ.* 580, 1268–1275. doi:10.1016/j.scitotenv.2016.12.088

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Wang, Hu, Ma, Li, Li, Liang, Chang and Kong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Introduction to the BioChemical Library (BCL): An Application-Based Open-Source Toolkit for Integrated Cheminformatics and Machine Learning in Computer-Aided Drug Discovery

Benjamin P. Brown^{1*}, Oanh Vu², Alexander R. Geanes², Sandeepkumar Kothiwale², Mariusz Butkiewicz², Edward W. Lowe Jr.², Ralf Mueller², Richard Pape², Jeffrey Mendenhall^{2*} and Jens Meiler^{3,4*}

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

N Sukumar,
Shiv Nadar University, India
Wenyi Wang,
Genentech, Inc., United States

*Correspondence:

Jens Meiler
jens@meilerlab.org
Jeffrey Mendenhall
jeffreymendenhall@gmail.com
Benjamin P. Brown
benjamin.p.brown17@gmail.com

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 10 December 2021

Accepted: 24 January 2022

Published: 21 February 2022

Citation:

Brown BP, Vu O, Geanes AR, Kothiwale S, Butkiewicz M, Lowe EW, Mueller R, Pape R, Mendenhall J and Meiler J (2022) Introduction to the BioChemical Library (BCL): An Application-Based Open-Source Toolkit for Integrated Cheminformatics and Machine Learning in Computer-Aided Drug Discovery. *Front. Pharmacol.* 13:833099. doi: 10.3389/fphar.2022.833099

¹Chemical and Physical Biology Program, Medical Scientist Training Program, Center for Structural Biology, Vanderbilt University, Nashville, TN, United States, ²Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, TN, United States, ³Department of Chemistry, Departments of Pharmacology and Biomedical Informatics, Center for Structural Biology, Vanderbilt University, Nashville, TN, United States, ⁴Institute for Drug Discovery, Leipzig University Medical School, Leipzig, Germany

The BioChemical Library (BCL) cheminformatics toolkit is an application-based academic open-source software package designed to integrate traditional small molecule cheminformatics tools with machine learning-based quantitative structure-activity/property relationship (QSAR/QSPR) modeling. In this pedagogical article we provide a detailed introduction to core BCL cheminformatics functionality, showing how traditional tasks (e.g., computing chemical properties, estimating druglikeness) can be readily combined with machine learning. In addition, we have included multiple examples covering areas of advanced use, such as reaction-based library design. We anticipate that this manuscript will be a valuable resource for researchers in computer-aided drug discovery looking to integrate modular cheminformatics and machine learning tools into their pipelines.

Keywords: drug discovery, drug design, cheminformatics, open-source, deep neural network, QSAR, biochemical library, BCL

INTRODUCTION

Computer-aided drug discovery (CADD) methods are routinely employed to improve the efficiency of hit identification and lead optimization (Macalino et al., 2015; Usha et al., 2017). The importance of *in silico* methods in drug discovery is exemplified by the multitude of cheminformatics tools available today. These tools frequently include capabilities for tasks such as high-volume molecule processing (Hassan et al., 2006; SciTegic, 2007), ligand-based (LB) small molecule alignment (Labute et al., 2001; Jain Ajay, 2004; Chan, 2017; Brown et al., 2019), conformer generation (Cappel et al., 2015; Kothiwale et al., 2015; Friedrich et al., 2017a, 2019), pharmacophore modeling (Hecker et al., 2002; Acharya et al., 2011; Vlachakis et al., 2015), structure-based (SB) protein-ligand docking (Friesner et al., 2004; Meiler and Baker, 2006; Davis and Baker, 2009; Hartmann et al., 2009; Morris

et al., 2009; Kaufmann and Meiler, 2012; Lemmon et al., 2012), and ligand design. Increasingly, modern drug discovery relies on customizable and target-specific machine learning-based quantitative structure-activity relationship (QSAR) and structure-property relationship (QSPR) modeling during virtual high-throughput screening (vHTS) (Lo et al., 2018; Vamathevan et al., 2019).

Frequently, building a drug discovery pipeline with all of these parts requires users to combine multiple different software packages into their workflow. This can be challenging because of different version requirements in package dependencies. Moreover, file- and data-type incompatibilities between packages can lead to errors and pipeline inefficiencies. Here, we describe the BioChemical Library (BCL) cheminformatics toolkit, a freely available academic open-source software package with tightly integrated machine learning-based QSAR/QSPR capabilities.

The BCL is an application-based software package programmed and compiled in C++. This means that BCL applications can be integrated into existing pipelines without the need for package dependency management (i.e., maintaining directory-dependent virtual environments, or keeping separate Miniconda environments for each task). In addition, BCL applications are modular and can be easily combined into complex protocols with simple Shell scripts. Output files from the BCL are primarily common file types that can also be read as input by other software packages. Its command-line usage will be familiar to users of the popular macromolecular modeling software ROSETTA (Kaufmann et al., 2010). The simple command line user interface (UI) makes it easy to create complex protocols without extensive coding or scripting experience. Our goal with this manuscript is to describe the core functionalities of the BCL cheminformatics toolkit and provide detailed examples for real use cases. At the end, we briefly discuss ongoing software developments that may be of interest to users.

MOLECULE PREPARATION AND PROCESSING

Fundamentals of BioChemical Library Command-Line Syntax

The first thing to complete after downloading and installing the BCL is to add the license file to the `/path/to/bcl` folder. We further recommend adding `<path>/<to>/bcl` to the `LD_LIBRARY_PATH` and `PATH` environment variables in the `csrc/.bashrc`. This allows users to access the BCL from any terminal window simply by typing `bcl.exe` into the command-line. For detailed setup instructions, read the appropriate operating system (OS)-specific ReadMe file in `bcl/installer/`.

The BCL is organized into application groups each of which contains multiple applications. To view the application groups and associated applications, run the BCL help command:

`bcl.exe help`

The BCL has application groups for cheminformatics, protein folding, machine learning, and other tasks (**Supplementary Table S1**). To isolate and view the applications associated with the application group molecule, for example, run the application group help command:

`bcl.exe molecule:Help`

Generally, the syntax to access a BCL application is as follows:

`bcl.exe application_group:Application`

The help menu for any application can similarly be accessed as

`bcl.exe application_group:Application --help`

These help options list the basic arguments and parameters available for each application. More detailed help options are also frequently available for individual application parameters. In this way, all of the documentation required to run the BCL can be readily accessed from the command line. The application groups composing the core of the BCL cheminformatics toolkit include the following: Molecule, Descriptor, and Model (**Table 1**).

Filtering

Molecules are input to the BCL in the MDL structure-data format (SDF) file. Often, molecules that are downloaded or converted from one source to another contain errors (e.g., incorrect bond order assignments, undesired protonation states/formal charge, etc.). Dataset sanitization is a critical component of computational chemistry and informatics projects. The BCL molecule: Filter application is the first step in correcting these errors or identifying molecules that cannot be easily and automatically corrected.

To see all of the options available in molecule:Filter, run the following command:

`bcl.exe molecule:Filter--help`

or view the supplementary material (**Supplementary Table S2,S3**).

For the following examples we will make use of a set of the Platinum Diverse Dataset, a subset of high-quality ligands in their protein-bound 3D conformations (Friedrich et al., 2017b).

```
bcl.exe molecule:Filter \
-input_filenames platinum_diverse_dataset_2017_01.sdf.gz \
-output_matched platinum_diverse_dataset_2017_01.matched.sdf.gz \
-output_unmatched platinum_diverse_dataset_2017_01.unmatched.sdf.gz \
-add_h -neutralize \
-defined_atom_types=simple \
-logger File platinum_diverse_dataset_2017_01.Filter.log
```


TABLE 1 | Overview of BCL application groups covered in this manuscript.

| Application Group | Typical Inputs | Typical Outputs |
|-------------------|--|--|
| Molecule | Molecules (.sdf) | Molecules (.sdf) |
| Descriptor | Descriptor sets Molecules (.sdf; GenerateDataset only) Dataset binary file (.bin) Dataset comma-separated file (.csv) | Dataset binary file (.bin) Dataset comma-separated file (.csv) |
| Model | Dataset binary file (.bin) Machine learning model(s) | Machine learning model(s) Predictions |

This command reads in the SDF `platinum_diverse_dataset_2017_01.sdf.gz`, saturates all molecules with hydrogen atoms, neutralizes any formal charges, checks to see whether the molecules have valid atom types (e.g., carbon atoms making five covalent bonds are not valid), and then checks to see whether the molecules have simple connectivity (e.g., whether they are part of a molecular complex, such as a salt). The neutralization flag identifies atoms with formal charge and tries to remove the formal charge. The default behavior allows modification of the protonation state of the atom (i.e., pH) and/or the bond order. Other options (more or less aggressive neutralization schemes) are also available and can be seen in the help menu. Adding hydrogen atoms and neutralizing charges are not required operations but are shown above to demonstrate the functionality.

All molecules that match the filter (i.e., molecules with defined atom types and are not part of molecular complexes) are output into `platinum_diverse_dataset_2017_01.matched.sdf`, and molecules that fail to pass the filters are output into `platinum_diverse_dataset_2017_01.unmatched.sdf`. In this case, all molecules pass the filter. This allows the user to review the molecules that failed the filter and choose to either fix them or continue without them.

The `molecule:Filter` application can also be used to separate molecules by property and/or substructure using the `compare_property_values` flag. For example, to filter out molecules that contain 10 or more rotatable bonds and a topological polar surface area (TPSA) less than 140 \AA^2 , the following command can be used:

```
bcl.exe molecule:Filter \
-input_filenames platinum_diverse_dataset_2017_01.sdf.gz \
-output_matched platinum_diverse_dataset_2017_01.veber_pass.sdf.gz \
-output_unmatched platinum_diverse_dataset_2017_01.veber_fail.sdf.gz \
-compare_property_values TopologicalPolarSurfaceArea less 140 \
NRotBond less_equal 10 \
-logger File platinum_diverse_dataset_2017_01.veber.log
```

Of 2,859 molecules, 395 were first filtered out for have a TPSA $\geq 140 \text{ \AA}^2$, and then an additional 84 molecules that had greater than 10 rotatable bonds were filtered out. Notice that the filters are applied sequentially, and molecules must pass both filters to

be output to the matched file. Alternatively, the any flag can be specified such that if a molecule meets any one of the filter criteria, then it is output to the matched file:

```
bcl.exe molecule:Filter \
-input_filenames platinum_diverse_dataset_2017_01.sdf.gz \
-output_matched platinum_diverse_dataset_2017_01.any_pass.sdf.gz \
-output_unmatched platinum_diverse_dataset_2017_01.any_fail.sdf.gz \
-compare_property_values TopologicalPolarSurfaceArea less 140 \
NRotBond less_equal 10 \
-any -logger File platinum_diverse_dataset_2017_01.any.log
```

In this example, 2,801 molecules passed at least one of the filters and only 58 were filtered out.

One may also filter based on substructure similarity. This is particularly useful if there are specific substructures that are desired or that need to be avoided. For example, aromatic amines are a well-known toxicophore and cannot be incorporated into potential druglike molecules; however, it is not uncommon to find these substructures in datasets. Here, we will filter a subset of DrugBank (Wishart et al., 2018) molecules to remove aniline-containing compounds:

```
bcl.exe molecule:Filter \
-input_filenames drugbank_nonexperimental.simple.sdf.gz \
-output_matched drugbank_nonexperimental.simple.anilines.sdf.gz \
-output_unmatched drugbank_nonexperimental.simple.clean.sdf.gz \
-contains_fragments_from aniline.sdf.gz \
-logger File drugbank_nonexperimental.simple.toxicity_check.log
```

In practice, we usually explicitly filter certain toxicophore substructures via graph search with the `MoleculeTotalToxicFragments` descriptor in conjunction with `compare_property_values` flag; however, this example illustrates the flexibility to filter by substructure similarity with `molecule:Filter`. In addition to the standard use cases presented here, `molecule:Filter` can identify molecules with clashes in 3D space, conformers outside of some tolerance value from a reference conformer, exact substructure matches, specific chemical properties, and more. Some of these filters will be further explored in other subsections.

Removing Redundancy

Another critical aspect of dataset sanitization is removing redundancy. This is especially important when preparing datasets for QSAR model training and testing. If molecules appear more than once in a dataset, then it is possible that they could appear simultaneously in the training and test sets, leading to an artificial inflation in test set performance.

The BCL application `molecule:Unique` can help with this task. It has four levels at which it can compare and differentiate molecules:

1. Constitutions—compares atom identities and connectivity disregarding stereochemistry;
2. Configurations—compares atom identities, connectivity, and stereochemistry;
3. Conformations—compares configurations as well as 3D conformations;
4. Exact—checks to see whether atom identities and order are equal with the same connectivities, bond orders, stereochemistry, and 3D coordinates.

The first time the BCL encounters a molecule in an SDF it will store it in memory. Any additional encounters with the same molecule (at the chosen level described above) will be marked as duplicate encounters. The default behavior is to output only the first encounter of each molecule. There are cases in which a molecule appears multiple times but has different MDL properties and/or property values. It may not be desirable to lose the stored properties on duplicate compounds. In such cases, the user can choose to merge the properties or overwrite the duplicate descriptors instead.

For example, one may want to see if any high-throughput screening (HTS) hits have activity on multiple targets. Previously, we published nine high-quality virtual HTS (vHTS) benchmark sets for QSAR modeling binary classification tasks (Butkiewicz et al., 2013). Here, we will take a look at the active compounds from each of those nine datasets and see if any of them have activity on multiple targets.

```
bcl.exe molecule:Unique \
-input_filenames 1798_actives.sdf.gz 1843_actives.sdf.gz \
2258_actives.sdf.gz 2689_actives.sdf.gz \
435008_actives.sdf.gz 435034_actives.sdf.gz \
463087_actives.sdf.gz 485290_actives.sdf.gz 488997_actives.sdf.
gz \
-compare Constitutions \
-output_dupes all_actives.dupes.sdf.gz \
-logger File all_actives.unique.log
```

The output file `all_actives.dupes.sdf.gz` contains 22 molecules that are active in at least two different datasets (note that each individual dataset was pre-processed to remove redundant molecules). If we want to merge the properties of these 22 compounds and isolate them from the rest of the actives, we can perform a second `molecule: Unique` with the `merge_descriptors` flag set, and then use `molecule:Filter` with the `contains` flag to isolate the duplicated compounds:

```
bcl.exe molecule:Unique \
-input_filenames 1798_actives.sdf.gz 1843_actives.sdf.gz \
2258_actives.sdf.gz 2689_actives.sdf.gz \
435008_actives.sdf.gz 435034_actives.sdf.gz \
463087_actives.sdf.gz 485290_actives.sdf.gz 488997_actives.sdf.
gz \
-compare Constitutions-merge_descriptors \
-output all_actives.unique_merged.sdf.gz \
-logger File all_actives.unique_merged.log
```

followed by

```
bcl.exe molecule:Filter \
-input_filenames all_actives.unique_merged.sdf.gz \
-contains all_actives.dupes.sdf.gz \
-output_matched all_actives.dupes_merged.sdf.gz \
-logger File all_actives.dupes_merged.log
```

When `merge_descriptors` is passed, all unique properties are included in the resultant output file. If the same property is present on duplicates, then the first observation of that property is stored on the output molecule. If `overwrite_descriptors` is passed instead of `merge_descriptors`, the last observation of a duplicate property is stored. By default, without either of these flags only the MDL properties on the first occurrence of a molecule are stored in the output.

It may be that some of the compounds in the previous example that have activity on multiple targets are actually stereoisomers. Here, the molecules were compared based on atom identity and connectivity (Constitutions). Iterative runs of `molecule:Unique` coupled with `molecule:Filter` can be used to identify such cases.

Sorting and Reordering

Sorting molecules is also useful during vHTS. After making predictions on a million compounds with a QSAR model, frequently users will want to identify some small top fraction of most probable hits for experimental testing. This can be readily achieved with `molecule:Reorder` (note—this example utilizes pseudocode for filenames):

```
bcl.exe molecule:Reorder \
-input_filenames < screened_molecules.sdf> \
-output < screened_molecules.best.sdf> -output_max 100 \
-sort <QSAR_Score> -reverse \
-logger File < screened_molecules.best.log>
```

In this example, the `reverse` flag indicates that the scores will be sorted from largest to smallest (default behavior is smallest to largest). Not more than 100 molecules will be output into the file `screened_molecules.best.sdf.gz` because of the `output_max` specification (the default behavior returns all molecules in the new order).

In the previous section, we demonstrated that the BCL could identify duplicate compounds at multiple levels of discrimination. One important note is that redundant molecules are excluded (i.e., sent to the `output_dupes` file) in the order in which they are observed in the original input. Often, the user may want to control this sequence by sorting the molecules according to some property. In these cases, `molecule:Reorder` can be used to do just that.

Finally, a general note on SDF input and output. Aromaticity is automatically detected when reading input files; however, output structures are Kekulized (represented as alternating single-double bonds) by default. To output an SDF that contains explicit aromatic bonds (achieved by labeling bond order as 4 in the MDL SDF), pass the `explicit_aromaticity` flag on the command line.

Making Fragments

The BCL application `molecule:Split` gives researchers a tool to derive fragments from starting small molecules to aid in

TABLE 2 | Fragment splits currently supported by the BCL.

| Molecule Split Implementation | Description | Customizations |
|-------------------------------------|---|---|
| Scaffolds | returns Murcko scaffolds of molecules (Bemis and Murcko, 1996) | None |
| Inverse Scaffold | returns the remaining components of a molecule after the Murcko scaffold is removed (Bemis and Murcko, 1996) | None |
| GADD Fragments | splits molecules into GA-based Drug Database fragments (Daylight Theory: SMILES) | None |
| Largest Common Substructure | splits molecules into their maximum common substructures relative to an input set | level of equivalence of element- and bond- type comparisons |
| ECFP Fragments | splits molecules into radial fingerprint fragments similar to those used for extended connectivity fingerprints (Rogers and Hahn, 2010) | bond distance from each reference atom |
| Linear Fragments | splits molecule into linear non-branching fingerprint fragments similar to Obabel FP2 fingerprints | bond distance from each reference atom |
| Rings | returns all ring components of molecules | None |
| Rings With Unsaturated Substituents | returns ring components of molecules along with their unsaturated substituents | None |
| Unbridged Aromatic Rings | returns unbridged aromatic ring components of molecules | None |
| Unbridged Rings | returns unbridged ring components of molecules | None |
| Chains | returns non-ring (chain) components of molecules | None |
| Rigid | splits a molecule into rigid components; defined by breaking non-ring, non-amide single-bonds to heavy atoms | None |
| Rigid Sans Amide | splits a molecule into rigid components; defined by breaking non-ring, non-amide single-bonds to heavy atoms | None |
| Isolate | splits a molecule with multiple disconnected parts (e.g., salt crystal) into component parts | None |
| Largest | splits a molecule with multiple disconnected parts (e.g., salt crystal) into component part and returns the largest component by molecular weight | None |

pharmacophore modeling, fragment-based drug discovery, and *de novo* drug design. There are many different types of fragments molecule:Split is able generate from whole molecule(s) (Table 2).

For example, we can derive the Murcko scaffold from the FDA-approved 3rd generation tyrosine kinase inhibitor (TKI) osimertinib (Ramalingam et al., 2017) as follows:

```
bcl.exe molecule:Split \
-input_filenames osimertinib. sdf.gz \
-output osi. murcko.sdf.gz \
-implementation Scaffolds
```

Alternatively, we could remove the Murcko scaffold and return the other components:

```
bcl.exe molecule:Split \
-input_filenames osimertinib. sdf.gz \
-output osi. inverse_scaffold.sdf.gz \
-implementation InverseScaffold
```

Substructure comparisons are described in more detail in Section 5.1.

Coordinate Information

The last application of interest for molecule processing is molecule: Coordinates molecule: Coordinates is a minor application that performs several convenience tasks. First, molecule: Coordinates can recenter all molecules in the input file(s) to the origin. Second, it can compute molecular centroids.

Third, molecule: Coordinates can compute statistics on molecular conformers.

For example, passing the statistics flag compute statistics on bond lengths, bond angles, and dihedral angles. Passing the dihedral_scores flag will compute a per-dihedral breakdown of the BCL 3D conformer score. The BCL 3D conformer score, or ConfScore, computes an amide non-planarity penalty in addition to a normalized dihedral score. Passing the amide_deviations and amide_penalties will output the amide deviations and penalties on a per-amide basis, respectively. This can be useful when comparing conformations obtained from conformation sampling algorithms, crystal structures, and/or molecular dynamics (MD) trajectory ensembles. See Section 4 for more information on conformer sampling.

COMPUTING MOLECULAR PROPERTIES

Computing molecular descriptors/properties is a critical component of cheminformatics model building. We use the term “properties” to refer to individual chemical features and “descriptors” to refer to combinations of properties, often used to train QSAR/QSPR models; however, the terms are often used interchangeably in the BCL. In conjunction with substructure-based comparisons, generating molecular descriptors is arguably the foundation of LB CADD. The BCL was designed with a modular descriptor interface and extensible property definitions framework. This allows both developers and users alike to write new descriptors for specific applications as needed. To see a list of

available predefined molecular properties, perform the following command:

```
bcl.exe molecule:Properties--help
```

The property interface is organized into two general categories: 1) Descriptors of Molecules, and 2) Descriptors of Atoms. As you will see throughout this section and **Section 6**, properties can be modified and recombined in a highly customizable fashion. See the **Supplementary Materials** for an example containing multiple custom property definitions, as well as for sample output from the molecule:Properties help menu options detailing available features.

Computing Whole Molecule Properties

As the names suggest, some descriptors are intrinsic to the whole molecule, while others are specific to atoms. For example, compute some whole molecule descriptors for the EGFR kinase inhibitor osimertinib:

```
bcl.exe molecule:Properties \
-input_filenames osimertinib. sdf.gz \
-output osi. mol_properties.sdf.gz \
-add Weight NRotBond NRings TopologicalPolarSurfaceArea \
-tabulate Weight NRotBond NRings TopologicalPolarSurfaceArea \
-output_table osi. mol_properties.table.txt
```

The flag add will add the specified properties to the SDF as MDL properties. The tabulate flag will output the properties for each molecule in row-column format in the file specified by output_table. There is also a statistics flag that will compute basic statistics for each of the specific descriptors across all the molecules in the input SDFs and output to output_histogram. The key observation regarding the output file is that the values for Weight, NRotBond, etc., are emergent properties of the whole molecule.

Computing Atomic Properties

Next, compute some atomic descriptors for osimertinib:

```
bcl.exe molecule:Properties \
-add_h--neutralize \
-input_filenames osimertinib. sdf.gz \
-output osi. atom_properties.sdf.gz \
-add Weight Atom_SigmaCharge Atom_TopologicalPolarSurfaceArea \
-tabulate Atom_SigmaCharge Atom_TopologicalPolarSurfaceArea \
-output_table osi. atom_properties.table.txt \
-statistics Atom_SigmaCharge Atom_TopologicalPolarSurfaceArea \
-output_histogram osi. atom_properties.hist.txt
```

Notice here that the statistics flag outputs statistics across each atom property rather than across each molecule property. This is also the behavior when there are multiple input molecules. Importantly, here we see that the output is an array of values for each property. The indices of the array correspond to the atom indices of the molecule.

Performing Operations on Descriptors

Each category of descriptors can further be modified by molecule-specific or atom-specific operations. For example, some whole molecule properties can be obtained by performing simple operations on the per-atom properties. TopologicalPolarSurfaceArea (whole molecule property) is the sum of Atom_TopologicalPolarSurfaceArea (atomic property) across the whole molecule.

```
bcl.exe molecule:Properties \
-add_h--neutralize \
-input_filenames osimertinib. sdf.gz \
-output osi. mol_properties.sdf.gz \
-add TopologicalPolarSurfaceArea \
"MoleculeSum (Atom_TopologicalPolarSurfaceArea)"
```

Check to verify that TopologicalPolarSurfaceArea and MoleculeSum (Atom_TopologicalPolarSurfaceArea) yield the same value for osimertinib.

Examples of additional operations include other basic statistics (mean, max, min, standard deviation, etc.), property radial distribution function (RDF), Coulomb force, and shape moment. See the help menu for additional options and details.

Combining Properties to Evaluate Druglikeness

In **Section 2.2** we discussed using the molecule:Filter application to remove molecules from a dataset that failed specific druglikeness criteria (e.g., TPSA $\geq 140 \text{ \AA}^2$). Several familiar druglikeness metrics come prepackaged in the BCL (i.e., Lipinski's Rule of 5 and Veber's Rule), as well as several others inspired by the literature and conventional medicinal chemistry practices. For each molecule in the Platinum Diverse dataset, count how many Lipinski and Veber violations there are. In addition, count as drug-like all molecules that have fewer than two Lipinski violations:

```
bcl.exe molecule:Properties \
-input_filenames platinum_diverse_dataset_2017_01. sdf.gz \
-output_table platinum_diverse_dataset_2017_01. druglike.txt \
-tabulate LipinskiViolations LipinskiViolationsVeber LipinskiDruglike
```

The property LipinskiViolations counts how many times a molecule violates one of Lipinski's Rules (≤ 5 hydrogen bond donors (HBD; $-\text{NH}$ and $-\text{OH}$ groups), ≤ 10 hydrogen bond acceptors (HBA; any $-\text{N}$ or $-\text{O}$), molecular weight (MW) < 500 Daltons, and water-octanol partition coefficient ($\log P$) < 5). The LipinskiViolationsVeber property computes the number of times a molecule violates Veber's Rule (infraction if TPSA $\geq 140 \text{ \AA}^2$ and/or number of rotatable bonds > 10). The LipinskiDruglike property is a Boolean that returns 1 if fewer than two Lipinski violations occur; 0 otherwise. There is no equivalent Boolean operator for Veber druglikeness; however, it is simple to implement one using the aforementioned operators.


```
bcl.exe molecule:Properties \
-input_filenames platinum_diverse_dataset_2017_01.sdf.gz \
-output_table platinum_diverse_dataset_2017_01.veber_druglike.
txt \
-tabulate "Define [VeberDruglike = Less (lhs =
LipinskiViolationsVeber, rhs = 1)]" VeberDruglike
```

This command makes use of the Define and Less operators to return 1 if there are no violations to Veber's Rule and 0 otherwise. New properties created with Define can also be passed to subsequent operators on the same line. For example, one could create a descriptor called VeberAndLipinskiDruglike by doing the following:

```
bcl.exe molecule:Properties \
-input_filenames platinum_diverse_dataset_2017_01.sdf.gz \
-output_table platinum_diverse_dataset_2017_01.veber_druglike.
txt \
-tabulate \
"Define [VeberDruglike = Less (lhs = LipinskiViolationsVeber,
rhs = 1)]" \
"Define [VeberAndLipinskiDruglike = Multiply (LipinskiDruglike,
VeberDruglike)]" \
VeberAndLipinskiDruglike
```

This new descriptor returns 1 if a molecule passes both druglikeness filters, and 0 otherwise.

Many metrics can be created using the BCL descriptor framework without modifying the source code. This can be useful to users who come across novel methods in the literature and wish to implement them in their own work. Take as an example a seminal work from Bickerton et al., which sought to quantify the chemical aesthetics of potential druglike compounds. Bickerton et al. asked 79 medicinal chemists at AstraZeneca to answer "would you undertake chemistry on this compound if it were a hit?" for ~200 compounds each, to which chemists replied either "yes" or "no" (Bickerton et al., 2012). They generated a regression function that yielded a quantitative estimate of druglikeness (QED) using eight chemical descriptors: molecular weight, logP, number of hydrogen bond acceptors, number of hydrogen bond donors, polar surface area, number of rotatable bonds, number of aromatic rings, and number of ALERTS (Bickerton et al., 2012).

Using the same dataset and descriptors as Bickerton et al. (generously provided in their **Supplemental Materials**), similar druglikeness metrics can be implemented in the BCL through the descriptor framework. One approach could be to use the operators described above to reproduce the algebraic expression described in Eq. 1 of Bickerton et al. with the parameters described in their **Supplemental Materials**. The algebra expressed in BCL notation can be saved to an external text file and passed to the command-line using standard shell script syntax (e.g., @File.txt in Bash). Because there are relatively few descriptors in the Bickerton et al. model, an alternative approach could be to create a classification model.

Here, we demonstrate the latter by (Eq. 1) generating a decision tree (DT) model and then 2) converting our DT into

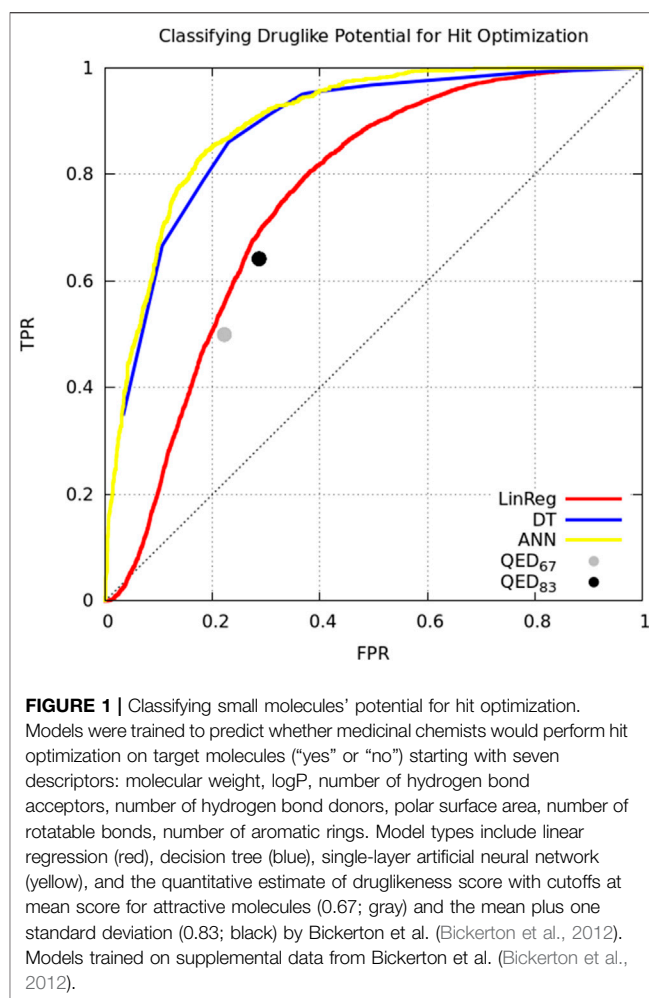


FIGURE 1 | Classifying small molecules' potential for hit optimization. Models were trained to predict whether medicinal chemists would perform hit optimization on target molecules ("yes" or "no") starting with seven descriptors: molecular weight, logP, number of hydrogen bond acceptors, number of hydrogen bond donors, polar surface area, number of rotatable bonds, number of aromatic rings. Model types include linear regression (red), decision tree (blue), single-layer artificial neural network (yellow), and the quantitative estimate of druglikeness score with cutoffs at mean score for attractive molecules (0.67; gray) and the mean plus one standard deviation (0.83; black) by Bickerton et al. (Bickerton et al., 2012). Models trained on supplemental data from Bickerton et al. (Bickerton et al., 2012).

a single logic statement to pass to the BCL descriptor interface. For comparison, we also generate linear regression (LinReg) and artificial neural network (ANN) models, and we include the original QED score. All models are trained to predict a chemist's verdict for each potential compound based on the descriptors used in Bickerton et al. (for details on model training and validation, see **Supplementary Methods**; for details on how to build machine learning models with the BCL, see **Section 7**).

Model classification performance is displayed as a receiver-operating characteristic (ROC) curve (**Figure 1**). Bickerton et al. found that the mean QED score for molecules that medicinal chemists found favorable was 0.67 (± 0.16 standard deviation). Taking the mean and mean plus standard deviation QED scores as cutoffs, we see that QED performs comparably to multiple linear regression. The ANN and DT perform better, but perhaps owing to the small number of and simple relation between variables there is no performance benefit of the ANN over the DT (**Figure 1**).

Now that we have our DT, we can reduce it to a readable if-else style format that can be converted into a BCL descriptor. Run the

script `SimplifyDecisionTree.py`, passing as an argument the DT model:

```
/path/to/bcl/scripts/machine_learning/analysis/
SimplifyDecisionTree.py \
/models/DT/model000000. model > DT. logic_summary.txt
```

We can see in the contents of `DT. logic_summary.txt` that the first thing the DT checks is whether the small molecule has less than two aromatic rings. Molecules with no aromatic rings are excluded, and molecules with one aromatic ring are subject to different criteria than molecules with two or more. Subsequent criteria are then evaluated. We can rewrite the logic summary as a descriptor and save it in a file called “`dt.obj`”. Then, we pass that file to `molecule:Properties` as a descriptor definition and use it to classify molecules:

```
bcl.exe molecule:Properties \
-add_h -neutralize \
-input_filenames platinum_diverse_dataset_2017_01. sdf.gz \
-output_table platinum_diverse_dataset_2017_01. dt_druglike.
txt \
-tabulate “Define (Hitlike = @dt.obj)” Hitlike
```

The “`dt.obj`” code object file is a plain text file that can be opened with any text editor. The syntax mimics the BCL command-line syntax. Code object files are a convenient way to write a long, multi-line BCL command-line that makes it easier to build and reuse feature sets.

On the topic of druglikeness, it is worth noting that additional advanced methods are also available to classify the chemical space of molecules in a dataset. In some cases, it is useful to identify potential drug-like compounds that not only fit the criteria discussed above but are also similar to some known class (es) of drugs. For example, when performing fragment-based combinatorial library design for kinase inhibitors, in addition to filtering out molecules that violate Veber’s rules, it may also be desirable to filter molecules that are not sufficiently chemically similar to existing kinase inhibitors. This can be accomplished by building and scoring against an applicability domain (AD) model. For further details on creating and using AD models in the BCL, see **Section 7.4.3**.

We have described multiple uses of the `molecule:Properties` application, placing special emphasis on how it can be utilized to build different types of druglikeness metrics. As it is fundamentally a tool to obtain information from small molecule chemical structures, `molecule:Properties` can also be used to help generate statistical potentials, chemical filters, QSAR/QSPR models, and more. Some of these use-cases will be explored in later sections.

SMALL MOLECULE CONFORMER GENERATION

Small molecule 3D conformer generation is a critical aspect of both SB and LB CADD because the biologically relevant

conformation of the molecule of interest is rarely known *a priori*. In SB molecular docking, small molecule flexibility is often represented through the inclusion of multiple discrete pre-generated conformers (Brylinski and Skolnick, 2008; Morris et al., 2009; Lemmon and Meiler, 2012; Combs et al., 2013; DeLuca et al., 2015). Small molecule conformations need to be sampled to arrive at the correct binding pose. Molecules that appear in binding pockets of substantially different proteins often bind in distinct modes for each protein, suggesting that the binding pose of the molecule need not be near the global energy minima of the molecule (Nicklaus et al., 1995; Boström et al., 1998; Perola and Charifson, 2004; Sitzmann et al., 2012; Friedrich et al., 2018). Likewise, in LB pharmacophore modeling, small molecules need to be flexibly aligned according to their chemical properties to identify the biologically relevant 3D features conferring bioactivity.

The BCL conformer generator, also called BCL:Conf, utilizes a fragment-based rotamer library derived from the crystallography open database (COD) to combine rotamers consisting of one or more dihedral angles according to a statistically-derived energy (Mendenhall et al., 2020). Clashes are dynamically resolved by iteratively identifying clashed atom pairs and rotating the central-most bonds between them without changing dihedral bins. In this way, conformational ensembles are stochastically generated according to likely rotamer combinations from the COD.

The BCL small molecule conformation sampler is a leader among general purpose small molecule conformer generation algorithms (Kothiwale et al., 2015; Mendenhall et al., 2020). In this section, we demonstrate how to use the BCL to generate global and local conformational ensembles and sample discrete rotamers within a molecule.

Generating Global Conformational Ensembles

Start by generating conformers of osimertinib with the default settings. Here, all that is needed is an input filename and an output filename:

```
bcl.exe molecule:ConformerGenerator \
-ensemble_filenames osimertinib. sdf.gz \
-conformers_single_file osimertinib. global_confs.sdf.gz
```

The `ensemble_filenames` argument is equivalent to the `input_filenames` argument used elsewhere (the difference is historical). The `conformers_single_file` argument is one of two output options. The other option is `conformers_separate_files`. As implied by the name, in the former case all conformers are output to a single file. In the latter case, if multiple molecules are input to `ensemble_filenames`, then a unique SDF will be written for the conformational ensembles of each of the input molecules [e.g., if the input SDF(s) contained 10 molecules, then `conformers_separate_files` would output 10SDFs each with a conformational ensemble of one of the input molecules].

By default, BCL:Conf will perform 8,000 conformer generation iterations, each of which rebuilds the molecule essentially from scratch (excepting rigid ring structures and bonds that do not

vary substantially in length or angle). Without any other options, the top conformations will be clustered, yielding the 100 best-scoring representatives of each different cluster. An unbiased view of the conformational space around the ligand can be obtained by setting the `skip_cluster` flag. For this application, it is advisable to lower the number of iterations to roughly double the number of desired conformations; the conformers are rebuilt from scratch at every iteration, so there is little gain from doing more iterations than conformers desired. The returned conformers are sorted by score. Number of iterations and final conformers can be specified with the `max_iterations` and `top_models` flags, respectively.

Conformations can be filtered to remove highly-similar conformations using the `conformation_comparer` flag (e.g., to standard RMSD, dihedral distance, etc.) and the tolerance for what constitutes an “identical” conformer increased from the default (0.0) to an arbitrarily large value (note that RMSD- and dihedral-based metrics have units of Å and degrees, respectively) (Kothiwal et al., 2015). For most applications, we recommend the use of SymmetryRMSD with a modest tolerance of 0.25 Å. By default, the tolerance is adjusted automatically to yield the desired number of clusters so as to best represent conformational space, however, a user-provided tolerance is treated as a minimal acceptable difference between clusters.

For high-throughput applications, we recommend reducing iterations from 8,000 down to 800 or even 250. BCL:Conf's speed is nearly linear in number of iterations. Generally, more iterations yield better performance, at a trade-off of slightly-faster than linear increase in time per conformation when clustering is used (Mendenhall et al., 2020).

Alternatively, if `conformation_comparer` is set to “RMSD 0.0”, then no filtering or clustering is specified, and BCL:Conf will perform `max_iterations` conformer generation iterations, randomly select `top_models` conformers, sort them from best to worst by score, and return them. This option is the fastest, and the ensembles returned are arguably the most Boltzmann-like. For a recent comparison of each set of parameters to one another and other conformer generation algorithms, please see Mendenhall et al., 2020 (Mendenhall et al., 2020). We recommend generating conformers with explicit hydrogen atoms added.

Generate conformers using two of the protocols described protocols. First, run

```
bcl.exe molecule:ConformerGenerator \
-add_h -ensemble_filenames osimertinib. sdf.gz \
-conformers_single_file osimertinib. symrmsd_cluster.conf.
sdf.gz \
-max_iterations 8,000 -top_models 25 \
-conformation_comparer SymmetryRMSD 0.25
```

Then,

```
bcl.exe molecule:ConformerGenerator \
-add_h -ensemble_filenames osimertinib. sdf.gz \
-conformers_single_file osimertinib. raw.conf.sdf.gz \
-max_iterations 8,000 -top_models 250 -skip_cluster
-conformation_comparer RMSD 0.0
```

Notice that the ensemble generated with the SymmetryRMSD comparer and clustering enabled occupies the densest part of the broader conformational space sampled in the raw distribution.

Generating Local Conformational Ensembles

Local sampling was implemented in the recent algorithmic improvements to BCL:Conf (Mendenhall et al., 2020). The idea is that sometimes users know or have predicted with some degree of certainty a chemically meaningful or bioactive pose of a small molecule, but additional refinement is needed. This is a common use case when modeling protein-ligand complexes starting with another ligand with some similarity to the ligand of interest (Bozhanova et al., 2021; Hanker et al., 2021). When using pre-generated conformers for docking or small molecule flexible alignment, it is unlikely that the best ligand conformer will be chosen and simultaneously have its position fully optimized in Cartesian space. Local sampling around an input conformer allows the user to refine ligand poses after an initial search.

Local sampling in the BCL is accomplished by restricting the rotamer search in one of four ways:

1. `-skip_rotamer_dihedral_sampling`—preserve input dihedrals to within 15-degrees of closest 30-degree bin (centered on 0°) in non-ring bonds.
2. `-skip_bond_angle_sampling`—preserve input conformer bond lengths and angles
3. `-skip_ring_sampling`—preserve input ring conformations
4. `-change_chirality`—by default, input chirality and isometry are preserved. Use this flag to allow for generation of enantiomers and stereoisomers.

These options are not mutually exclusive. Depending on how they are combined, different levels of sampling can be achieved. Moreover, they can be used in combination with any of the other options (e.g., conformation comparison type, clustering) described above. Generate local conformational ensembles of osimertinib by first placing all three restrictions:

```
bcl.exe molecule:ConformerGenerator \
-ensemble_filenames osimertinib. sdf.gz \
-conformers_single_file osimertinib. skip_all.local_conf.sdf.
gz \
-skip_rotamer_dihedral_sampling
-skip_bond_angle_sampling \
-skip_ring_sampling-skip_cluster
```

Next, apply only the `skip_rotamer_dihedral_sampling` and `skip_bond_angle_sampling` restrictions to generate a local ensemble:

```
bcl.exe molecule:ConformerGenerator \
-ensemble_filenames osimertinib. sdf.gz \
-conformers_single_file osimertinib. skip_dihed_ring.local_conf.
sdf.gz \
```

```
-skip_rotamer_dihedral_sampling
-skip_ring_sampling-skip_cluster
```

Both of the ensembles show less conformational diversity than the global conformational ensemble created in the previous section. Notice the relative sampling differences between each of the local conformation sampling protocols described.

Conformational Sampling of Substructures

Often times one may wish to only sample conformations of part of a molecule. For example, in docking congeneric ligand series, the core scaffold pose may be known with a high degree of confidence, and the goal is to optimize the pose of the rest of the molecule while keep the core scaffold fixed. Alternatively, crystal structures of protein-ligand complexes often have low or missing density for part of a bound ligand, and thus coordinate assignment may not be accurate. Discretely sampling specific small molecule rotamers thus becomes a useful task to perform.

In the BCL, this is accomplished by first assigning an MDL miscellaneous property named “SampleByParts” to the molecule(s) of interest. The value of the SampleByParts property corresponds to the 0-indexed atom indices of atoms in dihedrals that are allowed to be sampled by molecule:ConformerGenerator. By encoding this as a molecule-specific property, we avoid multiple command-line calls with different atom index specifications, allowing users to generate conformers more rapidly for multiple molecules and/or different independent rotamers within a molecule.

As an example, consider a crystal structure of epidermal growth factor receptor (EGFR) kinase in complex with osimertinib (PDB ID 4ZAU) (Yosaatmadja et al., 2015). This is the first publicly available crystal structure of the EGFR-osimertinib complex. In this structure, the solvent-exposed ethyldimethylamine substituent is missing density. We will sample alternative conformations of the ethyldimethylamine substituent than that which is proposed in the PDB ID 4ZAU. First, add the corresponding atom indices to the file osimertinib.sdf:

```
bcl.exe molecule:Properties \
-add "Define [SampleByParts = Constant (3,36,18,19,6,20,21)]"
SampleByParts \
-input_filenames osimertinib.sdf.gz-output \
osimertinib.sample_by_parts.sdf
```

Also, note that if you have many molecules for which you want to assign SampleByParts atom indices and you do not want to have to manually identify the relevant indices, you can also use the molecule:SetSampleByPartsAtoms application. This application sets SampleByParts indices based on comparison to user-supplied substructures. With the SampleByParts property defined in the SDF, generate global conformers as previously described:

```
bcl.exe molecule:ConformerGenerator \
-ensemble_filenames osimertinib.sample_by_parts.sdf.gz \
-conformers_single_file osimertinib.sample_by_parts.confs.sdf
gz \
-top_models 250 -cluster
```

Observe that sampling global conformers (i.e., sampling across dihedral bins allowing bond angle/length adjustment and ring conformer changes) with SampleByParts maintains the coordinates of all unspecified atoms. In this case, only dihedrals containing strictly the ethyldimethylamine atoms are sampled (Figure 2). Similarly, SampleByParts can be used in conjunction with the local sampling methods described above.

MOLECULE PROPERTY- AND SUBSTRUCTURE-BASED COMPARISONS

A critical component of LB CADD is molecular similarity analysis. Provided a set of molecules, we frequently want to know how similar each molecule is to a reference molecule(s). Fundamentally, this requires 1) defining what specifically will be compared between the molecules, and 2) defining the metric with which similarity will be measured. In the BCL, this is accomplished primarily through use of the molecule:Compare application. The command-line syntax of molecule:Compare differs from the syntax of other applications discussed so far. The SDF input files to molecule:Compare are passed as parameters instead of argument flags.

```
bcl.exe molecule:Compare < mandatory_parameter_one.sdf> \
<optional_parameter_two.sdf> -output < mandatory_output.file> \
```

This syntax strictly enforces two types of behavior:

1. If a single SDF is specified as a parameter, then all molecules in the file are compared with one another
2. If two SDFs are specified, then the molecule(s) in the second file will be compared against the molecule(s) in the first file.

Finally, it is worth noting that molecule:Compare’s performance scales approximately linearly with number of threads for costly comparisons. To enable threads, set -scheduler PThread <number_threads>. We suggest setting number_threads to number of physical cores on the device for maximum performance.

Defining Molecular Structures

The BCL encodes molecules as graphs where the edges are bonds, and the atoms are nodes. For substructure-based comparisons, we can define equivalence between bonds and atoms using various comparisons dubbed comparison types. For any substructure-based comparison between two or more molecules, some combination of atom and bond comparison types is required, which defines the equivalence class for the atoms and bonds, respectively. The default combination differs between tasks. For a summary of available atom and bond type comparisons, examine the help menu options of any comparer that utilizes substructures. For example,

```
bcl.exe molecule:Compare \
-method "LargestCommonSubstructureTanimoto (help)"
```

will display the default atom and bond comparison types for this comparison method as well as list the available comparison

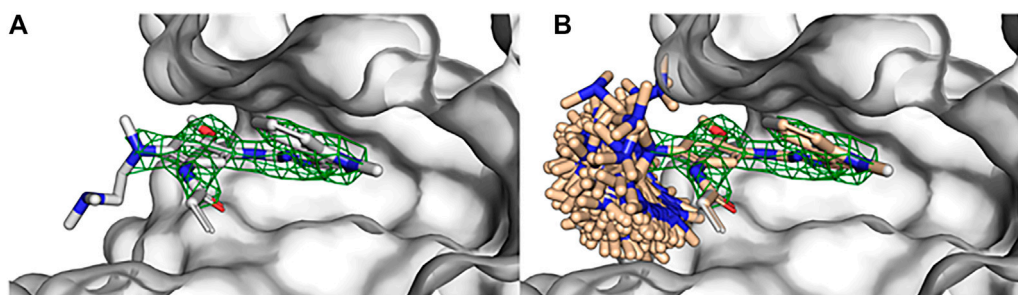


FIGURE 2 | Substructure sampling of small molecule rotamers with BCL:Conf. **(A)** Crystallographic structure of osimertinib bound to EGFR kinase (PDB ID 4ZAU) contains missing density of the ethyldimethylamine substituent of osimertinib. **(B)** Global conformational sampling of the osimertinib ethyldimethylamine substituent without perturbing the rest of the bound pose using BCL:Conf. Osimertinib electron density visualized with green mesh by importing the 2fo-fc map in PyMOL and contouring at 2σ .

types. For example, if atom type resolution occurs at AtomType, then an SP3 carbon would match another SP3 carbon but not an SP2. If the resolution is lowered to ElementType, then all carbon atoms can match one another independent of their orbital configuration. Similarly, bond type resolutions of BondOrder and BondOrderAmideWithIsometryOrAromaticWithRingness will yield different substructure matches.

Not all similarity comparisons occur at the structural/substructural level. A number of comparison metrics in the BCL occur between properties computed at the whole molecular, substructural, or atomic level. Further, distance-based comparisons between molecules that are constitutionally identical can also be made.

Similarity Scoring Between Constitutionally Unique Molecules

In cases where the similarity between unique molecules is desired there are broadly two approaches for measuring similarity: by substructure and by property. These are not mutually exclusive; depending on the desired resolution of the substructure comparisons, one can further measure property differences between substructures of different molecules.

One common substructure similarity metric is the Tanimoto coefficient (TC), expressed between two molecules as the ratio of matched-to-unmatched atoms:

$$TC = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \quad (1)$$

where A and B are the two molecules. The intersection of atoms in (Eq. 1) is the size of the largest common substructure under the specified comparison types. This is a specific formalism of the more general Tversky index when both α and β are equal to 1:

$$TC = \frac{|A \cap B|}{|A \cap B| + \alpha|A - B| + \beta|B - A|}, \quad (2)$$

The first-generation EGFR tyrosine kinase inhibitor gefitinib and the second-generation inhibitor afatinib are structurally very similar. Afatinib is modified from the gefitinib scaffold and

incorporates an acrylamide linker. Visualize the maximum common substructure (MCS) of afatinib and gefitinib using molecule:Split (Figure 3):

```
bcl.exe molecule:Split \
-implementation "LargestCommonSubstructure (file =
afatinib.sdf)" \
-input_filenames gefitinib.sdf.gz-output mcs_gef_afa.sdf.gz
```

Next, calculate the MCS TC of the gefitinib and afatinib:

```
bcl.exe molecule:Compare gefitinib.sdf.gz afatinib.sdf.gz \
-method LargestCommonSubstructureTanimoto-output
gef_afa_mcs_tani.txt
```

This method searches for the single largest common connected substructure as the intersection of two molecules and computes the TC. In this case, the MCS TC is approximately 0.48. Sometimes searching for a single connected substructure can be disadvantageous. For example, if the primary differences between molecules results from core substitutions bridging two otherwise identical halves, then the single largest common substructure approach will fail to account for the complete degree of similarity. Alternatively, the user can calculate the maximum common disconnected substructure (MCDS) TC:

```
bcl.exe molecule:Compare gefitinib.sdf.gz afatinib.sdf.gz \
-method
LargestCommonDisconnectedSubstructureTanimoto \
-output gef_afa_mcdis_tani.txt
```

As expected, the MCDS TC is greater than the MCS TC at approximately 0.86.

Distance-Based Scoring Between Constitutionally Identical Molecules

In Section 4 we demonstrated how the BCL can be used to generate small molecule conformational ensembles. One common way to measure the performance of small molecule

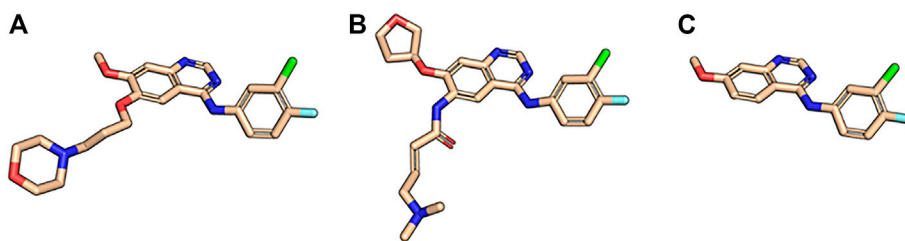


FIGURE 3 | Maximum common substructure between gefitinib and afatinib. **(A)** Afatinib (PDB ID 4G5J) and **(B)** gefitinib (PDB ID 4I22) in their binding mode 3D conformations next to **(C)** their maximum common substructure extracted with the BCL.

conformer generators is to measure how close we can recover biologically relevant conformations. We can do this in the BCL by measuring the RMSD or SymmetryRMSD of molecules in our conformational ensemble to the experimentally determined conformations. Generate a global ensemble of osimertinib:

```
bcl.exe molecule:ConformerGenerator \
-add_h -ensemble_filenames osimertinib. sdf.gz \
-conformers_single_file osimertinib. confs.sdf.gz \
-max_iterations 8,000 -top_models 50 -cluster \
-conformation_comparer SymmetryRMSD 0.25 -generate_3D
```

Note that we are generating the molecule completely *de novo* ignoring all information from input coordinates by using generate_3D. Measure the heavy-atom symmetric RMSD to the native conformation:

```
bcl.exe molecule:Compare osimertinib. sdf.gz osimertinib.
confs.sdf.gz \
-method SymmetryRMSD -logger File osi. sym_rmsd_native.log \
-output osi. sym_rmsd_native.txt -remove_h
```

On examination of osi. sym_rmsd_native.txt, we see that see that of our 25 generated conformers, 3 are less than 2.0 Å from the native conformer, and the best is approximately 0.66 Å from native. If we repeat this process for two additional TKIs, the first-generation inhibitor erlotinib and the second-generation inhibitor afatinib, we also see that we are able to obtain multiple conformers less than 1.0 Å from native.

In addition to RMSD-based metrics, molecule:Compare can also measure distance in the form of dihedral angle sums and dihedral distance bins. For additional information, examine the help menu options.

Largest Common Substructure Alignment

The BCL can be used to align small molecules according to their MCS. Unlike most of the examples in this section, this is accomplished through the molecule:AlignToScaffold application by passing three parameters:

```
bcl.exe molecule:AlignToScaffold <scaffold> <ensemble> <output>
```

For example, to align afatinib to gefitinib based on their MCS, use the following command:

```
bcl.exe molecule:AlignToScaffold gefitinib. sdf.gz afatinib.
sdf.gz \
afatinib.ats.sdf.gz \
```

Instead of aligning by MCS, the user may also align the target ensemble to the largest rigid component of the scaffold structure by passing the align_rigid flag. Moreover, if the user wants to define an alternative set of atoms to be aligned instead of the defaults, this can be accomplished by specifying those atoms for each the scaffold and target ensemble with align_scaffold_atoms and align_ensemble_atoms, respectively.

Property-Based Flexible Alignment

In addition to substructure-based alignment, we can also perform property-based alignment. Property-based alignment algorithms typically maximize the overlap or minimize the distance between molecular and/or atomic properties (Sliwoski et al., 2014). We have previously demonstrated that the performance of the BCL property-based alignment algorithm, also referred to as BCL: MolAlign, is on par with leading academic and commercial molecular alignment algorithms (Brown et al., 2019).

BCL:MolAlign combines the conformational sampling ability of BCL:Conf with the property framework described in Section 3 to minimize the property-distance between two molecules through flexible superimposition. The property-distance is computed between mutually-matching atom pairs that are dynamically updated with each iteration. Alignment pose sampling is accomplished through a series of moves that traverse the co-space defined by the relative position of the two molecules to one another (Brown et al., 2019). BCL: MolAlign can be used to perform alignments which can be classified as rigid (two molecules with fixed conformers), semi-flexible (one molecule with a fixed conformer, one molecule whose conformers are sampled), and fully-flexible (two molecules whose conformers are sampled).

To demonstrate how BCL:MolAlign can be used to perform each of these alignments, consider the classic problem of obtaining the crystallographic alignment of methotrexate (MTX) and dihydrofolic acid (DHF). This example is a good one because the intuitive heterocyclic overlap is not the correct one (Labute et al., 2001). Instead, alignment of the binding pockets of dihydrofolate reductase (DHFR) co-crystallized with MTX (PDB ID 1DLS) and DHF (PDB ID 1DHF) shows only partial heterocycle overlap and superimposition of the

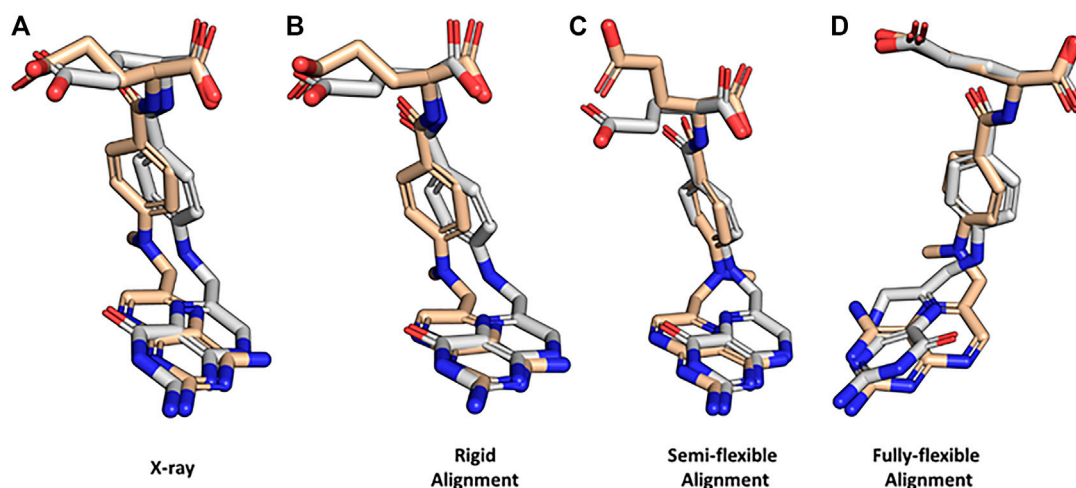


FIGURE 4 | Property-based alignment of dihydrofolic acid and methotrexate with BCL:MolAlign. **(A)** Superimposed crystallographic structures of dihydrofolic acid (DHF; PDB ID 1DHF) and methotrexate (MTX; PDB ID 1DLS) in complex with dihydrofolate reductase (DHFR). **(B)** Rigid alignment of DHF and MTX starting from the bioactive conformers from the crystal structures. **(C)** Flexible alignment of MTX (flexible) to DHF (rigid, bioactive conformer). **(D)** Fully flexible alignment of DHF and MTX. DHF is colored white and MTX is colored wheat. MTX was randomly rotated and translated prior to rigid alignment to DHF. All flexible alignments performed using generate_3D to remove bias from start coordinates.

heterocycle carbonyl in DHF and an aromatic hydrogen bond accepting nitrogen in MTX (**Figure 4A**). Perform a rigid alignment of MTX to DHF with the following command:

```
bcl.exe molecule:Compare mtx. perturbed.sdf.gz dhf. sdf.gz \
-add_h-neutralize \
-output mtx_dhf_rigid_rmsdx.output \
-logger File rigid_alignment.log -random_seed \
-method "PsiField \
(
output aligned mol a = mtx. rigid_aligned.sdf,
iterations = 1,000,
number outputs = 1
)"
```

The rigid alignment ranks the correct alignment mode as the top scoring alignment (**Figure 4B**). Rigid alignments are rarely useful for drug discovery because the bioactive conformation of the target small molecule is usually unknown; however, they provide a useful check for alignment scoring functions. Next, flexibly align MTX to the DHFR-binding pose of DHF:

```
bcl.exe molecule:Compare mtx. perturbed.sdf.gz dhf. sdf.gz \
-add_h-neutralize \
-output mtx_dhf_rigid_rmsdx.output \
-logger File semi-flexible_alignment.log -random_seed \
-method "PsiFlexField \
(
output_aligned_mol_a = mtx. semiflex_aligned.sdf,
rigid_mol_b = true,
number_flexible_trajectories = 3,
fraction_filtered_initially = 0.25,
```

```
fraction_filtered_iteratively = 0.50,
iterations = 400,
filter_iterations = 200,
refinement_iterations = 50,
conformer_pairs = 500,
number_outputs = 1,
sample_conformers = SampleConformations (
conformation_comparer = SymmetryRMSD,
generate_3D = 1,tolerance = 0.10,rotamer_library = cod,
max_iterations = 8,000,max_conformations = 50,
cluster = true)
)"
```

Here, we can see that BCL:MolAlign correctly determines the alignment of the heterocycles, central aromatic rings, and (partially) the acidic groups (**Figure 4C**). Note that rigid_mol_b is enabled, which fixes the pose of the second parameter molecule. For a detailed description of how each argument modifies the alignment algorithm, see Brown et al. (Brown et al., 2019). For performance considerations, we generally find that the number of conformer pairs is more critical to pose recovery than the numbers of iterations at each stage. For complex ligands with many rotational bonds, we recommend increasing max_conformations and conformer_pairs.

Fully-flexible alignment is useful when one is trying to recover pharmacophore features without knowing the binding pose of either molecule. Here, the goal is to align pharmacophore features of the molecules, not recover the native pose of the target molecule(s) by aligning to another molecule with a known binding mode. Perform a fully-flexible alignment of MTX and DHF.

```

bcl.exe molecule:Compare mtx. perturbed.sdf.gz dhf.
perturbed.sdf.gz \
-add_h-neutralize \
-output mtx_dhf_rigid_rmsdx.output \
-logger File fully-flexible_alignment.log \
-random_seed-scheduler PThread 8 \
-method "PsiFlexField \
( \
output_aligned_mol_a = mtx-dhf. fullflex_aligned.sdf, \
rigid_mol_b = false, \
number_flexible_trajectories = 5, \
fraction_filtered_initially = 0.25, \
fraction_filtered_iteratively = 0.50, \
iterations = 800, \
filter_iterations = 400, \
refinement_iterations = 100, \
conformer_pairs = 2,500, \
number_outputs = 1, \
sample_conformers = SampleConformations ( \
conformation_comparer = SymmetryRMSD, \
generate_3D = 1,tolerance = 0.10,rotamer_library = cod, \
max_iterations = 8,000,max_conformations = 50, \
cluster = true) \
)"

```

Fully-flexible alignment of MTX and DHF does not recover the most native-like conformations of MTX and DHF; however, it does recover correct alignments of the heterocycles, central aromatic rings, and acidic groups (**Figure 4D**). Notice that we increased the number of conformer pairs from 500 to 2,500 when we went from semi-flexible to fully-flexible alignment.

FEATURE GENERATION

The descriptor application group is the workhorse for molecule featurization. Similar to the molecule:Properties application, the descriptor application group provides command-line access to the internal descriptor framework. Unlike molecule, descriptor is dataset centric; its primary purpose is to generate, manipulate, and analyze feature datasets for QSAR/QSPR. In this section, we will demonstrate core applications in descriptor and how they can be utilized in QSAR/QSPR modeling.

Generating Simple Datasets From Molecules

Four specifications are required to generate feature datasets from small molecules:

1. The molecules for which to generate the features; these can be any valid SDF.
2. The types of features to generate; these are properties such as those described in **Section 3**. Typically, these are stored in a separate file and passed to the command-line at run-time; however, they can also be specified directly on the command-

line. Importantly, combining multiple descriptors for feature generation requires the use of the Combine descriptor.

3. The feature result label; this indicates the output(s) that models will train toward. This can be a constant value (i.e., if featurization is being done for some purpose other than model training), a property (e.g., LogP for a QSPR model), or another label (e.g., bioactivity label from experimental data).
4. The output filename; three output types are available. The BCL has a partial binary format with the ".bin" suffix that is used for all model training. Feature datasets can also be output with the ".csv" suffix for a comma-separated values (CSV) file. Moreover, ".csv" files and ".bin" files can be interconverted. In this way, features generated with the BCL can be used by other software, and vice versa. For inter-operability with Weka software, ".arff" format is also supported, with a limitation of only working with continuous variables.

Generate a simple feature dataset consisting of several scalar descriptors for a set of confirmed active M1 Muscarinic Receptor positive allosteric modulators (PAMs) and corresponding true negatives (Butkiewicz et al., 2013). The SDF corresponding to these compounds is 1798. combined.sdf. These molecules have been labeled with the MDL property "IsActive" such that the confirmed actives have a value of 1 and the negatives have a value of 0.

```

bcl.exe descriptor:GenerateDataset \
-source "SdfFile (filename = 1798. combined.sdf)" -id_labels
"String (M1)" \
-result_labels "Combine (IsActive)" \
-feature_labels "Combine (Weight, LogP,HbondDonor,
HbondAcceptor)" \
-output 1798. combined.scalars.bin

```

Binary files were designed for rapid non-consecutive reading and writing, but the interested reader will find that the file format consists of a textual header specifying the properties and their sizes followed by a simple binary output of all features. Dataset information and statistics can be obtained by calling descriptor:GenerateDataset compare. For example:

```

bcl.exe descriptor:GenerateDataset-compare 1798.
combined.scalars.bin

```

To better understand the binary file encodings, convert 1798. combined.scalars.bin to a CSV file:

```

bcl.exe descriptor:GenerateDataset \
-source "Subset (filename = 1798. combined.scalars.bin)" \
-output 1798. combined.scalars.csv

```

The first column of every row contains the ID label "M1" as specified when the binary file was generated. The next four columns contain the descriptors specified above: Weight, LogP, HbondDonor, and HbondAcceptor. The very last

column is the result value, which contains either 0 or 1 depending on the value in the SDF MDL property “IsActive”.

Convert CSV file back to a binary file:

```
bcl.exe descriptor:GenerateDataset \
-source "Csv(filename = 1798. combined.scalars.csv, number
result cols = 1, number id chars = 2)" \
-output 1798. combined.scalars.bin
```

CSV files do not contain all of the supplementary information contained within the partial binary file format. Thus, certain information needs to be provided directly. For example, we need to specify the number of characters that are part of the row ID label, otherwise the BCL will try to convert the string (or numerical) ID into feature values. ID labels therefore must be fixed-width. In addition, we need to tell the BCL how many of the columns are result values. By default, the BCL will assume that only the last column is the result label. By specifying number result cols = N, we tell the BCL to take the last N columns of the CSV as the result value(s).

Also notice that the feature and result label information is not informative after converting from CSV to binary. The values are transferred to the new file format, but the BCL obviously cannot know where those values came from. These must be manually specified.

```
bcl.exe descriptor:GenerateDataset \
-source "Csv(filename = 1798. combined.scalars.csv, number
result cols = 1, number id chars = 2)" \
-id_labels "String (M1)" \
-result_labels "Combine (IsActive)" \
-feature_labels "Combine (Weight, LogP, HbondDonor,
HbondAcceptor)" \
-output 1798. combined.scalars.bin
```

In this case, the feature labels are internal parsable properties of the BCL; however, when relabeling feature labels upon converting from CSV to binary format, the user can specify any labels so long as the total number of labels is consistent with the number of feature columns.

Modifying Datasets

After generating a dataset or importing a CSV file and converting it to binary format, feature datasets can be modified. The most frequent form of modification is randomization. Training a machine learning model, for example a neural network, often requires dataset randomization.

```
bcl.exe descriptor:GenerateDataset \
-source "Randomize [Subset (filename = 1798. combined.scalars.
bin)]" \
-output 1798. combined.scalars.rand.bin
```

Binary files are read by the “Subset” retriever. The Randomize operator is passed through the source flag and provided the dataset retriever option corresponding to the binary file.

Additional dataset operators can be classified by how they modify the dataset. For example, the PCA (principal components analysis) and EncodeByModel operators perform dimensionality reduction across feature (column) space, while the KMeans operator reduces dimensionality across molecule (row) space. Other operators are useful during model training and validation, such as Balanced, Chunks, and YScramble. Still others can be used to select particular ranges of rows from a dataset, such as Rows. Here, we will take a look at a few dataset operators. For full details on all available dataset operators, see the descriptor: GenerateDataset help menu.

Start by generating a dataset for the Kir2.1 inward rectifying potassium channel using the dataset compiled in Butkiewicz et al. (Butkiewicz et al., 2013) and the best performing LB descriptor set from Mendenhall and Meiler (Mendenhall and Meiler, 2016). This dataset contains 301,493 small molecules, 172 of which are confirmed active molecules. For each molecule, there will be 1,315 feature columns and 1 result column.

```
bcl.exe descriptor:GenerateDataset \
-source "SdfFile (filename = 1843. combined.sdf.gz)"
-scheduler PThread 8 \
-feature_labels MendenhallMeiler2015. Minimal.object \
-result_labels "Combine (IsActive)" \
-output 1843. Minimal.bin-logger File 1843. Minimal.log
```

Randomize the dataset:

```
bcl.exe descriptor:GenerateDataset \
-source "Randomize (Subset (filename = 1843. combined.bin))" \
-output 1843. combined.rand.bin-logger File 1843. Minimal.
rand.log
```

Note that we could have generated a randomized dataset with a single command by wrapping the SdfFile dataset retriever with Randomize; however, the Randomize dataset retriever is unable to support hyperthreading. Consequently, it is faster to generate larger datasets first using multiple threads and randomize them afterward. Next, perform PCA on the dataset using OpenCL to accelerate the calculation with a GPU. The flag `opencl` is optional and may not be supported on all platforms, but may provide a substantial speedup, depending on the GPU and dataset size:

```
bcl.exe descriptor:GeneratePCAEigenVectors \
-training "Subset (filename = 1843. Minimal.rand.bin)" \
-output_filename 1843. Minimal.PCs.dat-opencl \
-logger File 1843. Minimal.PCs.log
```

Finally, generate a new feature dataset accounting for 95% of the variance:

```
bcl.exe descriptor:GenerateDataset \
-source "PCA(dataset = Subset (filename = 1843. Minimal.
rand.bin), fraction = 0.95, filename = 1843. Minimal.PCs.
dat)" \
-output 1843. Minimal.rand.pca_095. bin-opencl \
-logger File 1843. Minimal.rand.pca_095. log
```

Performing PCA on the dataset has reduced the number of descriptors from 1,315 to 695. Alternatively, one could use EncodeByModel to reduce the number of feature columns using a pre-generated model. The following example utilizes pseudocode and a hypothetical pre-generated ANN with the MendenhallMeiler2015.Minimal.object features.

```
bcl.exe descriptor:GenerateDataset \
-source "EncodeByModel [storage = File (directory = /path/to/
model/directory, prefix = model),retriever = Subset
(filename=<my_binary_file.bin>)]" \
-output < my_encoded_binary_file.bin>
```

The input file < my_binary_file.bin > would have 1,315 descriptors from MendenhallMeiler2015.Minimal.object, and the output file < my_encoded_binary_file.bin > would have a number of descriptors corresponding to the number of neurons in the final hidden layer preceding the output layer of our hypothetical pre-generated ANN.

As a practical note, we have found that PCA-based dimensionality reduction useful for dataset visualization, but of limited value in improving model performance. Performance can often be recovered to that of the initial dataset when requiring at least 95% of the variance to be preserved, but performance improvement is rare from PCA, when using a regularized method such as dropout-ANNs.

Suppose you encoded the same original feature set using two different models and now want to combine the new encoded files for further training. This can readily be accomplished with the Combine operator.

```
bcl.exe descriptor:GenerateDataset \
-source "Combined [Subset (filename=<my_binary_file_1.
bin>), Subset (filename=<my_binary_file_2. bin>)]" \
-output < my_combined_binary_file.bin>
```

Next, instead of performing dimensionality reduction along the column (features) axis, we will reduce the dimensionality along the row (molecule) axis. Perform K-means clustering of the feature dataset to reduce our row number from 301,493 to 300.

```
bcl.exe descriptor:GenerateDataset \
-source "KMeans [dataset = Subset (filename = 1843.
combined.rand.bin), clusters = 300]" \
-output 1843. combined.rand.k300. bin \
-logger File 1843. combined.rand.k300. log
```

This form of dimensionality reduction is unlikely to be as useful for training a deep neural network (DNN); however, it can be useful in similarity analysis in low dimensional feature space. Some of the datasets generated in this section will be referenced again in **Section 7** to train classification machine learning models.

Small Molecule Autocorrelation Descriptors

As indicated in the previous section, the BCL can also compute signed autocorrelation functions. Autocorrelations are regularly used as features in cheminformatics machine learning models

(Sliwoski et al., 2014). When computed for atomic descriptors, such as Atom_SigmaCharge, the autocorrelations sum pairwise property products into distance bins by calculating the separation between molecule atom pairs in number of bonds (2DA) or Euclidean distance (3DA). Each distance bin is further separated into three sign-pair bins corresponding to property value sign of each atom in the pair (**Eq. 3**) (Sliwoski et al., 2015).

$$A(r_a, r_b) = \sum_j \sum_i^N \delta(r_a \leq r_{i,j} < r_b) P_i P_j, \quad (3)$$

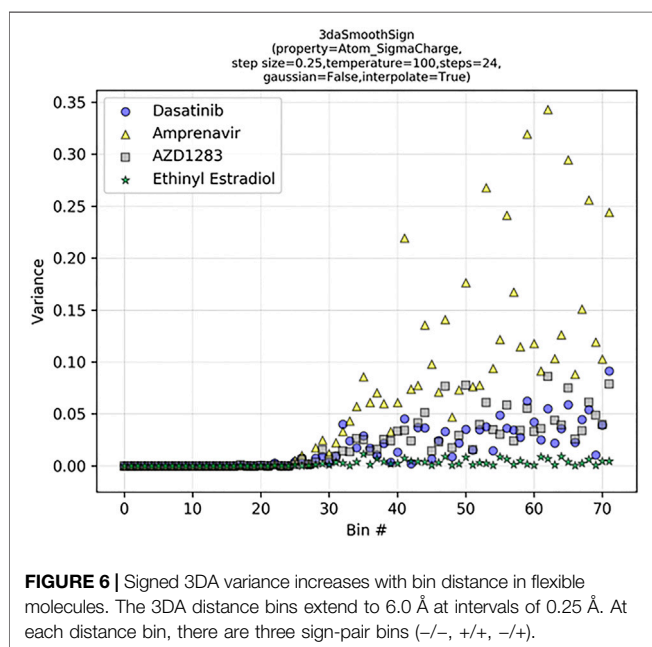
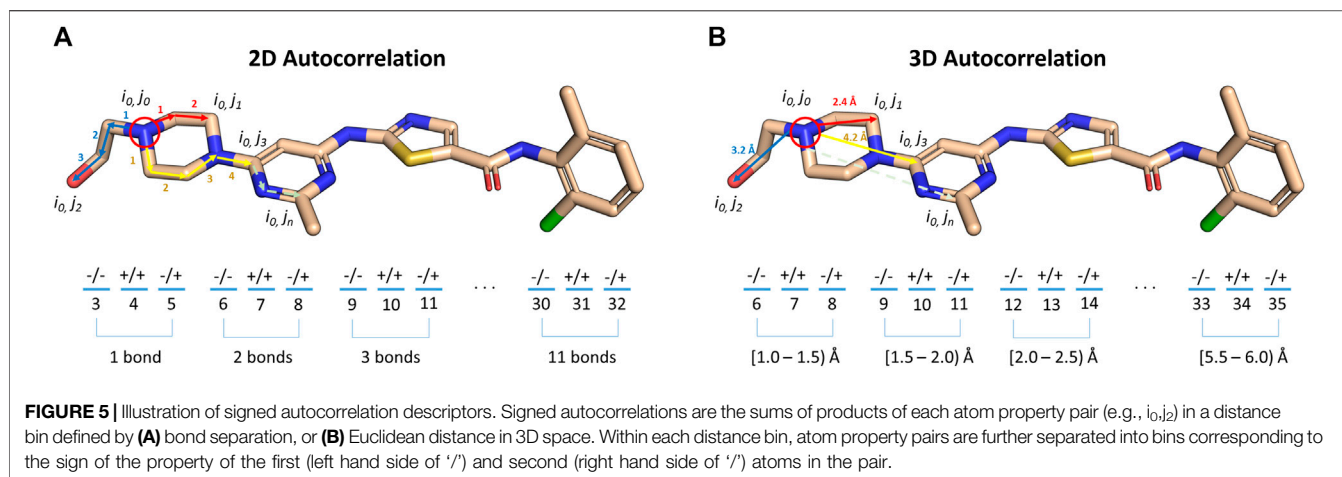
where r_a and r_b are the boundaries of the current distance interval, N is the total number of atoms in the molecule, $r_{(i,j)}$ is the distance between the two atoms being considered, δ is the Kronecker delta, and P is the property computed for each atom. 2DAs are conformation-independent, while 3DAs are conformation-dependent (**Figure 5**).

The "dasatinibs.sdf" file contains the coordinates and connectivity for two dasatinib molecules: one with 2D coordinates, the other with 3D coordinates. Compute the signed 2DA and 3DA for Atom_SigmaCharge on both dasatinib molecules.

```
bcl.exe descriptor:GenerateDataset \
-source "SdfFile (filename = dasatinibs.sdf)" \
-feature_labels "Combine (3daSmoothSign (property = Atom_
SigmaCharge))" \
-result_labels "Combine [Constant (999)]" -output dasatinibs.
3da.csv \
-logger File dasatinibs.3da.log
bcl.exe descriptor:GenerateDataset \
-source "SdfFile (filename = dasatinibs.sdf)" \
-feature_labels "Combine [2DASmoothSign (property = Atom_
SigmaCharge)]" \
-result_labels "Combine [Constant (999)]" -output dasatinibs.
2da.csv \
-logger File dasatinibs.2da.log
```

Upon examination of the tabulated 2DA and 3DA values for the two different dasatinib molecules, we observe that the 2DA contains the same values in both cases, while the 3DA contains unique values for the different conformers. To visualize the variance in each 3DA distance bin, we can tabulate the 3DAs for Atom_SigmaCharge on an ensemble of 3D conformations for several different molecules (**Figure 6**). Dasatinib is a TKI with 7 rotatable bonds, amprenavir is a HIV protease inhibitor with 12 rotatable bonds, AZD1283 is an antagonist of the P2Y12 receptor with 9 rotatable bonds, and ethinyl estradiol is a synthetic estradiol with only 1 rotatable bond that binds and activates estrogen receptors.

We can see that the variance in each descriptor column increases as a function of distance and number of rotatable bonds. In ethinyl estradiol there is little change in descriptor column variance as a function of distance. In contrast, molecules with increasing numbers of rotatable bonds display increasingly large variances at longer distance bins. This suggests that increasing conformational heterogeneity at longer distance bins leads to increased noise. Indeed, we have previously



found that extending LB 3DAs beyond approximately 6.0 Å generally results in reduced performance on QSAR classification tasks (Sliwoski et al., 2015), consistent with our example here (Figure 6). Importantly, however, at shorter distances where there is less conformational heterogeneity we are able to improve our performance with 3DAs even when the active conformation of the small molecule is unknown (Sliwoski et al., 2015; Mendenhall and Meiler, 2016). Moreover, models making predictions on molecules that are fairly rigid (e.g., steroid derivatives) may benefit from longer range distance bins.

It is also possible to use molecule:Properties to tabulate and compute statistics for molecules instead of plotting the CSV file data from descriptor:GenerateDataset. Here, we used descriptor:GenerateDataset to illustrate its usage. In practice, we do not just use a single 3DA or 2DA, but instead build sets of descriptors for feature and result

labels and store them as separate code object files. As mentioned previously, the code object file format is the same format as allowed on the command line.

MACHINE LEARNING ARCHITECTURES AND APPLICATIONS

The BCL supports multiple machine learning algorithms for QSAR/QSPR modeling. Among the methods available are ANNs (including DNNs and multitasking neural networks) (Dahl, 2014; Bharath et al., 2015; Mendenhall and Meiler, 2016; Xu et al., 2017), support vector machines (SVM) (Kawai et al., 2008; Ma et al., 2008; Mariusz et al., 2009), Kohonen networks (KN) (Kohonen, 1990; Korolev et al., 2003; Wang et al., 2005), restricted Boltzmann machines (RBM) (Le Roux and Bengio, 2008; Tijmen Tieleman, 2008), and decision trees (DT) (Mariusz et al., 2009; Sheridan, 2012; Butkiewicz et al., 2013). GPU acceleration is available for ANNs and SVMs through OpenCL (Munshi, 2008). The primary application group for machine learning in the BCL is model. To see the applications within model, check the help menu:

```
bcl.exe model:Help
```

Overview of BioChemical Library Model Training and Validation

Here, we will first introduce the user to the overall workflow involved in training, analyzing, and subsequently testing BCL machine learning models. The basic workflow for model training is the same for each machine learning method and can be completed via the model:Train application. To see the available machine learning methods, access the help options within model:Train.

```
bcl.exe model:Train --help
```

TABLE 3 | Machine learning model types.

| Model Name | Description |
|--|---|
| Applicability Domain Kohonen | A Kohonen map-based implementation to detect whether a point is within the applicability domain of a model. All nodes will use the same spline for computing applicability. This implies an assumption that the model in question has the most difficulty predicting things far from any node center, regardless of which node center it is |
| Applicability Measure Kohonen | A Kohonen map-based implementation to detect whether a point is within the applicability domain of a model. All nodes will have their own distance metric, which is valid if the model is capable of distinguishing between classes of features (e.g., if the model in question is a Kohonen map itself) |
| Decision Tree | A decision tree trained using one of several methods to partition feature indices |
| Kappa Nearest Neighbor Kohonen | A k-nearest-neighbor predictor; iteration optimizes k |
| Leverage | A Kohonen-network based predictor |
| Linear Regression | Computes the leverage matrix (projection or hat matrix), which allows identification of significant outliers that would likely substantially influence any simple linear model of system. A returned value >2 represents probable outliers, while greater than 3 represent definitive outliers. The average value is 1 for all values in the training set |
| Multiple Output Support Vector Machine | Performs multiple linear regression |
| Neural Network | A support vector machine with multiple outputs using sequential-minimal-optimization |
| Restricted Boltzmann Machine | A neural network with many customizable hyperparameters (e.g., hidden layer count, layer size, dropout type and fraction, transfer function, initialization with pre-generated models, learning rate, weight update/backpropagation scheme, etc.) |
| Support Vector Machine | A restricted Boltzmann machine neural network |
| | A support vector machine trained using sequential-minimal-optimization |

As of this writing, the available model types can be found in **Table 3**. The most reliable way to see available model types is via the help menu options of your version of the BCL.

To expose all options for a particular machine learning method, pass the algorithm name as the first parameter to the application with the help menu request:

```
bcl.exe model:Train "<training algorithm>(help)"
```

The following is a typical command-line format to train a model beginning with a pre-generated descriptor binary file:

```
bcl.exe model:Train < training algorithm> \
-max_minutes < maximum time of training in minutes> \
-max_iterations < maximum number of training iterations> \
-final_objective_function < performance metrics for model
evaluation> \
-feature_labels < names of descriptors> \
-training < training set> \
-monitoring < monitoring set> \
-independent < independent set> \
-storage_model < location in which to store the model> \
-opengl < enables GPU acceleration> \
-logger File < log file>
```

Model performance is evaluated with the user-specified objective function. The choice of objective function is typically related to the task being performed (e.g., classification vs regression) (**Table 4**).

BCL model:Train is designed to readily enable cross-validation. The application is flexible with respect to serialization of model predictions for each of the monitoring, independent, and training partitions as well as writing of the model itself. For example, in five-fold cross-validation, the dataset is split into five chunks. For each round of cross-

validation, the model is trained on four-fifths of the dataset, and the other fifth “independent” set is left out for testing. One of the chunks can additionally be specified as the monitoring dataset. The monitoring dataset can be used for early termination of the model training session to prevent overtraining (early termination is largely deprecated in favor of dropout to prevent earlier termination; we demonstrate it here to illustrate the syntax).

The initial dataset set is split into monitoring, independent, and training partitions with model:Train by assigning chunks with the dataset retriever responsible for binary format files, Subset. In the following pseudocode example, we will set the options to divide the training set into the following five chunks (0-indexed): chunks one to four will be used as the training set, and chunk 0 will be used as both the monitoring set and the independent set (this is appropriate only if the monitoring dataset is not being used for early termination).

```
-training "Subset (number chunks = 5, chunks = [1, 4],
filename=<my_dataset.bin>)"
-monitoring "Subset (number chunks = 5, chunks = [0],
filename=<my_dataset.bin>)"
-independent "Subset (number chunks = 5, chunks = [0],
filename=<my_dataset.bin>)"
```

Dataset partitioning is repeated for each round of cross-validation until each chunk takes a turn as the independent set. Then, the predictions of all the test sets are pooled together by the model:PredictionMerge application:

```
bcl.exe model:PredictionMerge \
-input_model_storage 'File (directory = /path/to/models/
,prefix = model)' \
-output < output_pooled_predictions>
```


TABLE 4 | Objective functions for machine learning models.

| Name | Prediction task | Formula |
|--|-----------------|--|
| Accuracy | Classification | $Accuracy = \frac{TP+TN}{P+TN}$ |
| AUC (Area under the receiver operating characteristic curve) | Classification | $TPR = \frac{TP}{FN+TP}$ $FPR = \frac{FP}{TN+FP}$ $AUC = \int TPR d(FPR)$ |
| LogAUC | Classification | $\log AUC = \frac{\int_{0.001}^{0.1} TPR d(\log(FPR))}{\int_{0.001}^{0.1} d(\log(FPR))}$ |
| MCC (Matthew's correlation coefficient) | Classification | $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ |
| PPV (Positive predictive value) | Classification | $PPV = \frac{TP}{TP+FP}$ |
| Enrichment factor | Classification | $EF(x\%) = \frac{PPV(x\%)}{PPV(100\%)}$ |
| MAE (Mean absolute error) | Regression | $MAE = \frac{1}{N} \sum_i f(x_i) - y_i $ |
| MAE_NMAD (MAE normalized by the mean absolute deviation) | Regression | $MAE_{NMAD} = \frac{MAE}{\frac{1}{N} \sum_i y_i - \bar{y} }$ |
| RMSD (Root-mean-square deviation) | Regression | $RMSD = \sqrt{\frac{1}{N} \sum_i (f(x_i) - y_i)^2}$ |
| NRMSD (RMSD normalized by the range) | Regression | $NRMSD = \frac{RMSD}{\max(y) - \min(y)}$ |
| RMSD_NSTD (RMSD normalized by the standard deviation) | Regression | $RMSD_{NSTD} = \frac{RMSD}{\text{Stdev}(y)}$ |

This command line averages predictions made on the same independent set, though other pooling operations are available (see help). Prediction performance is evaluated with the specified objective function on the pooled predictions using the model: ComputeStatistics application:

```
bcl.exe model:ComputeStatistics \
-input < output_pooled_predictions> \
-obj_function < performance_metric> \
-filename_obj_function < output_performance_metric_file>
```

Simplifying the Model Training and Validation Framework in Practice

To simplify model training, we have written a Python script “launch.py” to perform training and cross-validation with one command.

To see a list of model training operations (descriptor selection or scoring, for example):

```
/path/to/bcl/scripts/machine_learning/launch.py-h
```

To see the list of available flags for cross-validation, call

```
/path/to/bcl/scripts/machine_learning/launch.py-t cross-validation-h
```

The following pseudocode example generates a simple linear regression model:

```
/path/to/bcl/scripts/machine_learning/launch.py -t cross-validation \
--cross-validation 5 --local \
--learning-method LinearRegression (objective function = RMSD, \
```

```
solver = Cholesky (smoothing = 0)) \
--id linear_regression --final-objective-function RMSD \
--datasets < my_dataset.bin > --override-memory-multiplier: 1.25
```

More complex commands can be easily prepared inside of a configuration file to be passed to the “launch.py” script. A sample configuration file is available in the **Supplementary Material**.

```
bcl/trunk/scripts/machine_learning/launch.py-t
cross_validation \
--config-file config. example.ini
```

The “launch.py” script will automatically generate three new directories titled “log_files”, “results”, and “models”. Into each of those three directories a labeled directory (name specified with the id flag) is made. Model prediction output files and results of the final objective function are stored in the labeled directory within the “results” folder. Log files, commands, and autogenerated scripts are stored in the labeled directory within the “log_files” folder. Finally, final model details are stored in the labeled directory within the “models” folder.

In addition to running the training jobs locally, training can be run on a SLURM cluster using the slurm flag. In this way, large cross-validation jobs may leverage high-performance computing with minimal changes to the configuration. See additional configuration operations, such as slurm-host, using launch.py-t cross-validation-h.

Applying Models to Independent Test Sets for Virtual High-Throughput Screening

Note that in the above examples the training and test splits are derived from the same binary format file. This is not strictly

necessary, and the user can supply alternatively derived validation splits prepared in separate files. Moreover, using a dataset split as the independent test set is generally only useful for model validation. To apply trained model predictions to new molecules in a vHTS, either `model:Test` or `molecule:Properties` can be used. For example, if a model is trained and validated using five-fold cross-validation, then the merged prediction on an external test set can be made as follows with `model:Test`:

```
bcl.exe model:Test \
-retrieve_dataset "Subset (filename=<vHTS.test.bin>)" \
-storage_model "File (directory = /path/to/models/,prefix = model)" \
-average output < vHTS.model_test.csv> -logger File < vHTS.model_test.log>
```

Likewise, predictions can be made with `molecule:Properties` using the Prediction operators:

```
bcl.exe molecule:Properties-input_filenames < vHTS.test.sdf> \
-tabulate \
"Define {predicted_activity = PredictionMean [storage = File (directory = /path/to/models/,prefix = model)]}" predicted_activity \
"Define {local_ppv = PredictionInfo [predictor = File (directory = /path/to/models/,prefix = model),metrics (LocalPPV)]}" local_ppv \
"Define {XActive = Multiply [predicted_activity, Greater (lhs = local_ppv,rhs = 0.50)]}" XActive \
-output_table < vHTS.prop.test.csv> -logger File < vHTS.prop.test.log>
```

Notice that scoring new compounds via `molecule:Properties` allows multiple outcome metrics to be reported and modified on-the-fly, while scoring with `model:Test` just outputs the raw prediction values (and optionally just the mean with average). In this case, the output of `model:Test` is equivalent to “predicted_activity” from `molecule:Properties`. The property “XActive” is the “predicted_activity” score when the local PPV is greater than 0.5, and 0.0 otherwise. The localPPV metric calibrates model output values to local classification probability on the test sets. It is an estimate of the PPV at a singular model output value. This is in contrast to traditional PPV, which specifies the value of a prediction at, or above, a given output value (assuming positive parity). This metric assumes that the trained model prediction value varies monotonically with the actual prediction likelihood.

Supervised Learning

Training a Standard Artificial Neural Network to Classify Kir2.1 Positive Allosteric Modulators

ANNs are one of the most commonly employed classes of non-linear classifiers in QSAR modeling for LB-CADD due to their strong predictive power (Dahl, 2014; Xu et al., 2017; Vamathevan et al., 2019). To see all the options available to a neural network in the BCL, call

```
bcl.exe model:Train "NeuralNetwork (help)"
```

The BCL supports shallow and deep single- and multi-tasking neural networks. Transfer functions include linear, sigmoid, rectified linear, and leaky rectified linear. For a network with L hidden layers indexed $l \in (1 \dots L)$, forward propagation for $l \in (0 \dots L - 1)$ can be described as

$$\mathbf{z}^{(l+1)} = \mathbf{w}^{(l+1)} \mathbf{y}^l + \mathbf{b}^{(l+1)}, \quad (4)$$

$$\mathbf{y}^{(l+1)} = f(\mathbf{z}^{(l+1)}) \quad (5)$$

where \mathbf{y}^l is the output vector at layer l connected to the input vector $\mathbf{z}^{(l+1)}$ at layer $l+1$ by weights \mathbf{w} and biases \mathbf{b} , and f is the transfer function applied to each set of inputs into the $l+1$ layer. Correspondingly, the activation of a single neuron i in hidden layer $l+1$ can be represented as

$$z_i^{(l+1)} = \mathbf{w}_i^{(l+1)} \mathbf{y}^l + b_i^{(l+1)}, \quad (6)$$

$$y_i^{(l+1)} = f(z_i^{(l+1)}) \quad (7)$$

to yield the output $y_i^{(l+1)}$ from layer $l+1$. We have found that for classical QSAR tasks a simple mean-squared error (MSE) cost function is adequate.

Historically, overtraining in ANNs has been prevented by early termination of training when the monitoring dataset improvement rate or improvement scores fail to progress beyond a pre-determined extent. More recently, we have demonstrated that dropout is a better alternative to prevent model overtraining in QSAR tasks (Mendenhall and Meiler, 2016). The dropout approach has been described elsewhere in detail (Nitish et al., 2014). Briefly, during forward propagation each layer of the ANN is assigned a probability p according to which the output value y_i^l of each i neuron in the layer l will be independently set to zero (i.e., “dropped”).

$$z_i^{(l+1)} = \mathbf{w}^{(l+1)} (\mathbf{r}^l * \mathbf{y}^l) + \mathbf{b}_i^{(l+1)}, \quad (8)$$

Here, \mathbf{r}^l is a vector with the same dimensions as \mathbf{y}^l whose values are either 0 (at fraction p) or 1 (at fraction $1-p$) and multiplied elementwise by the values in \mathbf{y}^l . At the end of every training batch, \mathbf{r}^l is shuffled. If neurons are dropped with a probability p , then at test time the corresponding weights are scaled down by the factor $1-p$.

Train a shallow (single hidden layer) neural network to classify molecules as either active or inactive PAMs of Kir2.1 beginning with the randomized dataset we generated in **Section 6.2**:

```
launch.py -t cross-validation --local \
--datasets 1843.combined.rand.bin --id 1843.ann.1x32_005_025 \
--config-file config.example.ann.ini \
```

The configuration file specifies the learning method as follows:

```
learning-method: 'NeuralNetwork ( \
transfer function = Sigmoid, \
weight update = Simple (alpha = 0.50,eta = 0.05), \
dropout (0.05,0.25), \
```

```
objective function = % (objective-function)s, \
scaling = AveStd, steps per update = 1, hidden architecture (32), \
balance = True, balance target ratio = 0.10, \
shuffle = True, input dropout type = Zero \
)'
```

Note that we are asking for an ANN with one hidden layer composed of 32 neurons. The input and hidden layers will have 5 and 25% dropout, respectively. In addition, we have enabled class balancing. We have far fewer active (172) than inactive (301,321) compounds. Balancing oversamples the underrepresented (minor) class to achieve a ratio of (in this case) 0.10 with the most common class (major). The balance max repeats flag can also be set to specify the maximum number of times that a feature can be repeated. This does not lead to overtraining because of dropout. Batch size is controlled with the steps per update flag. The objective-function variable is defined in the configuration file as

```
"AucRocCurve (cutoff = 0.5,parity = 1,x_axis_log = 1, min fpr = 0.001, max fpr = 0.1)"
```

Additional variables, such as the maximum number of training iterations (20), number of rounds of cross-validation (5), monitoring dataset (independent set), etc. are also set in the configuration file.

As a comparison, train an additional ANN with the same parameters using the feature set whose dimensions were reduced with PCA in **Section 6.2**:

```
launch.py -t cross-validation --local \
--datasets 1843. combined.rand.pca_095. bin \
--id 1843. pca_095. ann.1x32_005_025 \
--config-file config. example.ann.ini \
```

The “launch.py” pipeline automatically generates a ROC curve for each model with and without a log scaled x-axis (**Figure 7**). The overall AUC is quite similar between the two methods (**Figures 7B,D**); however, the model trained with the PCA descriptors has worse early enrichment (logAUC = 0.39) than the model trained with the full descriptors (logAUC = 0.46) (**Figures 7A,C**).

Training a Deep, Multitasking Neural Network to Predict Solubility

Predicting physicochemical properties such as solubility is a challenging but critical component of lead compound optimization. Many substitutions to a candidate molecule may increase the potency or selectivity, but at the cost of worsening solubility, metabolic stability, or other properties. Therefore, it is advantageous to prioritize synthesis and evaluation of derivatives that are simultaneously predicted to be active and have a promising chemical profile. To do this, we need a target-agnostic QSPR model.

Dahl and colleagues demonstrated that multitask learning could improve the prediction of multiple outputs simultaneously if the training tasks are correlated (Dahl, 2014;

Xu et al., 2017). As an example of how such a model is trained with the BCL, we will train a deep neural network to simultaneously predict three measures relating to solubility: the water-octanol partition coefficient (logP), the aqueous solubility (logS), and the hydration free energy (i.e., the solvation free energy in water; $\Delta G_{\text{hydration}}$). Note that not the descriptors, model architecture, nor hyper-parameters have been optimized for performance. This can be seen as an “out of the box” model a user might create.

Molecules for training and validation are sourced from previously published databases (Syracuse Research Corporation, 1994; Edward W.; Lowe et al., 2011; Mobley and Guthrie, 2014; Wu et al., 2018) and combined with BCL molecule: Unique to remove redundant compounds (see **Supplementary Methods** for details). Note that we anticipate some additional error in predictions introduced by not averaging replicate experimental measurements of QSPR properties prior to removing redundancy. Generate three datasets: One with all of the unique compounds (Full), another that contains only those compounds with all three result labels (Dense), and one that contains all of the compounds minus those with all three result labels (Full-Dense). The following command generates the feature set for all of the compounds with three result labels encoded by MDL property labels:

```
bcl.exe descriptor:GenerateDataset \
-source "SdfFile (filename = all_logp_logs_dgsolv.sdf.gz)" \
-feature_labels VuMendenhallMeiler2019. Scalar_Mol2D.object \
-result_labels "Combine (LogP_actual, LogS_actual,dG_hydration_kcal-mol)" \
-output all_logp_logs_dgsolv.Scalar_Mol2D.bin \
-logger File all_logp_logs_dgsolv.Scalar_Mol2D.log \
-scheduler PThread 8 -compare
```

To generate the Dense feature set, add the `forbid_incomplete_records` flag. The two binary format files should contain 35,874 and 448 rows, respectively, and the third dataset should contain the difference between them, 35,426. The distribution of result values overlaps reasonably well between the Full and Dense datasets, with the exception of the LogS distributions (**Figure 8**).

Randomize the datasets before training the model. The configuration file `config. exmple.mdnn.ini` sets up the neural network architecture:

```
learning-method: "NeuralNetwork ( \
transfer function = Rectifier (0.05), \
weight update = Simple (alpha = 0.50,eta = 0.005), \
dropout (0.05,0.25, 0.05), \
objective function = % (objective-function)s, \
scaling = AveStd, steps per update = 10, hidden architecture (128,32), \
balance = False, shuffle = True, input dropout type = Zero \
)"
```

Note that our network contains 2 hidden layers with 128 and 32 neurons, respectively, with 5% dropout on the input layer, 25%

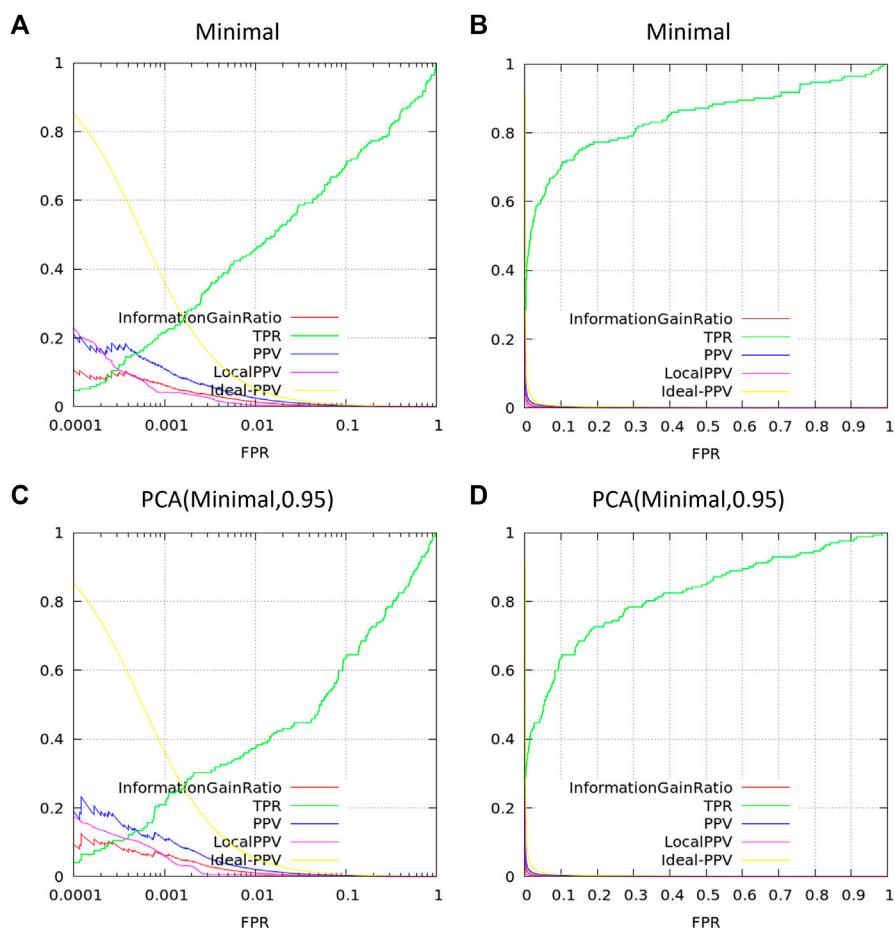


FIGURE 7 | ROC curve comparison Kir2.1 activity prediction models with different descriptors. Models were trained with either (A,B) the Minimal dataset containing 959 non-redundant standard LB descriptors (see Supplemental Data) on a log10 (A) or linear (B) x-axis, or PCA-modified LB descriptors accounting for 95% variance on a log10 (C) or linear (D) x-axis.

dropout on the first hidden layer, and 5% dropout on the second hidden layer. Our objective function will be MAE_NMAD since this is a regression task. We will perform five-fold cross validation (specified in the configuration file). Train the network:

```
launch.py -t cross-validation --local \
--datasets all_logp_logs_dgsolv.Scalar_Mol2D.rand.bin \
--id all_logp_logs_dgsolv.Scalar_Mol2D.2x256-32_005_025_005 \
--config-file config.example.mdnn.ini --just-submit
```

The just-submit flag sends the process to the background. Train the dense network as well; it should take less time since there are relatively few examples in the training sets. Check the log_merge.txt file in the corresponding “log_files” subdirectory to view the final objective function for each of the three result labels (Table 5).

In cases where the training set has small deviation from the mean value, MAE will be lower, which can be misleading. To

address this, we normalize MAE by MAD. Here, we see that the model trained on the Dense set of features learned LogP the best. However, this may be an artifact of the reduced training space. If we were to evaluate whether the Dense model was able to extrapolate beyond the very small training set, we would almost certainly see worse performance.

To illustrate this, evaluate the predictive power of our Dense model on molecules in our Full-Dense training set, and vice versa. This can be accomplished using either model: Test or molecule:Properties as described in Section 7.1. The results of this analysis are in Table 6. The model trained on the Full-Dense set does a good job predicting the QSPR properties for the Dense molecule set, achieving Pearson correlation coefficients between 0.82 and 0.99 for the three tasks. We see that the values we obtained in the internal random-split 5-fold cross validation (Table 5) agree with those obtained on the Dense set predictions (Table 6). In contrast, despite having the best five-fold cross-validation

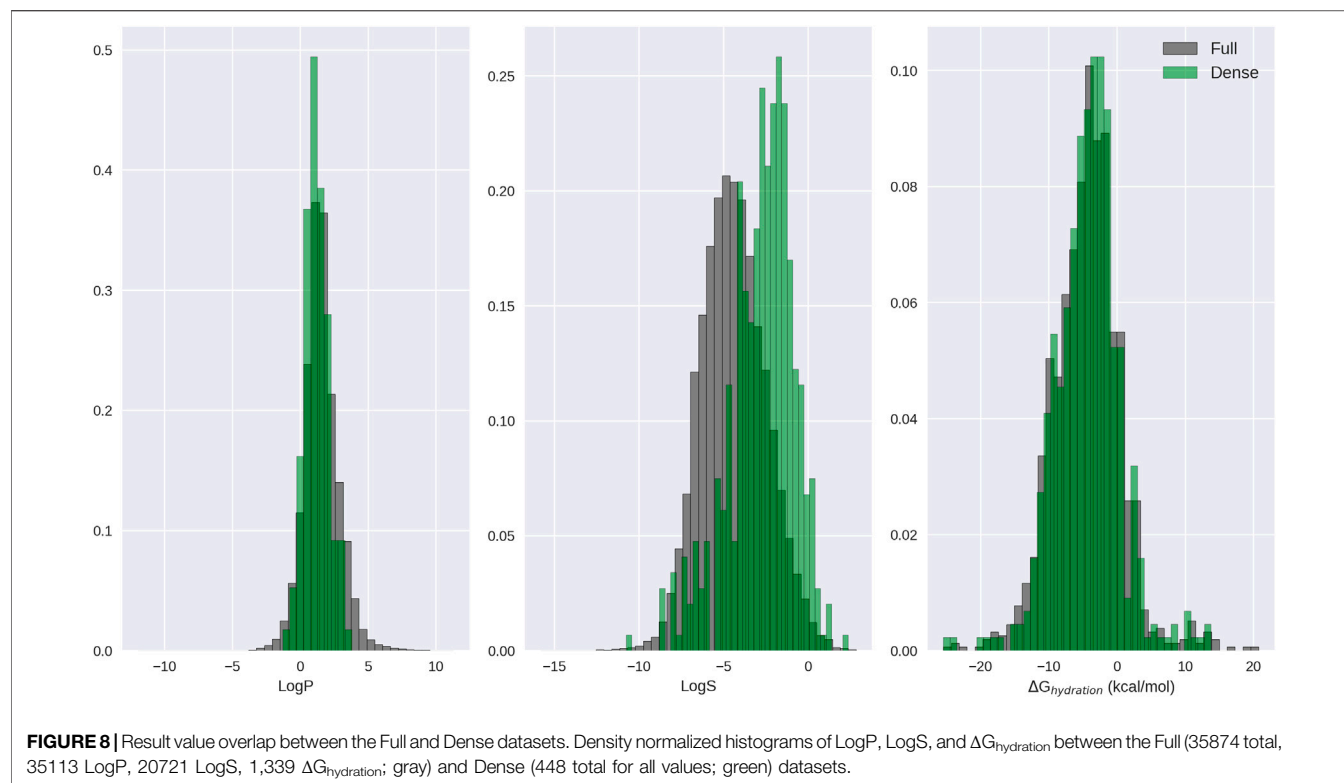


TABLE 5 | Five-fold cross validation results for multitask modeling of solubility prediction. These table values are automatically calculated and output in the log_merge.txt file in the corresponding subdirectory of the autogenerated “log_files” directory. The Full set consisted of 35,874 molecules (with 35113 LogP, 20721 LogS, and 1,339 $\Delta G_{\text{hydration}}$ result labels). The Dense set consisted of 448 molecules (with 448 LogP, 448 LogS, and 448 $\Delta G_{\text{hydration}}$ result labels). The Full–Dense set contained 35,428 molecules (with 34665 LogP, 20273 LogS, and 891 $\Delta G_{\text{hydration}}$ result labels).

| | | QSPR Prediction | | | | | | | | |
|-------------------|-----------------|-----------------|------|---------|------|------|---------|-------------------------------|------|---------|
| | | LogP | | | LogS | | | $\Delta G_{\text{hydration}}$ | | |
| | Analysis Metric | MAE | MAD | MAE/MAD | MAE | MAD | MAE/MAD | MAE | MAD | MAE/MAD |
| Model Feature Set | Full | 0.61 | 0.95 | 0.64 | 0.21 | 1.51 | 0.14 | 1.64 | 3.62 | 0.45 |
| | Dense | 0.20 | 0.68 | 0.29 | 0.24 | 1.51 | 0.16 | 1.53 | 3.58 | 0.43 |
| | Full–Dense | 0.51 | 0.97 | 0.53 | 0.23 | 1.51 | 0.15 | 2.03 | 3.85 | 0.53 |

performance (Table 5), the model trained on the Dense feature set performs extremely poorly at predicting quantitative QSPR properties of the Full–Dense molecule set (Table 6).

Taken together, these data suggest that there is likely a significant fraction of molecules in the Full–Dense set that occupy an area of feature space not represented in the 448 molecule Dense set. This is a good example that internal randomized cross-validation on a small training set is not an accurate predictor of external test set performance unless the external test set is within a similar domain of applicability (Tetko et al., 2008; Sheridan, 2012). Applicability domains in the BCL will be discussed in more detail in Section 7.5.

Training a Decision Tree

DT is a tree-based machine learning algorithm that partitions the dataset into smaller subsets as it develops. A DT starts from a root node, branches out to internal nodes, and ends at leaf nodes. To see the different options of a decision tree, call

```
bcl.exe model:Train “DecisionTree (help)”
```

The default option of the decision method chooses the features for data splitting with the maximum information gain, and its prediction performance is scored by accuracy.

```
learning-method: DecisionTree ( \
```

TABLE 6 | QSPR external test-set predictions. Results of predicting QSPR properties on the Dense dataset with the model trained on the Full-Dense feature set, and results of predicting QSPR properties on the Full-Dense dataset with the model trained on the Dense feature set. The table is organized such that the values indicate the performance of the model trained with the indicated set of descriptors on the alternate test set. The Dense set consisted of 448 molecules (with 448 LogP, 448 LogS, and 448 $\Delta G_{\text{hydration}}$ result labels). The Full-Dense set contained 35,428 molecules (with 34665 LogP, 20273 LogS, and 891 $\Delta G_{\text{hydration}}$ result labels).

| QSPR Prediction | | Model Feature Set | | | | | |
|-----------------|---------|-------------------|------|-------------------------------|--------|-------|-------------------------------|
| | | Full-Dense | | | Dense | | |
| | | LogP | LogS | $\Delta G_{\text{hydration}}$ | LogP | LogS | $\Delta G_{\text{hydration}}$ |
| Analysis Metric | MAE | 0.94 | 0.27 | 1.71 | 580.32 | 65.18 | 30.03 |
| | MAE/MAD | 1.37 | 0.18 | 0.48 | 599.81 | 43.20 | 7.80 |
| | R | 0.88 | 0.99 | 0.82 | 0.00 | -0.11 | -0.05 |
| | p | 0.89 | 0.99 | 0.88 | 0.48 | 0.88 | 0.75 |

```

objective function = Accuracy, \
partitioner = InformationGain, \
Activity cutoff = 0.5, \
nodes core = SplitRating, \
min split = 0 \
)

```

There are two factors that determine the order of features and their corresponding splitting values in dataset partitioning in a decision tree: partitioners and node scores. Four types of partitioners are currently implemented in the BCL: InformationGain, Gini, ROC, and Sequence. The first three options rate the feature to split the dataset by information gain, Gini index, and area under the curve of the local ROC curves (Ferri et al., 2002), respectively. The last option only allows splits that result in at least one pure node.

While the partitioner determines how to calculate the split rating of different configurations of dataset partition, the node score type dictates how to rank different combinations of feature order and their corresponding splitting values. Four types of node scores are currently implemented in the BCL: split rating (SplitRating), number of correct predictions before splitting (InitialNumIncorrect), split rating times initial number of correct predictions (RatingTimesInitialNumIncorrect), and sum of number of incorrect predictions before and after data splitting (InitialIncorrectPlusFinalCorrect). The users can also control the minimum number of incorrect classifications of a node by assigning a value to the min split flag.

A DT was employed in Section 3.4 to classify small molecules' potential for hit optimization. The BCL can convert DTs into descriptor files that can be used to help defined new properties. For more details, see Section 3.4.

Unsupervised Learning

Adjusting Tunable Parameters in a Self-Organizing Map

A self-organizing map (SOM), also commonly referred to as a Kohonen map, is an unsupervised learning method that is commonly used in clustering and dimensionality reduction. The SOM produces a low-dimensional (typically one to two dimensions), discretized representation of the input space of the training samples, called a map. This method applies

competitive learning to reach a solution, as opposed to conventional feed-forward neural networks, which utilize error-correction learning. To see the options available to a Kohonen map model, call

```
bcl.exe model:Train "Kohonen (help)"
```

Here is the typical configuration file setup to build a Kohonen map model:

```

learning-method: Kohonen (
shuffle = True, scaling = AveStd, map dimensions = (10, 10), \
steps per update = 0, radius = 7.5, length = 140, Neighbor
kernel = Bubble, \
Initializer = RandomlyChosenVectors, cutoff = 0.5, objective
function = RMSD \
)

```

Before training a Kohonen map, users may shuffle the training set (shuffle = True). Similar to the ANNs, there are two options for scaling the input: AveStd and MinMax. The former works best when the input descriptors are continuous, and the latter is ideal for sparse and/or discretized input data. Regarding the configuration of the SOM, the map dimensions option dictates the number of nodes, or neurons, in each direction of the map. Setting the steps per update flag (i.e., batch size) to 0 indicates that all training rows will be used for each iteration.

The initial radius of the neighborhood function, radius, is the maximum distance between the neighbor neuron and the best matching unit (BMU). Increasing the radius generally increases model quality at the expense of training time. In our experience, diminishing returns are met when the radius approaches 1/3 to 1/2 the total distance of the map. The number of iterations it takes for the radius to decrease to 0 in the given neighbor kernel function is given by length. The radius of the neighborhood is gradually reduced as the number of the iterations t increases, such that by $4 \times \text{length}$ the original radius is reduced to size 0:

$$\text{radius}_{t+1} = \text{radius}_{t=0} \left(1 - \frac{t+1}{4 \times \text{length}} \right), \quad (9)$$

Each iteration, the neurons compete by measuring their distances to the input dataset. The neuron j , with associated

weight vector w , with the lowest distance d to the randomly selected input vector x is the winner.

$$d_j(x) = \sqrt{\sum_i (x_i - w_{ji})^2}, \quad (10)$$

Iterations proceed for the entire batch size prior to updating neuron weights. The next step is updating the weights within the neighborhood of the winning node. There are two options for the neighbor kernel function: Bubble and Gaussian. The new weights are updated as

$$w_{ij}^{t+1} = w_{ij}^t + \alpha_j^t B_j(x_i^t - w_{ij}^t), \quad (11)$$

where the β is 0.8 for the winning node and 0.2 for other nodes in the neighborhood and learning rate α is $\exp(-\frac{(\text{distance to winner})^2}{2 \times \text{radius}^2})$ for the Gaussian kernel and 1 for the Bubble kernel. The Bubble kernel keeps the learning rate constant inside the neighborhood, while the Gaussian kernel reduces the learning rates for more distant nodes, at a substantial performance cost.

Finally, users can select one of the objective functions mentioned above to evaluate the prediction performance of the model. At test time, the model will assign an AD score for each external compound. This AD score is the normalized distance of that compound to the closest node of the training set. For instance, a tested molecule with an AD score of 0.90 is further from the closest node than 90% of other molecules in the training set. In other words, that molecule's feature space was not so well-represented in the training dataset.

Training a Self-Organizing Map Druglikeness Applicability Domain

We will use the BCL to build class-specific druglikeness applicability domain (AD) models from the structures of FDA approved drugs: 58 opioid receptor modulators and 82 kinase inhibitors (Wishart et al., 2018). From each set of molecules, 5 molecules are randomly removed from the training set for external validation. Training occurs on the remaining molecules. The AD models will be used to measure the similarity between external compounds and a “typical drug” targeting opioid receptors or kinases. Generate a configuration file for the AD called `AD.config` containing the following:

```
learning-method: "ApplicabilityDomainKohonen ( \
  shuffle = 0, map dimensions (% (cluster_num)s), steps per
  update = 0, \
  length = 140, radius = 7.5, neighbor kernel = Bubble, \
  initializer = RandomlyChosenVectors, scaling = AveStd, cutoff
  = 0.5, \
  share distance metric = True
)"
```

Note that the map dimensions are set by the `cluster_num` flag in the training command. Generate feature set for each molecule file using `descriptor:GenerateDataset`. Train the kinase set AD model:

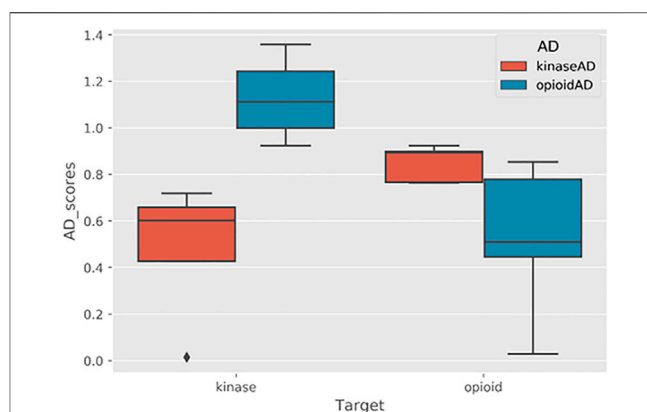


FIGURE 9 | Applicability domain models differentiate molecular structures targeting unique proteins. Each box plot represents AD scores of five drugs that target either kinases or opioid receptors. AD models trained on kinase and opioid training datasets are colored in red (legend: kinaseAD) and blue (legend: opioidAD), respectively.

```
launch.py -t cross_validation --config-file AD.config \
--datasets kinase.train.Scalar_UMol2D.bin \
--id kinase.Scalar_UMol2D.AD --max-iterations 200 \
--local --no-cross-validation --cluster_num 5
```

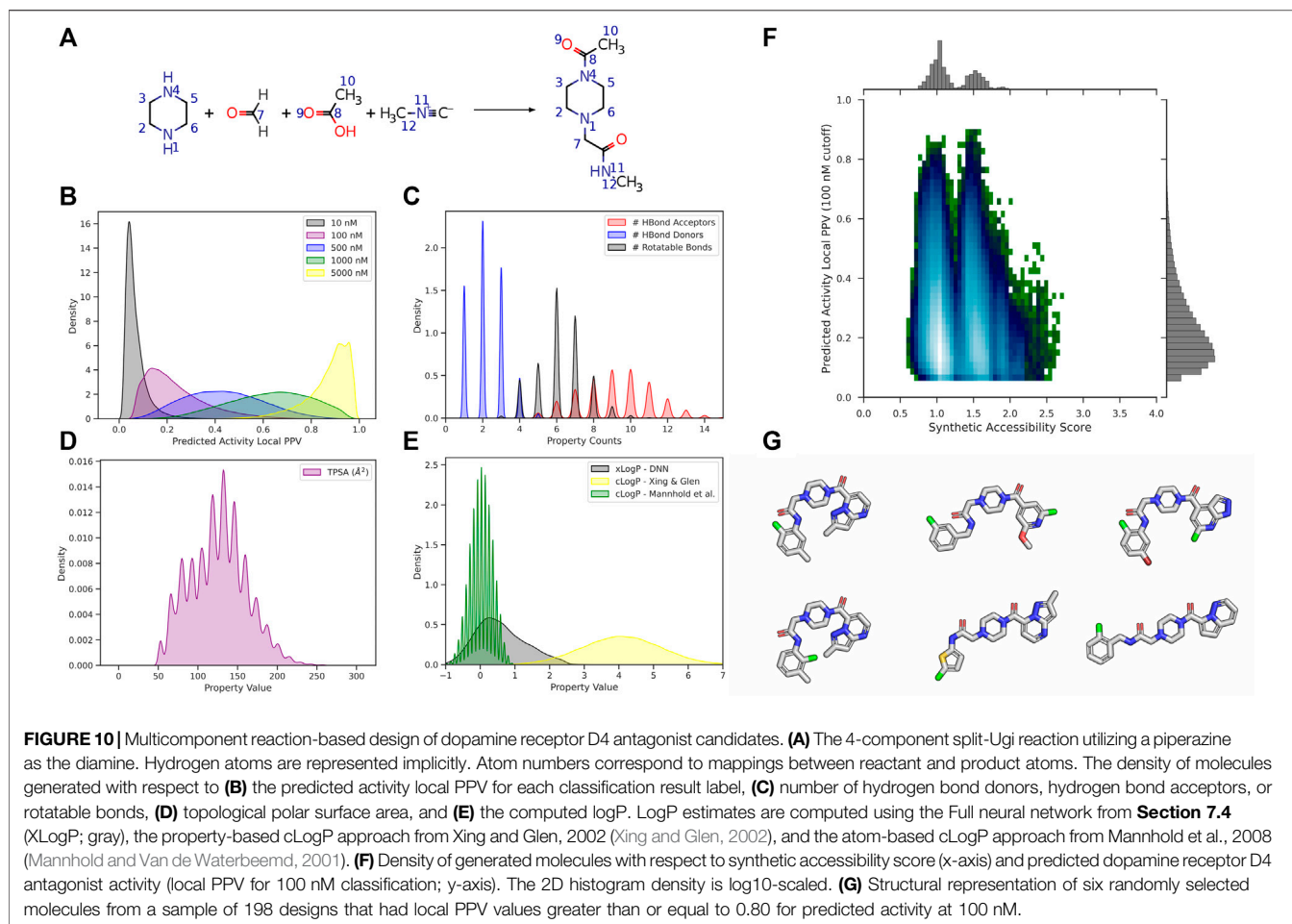
Afterward, train the opioid receptor set AD model. Next, we can evaluate the test sets with each AD model, beginning with the kinase inhibitor test set with the kinase inhibitor AD model:

```
bcl.exe model:Test -retrieve_dataset \
  "SdfFile (filename = kinase.test.sdf.gz)" \
  -storage_model \
  "File (directory = ./models/kinase_mol2d_scalar_AD, prefix =
  model)" \
  -output kinase_kinaseAD.test.out
```

The AD scores are listed in the output data file. The first two lines are the format name, and the dimension of the data table. The AD scores of five test compounds are stored in the second columns of the last 5 lines. We can see that our test set compounds from the FDA approved kinase inhibitor list have a shorter AD distance than our molecules in the opioid receptor test set, and vice versa (**Figure 9**). These scores represent the distance of each test compound to the feature space occupied by the training set FDA approved kinase inhibitors. In other words, they tell us how far we are from drug-like feature space for this group of inhibitors. The output AD scores are summarized in **Figure 9**.

DRUG DESIGN

Up to this point we have demonstrated vHTS predictions on pre-existing external datasets. Screening external datasets can be very valuable because of the ever-increasing number and availability of



public, commercial, and institutional small molecule repositories. Nevertheless, it is also frequently the case that computation can be applied to assist specific medicinal chemistry projects. For example, in silico drug design can conceivably be utilized for library design, hit explosion, or scaffold hopping. Here, we will demonstrate how to perform multicomponent reaction (MCR)-based drug design with the BCL.

Defining Reaction Files for Drug Design

Reaction-based drug design in the BCL proceeds according to user-defined MDL RXN (.rxn) files. There are a number of predefined reactions located in `bcl/rotamer_library/functional_reactions`. Reactions can be single-component intramolecular reactions, or multi-component intermolecular reactions of up to four unique reagents. Reactants must have their atoms mapped to corresponding atoms in the product(s). Atom mapping is required for substituents on the input reagents to be merged with the product(s).

The reaction design framework functions in part by performing substructure comparisons of candidate reagents to reactant structures drawn in the RXN file. Substructure matching occurs at a resolution of `ElementType` for atoms and `BondOrderOrAromatic` for bonds. If there are candidate reagents that collectively can match all reactant

positions in a reaction, then the reaction can proceed. Note that unlike input SDFs for molecule files, aromaticity must be shown explicitly in the RXN file to be interpreted. Also note that reactant matching will only match hydrogen atoms if they are drawn explicitly.

Executing Reaction Design

In this example, we will generate products according to a 4-component split-Ugi reaction utilizing piperazine as the diamine scaffold in all designs (**Figure 10A**).

```
bcl.exe molecule:React \
-starting_fragments piperazine.sdf -reagents reagents_le_20.sdf \
-reactions/rxns_dir/ -routine Random -repeats 9 -
ligand_based \
-fix_geometry -fix_ring_geometry -extend_adjacent_atoms 2 \
-output_filename ugi_products.sdf -logger File ugi_reaction.log
```

The individual molecule fragments passed via `starting_fragments` are treated as required reaction components. The `reactions` flag is given the path to a directory containing all RXN files the user wishes to include in the reaction. The `reagents` flag specifies candidate reactants with which the

starting_fragments molecules are reacted. Thus, for every entry in the SDF passed via starting_fragments, the molecule: React application will check to see if it is a valid reactant for any of the reactions in the directory specified by reactions; for those reactions that the current starting_fragments molecule is a valid reactant, the remaining possible reactant positions are fit against the molecule fragments provided via reagents.

The routine flag specifies how to continue with reaction sampling. Currently, there are two options, though additional options are under development. The default is Random, which will perform one valid reaction (if any exist) for each molecule in starting_fragments using a randomly selected reaction and reagents from the user input. By default, the Random routine will run one time; however, by specifying repeats users can increase the number of cycles. If the starting_fragments SDF contains 100 entries and repeats is set to 4, then the molecule: React application will run 500 times—one initial run for all entries followed by four repeats of all 100 entries. Alternatively, users may specify Exhaustive, which will enumerate all possible products from all given reactions and reagents for each starting_fragments molecule. Ongoing efforts to expand the reaction-based drug design framework include additional optimization routines, such as evolutionary fragment generation and simulated annealing, as well as mixed intra- and inter-reagent reactions. Other options are related to generation of 3D conformers for the product molecules and are explained in the help menu.

```
bcl.exe molecule:React --help
```

Analyzing Designs

For illustration purposes, we generated ~700,000 configurationally unique molecules with the split-Ugi reaction (Figure 10A). As our starting fragment, we used a solitary piperazine ring. For simplicity and to keep the size of the product library reasonably small, we also utilized a formaldehyde in the second reactant position (though another aldehyde is possible). We passed a collection of commercially available building block fragments, filtered such that the heavy atom count was less than or equal to 20, to fill positions three and four via our reagents flag. We analyzed the resulting library without any additional filtering (e.g., for druglikeness, predicted mutagenicity, Lipinski's rules, etc.).

Piperazine rings and related substructures are well-defined core components of dopamine receptor (DR) antagonists (Lindsley and Hopkins, 2017). Utilizing BCL commands described in previous sections along with publicly available PubChem Bioassays, we trained a single QSAR model to simultaneously predict dopamine receptor D4 (DRD4) antagonist activity at multiple thresholds (10, 100, 500, 1,000, and 5,000 nM). Subsequently, we employed this QSAR model to predict the DRD4 antagonist activity of our newly created library (Figure 10B).

As might be expected, there are a high density of molecules with a low (< 0.20) local PPV for activity at 10 nM; however, as the threshold for activity increases, the density of molecules that are identified as active increases (Figure 10B). We also quantified the number of HBDS, HBAs, and rotatable bonds in our

molecules (Figure 10C). Most compounds have fewer than 5 HBD and 10 rotatable bonds. Approximately half of the dataset contains 10 or more HBA, which would contribute to Lipinski's rules violations, though many FDA-approved molecules do not follow Lipinski rules strictly (DeGoey et al., 2018). Nevertheless, number of HBAs may be one criterion by which to filter out molecules from the library from further analysis.

We also estimated topological polar surface area (TPSA) (Figure 10D) and water-octanol partition coefficient (logP) (Figure 10E). More than half of the molecules have a TPSA less than 150 Å². One could also filter out molecules from the library that have TPSA greater than 150 Å² and/or greater than 10 rotatable bonds (Veber rules for druglikeness). We performed logP estimates with three unique methods: 1) the DNN we trained in Section 7.4.2; 2) a property-based metric from Xing and Glen, 2002 (Xing and Glen, 2002); and 3) an atom-based metric from Mannhold et al., 2008 (Mannhold and Van de Waterbeemd, 2001). Each of these metrics are available in the BCL as molecular properties and can be employed to characterize the solubility of candidate compound libraries.

Finally, we display predicted activity at 100 nM as a function of synthetic accessibility score (SAScore) (Ertl and Schuffenhauer, 2009) (Figure 10F). Encouragingly, the molecules predicted most likely to be active at 100 nM (local PPV ≥ 0.80) have SAScores below 2.0, well-within an acceptable range (Ertl and Schuffenhauer, 2009). Overall, the SAScores of the library are low, reflective of the reaction type and selected reagents (Figure 10F). We selected six random molecules with local PPV greater than 0.80 at the 100 nM activity cutoff for display (Figure 10G). These molecules are topologically similar to known antagonists of DRs, specifically DRD4; however, it is possible that this reaction produces a scaffold with an activity cliff (loss of protonation of the piperazine ring) (Berry et al., 2010; Lindsley and Hopkins, 2017).

DISCUSSION

The BCL is an academic research project made available for public use. As an academic research project, the BCL is under continuous development. Ongoing improvements are anticipated for many of the applications described here, including small molecule conformer sampling, small molecule flexible alignment, descriptor/feature generation, and additional machine learning architectures (e.g., random forest, gradient boosting, and convolutional neural networks), strategies, and pre-generated models. In addition, several new tools are currently under active development for tasks such as library design, *de novo* drug design, pharmacophore mapping, and more.

This manuscript has focused extensively on LB in silico drug discovery tools; however, we have also begun incorporating SB tools, such as deep learning-based protein-ligand interaction scoring (Brown et al., 2021). Two primary goals moving forward are 1) continuing to increase the accessibility of the BCL to other scientists, and 2) integrating the BCL with other state-of-the-art software packages to allow for more complex protocol design. To accomplish these goals in tandem, we are completing

scientific advances and software changes required to functionally integrate and compile the BCL in the Rosetta macromolecular modeling suite (Leman et al., 2020), enabling access to protocol development at the C++ (Rosetta applications), Python (PyRosetta), and XML (RosettaScripts) levels, in addition to the API described in this manuscript. We are also developing a graphical user interface (GUI) for the BCL LB drug discovery. The GUI will enable on-the-fly QSAR/QSPR calculations and druglikeness evaluation while the user is drawing molecules.

Our hope is that this manuscript will serve as a resource for those interested in utilizing the BCL for cheminformatics research. Several high level BCL applications can also be accessed via webserver for non-expert users. The webserver is available through the BCL Commons website at <http://www.meilerlab.org/bclcommons>. Example files mentioned throughout the manuscript are freely available on the Meiler Lab GitHub page.

The BCL can be downloaded freely from <http://www.meilerlab.org/bclcommons> and requires a supporting license from <http://meilerlab.org/servers/bcl-academic-license> that is free for academic and non-profit users, with commercial licenses available for a fee.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and

accession number(s) can be found below: <https://github.com/Meilerlab>.

AUTHOR CONTRIBUTIONS

Original code conceptualization and code development: JMei
Conceptualization (manuscript): BB, JMen, and JMei
Code contributions: All authors
Code review: All authors
Manuscript writing: BB, OV.

FUNDING

Work in the JMei laboratory is supported through the NIH (R01 DA046138, R01 GM099842) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, through SFB1423, project number 421152132). BB is supported through the NIH by a Ruth L. Kirschstein NRSA fellowship (F30DK118774). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.833099/full#supplementary-material>

REFERENCES

- Acharya, C., Coop, A., Polli, J. E., and Mackerell, A. D. (2011). Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr. Comput. Aided Drug Des.* 7, 10–22. doi:10.2174/157340911793743547
- Bemis, G. W., and Murcko, M. A. (1996). The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* 39, 2887–2893. doi:10.1021/jm9602928
- Berry, C. B., Locuson, C. W., Daniels, J. S., Lindsley, C. W., and Hopkins, C. R. (2010). “Discovery and Characterization of ML398, a Potent and Selective Chiral Morpholine Based Antagonist of the Dopamine 4 (D4) Receptor,” in *Probe Reports from the NIH Molecular Libraries Program* (Bethesda (MD): National Center for Biotechnology Information (US)).
- Bharath, R., Steven, K., Patrick, R., Dale Webster, D. K., and Vijay, P. (2015). *Massively Multitask Networks for Drug Discovery*. Ithaca, NY: arXiv: 1502.02072v1.
- Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. (2012). Quantifying the Chemical beauty of Drugs. *Nat. Chem.* 4, 90–98. doi:10.1038/nchem.1243
- Boström, J., Norrby, P. O., and Liljefors, T. (1998). Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput. Aided Mol. Des.* 12, 383–396.
- Bozhanova, N. G., Calcutt, M. W., Beavers, W. N., Brown, B. P., Skaar, E. P., and Meiler, J. (2021). Lipocalin B1c Is a Potential Heme-Binding Protein. *FEBS Lett.* 595, 206–219. doi:10.1002/1873-3468.14001
- Brown, B. P., Mendenhall, J., Geanes, A. R., and Meiler, J. (2021). General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps. *J. Chem. Inf. Model.* 61, 603–620. doi:10.1021/acs.jcim.0c01001
- Brown, B. P., Mendenhall, J., and Meiler, J. (2019). BCL:MolAlign: Three-Dimensional Small Molecule Alignment for Pharmacophore Mapping. *J. Chem. Inf. Model.* 59, 689–701. doi:10.1021/acs.jcim.9b00020
- Brylinski, M., and Skolnick, J. (2008). Q-dock: Low-Resolution Flexible Ligand Docking with Pocket-specific Threading Restraints. *J. Comput. Chem.* 29, 1574–1588. doi:10.1002/jcc.20917
- Butkiewicz, M., Lowe, E. W., Mueller, R., Mendenhall, J. L., Teixeira, P. L., Weaver, C. D., et al. (2013). Benchmarking Ligand-Based Virtual High-Throughput Screening with the PubChem Database. *Molecules* 18, 735–756. doi:10.3390/molecules18010735
- Cappel, D., Dixon, S. L., Sherman, W., and Duan, J. (2015). Exploring Conformational Search Protocols for Ligand-Based Virtual Screening and 3-D QSAR Modeling. *J. Comput. Aided Mol. Des.* 29, 165–182. doi:10.1007/s10822-014-9813-4
- Chan, S. L. (2017). MolAlign: an Algorithm for Aligning Multiple Small Molecules. *J. Comput. Aided Mol. Des.* 31, 523–546. doi:10.1007/s10822-017-0023-8
- Combs, S. A., Deluca, S. L., Deluca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., et al. (2013). Small-molecule Ligand Docking into Comparative Models with Rosetta. *Nat. Protoc.* 8, 1277–1298. doi:10.1038/nprot.2013.074
- Dahl, G. E. (2014). *Multi-task Neural Networks for QSAR Predictions*. Ithaca, NY: arXiv preprint arXiv:1406.1231.
- Davis, I. W., and Baker, D. (2009). RosettaLigand Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* 385, 381–392. doi:10.1016/j.jmb.2008.11.010
- DeGoey, D. A., Chen, H. J., Cox, P. B., and Wendt, M. D. (2018). Beyond the Rule of 5: Lessons Learned from AbbVie's Drugs and Compound Collection. *J. Med. Chem.* 61, 2636–2651. doi:10.1021/acs.jmedchem.7b00717

- DeLuca, S., Khar, K., and Meiler, J. (2015). Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLoS One* 10, e0132508. doi:10.1371/journal.pone.0132508
- Ertl, P., and Schuffenhauer, A. (2009). Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform* 1, 8. doi:10.1186/1758-2946-1-8
- Ferri, C., Flach, P., and Hernandez-Orallo, J. (2002). "Learning Decision Trees Using the Area under the ROC Curve," in *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002)* (Sydney, Australia: University of New South Wales).
- Friedrich, N. O., de Bruyn Kops, C., Flachsenberg, F., Sommer, K., Rarey, M., and Kirchmair, J. (2017a). Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* 57, 2719–2728. doi:10.1021/acs.jcim.7b00505
- Friedrich, N. O., Flachsenberg, F., Meyder, A., Sommer, K., Kirchmair, J., and Rarey, M. (2019). Conformer: A Novel Method for the Generation of Conformer Ensembles. *J. Chem. Inf. Model.* 59, 731–742. doi:10.1021/acs.jcim.8b00704
- Friedrich, N. O., Meyder, A., de Bruyn Kops, C., Sommer, K., Flachsenberg, F., Rarey, M., et al. (2017b). High-Quality Dataset of Protein-Bound Ligand Conformations and its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* 57, 529–539. doi:10.1021/acs.jcim.6b00613
- Friedrich, N. O., Simsir, M., and Kirchmair, J. (2018). How Diverse Are the Protein-Bound Conformations of Small-Molecule Drugs and Cofactors? *Front. Chem.* 6, 68. doi:10.3389/fchem.2018.00068
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 47, 1739–1749. doi:10.1021/jm0306430
- Hanker, A. B., Brown, B. P., Meiler, J., MarÅn, A., Jayanthan, H. S., Ye, D., et al. (2021). Co-occurring Gain-Of-Function Mutations in HER2 and HER3 Modulate HER2/HER3 Activation, Oncogenesis, and HER2 Inhibitor Sensitivity. *Cancer Cell* 39, 1099–e8. e8. doi:10.1016/j.ccell.2021.06.001
- Hartmann, C., Antes, I., and Lengauer, T. (2009). Docking and Scoring with Alternative Side-Chain Conformations. *Proteins* 74, 712–726. doi:10.1002/prot.22189
- Hassan, M., Brown, R. D., Varma-O'Brien, S., and Rogers, D. (2006). Cheminformatics Analysis and Learning in a Data Pipelining Environment. *Mol. Divers.* 10, 283–299. doi:10.1007/s11030-006-9041-5
- Hecker, E. A., Duraiswami, C., Andrea, T. A., and Diller, D. J. (2002). Use of Catalyst Pharmacophore Models for Screening of Large Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* 42, 1204–1211. doi:10.1021/ci020368a
- Jain, A. N. (2004). Ligand-based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* 47, 947–961. doi:10.1021/jm030520f
- Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H., and Meiler, J. (2010). Practically Useful: what the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry* 49, 2987–2998. doi:10.1021/bi902153g
- Kaufmann, K. W., and Meiler, J. (2012). Using RosettaLigand for Small Molecule Docking into Comparative Models. *PLoS One* 7, e50769. doi:10.1371/journal.pone.0050769
- Kawai, K., Fujishima, S., and Takahashi, Y. (2008). Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines. *J. Chem. Inf. Model.* 48, 1152–1160. doi:10.1021/ci7004753
- Kohonen, T. (1990). The Self-Organizing Map. *Proc. IEEE* 78, 1464–1480. doi:10.1109/5.58325
- Korolev, D., Balakin, K. V., Nikolsky, Y., Kirillov, E., Ivanenkov, Y. A., Savchuk, N. P., et al. (2003). Modeling of Human Cytochrome P450-Mediated Drug Metabolism Using Unsupervised Machine Learning Approach. *J. Med. Chem.* 46, 3631–3643. doi:10.1021/jm030102a
- Kothiwale, S., Mendenhall, J. L., and Meiler, J. (2015). BCL:Conf: Small Molecule Conformational Sampling Using a Knowledge Based Rotamer Library. *J. Cheminform.* 7, 47. doi:10.1186/s13321-015-0095-1
- Labute, P., Williams, C., Feher, M., Sourial, E., and Schmidt, J. M. (2001). Flexible Alignment of Small Molecules. *J. Med. Chem.* 44, 1483–1490. doi:10.1021/jm0002634
- Le Roux, N., and Bengio, Y. (2008). Representational Power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Comput.* 20, 1631–1649. doi:10.1162/neco.2008.04.07-510
- Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., et al. (2020). Macromolecular Modeling and Design in Rosetta: Recent Methods and Frameworks. *Nat. Methods* 17, 665–680. doi:10.1038/s41592-020-0848-2
- Lemmon, G., Kaufmann, K., and Meiler, J. (2012). Prediction of HIV-1 Protease/inhibitor Affinity Using RosettaLigand. *Chem. Biol. Drug Des.* 79, 888–896. doi:10.1111/j.1747-0285.2012.01356.x
- Lemmon, G., and Meiler, J. (2012). Rosetta Ligand Docking with Flexible XML Protocols. *Methods Mol. Biol.* 819, 143–155. doi:10.1007/978-1-61779-465-0_10
- Lindsley, C. W., and Hopkins, C. R. (2017). Return of D4 Dopamine Receptor Antagonists in Drug Discovery. *J. Med. Chem.* 60, 7233–7243. doi:10.1021/acs.jmedchem.7b00151
- Lo, Y. C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* 23, 1538–1546. doi:10.1016/j.drudis.2018.05.010
- Lowe, E. W., Butkiewicz, M., Spellings, M., Albert, O., and Meiler, J. (2011). *Comparative Analysis of Machine Learning Techniques for the Prediction of LogP*. IEEE.
- Ma, X. H., Wang, R., Yang, S. Y., Li, Z. R., Xue, Y., Wei, Y. C., et al. (2008). Evaluation of Virtual Screening Performance of Support Vector Machines Trained by Sparsely Distributed Active Compounds. *J. Chem. Inf. Model.* 48, 1227–1237. doi:10.1021/ci800022e
- Macalino, S. J., Gosu, V., Hong, S., and Choi, S. (2015). Role of Computer-Aided Drug Design in Modern Drug Discovery. *Arch. Pharm. Res.* 38, 1686–1701. doi:10.1007/s12272-015-0640-5
- Mannhold, R., and Van de Waterbeemd, H. (2001). Substructure and Whole Molecule Approaches for Calculating Log P. *J. Comput. Aided Mol. Des.* 15, 337–354. doi:10.1023/a:1011107422318
- Mariusz, B., Ralf, M., Danilo, S., Eric, D., Jens, M., and Kai, C. (2009). *Application of Machine Learning Approaches on Quantitative Structure Activity Relationships*. best student paper in IEEE symposium in CIBCB.
- Meiler, J., and Baker, D. (2006). ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins* 65, 538–548. doi:10.1002/prot.21086
- Mendenhall, J., and Meiler, J. (2016). Improving Quantitative Structure-Activity Relationship Models Using Artificial Neural Networks Trained with Dropout. *J. Comput. Aided Mol. Des.* 30, 177–189. doi:10.1007/s10822-016-9895-2
- Mendenhall, J., Brown, B. P., Kothiwale, S., and Meiler, J. (2020). BCL:Conf: Improved Open-Source Knowledge-Based Conformation Sampling Using the Crystallography Open Database. *J. Chem. Inf. Model.* 61, 189–201. doi:10.1021/acs.jcim.0c01140
- Mobley, D. L., and Guthrie, J. P. (2014). FreeSolv: a Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput. Aided Mol. Des.* 28, 711–720. doi:10.1007/s10822-014-9747-x
- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* 30, 2785–2791. doi:10.1002/jcc.21256
- Munshi, A. (2008). *OpenCL: Parallel Computing on the GPU and CPU*. Tutorial: SIGGRAPH.
- Nicklaus, M. C., Wang, S., Driscoll, J. S., and Milne, G. W. (1995). Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* 3, 411–428. doi:10.1016/0968-0896(95)00031-b
- Nitish, S., Geoffrey, H., Alex, K., Ilya, S., and Ruslan, S. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Machine Learn. Res.* 15, 1929–1958.
- Perola, E., and Charifson, P. S. (2004). Conformational Analysis of Drug-like Molecules Bound to Proteins: an Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* 47, 2499–2510. doi:10.1021/jm030563w
- Ramalingam, S. S., Yang, J. C.-H., Lee, C. K., Kurata, T., Kim, D.-W., John, T., et al. (2018). Osimertinib as First-Line Treatment of EGFR Mutation-Positive

- Advanced Non-small-cell Lung Cancer. *Jco* 36, 841–849. doi:10.1200/JCO.2017.74.7576
- Rogers, D., and Hahn, M. (2010). Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi:10.1021/ci100050t
- SciTegic (2007). *Pipeline Pilot - Streamlines the Integration and Analysis of Vast Quantities of Data Flooding the Research Informatics World*. Springer.
- Sheridan, R. P. (2012). Three Useful Dimensions for Domain Applicability in QSAR Models Using Random forest. *J. Chem. Inf. Model.* 52, 814–823. doi:10.1021/ci300004n
- Sitzmann, M., Weidlich, I. E., Filippov, I. V., Liao, C., Peach, M. L., Ihlenfeldt, W. D., et al. (2012). PDB Ligand Conformational Energies Calculated Quantum-Mechanically. *J. Chem. Inf. Model.* 52, 739–756. doi:10.1021/ci200595n
- Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational Methods in Drug Discovery. *Pharmacol. Rev.* 66, 334–395. doi:10.1124/pr.112.007336
- Sliwoski, G., Mendenhall, J., and Meiler, J. (2015). Autocorrelation Descriptor Improvements for QSAR: 2DA_Sign and 3DA_Sign. *J. Comput. Aided Mol. Des.* 30, 209–217. doi:10.1007/s10822-015-9893-9
- Syracuse Research Corporation (1994). *Physical/Chemical Property Database*. Syracuse, NY: PHYSPROP.
- Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., et al. (2008). Critical Assessment of QSAR Models of Environmental Toxicity against tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* 48, 1733–1746. doi:10.1021/ci800151m
- Tijmen Tieleman (2008). “Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient,” in *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008)* (Helsinki, Finland: DBLP).
- Usha, T., Shanmugarajan, D., Goyal, A. K., Kumar, C. S., and Middha, S. K. (2017). Recent Updates on Computer-Aided Drug Discovery: Time for a Paradigm Shift. *Curr. Top. Med. Chem.* 17, 3296–3307. doi:10.2174/1568026618666180101163651
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* 18, 463–477. doi:10.1038/s41573-019-0024-5
- Vlachakis, D., Fakourelis, P., Megalooikonomou, V., Makris, C., and Kossida, S. (2015). DrugOn: a Fully Integrated Pharmacophore Modeling and Structure Optimization Toolkit. *PeerJ* 3, e725. doi:10.7717/peerj.725
- Wang, Y. H., Li, Y., Yang, S. L., and Yang, L. (2005). Classification of Substrates and Inhibitors of P-Glycoprotein Using Unsupervised Machine Learning Approach. *J. Chem. Inf. Model.* 45, 750–757. doi:10.1021/ci050041k
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi:10.1093/nar/gkx1037
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: a Benchmark for Molecular Machine Learning. *Chem. Sci.* 9, 513–530. doi:10.1039/c7sc02664a
- Xing, L., and Glen, R. C. (2002). Novel Methods for the Prediction of logP, pK(a), and logD. *J. Chem. Inf. Comput. Sci.* 42, 796–805. doi:10.1021/ci010315d
- Xu, Y., Ma, J., Liaw, A., Sheridan, R. P., and Svetnik, V. (2017). Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* 57, 2490–2504. doi:10.1021/acs.jcim.7b00087
- Yosaatmadja, Y., Silva, S., Dickson, J. M., Patterson, A. V., Smaill, J. B., Flanagan, J. U., et al. (2015). Binding Mode of the Breakthrough Inhibitor AZD9291 to Epidermal Growth Factor Receptor Revealed. *J. Struct. Biol.* 192, 539–544. doi:10.1016/j.jsb.2015.10.018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Brown, Vu, Geanes, Kothiwale, Butkiewicz, Lowe, Mueller, Pape, Mendenhall and Meiler. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Combined Machine Learning and GRID-Independent Molecular Descriptor (GRIND) Models to Probe the Activity Profiles of 5-Lipoxygenase Activating Protein Inhibitors

Hafiza Aliza Khan and Ishrat Jabeen *

Research Centre for Modelling and Simulation (RCMS), NUST Interdisciplinary Cluster for Higher Education (NICHE), National University of Sciences and Technology (NUST), Islamabad, Pakistan

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Sajjad Gharaghani,
University of Tehran, Iran
Eslam Pourbasheer,
University of Mohaghegh Ardabili, Iran

*Correspondence:

Ishrat Jabeen
ishrat.jabeen@rcms.nust.edu.pk

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 30 November 2021

Accepted: 03 January 2022

Published: 01 March 2022

Citation:

Khan HA and Jabeen I (2022)
Combined Machine Learning and
GRID-Independent Molecular
Descriptor (GRIND) Models to Probe
the Activity Profiles of 5-Lipoxygenase
Activating Protein Inhibitors.
Front. Pharmacol. 13:825741.
doi: 10.3389/fphar.2022.825741

Leukotrienes (LTs) are pro-inflammatory lipid mediators derived from arachidonic acid (AA), and their high production has been reported in multiple allergic, autoimmune, and cardiovascular disorders. The biological synthesis of leukotrienes is instigated by transfer of AA to 5-lipoxygenase (5-LO) via the 5-lipoxygenase-activating protein (FLAP). Suppression of FLAP can inhibit LT production at the earliest level, providing relief to patients requiring anti-leukotriene therapy. Over the last 3 decades, several FLAP modulators have been synthesized and pharmacologically tested, but none of them could be able to reach the market. Therefore, it is highly desirable to unveil the structural requirement of FLAP modulators. Here, in this study, supervised machine learning techniques and molecular modeling strategies are adapted to vaticinate the important 2D and 3D anti-inflammatory properties of structurally diverse FLAP inhibitors, respectively. For this purpose, multiple machine learning classification models have been developed to reveal the most relevant 2D features. Furthermore, to probe the 3D molecular basis of interaction of diverse anti-inflammatory compounds with FLAP, molecular docking studies were executed. By using the most probable binding poses from docking studies, the GRIND model was developed, which indicated the positive contribution of four hydrophobic, two hydrogen bond acceptor, and two shape-based features at certain distances from each other towards the inhibitory potency of FLAP modulators. Collectively, this study sheds light on important two-dimensional and three-dimensional structural requirements of FLAP modulators that can potentially guide the development of more potent chemotypes for the treatment of inflammatory disorders.

Keywords: 5-lipoxygenase activating protein (FLAP) inhibitors, machine learning, molecular docking, grind, leukotrienes (LTs)

1 INTRODUCTION

The 5-LO pathway is responsible for the biological synthesis of leukotrienes (LTs) using arachidonic acid (AA) predominately by inflammatory cells like polymorphonuclear leukocytes, activated macrophages, and mast cells upon arrival of immunologic and non-immunologic stimuli (Hedi and Norbert 2004). Activation of leukocytes results in translocation of cytosolic protein

phospholipase A2 (PLA2) to membrane where it selectively hydrolyzes the sn-2 acyl bond of membrane phospholipids to release AA and lysophosphatidic acid. An integral membrane protein called FLAP (5-lipoxygenase activating protein) uptakes the AA and efficiently transfers it to the active site of 5-lipoxygenase (5-LO) enzyme, which catalyzes a series of reactions at a single active site (Peters-Golden 1998; Peters-Golden and Brock 2003). In the first step, in a calcium- and ATP-dependent reaction, AA is converted to a 5-lipoxygenase-specific hydroperoxide intermediate (5-HPETE), while in the second step, 5-LO performs synthase reaction for conversion of 5-HPETE to the epoxide intermediate, leukotriene A4 (LTA4) (Woods et al., 1995; Smyrniotis et al., 2014). LTA4 acts as a common precursor for biosynthesis of chemoattractant leukotriene B4 (LTB4) by a zinc-bound LTA4 hydrolase (LTA4H) and bronchoconstrictive cysteinyl leukotrienes (CysLTs or LTC4) with the help of LTC4 synthase (LTC4S) (Jakschik and Kuo 1983; Haeggström 2000). Both LTB4 and CysLTs are physiologically active final products of the 5-LO pathway and are exported out of the cell through specific transport proteins while extracellular peptidases metabolize LTC4 to LTD4, which is converted into LTE4 depending on type of inflammatory signal and cell demand (Jedlitschky and Keppler 2002). After export, LTs bind with respective G-protein-coupled receptors, e.g., LTB4 binds with BLT1 and BLT2, whereas CysLTs activate CysLT1 and CysLT2 receptors to incite further proinflammatory signaling cascades (Ghosh et al., 2016).

Since high levels of LTs have been reported in the pathophysiology of a wide range of inflammatory, cardiovascular, and autoimmune disorders, FLAP has become the focus of immense research because LT production can be stopped at the earliest level (Folco et al., 2000; Liu and Yokomizo 2015). Over the course of the last 3 decades, several FLAP modulators have been proposed including first generation of derivatives of indoles and quinolines for asthma treatment (Evans et al., 1991; Frenette et al., 1999). These inhibitors such as MK-886, MK-591, and BAY-X-1005 demonstrated efficiency in clinical trials in patients with inflammatory diseases in the mid-1990s but were not brought to market due to poor pharmacokinetics (Friedman et al., 1993; Diamant et al., 1995; Dahlén et al., 1997). Revelation of SAR data along with crystal structure expedites the drug discovery quest against FLAP, leading to the second generation consisting of derivatives of diarylalkanes, biaryl amino-heteroarenes, and benzimidazoles, proposed with renewed interest for treatment of cardiovascular diseases (Lemurell et al., 2015; Macdonald et al., 2008; N.; Stock et al., 2010). Moreover, several inhibitors proved to be promising readouts for preclinical and clinical studies such as AM103, AM803, BI665915, AZD5718, and AZD6642 and have been shown to ameliorate inflammation-related diseases (Bain et al., 2010; Lorrain et al., 2010; Antoniu, 2014; Ericsson et al., 2020). However, despite several practices, not a single inhibitor has won the race to the market as a drug to date. Therefore, development of more potent chemical entities against FLAP is highly desirable to provide relief to patients suffering from inflammatory disorders.

Mostly FLAP modulators were synthesized and pharmacologically tested and optimized through SAR (structure–activity relationship) studies. Some candidates were also identified by virtual screening from a ligand-based pharmacophore built upon smaller datasets (Banoglu et al., 2012; Temml et al., 2017; Olgac et al., 2020). Here, in this study, advanced machine learning (ML) techniques along with classical modeling strategies are adapted to shed light on important 2D and 3D anti-inflammatory properties of a diverse set of inhibitors targeting FLAP. For this reason, ML models based on most relevant 2D descriptors or features have been constructed. Further molecular docking was performed to establish a binding hypothesis of each class of inhibitors within the FLAP binding cavity followed by common scaffold clustering to obtain the most probable 3D binding solutions. The most probable 3D binding poses were utilized for GRID-independent molecular descriptor analysis (GRIND) to probe the important 3D binding features and associated mutual distances in active FLAP modulators.

2 MATERIALS AND METHODS

2.1 Machine Learning Modeling

2.1.1 Dataset Preparation

All compounds having activity values in IC_{50} against FLAP were retrieved from the ChEMBL database under target ID ChEMBL4550 followed by removal of compounds with similar canonical smiles resulting in a dataset of 658 compounds. The IC_{50} of the finalized 658 compounds ranged from 0.3 to 22,500 nM. Furthermore, the highly active and least active compounds were distinguished by the application of activity threshold, i.e., compounds having $IC_{50} < 10$ nM were categorized as highly active while compounds having $IC_{50} > 70$ nM were categorized as least active considering that FLAP inhibitors that have entered clinical trials usually possess values < 10 nM (Gür, Çalışkan, and Banoglu 2018). Compounds with IC_{50} values in between >10 nM and <70 nM were labeled as intermediates and were removed. For ML classification model development, highly active compounds were labeled as one, while least active ones were labeled as 0. The final dataset was composed of 503 (253 highly actives and 250 least actives) compounds and was randomly divided into a training set (402 compounds: 201 highly actives and 201 least actives) and a test set (101 compounds: 52 highly actives and 49 least actives) by a ratio of 80% and 20% respectively using *train_test_split* function (*random_state* = 42) of *model_selection* library from the *scikit-learn* Python package (Pedregosa et al., 2012). Additionally, it was ensured that the ratio of the highly active to weakly active inhibitors remained equal in the training and test set.

2.1.2 Computation of 2D Chemical Descriptors

Initially, 4,179 2D descriptors were calculated using *alvaDesc* tool version 2.0.8 (Mauri 2020). The descriptors can be divided into 21 categories named constitutional indices, ring descriptors, topological indices, walk and path counts, connectivity indices,

information indices, 2D-matrix based descriptors, 2D autocorrelations, burden eigenvalues, P_VSA like descriptors, ETA indices, edge adjacency indices, fractional group counts, atom-centered fragments, atom-type-estate indices, pharmacophore descriptors, 2D atom pairs, charge descriptors, molecular properties, drug-like indices, MDE descriptors, and chirality descriptors. Descriptors with null values and variance near zero were removed. For the remaining 2,352 descriptors, Pearson autocorrelation coefficient was calculated and autocorrelated descriptors along with low dependency (correlation) on the target variable (inhibitory potency, IC₅₀) were discarded, resulting in a set of 442 descriptors. The final 442 features were further applied to train the ML models.

2.1.3 Machine Learning Modeling

For this study, six supervised ML classification models named support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), decision tree (DT), logistic regression (LR), and gradient boost decision tree (GBDT) were developed. SVM, RF, MLP, DT, and LR were generated using the scikit-learn Python package whereas the GBDT was built by the XGBoost Python package (Cortes, Vapnik, and Saitta 1995; Liaw and Wiener 2002; Haykin, 2009; Quinlan 1986; T.; Chen and Guestrin 2016; McCullagh and Nelder 1989). To select the most relevant features from the set of 442 descriptors, the RFECV (Recursive Feature Elimination and Cross-Validation Selection) algorithm of scikit-learn was used (Guyon et al., 2002). Recursive feature elimination (RFE) is a wrapper-type feature selection that works by eliminating *n* features from a model by fitting the model multiple times and, at each step, removing the weakest features, determined by either the *coef_* (SVM and LR) or *feature_importances_* (RF, DT, and XGBoost) attribute of the fitted model (Guyon et al., 2002). Since there is no attribute available to estimate feature importance in MLP, XGBoost was used as the base estimator. The cross-validation (*cv*) parameter of RFECV was set at fivefold and was done by using the RepeatedStratifiedKFold method of the model_selection library from scikit-learn (Zeng and Martinez 2010). The GirdSearchCV library in scikit-learn was used to tune hyperparameters of the estimators based on a 10-fold cross-validation Matthews Correlation Coefficient (MCC). This process was repeated ten times. Moreover, to assemble data transformer (RFECV) and hyperparameter tuner (GirdSearchCV) with simultaneous cross-validation while setting different parameters, the pipeline module of scikit-learn was used.

An SVM constructs a maximum marginal hyperplane with the help of a kernel function to map the non-linear problem in multidimensional space for data separation. The performance of the SVM model is controlled by parameters such as kernel, capacity parameter (*C*), and gamma. Kernel represents sample distribution in the mapping space, *C* controls the trade-off between smooth decision boundary, and gamma controls the extent of curvature in decision boundary (Nekoei, Mohammadhosseini, and Pournbasheer 2015; Pournbasheer et al., 2017). For this project, linear kernel was utilized while all parameters were set at their default values except for tuning of

penalty parameter (*C*) (Chang and Lin, 2021). MLP is a feedforward artificial neural network and is trained using back propagation algorithm. It has an activation function that forms a linear combination according to weights of inputs to decide the output. The MLP model was controlled by tuning the following parameters: the number of neurons (*hidden_layer_sizes*) and activation function (*activation*), while the rest of the parameters were set at their default values (Glorot and Bengio, 2021). An LR model predicts a dependent data variable by analyzing the relationship through logic functions between one or more existing independent variables. It was controlled by tuning the following parameters: the way of regularization (*penalty*), strength of regularization (*C*), tolerance for stopping criteria (*tol*), and algorithm of optimization (*solver*), whereas other parameters not mentioned were set at their default values (Fan et al., 2008). A DT classifies data by splitting them into source nodes and then multiple child nodes using statistical probability. The DT model was optimized by tuning the following parameters: quality of split (*criterion*), split at each node (*splitter*), and number of features for the best split (*max_features*). The remaining parameters were set as their default values (Brieman and Olshen 2012). An RF builds multiple decision trees and merges them together to get an accurate and stable prediction. The RF model was controlled by tuning the following parameters: number of trees (*n_estimators*), quality of a split (*criterion*), features for the best split (*max_features*), and the minimum number of samples required for splitting (*min_samples_split*); the other parameters not mentioned were set at their default values (Breiman 2001). XGBoost is an ensemble tree method that applies the principle of boosting weak learners using the gradient descent architecture. For this project, gradient boost tree (GBDT) has been implemented, which uses decision trees as weak classifiers. The XGBoost model was controlled by tuning the following parameters: the maximum depth of a tree (*max_depth*), the number of the tree (*n_estimators*), minimum loss reduction required for partition on a node (*gamma*), minimum sum of instance weight needed to generate a child node (*min_child_weight*), strength of L1 regularization (*reg_alpha*), and learning rate (*learning_rate*). The other parameters not mentioned were set at their default values (T. Chen and Guestrin 2016).

The repeated stratified 5-fold cross-validation was used on the training set to select and evaluate the robustness of models, and the test set was used to evaluate the performance of models. Evaluation parameters include classification accuracy (ACC), true positive rate or sensitivity (SE), true negative rate or specificity (SP), and Matthews correlation coefficient (MCC) as mentioned in (Eqs 1–4) below:

$$\text{True Positive Rate (Sensitivity)} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{True Negative Rate (Specificity)} = \frac{TN}{TN + FN} \quad (2)$$

$$\text{Classification Accuracy (ACC)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Matthews Correlation Coefficient (MCC)} \\ = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

2.2 Molecular Modeling

2.2.1 Calculation of Lipophilic Efficiency (LipE) and cLogP

To estimate the druglikeness of the initially finalized 658 FLAP inhibitors (section 2.1.1), LogP was calculated by using Bio-Loom software (BioByte - Bio-Loom, 2021) followed by computation of LipE with the following equation:

$$\text{LipE} = \text{pIC}_{50} - \text{cLogP} \quad (5)$$

Briefly, lipophilicity or cLogP strongly impacts membrane passive permeability, which is required for oral absorption and access of the drug to intracellular compartments and tissue penetration (Arnott and Planey 2012). Lipophilic efficiency (LipE) is defined as normalization of the pIC_{50} with respect to cLogP of the compound. Previously, Leeson et al. proposed that an ideal drug candidate should have a LipE value greater than five, which is obtained in case of high potency and low lipophilicity (Leeson and Springthorpe 2007). For the application of molecular modeling techniques (Docking and 3D QSAR GRIND), LipE and cLogP filter were used; i.e., compounds having LipE greater than one and cLogP greater than two were selected. The new dataset of compounds having LipE value greater than one and cLogP greater than two was divided into a training set (80%) and a test set (20%) by using the *train_test_split* function (*random_state* = 42) of the *model_selection* library from the *scikit-learn* Python package. Both training and test datasets were further employed in molecular modeling studies (docking studies and GRIND modeling).

2.2.1 Molecular Docking and Pose Analysis

To explore the binding interactions of structurally diverse FLAP inhibitors, and to obtain the most probable 3D binding conformations of ligands for GRIND analysis, inhibitors having LipE value greater than one and cLogP greater than two were docked into the binding pocket of the FLAP structure retrieved from the Protein Data Bank (PDB ID: 2Q7M) (Ferguson et al., 2007). Protein structure was prepared by energy minimization through the Amber99 force field of MOE (A. A. Chen and Pappu 2007). The energy-minimized structure was imported into GOLD software (version 5.6.1) (Jones et al., 1997) followed by determination of x, y, z coordinates around the single-solvent accessible point present in the center of the active site. The binding site area was kept at 12 Å radius, which included all important amino acid residues reported by previous studies. A total of 100 conformations for each ligand were generated, and GOLD fitness scoring function was used to rank each pose of ligands with subsequent energy minimization of ligand-protein docking complexes using LigX implemented in software MOE. Gold score fitness scoring function was calculated as:

$$\text{Fitness} = S_{(hb)_{ext}} + 1.3750 * S_{(vdw)_{ext}} + S_{(hb)_{int}} + 1.0000 * S_{(int)} \quad (6)$$

Based on structural similarity, common scaffold clustering (CSC) as proposed by Jabeen et al. (2012) was conducted to reduce the conformational space. For this purpose, RMSD matrix was generated through agglomerative hierarchical cluster analysis, and clusters with maximum docked ligands were selected for ligand-protein interaction profiling. Common interactions between each class were sorted out and binding hypothesis was generated for each class with respect to interaction pattern and position in binding pocket. Conformations from selected clusters were further utilized in GRIND analysis as training set.

2.2.3 Grid Independent Molecular Descriptors Analysis

Selected 3D molecular conformations of ligands obtained from clusters containing maximum docked ligands along with their inhibitory potencies (pIC_{50}) were imported in Pentacle software version 1.06 to construct the GRIND model (Pastor et al., 2000). Calculation of molecular interaction fields (MIFs) was done by use of different probes, namely, N1, O, DRY, and TIP, where N1 (amide N) represents a hydrogen bond donor, O (sp^2 carbonyl O) denotes a hydrogen bond acceptor, DRY indicates a hydrophobic region, and TIP stands for steric hotspots within the virtual receptor site. A GRID was used to iteratively place these probes to calculate the total energy by addition of Lennard-Jones potential energy (Elj), hydrogen bond energies (Ehb), and the electrostatic energy (Eel), whereas with the help of the following equation, total interaction energy at each node was calculated:

$$E_{xyz} = \sum E_{hb} + \sum E_{lj} + \sum E_{el} \quad (7)$$

AMANDA algorithm was used to extract the most relevant and significant MIFs along with evaluation of structural characteristics of the dataset explained by GRIND descriptors (Durán, Martínez, and Pastor 2008). The default GRID space of 0.5 and the energy cutoff values, which are -4.2, -2.6, -0.5, and -0.74 for N1, O, DRY, and TIP, respectively, were used for discretization of MIFs, while nodes that did not meet the energy cutoff were discarded. The next encoding step involves CLACC algorithm that aided in selection of consistent nodes by adjustment of compounds according to their moment of inertia. The values obtained from encoding consist of a consistent set of variables whose values were directly represented in the form of correlogram plots. The final GRIND model with PLS (partial least square) analysis using LOO (leave one-out) method with statistically significant R^2 , q^2 , and standard error values (SDEP) was built on the training set followed by evaluation with the test set (section 2.2.1). Additionally, r^2_m metrics (r^2_m , Delta r^2_m) was also generated for validation purposes according to the previously published studies (Roy et al., 2013; Gajo et al., 2016).

TABLE 1 | The layout of prediction performances of machine learning models assessed by stratified 5-fold cross-validation for the training set and test set.

| Classifier | Training set (n = 402) | | | | CV5 of training set (n = 402) | | | | Test set (n = 101) | | | |
|------------------------------|------------------------|------|------|------|-------------------------------|------|------|------|--------------------|------|------|------|
| | SE | SP | ACC | MCC | SE | SP | ACC | MCC | SE | SP | ACC | MCC |
| XGBoost (GBDT) | 0.99 | 0.99 | 0.99 | 0.98 | 0.91 | 0.89 | 0.90 | 0.81 | 0.89 | 0.91 | 0.90 | 0.80 |
| Random forest (RF) | 0.99 | 1.00 | 1.00 | 0.99 | 0.85 | 0.90 | 0.87 | 0.75 | 0.94 | 0.88 | 0.91 | 0.82 |
| Decision tree (DT) | 0.88 | 0.94 | 0.91 | 0.82 | 0.83 | 0.83 | 0.83 | 0.66 | 0.83 | 0.84 | 0.84 | 0.68 |
| Support vector machine (SVM) | 0.96 | 0.98 | 0.97 | 0.93 | 0.84 | 0.77 | 0.80 | 0.61 | 0.75 | 0.80 | 0.78 | 0.56 |
| Logistic regression (LR) | 0.82 | 0.88 | 0.85 | 0.69 | 0.84 | 0.75 | 0.79 | 0.59 | 0.85 | 0.87 | 0.86 | 0.72 |
| Multilayer perceptron (MLP) | 0.79 | 0.81 | 0.80 | 0.60 | 0.72 | 0.78 | 0.75 | 0.50 | 0.70 | 0.71 | 0.70 | 0.40 |

3 RESULTS

3.1 Machine Learning Models

Six ML models were developed by different algorithms (SVM, LR, MLP, DT, RF, and GBDT) using two-dimensional structural features of FLAP inhibitors. The performance of these models on 5-fold repeated stratified cross-validation is explained in **Table 1**. The cross-validation accuracy of the training set ranged between 0.90 and 0.75, and the MCC ranged from 0.81 to 0.50. The prediction accuracy and MCC values of the test set ranged from 0.90 to 0.70 and 0.80 to 0.40, respectively. MCC is often used as a measure of quality of binary classification models. Two models (XGBoost and RF) exhibited an MCC value of >0.7 on training and test sets, which means these two algorithms have a relatively good ability to predict whether a compound was a highly active or a least active FLAP modulator. In terms of the best model, XGBoost outperformed all and the accuracy and MCC values were observed as 0.90 and 0.81, respectively. Additionally, a previous fingerprint-based ML study on FLAP modulators stated that the reliability of predicted results depends mainly on the compounds themselves rather than algorithms or fingerprints (Tu et al., 2020).

The lowest performance was shown by the MLP model with an accuracy value of 0.75 and an MCC value of 0.50. For the best model (XGBoost), RFECV curve jumps to a maximum accuracy when the 46 informative features are captured with feature importance values ranging from 0.01 to 0.4. These 46 features mainly belong to eight descriptor categories named topological indices, 2D matrix-based descriptors, 2D autocorrelations, P_VSA-like descriptors, edge adjacency indices, atom-type E-state indices, pharmacophore descriptors, and molecular properties descriptors. All 46 captured features of the best-performing model (XGBoost) along with description and feature importance values are given in **Supplementary Table S1**. Additionally, 84, 126, 89, and 90 features have been captured by RF, DT, SVM, and LR, respectively, and RFECV curves for all models with optimal number of selected features are illustrated in **Supplementary Figure S1**. We anticipate that these 46 2D descriptors have the largest impact to differentiate between highly active and least active FLAP inhibitors. Additionally, the tuned hyperparameters for each model can be found in **Supplementary Table S2**.

3.2 LipE and cLogP Calculation

LipE and cLogP demonstrate the druggability of a compound in lead optimization programs to evaluate the potential for better *in*

vivo efficacy and safety. A graph between pIC₅₀ and cLogP along with LipE values of the compounds in the training set is shown in **Supplementary Figure S2**. In the current dataset of 658 FLAP inhibitors, only 238 compounds out of 658 demonstrated LipE value greater than five, which is the optimal threshold with cLogP values in the range of −0.27 to 3.78. Moreover, only 136 compounds showed a cLogP value between optimal range of 2–3.5 as proposed by Leeson and Springthorpe (2007). Additionally, 349 compounds out of 658 exhibited values of LipE less than 1 (cLogP = 4.3–10.19) while the cLogP range for 309 compounds having a LipE value greater than one was observed as 0.27–7.88. Interestingly, several potent FLAP inhibitors such as MK-886 (pIC₅₀ = 8.65 cLogP = 8.58, LipE = 0.07), MK-591 (pIC₅₀ = 9.30 cLogP = 8.82, LipE = 0.48), AM-643 (pIC₅₀ = 8.69, cLogP = 7.72, LipE = 0.97), AM-679 (pIC₅₀ = 8.65, cLogP = 7.98, LipE = 0.67), AM-803 (pIC₅₀ = 8.53, cLogP = 8.97, LipE = 0.43), and BRP-7 (pIC₅₀ = 6.50, cLogP = 7.23, LipE = 0.72) displayed significantly low values of LipE. It seems that increase in potency of these compounds might be due to increase in lipophilicity. On the other hand, other FLAP modulators such as BI665915 (pIC₅₀ = 8.76, cLogP = 2.14, LipE = 6.62) and AZD6642 (pIC₅₀ = 8.31, cLogP = 1.72, LipE = 6.62) showed relatively high values of LipE.

Herein, a dataset of 187 compounds having LipE value greater than one and cLogP greater than two was selected for further application of molecular modeling studies as all FLAP inhibitors in clinical trials possess high values of lipophilicity (cLogP). The dataset of 187 compounds was subsequently divided into a training set (151 compounds, **Supplementary Table S3**) and a test set (36 compounds, **Supplementary Table S4**). Docking-guided GRIND analysis was performed on the training set followed by evaluation of the final GRIND model with the test set.

3.3 Molecular Docking and SAR-Guided Pose Analysis

The selected dataset of 187 compounds mainly consists of already published indoles, biaryl bicycloheptanes, oxadiazole, and benzimidazole-based compounds. The dataset was further divided into a training set (151 compounds) and a test set (36 compounds) and based on common scaffolds; the training set was classified into six distinct classes. Common scaffold along with activity, lipophilicity, and lipophilic efficiency ranges of the six classes is depicted in **Figure 1**. Furthermore, a binding hypothesis of each class within the FLAP binding cavity was established. The

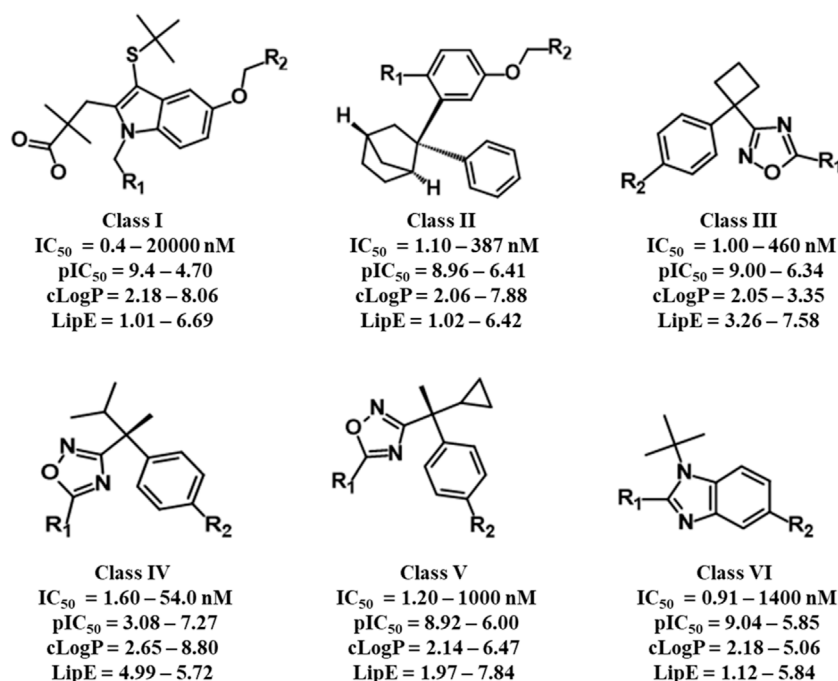


FIGURE 1 | Common scaffolds of six classes of FLAP inhibitors used for common scaffold clustering to obtain the most probable 3D binding poses for employment in GRIND studies.

distributions of compounds in each class along with common scaffolds are depicted in **Supplementary Table S5**.

All datasets of the selected 187 compounds were docked into the FLAP binding pocket, which included an area of 12 Å selected by assigning x (65.7018), y (58.7512), and z (36.4565) coordinates between chains B and C near previously known interacting amino acid residues (B-F123, B-L120, B-I119, B-R117, B-K116, B-G115, B-F114, B-I113, B-Y112, B-T66, B-A63, B-D62, C-V61, C-C60, C-Q58, C-N57, C-H28, C-A27, C-F25, C-G24, C-N23, and C-V21) (Mancini et al., 1994; Ma et al., 2008). To remove any biases in the docking protocol, 100 poses per ligand were generated using the GOLD score fitness function. Further docking solutions were inspected by algoromatics hierarchical cluster analysis based on root mean square deviation (RMSD) at 3.5 Å of the heavy atoms around a common scaffold. To follow the idea of similar binding mode for similar compounds, only those clusters that comprised the maximum number of docked ligands were selected. Overall, one cluster of binding conformations of compounds in all classes have been identified that contained the maximum number of docked ligands. The final selected cluster of each class and details of common scaffold clustering are depicted in **Supplementary Table S5**. Briefly, 26 out of 32 compounds for class I, 12 out of 20 for class II, 32 out of 35 for class III, 10 out of 10 for class IV, 13 out of 18 for class V, and 26 out of 36 for class VI were clustered out. Interestingly, the binding position of all final clusters was the same, and they bind between helix 4 ($\alpha 4$) and helix 2 ($\alpha 2$) of chain B and helix 1 ($\alpha 1$) and helix 2 ($\alpha 2$) of chain C, but a distinct binding pattern was observed for each class. The

binding region between chains B and C occupied by all generated poses of 187 ligands is shown in **Figure 2A**.

Briefly, class I compounds are derivatives of indole with dimethyl butanoic acid and S-tert-butyl substituents at positions two and three, respectively, as displayed in **Figure 1**, while R_1 at position one and R_2 at position five are generally occupied by heterogeneous 6-membered cyclic rings. The binding solutions for final cluster (cluster 1, **Supplementary Table S5**) of compounds in class I showed that dimethylbutanoic acid makes π -H-bond interactions with C-H28 and C-Val21, S-tert-butyl makes π -H-bond interactions with B-L120 and B-F123, while the indole scaffold is primarily involved in making π -H-bond interactions with C-G24 (**Figure 2B**). The R_1 substituents show hydrogen bonding with B-D62 and C-N23 and π -H interactions contact with B-A63 and C-N23, whereas N of the pyridine ring of R_2 shows a strong hydrogen bond with B-A63 while R_2 substituents show hydrogen bonding with B-R117 and B-K116 (**Figure 2B**). Overall, compounds of class 1 displayed a positive trend ($R^2 = 0.57$) between lipophilicity and inhibitory potency (**Supplementary Figure S3**) and exhibit a distinct SAR pattern. For instance, compound **1** ($IC_{50} = 0.4 \text{ nM}$, **Supplementary Table S3**) having the highest activity value ($cLogP = 8.06$, $LipE = 1.34$) among all the datasets contains 5-methylpyridine at R_1 and para-fluoro-2-phenylpyridine at R_2 as depicted in **Supplementary Table S6**. The final docking solution of compound **1** reveals that the pyridine ring present at R_1 shows a π -H-bond bonding interaction with the $-NH_2$ group of B-R117 (**Figure 4**). Compound **98** ($IC_{50} = 9.0 \text{ nM}$, **Supplementary Table S6**) has a similar structure to compound **1** except for the absence

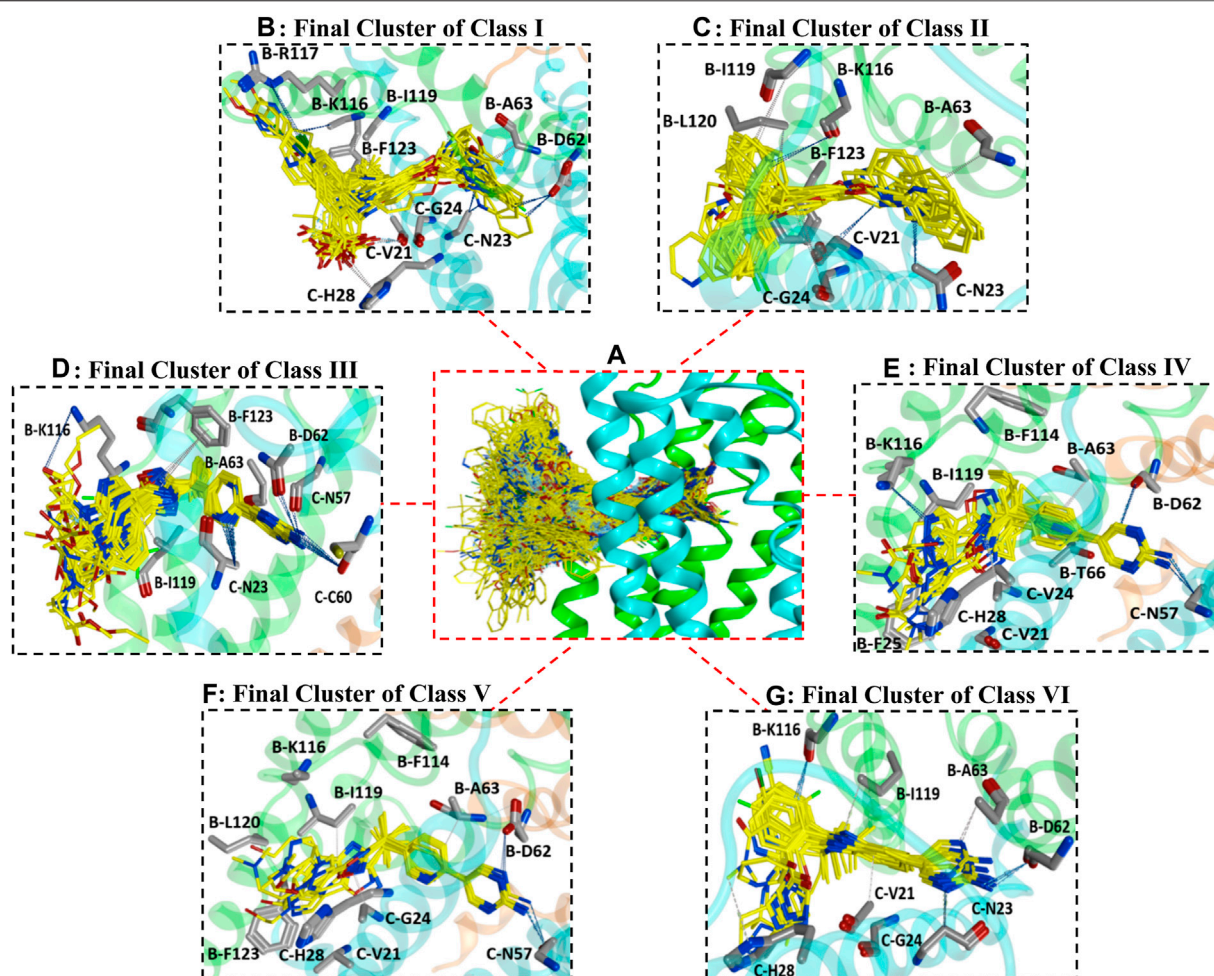


FIGURE 2 | (A) illustrates the binding positions and chemical space occupied by all generated poses of 187 FLAP antagonists between chains B and C of the FLAP binding cavity. Chain B is shown in green color, chain C is depicted in blue color, while chain A is depicted in orange color. **(B–G)** represents binding poses of maximum docked ligands in final clusters from class I to VI respectively obtained from common scaffold-based clustering.

of terminal 5-methyl on R_1 and the absence of fluorine on R_2 , rendering it low lipophilic ($c\text{LogP} = 3.35$, $\text{LipE} = 4.69$) and less active.

A study by Stock et al. also established that terminal 5-methyl on pyridine at R_1 significantly increases the inhibitory potency of compounds against FLAP (N. Stock et al., 2010; N. S. Stock, 2011). Interestingly, the hydrogen bonding between $-\text{NH}_2$ of B-R117 and nitrogen of the pyridine ring of R_2 of compound **98** has also been observed in the final docking solutions. However, the pyridine ring at R_1 did not seem to be involved in making any clear interactions. It was observed that the absence of terminal methyl on the pyridine ring of R_1 in compound **98** might reduce the exposure of pyridine ring to amino acids inside the FLAP binding cavity (Figure 3), leading to a substantial decrease in inhibitory potency of compound **98**. Moreover, the high LipE value of compound **98** as compared to compound **1** could be attributed only to its low logP (o/w) without an increase in biological activity.

Class II of FLAP antagonists contains 2,2-biaryl bicycloheptane as a common scaffold having diverse substituents at position 2 (R_1) and quinoline moiety at position 5 (R_2) of the exo aryl group (Figure 1). Ligand–protein interaction profiling of the final cluster (cluster 3, Supplementary Table S5) shows that the common scaffold orients itself towards the outer side-facing membrane and makes π -H interactions with B-L120, B-I119, and C-V21, whereas the quinoline moiety occupied the inside of the FLAP binding cavity and shows π -H interactions with B-A63, C-G24, and C-N23 (Figure 2D). Generally, R_1 is involved in making hydrogen bonds with B-F123 and B-K116 amino acid residues. Overall, a slight positive correlation ($R^2 = 0.27$) has been observed between inhibitory potency and lipophilicity for class III (Supplementary Figure S3). Moreover, a distinct SAR pattern was observed among compounds of class III. For instance, compound **10** ($\text{IC}_{50} = 1.1 \text{ nM}$, Supplementary Table S3), being the most potent and lipophilic ($c\text{LogP} = 7.88$, $\text{LipE} = 1.07$) member of this class, possesses oxadiazole-2-thione at R_1 ,

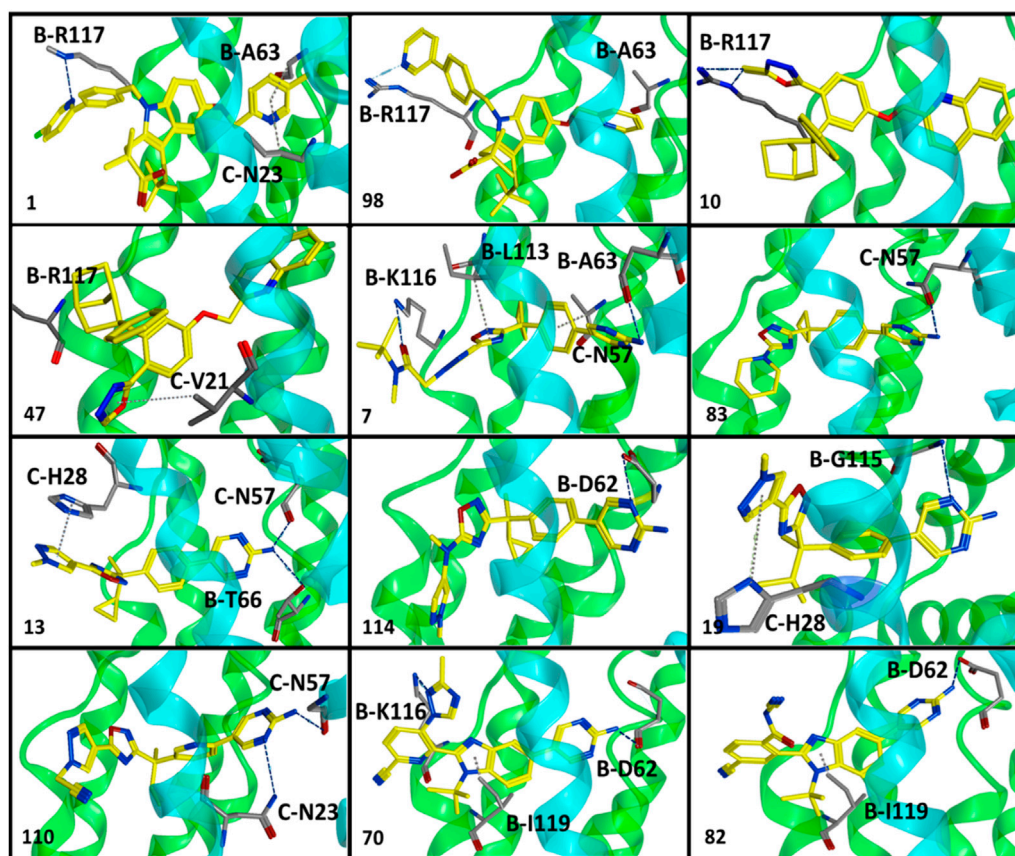


FIGURE 3 | Optimal binding poses of compounds displaying a distinct SAR pattern from all six classes of FLAP modulators. These poses were obtained from clusters with maximum docked ligands (common scaffold clustering) and were further employed for GRID-independent molecular descriptor (GRIND) analysis. Chain B is shown in green while chain C is depicted in blue color.

and absence of thiol substituent of oxadiazole at R_1 (**Supplementary Table S6**) resulted in approximately threefold decrease in inhibitory potency of compound **47** ($IC_{50} = 2.9$ nM, $cLogP = 7.37$, $LipE = 1.16$). The lipophilicities and $LipE$ values of compounds **10** and **47** are relatively the same and the difference in inhibitory potencies might be due to a distinct binding pattern. The final docking solution of compound **10** reveals the presence of two hydrogen bonds between the terminal sulfur of the oxadiazole-2-thione group at R_1 and $-NH_2$ of B-R117 (**Figure 3**). In compound **47**, only a π -H-bond interaction was found between the oxadiazole ring of R_1 and C-V21 that might be not very favorable and contribute to its low inhibitory. The positive contribution of negative ionizable moieties at the oxadiazole ring of R_1 towards inhibitory potency for class II FLAP antagonists is also evident from previous SAR studies (Chu et al., 2012).

Classes III, IV, and V are cyclobutylbenzene, cyclopropylethylbenzene, and dimethylpropylbenzene derivatives of oxadiazole, respectively (**Figure 1**). Unlike other three classes, no positive correlation between lipophilicity and inhibitory potency was observed for classes III, IV, and V (**Supplementary Figure S3**). It means that the difference in inhibitory potency might be due to the distinct interaction

pattern and $LipE$ values. All compounds of classes IV, V, and VI contain diverse substituents at R_1 and pyrimidinamine at R_2 (**Supplementary Table S6**), which occupies the inside of the FLAP binding cavity (**Figures 2E–G**). The final cluster of class III (cluster 3, **Supplementary Table S5**) reveals that the common scaffold shows π - π stacking with B-F114 and π -H-bond interactions with B-A63 (**Figure 2E**). R_2 forms hydrogen bonding with B-D62, C-C60, and π -H-bond interactions with C-N57 and C-N23, while R_1 seems to be involved in making hydrogen bonds with B-K116 and π -H interactions with B-I119. Compound **7** ($IC_{50} = 1.0$ nM, $cLogP = 3.13$, $LipE = 5.87$, **Supplementary Table S3**), being the most potent compound of class III, contains N-tert-butyl methylacetamide at the pyrazole ring of R_1 compared to compound **83** ($IC_{50} = 6.5$ nM, $cLogP = 2.68$, $LipE = 5.50$, **Supplementary Table S3**), which contains only piperidine ring at R_1 , resulting in a twofold decrease in its inhibitory potency (**Supplementary Table S6**). The final docking solution of compound **7** reveals that the carbonyl group of N-tert-butyl methylacetamide at R_1 forms a strong hydrogen bond with $-NH_2$ of B-K116 (**Figure 3**). However, for compound **83**, no interaction was observed between the piperidine ring of R_1 and amino acid residues of the FLAP binding cavity. The difference in the binding interaction

pattern of compounds **7** and **83** might be solely responsible for the difference in inhibitory potencies of both compounds as lipophilicities and LipE values are not significantly different. For class IV, ligand–protein interaction profiling of the final cluster (cluster 4, **Supplementary Table S2**) suggests that the common scaffold makes π -H interactions with B-T66, B-A63, C-G24, B-I119, and C-Val21, and π - π stacking with B-F114 (**Figure 2F**). Amino acid residues such as B-D62 and C-N57 present inside the FLAP binding cavity shows hydrogen bonding with R₂ whereas R₁ makes hydrogen bonds with B-K116 and π - π interactions with C-H28 and C-F25. Compounds **13** and **114** of class IV were selected to evaluate binding poses due to the distinct SAR pattern (**Supplementary Table S6**). Compound **13** (IC₅₀ = 1.3 nM, cLogP = 2.36, LipE = 6.52, **Supplementary Table S3**), being the most active from class IV, contains the terminal methyl at the pyrazole ring of R₁, while compound **114** (IC₅₀ = 29.0 nM, cLogP = 4.86, LipE = 2.67, **Supplementary Table S3**) contains N,1-dimethylpyrazol-4-amine at R₁, resulting in a twofold decrease in inhibitory potency. The final docking solution of compound **13** reveals that the pyrazole ring of R₁ is involved in making π - π stacks with C-H28, whereas no significant interaction was observed for terminal methyl (**Figure 3**). For compound **114**, the final binding pose suggests that the terminal pyrazole ring at R₁ is unable to show any interactions that might contribute to low inhibitory potency. The significantly low LipE value of compound **114** as compared to compound **13** suggests that gain in activity of compound **13** might be due to its distinct interaction pattern. Similarly, the ligand–protein interaction analysis of the final cluster (cluster 5, **Supplementary Table S5**) of class V points out that the scaffold makes π -H-bond interactions with B-L120, B-I119, B-A63, C-G24, and C-Val21, and π - π stacks with B-F123 and B-F114 (**Figure 2G**). R₂ is involved in making hydrogen bonds with B-D62 and C-N57 while R₁ makes hydrogen bonds with B-K116 and π - π contact with C-H28. Compound **19** (IC₅₀ = 1.6 nM, cLogP = 3.08, LipE = 5.71, **Supplementary Table S3**) contains terminal methyl at the pyrazole ring of R₁, whereas compound **110** (IC₅₀ = 23 nM, cLogP = 2.65, LipE = 4.98, **Supplementary Table S3**) possesses acetonitrile at the pyrazole ring of R₁ (**Supplementary Table S6**) and approximately two orders of magnitude decrease of inhibitory potency was observed for compound **110** as compared to compound **19**. The final binding pose of compound **19** reveals that the pyrazole ring of R₁ is involved in making π - π contact with C-H28 while terminal methyl could not make any interactions (**Figure 3**). The absence of interactions between acetonitrile at R₁ of compound **110** and amino acid residues of the FLAP binding pocket was likely the reason for the two orders of magnitude decrease in inhibitory potency of compound **110** as LipE and lipophilicity values of both compounds do not differ appreciably. Overall, compounds of classes III, IV, and V displayed better LipE values, but the high inhibitory potencies of highly active compounds are due to strong interactions among particular functional groups and amino acids of the FLAP binding cavity.

Class VI FLAP inhibitors are benzimidazole derivatives (**Figure 1**) having diverse substituents at R₁ and pyrimidinamine at R₂ around the benzimidazole scaffold. The compounds of class VI did not exhibit any correlation between

activity and lipophilicity (**Supplementary Figure S3**). The ligand–protein interaction profile of the final cluster indicated that the pyrimidinamine group orients itself towards the inner side of the FLAP binding cavity and is involved in making hydrogen bonds with B-D62 and C-N23, whereas the common scaffold occupies the between chains B and C and forms π -H bonding with B-I119, C-G24, and C-V21. The diverse R₁ is involved in making hydrogen bonds with B-K116 and π - π interactions with C-H28. Compounds of class VI showed a distinct SAR pattern; e.g., in compound **70** (IC₅₀ = 4.2 nM, cLogP = 2.54, LipE = 5.84), the pyridine moiety of R₁ contains methyl triazole at position three and its replacement with acetonitrile in compound **82** resulted in two orders of magnitude decrease in the inhibitory potency of compound **82** (IC₅₀ = 6.09 nM, cLogP = 2.18, LipE = 6.04). The selected binding pose of compound **70** indicated that the triazole ring is making hydrogen bonds with B-K116 whereas no interaction was observed between acetonitrile and amino acid residues of the FLAP binding cavity in the case of compound **82** (**Figure 3**). The compounds of class VI did not show any correlation with LipE, which means that the difference in binding interactions is the main driving factor behind the difference in activity.

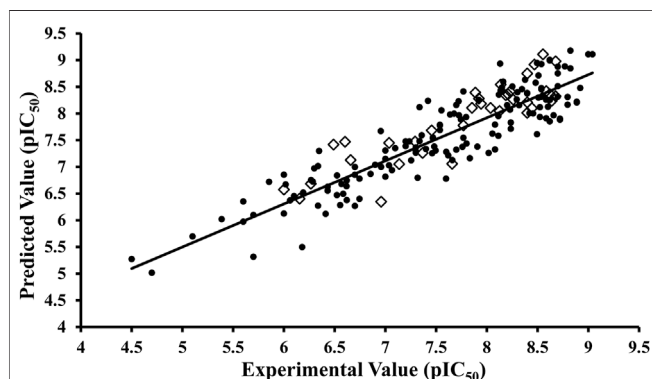
Overall, our criteria for the selection of compounds for molecular modeling studies were cLogP and LipE. However, our results indicate that only cLogP contributes slightly positively towards inhibitory potency for classes I and II, whereas for compounds of classes III, IV, V, and VI, the difference in interaction pattern might be exclusively responsible for the difference in inhibitory potency, as in these classes of FLAP inhibitors, the high LipE values were maintained due to loss in lipophilicity. In addition, our docking results suggest that heterocyclic moieties are involved in making π -H interactions with hydrophobic amino acid residues of the FLAP binding cavity. Therefore, the presence of pyridine, pyrimidine, pyridazine, pyrazole, triazole, and oxadiazole rings moderately increases not only the lipophilicity but also the inhibitory potency. Moreover, an increase or decrease in LipE values of FLAP inhibitors does not alter the inhibitory potencies in either way. Further docking poses obtained from multiple clusters with maximum docked ligands were employed to generate the vGRIND model.

3.4 GRID-Independent Molecular Descriptors Analysis

The selected binding poses of 151 (**Supplementary Table S3**) compounds of the training set obtained through common scaffold clustering of docking poses along with their inhibitory activity (pIC₅₀) values were implied in the pentacle v 1.07 software package that utilizes special alignment independent GRIND descriptors to develop a 3D-QSAR model. To correlate the inhibitory potencies with 3D structural features and to derive the most important pharmacophoric features of our training set, a partial least square model was developed on five principal components using the leave-one-out (LOO) cross-validation method, resulting in initial models with satisfactory values of variables. The inconsistent nodes were removed by one-

TABLE 2 | Statistical parameters obtained before and after application of fractional factorial design (FFD) on final GRIND model.

| Fractional factorial design cycle (FFD) | | | | | | | | | | | |
|---|-------|--------------------|------|---------|---------------|--------------|-------|--------------------|------|---------|---------------|
| Complete variable | | | | | | FFD1 | | | | | |
| Datasets | R^2 | q^2_{Loo} | SDEP | r^2_m | Delta r^2_m | Datasets | R^2 | q^2_{Loo} | SDEP | r^2_m | Delta r^2_m |
| Training set | 0.71 | 0.60 | 0.49 | 0.703 | 0.004 | Training set | 0.82 | 0.66 | 0.47 | 0.775 | 0.001 |
| Test set | 0.63 | 0.58 | 0.49 | 0.517 | 0.028 | Test set | 0.77 | 0.64 | 0.47 | 0.686 | 0.012 |

**FIGURE 4** | Activity interactive graph plot between predicted and actual experimental activity values. The graph plot displays separate data series for training (filled circles) and test (rhombus) set. R^2 for training set was observed as 0.82 and 0.77 for test set.

time application of the fractional factorial design (FFD) variable selection algorithm. The final GRIND model was obtained with good values of performance measures, $q^2 = 0.66$ and $R^2 = 0.82$, while the standard error of prediction (SDEP) was 0.47. The before and after FFD application statistics along with the r^2_m metric is shown in **Table 2**. The difference between actual and predicted activity values was less than one log unit for all 151 inhibitors of the training set as shown in **Figure 4**. The test set (**Supplementary Table S4**) was used for the evaluation of the final GRIND model, which predicted inhibitory potencies of test set compounds with a difference of less than one log unit for all compounds between experimental and predicted pIC_{50} values with R^2 value observed as 0.77 (**Figure 4**).

A PLS coefficient correlogram of the GRIND variables is shown in **Figure 5A** and describes important 3D structural features that directly/inversely correlate with the inhibitory potencies of the training set compounds. The PLS coefficient correlogram depicts that DRY-DRY, DRY-N1, DRY-TIP, and N1-TIP pair of probes positively contribute towards the inhibitory potency of chemically diverse FLAP inhibitors whereas no inverse contribution was observed by any variable. These variables are located at a certain distance within active inhibitors between substitutions at R_1 , R_2 , and common scaffolds.

More explicitly, the DRY-DRY correlogram in **Figure 5A** shows the presence of two hydrophobic contours (HYD1 and HYD2) at a mutual distance of 16.00–16.40 Å in a virtual receptor

site of highly active FLAP inhibitors $pIC_{50} > 7.5$. For class I, the distance is present between the pyridine ring of R_1 and the phenyl ring of R_2 ; for class II, it is observed between the quinoline group and the endo aryl moiety of the common scaffold; for classes III, IV, and V, it is present between the pyrazole ring at R_1 position and the phenyl ring of the common scaffold; and for class VI, it was observed between phenyl of the common scaffold and pyrimidinamine of R_2 (**Figures 5B–E**). Furthermore, the backstage projection of the actual FLAP structure onto the identified hotspots revealed the presence of complementary hydrophobic amino acid residues such as B-A63, B-I119, and B-L120. This further strengthened our docking outcomes as all of these amino acid residues are involved in making extensive π -H interactions with dataset compounds. Additionally, a recent pharmacophore study of Olgac et al. revealed that four hydrophobic features are important in most potent indole- and oxadiazole-based FLAP inhibitors (Olgac et al., 2020).

Similarly, DRY-N1 (**Figure 5A**) explicates the positive contribution of one hydrophobic (HYD3) and one hydrogen bond acceptor (HBA1) at a mutual distance of 16.40–16.80 Å within active FLAP modulators. Interestingly, this distance was observed in all highly active FLAP modulators $pIC_{50} > 7.5$ and absent in all less-active compounds $pIC_{50} < 7.5$. Briefly, for class I, it is observed between the terminal negative ionizable moiety present at R_2 and the pyridine ring of R_1 ; in class II, it is observed between the quinoline group and pyrazole ring; for classes III, IV, and V, it is present between the pyrimidinamine group of R_2 and oxadiazole ring; and for class VI, it was observed between pyrimidinamine and pyridine of R_1 as displayed in **Figures 5B–E**. Projecting actual FLAP structure onto the identified virtual hotspots revealed the presence of hydrophobic amino acids B-F114, B-A63, and C-G24 and complementary amide groups in the B-K116 and B-R117 amino acid residues within the FLAP binding cavity that further complements the accuracy of our model. These results further reinforce our docking outcomes, which demonstrated the importance of B-A63 and B-K116 for the hydrophobic and hydrogen bonding interactions within the FLAP binding cavity. These outcomes are also in accord with another pharmacophore-based study that demonstrated the importance of hydrophobic and hydrogen bond acceptor features in the highly active indole- and biaryl bicycloheptane-based FLAP inhibitors (Temml et al., 2017).

Moreover, DRY-TIP correlogram (**Figure 5A**) portrays the presence of one hydrophobic (HYD4) and one shape-based feature (TIP1) that positively contribute towards the inhibitory potency of FLAP inhibitors. For the sharpest peak, the two

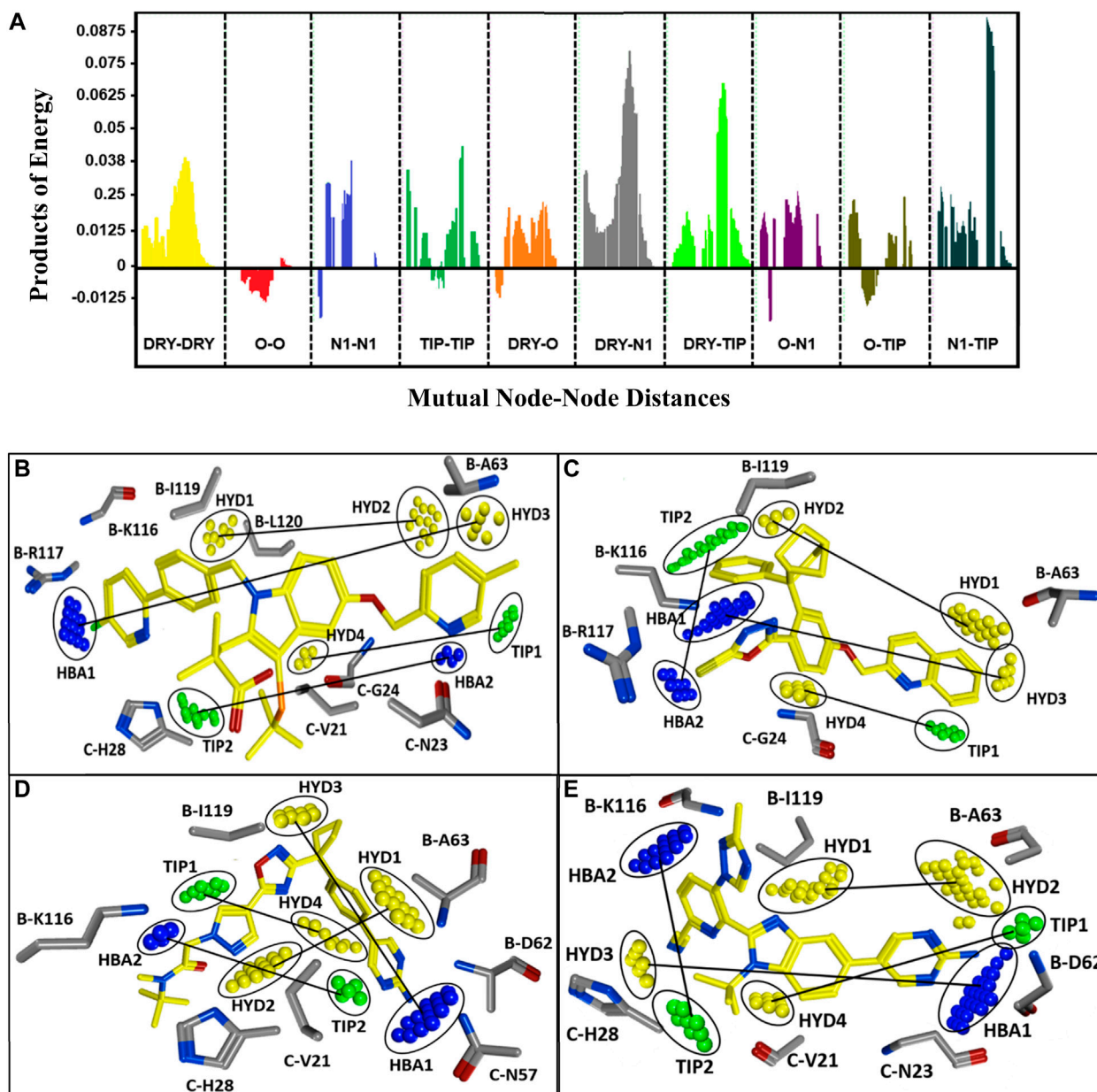


FIGURE 5 | (A) Correlation plot of PLS coefficients representing the pair of probes contributing positively (peaks above 0) or negatively (peaks below 0) towards the inhibitory potencies of training set compounds. The positive contribution towards pIC_{50} of FLAP inhibitors has been depicted by DRY-DRY (two hydrophobic), DRY-N1 (one hydrophobic and one hydrogen bond acceptor), DRY-TIP (one hydrophobic and one steric), and N1-TIP (one hydrogen bond acceptor and one steric) variables. The variables are present in all highly active FLAP compounds and are located at mutual distances of 16.00–16.40 Å, 16.40–16.80 Å, 18.00–18.40 Å, and 17.20–17.60 Å, respectively. **(B)** The identified hotspots on most active indole-based FLAP inhibitor (compound 1) of training set with projection of actual FLAP structure. Hydrophobic features are depicted in yellow, hydrogen bond acceptors are in blue, while steric hotspots are depicted in green color. The two hydrophobic hotspots (HYD1 and HYD2) are located between two aromatic moieties, one hydrophobic (HYD3) and one hydrogen bond acceptor feature (HBA1) are present between aromatic rings and terminal negative ionizable substitution, one hydrophobic (HYD4) and steric feature (TIP1) can be spotted between aromatic ring and indole scaffold, while one hydrogen bond acceptor (HBA2) and one steric (TIP2) hotspot are present between dimethylbutanoic acid and pyridine ring. **(C)** The most active compound (compound 10) of class II with mapping of complemented amino acids on the recognized contours. **(D)** The most active of class III (compound 7), which is also the most active compound from oxadiazole-based FLAP antagonists (classes III, IV, and V) and mapped hotspots along with projection of complementary amino acids of FLAP binding cavity. Due to high structural similarity, the features were also observed at the same positions in all active compounds of classes IV and V. **(E)** The compound (70) from class VI with identified hotspots and corresponding amino acids.

contours are present at a mutual distance of 18.00–18.40 Å, indicating the distance between indole scaffold and R₁ for class I; exoaryl of common scaffold and quinoline for class II; phenyl ring of scaffold and pyrazole ring of R₁ for classes III, IV, and V; and pyrazole of common scaffold and pyrimidinamine for class VI (Figures 5B–E). Two identified contours (HYD4 and TIP) were mapped on the actual FLAP binding site, and interestingly, the hydrophobic region in all active compounds is complementary to hydrophobic amino acids C-V21, C-G24, and C-H28. It is also in accordance with our docking findings where many compounds of our dataset are involved in making π -H interactions and π - π stacking with these amino acids. The green contour elucidates a steric hotspot region, and it defines the 3D molecular shape of FLAP inhibitors.

The last selected peak N1-TIP (Figure 5A) represents the presence of one hydrogen bond acceptor (HBA2) and one shape-based feature (TIP2) at a mutual distance of 17.20–17.60 Å within highly active FLAP inhibitors. The two features at this distance contributes positively towards the inhibitory potency of compounds against FLAP. The hydrogen bond acceptor hotspot in the virtual receptor site of FLAP is complemented by the presence of B-K116, B-D62, C-N57, and C-N23 amino acids in the actual receptor site when we mapped the FLAP structure onto the identified N1 (HBA2) hotspot. The -NH₂ and carbonyl groups of these amino acids are involved in making hydrogen bonds with active FLAP modulators as evident from our docking studies and pose analysis. Moreover, these features have been observed in all active FLAP inhibitors pIC₅₀ > 7.5 while they are absent in less active compounds pIC₅₀ < 7.5. For class I, the distance is present between the pyrimidine ring at R₂ position and dimethylbutanoic acid; for class II it is observed between substituents at the pyrazole ring of R₁ and endo aryl of scaffold; for classes III, IV, and V, it is present between the pyrimidine amine of R₂ and pyrazole ring at R₁ position; and for class VI, it is present between the triazole of R₁ and tertbutyl of the common scaffold (Figures 5B–E). The TIP probe signifies the importance of a steric hotspot at a distance of 17.20–17.60 Å from the hydrogen bond acceptor feature.

Generally, our study provided a deeper understanding of three-dimensional requirements of diverse inhibitor binding within the FLAP binding cavity by mapping the mutual distances of important pharmacophoric features (four hydrophobic, two hydrogen bond acceptor, and two steric features) as well as the complementary distances of the important interacting amino acid residues (B-L120, B-I119, B-R117, B-K116, B-F114, B-A63, B-D62, C-H28, C-G24, C-N23, and C-V21). Previous docking studies also revealed that highly potent FLAP modulators result in π - π stacking with C-H28, hydrophobic interactions with B-L120, B-I119, and hydrogen bonding with B-R117, B-K116, and B-D62 (Ma et al., 2008). Overall, the binding hypothesis generated for each class within the FLAP binding cavity was complementary with our GRIND model, which predicted the inhibitory potencies of validation and test sets with reasonable accuracy, indicating the fitness of our model. Based on our current findings, we suggest that the high inhibitory potency of a compound against FLAP can be achieved by (1) increasing the hydrogen bond acceptor

strength on at least one substitution position (R₁ or R₂); (2) insertion of heterocyclic moieties such as pyridine, pyrimidine, pyridazine, pyrazole, and triazole at each side of the common scaffold to increase hydrophobic strength; and (3) maintaining a distance of 16.00–16.40 Å between two hydrophobic groups (aromatic rings) and 16.40–16.80 Å between hydrophobic and hydrogen bond acceptor groups.

4 DISCUSSION

Since high levels of leukotrienes have been reported in multiple pathophysiological conditions in the past 3 decades, leukotriene synthesis pathway has been targeted at many levels while FLAP has received the greatest focus because it initiates the biological synthesis of leukotrienes *via* leukotriene synthesis pathway (Massoumi and Sjölander 2007; Bryda and Wątroba 2018; Jo-Watanabe, Okuno, and Yokomizo 2019). Several practices have been made to propose potent FLAP modulators, and many of them have shown good clinical efficacy. However, not a single molecule could be able to change into the status of “drug”. The focus of the present study is to unveil the two- and three-dimensional structural requirements of FLAP modulators.

First to demonstrate the important two-dimensional structural features, supervised ML approach was adapted over classical 2D QSAR modeling. The preference was made for two reasons: (1) to escape the alignment step as FLAP modulators are highly diverse in nature, and (2) evidence from the past strengthens the adaptation of ML for quantitative structure–activity relationship studies (Tsou et al., 2020; Gupta et al., 2021). We developed multiple ML models including XGboost (GBDT), random forest (RF), decision tree (DT), support vector machine (SVM), logistic regression (LR), and multilayer perceptron (MLP), and in comparison, XGBoost and RF were able to classify our training set and predict the test set with significant classification and prediction accuracies. Moreover, recursive feature elimination with cross-validation (RFECV) captured relevant features or 2D descriptors, which are mainly participating in the classification of highly active and least active FLAP inhibitors.

Further molecular modeling studies were performed to vaticinate the important three-dimensional pharmacophoric features instead of ML. The preference was made because (1) three-dimensional structural properties are highly dependent on binding poses, and (2) the GRIND model not only explains the important molecular interaction fields but also distances between them along with important amino acid residues by creating a virtual receptor site (Shafi and Jabeen 2017). Before implication of molecular modeling strategies, the dataset was first subjected to calculation of LipE and cLogP. The purpose of LipE-based lead optimization is to improve LipE while maintaining an appropriate range of logP for the optimization of potency and ADME properties. The increased potency of a compound with eque-LipE to the reference ligand demonstrates that the increase in lipophilicity alone is responsible for the increased potency, although other factors associated with the specific structural change cannot be ruled out. Also, an increase in LipE of a

compound suggests that an increase in potency is beyond lipophilicity increases alone and other factors such as transport to the target and hydrogen bonding strength within the protein binding site could be associated with this response. In total, 238 out of 658 demonstrated the LipE value greater than five, which is the optimal threshold, and only 136 demonstrated cLogP between the optimal range of 2–3.5. Herein, we selected 187 compounds having LipE greater than one and cLogP greater than two because FLAP is an integral membrane protein, which means that compounds should possess a high lipophilicity value for efficient binding. Also, FLAP modulators in clinical trials usually possess lipophilicity >3. The dataset of 187 compounds was divided into six distinct classes based on the common scaffold with subsequent docking into the FLAP binding pocket. Our docking results indicated that the FLAP binding pocket can cater diverse anti-inflammatory compounds and they bind between chains B and C. The ligand–protein interaction profile of selected FLAP modulators revealed that mostly B-R117, BK-116, C-N57, C-N23, and B-D62 FLAP amino acid residues are involved in making hydrogen bonds; B-A63, B-L119, B-L120, B-V21, and C-G24 make π -H-bond interactions, whereas C-H28 is involved in forming π - π contact with FLAP modulators. Also, for classes I and II, a moderate correlation was observed between lipophilicity and inhibitory potency, whereas for compounds of classes III, IV, V, and VI, an increase or decrease in lipophilicity or LipE did not alter the inhibitory potency in either way or *vice versa*.

To select the most probable binding poses, common scaffold clustering was performed because using GRID-independent molecular descriptors, analysis of 3D structural features is highly dependent on 3D confirmations of the molecules (Pastor et al., 2000). Multiple clusters at 3.5 Å RMSD were generated and binding poses from clusters with maximum number of docked ligands were further used to build the GRIND model. The reliability of binding pose selection *via* common scaffold clustering for generation of the GRIND model can be explained by satisfactory statistical results obtained for the final GRIND model. Furthermore, the model signifies the positive contribution of four hydrophobic, two hydrogen bond acceptor, and two steric features towards the inhibitory potency of FLAP modulators. The identified hotspots or pharmacophoric features were successfully mapped onto the highly active FLAP modulators followed by projection of the actual receptor site, which revealed the presence of corresponding amino acid residues. Overall, our GRIND model suggested that (1) two hydrophobic features should be present at a mutual distance of 16.00–16.40 Å, (2) one hydrophobic and hydrogen bond acceptor feature should be present at a distance of 16.40–16.80 Å, (3) the distance between hydrophobic and steric feature should be 18.00–18.40 Å, and (4) it should be 17.20–17.60 Å between hydrogen bond acceptor and steric features. The importance of hydrophobic and hydrogen bond acceptor features has also been demonstrated by previous studies (Temml et al., 2017; Olgac et al., 2020).

Based on these findings, further analyses will focus on virtual screening from both ML and GRIND models followed by selection of common compounds. The common hits can further be structurally tuned and optimized. The ML and GRIND model will allow internal inspection of FLAP modulators, before validating them using predictions on vendor libraries, purchase, and testing.

5 CONCLUSION

The current study deals with the development of ML models and a GRIND model on a diverse series of FLAP inhibitors. First of all, our ML models signify some important 2D descriptors, and the best-performing model (XG-Boost) has successfully classified the active and inactive compounds present in our training set exhibiting 91% overall classification accuracy. The subsequent screening of test set from the model resulted in 90% prediction accuracy, which further accentuates the efficiency of the model. Secondly, docking studies reveal that hydrogen bonding and hydrophobic interactions are critical for binding of FLAP inhibitors. Further common scaffold-based clustering revealed the optimal binding mode of structurally diverse inhibitors and aided in determination of their molecular basis of interaction within the FLAP binding cavity. Thirdly, the most probable binding poses were utilized for GRIND model development, which showed valid statistical results having an R^2 of 0.82 and a q^2 of 0.66. Additionally, the GRIND model predicted all compounds of training and test set with an activity difference of less than one log unit. Overall, our GRIND model illustrated that four hydrophobic, two hydrogen bond acceptor, and two steric features are critical for achieving high inhibitory potency against FLAP. All the features were successfully complemented by the docking studies highlighting the significance of respective amino acid residues such as B-L120, B-L119, B-A63, C-H28, C-G24, and C-V21 for hydrophobic interactions and B-R117, B-K116, D-62, and C-N57 for hydrogen bonding. In general, application of ML, docking analysis, common scaffold clustering, and GRIND modeling to predict the 2D structural requirements as well as the 3D molecular basis of interaction of diverse FLAP inhibitors could potentially guide the development of more potent chemotypes for the treatment of inflammatory disorders requiring anti-leukotriene therapy.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization, HK and IJ; methodology, HK and IJ; software, IJ; validation, HK and IJ; formal analysis, HK; investigation, HK; resources, IJ; data curation, HK; writing—original draft preparation, HK; writing—review and editing, IJ; visualization, HK; supervision, IJ; project administration, IJ.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.825741/full#supplementary-material>

REFERENCES

- Antoniou, S. A. (2014). Targeting 5-Lipoxygenase-Activating Protein in Asthma and Chronic Obstructive Pulmonary Disease. *Expert Opin. Ther. Targets* 18 (11), 1285–1292. doi:10.1517/14728222.2014.945425
- Arnott, J. A., and Planey, S. L. (2012). The Influence of Lipophilicity in Drug Discovery and Design. *Expert Opin. Drug Discov.* 7, 863–875. doi:10.1517/17460441.2012.714363
- Bain, G., King, C. D., Rewolinski, M., Schaab, K., Santini, A. M., Shapiro, D., et al. (2010). Pharmacodynamics and Pharmacokinetics of AM103, a Novel Inhibitor of 5-Lipoxygenase-Activating Protein (FLAP). *Clin. Pharmacol. Ther.* 87 (4), 437–444. doi:10.1038/clpt.2009.301
- Banoglu, E., Çalişkan, B., Luderer, S., Eren, G., Özkan, Y., Altenhofen, W., et al. (2012). Identification of Novel Benzimidazole Derivatives as Inhibitors of Leukotriene Biosynthesis by Virtual Screening Targeting 5-Lipoxygenase-Activating Protein (FLAP). *Bioorg. Med. Chem.* 20 (12), 3728–3741. doi:10.1016/j.bmc.2012.04.048
- BioByte - Bio-Loom (2021). BioByte - Bio-Loom. Available at: <http://biobyte.com/bb/prod/bioloom.html> (Accessed November 10, 2020).
- Breiman, L. (2001). Random Forests. *Machine Learn.* 45 (1), 5–32. doi:10.1023/A:1010933404324
- Brieman, F., and Olshen, S. (2012). Classification A Nd R Egression Trees. *Chapman & Hall/Crc* 66, 37–39.
- Bryda, J., and Wątroba, S. (2018). The Proinflammatory Role of Lipoxygenases in Rheumatoid Arthritis. *J. Pre Clin. Clin. Res.* 12 (4), 129–134. doi:10.26444/JPCR/99597
- Chang, C.-C., and Lin, C.-J. 2021 “LIBSVM: A Library for Support Vector Machines.”
- Chen, A. A., and Pappu, R. V. (2007). Parameters of Monovalent Ions in the AMBER-99 Forcefield: Assessment of Inaccuracies and Proposed Improvements. *J. Phys. Chem. B* 111 (41), 11884–11887. doi:10.1021/jp0765392
- Chen, T., and Guestrin, C. (2016). “XGBoost.” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13–17: 785–794. doi:10.1145/2939672.2939785
- Chu, L., Armstrong, H. M., Chang, L. L., Cheng, A. F., Colwell, L., Cui, J., et al. (2012). Evaluation of Endo- and Exo-Aryl-Substitutions and Central Scaffold Modifications on Diphenyl Substituted Alkanes as 5-Lipoxygenase Activating Protein Inhibitors. *Bioorg. Med. Chem. Lett.* 22 (12), 4133–4138. doi:10.1016/j.bmcl.2012.04.064
- Cortes, C., Vapnik, V., and Saitta, L. (1995). Support-Vector Networks. *Mach Learn.* 20 (3), 273–297. doi:10.1007/BF00994018
- Dahlén, B., Kumlin, M., Ihre, E., Zetterström, O., and Dahlen, S. E. (1997). Inhibition of Allergen-Induced Airway Obstruction and Leukotriene Generation in Atopic Asthmatic Subjects by the Leukotriene Biosynthesis Inhibitor BAYx 1005. *Thorax* 52 (4), 342–347. doi:10.1136/thx.52.4.342
- Diamant, Z., Timmers, M. C., van der Veen, H., Friedman, B. S., De Smet, M., Depré, M., et al. (1995). The Effect of MK-0591, a Novel 5-Lipoxygenase Activating Protein Inhibitor, on Leukotriene Biosynthesis and Allergen-Induced Airway Responses in Asthmatic Subjects *In Vivo*. *J. Allergy Clin. Immunol.* 95 (1–2), 42–51. doi:10.1016/S0091-6749(95)70151-6
- Durán, A., Martínez, G. C., and Pastor, M. (2008). Development and Validation of AMANDA, a New Algorithm for Selecting Highly Relevant Regions in Molecular Interaction Fields. *J. Chem. Inf. Model.* 48 (9), 1813–1823. doi:10.1021/ci800037t
- Ericsson, H., Nelander, K., Heijer, M., Kjaer, M., Lindstedt, E. L., Albayaty, M., et al. (2020). Phase 1 Pharmacokinetic Study of AZD5718 in Healthy Volunteers: Effects of Coadministration with Rosuvastatin, Formulation and Food on Oral Bioavailability. *Clin. Pharmacol. Drug Dev.* 9 (3), 411–421. doi:10.1002/cpdd.756
- Evans, J. F., Lévellé, C., Mancini, J. A., Prasit, P., Thérien, M., Zamboni, R., et al. (1991). 5-Lipoxygenase-Activating Protein Is the Target of a Quinoline Class of Leukotriene Synthesis Inhibitors. *Mol. Pharmacol.* 40 (1), 22–27.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *J. Machine Learn. Res.* 9, 1871–1874.
- Ferguson, A. D., McKeever, B. M., Xu, S., Wisniewski, D., Miller, D. K. Ting. Ting. Yamin, Yamin, T. T., et al. (2007). Crystal Structure of Inhibitor-Bound Human 5-Lipoxygenase-Activating Protein. *Science* 317 (5837), 510–512. doi:10.1126/science.1144346
- Folco, G., Rossoni, G., Buccellati, C., Berti, F., Maclouf, J., and Sala, A. (2000). Leukotrienes in Cardiovascular Diseases. *Am. J. Respir. Crit. Care Med.* 161 (2 II), S112–S116. doi:10.1164/ajrccm.161.supplement_1.lta-22
- Frenette, R., Hutchinson, J. H., Léger, S., Thérien, M., Brideau, C., Chan, C. C., et al. (1999). Substituted Indoles as Potent and Orally Active 5-Lipoxygenase Activating Protein (FLAP) Inhibitors. *Bioorg. Med. Chem. Lett.* 9 (16), 2391–2396. doi:10.1016/S0960-894X(99)00399-6
- Friedman, B. S., Bel, E. H., Buntinx, A., Tanaka, W., Han, Y. H., Shingo, S., et al. (1993). Oral Leukotriene Inhibitor (MK-886) Blocks Allergen-Induced Airway Responses. *Am. Rev. Respir. Dis.* 147 (4), 839–844. doi:10.1164/ajrccm/147.4.839
- Gajo, G. C., de Assis, T. M., Assis, L. C., Leticia, C. A., Ramalho, T. C., da Cunha, E. F. F., et al. (2016/2016). Quantitative Structure-Activity Relationship Studies for Potential Rho-Associated Protein Kinase Inhibitors. *J. Chem.* 2016, 1–12. doi:10.1155/2016/9198582
- Ghosh, A., Chen, F., Thakur, A., and Hong, H. (2016). Cysteinyl Leukotrienes and Their Receptors: Emerging Therapeutic Targets in Central Nervous System Disorders. *CNS Neurosci. Ther.* 22 (12), 943–951. doi:10.1111/cns.12596
- Glorot, X., and Yoshua, B. (2021) Understanding the Difficulty of Training Deep Feedforward Neural Networks.
- Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., and Kumar, P. (2021). Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery. *Mol. Divers.* 25 (April), 1315–1360. doi:10.1007/S11030-021-10217-3
- Gür, Z. T., Çalişkan, B., Banoglu, E., and Banoglu, Erden. (2018). Drug Discovery Approaches Targeting 5-Lipoxygenase-Activating Protein (FLAP) for Inhibition of Cellular Leukotriene Biosynthesis. *Eur. J. Med. Chem.* 153, 34–48. doi:10.1016/j.ejmech.2017.07.019
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002/2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learn.* 46 (1 461), 389–422. doi:10.1023/A:1012487302797
- Haeggström, J. Z. (2000). “Structure, Function, and Regulation of Leukotriene A4 Hydrolase” *American Journal of Respiratory and Critical Care Medicine*. American Lung Association. doi:10.1164/ajrccm.161.supplement_1.lta-6
- Haykin, S. (2009). *Neural Networks and Learning Machines*. Third Edition. New York, Boston San, Francisco London, Toronto Sydney, Tokyo Singapore, Madrid Mexico.
- Hedi, H., and Norbert, G. (2004). 5-Lipoxygenase Pathway, Dendritic Cells, and Adaptive Immunity. *J. Biomed. Biotechnol.* 2004, 99–105. doi:10.1155/S110724304310041
- Jabeen, I., Pleban, K., Rinner, U., Chiba, P., and Ecker, G. F. (2012). Structure-Activity Relationships, Ligand Efficiency, and Lipophilic Efficiency Profiles of Benzophenone-type Inhibitors of the Multidrug Transporter P-Glycoprotein. *J. Med. Chem.* 55 (7), 3261–3273. doi:10.1021/jm201705f
- Jakschik, B. A., and Kuo, C. G. (1983). Characterization of Leukotriene A4 and B4 Biosynthesis. *Prostaglandins* 25 (6), 767–782. doi:10.1016/0090-6980(83)90002-3
- Jedlitschky, G., and Keppler, D. (2002). “Transport of Leukotriene C4 and Structurally Related Conjugates” *Vitamins and Hormones*. Academic Press. doi:10.1016/s0083-6729(02)64005-1
- Jo-Watanabe, A., Okuno, T., and Yokomizo, T. (2019). The Role of Leukotrienes as Potential Therapeutic Targets in Allergic Disorders. *Int. J. Mol. Sci.* 20 (14). doi:10.3390/IJMS20143580
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* 267 (3), 727–748. doi:10.1006/jmbi.1996.0897
- Leeson, P. D., and Springthorpe, B. (2007). The Influence of Drug-like Concepts on Decision-Making in Medicinal Chemistry. *Nat. Rev. Drug Discov.* 6 (11), 881–890. doi:10.1038/nrd2445
- Lemurell, M., Ulander, J., Winiwarter, S., Dahlén, A., Davidsson, Ö., Emténäs, H., et al. (2015). Discovery of AZD6642, an Inhibitor of 5-Lipoxygenase Activating Protein (FLAP) for the Treatment of Inflammatory Diseases. *J. Med. Chem.* 58 (2), 897–911. doi:10.1021/jm501531v
- Liaw, A., and Wiener, M. (2002). Classification and Regression by. *RandomForest* 2 (3).

- Liu, M., and Yokomizo, T. (2015). The Role of Leukotrienes in Allergic Diseases. *Allergol. Int.* 64, 17–26. doi:10.1016/j.alit.2014.09.001
- Lorrain, D. S., Bain, G., Charles, C., Correa, L. D., Santini, A. M., Chapman, C., et al. (2010). Pharmacology of AM803, a Novel Selective Five-Lipoxygenase-Activating Protein (FLAP) Inhibitor in Rodent Models of Acute Inflammation. *Eur. J. Pharmacol.* 640 (1–3), 211–218. doi:10.1016/j.ejphar.2010.05.003
- Ma, X., Zhou, L., Zuo, Z.-L., Liu, J., Yang, M., and Wang, R.-W. (2008). Molecular Docking and 3-D QSAR Studies of Substituted 2,2-Bisaryl-Bicycloheptanes as Human 5-Lipoxygenase-Activating Protein (FLAP) Inhibitors. *QSAR Comb. Sci.* 27 (9), 1083–1091. doi:10.1002/qsar.200810053
- Macdonald, D., Brideau, C., Chan, C. C., Falgoutyret, J. P., Falgoutyret, J. P., Frenette, R., et al. (2008). Substituted 2,2-Bisaryl-Bicycloheptanes as Novel and Potent Inhibitors of 5-Lipoxygenase Activating Protein. *Bioorg. Med. Chem. Lett.* 18 (6), 2023–2027. doi:10.1016/j.bmcl.2008.01.105
- Mancini, J. A., Coppolino, M. G., Klassen, J. H., Klassen, J. H., Charleson, S., and Vickers, P. J. (1994). The Binding of Leukotriene Biosynthesis Inhibitors to Site-Directed Mutants of Human 5-Lipoxygenase-Activating Protein. *Life Sci.* 54 (9), PL137–42. doi:10.1016/0024-3205(94)00872-8
- Massoumi, R., and Sjölander, A. (2007). The Role of Leukotriene Receptor Signaling in Inflammation and Cancer. *ScientificWorldJournal* 7 (September), 1413–1421. doi:10.1100/TSW.2007.200
- Mauri, A. (2020). AlvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. *Methods Pharmacol. Toxicol.* 801–820. doi:10.1007/978-1-0716-0150-1_32
- McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 511.
- Nekoei, M., Mohammadhosseini, M., and Pourbasheer, E. (2015/2015). QSAR Study of VEGFR-2 Inhibitors by Using Genetic Algorithm-Multiple Linear Regressions (GA-MLR) and Genetic Algorithm-Support Vector Machine (GA-SVM): A Comparative Approach. *Med. Chem. Res.* 24 (7), 3037–3046. doi:10.1007/S00044-015-1354-4
- Olga, A., Carotti, A., Kretzer, C., Zergiebel, S., Seeling, A., Garscha, U., et al. (2020). Discovery of Novel 5-Lipoxygenase-Activating Protein (FLAP) Inhibitors by Exploiting a Multistep Virtual Screening Protocol. *J. Chem. Inf. Model.* 60 (3), 1737–1748. doi:10.1021/acs.jcim.9b00941
- Pastor, M., Cruciani, G., McLay, I., Pickett, S., and Clementi, S. (2000). GRIND-Independent Descriptors (GRIND): A Novel Class of Alignment-independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* 43 (17), 3233–3243. doi:10.1021/jm000941m
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Bertrand, T., Grisel, O., et al. (2012). Scikit-Learn: Machine Learning in Python. *J. Machine Learn. Res.* 12 (January), 2825–2830.
- Peters-Golden, M. (1998). Cell Biology of the 5-Lipoxygenase Pathway. *Am. J. Respir. Crit. Care Med.* 157 (6 Pt 1), S227–S228. doi:10.1164/ajrccm.157.6.mar4
- Peters-Golden, M., and Brock, T. G. (2003). 5-Lipoxygenase and FLAP. *Prostaglandins, Leukot. Essent. Fatty Acids* 69 (2–3), 99–109. doi:10.1016/S0952-3278(03)00070-X
- Pourbasheer, E., Vahdani, S., Malekzadeh, D., Aalizadeh, R., and Ebadi, A. (2017). QSAR Study of 17 β -HSD3 Inhibitors by Genetic Algorithm-Support Vector Machine as a Target Receptor for the Treatment of Prostate Cancer. *Iran J. Pharm. Res.* 16 (3), 966–980.
- Quinlan, J. R. (1986). Induction of Decision Trees. *Mach. Learn.* 1 (1), 81–106. doi:10.1007/BF00116251
- Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., and Das, R. N. (2013). Some Case Studies on Application of "r(m)2" Metrics for Judging Quality of Quantitative Structure-Activity Relationship Predictions: Emphasis on Scaling of Response Data. *J. Comput. Chem.* 34 (12), 1071–1082. doi:10.1002/JCC.23231
- Shafi, T., and Jabeen, I. (2017). Grid-Independent Descriptors (GRIND) Analysis and SAR Guided Molecular Docking Studies to Probe Selectivity Profiles of Inhibitors of Multidrug Resistance Transporters ABCB1 and ABCG2. *Curr. Cancer Drug Targets* 17 (2), 177–190. doi:10.2174/1568009616666160901094140
- Smyrniotis, C. J., Barbour, S. R., Xia, Z., Hixon, M. S., and Holman, T. R. (2014). ATP Allosterically Activates the Human 5-Lipoxygenase Molecular Mechanism of Arachidonic Acid and 5(S)-Hydroperoxy-6(E),8(Z),11(Z),14(Z)-Eicosatetraenoic Acid. *Biochemistry* 53 (27), 4407–4419. doi:10.1021/bi401621d
- Stock, N., Bacceti, C., Bain, G., Chapman, C., Correa, L., Darlington, J., et al. (2010). 5-Lipoxygenase-activating Protein Inhibitors. Part 3: 3-[3-Tert-Butylsulfanyl-1-[4-(5-Methoxy-Pyrimidin-2-Yl)-Benzyl]-5-(5-Methyl-Pyridin-2-Yl-methoxy)-1h-Indol-2-Yl]-2,2-Dimethyl-Propionic Acid (AM643)-A Potent FLAP Inhibitor Suitable for Topical Administration. *Bioorg. Med. Chem. Lett.* 20 (15), 4598–4601. doi:10.1016/j.bmcl.2010.06.011
- Stock, N. S., Bain, G., Zunic, J., Li, Y., Ziff, J., Roppe, J., et al. (2011). 5-Lipoxygenase-Activating Protein (FLAP) Inhibitors. Part 4: Development of 3-[3-Tert-Butylsulfanyl-1-[4-(6-Ethoxypyridin-3-Yl)Benzyl]-5-(5-Methylpyridin-2-Yl-methoxy)-1 H -Indol-2-Yl]-2,2-Dimethylpropionic Acid (AM803), a Potent, Oral, once Daily FLAP in. *J. Med. Chem.* 54 (23), 8013–8029. doi:10.1021/jm2008369
- Temml, V., Garscha, U., Romp, E., Schubert, G., Gerstmeier, J., Kutil, Z., et al. (2017). Discovery of the First Dual Inhibitor of the 5-Lipoxygenase-Activating Protein and Soluble Epoxide Hydrolase Using Pharmacophore-Based Virtual Screening. *Sci. Rep.* 7 (January), 42751–42813. doi:10.1038/srep42751
- Tsou, L. K., Yeh, S.-H., Ueng, S.-H., Chang, C.-P., Song, J.-S., Wu, M.-H., et al. (2020). Comparative Study between Deep Learning and QSAR Classifications for TNBC Inhibitors and Novel GPCR Agonist Discovery. *Sci. Rep.* 10, 1–11. doi:10.1038/s41598-020-73681-1
- Tu, G., Qin, Z., Huo, D., Zhang, S., and Yan, A. (2020). Fingerprint-Based Computational Models of 5-Lipo-Oxygenase Activating Protein Inhibitors: Activity Prediction and Structure Clustering. *Chem. Biol. Drug Des.* 96 (3), 931–947. doi:10.1111/cbdd.13657
- Woods, J. W., Coffey, M. J., Brock, T. G., Singer, I. I., Singer, I. I., and Peters-Golden, M. (1995). 5-Lipoxygenase Is Located in the Euchromatin of the Nucleus in Resting Human Alveolar Macrophages and Translocates to the Nuclear Envelope upon Cell Activation. *J. Clin. Invest.* 95 (5), 2035–2046. doi:10.1172/JCI117889
- Zeng, X., and Martinez, T. R. (2010). Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation 12, 1–12. doi:10.1080/095281300146272

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Khan and Jabeen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Impact of the Secondary Binding Pocket on the Pharmacology of Class A GPCRs

Attila Egyed, Dóra Judit Kiss and György M. Keserű*

Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Budapest, Hungary

OPEN ACCESS

Edited by:

Leonardo L. G. Ferreira,
University of São Paulo, Brazil

Reviewed by:

Yinglong Miao,
University of Kansas, United States
Marcel Bermudez,
Freie Universität Berlin, Germany

*Correspondence:

György M. Keserű
keseru.gyorgy@ttk.hu

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 03 January 2022

Accepted: 01 February 2022

Published: 09 March 2022

Citation:

Egyed A, Kiss DJ and Keserű GM
(2022) The Impact of the Secondary
Binding Pocket on the Pharmacology
of Class A GPCRs.
Front. Pharmacol. 13:847788.
doi: 10.3389/fphar.2022.847788

G-protein coupled receptors (GPCRs) are considered important therapeutic targets due to their pathophysiological significance and pharmacological relevance. Class A receptors represent the largest group of GPCRs that gives the highest number of validated drug targets. Endogenous ligands bind to the orthosteric binding pocket (OBP) embedded in the intrahelical space of the receptor. During the last 10 years, however, it has been turned out that in many receptors there is secondary binding pocket (SBP) located in the extracellular vestibule that is much less conserved. In some cases, it serves as a stable allosteric site harbouring allosteric ligands that modulate the pharmacology of orthosteric binders. In other cases it is used by bitopic compounds occupying both the OBP and SBP. In these terms, SBP binding moieties might influence the pharmacology of the bitopic ligands. Together with others, our research group showed that SBP binders contribute significantly to the affinity, selectivity, functional activity, functional selectivity and binding kinetics of bitopic ligands. Based on these observations we developed a structure-based protocol for designing bitopic compounds with desired pharmacological profile.

Keywords: GPCR (G-protein coupled receptor), allosteric, bitopic, selectivity, functional selectivity

INTRODUCTION

G-protein coupled receptors (**Figure 1**) are among the most popular targets for drug discovery and the development of novel therapeutic and pharmacological tools. One third of the drugs currently approved by the Food and Drug Administration affects one of the GPCRs (Sriram and Insel, 2018). They are critical in signal transduction of hormones and neurotransmitters, and consequently are pharmacological targets for many diseases (Overington et al., 2006). Furthermore, studying these receptors may help to elucidate the signaling mechanisms in cells, as they play a crucial role in the regulation of both central and peripheral neurological and physiological processes. Detailed understanding of these processes facilitates the development of more targeted therapies (Christopoulos, 2014).

GPCRs have multiple ligand binding sites, the orthosteric binding pocket and a generally separated less conserved allosteric secondary binding pocket (Christopoulos, 2014). Basically, the endogenous ligand binds to the OBP. SBPs are found in both the extracellular and intracellular parts of the receptor (**Figure 2**), some of these binding sites are well separated from the OBP while others may have extended binding pocket-like features such as the 5-HT_{2A} aripiprazole structure (PDB: 7VOE) (Chen et al., 2021).

These secondary binding sites have become key to achieve the right subtype selectivity and functionality. Therefore, a lot of effort was given to the research of allosteric binding sites and allosteric modulators. A large number of allosteric modulators of GPCRs that bind to the

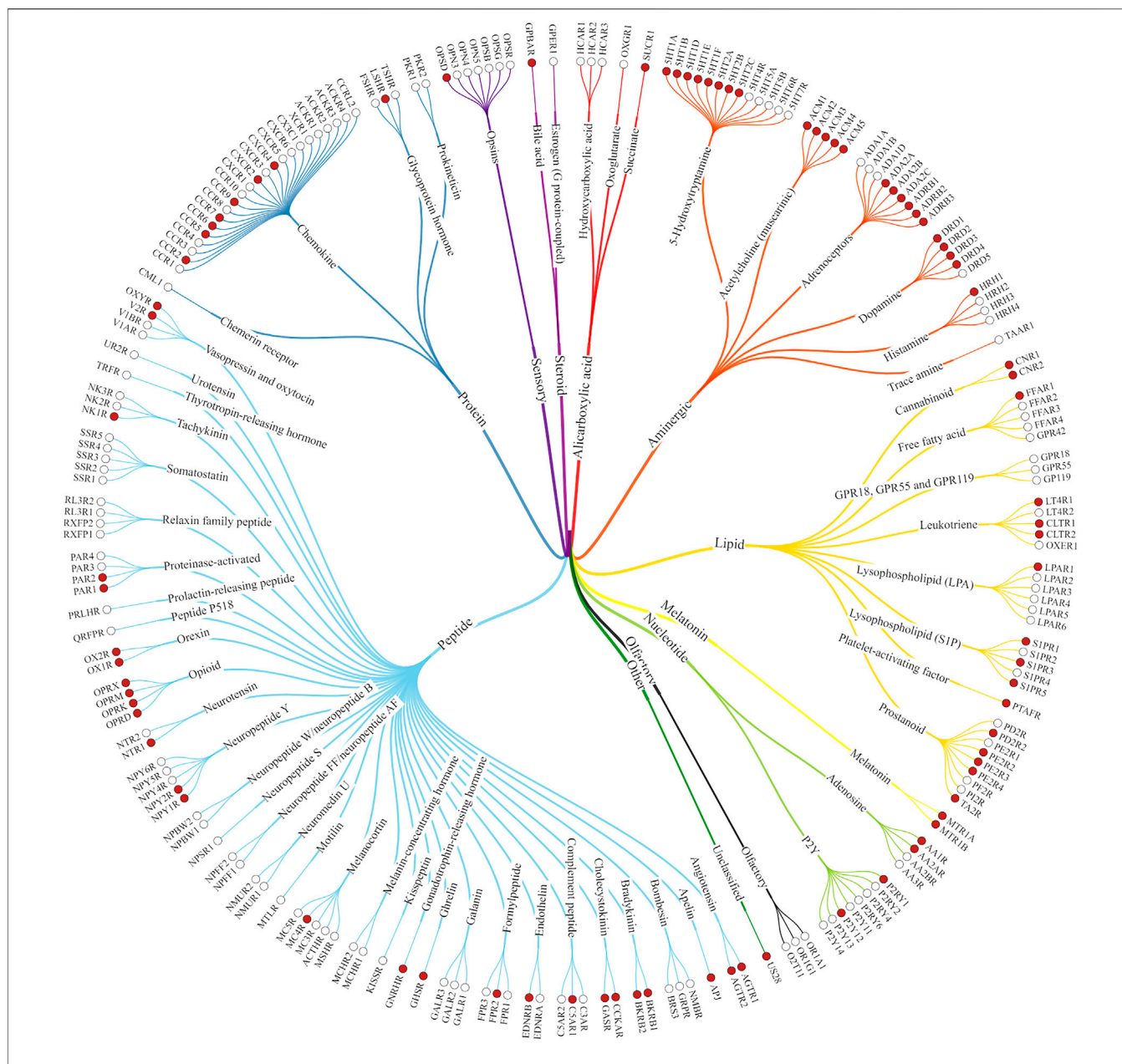


FIGURE 1 | Class A GPCRs. The structures of the receptors marked with red dots have already been solved experimentally (Kooistra et al., 2021).

extracellular or intracellular domains were identified. The combination of a primary pharmacophore (PP) binding to the OBP and a secondary pharmacophore (SP) binding to the SBP resulted in bitopic compounds (**Figure 2C**) that combine the pharmacological properties of both types of ligands defining a new unique pharmacological profile. One of the first published bitopic molecules of this type is methoctramine that acts as an antagonist at the muscarinic receptor M_2R (Melchiorre et al., 1987).

In this review we would like to give only a brief insight into class A GPCR structures and the world of allosteric modulators as several reviews have been published in the field. Mainly, we

discuss in detail the recent advances in bitopic ligands, while we close the review with an outlook towards the design approaches in the field.

LIGAND BINDING POCKET REVEALED BY EXPERIMENTAL STRUCTURES

Recent advances in X-ray crystallography and cryo electron microscopy provided many new structures of GPCRs complexed with allosteric ligands. As of early December 2021, 57 GPCR structures containing allosteric ligands have been found

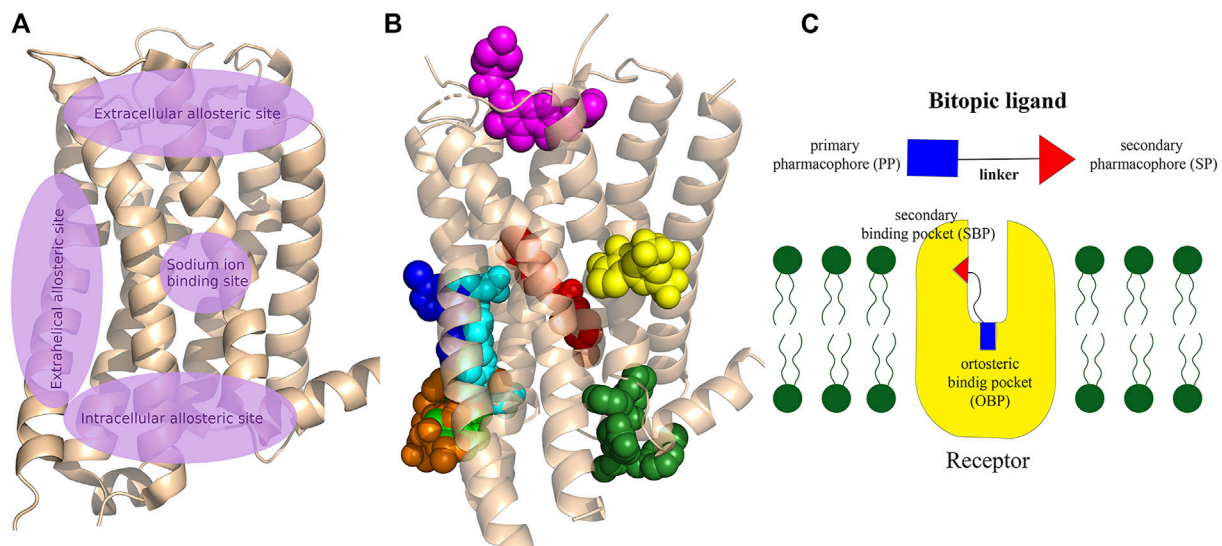


FIGURE 2 | (A) Schematic representation of the main allosteric sites in Class A GPCRs. The OBP, where the endogenous ligands bind to the receptor, is located between the extracellular allosteric site and the sodium binding site, deep in the crevice of the receptor formed by the transmembrane helices. Some allosteric sites are clearly separated from OBP, while others can be considered as an expansion of the orthosteric pocket. **(B)** Visualisation of allosteric binding sites for some important compounds related to the review: mevidalen in the D₁R (green, PDB code: 7LJD), AP8 in FFAR1 (cyan, PDB code: 5TZY), ORG27569 in CB₁ (red, PDB code: 6KQI), MIPS521 in A₁R (yellow, PDB code: 7LD3), LY2119620 in M₂R (magenta, PDB code: 4MQT), Cmpd-15PA in β_2 AR (dark green, PDB code: 5X7D), AS408 in β_2 AR (dark blue, PDB code: 6OBA), cmpd-6fa in β_2 AR (orange, PDB code: 8N48). Cholesterol was shown to bind to extrahelical binding sites to different TMs that could not be depicted on the figure to maintain clarity. For details please see the recent review of Jakubik and El-Fakahany (2021) and for a review of the allosteric sites at the receptor–lipid bilayer interface please see Wang et al. (2021) **(C)** Schematic structure of a bitopic compound. The primary pharmacophore that binds to the OBP is linked through a linker to the secondary pharmacophore binding to the SBP.

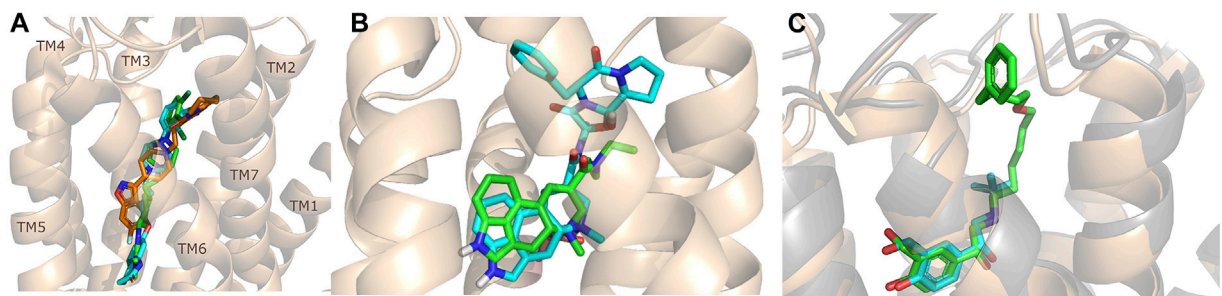


FIGURE 3 | Structures of some important bitopic compounds. (A) The unusual "upside-down" binding mode of cariprazine (green) and aripiprazole (cyan) in the inactive 5-HT_{2A} structure. Risperidone (orange) is shown as a reference to highlight the cryptic pocket opened up by aripiprazole and cariprazine. **(B)** The aligned LSD (Wacker et al., 2017) and ergotamine (Wacker et al., 2013) 5-HT_{2B} structure highlighting that the introduction of an SP can influence the binding mode of the PP. The figure was reproduced from **Supplementary Figure S7** of our paper (Egyed, A et al. Controlling Receptor Function from the Extracellular Vestibule of G-Protein Coupled Receptors. Chem. Commun. 2020, 56 (91), 14167–14170) (Egyed et al., 2020). **(C)** The binding mode of salbutamol (cyan) and salmeterol (green) (Masureel et al., 2018) in the β_2 R highlighting the important role of ECL2 as discussed in more detail in the binding kinetic section of this review.

in GPCRdb (Kooistra et al., 2021), these structures cover 20 receptor types and three different states; active, inactive and intermediate. Among allosteric ligands, examples of positive (PAM) and negative allosteric modulators (NAM) can be found. The collection of the published GPCR structures with allosteric ligands is available in the supporting information (Kooistra et al., 2021) (**Supplementary Table S1**). In addition, a significant number of active structures have become accessible, which may provide more information on the mechanism of receptor activation and offer considerable support for drug

design, although few of these are allosteric ligands. Among them, 35 active aminergic GPCR structures have been published in the last 2 years (**Supplementary Table S2**) (Kooistra et al., 2021). These include 7 serotonin (Kim et al., 2020; Peiyu Xu et al., 2021a; Huang et al., 2021) (5-HT₁), 15 dopamine (Zhuang et al., 2021a; Xiao et al., 2021; Zhuang et al., 2021b; Yin et al., 2020; Peiyu Xu et al., 2021b) (DR), 1 histamine (Xia et al., 2021) (HR), 1 muscarinic (Staus et al., 2020) (MR) and 11 adrenergic (Lee et al., 2020; Fan Yang et al., 2021; Yuan et al., 2020; Su et al., 2020; Xinyu Xu et al., 2021; Zhang et al., 2020;

Nagiri et al., 2021) (AR) receptor structures. Out of these complexes, 20 structures contain allosteric modulators but not obviously in the SBP, while 10 were co-crystallized with bitopic ligands bound both the OBP and the SBP. The discussion of the structures in detail is out of scope of this review, however we highlight here the new cariprazine and aripiprazole bound 5-HT_{2A} structures (**Figure 3A**). (Chen et al., 2021) Interestingly, both compounds display an unexpected binding mode with their secondary binding motif exploring a binding pocket deep in the receptor instead of engaging with the extracellular secondary binding pocket. In the dopamine D₂ and D₃ receptors (D₂R, D₃R) the docking positions of aripiprazole so far have shown that 4-(2,3-dichlorophenyl)piperazine PP is located roughly parallel to the membrane plane and close to S5.42 and F6.51. The dihydroquinoline secondary pharmacophore is located at the junction of transmembrane helices (TM) 1, TM2, TM7 or TM3, TM5 and extracellular loop (ECL) 2. However, in the 5-HT_{2A} crystal structures of aripiprazole and cariprazine the ligands are located in an “upside-down” binding mode. The 2,3-dichlorophenyl PP occupies the orthosteric site and faces the extracellular region, but the dihydroquinoline SP vertically penetrates the hydrophobic pocket formed between TM5 and TM6 and interacts with residues L247^{5.51}, V333^{6.45} and C337^{6.49} and forms π - π interactions with residues F332^{6.44} and W338^{6.48}. Upon binding of aripiprazole, a conformational rearrangement occurs resulting in an increase in the size of the binding pocket. Induced docking with D₂R was used to reproduce the “upside-down” binding pose of aripiprazole and cariprazine. Compared with the rigid docking, a much lower binding energy was calculated in the induced-fit docking, indicating that the upside-down binding mode represents a more stable conformation of D₂R (Chen et al., 2021).

ALLOSTERIC MODULATORS IN THE CLASS A GPCR FIELD

Allosteric binding sites (**Figures 2A,B**) have attracted increasing interest in order to develop more selective agents with fewer side effects (Congreve et al., 2017a; Chan et al., 2019). Allosteric sites are typically less conserved than orthosteric pockets and therefore they could provide greater selectivity and better control over the dynamical equilibrium of the receptor. Following the classic structural architecture of a class A GPCR, the orthosteric binding pocket is formed by the transmembrane helices while the extracellular loops and the N-terminus of the peptide chain define the secondary binding domain. It should be mentioned, however, that there are other allosteric sites (e.g., extrahelical sites at the protein-membrane interface, intracellular sites at the signalling domain or intrahelical sodium site) available. Allosteric ligands can modify the biological response, they can stabilise the active or inactive conformation that is potentially linked to biased signalling or partial agonism (Wakefield et al., 2019). Based on spectroscopic and structural studies, conformational changes in the receptor govern the activation of signalling pathways. Characterization of interactions with intracellular partners guiding the allosteric process is a major

challenge and can only be fully understood by using a combination of different methodologies (Liu et al., 2012; Masureel et al., 2018; Frei et al., 2020). Most allosteric modulators have been discovered serendipitously by high throughput screening (HTS) campaigns (Bian et al., 2020). Due to the vastness of the topic and the number of reviews published in the last years, we will only provide a brief insight into the world of allosteric modulators.

The tissue distribution and relative expression of the four adenosine receptor (AR) subtypes A₁R, A_{2A}R, A_{2B}R and A₃R regulate the physiological effects of endogenous adenosine. Adenosine receptors are expressed in most tissues and major organs, including brain, heart, kidney, skin, adipose tissue, immune cells, lung and liver. The four adenosine receptor subtypes can be broadly classified into two classes. Baressi et al. described a type of A_{2B}R allosteric modulators with good selectivity over the other subtypes, these compounds contain a 1,3-substituted indole unit (Barresi et al., 2021a; Barresi et al., 2021b). Lu et al. established a fragment screening method using mass spectrometry to screen GPCR ligands, identifying an A_{2A}R NAM. Fg754 (**Figure 4**) contains a specific acetidine moiety that forms bonds in the sodium ion pocket. Based on molecular dynamics (MD) simulations, it may overlap with the orthosteric binding site, probably acting in a mixed mode. The compound could thus be a new starting point for the development of allosteric modulators or bitopic compounds (Yan Lu et al., 2021). The A₁R and A₃R preferentially bind to G_{i/o} proteins to inhibit adenylate cyclase activity, while the A_{2A}R and A_{2B}R preferentially bind to G_s proteins to stimulate adenylate cyclase activity. Like other GPCRs, adenosine receptors can interact with different G-protein subtypes. In addition, A_{2B}R has been suggested to couple to both G_{i/o} and G_q proteins (Linden et al., 1999; Gao et al., 2018), while A₁R has been shown to couple to both G_s and G_q proteins (Cordeaux et al., 2004). In addition to G_q signalling, G_q dimers released following G protein activation can interact with effector proteins to modulate intracellular signalling. Beside G-protein-dependent signalling, adenosine receptors can also signal through G-protein-independent effectors. One of the best described G-protein-independent pathway is initiated following recruitment of arrestin adaptor proteins (β -arrestin1 and β -arrestin2). This process is typically preceded by G-protein coupled receptor kinase mediated phosphorylation, but recent studies have shown the possibility of phosphorylation independent β -arrestin recruitment for several GPCRs, including A₃R. Arrestin recruitment has been investigated primarily in A_{2B}R and A₃R and there is limited evidence that A₁R or A_{2A}R can recruit β -arrestin. McNeill et al. have discussed in detail the effects of allosteric modulators belonging to different subtypes on distorted signalling, which will not be discussed in detail below (McNeill et al., 2021).

Free fatty acids may act as signalling molecules at FFA receptors (FFARs). Free fatty acids of different chain lengths and saturation states activate FFARs as endogenous agonists by binding at the orthosteric receptor site. Following FFAR deorphanisation, a number of ligands targeting allosteric sites on FFARs have been identified with the aim of developing drugs for metabolic, (auto)inflammatory, infectious, endocrine,

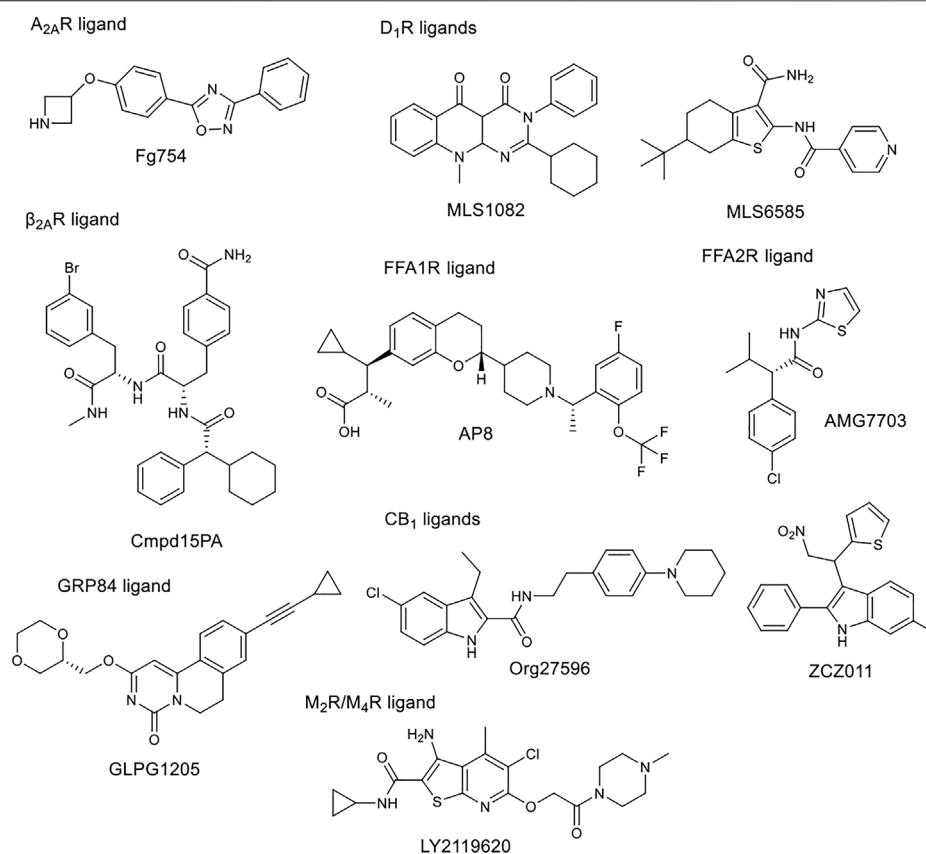


FIGURE 4 | Chemical structure of selected allosteric modulators.

cardiovascular and renal diseases. In 2021, Grundmann et al. published a detailed review (Grundmann et al., 2021) on free fatty acid receptors, describing in detail the subtypes (FFAR1, FFAR2, FFAR3, FFAR4, GPR84), their function, structures and outlined the importance and challenges of allosteric modulators. FFAR1 is the most studied subtype. Although the biology of the receptors is still largely elusive, a large body of research evidence has accumulated around ligand-receptor interactions and their associated signalling capabilities. At least three distinct groups of FFAR1-activating ligands can be distinguished: 1) endogenous/orthosteric agonists (long-chain fatty acids), partial allosteric agonists (fasiglifam, MK-8666, AM 837), and full allosteric agonists (AM 1638, AP8) (Figure 2B, Figure 4). These groups differ not only in their apparent binding sites (Figure 2B) on the receptor, but also in their ability to induce different downstream signalling pathways of FFAR1, ultimately leading to different results in the phenotype of the FFA1 receptor *in vivo*. New results on allosteric FFAR2 ligands (AMG 7703, AZ1729, Compound 58) (Figure 4), show promising pharmacological properties and have generated new interest in this target, considering new allosteric modalities. GLPG1205 (Figure 4), an antagonist and negative allosteric modulator of GPR84, showed promising preclinical results in models of idiopathic pulmonary fibrosis, but was later discontinued from development. Allosteric targeting of small-, medium-, and long-chain fatty acid receptors is a

promising approach to address a variety of therapeutic areas, demonstrating the biological diversity and drug target attractiveness of members of this receptor family (Grundmann et al., 2021).

The cannabinoid receptor type 1 (CB₁) was first discovered as the main target for Δ^9 -tetrahydrocannabinol (THC), the psychoactive compound in Cannabis. CB₁ was first identified in rat and later cloned from a human brain cDNA library. Widely known CB₁ agonists are synthetic cannabinoids and THC analogues, such as HU-210 (Howlett et al., 1990), CP55940 (Kapur et al., 2009), and WIN55212 (Felder et al., 1995). The CB₁ receptor preferentially binds a G_i protein and its activation leads to a decrease in cyclic adenosine monophosphate (cAMP) levels in cells. Other signalling pathways have also been investigated, focusing primarily on ERK1/2 phosphorylation. ERK signalling is hypothesised to play a role in cocaine addiction, and together with cAMP, to be an important regulator of synaptic plasticity, memory and learning. Inhibition of CB₁ proved effective in the treatment of obesity with antagonists or inverse agonists, but they were later withdrawn from the market due to adverse psychiatric side effects (anxiety, suicidal ideation). Several new strategies to avoid potential side effects have been analysed, one of them being the development of allosteric modulators. Leo and Abood reviewed the physiological and pathophysiological roles of CB₁,

described the signalling mechanisms, and investigated CB₁ biased signaling (Leo and Abood, 2021). Based on agonist-bound solvated molecular structures and biased allosteric modulators they look at possible molecular mechanisms of CB₁ signalling. Mielnik et al. present the *in vitro* and *in vivo* profiles of several NAMs (Org27569, PSNCBAM-1, ABM300, Pepcan-12, Pregnenolone, and cannabidiol) and PAMs (ZCZ011, GAT211, Lipoxin A4, LDK1258) in detail (**Figure 2B**, **Figure 4**). They concluded that CB₁ PAMs in anxiety and depression while CB₁ NAMs—in combination with cannabidiol—in psychosis could be promising (Mielnik et al., 2021).

Che and Roth have provided a detailed summary of the pharmacology, ligands (orthosteric, allosteric), and structures of opioid receptors (OR) (Che and Roth, 2021). Activating μ -opioid receptor (MOR) causes serious side effects, which are the root of the current opioid crisis. In their review, potential strategies and targets for developing opioid alternatives were discussed. Separately, they list OR biased agonists, allosteric modulators, multitarget ligands and peripherally restricted ligands. The complexity of signalling pathways should be considered in the therapeutic potential of biased agonists, and allosteric modulators are alternative means to modulate more precisely the action of endogenous or exogenous ligands. As opioid receptors are widely expressed in the peripheral system, the use of ligands restricted to this system would avoid central nervous system induced side effects. Simultaneous targeting of multiple opioid and non-opioid receptors may result in safer analgesics (Che and Roth, 2021).

The family of aminergic GPCRs includes adrenergic, dopamine, serotonin, histamine, muscarinic and trace amine receptors. These receptors have several similarities, they bind monoamine neurotransmitters, acetylcholine, or trace amines. They share common features in sequence, structure and function. Ergotamine (**Figure 3B**) can bind to 22 aminergic receptors with K_i values less than 1 μ M (Peng et al., 2018). Other examples can be found in the literature, such as chlorpromazine, clozapine, thioridazine, olanzapine which have good affinity for several aminergic GPCRs (Roth et al., 2004). On the other hand, it would be important to produce drugs that have subtype and functional selectivity to avoid side effects.

In the field of adrenergic receptors, Wu and co-workers have discussed in detail the binding of endogenous ligands to different receptors, the mechanism of β -adrenergic and α_2 receptor attenuation, distorted signal transduction, subtype selectivity, and selectivity between the main types. Insights into the allosteric modulation of β_{2A} R were provided. They also reported on the results obtained with different modalities. The cholesterol binding site was recently described in detail by Sarkar and Chattopadhyay (2020). The arrangement of the 7 TMs in each class of GPCRs results in a groove at the lipid interface formed by TM3/4/5, and in β_{2A} R, to this site the binding of PAMs and NAMs were identified. GPCRs use the cytoplasmic surface to interact with intracellular partners with small molecules binding at this site discovered primarily in chemokine receptors. Only Cmpd15PA (**Figure 2B**, **Figure 4**) in β_{2A} R targets this site outside the chemokine subfamily. These small molecules are all NAMs. Cmpd15PA has little interaction with the G protein, but stabilizes

the receptor inactive state through extensive interactions with TM1, TM2, TM6, TM7, H8 and intracellular loop 1 (Wu et al., 2021).

The five dopamine receptor subtypes (D1–5) are activated by the endogenous catecholamine dopamine. The D1-like family comprises dopamine D₁ and D₅ receptors that mainly couple to the G_s G-protein and thereby stimulate cAMP production. The D₂-like family includes D₂, D₃, and D₄ receptors, that couple to G_{i/o} G-proteins and attenuate cAMP production (British Pharmacological Society, 2021). Fasciani et al. have presented allosteric modulators of the DR, the bitopic compound SB269652 has been analysed in detail. Mao et al. describe the role of different dopamine receptor allosteric modulators in the treatment of Parkinson's disease. DR allosteric modulators represent an alternative and promising strategy for drug discovery of GPCRs with high selectivity and low side effects (Mao et al., 2020). Like many other receptors, the classical approach to D₁R is the development of orthosteric ligands, but this has several drawbacks from a therapeutic point of view. D₁R agonists have narrow therapeutic window, can induce seizures and hypotensive side effect. PAMs are a more useful approach because they potentiate the effect of endogenous dopamine, the available dopamine level provides a natural ceiling effect for PAM activity, and endogenous spatial and temporal regulation of dopamine-mediated stimulation is maintained. To date, seven D₁R PAM structural classes have been discovered. Two of these (MLS1082 and MLS6585) were discovered in 2018 by Luderman and colleagues using HTS (Luderman et al., 2018) (**Figure 4**). Subsequently, MLS1082 was investigated in a SAR study and they identified several analogues that enhanced dopamine-induced D₁R activation (Luderman et al., 2021).

There are five subtypes of the muscarinic acetylcholine receptor. The different subtypes show high degree of homology in the transmembrane domains. In recent years, the structures of all five have been resolved by X-ray crystallography (Vuckovic et al., 2019; Thal et al., 2016; Kruse et al., 2013; Kruse et al., 2012; Haga et al., 2012). In a review, Jakunik and El-Fakahany provide a detailed analysis of allosteric adhesion, the molecular mechanisms of action, and present specific modulators. The diversity of the effects of allosteric modulators and the studies on them will greatly influence the development of new therapies. Selective PAMs (LY2119620) (**Figure 2B**, **Figure 4**), which have therapeutic potential in the treatment of Alzheimer's disease or schizophrenia, show encouraging results (Conn et al., 2009; Bock et al., 2018; Jakubik and El-Fakahany, 2020).

Biochemically, 5-hydroxytryptamine (5-HT) is derived from the amino acid tryptophan, undergoing hydroxylation and decarboxylation processes that are catalyzed by tryptophan hydroxylase and aromatic L-amino acid decarboxylase, respectively. As a biogenic amine, 5-HT plays important roles in cardiovascular function, bowel motility, platelet aggregation, hormone release and psychiatric disorders. 5-HT achieves its physiological functions by targeting various 5-HT receptors (5-HTRs), which are composed of six classes (5-HT₁, 5-HT₂, 5-HT₄, 5-HT₅, 5-HT₆, and 5-HT₇ receptors, a total of 13 subtypes) and a class of cation-selective ligand-gated ion channels, the 5-HT₃

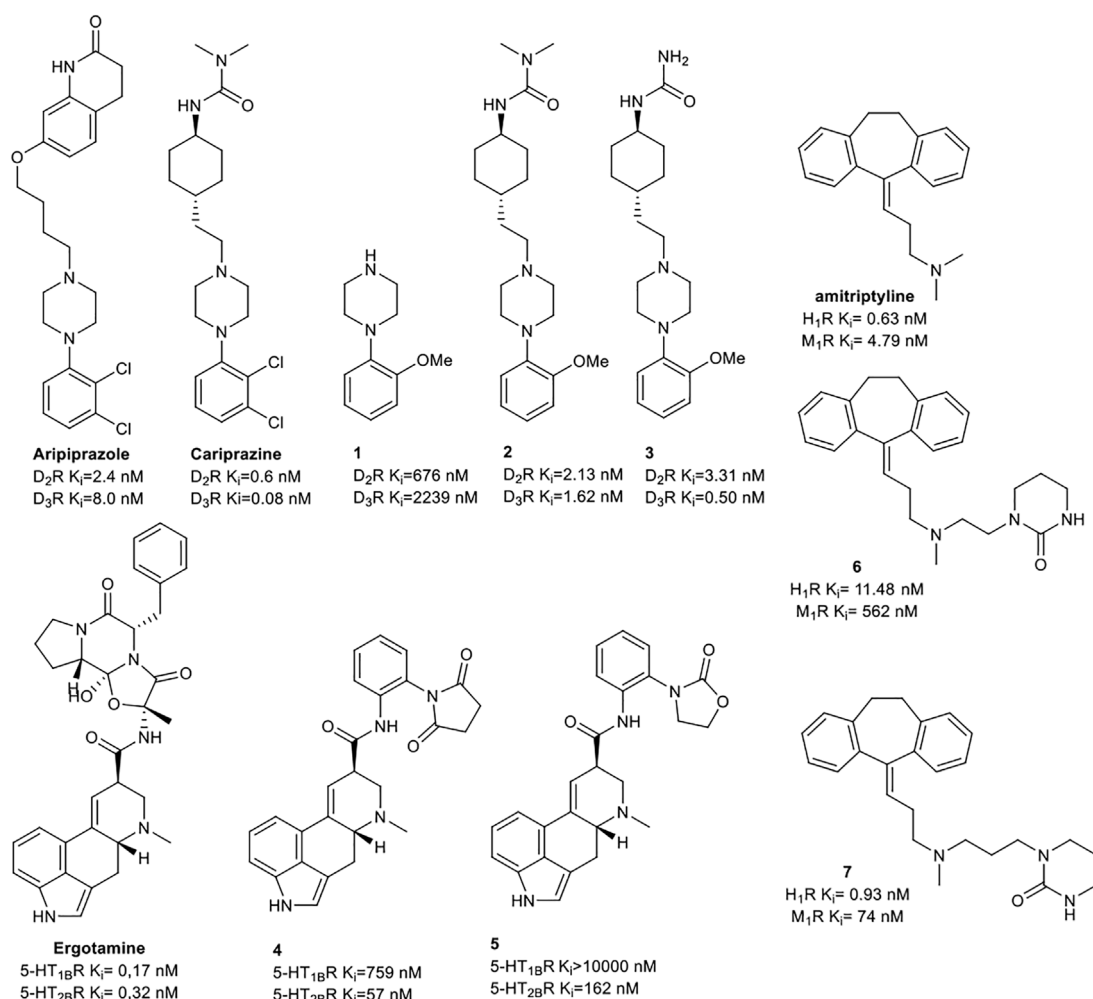


FIGURE 5 | Designed bitopic ligands and the reference compounds in the study of Keserü et al. (Egyed et al., 2021).

receptor. Barnes et al. have published a review (Barnes et al., 2021) detailing each subtype, describing their functions and pharmacology one by one and discuss known allosteric ligands. They find that 5-HT receptors are less involved in allosteric modulation than other GPCRs (e.g., muscarinic, GABA), with the possible exception of 5-HT₃R. However, from some structures with ergoline, it becomes clear that, in addition to the classical OBP, some 5-HT receptors have an extended binding site very similar to that described for muscarinic allosteric ligands. Such molecular targets may offer attractive strategies for new therapies (Barnes et al., 2021).

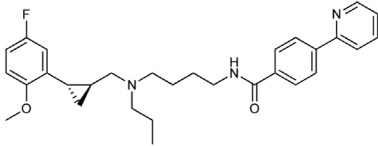
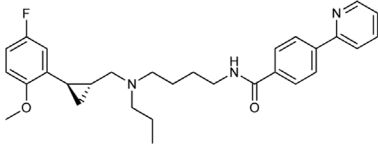
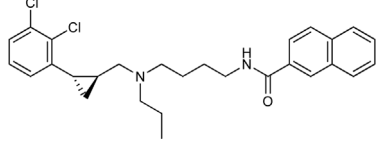
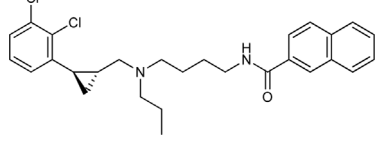
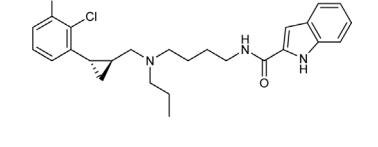
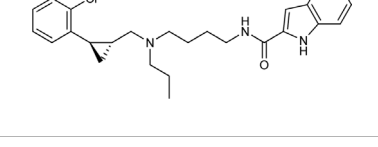
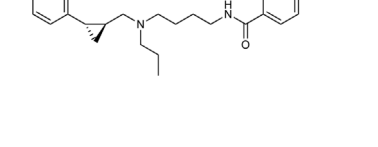
BITOPIC LIGANDS TO STUDY SELECTIVITY AND FUNCTIONAL SELECTIVITY OF CLASS A GPCRS

As outlined in the introduction, our primary focus is on bitopic compounds in this review. These compounds combine the efficiency of orthosteric ligands and the diversity of allosteric

SPs by interacting with both binding sites simultaneously. This gives bitopic ligands an advantage over allosteric modulators, as the latter need an orthosteric ligand to exert their effect. This may be important in cases where endogenous substrate depletion contributes to the pathogenesis of disease, such as in Parkinson's and Alzheimer's diseases, but there are further examples in metabolic disorders. The key structural moieties of bitopic compounds (PP, SP and linker, depicted on **Figure 3C**) have different roles. PP is classically considered to be responsible for functionality while SP can modulate binding affinity, selectivity as well as functional character and efficacy. The linker connects the two pharmacophores and may be responsible for the optimal binding poses by positioning the pharmacophores and affecting the pharmacology profile (Bethany et al., 2019).

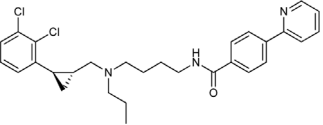
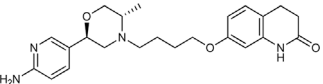
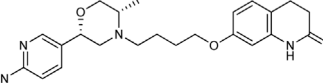
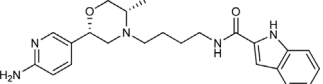
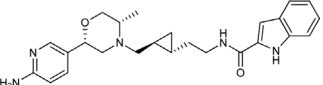
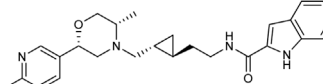
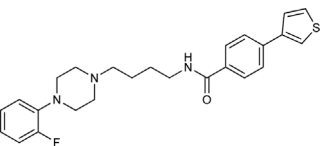
In the design of bitopic compounds, the desired orthosteric binding motif should have high affinity for the selected receptor and ideally, the SP should provide high subtype selectivity while maintaining or even increasing affinity. In the case of a linker, the choice of attachment points and length

TABLE 1 | Selected compounds from DR related selectivity studies (Battiti et al., 2019; Tan et al., 2020; Lee et al., 2021).

| Cmpd | Structure | K _i (nM) | | | | | |
|---------------------------------------|---|---------------------|------------------|------------------|------------------|------------------|--------------------|
| | | D ₁ R | D ₂ R | D ₃ R | D ₄ R | D ₅ R | 5-HT _{2C} |
| (1 <i>S</i> ,2 <i>S</i>)- 17a |  | 1,071 | 1,230 | 3.8 | 851 | >5,000 | 50.1 |
| (1 <i>R</i> ,2 <i>R</i>)- 17b |  | 4,898 | 1,349 | 4.1 | 575 | >5,000 | 1,122 |
| (1 <i>S</i> ,2 <i>S</i>)- 18a |  | 1,047 | 1,148 | 20.8 | 776 | >5,000 | 138 |
| (1 <i>R</i> ,2 <i>R</i>)- 18b |  | 1,288 | 676 | 4.4 | 813 | >5,000 | 513 |
| (1 <i>S</i> ,2 <i>S</i>)- 19a |  | 1,122 | 992 | 12.8 | 676 | >5,000 | 61.7 |
| (1 <i>R</i> ,2 <i>R</i>)- 19b |  | 1,380 | 537 | 2.2 | 1,047 | >5,000 | 513 |
| (1 <i>S</i> ,2 <i>S</i>)- 20a |  | 2344 | 1,023 | 5.3 | 912 | >5,000 | 44.7 |

(Continued on following page)

TABLE 1 | (Continued) Selected compounds from DR related selectivity studies (Battiti et al., 2019; Tan et al., 2020; Lee et al., 2021).

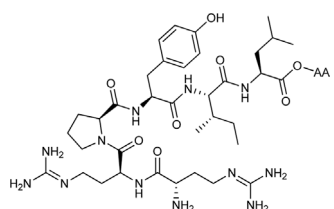
| Cmpd | Structure | K _i (nM) | | | | | |
|---------------------------------------|---|--------------------------------------|--------------------------------------|--------------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | | D ₁ R | D ₂ R | D ₃ R | D ₄ R | D ₅ R | 5-HT _{2C} |
| (1 <i>R</i> ,2 <i>R</i>)- 20b |  | 1,349 | 550 | 1.5 | 676 | >5,000 | 417 |
| Cmpd | Structure | D ₂ R K _i (nM) | D ₃ R K _i (nM) | D ₄ R K _i (nM) | D ₂ R/D ₃ R | D ₄ R/D ₃ R | |
| 24 |  | 2600 | 24200 | ND | 0.110 | ND | |
| 25 |  | 34.6 | 31.2 | ND | 1.1 | ND | |
| 27 |  | 134 | 5.96 | 357 | 22.5 | 59.9 | |
| 28a |  | 87.8 | 1.85 | 286 | 47.5 | 155 | |
| 28b |  | 831 | 282 | 2930 | 2.95 | 10.4 | |
| 39 |  | 648 | 1.4 | - | 467 | - | |

must be appropriate, and the linker must be moderately flexible to allow the pharmacophores to bind properly. For agonists, it is important that the linker does not interfere with conformational changes induced by receptor activation (Valant et al., 2012; Lane et al., 2013; Fronik et al., 2017; Bethany et al., 2019). Reinecke et al. published a review on bitopic compounds in 2019, summarizing the new bitopic compounds that have been published in the last 5 years (Bethany et al., 2019). Here we therefore focus on

compounds published in 2020–21, with a contextual analysis of previously published compounds where appropriate. In the following subsections, we discuss subtype selectivity and functional selectivity results separately.

Receptor and Subtype Selectivity

Receptor and subtype selectivity is an important criterion for minimizing side effects, therefore tremendous efforts go into the development of compounds with designed binding profile.

TABLE 2 | NTS1 and NTS2 receptor-binding data for bitopic ligands (Kling et al., 2019).


| Cmpd | NT (8-13)-AA | K_i (nM) | | NTS2/NTS1 | IP acc. Assay | |
|-----------------|-----------------------|-------------------------------|-------------------------------|-----------|-------------------------------|----------------------|
| | | NTS ₁ nM \pm SEM | NTS ₂ nM \pm SEM | | EC ₅₀ nM \pm SEM | Efficacy % \pm SEM |
| NT(8-13) | | 0.24 \pm 0.048 | 1.2 \pm 0.25 ^[h] | 5.0 | 0.74 \pm 0.20 | 100% |
| 51 | NT (8-13)-Gly-OH | 6.8 \pm 4.5 | 53 \pm 21 | 7.8 | 18 \pm 4 | 98 \pm 2% |
| 52 | NT (8-13)-Ser-OH | 3.3 \pm 1.7 | 58 \pm 28 | 18 | 37 \pm 16 | 98 \pm 5% |
| 53 | NT (8-13)-Phe-OH | 0.91 \pm 0.49 | 12 \pm 4.0 | 13 | 150 \pm 22 | 100 \pm 5% |
| 54 | NT (8-13)-Tyr-OH | 1.3 \pm 0.38 | 34 \pm 9.4 | 26 | 110 \pm 26 | 95 \pm 10% |
| 55 | NT (8-13)-hTyr-OH | 1.5 \pm 0.65 | 37 \pm 9.1 | 25 | 24 \pm 5 | 92 \pm 8% |
| 56 | NT (8-13)-meta-Tyr-OH | 2.1 \pm 0.4 | 44 \pm 23 | 21 | 34 \pm 7 | 94 \pm 4% |

Keserü et al. have developed a fragment based docking protocol to design specific receptor ligands. Based on the docking results, they have synthesized several compounds and demonstrated the usefulness of the method for the designing D₂/D₃, 5-HT_{1B}/5-HT_{2B} and H₁/M₁ receptor ligands with improved selectivity (Figure 5). In the first two cases, the selectivity of the PP was reversed using the SP moiety, while in the third case, a selective compound was designed and synthesized for a receptor pair with very similar PP (Egyed et al., 2021).

The importance of bitopic compounds in the inhibition of dopamine receptors is demonstrated by second and third generation antipsychotics, including aripiprazole (Burris et al., 2002) and cariprazine (Ágai-Csongor et al., 2012). 2,3-dichlorophenyl-piperazine, that serves as PP in these compounds was changed to 2-methoxyphenylpiperazine (**1**) PP. Although this PP exhibits weak D₂R selectivity, combined with a suitable SP group (**2,3**) its profile has been changed to mild D₃R selectivity. The efficacy of this methodology was further tested on serotonin receptors. The LSD-like PP of ergotamine (Figure 3B) did not show subtype selectivity at the two selected serotonin receptors, but the designed compounds (**4,5**) with the modified SP already had significantly higher affinity at 5-HT_{2B}R over 5-HT_{1B}R. Although ergotamine was more potent, compounds **4** and **5** had much greater selectivity over it. Among the first-generation antihistamines, muscarinic acetylcholine M₁ activity was a major problem due to side effects. Therefore, huge efforts were dedicated to the development of compounds with significant H₁R receptor selectivity. Starting from amitriptyline having only 7-fold selectivity, bitopic compounds (**6,7**) were designed that demonstrated 50–80 fold selectivity over M₁R (Egyed et al., 2021). The proposed protocol detailed in the design section of this review may be applied to other targets to achieve designed selectivity with bitopic compounds.

Tan et al. have exploited the basic 2-phenylcyclopropylmethylamine (PCPMA) scaffold (**8, 9**), whose analogues are known 5-HT_{2C}R agonists (Cheng et al.,

2015; Cheng et al., 2016a; Cheng et al., 2016b; Zhang et al., 2017), to design new bitopic compounds (Tan et al., 2020) (Supplementary Table S3). Here we discuss only a subset of these compounds. As secondary pharmacophore, 1,2,4-triazolylthiol ethers were used and a propyl chain was employed as a linker. The introduction of SP alone improved D₃R activity 3-fold. A major leap forward was the realization that the alkyl side chain introduced on the amino group of PCPMA significantly improves subtype selectivity and D₃R affinity. Next, they investigated the substituents of the aromatic ring of PCPMA. First, the ortho positioned 2-fluoroethoxy group was changed, whereby methoxy was found to be the optimal one, thus significantly improving the D₃R affinity. The replacement of the fluorine atom by chlorine resulted in a moderate selectivity towards D₂R, D₄R, 5-HT_{2C}R and a strong selectivity towards D₁R and D₅R (**10–12**). As these results could only approximate the values of the reference compound **13** (BP-897) (Supplementary Table S3) the strategy was changed and a butylene linker was used instead of the propylene group, the SP was replaced by other aromatic rings (naphthyl, indolyl, and 4-pyridylphenyl) and an amide bond between the linker and the SP was introduced instead of thioether (**14–20**). For these compounds, only N-alkyl substituted variants have been prepared and the effects of several PPs have been investigated. When examining the racemic compounds, the compound containing 4-pyridylphenyl SP and dichlorophenyl PP (**20**) has more than 1000-fold selectivity towards the other DRs, with milder but still significant selectivity in the range of **17, 18**, and **19** (Tan et al., 2020) (Table 1).

Battiti and co-workers performed a SAR analysis combining two PPs for the synthesis of bitopic compounds; one is a selective dopamine agonist PF-592379 (Allerton et al., 2005; Ackley, 2008) and the other is PD-128907, which is a D₂R/D₃R agonist. (Supplementary Table S4) They concluded that the structural features of PD-128907 avoided the construction of bitopic compound. Therefore, they focused to PF-592379 to synthesize D₂R/D₃R active bitopic compounds. Here we discuss a

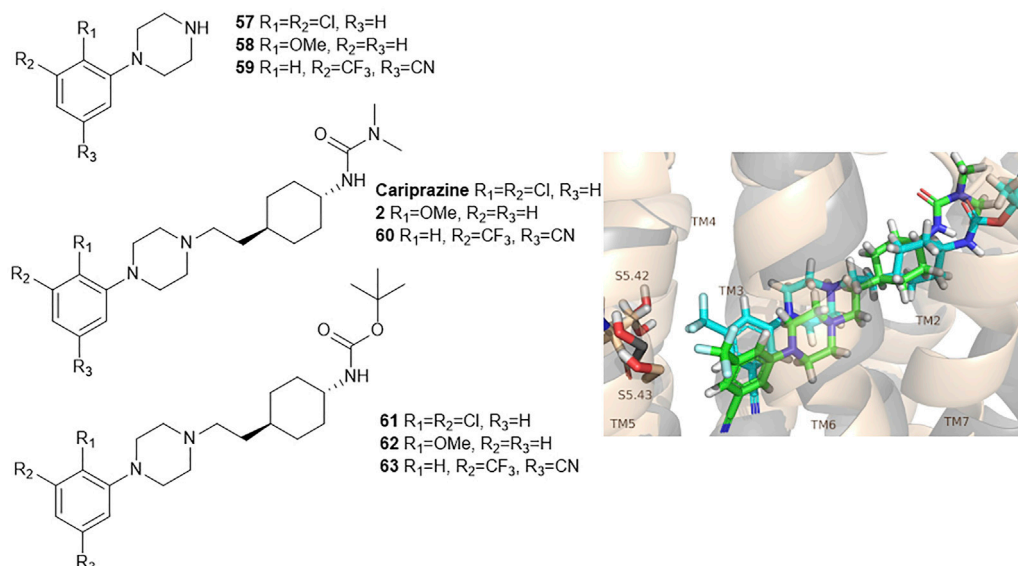


FIGURE 6 | D₂R and D₃R ligands with designed functional profile (Egyed et al., 2020). The binding mode of compound **60** and **63** was extracted from the MD simulations. The simulations revealed that the SP motif influence the position of the PP and that might be linked to the observed different functional profile. The figure representing the binding mode is reproduced from the TOC Figure of our original article Egyed, A et al. Controlling Receptor Function from the Extracellular Vestibule of G-Protein Coupled Receptors. Chem. Commun. 2020, 56 (91), 14167–14170.

TABLE 3 | Functional activities (pIC₅₀ or pEC₅₀ and maximal efficacy (E_{max}) values with s.d. values in parentheses) measured for the G-protein mediated and β-arrestin mediated pathway of the hD₂ and hD₃ receptor (Egyed et al., 2020).

| hD ₂ R | G-protein mediated pathway | | | β-Arrestin mediated pathway | | |
|-------------------|---|--|--|---|---|--|
| | H | SP 1 | SP 2 | H | SP 1 | SP 2 |
| PP 1 | 57 pEC ₅₀ < 4.3 uM E _{max} = 45.6% (3) partial agonist | Cariprazine pEC ₅₀ = 8.85 (0.1) Gao et al. (2014) E _{max} = 77.4% (7) partial agonist | 61 pEC ₅₀ = 8.64 (0.22) E _{max} = 99.4% (2) full agonist | pEC ₅₀ = 3.85 (0.12) E _{max} = 7% (1) partial agonist | pEC ₅₀ = 9.69 Gao et al. (2014) E _{max} = 13.9% partial agonist | pEC ₅₀ = 8.40 (0.17) E _{max} = 26% (2) partial agonist |
| PP 2 | 58 pIC ₅₀ = 6.4 (1.0) Newman et al. (2012) E _{max} = 14% (1) partial agonist | 2 pEC ₅₀ = 8.62 (0.07) E _{max} = 82.7% (3) partial agonist | 62 pIC ₅₀ = 8.42 (0.18) E _{max} = 78.7% (4) partial agonist | pIC ₅₀ = 5.03 (0.12) antagonist | pIC ₅₀ = 8.08 (0.05) antagonist | pIC ₅₀ = 7.63 (0.10) antagonist |
| PP 3 | 59 pIC ₅₀ = 4.72 (0.78) antagonist | 60 pIC ₅₀ = 6.10 (0.13) antagonist | 63 EC ₅₀ > 50 uM E _{max} = 25.4% (4) partial agonist | pIC ₅₀ = 5.89 (0.13) antagonist | pIC ₅₀ = 7.71 (0.10) antagonist | pIC ₅₀ = 7.23 (0.12) antagonist |
| hD ₃ R | G-protein mediated pathway | | | β-arrestin mediated pathway | | |
| | H | SP 1 | SP 2 | H | SP 1 | SP 2 |
| PP 1 | pEC ₅₀ = 7.50 (0.34) E _{max} = 72% (12) partial agonist | pEC ₅₀ = 8.58 Kiss et al. (2010) E _{max} = 27% Kiss et al. (2010) partial agonist | pEC ₅₀ = 8.09 (0.13) E _{max} = 94% (7) full agonist | 30% (5) in 80 μM partial agonist | pEC ₅₀ = 8.32 Frank et al. (2018) E _{max} = 32% Frank et al. (2018) partial agonist | pEC ₅₀ = 8.42 (0.21) E _{max} = 61% (6) partial agonist |
| PP 2 | pEC ₅₀ = 6.12 (0.17) E _{max} = 11% (4) partial agonist | pEC ₅₀ = 8.43 (0.51) E _{max} = 11% (3) partial agonist | pEC ₅₀ = 8.63 (0.13) E _{max} = 15% (6) partial agonist | pIC ₅₀ = 4.83 (0.30) antagonist | pIC ₅₀ = 7.92 (0.10) antagonist | pIC ₅₀ = 7.52 (0.20) antagonist |
| PP 3 | pIC ₅₀ = 5.01 (0.17) antagonist | pIC ₅₀ = 7.56 (0.23) antagonist | pEC ₅₀ = 7.53 (0.34) E _{max} = 15% (3) partial agonist | pIC ₅₀ = 5.44 (0.15) antagonist | pIC ₅₀ = 8.04 (0.32) antagonist | pIC ₅₀ = 7.86 (0.21) antagonist |

The bold values indicate the number of compounds.

representative example for different SPs (**Supplementary Table S4**). D₂R and D₃R binding data clearly show that the (S,S) enantiomer of the PP is more favourable for receptor binding. The (S,S) enantiomer already plays a prominent role in PP (**22**, **23**), with a 3-fold activity difference between the enantiomers.

The same effect can be observed when using tetrahydroisoquinoline (**24,25**) or indole (**26,27**) SP, although here the difference in activity at the D₃ receptor is about 100-fold (**Table 1**, **Supplementary Table S4**). Compound **27** show a 22.5-fold subtype selectivity towards D₃R that is due to the SP moiety.

TABLE 4 | Functional Data of compounds at D₃R and 5-HT_{2C} (All compounds were tested as HCl salts. For agonist activity, Emax values are shown in brackets. NT, not tested.).

| Cmpd | D ₃ R Gi | D ₃ R Tango | 5-HT _{2C} G _q (Ca ²⁺) |
|-------------|--------------------------------------|---------------------------------------|---|
| (1R,2R)-17b | EC ₅₀ = 3.58 nM (77.9%b) | EC ₅₀ = 126.4 nM (50.2%) | antagonist IC ₅₀ = 14.5 μM |
| (1S,2S)-17a | no agonism; antagonist: Ki = 16.7 nM | NT | antagonist IC ₅₀ = 0.86 μM |
| (1R,2R)-18b | EC ₅₀ = 177.5 nM (71.7%) | 9.2% at 3 μM | antagonist IC ₅₀ = 16.1 μM |
| (1S,2S)-18a | EC ₅₀ = 99.2 nM (83.4%) | 44.4% at 3 μM | agonist EC ₅₀ = 3538 nM (30.3%) |
| (1R,2R)-19b | EC ₅₀ = 87.0 nM (40.7%) | <5% at 3 μM | antagonist: IC ₅₀ > 30 μM |
| (1S,2S)-19a | EC ₅₀ = 142.8 nM (63.4%) | EC ₅₀ = 1,000.2 nM (27.1%) | agonist EC ₅₀ = 2549 nM (44.2%) |
| (1R,2R)-20b | EC ₅₀ = 12.5 nM (68.1%) | 3.1% at 3 μM | antagonist IC ₅₀ = 10.1 μM |
| (1S,2S)-20a | EC ₅₀ = 29.6 nM (96.2%) | EC ₅₀ = 11086 nM (119.1%) | agonist EC ₅₀ = 738.3 nM (51.9%) |

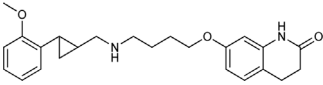
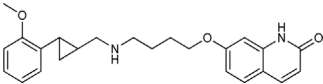
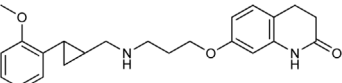
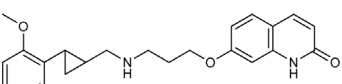
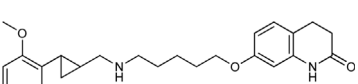
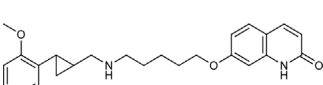
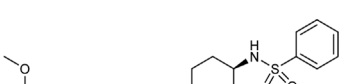

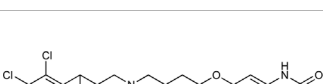
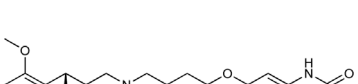
Next the authors investigated the effect of the linker. Changing the original cyclopropylethyl linker (**Supplementary Table S4**) for the racemic derivative (**rac-trans-28**) resulted in 37.3-fold selectivity towards D₃R. Separating the enantiomers, (1S, 2R)-trans-cyclopropyl stereochemistry (**28a**) showed D₃R K_i of 1.85 nM and an unprecedented 47.5-fold selectivity for D₃R over D₂R (D₂R K_i = 87.8 nM), while the other enantiomer (**28b**) has much weaker activity coupled with poor selectivity (Battiti et al., 2019). Finally, two additional linkers were used (**31**, **32**) that are widely used among D₃R bitopic compounds including several high selectivity partial agonists or antagonists (Kumar et al., 2016; Michino et al., 2017; Verma et al., 2018). Compound **31** showed reduced affinity compared to **28a**, inferring that the hydroxyl group on the linker is optimal for antagonism but cannot be directly transferred to the agonist binding mode due to different receptor conformations in the active and inactive states. Compound **32** shows good affinity but neither affinity nor selectivity reaches that of **28a**. Compounds **31**, **32**, **28a** were tested at MOR. For **32** there is a decrease in affinity at the dopamine receptor but the weak subtype selectivity is retained, however there is a 22.9-fold increase in activity at the MOR receptor (**Supplementary Table S4**) (Battiti et al., 2020). The same group synthesized a number of eticlopride analogues using different SPs in the 2-N or 4-C position of pyrrolidine *via* glycerol (Battiti et al., 2020). They found that O-alkylated analogues had better affinity for D₂ and D₃ receptors than the N-substituted derivatives. In BRET assays, these compounds exhibited antagonist or very weak partial agonist behaviour. Docking studies revealed that the SPs of the O-alkylated analogues form aromatic stacking interactions with conserved residues His6.55 and Tyr7.35 both in the D₂ and D₃ receptors, while the SPs of the N-alkylated derivatives extend towards the extracellular site that is less conserved (Shaik et al., 2021).

N-phenylpiperazine analogues were used extensively for constructing bitopic ligands against dopamine receptors. Lee et al. synthesized and evaluated a series of N-phenylpiperazine analogues substituted with 3-thiophen and 4-thiazolylphenylfluoride (**Supplementary Table S5**). They identified several ligands that bind with high affinity to D₃R and exhibit considerable selectivity towards D₂R. Comparison of the binding results of compounds **33–38** and **39–44** suggests that **39–44** binds to D₃R but not to D₂R. The replacement of the thiophene ring by a thiazole ring (**45–50**) led to a decrease in receptor binding selectivity. Compound **39** (**Table 1**) possessed

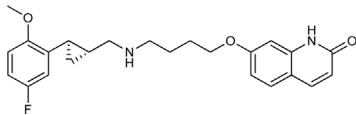
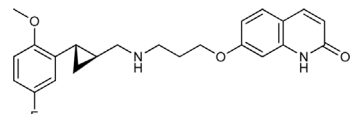
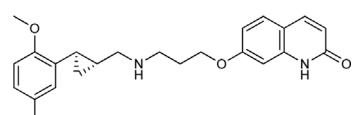
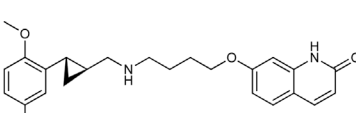
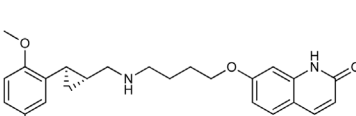
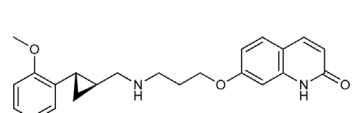
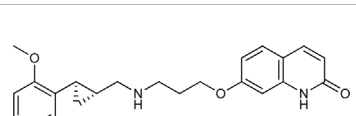
the highest D₃R affinity (K_i = 1.4 nM) and 450-fold selectivity that nominated this compound for *in vivo* testing. Intraperitoneal administration of **39** led to a significant reduction in DOI-dependent head twitch response in mice and a reduction in AIM scores in dyskinetic hemiparkinsonian rats. These data suggest that compound **39** is able to cross the blood-brain barrier and achieves therapeutic concentrations (Lee et al., 2021).

Starting from the 5-HT_{2A} receptor-bound structure of aripiprazole and cariprazine Chen et al. designed D₂/D₃ receptor ligands with no significant 5-HT_{2A} affinity (Chen et al., 2021). The authors suggested that the unusual “upside-down” binding mode (**Figure 3A**) might affect the observed selectivity. According to the structural rearrangements, the location of the SP of aripiprazole in the exosite is important for its signal transduction efficiency. In the interest of identifying residues critical for efficacy, the exosite sequence of the 5-HT_{2A} and D₂ receptors was aligned, with an important difference between the two found at position 5.51, which is Leu in 5-HT_{2A}R and Phe in D₂R. Mutations in D₂R demonstrated that substitution of F202^{5.51} with Leu or Ala reduces the G-protein activity and β-arrestin2 recruitment of aripiprazole. In addition, a derivative of aripiprazole substituted with benzothiazole for the dihydroquinoline ring of D₂R had reduced efficiencies of both G protein activity and β-arrestin2 recruitment. Substitution of L247^{5.51}F in 5-HT_{2A}R did not increase the efficacy of aripiprazole. The results suggest that aripiprazole may stabilize different conformations of TM5 and TM6 between the two receptors. Alignment of 5-HT_{2A}R and D₂R structures (active and inactive) shows that activation of 5-HT_{2A}R requires a larger downstream swing of W6.48 from the CWxP motif than that observed for D₂R activation. In the 5-HT_{2A}R, dihydroquinoline is located deeper in the binding pocket interacting with W336^{6.48}, restricting its movement, whereas it can move gently upon D₂R receptor activation. Similar observations were made for cariprazine. Here, the dynamic coupling between F/L5.51, W6.48 and the PIF motif by the exosite may partly explain why the compounds tested have different efficacies at 5-HT_{2A}R and D₂R receptors. Compared with inactive and active D₂R constructs, the 5-HT_{2A}R-aripiprazole complex in the extracellular compartment shows inward movement of TM6, TM7, and ECL2 toward the seven transmembrane cores. These rearrangements suggest that the 5HT_{2A}R affinity of the bitopic compounds can be reduced by increasing the size of

TABLE 5 | Pharmacological profiling of compounds (D2R binding and functional activity) (Yan et al., 2021).

| Cmpd | Structure | D ₂ R binding K _i nM (pK _i ±SEM) | D ₂ R G _{ai1} BRET EC ₅₀ nM (E _{max} %) (pEC ₅₀ ± SEM) | D ₂ R β-arrestin2 BRET EC ₅₀ nM (E _{max} %) (pEC ₅₀ ± SEM) |
|-------------|---|--|--|---|
| 64 |  | 61.9 (7.21 ± 0.04) | 49.0 (25 ± 2%) (7.31 ± 0.09) | 67.6 (30 ± 1%) (7.17 ± 0.07) |
| 65 |  | 59.9 (7.22 ± 0.13) | 26.3 (52 ± 1%) (7.58 ± 0.08) | 32.4 (53 ± 2%) (7.49 ± 0.14) |
| 66 |  | 125.7 (6.90 ± 0.08) | 9.30 (58 ± 3%) (8.03 ± 0.01) | 10.0 (52 ± 1) (8.00 ± 0.11) |
| 67 |  | 155.7 (6.81 ± 0.03) | 11.2 (65 ± 3%) (7.95 ± 0.04) | 7.08 (60 ± 1%) (8.15 ± 0.12) |
| 68 |  | 259.2 (6.59 ± 0.05) | 891.2 (12 ± 1%) (6.05 ± 0.42) | 416.9 (14 ± 4%) (6.38 ± 0.64) |
| 69 |  | 217.8 (6.66 ± 0.08) | 77.6 (18 ± 1%) (7.11 ± 0.12) | 190.6 (19 ± 1%) (6.72 ± 0.49) |
| 70 |  | 977.2 (6.01 ± 0.11) | 8.45 (68 ± 1%) (8.07 ± 0.11) | 9.49 (16 ± 1%) (8.02 ± 0.06) |
| 71 |  | 244.3 (6.61 ± 0.07) | 34.8 (51 ± 5%) (7.46 ± 0.10) | 94.0 (39 ± 4%) (7.03 ± 0.20) |
| 72 |  | 128.1 (6.89 ± 0.112) | 14.73 (66 ± 3%) (7.83 ± 0.12) | 27.6 (33 ± 1%) (7.56 ± 0.09) |
| (1S,2S)-73a |  | 20.8 (7.68 ± 0.06) | 9.43 (29 ± 3%) (8.03 ± 0.05) | 3.63 (18 ± 1%) (8.44 ± 0.17) |

(Continued on following page)

| Cmpd | Structure | D ₂ R binding K _i nM (pK _i ±SEM) | D ₂ R G _{αi1} BRET EC ₅₀ nM (E _{max} %) (pEC ₅₀ ± SEM) | D ₂ R β-arrestin2 BRET EC ₅₀ nM (E _{max} %) (pEC ₅₀ ± SEM) |
|-------------|---|--|--|---|
| (1R,2R)-73b |  | 43.8 (7.36 ± 0.07) | 12.9 (13 ± 3%) (7.89 ± 0.14) | 1.86 (10 ± 2%) (8.71 ± 0.15) |
| (1S,2S)-74a |  | 6.58 (8.18 ± 0.04) | 4.12 (55 ± 2%) (8.39 ± 0.08) | 4.66 (29 ± 1%) (8.33 ± 0.15) |
| (1R,2R)-74b |  | 362.5 (6.44 ± 0.07) | 62.0 (7 ± 1%) (7.21 ± 0.16) | 14.7 (17 ± 1%) (7.83 ± 0.12) |
| (1S,2S)-75a |  | 11.5 (7.94 ± 0.07) | 8.9 (40 ± 2%) (8.05 ± 0.04) | 2.50 (20 ± 1%) (8.60 ± 0.10) |
| (1R,2R)-75b |  | 30.1 (7.52 ± 0.02) | NT | NT |
| (1S,2S)-76a |  | 12.8 (7.89 ± 0.05) | 3.41 (71 ± 3%) (8.47 ± 0.08) | 8.30 (47 ± 2%) (8.08 ± 0.06) |
| (1R,2R)-76b |  | 317.0 (6.50 ± 0.04) | 197.2 (41 ± 5%) (6.71 ± 0.05) | 70.1 (18 ± 3%) (7.15 ± 0.15) |

Kling et al. investigated the neurotensin receptor type (NTS) 1 receptor crystal structures (White et al., 2012; Egloff et al., 2014) and found that an allosteric binding site was saturated at the C-terminus of NT (8–13). Following sequence analysis, they confirmed that there is a difference between NTS₁R (Arg149^{3,32}) and NTS₂R (His115^{3,32}) that may allow for subtype selectivity. Several bitopic ligands of type NT (8–13)

TABLE 6 | Binding and functional data for enantiomer selective ligands (Yan et al., 2021).

| Cmpd | K_i , nM (pK _i ±SEM) | | | | | |
|---------------|--|--|--|--|--|--|
| | D ₁ R | D ₂ R | D ₃ R | D ₄ R | 5-HT _{1A} | 5-HT _{2A} |
| (1S,2S)-73a | >10,000 | 20.8 (7.68 ± 0.06) | 73.6 (7.13 ± 0.26) | 122.3 (6.91 ± 0.23) | 34.5 (7.46 ± 0.30) | 1.411 (5.85 ± 0.18) |
| (1S,2S)-74a | >10,000 | 6.58 (8.18 ± 0.04) | 22.6 (7.65 ± 0.33) | 304.6 (6.52 ± 0.34) | 19.0 (7.72 ± 0.16) | 519.6 (6.28 ± 0.10) |
| (1S,2S)-75a | >10,000 | 11.5 (7.94 ± 0.07) | 37.6 (7.43 ± 0.29) | 373.0 (6.43 ± 0.21) | 30.3 (7.52 ± 0.05) | 2093 (5.68 ± 0.10) |
| (1S,2S)-76a | >10,000 | 12.8 (7.89 ± 0.05) | 33.9 (7.47 ± 0.28) | 604.0 (6.22 ± 0.09) | 32.8 (7.48 ± 0.13) | 1,160 (5.94 ± 0.12) |
| Aripiprazole | 1,146 (5.94 ± 0.06) | 2.13 (8.67 ± 0.03) | 4.02 (8.40 ± 0.10) | 100.8 (7.00 ± 0.19) | 13.3 (7.88 ± 0.01) | 39.6 (7.40 ± 0.03) |
| caripiprazine | 3414 (5.47 ± 0.11) | 1.45 (8.84 ± 0.07) | 0.27 (9.57 ± 0.21) | 507.0 (6.30 ± 0.16) | 4.01 (8.40 ± 0.06) | 219.4 (6.66 ± 0.05) |
| haloperidol | NT | 6.33 (8.20 ± 0.08) | 22.7 (7.64 ± 0.18) | 26.3 (7.58 ± 0.07) | NT | NT |
| LE300 | 2.93 (8.53 ± 0.13) | NT | NT | NT | NT | NT |
| 5-HT | NT | NT | NT | NT | 6.50 (8.19 ± 0.19) | 79.1 (7.10 ± 0.07) |
| | | | | | | |
| Cmpd | D ₂ R G _α o _a BRET EC ₅₀ , nM (E _{max} %) | | | D ₃ R G _α o _a BRET EC ₅₀ , nM (E _{max} %) | | |
| | D ₂ R G _α o _a BRET EC ₅₀ , nM (E _{max} %) | D ₂ R G _α o _a BRET EC ₅₀ , nM (E _{max} %) | D ₂ R G _α o _a BRET EC ₅₀ , nM (E _{max} %) | D ₃ R G _α o _a BRET EC ₅₀ , nM (E _{max} %) | D ₃ R G _α o _a BRET EC ₅₀ , nM (E _{max} %) | 5-HT _{1A} G _α o _a BRET EC ₅₀ , nM (E _{max} %) |
| (1S,2S)-73a | 7.18 (44 ± 2%) (8.14 ± 0.25) | 5.14 (19 ± 4%) (8.29 ± 0.34) | 52.91 (17 ± 5%) (7.28 ± 0.21) | 95.94 (58 ± 2%) (7.02 ± 0.13) | 95.94 (58 ± 2%) (7.02 ± 0.13) | 95.94 (58 ± 2%) (7.02 ± 0.13) |
| (1S,2S)-74a | 1.60 (66 ± 3%) (8.80 ± 0.13) | 117.6 (23 ± 7%) (6.93 ± 0.21) | 9.84 (18 ± 3%) (8.01 ± 0.07) | 51.96 (49 ± 2%) (7.28 ± 0.10) | 51.96 (49 ± 2%) (7.28 ± 0.10) | 51.96 (49 ± 2%) (7.28 ± 0.10) |
| (1S,2S)-75a | 6.45 (57 ± 2%) (8.19 ± 0.11) | 97.65 (31 ± 2%) (7.01 ± 0.20) | 37.35 (23 ± 4%) (7.43 ± 0.11) | 45.43 (38 ± 2%) (7.34 ± 0.27) | 45.43 (38 ± 2%) (7.34 ± 0.27) | 45.43 (38 ± 2%) (7.34 ± 0.27) |
| (1S,2S)-76a | 2.03 (77 ± 2%) (8.69 ± 0.07) | 129.3 (54 ± 5%) (6.89 ± 0.05) | 12.89 (46 ± 1%) (7.89 ± 0.13) | 58.75 (45 ± 3%) (7.23 ± 0.11) | 58.75 (45 ± 3%) (7.23 ± 0.11) | 58.75 (45 ± 3%) (7.23 ± 0.11) |
| Quinpirole | 1.18 (97 ± 2%) (8.93 ± 0.02) | 1.97 (100 ± 2%) (8.71 ± 0.02) | 3.31 (101 ± 1%) (8.48 ± 0.08) | NT | NT | NT |
| 5-HT | NT | NT | NT | 6.29 (98 ± 2%) (8.20 ± 0.09) | 6.29 (98 ± 2%) (8.20 ± 0.09) | 6.29 (98 ± 2%) (8.20 ± 0.09) |

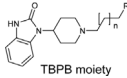
were synthesized (Table 2) and compounds (51–56) showed a promising trend in the NTS₁R selectivity. The best compound (54) has K_i value of 1.3 nM associated with 26-fold selectivity towards NTS₂R. Homology modelling and MD simulations confirmed that the compounds bind in a bitopic mode, with NT (8–13) occupying the orthosteric binding site and the amino acid extension occupying the secondary binding site. These results provide a promising starting point for the design of NTS₁R selective agonists (Kling et al., 2019).

Functional Selectivity

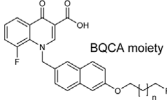
Advances in GPCR structural biology and pharmacology have opened up new opportunities for functional drug design. Modulation of GPCRs through allosteric binding sites can alter receptor structure, dynamics and function, resulting in increased spatial and temporal variation. One important aspect of these changes is functional selectivity or otherwise termed biased signalling. Biased signalling can contribute to the enhancement of the intended effect, but can also cause side effects, so one of the most intriguing areas of current research is investigating the functional character of the ligands in different signalling pathways (Hauser et al., 2017).

Egyed et al. reported a systematic study exploring the extracellular SBP to fine-tune the functional profile of D₂R and D₃R ligands. Introduction of the SP increased affinity at both D₂ and D₃ receptors for each ligand. The study demonstrated that the G_{i/o} and β-arrestin pathways can be specifically modulated from the extracellular vestibule incorporating different SPs to the ligands. Molecular dynamics simulations revealed that G-protein signalling could be linked to the orientation of the PP that is influenced by the SBP binding part of the bitopic compounds (Figure 6). Three PPs and two SPs (Figure 6) were tested using an ethylcyclohexyl linker in analogy to caripiprazine. In the G_{i/o}-mediated signalling pathway, dichlorophenylpiperazine (57) (PP 1) was a partial agonist on both D₂R and D₃R (Table 3). Application of N,N-dimethylurea (SP 1) (caripiprazine) also resulted in a partial agonist with significantly increased potency (D₂R pEC₅₀ = 8.85 nM, E_{max} = 77.4%, D₃R pEC₅₀ = 8.58 nM E_{max} = 27%). The use of the OtBu motif (SP 2) (61) led to a full agonist, the potency on D₂R was superior to that on D₃R. For 2-methoxyphenylpiperazine (2, 58, 62) (PP 2), no prominent change was observed, all were partial agonists. The 3-(piperazin-1-yl)-5-(trifluoromethyl)benzonitrile (59) (PP 3) with the N,N-dimethylurea SP (60), showed antagonist effects on the G protein coupled signalling pathway of D₂R and D₃R, with an increase in potency. Interestingly, incorporating SP 2 (63) turned the function of PP to a weak partial agonist at both receptors. These results suggest that PP and SP affect functionality together. In the β-arrestin signalling pathway, compounds with SP 2 achieve the largest increase in E_{max} values, while this was lower for caripiprazine. (Table 3). This suggests that caripiprazine shows a significant bias towards the G-protein controlled pathway on D₂R. In all cases, the bitopic compounds with 2-methoxyphenylpiperazine PP (2, 58, 62) exhibited antagonist behaviour in contrast to the partial agonism observed in the G-protein coupled signalling pathway. The antagonistic behaviour of 59 was also preserved

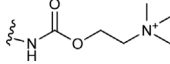
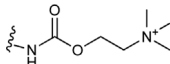
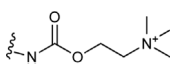
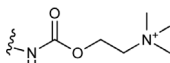
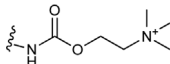
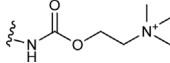
TABLE 7 | Potency and efficacy induced by muscarinic agonists bitopic compounds HEK293t cells overexpressing the M1 receptor (Schramm et al., 2019).



TBPB moiety



BQCA moiety

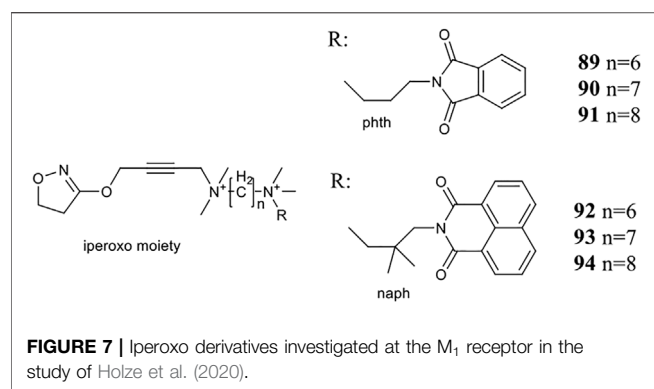
| Cmpd | N | R | pEC ₅₀ nM ± SEM | % E _{max} ±SEM |
|-----------|---|---|----------------------------|-------------------------|
| CCh | | | 6.97 ± 0.03 | 99 ± 1 |
| TBPB | | | 7.32 ± 0.02 | 83 ± 1 |
| BQCA | | | 7.20 ± 0.03 | 90 ± 1 |
| 77 (TBPB) | 1 |  | n.d. | n.d. |
| 78 (TBPB) | 3 |  | 5.09 ± 0.24 | 12 ± 2 |
| 79 (TBPB) | 6 |  | n.d. | n.d. |
| 80 (BQCA) | 1 |  | 5.89 ± 0.01 | 66 ± 0.5 |
| 81 (BQCA) | 3 |  | 6.67 ± 0.02 | 78 ± 1 |
| 82 (BQCA) | 6 |  | 6.62 ± 0.03 | 28 ± 0.5 |
| 83 (TBPB) | 1 | H | 6.05 ± 0.01 | 99 ± 1 |
| 84 (TBPB) | 3 | H | 6.42 ± 0.01 | 97 ± 1 |
| 85 (TBPB) | 6 | H | 7.38 ± 0.04 | 98 ± 2 |
| 86 (BQCA) | 1 | H | 5.82 ± 0.02 | 35 ± 1 |
| 87 (BQCA) | 3 | H | n.d. | n.d. |
| 88 (BQCA) | 6 | H | n.d. | n.d. |

in the β -arrestin signalling pathway; following the previous trends introduction of any SP led to an increase in pIC₅₀ values here as well. In general, the efficacy data measured at both receptors followed similar trends in both modalities as the receptor affinities (Egyed et al., 2020).

High affinity binders, such as **39**, **40**, **41**, **42**, **49** (**Supplementary Table S5**) were also tested for their efficacy on D₃R, both by examining forskolin-dependent inhibition of adenylyl cyclase and by measuring β -arrestin binding.

Compounds **42** and **49** were found antagonists in both assays. Compound **41** display functional selectivity, being a weak partial agonist in the adenylyl cyclase assay and a very weak partial agonist/antagonist in the β -arrestin binding assay. Compounds **39** and **40** exhibit weak partial agonism in both the adenylyl cyclase inhibition and β -arrestin binding assays (Lee et al., 2021).

Investigating pure enantiomeric forms of compounds **17–20** (**Supplementary Table S3**) Tan et al. showed that the (R,R) enantiomers (**17b–20b**) have a better affinity for D₃R than (S,S)



(**17a-20a**), with the exception of compound **17**, which had an identical affinity for both of the enantiomers (**17a**, **17b**) (Tan et al., 2020). The (R,R) isomers (**17b-20b**) showed weaker affinity (3–20-fold) towards 5-HT_{2C}R than their (S,S) counterparts (**17a-20a**). The data suggest that D3R is less sensitive to conformational changes than the 5-HT_{2C} receptor. Functional studies were also performed with the **17a,b-20a,b** (Table 4). Compounds **18–20** were all full or partial agonist on D₃ receptors, whereas for 5-HT_{2C}R the (S,S) enantiomers (**18a-20a**) are weak partial agonists, whereas the (R,R) enantiomers (**18b-20b**) are weak antagonists. Compared to the binding assay, functional results indicate greater selectivity towards D₃R. Furthermore, these compounds showed only very weak partial agonism at 5-HT_{2A}R and no affinity at 5-HT_{2B}R. The two enantiomers of compound **17** exhibit opposite behaviour, while (**1R,2R**)-**17b** was a potent agonist (EC_{50} = 3.6 nM, E_{max} = 77.9%), (**1S,2S**)-**17a** was an antagonist on D₃R with a K_i of 16.7 nM, and both derivatives were weak antagonists with micromolar activity on 5-HT_{2C} receptor. Docking studies suggested a difference between the two compounds (**17a,17b**) in the orientation of PP. In the case of the agonist (**1R,2R**)-**17b**, the

2-methoxy group is deep in the OBP and forms hydrophobic interactions with residues C114^{3,36}, S196^{5,46}, and F346^{6,52}. In the case of the antagonist (**1S,2S**)-**17a**, the 2-methoxy group flips out to the extracellular side and the cyclopropane linker between the benzene ring and the protonated N overlays perfectly with the amide linker of eticlopride, which is not present in the agonist. Compounds (**1S,2S**)-**17a**, (**1R,2R**)-**18b**, (**1R,2R**)-**19b**, and (**1R,2R**)-**20b** were inactive in the Tango assay on D₃R, indicating their preference for the G-protein signalling pathway. For further profiling (**1R,2R**)-**17b** and (**1R,2R**)-**19b** were tested on 29 other aminergic GPCRs that confirmed their good selectivity for D₃R (Tan et al., 2020).

Yan et al. also used PCPMA analogues as PP, with propyl, butyl, pentyl, or cyclohexylethyl linkers, and SP groups taken from aripiprazole, brexpiprazole, and cariprazine, respectively. The synthesized library was measured in D₂R binding, D₂R G_i and D₂R β -arrestin BRET assays (Table 5). The starting compound (**64**) exhibits good affinity (K_i = 61.9 nM) and partial agonist activity in both G_i (EC_{50} = 49.0 nM, E_{max} = 25%) and β -arrestin (EC_{50} = 67.6 nM, E_{max} = 30%) BRET assays. In comparison, replacement of SP with quinolone (**65**) increased the potency two-fold with unchanged binding. Changing the linker to propyl (**66,67**) led to a small decrease in binding affinity but an increase in efficacy (~10 nM EC_{50} values and E_{max} values higher than 50%). Lengthening the linker to 5C units (**68,69**) led to a decrease in binding affinity and functional activity. The cariprazine-like SP (dimethylamine) and linker (cyclohexyl) with this PP did not show significant activity. The best compound from this series (**70**) has very potent partial agonist character in both G_i BRET (EC_{50} = 8.45 nM, E_{max} = 68%) and β -arrestin2 recruitment assays (EC_{50} = 9.49 nM, E_{max} = 16%), with a much lower E_{max} in the latter. The significant difference between binding affinity and potency for many of these compounds likely reflects the use of an antagonist radioligand [³H]-N-methylspiperone] in the competitive

TABLE 8 | Binding affinities and functional efficacies of NAQ and NCQ (Wang et al., 2020).

| Cmpd | K _i (nM±SEM) | | | MOR vs. KOR | MOR vs. DOR | MOR (³⁵ S) GTPγS binding | |
|------------|-------------------------|--------------|----------------|-------------|-------------|--------------------------------------|-----------------------------------|
| | MOR | KOR | DOR | | | EC ₅₀ (nM±SEM) | E _{max} of DAMGO % ± SEM |
| <p>NAQ</p> | 0.55 ± 0.15 | 26.45 ± 5.22 | 132.50 ± 27.01 | 48 | 241 | 4.36 ± 0.72 | 15.83 ± 2.53 |
| <p>NCQ</p> | 0.55 ± 0.01 | 22.20 ± 2.10 | 33.90 ± 0.50 | 40 | 62 | 1.74 ± 0.13 | 51.00 ± 0.40 |

binding assay, from which an agonist ligand tends to show much lower apparent binding affinity. Attempts have been made to use several PPs but these have been shown to give significantly worse results than the methoxy derivative. In the case of isoquinoline and tetrahydroisoquinoline SP, it was not practical to use the dichlorophenyl motif in the PP (**71,72**). The best results were obtained with derivatives containing halogen in the meta position on the phenyl group of PP and methoxy in the ortho position (**73a,b-76a,b**). Pure forms of the enantiomers were also investigated. The majority of the fluorinated derivatives (**(1S,2S)-42a**, **(1R,2R)-73b**, **(1S,2S)-74a**) showed K_i values below 50 nM on binding assay and EC_{50} values below 20 nM in both G_i and β -arrestin2 BRET assays. The same trend was observed for the chlorinated derivatives [**(1S,2S)-75a**, **(1S,2S)-76a**]. Higher E_{max} was observed for the halogenated derivatives in the G_i signal transduction than in the β -arrestin. After separation of the enantiomers, it was confirmed that the (S,S)-isomers were more efficient in D_2R binding and functional assay. The (R,R) compounds exhibit partial agonist behaviour and the E_{max} values are higher for G_i signaling. The selectivity of the compounds [**(1S,2S)-73a**, **(1S,2S)-74a**, **(1S,2S)-75a**, **(1S,2S)-76a**] was investigated on D_1R , D_2R , D_4R , D_5R , 5-HT_{1A}R, 5-HT_{2A}R, and 5-HT_{2C}R, with low selectivity observed towards the D_3 receptor and potent activity on the 5-HT_{1A} receptor, and good or acceptable selectivity on the other receptors (**Table 6**). In the case of D_3R , these compounds showed weak partial agonist activity in both G_o and β -arrestin2 BRET assays, albeit with different efficacies. For the 5-HT_{1A} receptor, all four compounds (**(1S,2S)-73a**, **(1S,2S)-74a**, **(1S,2S)-75a**, **(1S,2S)-76a**) were similar partial agonists in G_i BRET assays. The lack of selectivity over D_3R and 5-HT_{1A}R should not be a concern for these compounds, as both D_3R and 5-HT_{1A}R have been shown to be involved in the therapeutic effects of some antipsychotics. Overall, these four compounds have shown an interesting pharmacological profile (Yan et al., 2021).

Schramm et al. investigated the effect of bitopic compounds on muscarinic acetylcholine receptors. Carbachol (CCh) PP was cross-linked to allosteric ligands by linkers of different lengths (1C, 3C, 5C, 8C). The benzoimidazole-piperidine moiety of TBPB [1-(1'-(2-tolyl)-1,4'-bipiperidin-4-yl)-1H-benzo(d)imidazol-2(3H)-one], a known selective bitopic M_1R agonist, and BQCA (benzyl quinolone carboxylic acid) derivatives, that are PAMs, were used as allosteric modulators (**Table 7**). It was found that BQCA-CCh bitopic compounds act as agonists. The highest potency and efficacy was observed for the compound containing BQCA moiety **81**. Comparing with reference compound **86**, which does not contain a CCh moiety but only the linker, revealed that the CCh moiety provides some of the agonist activity. In contrast, the TBPB-CCh bitopic ligand (**78**) showed partial agonism, while the reference **84** was a full agonist. The binding mode of **81** was investigated by docking to an active receptor model. The ammonium group of the CCh moiety forms a charge-assisted hydrogen bond with D105^{3,32}, while the carbamate carbonyl group serves as a hydrogen bond acceptor for the hydroxyl group of Y408^{7,43}. This is different from the carbachol binding mode, in which the carbamate structure has a

different orientation. The BQCA moiety, located in the region of the extracellular loop, is stabilized by hydrophobic contacts with L174^{ECL2} and Y179^{ECL2} and a charge-assisted H-bond with K392^{ECL3}. They concluded that partial agonism through bitopic compounds can be achieved not only by quenching orthosteric receptor activation by an allosteric moiety as in **81** but also by quenching bitopic activation of the receptor by an orthosteric moiety such as CCh in **78** (Schramm et al., 2019).

Holze et al. have shown that allosteric coupling of the M_1R can induce conformational changes that affect intracellular signalling. They investigated two groups of M_1R bitopic agonists and varied the length of the linker. Iperoxo, a known agonist, was selected as the PP motif, while two negative allosteric modulators, phtp (**89-91**) and naph (**92-94**), were incorporated as SP. (**Figure 7**) The latter differs from the phtp derivative in two main respects: naph contains a larger and branched aliphatic linker. The two pharmacophores were linked by alkyl chains of different length (6-8C) (**89-94**). While the ligand affinities for the allosteric binding site were very similar within a ligand set, the ligand affinities for the orthosteric binding site depended on the length of the linker, where increasing linker length was correlated with increasing ligand affinity. From this information, it was concluded that the same binding mode was adopted by iperexo in a series of bitopic compounds driven by its high affinity, and this was confirmed by MD simulations. Therefore, a series of bitopic ligands differing only in the length of the linker may be suitable to investigate the effect of allosteric coupling on signal transduction with subnanometer accuracy. Whereas the longest bitopic agonist, **91**, was able to stimulate all three G-protein families, **90** activated $G_{q/11}$ and G_s proteins, **89** promoted signal transduction only via $G_{q/11}$. **93** and **94** only activated $G_{q/11}$ protein signalling, while **92** did not activate any signalling pathway, unlike **89**. None of the naph-based ligands were able to activate G_s and $G_{i/o}$ signalling. These data suggest that different G-proteins show different sensitivities to M_1R activation by these bitopic compounds. While $G_{q/11}$ coupling is conserved in almost all bitopic ligands, G_s signalling is promoted only by two members of the phtp series. $G_{i/o}$ activation is particularly sensitive to the bitopic ligand structure with only **91** showing weak $M_1R/G_{i/o}$ coupling among the compounds tested. MD simulations show that binding of iperexo results in a complete contraction of the extracellular parts of the ligand binding pocket. In contrast, the bitopic ligands of the phtp series bind in such a way that they sterically inhibit the closure of the binding pocket. The extent of the conformational interference depends on the length of the linker and hence the position of the allosteric building block. Since the phtp part of **89** is located close to the orthosteric binding site, it inhibits closure, resulting in a more open extracellular conformation. Elongation of the linker with additional methylene groups allowed for subnanometer regulation of the position of the allosteric building block, thereby progressively reducing the closure of the binding pocket, ultimately resulting in greater G-protein binding capacity. FRET measurements have demonstrated that the more closed ligand-binding pocket is associated with greater receptor conformational changes at the G-protein binding surface *via* an allosteric coupling mechanism. Consistent with

this idea, **92**, a bitopic ligand with a branched and larger allosteric motif, did not induce conformational changes in M₁R (Holze et al., 2020).

Wang et al. investigated two naltrexone derivatives substituted with isoquinoline at MOR. The isoquinoline moiety of these bitopic compounds is the SP that interacts with the allosteric site of MOR, and the epoxymorphinan moiety is the PP (**Table 8**). **NAQ** has a high affinity for MOR ($K_i = 0.55$ nM) and high selectivity for κ -opioid receptor (KOR) (48-fold) and δ -opioid receptor (DOR) (241-fold). Compared to DAMGO, it acts as a MOR antagonist in the 35 S-GTP [γ S]-binding assay with CHO cell lines expressing MOR. It showed less significant withdrawal effects compared to the well-known opioid antagonists naloxone and naltrexone. Similar properties were observed for the compound **NCQ** ($K_i = 0.55$, 40-fold KOR, 62-fold DOR selectivity), which shares the same PP part as **NAQ** and differs only in the SP. **NCQ** contains a methoxy at position 1 and a chloro functional group at position 4 of isoquinoline. However, in 35 S-GTP (γ S)-binding assay, **NCQ** behaved as a partial agonist. MD simulations and free energy calculations proposed that the allosteric part of **NAQ** and **NCQ** bind differently in the inactive structure and in the active structure, respectively. Docking studies have shown that the SP parts of **NAQ** and **NCQ** may occupy two different subdomains of the allosteric site of MOR, named ABD1 and ABD2. MD simulations were performed with three poses (**NAQ** inactive, **NCQ** active and inactive) obtained from the docking calculations and showed that the SP part of **NAQ** was bound to ABD1 in the inactive MOR. Although the SP motif occupied an allosteric site, no significant modulatory effect was observed on the binding of the PP, similar to the function of a silent allosteric modulator. In the inactive and active MOR the SP of **NCQ** showed positive allosteric modulation through binding to ABD2. Molecular modelling combined with interaction energy and distance analyses unravelled the molecular mechanisms of allosteric modulation of **NAQ** and **NCQ** and emphasized the importance of the chlorine and methoxy substituents of the isoquinoline ring for the allosteric modulatory function of **NCQ** (Wang et al., 2020).

Binding Kinetics

Although ligand-receptor binding kinetics might have a fundamental role in the development of drug candidates, it is still often overlooked in the early phase of drug discovery. In line with the increased interest in the field, more and more kinetics data (among others association and dissociation rate, residence time, etc.) have been published in the literature, however the magnitude still lags behind the amount of affinity and selectivity data available especially regarding only the allosteric and bitopic ligands. Furthermore, the interpretation of the kinetic data might be hindered by the probe dependence as observed in a prototypical competitive radioligand binding assay for H₁ receptor antagonists, although that aspect is often not considered (Bosma et al., 2019). In line with the relatively limited amount of recent papers, first we refer the readers to recent general review articles on binding kinetics (Sykes et al., 2019; Hoare et al., 2020; Rafael et al., 2020; van der Velden et al.,

2020). Very recently a book chapter collecting available kinetic data of GPCR ligands together with experimental evidence for properties that influence the residence time were published (Potterton et al., 2022). The repository enables researchers to analyse the relationship between the structure and the kinetic parameters as well as provides data for the development of predictive algorithms. The authors also outline machine learning workflows to predict residence time. Sykes et al. reviewed recently the literature related to the binding kinetics of GPCR ligands (Sykes et al., 2019). They discussed the theoretical aspects, the experimental methods and their limitations, detailed several factors influencing binding kinetics among others they explored the role of allosteric modulators, that by definition act through the modulation of the binding kinetics of the endogenous or orthosteric ligands. The authors also discuss some molecular level features including shielding the hydrogen bonds from water that affects the binding kinetics.

Although shielding the hydrogen bonds was thought to decrease residence time, in a recent case study on CCR2 receptor, MD simulations of Magarkar et al. suggested that even shielding an intra protein hydrogen bond can enhance the residence time of ligands through the preservation of the binding site rigidity (Magarkar et al., 2019). The ECL2 loop, that is regularly engaged with bitopic compounds, was also proposed to modulate the binding kinetics (Sykes et al., 2019; van der Velden et al., 2020). Already one of the seminal works in the field of modelling the binding pathway to GPCRs, which investigated the binding of three antagonists and an agonists to the β_2 -adrenoreceptor and one agonist to the β_1 -adrenoreceptor with MD simulations, highlighted the role of the ECL2 loop and the extracellular vestibule. Interestingly, even the highest barrier of binding often corresponds to the association with the extracellular vestibule even though the binding requires conformational change of the receptor and the ligand has to enter through a narrow passage (Dror et al., 2011). In several receptors, ECL2 were proposed to function as a lid facilitating the entrance and exit of the ligands (Thomas et al., 2016; Wacker et al., 2017; Frank et al., 2018). One of these studies investigated the binding kinetics of cariprazine and aripiprazole. As a prototypical bitopic compounds we exemplify here the effect of the SBP on the binding kinetics through them (Frank et al., 2018). At the D₃ receptor, aripiprazole exhibits a slow monophasic dissociation, while cariprazine displays a rapid biphasic behaviour. Interestingly, in the D₂ receptor both compounds display a slow dissociation. These differences may influence the *in vivo* action of the drugs. Interactions with ECL2 residues influence the residence time in other receptors like in the β_2 and A_{2A} receptors, as well (Guo et al., 2016; Masureel et al., 2018). Gaussian accelerated molecular dynamics revealed the role of the ECL2 loop in the formation of allosteric sites for PAMs in the adenosine A₁ receptor (Miao et al., 2018) and unveiled an intermediate binding site between ECL2 and TM1 for caffeine in the adenosine A_{2A} receptor. The authors analysed the effect of more general features like physicochemical properties of the ligand (e.g., lipophilicity) and close contact residue numbers on the drug-receptor dissociation.

Van der Velden et al. summarized structural considerations in relation to binding kinetics presenting the results through four case studies (van der Velden et al., 2020). They showcased the role of the ECL2 loop in the regulation of the ligand kinetics through tiotropium binding to the M₃R and M₂R receptors (Kruse et al., 2012; Tautermann et al., 2013). The more open, flexible ECL2 loop conformation was linked to the shorter residence time observed in the M₂R receptor. Through the example of ZM241385, an A_{2A} receptor antagonists they highlighted the role of molecular dynamics and mutation experiments in providing structural background for observed kinetics behaviour (Guo et al., 2016). Another example was focused on the β_2 adrenoreceptor. Salmeterol, a bitopic compound displays a 5–7 fold higher residence time compared to salbutamol and epinephrin, both binding only to the orthosteric site (Figure 3C). As salmeterol and salbutamol share the orthosteric binding motif, the interactions in the extracellular site are linked to the increased residence time (Masareel et al., 2018). They also discussed other aspects, like the effect of natural receptor variants, ligand variants and probe dependency.

Riddy et al. investigated the binding kinetics of H₃ receptor antagonists/inverse agonists (Riddy et al., 2019). Although the binding mode of the compounds were not investigated experimentally, they likely form interactions outside the orthosteric pocket, too therefore can be considered bitopic. The different pharmacological profile and the residence time of the compounds might be linked to their preclinical and clinical efficacy. Furthermore, H₃ and off-target sigma-1 receptor occupancy may contribute to paradoxical efficacy of some compounds. In the study of Pedersen et al. (2020) the differential binding kinetics profile of the agonists were not linked to the functional bias, as the bias profile of the selected agonists were not time-dependent and despite the difference in their binding kinetic properties they can display the same degree of bias.

Bitopic compounds and allosteric modulators may directly bind to the secondary binding pocket, however, during the association and dissociation process the secondary site plays a crucial role for the appropriate positioning of all compounds. While experiments rarely shed light on the structural details of binding, molecular dynamics simulations can explore the atomistic process and are useful to predict residence time (Potterton et al., 2019; Decherchi and Cavalli, 2020; Lamim Ribeiro et al., 2020; Salmaso and Jacobson, 2020; Bekker et al., 2021; Kokh and Wade, 2021). Ribeiro and co-workers recently used machine learning and infrequent metadynamics to efficiently predict kinetic rates, transient conformational states, and molecular determinants of drug dissociation on the MOR (Lamim Ribeiro et al., 2020). While both investigated compounds bind to the orthosteric pocket, the transient conformational state for the dissociation was identified around the secondary binding pocket suggesting a key role of the secondary site in the association/dissociation process. In dynamic docking simulations Bekker et al. investigated β_2 -adrenoreceptor antagonists identifying several stable and metastable conformational states for the compounds along their association/dissociation path (Bekker et al., 2021). Based on

these simulations they propose a way to develop allosteric modulators to inhibit the receptor by blocking the path of the endogenous ligand to the orthosteric site. Metastable binding sites play a crucial role in the study of Gaiser et al. as well (Gaiser et al., 2019). They developed homobivalent bitopic ligands for β_2 AR to target the OBP and a previously identified metastable binding site as an allosteric site. Among others the residence time of orthosteric and bitopic A_{2A} receptor binders was predicted with ensemble based steered molecular dynamics (Potterton et al., 2019). Analysis of the pathways revealed dominant interactions, residues influencing the dissociation time and the calculations proposed that changes in water-ligand energy from the ligand in the binding pocket to the extracellular vestibule was the main factor in the determination of residence time. While hydrophilic ligands are expected to access the orthosteric binding site, that is deeply embedded in the center of the receptor, from the aqueous phase, hydrophobic compounds were proposed to entry through lipid pathways. The examples detailed in this part explore the traditional pathway, however cholesterol and other ligands might enter the receptor from the membrane. As an exciting study we refer to the work of Guixà-González et al. who investigated the cholesterol access to the A_{2A}R with combined computational and experimental methods. They showed that cholesterol's impact on A_{2A}R-binding affinity goes beyond pure allosteric modulation and unveils a new interaction mode between cholesterol and the A_{2A}R (Guixà-González et al., 2017). Similar findings were collected and analysed in a recent review dedicated to the role of the lipid bilayer in the binding of the ligands to the orthosteric and allosteric sites (Szenk et al., 2019). Even though in this review we focused mainly on the secondary binding pocket in the extracellular vestibule that is accessible through the aqueous phase, some allosteric sites on the receptor surface can only be targeted through the membrane fortifying that investigation of the binding pathways through the membrane is also crucial.

DESIGN APPROACHES FOR ALLOSTERIC AND BITOPIC COMPOUNDS

During the previous sections we often pointed out the value of computational approaches in the investigation of both allosteric and bitopic compounds. Due to the tremendous number of studies a comprehensive overview of the computational approaches to design allosteric (Wold et al., 2019; Chatzigoulas and Cournia, 2021) and bitopic ligands (Newman et al., 2016, 2020; Fronik et al., 2017) for GPCRs warrant a separate review (Basith et al., 2018; Raschka and Kaufman, 2020; Ballante et al., 2021), we could only highlight here a few important studies to draw attention towards their usefulness in drug discovery settings (Dehua Yang et al., 2021).

Allosteric sites are less conserved and therefore they can be exploited to design ligands with high selectivity and modalities that could not be achieved from the orthosteric site. The increasing number of experimental GPCR structures urges the use of structure-based methods. However, the identification of the allosteric sites remains challenging as they often form fully

only in the presence of an allosteric ligand following an induced fit mechanism. Nevertheless, several computational approaches were developed to facilitate the spotting of new allosteric sites like Allosite (Huang et al., 2013), AlloFinder (Huang et al., 2018), ExProSE (Greener et al., 2017), Fpocket (Le Guilloux et al., 2009; Schmidtke et al., 2010), FTmap (Brenke et al., 2009; Kozakov et al., 2015), GRID (Goodford, 1985), LIGSITE^{cs} (Huang and Schroeder, 2006), SiteMap (Halgren, 2009) and MixMD (Ghanakota and Carlson, 2016). FTMap and FTSite was recently shown to perform well on identifying GPCR allosteric sites with limitations on those occurring on the protein-membrane interface that could be attributed to the development of the program originally for soluble globular proteins (Wakefield et al., 2019).

Even after the identification of the allosteric site, simple docking might not always be successful due to induced fit binding. Furthermore, allosteric modulators are prone to “steep” SAR, obscure relationship between the binding affinity and functional effect and slow kinetics (on and/or off rates) that hinders their discovery and design (Congreve et al., 2017b). Huang et al. developed a protocol combining homology modelling and docking to find novel allosteric modulators of the orphan GPR68 and GPR65 receptors (Huang et al., 2015). They generated over three thousand homology models and docked their experimentally validated active compound lorazepam and decoy compounds to identify putative binding sites. They optimized the binding site around the bound ligand and redocked the ligand and the decoys again until a stable docking mode emerged. That plausible binding site was utilized to dock over 3.1 million lead-like compounds. From the selected 17 hits four increased cAMP production. Docking close analogues of the hit compounds lead to another 25 compounds for testing among them 13 with higher activity than the reference compound lorazepam. Similar protocol was utilized for the GPR65 receptor as well showing that the protocol might be applied to a broader field. While this protocol might be applied to several—even orphan—GPCRs it requires at least one experimentally determined known active compound that might be hard to get for other orphan GPCRs and close enough homology to templates that warrant the homology modelling. Nevertheless, this is a great example how the combined experimental and computational approaches can lead to the identification of novel allosteric modulators even for orphan GPCR targets. Miao et al. focused on the identification of novel, chemically diverse allosteric modulators of the M₂ receptor (Miao et al., 2016). The authors used accelerated molecular dynamics to account for receptor flexibility and to generate an ensemble of structures for docking. After retrospective validation virtual screening coupled with induced fit docking (IFD) was applied to select compounds targeting the IXO-nanobody-bound active and the QNB-bound inactive M₂ mAChR for testing. The method successfully identified both positive and negative allosteric modulators and clearly demonstrate that accounting for receptor flexibility is a key in the discovery of allosteric modulators. Nevertheless, for less flexible binding sites even simple docking protocols might be plausible as demonstrated by Korczynska et al. identifying a

positive allosteric modulator that potentiates antagonist binding leading to subtype selectivity at the M₂ muscarinic acetylcholine receptor (Korczynska et al., 2018). Since allosteric modulators are often small and rigid compounds, fragment based approaches (Keserű et al., 2016) emerge as a plausible choice for the design that is supported by several successful application (Christopher et al., 2015; Orgován et al., 2019). Furthermore, covalent approaches should not be overlooked either to aid structurally informed rational design (Lu and Zhang, 2017; Bian et al., 2020; Wenchao Lu et al., 2021).

Bitopic compounds are in the forefront of drug development for GPCRs as they can combine the advantages of targeting the orthosteric and a secondary site (Newman et al., 2016, 2020; Fronik et al., 2017). Fragment based methods are often applied to design novel bitopic compounds (Vass et al., 2014; Egyed et al., 2021). Recently our group have developed a computational protocol to design specific, selective receptor ligands (Egyed et al., 2021). First fragments were docked to the orthosteric binding site of the receptors available in experimental structures (D₃: PDB ID: 3PBL (Chien et al., 2010), 5-HT_{1B}: PDB ID: 4IAQ (Wang et al., 2013), 4IAR (Wang et al., 2013); 5-HT_{2B}: PDB ID: 4IB4 (Wacker et al., 2013), 4MC3 (Liu et al., 2013); H₁: PDB ID 3RZE (Shimamura et al., 2011) and M₁: PDB ID: 5CXV (Thal et al., 2016)), or a homology model in case of the D₂ receptor. Then, virtual fragment screening was performed against the secondary binding site of the combined protein-ligand complex. The identified SBP fragment was then linked to the OBP core by a linker. As a control, the resulting bitopic compounds were docked back into the initial crystal structure. This protocol has been validated by designing selective D₂/D₃, 5-HT_{1B}/5-HT_{2B} and H₁/M₁ receptors. Docking-based fragment evolution approach utilizes the same methodology as exemplified on the design of β_1 and β_2 receptor bitopic compounds (Chevallard et al., 2021). The fragment evolution protocol merges fragment growing with a matrix-based strategy that was originally implemented for potency optimization (Chevallard et al., 2019). First, possible OBP fragments were docked and they were evaluated using the concept of ligand efficiency. Next, fragment growing surrogates suitable for reactive alkylation were defined and docked to the secondary binding pocket. Surrogates that overlap with the core OBP fragment or was marked favourably in both receptors were removed from the top ranked compounds, the remaining top surrogates were kept for further investigation. The OBP fragments were reacted *in silico* with the surrogates, the resulting compounds were docked into the receptors to ensure pose fidelity. Based on these calculations the best surrogates were selected as secondary binding motif for the β_1 and β_2 receptor, respectively. The approach was validated by the synthesis and experimental evaluation of the designed compounds. Classical docking and virtual screening approaches could be also utilized for the development of bitopic compounds (Cao et al., 2018) and even to develop fluorescent GPCR probes (Prokop et al., 2021). We highlight here a study that utilized structure guided design of GPCR polypharmacology (Kampen et al., 2021). Kampen et al. aimed to design dual A_{2A}/D₂ bitopic compounds that was very challenging due to the significantly

different binding sites of the receptors. First, docking based structural analysis confirmed that dual-target ligands of the A_{2A}R and D₂R could be obtained by targeting the orthosteric and secondary pockets. Then, they designed potential dual targeting virtual chemical libraries that could be rapidly synthesized. The prepared libraries were screened virtually with docking on the A_{2A} and the D₂ receptor to select hits. From them one promising compound was selected and developed further with SAR investigations.

Discussing the recent advances in the allosteric and bitopic field we pointed out several times the usefulness of MD based methods. These simulations can explore the differences in the interaction patterns of congeneric molecules more sensitively compared to docking that could be important to understand the different functional outcome of these ligands (Egyed et al., 2020) and to design compounds with specific pharmacological profile (McCorvy et al., 2018) and they might reveal cryptic pockets opened by ligands (Ferruz et al., 2018) that might be overlooked in simple docking calculations. Mutation studies combined with extensive molecular dynamics modelling the dissociation of the ligands was utilized to clarify the structural basis of the long duration of action and kinetic selectivity of tiotropium for the M₃ receptor (Tautermann et al., 2013). A similar study aimed to clarify the molecular determinants of the bitopic binding mode of a negative allosteric modulator of the dopamine D₂ receptor (Draper-Joyce et al., 2018). MDs combined with docking linked the degree of closure of the extracellular loop region to the extent of ligand bias and highlighted the importance of the appropriate receptor conformation for virtual screening at the 5-HT_{2B} receptor (Denzinger et al., 2020). A similar concept was presented by Bermudez et al. proposing that agonists with extended binding modes selectively interfere with binding pocket closure and through divergent allosteric coupling that leads to ligand bias (Bermudez and Bock, 2019).

The structure-based methods clearly benefit from the increase of published GPCR structures, especially that more and more active structures are available, however the design still remains challenging. Nevertheless, with more template available for homology modelling and the publication of AlphaFold (Jumper et al., 2021) facilitate the structure-based methods for targets previously out of scope for these methods broadening the applicability spectrum. While we mainly highlighted structure based approaches classical ligand based methods and cheminformatics also contribute to the development of bitopic GPCR ligands (Basith et al., 2018; James and Heifetz, 2018; Raschka and Kaufman, 2020).

REFERENCES

- Ackley, M. A. (2008). Morpholine Dopamine Agonists for the Treatment of Pain. WO 087512 A1.
- Ágai-Csongor, É., Domány, G., Nógrádi, K., Galambos, J., Vágó, I., Keserű, G. M., et al. (2012). Discovery of Cariprazine (RGH-188): A Novel Antipsychotic Acting on Dopamine D3/D2 Receptors. *Bioorg. Med. Chem. Lett.* 22 (10), 3437–3440. doi:10.1016/j.bmcl.2012.03.104
- Allerton, C. M. N., Cook, A. S., Hepworth, D., and Miller, D. C. (2005). Aminopyridine Derivatives as Selective Dopamine D3 Agonists. WO 115985 A1.

CONCLUSION

GPCRs are one of the largest families of receptors and are among the most targeted proteins for drug discovery. One of the major challenges in the field is the identification of subtype and functionally selective compounds with high potency, designed efficacy and appropriate binding kinetics profile, which are essential to avoid side effects. The secondary binding pocket plays a prominent role in achieving selectivity, while orthosteric ligands are mainly responsible for affinity and functional activity. Bitopic compounds combine the properties of orthosteric and allosteric pharmacophores. With the continuous expansion of available GPCR structures, the secondary binding sites of the receptors are becoming better understood, allowing the construction of complex ligands with designed pharmacological profile. In this review, we have provided an insight into allosteric modulators of class A GPCRs and a detailed review of bitopic compounds that have been released in the last years. We have highlighted the influence of the secondary site in affinity, selectivity, functional selectivity and binding kinetics. The increasing amount of pharmacological data and new structures together with appropriate modelling tools can contribute to the design of allosteric and bitopic drug candidates with an optimized pharmacology profile and thus accelerating the drug discovery against diseases with high unmet medical need.

AUTHOR CONTRIBUTIONS

AE and DK wrote the first draft of the paper and prepared the Figures. GK developed the concept of the paper and contributed to write the manuscript.

FUNDING

This work was supported by a grant from the National Brain Research Program of Hungary (2017-1.2.1-NKP-2017-00002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.847788/full#supplementary-material>

- Ballante, F., Kooistra, A. J., Kampen, S., de Graaf, C., and Carlsson, J. (2021). Structure-Based Virtual Screening for Ligands of G Protein-Coupled Receptors: What Can Molecular Docking Do for You. *Pharmacol. Rev.* 73 (4), 1698–1736. doi:10.1124/pharmrev.120.000246
- Barnes, N. M., Ahern, G. P., Becamel, C., Bockaert, J., Camilleri, M., Chaumont-Dubel, S., et al. (2021). International Union of Basic and Clinical Pharmacology. CX. Classification of Receptors for 5-Hydroxytryptamine; Pharmacology and Function. *Pharmacol. Rev.* 73 (1), 310–520. doi:10.1124/pr.118.015552
- Barresi, E., Martini, C., Da Settimo, F., Greco, G., Taliani, S., Giacomelli, C., et al. (2021a). Allosterism vs. Orthosterism: Recent Findings and Future Perspectives on A2B AR Physio-Pathological Implications. *Front. Pharmacol.* 12, 384. doi:10.3389/fphar.2021.652121

- Barresi, E., Giacomelli, C., Marchetti, L., Baglini, E., Salerno, S., Greco, G., et al. (2021b). Novel Positive Allosteric Modulators of A2B Adenosine Receptor Acting as Bone Mineralisation Promoters. *J. Enzyme Inhib. Med. Chem.* 36 (1), 286–294. doi:10.1080/14756366.2020.1862103
- Basith, S., Cui, M., Macalino, S. J. Y., Park, J., Clavio, N. A. B., Kang, S., et al. (2018). Exploring G Protein-Coupled Receptors (GPCRs) Ligand Space via Cheminformatics Approaches: Impact on Rational Drug Design. *Front. Pharmacol.* 9, 128. doi:10.3389/fphar.2018.00128
- Battiti, F. O., Cemaj, S. L., Guerrero, A. M., Shaik, A. B., Lam, J., Rais, R., et al. (2019). The Significance of Chirality in Drug Design and Synthesis of Bitopic Ligands as D3 Receptor (D3R) Selective Agonists. *J. Med. Chem.* 62 (13), 6287–6314. doi:10.1021/acs.jmedchem.9b00702
- Battiti, F. O., Newman, A. H., and Bonifazi, A. (2020). Exception that Proves the Rule: Investigation of Privileged Stereochemistry in Designing Dopamine D3R Bitopic Agonists. *ACS Med. Chem. Lett.* 11 (10), 1956–1964. doi:10.1021/acsmchemlett.9b00660
- Bekker, G. J., Araki, M., Oshima, K., Okuno, Y., and Kamiya, N. (2021). Accurate Binding Configuration Prediction of a G-Protein-Coupled Receptor to its Antagonist Using Multicanonical Molecular Dynamics-Based Dynamic Docking. *J. Chem. Inf. Model.* 61 (10), 5161–5171. doi:10.1021/acs.jcim.1c00712
- Bermudez, M., and Bock, A. (2019). Does Divergent Binding Pocket Closure Drive Ligand Bias for Class A GPCRs. *Trends Pharmacol. Sci.* 40 (4), 236–239. doi:10.1016/j.tips.2019.02.005
- Bethany, A. R., Huiqun, W., and Yan, Z. (2019). Recent Advances in the Drug Discovery and Development of Dualsteric/Bitopic Activators of G Protein-Coupled Receptors. *Curr. Top. Med. Chem.* 19 (26), 2378–2392. doi:10.2174/1568026619666191009164609
- Bian, Y., Jun, J. J., Cuyler, J., and Xie, X. Q. (2020). Covalent Allosteric Modulation: An Emerging Strategy for GPCRs Drug Discovery. *Eur. J. Med. Chem.* 206, 112690. doi:10.1016/j.ejmech.2020.112690
- Bock, A., Schrage, R., and Mohr, K. (2018). Allosteric Modulators Targeting CNS Muscarinic Receptors. *Neuropharmacology* 136 (Pt C), 427–437. doi:10.1016/j.neuropharm.2017.09.024
- Bosma, R., Stoddart, L. A., Georgi, V., Bouzo-Lorenzo, M., Bushby, N., Inkoom, L., et al. (2019). Probe Dependency in the Determination of Ligand Binding Kinetics at a Prototypical G Protein-Coupled Receptor. *Sci. Rep.* 9 (1), 7906. doi:10.1038/s41598-019-44025-5
- Brenke, R., Kozakov, D., Chuang, G. Y., Beglov, D., Hall, D., Landon, M. R., et al. (2009). Fragment-based Identification of Druggable 'hot Spots' of Proteins Using Fourier Domain Correlation Techniques. *BIOINFORMATICS* 25 (5), 621–627. doi:10.1093/bioinformatics/btp036
- British Pharmacological Society (2021). Dopamine Receptors – IUPHAR Review 13 - Beaulieu - 2015 - British Journal of Pharmacology - Wiley Online Library. Available at: <https://bpspubs.onlinelibrary.wiley.com/doi/10.1111/bph.12906> (accessed 12 2021, 19).
- Burris, K. D., Molski, T. F., Xu, C., Ryan, E., Tottori, K., Kikuchi, T., et al. (2002). Aripiprazole, a Novel Antipsychotic, Is a High-Affinity Partial Agonist at Human Dopamine D2 Receptors. *J. Pharmacol. Exp. Ther.* 302 (1), 381–389. doi:10.1124/jpet.102.033175
- Cao, Y., Sun, N., Zhang, J., Liu, Z., Tang, Y. Z., Wu, Z., et al. (2018). Design, Synthesis, and Evaluation of Bitopic Arylpiperazine-Phthalimides as Selective Dopamine D3 Receptor Agonists. *MedChemComm* 9 (9), 1457–1465. doi:10.1039/C8MD00237A
- Chan, H. C. S., Li, Y., Dahoun, T., Vogel, H., and Yuan, S. (2019). New Binding Sites, New Opportunities for GPCR Drug Discovery. *Trends Biochem. Sci.* 44 (4), 312–330. doi:10.1016/j.tibs.2018.11.011
- Chatzigeorgoulas, A., and Cournia, Z. (2021). Rational Design of Allosteric Modulators: Challenges and Successes. *Wires Comput. Mol. Sci.* 11 (6), e1529. doi:10.1002/wcms.1529
- Che, T., and Roth, B. L. (2021). Structural Insights Accelerate the Discovery of Opioid Alternatives. *Annu. Rev. Biochem.* 90, 739–761. doi:10.1146/annurev-biochem-061620-044044
- Chen, Z., Fan, L., Wang, H., Yu, J., Lu, D., Qi, J., et al. (2021). Structure-Based Design of a Novel Third-Generation Antipsychotic Drug Lead with Potential Antidepressant Properties. *Nat. Neurosci.* 25, 39–49. doi:10.1038/s41593-021-00971-w
- Cheng, J., Giguère, P. M., Onajole, O. K., Lv, W., Gaisin, A., Gunosewoyo, H., et al. (2015). Optimization of 2-Phenylcyclopropylmethylamines as Selective Serotonin 2C Receptor Agonists and Their Evaluation as Potential Antipsychotic Agents. *J. Med. Chem.* 58 (4), 1992–2002. doi:10.1021/jm5019274
- Cheng, J., Giguère, P. M., Schmerberg, C. M., Pogorelov, V. M., Rodriguez, R. M., Huang, X. P., et al. (2016a). Further Advances in Optimizing (2-Phenylcyclopropyl)methylamines as Novel Serotonin 2C Agonists: Effects on Hyperlocomotion, Prepulse Inhibition, and Cognition Models. *J. Med. Chem.* 59 (2), 578–591. doi:10.1021/acs.jmedchem.5b01153
- Cheng, J., McCorvy, J. D., Giguère, P. M., Zhu, H., Kenakin, T., Roth, B. L., et al. (2016b). Design and Discovery of Functionally Selective Serotonin 2C (5-HT_{2C}) Receptor Agonists. *J. Med. Chem.* 59 (21), 9866–9880. doi:10.1021/acs.jmedchem.6b01194
- Chevillard, F., Stotani, S., Karawajczyk, A., Hristeva, S., Pardon, E., Steyaert, J., et al. (2019). Interrogating Dense Ligand Chemical Space with a Forward-Synthetic Library. *Proc. Natl. Acad. Sci. U. S. A.* 116 (23), 11496–11501. doi:10.1073/pnas.1818718116
- Chevillard, F., Kelemen, Á., Baker, J. G., Aranyodi, V. A., Balzer, F., Kolb, V., et al. (2021). Fragment Evolution for GPCRs: The Role of Secondary Binding Sites in Optimization. *Chem. Commun.* 57 (81), 10516–10519. doi:10.1039/D1CC04636E
- Chien, E. Y., Liu, W., Zhao, Q., Katritch, V., Han, G. W., Hanson, M. A., et al. (2010). Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* 330 (6007), 1091–1095. doi:10.1126/science.1197410
- Christopher, J. A., Aves, S. J., Bennett, K. A., Doré, A. S., Errey, J. C., Jazayeri, A., et al. (2015). Fragment and Structure-Based Drug Discovery for a Class C GPCR: Discovery of the MGLu5 Negative Allosteric Modulator HTL14242 (3-Chloro-5-[6-(5-Fluoropyridin-2-Yl)Pyrimidin-4-Yl]Benzonitrile). *J. Med. Chem.* 58 (16), 6653–6664. doi:10.1021/acs.jmedchem.5b00892
- Christopoulos, A. (2014). Advances in G Protein-Coupled Receptor Allostery: From Function to Structure. *Mol. Pharmacol.* 86 (5), 463–478. doi:10.1124/mol.114.094342
- Congreve, M., Oswald, C., and Marshall, F. H. (2017). Applying Structure-Based Drug Design Approaches to Allosteric Modulators of GPCRs. *Trends Pharmacol. Sci.* 38 (9), 837–847. doi:10.1016/j.tips.2017.05.010
- Congreve, M., Oswald, C., and Marshall, F. H. (2017). Applying Structure-Based Drug Design Approaches to Allosteric Modulators of GPCRs. *Trends Pharmacol. Sci.* 38, 837–847. doi:10.1016/j.tips.2017.05.010
- Conn, P. J., Jones, C. K., and Lindsley, C. W. (2009). Subtype-Selective Allosteric Modulators of Muscarinic Receptors for the Treatment of CNS Disorders. *Trends Pharmacol. Sci.* 30 (3), 148–155. doi:10.1016/j.tips.2008.12.002
- Cordeaux, Y., IJzerman, A. P., and Hill, S. J. (2004). Coupling of the Human A1 Adenosine Receptor to Different Heterotrimeric G Proteins: Evidence for Agonist-specific G Protein Activation. *Br. J. Pharmacol.* 143 (6), 705–714. doi:10.1038/sj.bjp.0705925
- Decherchi, S., and Cavalli, A. (2020). Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chem. Rev.* 120 (23), 12788–12833. doi:10.1021/acs.chemrev.0c00534
- Denzinger, K., Nguyen, T. N., Noonan, T., Wolber, G., and Bermudez, M. (2020). Biased Ligands Differentially Shape the Conformation of the Extracellular Loop Region in 5-HT_{2B} Receptors. *Ijms* 21 (24), 9728. doi:10.3390/ijms21249728
- Draper-Joyce, C. J., Michino, M., Verma, R. K., Klein Herenbrink, C., Shonberg, J., Kopinathan, A., et al. (2018). The Structural Determinants of the Bitopic Binding Mode of a Negative Allosteric Modulator of the Dopamine D2 Receptor. *Biochem. Pharmacol.* 148, 315–328. doi:10.1016/j.bcp.2018.01.002
- Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., et al. (2011). Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc. Natl. Acad. Sci. U. S. A.* 108 (32), 13118–13123. doi:10.1073/pnas.1104614108
- Egloff, P., Hillenbrand, M., Klenk, C., Batyuk, A., Heine, P., Balada, S., et al. (2014). Structure of Signaling-Competent Neurotensin Receptor 1 Obtained by Directed Evolution in Escherichia Coli. *Proc. Natl. Acad. Sci. U. S. A.* 111 (6), E655–E662. doi:10.1073/pnas.1317903111
- Egyed, A., Domány-Kovács, K., Koványi, B., Hórti, F., Kurkó, D., Kiss, D. J., et al. (2020). Controlling Receptor Function from the Extracellular Vestibule of G-Protein Coupled Receptors. *Chem. Commun. (Camb)* 56 (91), 14167–14170. doi:10.1039/D0CC005532H

- Egyed, A., Kelemen, Á. A., Vass, M., Visegrády, A., Thee, S. A., Wang, Z., et al. (2021). Controlling the Selectivity of Aminergic GPCR Ligands from the Extracellular Vestibule. *Bioorg. Chem.* 111, 104832. doi:10.1016/j.bioorg.2021.104832
- Felder, C. C., Joyce, K. E., Briley, E. M., Mansouri, J., Mackie, K., Blond, O., et al. (1995). Comparison of the Pharmacology and Signal Transduction of the Human Cannabinoid CB1 and CB2 Receptors. *Mol. Pharmacol.* 48 (3), 443–450.
- Ferruz, N., Doerr, S., Vanase-Frawley, M. A., Zou, Y., Chen, X., Marr, E. S., et al. (2018). Dopamine D3 Receptor Antagonist Reveals a Cryptic Pocket in Aminergic GPCRs. *Sci. Rep.* 8, 897. doi:10.1038/s41598-018-19345-7
- Frank, A., Kiss, D. J., Keserü, G. M., and Stark, H. (2018). Binding Kinetics of Cariprazine and Aripiprazole at the Dopamine D3 Receptor. *Sci. Rep.* 8, 12509. doi:10.1038/s41598-018-30794-y
- Frei, J. N., Broadhurst, R. W., Bostock, M. J., Solt, A., Jones, A. J. Y., Gabriel, F., et al. (2020). Conformational Plasticity of Ligand-Bound and Ternary GPCR Complexes Studied by 19F NMR of the β 1-adrenergic Receptor. *Nat. Commun.* 11 (1), 669. doi:10.1038/s41467-020-14526-3
- Fronik, P., Gaiser, B. I., and Sejer Pedersen, D. (2017). Bitopic Ligands and Metastable Binding Sites: Opportunities for G Protein-Coupled Receptor (GPCR) Medicinal Chemistry. *J. Med. Chem.* 60 (10), 4126–4134. doi:10.1021/acs.jmedchem.6b01601
- Gaiser, B. I., Danielsen, M., Marcher-Rørsted, E., Røpke Jørgensen, K., Wróbel, T. M., Frykman, M., et al. (2019). Probing the Existence of a Metastable Binding Site at the β 2-Adrenergic Receptor with Homobivalent Bitopic Ligands. *J. Med. Chem.* 62 (17), 7806–7839. doi:10.1021/acs.jmedchem.9b00595
- Gao, Y., Peterson, S., Masri, B., Houghland, M. T., Adham, N., Gyertyán, I., et al. (2014). Cariprazine Exerts Antimanic Properties and Interferes with Dopamine D2 Receptor β -arrestin Interactions. *Pharmacol. Res. Perspect.* 3 (1), e00073. doi:10.1002/prp2.73
- Gao, Z. G., Inoue, A., and Jacobson, K. A. (2018). On the G Protein-Coupling Selectivity of the Native A2B Adenosine Receptor. *Biochem. Pharmacol.* 151, 201–213. doi:10.1016/j.bcp.2017.12.003
- Ghanakota, P., and Carlson, H. A. (2016). Moving beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *J. Phys. Chem. B* 120 (33), 8685–8695. doi:10.1021/acs.jpcc.6b03515
- Goodford, P. J. (1985). A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* 28 (7), 849–857. doi:10.1021/jm00145a002
- Greener, J. G., Filippis, I., and Sternberg, M. J. E. (2017). Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints. *Structure* 25 (3), 546–558. doi:10.1016/j.str.2017.01.008
- Grundmann, M., Bender, E., Schamberger, J., and Eitner, F. (2021). Pharmacology of Free Fatty Acid Receptors and Their Allosteric Modulators. *Int. J. Mol. Sci.* 22 (4), 1763. doi:10.3390/ijms22041763
- Guixà-González, R., Albasanz, J. L., Rodríguez-Espigares, I., Pastor, M., Sanz, F., Martí-Solano, M., et al. (2017). Membrane Cholesterol Access into a G-Protein-Coupled Receptor. *Nat. Commun.* 8 (1), 14505. doi:10.1038/ncomms14505
- Guo, D., Pan, A. C., Dror, R. O., Mocking, T., Liu, R., Heitman, L. H., et al. (2016). Molecular Basis of Ligand Dissociation from the Adenosine A2A Receptor. *Mol. Pharmacol.* 89 (5), 485–491. doi:10.1124/mol.115.102657
- Haga, K., Kruse, A. C., Asada, H., Yurugi-Kobayashi, T., Shiroishi, M., Zhang, C., et al. (2012). Structure of the Human M2 Muscarinic Acetylcholine Receptor Bound to an Antagonist. *Nature* 482 (7386), 547–551. doi:10.1038/nature10753
- Halgren, T. A. (2009). Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* 49 (2), 377–389. doi:10.1021/ci800324m
- Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B., and Gloriam, D. E. (2017). Trends in GPCR Drug Discovery: New Agents, Targets and Indications. *Nat. Rev. Drug Discov.* 16 (12), 829–842. doi:10.1038/nrd.2017.178
- Hoare, S. R. J., Tewson, P. H., Quinn, A. M., Hughes, T. E., and Bridge, L. J. (2020). Analyzing Kinetic Signaling Data for G-Protein-Coupled Receptors. *Sci. Rep.* 10 (1), 12263. doi:10.1038/s41598-020-67844-3
- Holze, J., Bermudez, M., Pfeil, E. M., Kauk, M., Bödefeld, T., Irmen, M., et al. (2020). Ligand-Specific Allosteric Coupling Controls G-Protein-Coupled Receptor Signaling. *ACS Pharmacol. Transl. Sci.* 3 (5), 859–867. doi:10.1021/acspctsci.0c00069
- Howlett, A. C., Champion, T. M., Wilken, G. H., and Mechoulam, R. (1990). Stereochemical Effects of 11-OH-Delta 8-Tetrahydrocannabinol-Dimethylheptyl to Inhibit Adenylate Cyclase and Bind to the Cannabinoid Receptor. *Neuropharmacology* 29 (2), 161–165. doi:10.1016/0028-3908(90)90056-w
- Huang, B., and Schroeder, M. (2006). LIGSITEcsc: Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation. *BMC Struct. Biol.* 6, 19. doi:10.1186/1472-6807-6-19
- Huang, M., Song, K., Liu, X., Lu, S., Shen, Q., Wang, R., et al. (2018). AlloFinder: A Strategy for Allosteric Modulator Discovery and Allosterome Analyses. *Nucleic Acids Res.* 46, W451–W458. doi:10.1093/nar/gky374
- Huang, S., Xu, P., Tan, Y., You, C., Zhang, Y., Jiang, Y., et al. (2021). Structural Basis for Recognition of Anti-migraine Drug Lasmiditan by the Serotonin Receptor 5-HT1F-G Protein Complex. *Cell Res* 31 (9), 1036–1038. doi:10.1038/s41422-021-00527-4
- Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., et al. (2013). Allosite: A Method for Predicting Allosteric Sites. *Bioinformatics* 29 (18), 2357–2359. doi:10.1093/bioinformatics/btt399
- Huang, X. P., Karpiak, J., Kroeze, W. K., Zhu, H., Chen, X., Moy, S. S., et al. (2015). Allosteric Ligands for the Pharmacologically Dark Receptors GPR68 and GPR65. *Nature* 527 (7579), 477–483. doi:10.1038/nature15699
- Jakubik, J., and El-Fakahany, E. E. (2020). Current Advances in Allosteric Modulation of Muscarinic Receptors. *Biomolecules* 10 (2), 325. doi:10.3390/biom10020325
- Jakubik, J., and El-Fakahany, E. E. (2021). Allosteric Modulation of GPCRs of Class A by Cholesterol. *Ijms* 22 (4), 1953. doi:10.3390/ijms22041953
- James, T. (2018). “Cheminformatics in the Service of GPCR Drug Discovery,” in *Computational Methods for GPCR Drug Discovery*. Editor A. Heifetz (New York, NY: Springer), 395–411. *Methods in Molecular Biology*. doi:10.1007/978-1-4939-7465-8_20
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2
- Kampen, S., Duy Vo, D., Zhang, X., Panel, N., Yang, Y., Jaiteh, M., et al. (2021). Structure-Guided Design of G-Protein-Coupled Receptor Polypharmacology. *Angew. Chem. Int. Ed.* 60 (33), 18022–18030. doi:10.1002/anie.202101478
- Kapur, A., Zhao, P., Sharir, H., Bai, Y., Caron, M. G., Barak, L. S., et al. (2009). Atypical Responsiveness of the Orphan Receptor GPR55 to Cannabinoid Ligands. *J. Biol. Chem.* 284 (43), 29817–29827. doi:10.1074/jbc.M109.050187
- Keserü, G. M., Erlanson, D. A., Ferenczy, G. G., Hann, M. M., Murray, C. W., and Pickett, S. D. (2016). Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia. *J. Med. Chem.* 59, 8189–8206. doi:10.1021/acs.jmedchem.6b00197
- Kim, K., Che, T., Panova, O., DiBerto, J. F., Lyu, J., Krumm, B. E., et al. (2020). Structure of a Hallucinogen-Activated Gq-Coupled 5-HT2A Serotonin Receptor. *Cell* 182 (6), 1574–e19. e19. doi:10.1016/j.cell.2020.08.024
- Kiss, B., Horváth, A., Némethy, Z., Schmidt, E., Laszlovszky, I., Bugovics, G., et al. (2010). Cariprazine (RGH-188), a Dopamine D(3) Receptor-Preferring, D(3)/D(2) Dopamine Receptor Antagonist-Partial Agonist Antipsychotic Candidate: *In Vitro* and Neurochemical Profile. *J. Pharmacol. Exp. Ther.* 333 (1), 328–340. doi:10.1124/jpet.109.160432
- Kling, R. C., Burchardt, C., Einsiedel, J., Hübner, H., and Gmeiner, P. (2019). Structure-Based Exploration of an Allosteric Binding Pocket in the NTS1 Receptor Using Bitopic NT(8-13) Derivatives and Molecular Dynamics Simulations. *J. Mol. Model.* 25 (7), 193. doi:10.1007/s00894-019-4064-x
- Kokh, D. B., and Wade, R. C. (2021). G Protein-Coupled Receptor-Ligand Dissociation Rates and Mechanisms from rTAMD Simulations. *J. Chem. Theor. Comput.* 17 (10), 6610–6623. doi:10.1021/acs.jctc.1c00641
- Kooistra, A. J., Mordalski, S., Pándy-Szekeres, G., Esguerra, M., Mamyrbekov, A., Munk, C., et al. (2021). GPCRdb in 2021: Integrating GPCR Sequence, Structure and Function. *Nucleic Acids Res.* 49 (D1), D335–D343. doi:10.1093/nar/gkaa1080
- Korczynska, M., Clark, M. J., Valant, C., Xu, J., Moo, E. V., Albold, S., et al. (2018). Structure-Based Discovery of Selective Positive Allosteric Modulators of

- Antagonists for the M2 Muscarinic Acetylcholine Receptor. *Proc. Natl. Acad. Sci. U S A* 115 (10), E2419–E2428. doi:10.1073/pnas.1718037115
- Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat. Protoc.* 10 (5), 733–755. doi:10.1038/nprot.2015.043
- Kruse, A. C., Hu, J., Pan, A. C., Arlow, D. H., Rosenbaum, D. M., Rosemond, E., et al. (2012). Structure and Dynamics of the M3 Muscarinic Acetylcholine Receptor. *Nature* 482 (7386), 552–556. doi:10.1038/nature10867
- Kruse, A. C., Ring, A. M., Manglik, A., Hu, J., Hu, K., Eitel, K., et al. (2013). Activation and Allosteric Modulation of a Muscarinic Acetylcholine Receptor. *Nature* 504 (7478), 101–106. doi:10.1038/nature12735
- Kumar, V., Bonifazi, A., Ellenberger, M. P., Keck, T. M., Pommier, E., Rais, R., et al. (2016). Highly Selective Dopamine D3 Receptor (D3R) Antagonists and Partial Agonists Based on Eticlopride and the D3R Crystal Structure: New Leads for Opioid Dependence Treatment. *J. Med. Chem.* 59 (16), 7634–7650. doi:10.1021/acs.jmedchem.6b00860
- Lamim Ribeiro, J. M., Provasi, D., and Filizola, M. (2020). A Combination of Machine Learning and Infrequent Metadynamics to Efficiently Predict Kinetic Rates, Transition States, and Molecular Determinants of Drug Dissociation from G Protein-Coupled Receptors. *J. Chem. Phys.* 153 (12), 124105. doi:10.1063/5.0019100
- Lane, J. R., Sexton, P. M., and Christopoulos, A. (2013). Bridging the Gap: Bitopic Ligands of G-Protein-Coupled Receptors. *Trends Pharmacol. Sci.* 34 (1), 59–66. doi:10.1016/j.tips.2012.10.003
- Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* 10, 168. doi:10.1186/1471-2105-10-168
- Lee, B., Taylor, M., Griffin, S. A., McInnis, T., Sumien, N., Mach, R. H., et al. (2021). Evaluation of Substituted N-Phenylpiperazine Analogs as D3 vs. D2 Dopamine Receptor Subtype Selective Ligands. *Molecules* 26 (11), 3182. doi:10.3390/molecules26113182
- Lee, Y., Warne, T., Nehmé, R., Pandey, S., Dwivedi-Agnihotri, H., Chaturvedi, M., et al. (2020). Molecular Basis of β -arrestin Coupling to Formoterol-Bound β 1-adrenoceptor. *Nature* 583 (7818), 862–866. doi:10.1038/s41586-020-2419-1
- Leo, L. M., and Abood, M. E. (2021). CB1 Cannabinoid Receptor Signaling and Biased Signaling. *Molecules* 26 (17), 5413. doi:10.3390/molecules26175413
- Linden, J., Thai, T., Figler, H., Jin, X., and Robeva, A. S. (1999). Characterization of Human A(2B) Adenosine Receptors: Radioligand Binding, Western Blotting, and Coupling to G(q) in Human Embryonic Kidney 293 Cells and HMC-1 Mast Cells. *Mol. Pharmacol.* 56 (4), 705–713.
- Liu, J. J., Horst, R., Katritch, V., Stevens, R. C., and Wüthrich, K. (2012). Biased Signaling Pathways in β 2-adrenergic Receptor Characterized by 19F-NMR. *Science* 335 (6072), 1106–1110. doi:10.1126/science.1215802
- Liu, W., Wacker, D., Gati, C., Han, G. W., James, D., Wang, D., et al. (2013). Serial Femtosecond Crystallography of G Protein-Coupled Receptors. *Science* 342 (6165), 1521–1524. doi:10.1126/science.1244142
- Lu, S., and Zhang, J. (2017). Designed Covalent Allosteric Modulators: An Emerging Paradigm in Drug Discovery. *Drug Discov. Today* 22 (2), 447–453. doi:10.1016/j.drudis.2016.11.013
- Lu, Y., Liu, H., Yang, D., Zhong, L., Xin, Y., Zhao, S., et al. (2021). Affinity Mass Spectrometry-Based Fragment Screening Identified a New Negative Allosteric Modulator of the Adenosine A2A Receptor Targeting the Sodium Ion Pocket. *ACS Chem. Biol.* 16 (6), 991–1002. doi:10.1021/acscchembio.0c00899
- Lu, W., Kostic, M., Zhang, T., Che, J., Patricelli, M. P., Chouchani, E. T., et al. (2021). Fragment-Based Covalent Ligand Discovery. *RSC Chem. Biol.* 2 (2), 354–367. doi:10.1039/D0CB00222D
- Luderman, K. D., Conroy, J. L., Free, R. B., Southall, N., Ferrer, M., Sanchez-Soto, M., et al. (2018). Identification of Positive Allosteric Modulators of the D1 Dopamine Receptor that Act at Diverse Binding Sites. *Mol. Pharmacol.* 94 (4), 1197–1209. doi:10.1124/mol.118.113175
- Luderman, K. D., Jain, P., Benjamin Free, R., Conroy, J. L., Aubé, J., Sibley, D. R., et al. (2021). Development of Pyrimidone D1 Dopamine Receptor Positive Allosteric Modulators. *Bioorg. Med. Chem. Lett.* 31, 127696. doi:10.1016/j.bmcl.2020.127696
- Magarkar, A., Schnapp, G., Apel, A. K., Seeliger, D., and Tautermann, C. S. (2019). Enhancing Drug Residence Time by Shielding of Intra-protein Hydrogen Bonds: A Case Study on CCR2 Antagonists. *ACS Med. Chem. Lett.* 10 (3), 324–328. doi:10.1021/acsmchemlett.8b00590
- Mao, Q., Qin, W. Z., Zhang, A., and Ye, N. (2020). Recent Advances in Dopaminergic Strategies for the Treatment of Parkinson's Disease. *Acta Pharmacol. Sin.* 41 (4), 471–482. doi:10.1038/s41401-020-0365-y
- Masureel, M., Zou, Y., Picard, L. P., van der Westhuizen, E., Mahoney, J. P., Rodrigues, J. P. G. L. M., et al. (2018). Structural Insights into Binding Specificity, Efficacy and Bias of a β 2AR Partial Agonist. *Nat. Chem. Biol.* 14 (11), 1059–1066. doi:10.1038/s41589-018-0145-x
- McCorvy, J. D., Butler, K. V., Kelly, B., Rechsteiner, K., Karpiak, J., Betz, R. M., et al. (2018). Structure-inspired Design of β -arrestin-biased Ligands for Aminergic GPCRs. *Nat. Chem. Biol.* 14 (2), 126–134. doi:10.1038/nchembio.2527
- McNeill, S. M., Baltos, J. A., White, P. J., and May, L. T. (2021). Biased Agonism at Adenosine Receptors. *Cell. Signal.* 82, 109954. doi:10.1016/j.cellsig.2021.109954
- Melchiorre, C., Cassinelli, A., and Quaglia, W. (1987). Differential Blockade of Muscarinic Receptor Subtypes by Polymethylene Tetraamines. Novel Class of Selective Antagonists of Cardiac M-2 Muscarinic Receptors. *J. Med. Chem.* 30 (1), 201–204. doi:10.1021/jm00384a034
- Miao, Y., Bhattarai, A., Nguyen, A. T. N., Christopoulos, A., and May, L. T. (2018). Structural Basis for Binding of Allosteric Drug Leads in the Adenosine A1 Receptor. *Sci. Rep.* 8 (1), 16836. doi:10.1038/s41598-018-35266-x
- Miao, Y., Goldfeld, D. A., Moo, E. V., Sexton, P. M., Christopoulos, A., McCammon, J. A., et al. (2016). Accelerated Structure-Based Design of Chemically Diverse Allosteric Modulators of a Muscarinic G Protein-Coupled Receptor. *Proc. Natl. Acad. Sci. U S A* 113 (38), E5675–E5684. doi:10.1073/pnas.1612353113
- Michino, M., Boateng, C. A., Donthamsetti, P., Yano, H., Bakare, O. M., Bonifazi, A., et al. (2017). Toward Understanding the Structural Basis of Partial Agonism at the Dopamine D3 Receptor. *J. Med. Chem.* 60 (2), 580–593. doi:10.1021/acs.jmedchem.6b01148
- Mielnik, C. A., Lam, V. M., and Ross, R. A. (2021). CB1 Allosteric Modulators and Their Therapeutic Potential in CNS Disorders. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 106, 110163. doi:10.1016/j.pnpbp.2020.110163
- Nagiri, C., Kobayashi, K., Tomita, A., Kato, M., Kobayashi, K., Yamashita, K., et al. (2021). Cryo-EM Structure of the β 3-adrenergic Receptor Reveals the Molecular Basis of Subtype Selectivity. *Mol. Cell* 81 (15), 3205–e5. doi:10.1016/j.molcel.2021.06.024
- Newman, A. H., Battiti, F. O., and Bonifazi, A. (2016/2020). 2016 Philip S. Portoghesi Medicinal Chemistry Lectureship: Designing Bivalent or Bitopic Molecules for G-Protein Coupled Receptors. The Whole Is Greater Than the Sum of its Parts. *J. Med. Chem.* 63 (5), 1779–1797. doi:10.1021/acs.jmedchem.9b01105
- Newman, A. H., Beuming, T., Banala, A. K., Donthamsetti, P., Pongetti, K., LaBounty, A., et al. (2012). Molecular Determinants of Selectivity and Efficacy at the Dopamine D3 Receptor. *J. Med. Chem.* 55 (15), 6689–6699. doi:10.1021/jm300482h
- Orgován, Z., Ferenczy, G. G., and Keserű, G. M. (2019). Fragment-Based Approaches for Allosteric Metabotropic Glutamate Receptor (mGluR) Modulators. *Ctmc* 19 (19), 1768–1781. doi:10.2174/1568026619666190808150039
- Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How Many Drug Targets Are There. *Nat. Rev. Drug Discov.* 5 (12), 993–996. doi:10.1038/nrd2199
- Pedersen, M. F., Wróbel, T. M., Märcher-Rørsted, E., Pedersen, D. S., Möller, T. C., Gabriele, F., et al. (2020). Biased Agonism of Clinically Approved μ -opioid Receptor Agonists and TRV130 Is Not Controlled by Binding and Signaling Kinetics. *Neuropharmacology* 166, 107718. doi:10.1016/j.neuropharm.2019.107718
- Peng, Y., McCorvy, J. D., Harpsøe, K., Lansu, K., Yuan, S., Popov, P., et al. (2018). 5-HT2C Receptor Structures Reveal the Structural Basis of GPCR Polypharmacology. *Cell* 172 (4), 719–e14. doi:10.1016/j.cell.2018.01.001
- Potterton, A., Hussein, F. S., Southey, M. W. Y., Bodkin, M. J., Heifetz, A., Coveney, P. V., et al. (2019). Ensemble-Based Steered Molecular Dynamics Predicts Relative Residence Time of A2A Receptor Binders. *J. Chem. Theor. Comput.* 15 (5), 3316–3330. doi:10.1021/acs.jctc.8b01270
- Potterton, A., Heifetz, A., and Townsend-Nicholson, A. (2022). Predicting Residence of GPCR Ligands with Machine Learning. *Methods Mol. Biol. Clifton NJ* 2390, 191–205. doi:10.1007/978-1-0716-1787-8_8

- Prokop, S., Ábrányi-Balogh, P., Barti, B., Vámosi, M., Zöldi, M., Barna, L., et al. (2021). PharmacOSTORM Nanoscale Pharmacology Reveals Cariprazine Binding on Islands of Calleja Granule Cells. *Nat. Commun.* 12 (1), 6505. doi:10.1038/s41467-021-26757-z
- Rafael, F., Josema, C., and Enric, I. C. (2020). The Kinetic Component in Drug Discovery: Using the Most Basic Pharmacological Concepts to Advance in Selecting Drugs to Combat CNS Diseases. *Curr. Neuropharmacol.* 18 (3), 250–257. doi:10.2174/1570159X17666191001144309
- Raschka, S., and Kaufman, B. (2020). Machine Learning and AI-Based Approaches for Bioactive Ligand Discovery and GPCR-Ligand Recognition. *Methods* 180, 89–110. doi:10.1016/j.ymeth.2020.06.016
- Riddy, D. M., Cook, A. E., Shackelford, D. M., Pierce, T. L., Mocaer, E., Mannoury la Cour, C., et al. (2019). Drug-Receptor Kinetics and Sigma-1 Receptor Affinity Differentiate Clinically Evaluated Histamine H3 Receptor Antagonists. *Neuropharmacology* 144, 244–255. doi:10.1016/j.neuropharm.2018.10.028
- Roth, B. L., Sheffler, D. J., and Kroeze, W. K. (2004). Magic Shotguns versus Magic Bullets: Selectively Non-selective Drugs for Mood Disorders and Schizophrenia. *Nat. Rev. Drug Discov.* 3 (4), 353–359. doi:10.1038/nrd1346
- Salmaso, V., and Jacobson, K. A. (2020). In Silico Drug Design for Purinergic GPCRs: Overview on Molecular Dynamics Applied to Adenosine and P2Y Receptors. *Biomolecules* 10 (6), 812. doi:10.3390/biom10060812
- Sarkar, P., and Chattopadhyay, A. (2020). Cholesterol Interaction Motifs in G Protein-Coupled Receptors: Slippery Hot Spots. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 12 (4), e1481. doi:10.1002/wsbm.1481
- Schmidtke, P., Le Guilloux, V., Maupetit, J., and Tufféry, P. (2010). Fpocket: Online Tools for Protein Ensemble Pocket Detection and Tracking. *Nucleic Acids Res.* 38, W582–W589. doi:10.1093/nar/gkq383
- Schramm, S., Agnetta, L., Bermudez, M., Gerwe, H., Irmen, M., Holze, J., et al. (2019). Novel BQCA- and TBPB-Derived M1 Receptor Hybrid Ligands: Orthosteric Carbachol Differentially Regulates Partial Agonism. *ChemMedChem* 14 (14), 1349–1358. doi:10.1002/cmdc.201900283
- Shaik, A. B., Boateng, C. A., Battiti, F. O., Bonifazi, A., Cao, J., Chen, L., et al. (2021). Structure Activity Relationships for a Series of Eticlopride-Based Dopamine D2/D3 Receptor Bitopic Ligands. *J. Med. Chem.* 64 (20), 15313–15333. doi:10.1021/acs.jmedchem.1c01353
- Shimamura, T., Shiroishi, M., Weyand, S., Tsujimoto, H., Winter, G., Katritch, V., et al. (2011). Structure of the Human Histamine H1 Receptor Complex with Doxepin. *Nature* 475 (7354), 65–70. doi:10.1038/nature10236
- Sriram, K., and Insel, P. A. (2018). G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs. *Mol. Pharmacol.* 93 (4), 251–258. doi:10.1124/mol.117.111062
- Staus, D. P., Hu, H., Robertson, M. J., Kleinhenz, A. L. W., Wingler, L. M., Capel, W. D., et al. (2020). Structure of the M2 Muscarinic Receptor-β-Arrestin Complex in a Lipid Nanodisc. *Nature* 579 (7798), 297–302. doi:10.1038/s41586-020-1954-0
- Su, M., Zhu, L., Zhang, Y., Paknejad, N., Dey, R., Huang, J., et al. (2020). Structural Basis of the Activation of Heterotrimeric Gs-Protein by Isoproterenol-Bound β1-Adrenergic Receptor. *Mol. Cell* 80 (1), 59–e4. doi:10.1016/j.molcel.2020.08.001
- Sykes, D. A., Stoddart, L. A., Kilpatrick, L. E., and Hill, S. J. (2019). Binding Kinetics of Ligands Acting at GPCRs. *Mol. Cell. Endocrinol.* 485, 9–19. doi:10.1016/j.mce.2019.01.018
- Szlenk, C. T., Gc, J. B., and Natesan, S. (2019). Does the Lipid Bilayer Orchestrate Access and Binding of Ligands to Transmembrane Orthosteric/Allosteric Sites of G Protein-Coupled Receptors. *Mol. Pharmacol.* 96 (5), 527–541. doi:10.1124/mol.118.115113
- Tan, L., Zhou, Q., Yan, W., Sun, J., Kozikowski, A. P., Zhao, S., et al. (2020). Design and Synthesis of Bitopic 2-Phenylcyclopropylmethylamine (PCPMA) Derivatives as Selective Dopamine D3 Receptor Ligands. *J. Med. Chem.* 63 (9), 4579–4602. doi:10.1021/acs.jmedchem.9b01835
- Tautermann, C. S., Kiechle, T., Seeliger, D., Diehl, S., Wex, E., Banholzer, R., et al. (2013). Molecular Basis for the Long Duration of Action and Kinetic Selectivity of Tiotropium for the Muscarinic M3 Receptor. *J. Med. Chem.* 56 (21), 8746–8756. doi:10.1021/jm401219y
- Thal, D. M., Sun, B., Feng, D., Nawaratne, V., Leach, K., Felder, C. C., et al. (2016). Crystal Structures of the M1 and M4 Muscarinic Acetylcholine Receptors. *Nature* 531 (7594), 335–340. doi:10.1038/nature17188
- Thomas, T., Fang, Y., Yuriev, E., and Chalmers, D. K. (2016). Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D2 and D3 Receptors. *J. Chem. Inf. Model.* 56 (2), 308–321. doi:10.1021/acs.jcim.5b00457
- Valant, C., Robert Lane, J., Sexton, P. M., and Christopoulos, A. (2012). The Best of Both Worlds? Bitopic Orthosteric/allosteric Ligands of G Protein-Coupled Receptors. *Annu. Rev. Pharmacol. Toxicol.* 52 (1), 153–178. doi:10.1146/annurev-pharmtox-010611-134514
- van der Velden, W. J. C., Heitman, L. H., and Rosenkilde, M. M. (2020). Perspective: Implications of Ligand-Receptor Binding Kinetics for Therapeutic Targeting of G Protein-Coupled Receptors. *ACS Pharmacol. Transl. Sci.* 3 (2), 179–189. doi:10.1021/acspstsci.0c00012
- Vass, M., Agai-Csöngör, E., Hórti, F., and Keserű, G. M. (2014). Multiple Fragment Docking and Linking in Primary and Secondary Pockets of Dopamine Receptors. *ACS Med. Chem. Lett.* 5 (9), 1010–1014. doi:10.1021/ml500201u
- Verma, R. K., Abramyan, A. M., Michino, M., Free, R. B., Sibley, D. R., Javitch, J. A., et al. (2018). The E2.65A Mutation Disrupts Dynamic Binding Poses of SB269652 at the Dopamine D2 and D3 Receptors. *PLOS Comput. Biol.* 14 (1), e1005948. doi:10.1371/journal.pcbi.1005948
- Vuckovic, Z., Gentry, P. R., Berizzi, A. E., Hirata, K., Varghese, S., Thompson, G., et al. (2019). Crystal Structure of the M5 Muscarinic Acetylcholine Receptor. *Proc. Natl. Acad. Sci. U. S. A.* 116 (51), 26001–26007. doi:10.1073/pnas.1914446116
- Wacker, D., Wang, C., Katritch, V., Han, G. W., Huang, X. P., Vardy, E., et al. (2013). Structural Features for Functional Selectivity at Serotonin Receptors. *Science* 340 (6132), 615–619. doi:10.1126/science.1232808
- Wacker, D., Wang, S., McCorry, J. D., Betz, R. M., Venkatakrishnan, A. J., Levit, A., et al. (2017). Crystal Structure of an LSD-Bound Human Serotonin Receptor. *Cell* 168 (3), 377–e12. doi:10.1016/j.cell.2016.12.033
- Wakefield, A. E., Mason, J. S., Vajda, S., and Keserű, G. M. (2019). Analysis of Tractable Allosteric Sites in G Protein-Coupled Receptors. *Sci. Rep.* 9 (1), 6180. doi:10.1038/s41598-019-42618-8
- Wang, C., Jiang, Y., Ma, J., Wu, H., Wacker, D., Katritch, V., et al. (2013). Structural Basis for Molecular Recognition at Serotonin Receptors. *Science* 340 (6132), 610–614. doi:10.1126/science.1232807
- Wang, H., Reinecke, B. A., and Zhang, Y. (2020). Computational Insights into the Molecular Mechanisms of Differentiated Allosteric Modulation at the Mu Opioid Receptor by Structurally Similar Bitopic Modulators. *J. Comput. Aided Mol. Des.* 34 (8), 879–895. doi:10.1007/s10822-020-00309-x
- Wang, Y., Yu, Z., Xiao, W., Lu, S., and Zhang, J. (2021). Allosteric Binding Sites at the Receptor-Lipid Bilayer Interface: Novel Targets for GPCR Drug Discovery. *Drug Discov. Today* 26 (3), 690–703. doi:10.1016/j.drudis.2020.12.001
- White, J. F., Noinaj, N., Shibata, Y., Love, J., Kloss, B., Xu, F., et al. (2012). Structure of the Agonist-Bound Neurotensin Receptor. *Nature* 490 (7421), 508–513. doi:10.1038/nature11558
- Wold, E. A., Chen, J., Cunningham, K. A., and Zhou, J. (2019). Allosteric Modulation of Class A GPCRs: Targets, Agents, and Emerging Concepts. *J. Med. Chem.* 62 (1), 88–127. doi:10.1021/acs.jmedchem.8b00875
- Wu, Y., Zeng, L., and Zhao, S. (2021). Ligands of Adrenergic Receptors: A Structural Point of View. *Biomolecules* 11 (7), 936. doi:10.3390/biom11070936
- Xia, R., Wang, N., Xu, Z., Lu, Y., Song, J., Zhang, A., et al. (2021). Cryo-EM Structure of the Human Histamine H1 Receptor/Gq Complex. *Nat. Commun.* 12 (1), 2086. doi:10.1038/s41467-021-22427-2
- Xiao, P., Yan, W., Gou, L., Zhong, Y. N., Kong, L., Wu, C., et al. (2021). Ligand Recognition and Allosteric Regulation of DRD1-Gs Signaling Complexes. *Cell* 184 (4), 943–e18. doi:10.1016/j.cell.2021.01.028
- Xu, P., Huang, S., Zhang, H., Mao, C., Zhou, X. E., Cheng, X., et al. (2021a). Structural Insights into the Lipid and Ligand Regulation of Serotonin Receptors. *Nature* 592 (7854), 469–473. doi:10.1038/s41586-021-03376-8
- Xu, P., Huang, S., Mao, C., Krumm, B. E., Zhou, X. E., Tan, Y., et al. (2021b). Structures of the Human Dopamine D3 Receptor-Gi Complexes. *Mol. Cell* 81 (6), 1147–e4. doi:10.1016/j.molcel.2021.01.003
- Xu, X., Kaindl, J., Clark, M. J., Hübner, H., Hirata, K., Sunahara, R. K., et al. (2021). Binding Pathway Determines Norepinephrine Selectivity for the Human β1AR over β2AR. *Cel Res* 31 (5), 569–579. doi:10.1038/s41422-020-00424-2
- Yan, W., Fan, L., Yu, J., Liu, R., Wang, H., Tan, L., et al. (2021). 2-Phenylcyclopropylmethylamine Derivatives as Dopamine D2 Receptor Partial Agonists: Design, Synthesis, and Biological Evaluation. *J. Med. Chem.* 64 (23), 17239–17258. doi:10.1021/acs.jmedchem.1c01327

- Yang, D., Zhou, Q., Labroska, V., Qin, S., Darbalaei, S., Wu, Y., et al. (2021). G Protein-Coupled Receptors: Structure- and Function-Based Drug Discovery. *Sig Transduct Target. Ther.* 6 (1), 1–27. doi:10.1038/s41392-020-00435-w
- Yang, F., Ling, S., Zhou, Y., Zhang, Y., Lv, P., Liu, S., et al. (2021). Different Conformational Responses of the β 2-adrenergic Receptor-Gs Complex upon Binding of the Partial Agonist Salbutamol or the Full Agonist Isoprenaline. *Natl. Sci. Rev.* 8 (9), nwaa284. doi:10.1093/nsr/nwaa284
- Yin, J., Chen, K. M., Clark, M. J., Hijazi, M., Kumari, P., Bai, X. C., et al. (2020). Structure of a D2 Dopamine Receptor-G-Protein Complex in a Lipid Membrane. *Nature* 584 (7819), 125–129. doi:10.1038/s41586-020-2379-5
- Yuan, D., Liu, Z., Kaindl, J., Maeda, S., Zhao, J., Sun, X., et al. (2020). Activation of the α 2B Adrenoceptor by the Sedative Sympatholytic Dexmedetomidine. *Nat. Chem. Biol.* 16 (5), 507–512. doi:10.1038/s41589-020-0492-2
- Zhang, G., Cheng, J., McCorvy, J. D., Lorello, P. J., Caldarone, B. J., Roth, B. L., et al. (2017). Discovery of N-Substituted (2-Phenylcyclopropyl)Methylamines as Functionally Selective Serotonin 2C Receptor Agonists for Potential Use as Antipsychotic Medications. *J. Med. Chem.* 60 (14), 6273–6288. doi:10.1021/acs.jmedchem.7b00584
- Zhang, Y., Yang, F., Ling, S., Lv, P., Zhou, Y., Fang, W., et al. (2020). Single-particle Cryo-EM Structural Studies of the β 2AR-Gs Complex Bound with a Full Agonist Formoterol. *Cell Discov* 6 (1), 45–5. doi:10.1038/s41421-020-0176-9
- Zhuang, Y., Xu, P., Mao, C., Wang, L., Krumm, B., Zhou, X. E., et al. (2021a). Structural Insights into the Human D1 and D2 Dopamine Receptor Signaling Complexes. *Cell* 184 (4), 931–e18. e18. doi:10.1016/j.cell.2021.01.027
- Zhuang, Y., Krumm, B., Zhang, H., Zhou, X. E., Wang, Y., Huang, X.-P., et al. (2021b). Mechanism of Dopamine Binding and Allosteric Modulation of the Human D1 Dopamine Receptor. *Cel Res* 31 (5), 593–596. doi:10.1038/s41422-021-00482-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Egyed, Kiss and Keserű. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Limits of Prediction for Machine Learning in Drug Discovery

Modest von Korff* and Thomas Sander

Idorsia Pharmaceuticals Ltd., Allschwil, Switzerland

In drug discovery, molecules are optimized towards desired properties. In this context, machine learning is used for extrapolation in drug discovery projects. The limits of extrapolation for regression models are known. However, a systematic analysis of the effectiveness of extrapolation in drug discovery has not yet been performed. In response, this study examined the capabilities of six machine learning algorithms to extrapolate from 243 datasets. The response values calculated from the molecules in the datasets were molecular weight, cLogP, and the number of sp³-atoms. Three experimental set ups were chosen for response values. Shuffled data were used for interpolation, whereas data for extrapolation were sorted from high to low values, and the reverse. Extrapolation with sorted data resulted in much larger prediction errors than extrapolation with shuffled data. Additionally, this study demonstrated that linear machine learning methods are preferable for extrapolation.

OPEN ACCESS

Edited by:

Leonardo L. G. Ferreira,
University of São Paulo, Brazil

Reviewed by:

Victor Kuz'min,
National Academy of Sciences of
Ukraine (NAN Ukraine), Ukraine
Muthukumarasamy Karthikeyan,
National Chemical Laboratory (CSIR),
India

*Correspondence:

Modest von Korff
modest.korff@idorsia.com

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 09 December 2021

Accepted: 10 February 2022

Published: 10 March 2022

Citation:

von Korff M and Sander T (2022) Limits
of Prediction for Machine Learning in
Drug Discovery.
Front. Pharmacol. 13:832120.
doi: 10.3389/fphar.2022.832120

Keywords: machine learning, drug discovery, extrapolation, data set, PLS (partial least square), Gaussian regression, random forest, support vector regression

INTRODUCTION

In drug discovery, new molecules undergo clinical trials in human subjects only after numerous checks for safety and potency in biological test systems. Often, a drug suitable for oral administration is desired, i.e., a molecule that can cross cellular membranes separating the gastrointestinal system and blood vessels. After absorption, blood vessels distribute the molecule throughout the organism and to its site of action. Blood contains many proteins that bind a substantial fraction of any compound. During distribution, molecules pass through the liver, which contains enzymes able to metabolize many types of chemical substances, thus reducing the concentration of the active drug (clearance). An important measure used in the optimization of a bioactive molecule is plasma exposure after oral administration, often expressed as “area under the curve” (AUC), i.e., the concentration of the active molecule in blood plasma integrated over time. Bioavailability depends on multiple properties of the molecule including cell layer permeability and clearance in the liver. When a molecule reaches the target protein, it must bind in such a way that it has the desired effect. A specific assay is usually developed to measure the effect of the molecule on the target protein. At present, it is still not possible to design a successful drug, fulfilling all necessary requirements, without biological tests. However, biological testing requires time and resources, which limit the number of compounds that can be explored. Medicinal chemists require quantitative models allowing prioritization the most promising molecules for biological testing.

Related Work

The use of quantitative structure-activity relationships (QSAR) is essential in drug discovery and has been investigated in multiple publications (Gramatica, 2007; Cherkasov et al., 2014). Recently, huge

efforts has been undertaken to find appropriate meta-parameter for QSAR models (Olier et al., 2018). It is well known that statistical models lose their predictive power when they are outside the range of calibration. Outside the calibration range, confidence intervals become infinite. These limits have been previously discussed for QSAR from (Tong et al., 2005) and were formulated in OECD policies for the validation of QSAR models (Member countries, 2004; OECD, 2014). Closely related to the calibration range is the term applicability domain. The term applicability domain is used in cheminformatics for quantitative structure activity models. The OECD guideline demands to consider the applicability domain but does not give a binding definition. By Roy et al. the application domain was defined as “The AD is a theoretical region in chemical space encompassing both the model descriptors and modeled response which allows one to estimate the uncertainty in the prediction of a particular compound based on how similar it is to the training compounds employed in the model development” (Roy et al., 2015). If the predicted molecules are similar to the training molecules in descriptor space, they are in the application domain (Jaworska et al., 2005). In drug discovery, the modeled response are molecular properties which the medicinal chemists aim to optimise. So, the properties medicinal chemists would like to predict are often outside the range of response values, which were already covered by experiments. At the start of a drug discovery project, a few molecules are usually identified which show modest activity at the target protein site. Starting compounds are modified by medicinal chemists to improve their properties. By adding all available information into the new compounds, they improve their characteristics over time. The next compound is often designed with the aim to show a lower binding constant to the target protein. Usually, this compound is similar to the already synthesized compounds and therefore in the applicability domain. During this optimization process, the desired response values are outside the range of the available response values. A model that aims to support the medicinal chemist in his work needs the capability of extrapolation. Recently, the use of extrapolation through machine learning, to assess the bioactivity of a molecule in drug discovery, has been evaluated (Cortés-Ciriano et al., 2018). Extrapolation outside the upper limits of the measured value range is wanted for the plasma exposure after oral administration. The plasma exposure should be as high as possible, but in a drug discovery project it is often too low. Additionally, frequently the majority of available response values are far away from the desired value range.

Our Work

The missing information in QSAR literature about differences between the errors of interpolation and extrapolation triggered a question. How effective can extrapolation of response values for chemical molecules be? To answer this question, we decided to use organic molecule datasets with calculated physicochemical properties. The physicochemical properties were used as response values in this study and were calculated from the molecular structure. Mathematically, a molecule is represented as a small graph with colored edges and colored nodes. This

molecular graph cannot directly be used as input for the applied machine learning methods. The graph must be transformed into a vector, a chemical descriptor. Machine learning creates models that relate descriptor vectors to the corresponding response values. With our setup a fully correct machine learning model was theoretically possible. The complete information needed to predict the response values was enclosed in the molecular structure. If this information is transferred to the descriptor vector and the machine learning algorithm constructs a perfect fitting model, a correct prediction will result. This model is “semi-mechanistic”, which is covered by the OECD guideline “When the AD is defined in more mechanistic terms, the (Q)SAR can predict reliably beyond the physicochemical and response space of the training set”. In our experimental setup the used response values allowed the machine learning algorithms to create such “semi-mechanistic” models.

METHODS

Datasets

For the construction of our molecule datasets, the size and structure of typical datasets in drug discovery was considered. In a drug discovery project, the molecules usually show a high similarity. New molecules are derived from a starting molecule that is explored by medicinal chemists. The newly synthesized molecules are similar to the starting molecule, but ideally have the desired features. We mimicked this process by taking a known drug molecule and removing randomly peripheral non hydrogen atoms. The removed atom was replaced with an appropriate number of hydrogen atoms. Rings were also randomly cut. Three top selling drug molecules were chosen: apixaban, rosuvastatin, and sofosbuvir (**Figure 1**). From each molecule, three sets, $S_{apix,1-3}$, $S_{rosu,1-3}$ and $S_{sofo,1-3}$, of about 300 molecules each were created. Consequently, nine datasets were constructed from three blockbuster drugs. Similar molecules are needed for successful machine learning models in QSAR (Netzeva et al., 2005). The similarity of test- and training molecules was guaranteed by our molecule degradation approach.

Dependent Variables

Dependent variables and response variables were calculated for each molecular structure. The simplest dependent variable in this study was molecular weight, which was calculated from the corresponding molecular formula. The logP value, the logarithm of the 1-octanol/water partition coefficient, is a more sophisticated variable which estimates the distribution of a drug based on an octanol/water system. The cLogP value assesses the permeation of a molecule from the gastrointestinal tract into blood vessels, and it is an important measure in drug discovery. Here, a fragmental approach from DataWarrior (Sander et al., 2015) was used to calculate the cLogP. This fragmental approach was developed for the OSIRIS Property Explorer (OsirisP) and successfully benchmarked in a large study with 90,000 compounds (Mannhold et al., 2009). In this independent examination, OsirisP ranked between the top logP calculation methods. An improved version of the Osiris logP calculator was implemented in DataWarrior in 2014. This

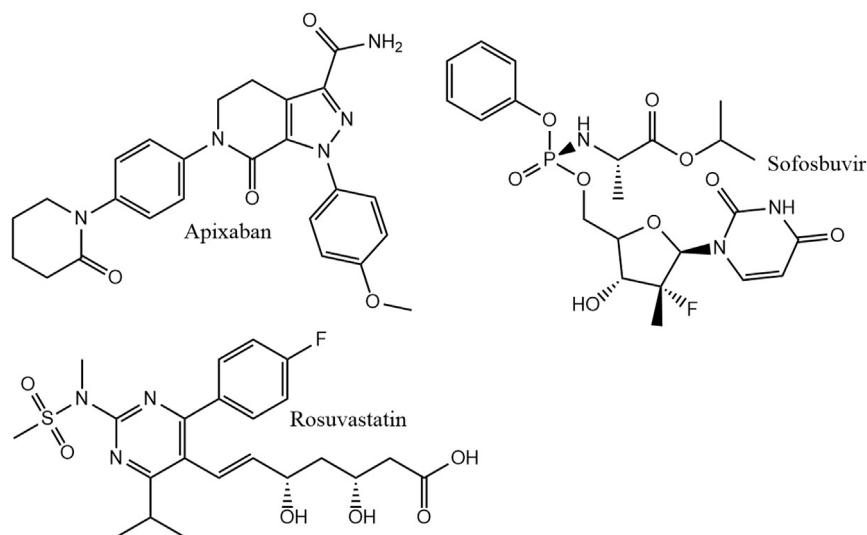


FIGURE 1 | Seed molecules for dataset generation.

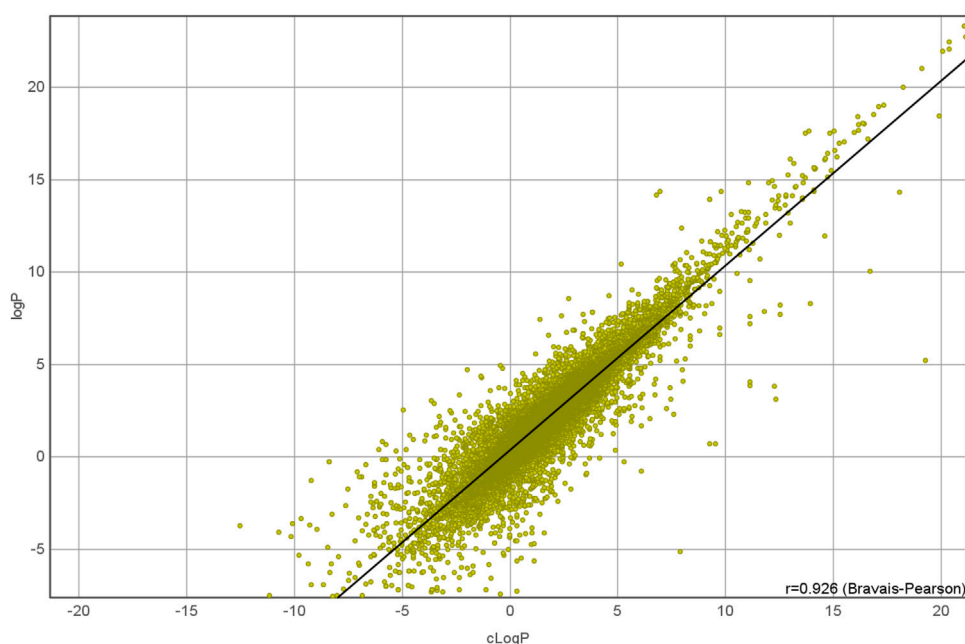


FIGURE 2 | Comparison of 25,000 experimental logP values with DataWarrior calculated logP.

updated OsirisP calculation is implemented as increment system adding contributions of every atom based on its atom type. OsirisP distinguishes around 400 atom types. This includes hybridisation state, ring membership, aromaticity, and additionally to the older version charges. More than 5,000 compounds with experimentally determined logP values were used as training set to calculate the increments. A recent comparison with 25,000 experimental logP values is given in **Figure 2**. The strong relation between the experimental and the calculated logP is shown by a correlation coefficient of 0.93.

However, this strong correlation is not needed for our experimental setup. Important for the experiment is the linear dependency between the molecular structure and the calculated logP values. Theoretically, this linear dependency allows linear regression methods like partial least square regression a perfect fit of dependent and independent variables.

A third response variable was the number of sp³-carbon atoms in a molecule, where each sp³-carbon atom has 4 neighboring atoms. In early drug discovery, the number of sp³-carbon atoms is used to choose molecules for high throughput screening in

TABLE 1 | Summary of the response values for all datasets. The first column indicates the property and the other columns the minimum, maximum, average, standard deviation, and median values.

| | min | max | avr | sdv | median |
|--------------|------|-----|-----|-----|--------|
| Apixaban | | | | | |
| MW | 110 | 434 | 288 | 101 | 294 |
| sp3-atoms | 2 | 19 | 10 | 3 | 10 |
| cLogP | -2.5 | 6.8 | 2.1 | 1.6 | 2.2 |
| Rosuvastatin | | | | | |
| MW | 60 | 453 | 255 | 115 | 256 |
| sp3-atoms | 0 | 17 | 8 | 4 | 8 |
| cLogP | -1.0 | 6.0 | 2.6 | 1.3 | 2.6 |
| Sofosbuvir | | | | | |
| MW | 110 | 500 | 317 | 108 | 317 |
| sp3-atoms | 5 | 24 | 16 | 4 | 16 |
| cLogP | -4.8 | 4.8 | 0.8 | 1.2 | 0.9 |

biological assays. For every molecule in the nine datasets $S_{apix,1-3}$, $S_{rosu,1-3}$ and $S_{sofo,1-3}$, the three dependent variables were calculated. By considering independent and dependent variables, a set of 27 datasets was obtained. A summary of the obtained values is given in Table 1.

Descriptors

A molecular graph is inappropriate input for most machine learning algorithms. Molecular descriptors are used in cheminformatics to describe molecular structure in algebraic form (Todeschini and Consonni, 2008). For a descriptor, a molecular graph is usually converted into a vector, which is the input for machine learning. The transformation from a molecule into a vector is one directional and comes with a loss of information. The molecular structure can not be recovered from the vector. Different transformations result in different losses of information. For this reason, three different topological molecule descriptors were chosen.

Fragment Fingerprint Descriptor

The fragment fingerprint is a dictionary based descriptor with a length of 512 bits. Each bit represents a substructure fragment. The dictionary of 512 substructures was created by a computational procedure, which had been optimized to achieve two goals: 1) any of these fragments should occur frequently in organic molecule structures and 2) each fragment should be linearly independent with regard to their substructure-match-pattern in diverse organic compound sets. To generate a descriptor vector, the molecular structure is searched for any of the substructures in the dictionary. For any match, the corresponding bit of the vector is set to 1. Any molecular structure is represented by a binary vector of length 512. The fragment fingerprint descriptor belongs to the same class as the “MDL structure keys” (McGregor and Pallai, 1997), which have recently been shown to outperform 3D descriptors in virtual screening (Nettles et al., 2006).

Path Fingerprint Descriptor

The path fingerprint is a molecular graph path walking fingerprint descriptor. All distinguishable paths with up to 7

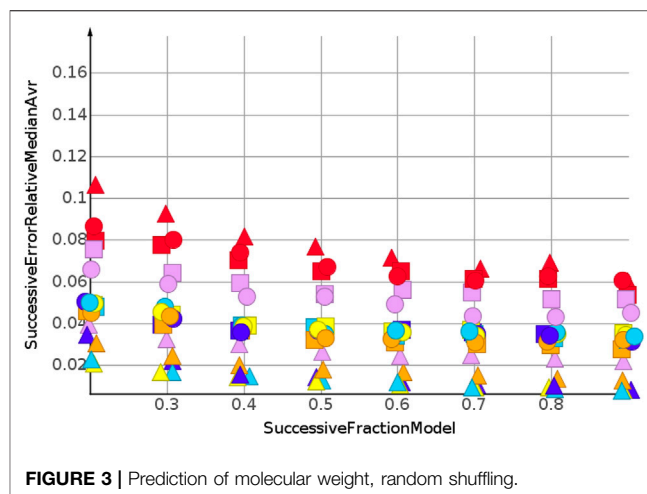


FIGURE 3 | Prediction of molecular weight, random shuffling.

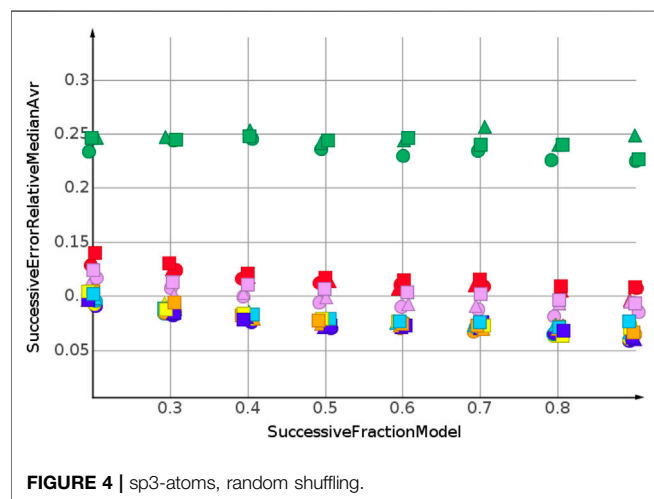
atoms are hashed into a descriptor vector of 512 bits. This descriptor is conceptually similar to ChemAxCFp, the chemical fingerprints from ChemAxon (ChemAxon, 1998) and to the Daylight fingerprints (Daylight, 1998).

Skeleton Spheres Descriptor

The skeleton spheres descriptor is a vector of integers which counts the occurrence of different substructures in a molecule. Five circular layers with increasing bond distance are located for each atom in the molecule. Hydrogen atoms are not considered. This results in four fragments starting with the naked central atom, adding one layer at a time. Every fragment is encoded as a canonical string (id-code), similar to the generation of canonical SMILES (Weininger et al., 1989). The canonical id-code includes the stereochemistry of the encoded fragment, which is a feature missing in other molecular descriptors. The id-code is then assigned to one of 1,024 fields in a vector. Therefore, the hash value of the id-code is calculated and the corresponding value in the vector is increased by one. The Hashlittle algorithm (Jenkins, 2006) is used as a binning function, which takes a text string as input and returns an integer value between 0 (inclusive) and 1,024 (exclusive). In preliminary experiments, this hash function showed a good uniform distribution of the generated hash values. To consider the molecular scaffold without the influence of the heteroatoms, the whole calculation is repeated while replacing the hetero atoms with carbon. The resulting hash values are used to increment the corresponding fields in the vector. By adding this skeleton information to the descriptor vector, the similarity calculation between two descriptor vectors becomes a bit insensitive to the exact position of the heteroatoms in two molecules. This directs the similarity value towards the perception of similarity by medicinal chemists. For medicinal chemists, the exact position of a hetero atom is not as discriminating as it would be for the spheres descriptor without the skeleton coding part. The additional consideration of the scaffold information and the use of a histogram instead of a binary vector distinguishes the skeleton spheres descriptor from other circular fingerprints. (Glem et al., 2006).

TABLE 2 | Prediction of molecular weight, random shuffling, skeleton spheres descriptor. The first column indicates the machine learning algorithm. The first row is the fraction of data used for model construction. The other values are the relative errors of the test data.

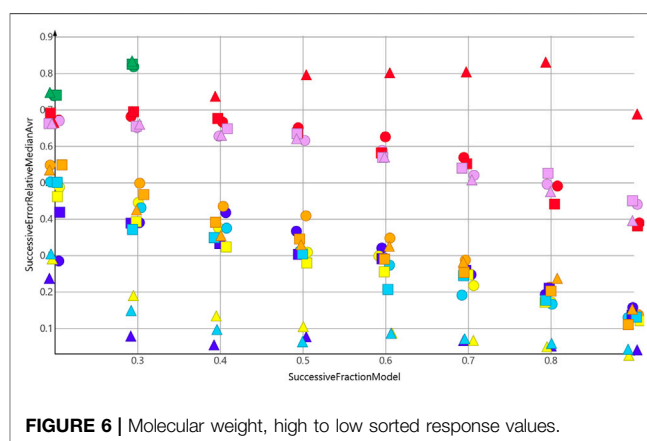
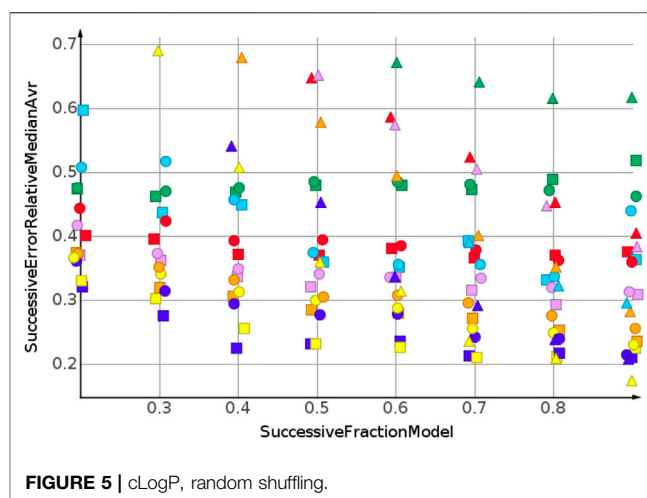
| | Fraction of train data | | | | | | |
|------|------------------------|------|------|------|------|------|------|
| | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 |
| GPR | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| Med | 0.29 | 0.30 | 0.30 | 0.29 | 0.29 | 0.29 | 0.28 |
| PLS | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| PLSP | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| RFR | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| SVM | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 |
| kNN | 0.09 | 0.08 | 0.08 | 0.07 | 0.06 | 0.07 | 0.06 |



Each of the nine molecule sets $S_{apix,1-3}$, $S_{rosu,1-3}$, and $S_{sofo,1-3}$ was compiled into three descriptor sets fragment fingerprint, path fingerprint and skeleton spheres.

Dataset Construction

A dataset D contains a matrix X and a vector y . Every row in the matrix X represents a molecule by one of the three descriptors, fragment fingerprint, path fingerprint, and skeleton spheres. Corresponding to a row i in X is a response value i in y . Three response values, molecular weight, cLogP and sp3-carbon, were available for each row in X . In drug discovery projects, the optimization process aims for response values outside the range of response values initially obtained. To assess the predictive power of a machine learning tool in a drug discovery project, we sorted the compounds by their response values. One dataset contained the ascending response values, a second the descending values and a third dataset was compiled from the shuffled response values. Summarizing the data set up, nine sets with molecules, each set compiled three descriptors, gave 27 descriptor matrices X . Three different response values, molecular weight, cLogP, and the number of sp3-carbon atoms were sorted according to ascending, descending or shuffled data. Combined with the 27 X matrices, a total of 243 datasets were obtained. The molecules



together with the descriptors and the calculated response values are available from (Korff, 2021). Each of these datasets underwent the successive regression procedure, as described in the next two paragraphs.

Machine Learning Techniques

Six modeling techniques were applied to construct regression models for the datasets: k next neighbor regression (kNN), partial least square regression (PLS), partial least square regression with power transformation (PLSP), random forest regression (RFR), Gaussian process regression (GPR), and support vector (SVM) regression. All parameters for these machine learning models were optimized by an exhaustive search. The median model was used as a baseline model. Any successful machine learning model should be significantly better than the baseline model. Also easy to calculate was the k next neighbor model for regression. In this model, the k next neighbors in the training set were screened for the query descriptor vector. The predicted \hat{y} value was the average of the corresponding y values weighted by similarity. Partial least square regression (PLSR) is a multivariate linear regression technique (De Jong, 1993), which only requires the number of factors as the input parameter. PLSR with power transformation includes a Box Cox transformation and is often used to model

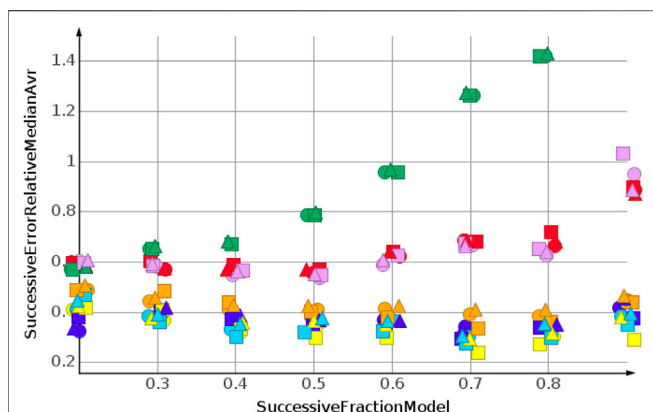


FIGURE 7 | Number of sp³-atoms, high to low sorted response values.

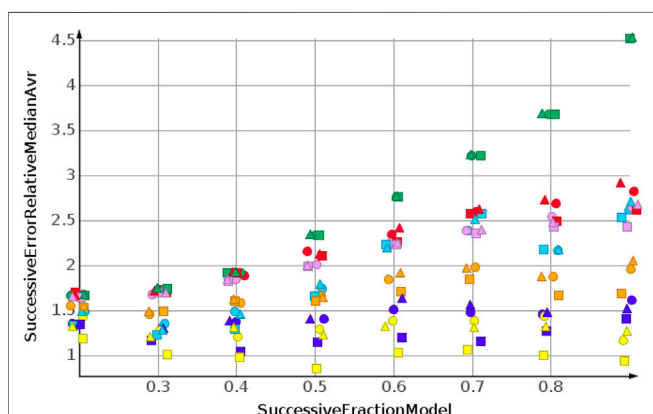


FIGURE 8 | cLogP, high to low sorted response values.

biological data, which are notoriously not normally distributed (Sakia, 1992). For random forests and Gaussian process regression, we used the implementation from Haifeng Li. (Li, 2021). Random forest regression was only included because it is frequently used for models in drug discovery. Random forests base on decision trees and are not capable of extrapolation. The Java program library libsvm was used for the support vector machine regression (Chang and Lin, 2011). Details for meta-parameter search: *k*NN: *k* from 1 to 9, step 1. PLSR: factors from 1 to 31, step 1. PLSR power transformation: factors like PLSR; λ from 0.05 to 2, step 0.05. Gaussian process regression: λ 0.001, 0.025 ... 1, ... 10,10,000. Random forest: trees 50, 100, 250, 500, 1,000; Maximum number leaf nodes from 2 to 54 step 2. Mtry: 0.15, 0.333, 0.45. Maximum node size from 2 to 54, step 2. Support vector regression: (Smola and Schölkopf, 2004) Epsilon regression, RBF kernel, power of 2 rule for: C from 2 to 5 to 215; ϵ from 2 to 10 to 26; γ 1/(number of fields in the descriptor). Details for the objective function are given in the next section.

Successive Regression

A two-step process was implemented to ensure an unbiased estimation for the extrapolation power of a model. The first step was the selection of one meta-parameter set for every machine

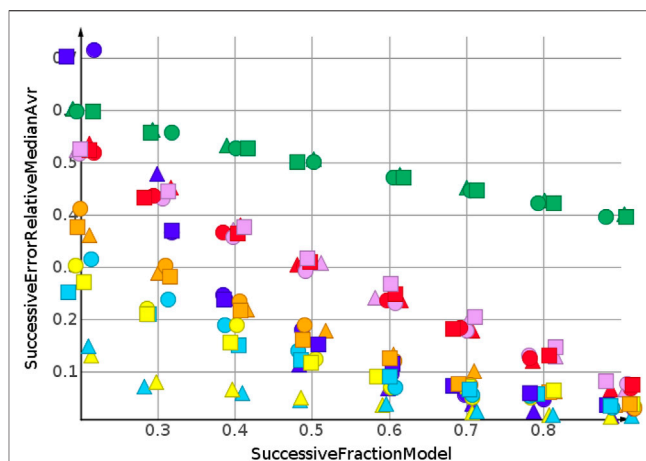


FIGURE 9 | Molecular weight, low to high sorted response values.

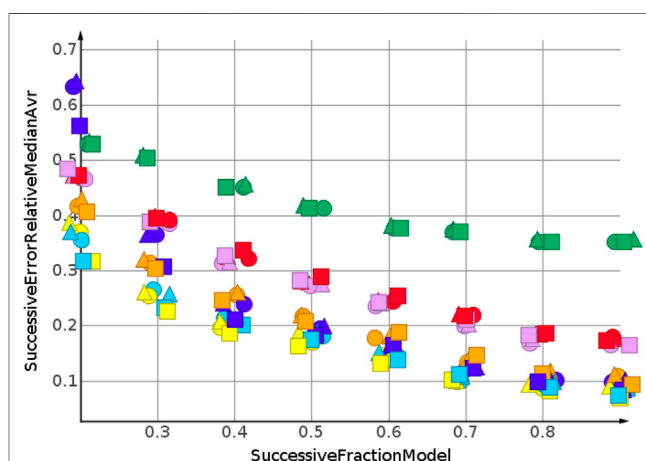


FIGURE 10 | Number of sp³-atoms, low to high sorted response values.

learning technique. The algorithm started with the first 20% of the molecule descriptors $X_{0,0.2}$, $y_{0,0.2}$ together with the measured response values to determine the meta parameters of the machine learning models *via* an exhaustive search. An eleven-fold Monte Carlo cross validation was employed to split all data into the training and validation datasets (Xu and Liang, 2001). A left out fraction of 25% was chosen as the size of the validation dataset. With this set up, the average error for all meta-parameter sets was calculated. For each machine learning technique *t*, the meta-parameter set $M_{\min,t}$ was chosen that showed the minimum average error. This meta-parameter set was used to construct a model from all data in $X_{0,0.2}$, $y_{0,0.2}$. In the second step, an independent test set was compiled from the next 10% of data, $X_{0.3}, y_{0.3}$. The average prediction error of $\hat{y}_{0.3}$ gave an unbiased estimator for the model, because the machine learning algorithm $M_{\min,t,0.2}$ had not seen these data before prediction. Subsequently, step one was repeated, this time with the dataset $X_{0.3}, y_{0.3}$. So, the former test data were added to $X_{0,0.2}, y_{0,0.2}$. The meta parameter for the machine learning algorithms $M_{\min,t,0.3}$ were now determined

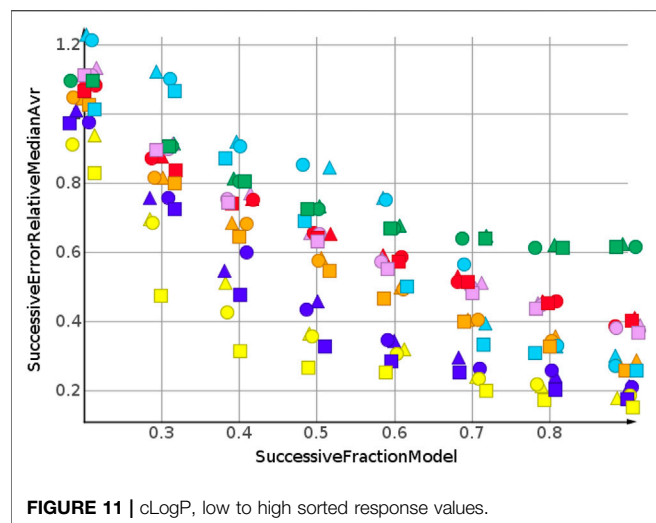


FIGURE 11 | cLogP, low to high sorted response values.

with $X_{0,0.3}$, $y_{0,0.3}$. So, the prediction was done for $y_{0.4}$. This process was repeated eight times, up to a model size with $X_{0,0.9}$, $y_{0,0.9}$ and a prediction for $y_{1.0}$. Using this method, we assessed the extrapolation power of the machine learning method together with the applied molecular descriptor for the sorted response data. The 10% test set, with higher or lower response values than the training set, was an unbiased estimator of the model's quality for extrapolation. As a quality measure for prediction, we used the relative error.

Technical Details

The source code was implemented in Java 1.8. The calculations were done on a SuperMicro computer with 176 processor cores. Meta-parameter calculation and test data prediction took approximately 72 h for all datasets. Data visualization was done with DataWarrior (Sander et al., 2015), an open source tool for data visualization and evaluation (Sander, 2021).

RESULTS

The successive regression procedure was applied to all 243 datasets. In the following, the results for nine datasets with the molecular structures $S_{apix,1-3}$, $S_{rosu,1-3}$ and $S_{sofo,1-3}$ are summarized by their median relative error. No extrapolation was needed for the shuffled datasets. **Figure 3** and **Table 2** show the machine learning results for the prediction for the shuffled data for three descriptors and three properties. The three descriptors, fragment fingerprint, path fingerprint and skeleton spheres, are indicated by shape. Circles, squares and triangles indicate fragment fingerprint, path fingerprint and skeleton spheres descriptors, respectively. A color code was used for the machine learning algorithms. Green indicated our base line model, which was the prediction by median, *k*NN regression in red, Gaussian process regression in blue, partial least square regression in yellow, partial least square regression with power transformation in light blue, random forest regression in magenta, and support vector regression in orange. All results are available as Data Warrior files (Korff, 2021).

TABLE 3 | Summary of the best results for the machine learning techniques. Rank count for the top three ranks. The ranks were calculated from all descriptors, predicted properties, and fractions of training data. The columns show the three different orientations of the response data: shuffled, sorted from high to low and from low to high.

| ML method | Shuffled | low2high | high2low |
|-----------------------------|----------|----------|----------|
| Gaussian process regression | 9 | 6 | 1 |
| KNN regression | 0 | 0 | 0 |
| Median | 0 | 0 | 0 |
| PLS | 8 | 13 | 18 |
| PLS Power | 3 | 6 | 6 |
| Random Forest regression | 0 | 0 | 0 |
| SVM regression | 7 | 2 | 2 |

TABLE 4 | Summary of the best results for the three descriptors. Rank count for the top three ranks. The ranks were calculated from all methods, predicted properties, and fractions of training data.

| Descriptor | Shuffled | low2high | high2low |
|-------------|----------|----------|----------|
| FragFp | 17 | 16 | 13 |
| PathFp | 16 | 22 | 26 |
| SkelSpheres | 30 | 25 | 24 |

TABLE 5 | Summary of the best results for the three response datasets. Rank count for the top three ranks. The ranks were calculated from all descriptors, methods, and fractions of training data.

| Response value | Shuffled | low2high | high2low |
|----------------|----------|----------|----------|
| MW | 54 | 47 | 49 |
| cLogP | 0 | 0 | 0 |
| sp3-Atoms | 9 | 16 | 14 |

For almost all models, the relative error for predicted molecular weight was less than 10%. For the majority of predictions, relative error was less than or equal 5%. No preference for any of the descriptors was observed, as indicated in **Figure 3**. A higher separation was shown by the machine learning techniques. The error for the median model is not shown in **Figure 3**. A relative error of approximately 30% was observed for all fractions of the model. Three machine learning models performed equally well. Gaussian process regression, partial least square regression and partial least square regression with power transformation showed a relative error below 3%. These results were obtained together with the skeleton spheres descriptor.

Figure 4 shows the results for the sp3-atoms with the shuffled data. The results were similar to the predicted molecular weight in 3. Relative error was higher than for the molecular weight prediction, but all models were better than the median model. In contrast to the molecular weight prediction, all three descriptors performed equally well for the models with the lowest error.

The predictions for cLogP, draw a different picture than the predictions for molecular weight and number of sp3-atoms, **Figure 5**. Only one model showed a relative error below 20%.

Many models were worse than the median model, indicated in green. The best performing machine learning models were partial least square regression and Gaussian process regression.

The prediction for shuffled data did not require extrapolation. The data range of the response values is covered by the training data. To simulate the requirements of drug discovery, the datasets were sorted by their response values. In the following, we discuss the results for sorting from high to low response values. This experimental set up forced the machine learning algorithms into extrapolation. The range of predicted response values was always outside the range of the training data. **Figure 6** shows the results for the prediction of molecular weight. The data were sorted from high to low.

As for the molecular weight prediction for shuffled data, the skeleton spheres descriptor together with partial least square regression, partial least square regression with power transformation and Gaussian process regression delivered the most predictive models. The range of the relative errors was very large, below 10% for the best models up to 50% for the *k*NN models with the skeleton spheres descriptor, depicted in red triangles. But, relative error was higher for all predictions than for the shuffled data. For the shuffled data, only one prediction was above a relative error of 0.1, with the *k*NN model at a fraction of 0.2. For the high to low sorted molecular weight data the majority of predictions showed a relative error above 0.1.

Two trends were observed for the prediction of number of sp³-atoms, **Figure 7**. The relative error of the median prediction increased with an increasing fraction of data used to construct the models. This also happened with the relative error for the *k*NN models, in red, and the random forest, in magenta. The relative error for Gaussian process regression, partial least squares, partial least squares with power transformation and support vector regression remained almost constant. As for molecular weight, the predictions for high to low sorted data had a much higher relative error than predictions for sorted data.

Figure 8 shows the results for cLogP. Data were sorted from high to low. Curve progression was similar to the curve progression of the relative error for the sp³-atom number prediction. However, the values for the relative error are much higher. Only four predictions had relative errors less than 100%.

When molecular weight values were sorted from low to high, the values to be predicted were higher than the values used for model construction. For the molecular weight prediction, the results are depicted in **Figure 9**. The skeleton spheres descriptor resulted in models with the lowest relative error. In the figures, the skeleton spheres descriptor is indicated by triangles. For the prediction of the number of sp³-atoms in **Figure 10** the models constructed from the path fingerprint were better than the models constructed from the skeleton spheres descriptor. As for the high to low sorted values in **Figure 7**, the path fingerprint was the best performing descriptor. Also, for cLogP value prediction, given in **Figure 11**, the path fingerprint was the best performing descriptor.

For each of the experimental set ups, including 243 individual datasets, all machine learning algorithms

outperformed median predictions, which were used as baseline controls. *k*NN regression and random forest regression were very similar in their prediction quality. These two algorithms were outperformed by support vector regression. The best performing machine learning algorithms were Gaussian process regression, partial least square regression and partial least square regression with power transformation. Together with the path fingerprint and the skeleton spheres descriptor, the best results were obtained. The relative errors for the successive predictions were lower for the low to high sorted values than for the high to low sorted values. This was caused by numeric effects, the absolute prediction error for big values results in a lower relative error than the same absolute prediction error for small values.

RESULTS SUMMARY AND CONCLUSION

All results are summarized in **Tables 3–5**. The figure of merit was the rank of the median error. For every successive fraction of test data, a median error was calculated from the nine molecule datasets $S_{apix,1-3}$, $S_{rosu,1-3}$ and $S_{sofo,1-3}$. By using the ranks of the errors, a bias was prevented, which would have been otherwise introduced by the error dependency on the fraction of training data. Because, a higher fraction of training data generally results in better models. This would have resulted in a bias if the median would have been used. By using the ranks the results for different fractions of training could be combined. In **Tables 3–5**, the frequency of the top three ranks is given. This means, the rank count increased by one, if the corresponding error belonged to the three lowest errors for the given conditions. Results for the machine learning algorithms are provided in **Table 3**. For shuffled response data, Gaussian process regression delivered the highest number of top models (9) for prediction. For extrapolation, for high to low sorted and for low to high sorted data, the partial least square regression outperformed the other machine learning algorithms. That the linear method outperformed the non-linear method is in accordance with the results from (Cortés-Ciriano et al., 2018), where the linear method, ridge regression, also outperformed the non-linear method, random forest.

Results for the descriptors are provided in **Table 4**. In total, the skeleton spheres descriptor outperformed the other two descriptors. However, the path fingerprint slightly outperformed the skeleton spheres descriptor for extrapolation for the high to low sorted response values. **Table 5** presents the rank counts for the most accurately predicted response values. As expected, the best models were obtained for molecular weight, followed by the number of sp³-atoms.

The purpose of this study was to examine the difference between prediction in the range of the training response values and extrapolation outside the training response values. It must be considered that the molecules in each dataset were derived from a single molecule. Consequently, there was a high similarity between molecules in a dataset. All molecules in this examination were in the domain of applicability. They were similar to the training molecules in descriptor space.

Nevertheless, the differences between the relative errors for the shuffled data and sorted data were striking. Even for molecular weight, with a very low error for shuffled data, the extrapolation for high to low sorted data became much more difficult. This was unexpected, because molecular weight depends solely on the molecular formula and does not need any molecular graph dependent feature. In addition, the relation between the molecular formula and molecular weight is strictly linear. cLogP values were hardest to predict. Prediction was achieved with a moderate error for shuffled data using linear regression techniques. However, after sorting the response values from high to low and successively extrapolating the lower values, no meaningful prediction for cLogP was possible. None of the machine learning algorithms were able to extrapolate cLogP values for high to low sorted data. This result was unexpected because the cLogP model is an incremental model that relies on substructure contributions to the overall cLogP. Therefore the contributions are linear and theoretically can be modelled by linear regression with chemical descriptors. We had expected, that the linear regression algorithms would be able to create “semi-mechanistic” models with more predictive power. There is a high demand in drug discovery for extrapolation of molecular features. The results of this study show large differences in prediction quality between interpolation and extrapolation. This demonstrates that any model used for extrapolation should be validated with extrapolation. For this validation, we suggest the successive prediction as described in this

contribution. We suggest to add the prediction of calculated values as reference standard to all publications in cheminformatics when regression methods are applied. Partial least square regression was by far the most successful extrapolation method. The successful extrapolation of molecular features show that partial least square regression is capable of providing meaningful models for extrapolation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Both authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.832120/full#supplementary-material>

REFERENCES

- Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol. (Tist)*. 2, 27. doi:10.1145/1961189.1961199
- ChemAxon (1998). ChemAxon Chemical Hashed Fingerprint. Available at: <https://docs.chemaxon.com/display/docs/chemical-hashed-fingerprint.md%relax>.
- Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). QSAR Modeling: where Have You Been? Where Are You Going to? *J. Med. Chem.* 57, 4977–5010. doi:10.1021/jm4004285
- Cortés-Ciriano, I., Firth, N. C., Bender, A., and Watson, O. (2018). Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening. *J. Chem. Inf. Model.* 58, 2000–2014. doi:10.1021/acs.jcim.8b00376
- Daylight (1998). Daylight Fingerprints. Available at: <https://www.daylight.com/dayhtml/doc/theory/theory.finger.html>.
- De Jong, S. (1993). SIMPLS: an Alternative Approach to Partial Least Squares Regression. *Chemometrics Intell. Lab. Syst.* 18, 251–263. doi:10.1016/0169-7439(93)85002-x
- Glem, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., and Smith, J. (2006). Circular Fingerprints: Flexible Molecular Descriptors with Applications from Physical Chemistry to ADME. *IDrugs*. 9, 199–204.
- Gramatica, P. (2007). Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* 26, 694–701. doi:10.1002/qsar.200610151
- Jaworska, J., Nikolova-Jeliazkova, N., and Aldenberg, T. (2005). QSAR Applicability Domain Estimation by Projection of the Training Set Descriptor Space: a Review. *Altern. Lab. Anim.* 33, 445–459. doi:10.1177/026119290503300508
- Jenkins, R. (2006). Hash Functions and Block Ciphers. Available at: <http://burtleburtle.net/bob/c/lookup3.c>.
- Datasets Korff, M. (2021). Available at: <https://github.com/Actelion/openchemlib/tree/master/examples/MachineLearning/LimitsOfPrediction>.
- Li, H. (2021). Smile - Statistical Machine Intelligence and Learning Engine. Available at: <http://haifengl.github.io/>.
- Mannhold, R., Poda, G. I., Ostermann, C., and Tetko, I. V. (2009). Calculation of Molecular Lipophilicity: State-Of-The-Art and Comparison of Log P Methods on More Than 96,000 Compounds. *J. Pharm. Sci.* 98, 861–893. doi:10.1002/jps.21494
- McGregor, M. J., and Pallai, P. V. (1997). Clustering of Large Databases of Compounds: Using the MDL “Keys” as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* 37, 443–448. doi:10.1021/ci960151e
- Member countries, O. (2004). Validation of (Q)SAR Models. Available at: <https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm>.
- Nettles, J. H., Jenkins, J. L., Bender, A., Deng, Z., Davies, J. W., and Glick, M. (2006). Bridging Chemical and Biological Space: “target Fishing” Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* 49, 6802–6810. doi:10.1021/jm060902w
- Netzeva, T. I., Worth, A., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., et al. (2005). Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* 33, 155–173. doi:10.1177/026119290503300209
- OECD (2014). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. *OECD Series on Testing and Assessment*. Paris, France: OECD Publishing, p 154. doi:10.1787/9789264085442-en.
- Olier, I., Sadawi, N., Bickerton, G. R., Vanschoren, J., Grosan, C., Soldatova, L., et al. (2018). Meta-QSAR: a Large-Scale Application of Meta-Learning to Drug Design and Discovery. *Mach. Learn.* 107, 285–311. doi:10.1007/s10994-017-5685-x
- Roy, K., Kar, S., and Ambure, P. (2015). On a Simple Approach for Determining Applicability Domain of QSAR Models. *Chemometrics Intell. Lab. Syst.* 145, 22–29. doi:10.1016/j.chemolab.2015.04.013
- Sakia, R. M. (1992). The Box-Cox Transformation Technique: a Review. *The Statistician*. 41, 169–178. doi:10.2307/2348250
- Sander, T., Freyss, J., von Korff, M., and Rufener, C. (2015). DataWarrior: an Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* 55, 460–473. doi:10.1021/ci500588j

- Sander, T. (2021). DataWarrior. Available at: <http://www.openmolecules.org/datawarrior>.
- Smola, A. J., and Schölkopf, B. (2004). A Tutorial on Support Vector Regression. *Stat. Comput.* 14, 199–222. doi:10.1023/b:stco.0000035301.49549.88
- Todeschini, R., and Consonni, V. (2008). Handbook of Molecular Descriptors. *John Wiley & Sons*. 11. doi:10.1002/9783527613106
- Tong, W., Hong, H., Xie, Q., Shi, L., Fang, H., and Perkins, R. (2005). Assessing QSAR Limitations - A Regulatory Perspective. *Cad.* 1, 195–205. doi:10.2174/1573409053585663
- Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* 29, 97–101. doi:10.1021/ci00062a008
- Xu, Q.-S., and Liang, Y.-Z. (2001). Monte Carlo Cross Validation. *Chemometrics Intell. Lab. Syst.* 56, 1–11. doi:10.1016/s0169-7439(00)00122-2

Conflict of Interest: Both authors are employed by Idorsia Pharmaceuticals Ltd.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 von Korff and Sander. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



From Data to Knowledge: Systematic Review of Tools for Automatic Analysis of Molecular Dynamics Output

Hanna Baltrukevich^{1,2} and Sabina Podlowska^{1*}

¹Maj Institute of Pharmacology, Polish Academy of Sciences, Kraków, Poland, ²Faculty of Pharmacy, Chair of Technology and Biotechnology of Medical Remedies, Jagiellonian University Medical College in Krakow, Kraków, Poland

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Mohamed AbdElAziz Khamis Omar,
Egypt-Japan University of Science
and Technology, Egypt
Rafael Doležal,
University of Hradec Králové, Czechia

*Correspondence:

Sabina Podlowska
smusz@if-pan.krakow.pl

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 27 December 2021

Accepted: 26 January 2022

Published: 10 March 2022

Citation:

Baltrukevich H and Podlowska S
(2022) From Data to Knowledge:
Systematic Review of Tools for
Automatic Analysis of Molecular
Dynamics Output.
Front. Pharmacol. 13:844293.
doi: 10.3389/fphar.2022.844293

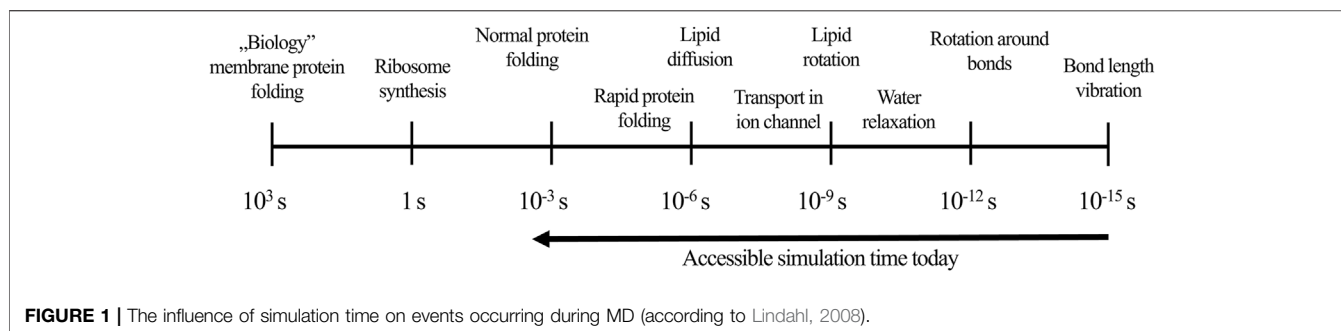
An increasing number of crystal structures available on one side, and the boost of computational power available for computer-aided drug design tasks on the other, have caused that the structure-based drug design tools are intensively used in the drug development pipelines. Docking and molecular dynamics simulations, key representatives of the structure-based approaches, provide detailed information about the potential interaction of a ligand with a target receptor. However, at the same time, they require a three-dimensional structure of a protein and a relatively high amount of computational resources. Nowadays, as both docking and molecular dynamics are much more extensively used, the amount of data output from these procedures is also growing. Therefore, there are also more and more approaches that facilitate the analysis and interpretation of the results of structure-based tools. In this review, we will comprehensively summarize approaches for handling molecular dynamics simulations output. It will cover both statistical and machine-learning-based tools, as well as various forms of depiction of molecular dynamics output.

Keywords: molecular dynamics, machine learning, structure-based drug design, clustering, data dimensionality reduction, interaction fingerprints

INTRODUCTION

Structure-based drug design is becoming an indispensable part of virtual screening campaigns, due to the expanding possibilities of carrying out experiments from this path. It is related both to the achievements in the field of crystallography (expressed by the increasing number of deposited crystal structures), but also to the availability of the computational power and more efficient computational algorithms. Structure-based tools, with their key representatives—docking and molecular dynamics simulations—are a great source of information on the possible interaction schemes occurring between ligand and target receptors (Yang, 2014; Wang et al., 2018).

Molecular docking is a technique that aims to predict the optimal binding mode(s) of a ligand in the respective receptor (Morris and Lim-Wilby, 2008; Guedes et al., 2014; Ferreira et al., 2015). As the docking methodology relies on minimizing free energy of the ligand-receptor complex, the obtained structure can constitute a good starting point for more detailed analysis of ligand-protein interactions during molecular dynamics (MD) simulations (Santos et al., 2019; Wang et al., 2019). Moreover, as most docking tools provide limited flexibility of the target, MD can explore conformational space and generate an ensemble of receptor conformations, which could further be



used during screening of chemical databases (Amaro et al., 2018; Acharya et al., 2020). The so-called ensemble sampling has not only increased the hit rate and, thus, improved the quality of virtual screening, but has also allowed efficient docking to the so-called “difficult protein targets” (Fu et al., 2014; Ellingson et al., 2015; Uehara and Tanaka, 2017; Bhattarai et al., 2020).

MD is an approach that relies on simulating dynamical changes of the system and capturing its evolution in time. MD offers an insight into the movement of the ligand-receptor complex at an atomistic level. Furthermore, it enables quantitative estimation of parameters that cannot be established in wet-lab experiments, e.g., values of torsional angles to describe flexibility, solvent accessible surface area to predict stability, or change in the entropy for distinct structures, such as water molecule in particular location (Ferreira et al., 2015; Leimkuhler and Matthews, 2016; Hollingsworth and Dror, 2018). The basis of the classical MD methodology is solving the Newton’s motion equations for each atom in the system, where the potential energy and forces of interacting particles are from the force-field definitions (Sutmann, 2002; Lindahl, 2008). These approximations are necessary to balance between the required accuracy and optimal speed of simulations’ performance. Moreover, MD timestep should be very small—1–10 fs – in order to minimize errors related to the potential energy estimation (Binder et al., 2004; Leimkuhler and Matthews, 2016). Huge numbers of timesteps, which are required for even relatively short simulations, contribute to the consumption of a great amount of computational resources. Fortunately, due to the increasing computational power and possibilities to perform simulations with the use of graphical processing units (GPU), MD simulations reached a millisecond time scale allowing to investigate events such as protein folding (Figure 1; Lindahl, 2008).

Thus, the amount of data produced by MD has dramatically increased over recent years and is far beyond the accessibility of manual analysis. For this reason, it is crucial to develop automatic tools for post-processing of such data. Great numbers of approaches are offered specifically by the software for MD simulations. Nevertheless, a lot of new independent methods for automated analysis have appeared recently, which are based on various statistical methods and machine learning (ML).

ML approaches are nowadays used at each stage of the drug design process and development (Ballester, 2019; Vamathevan

et al., 2019; Patel et al., 2020). Their most common application involves the evaluation of compound potential bioactivity in ligand-based virtual screening (Melville et al., 2009; Carpenter and Huang, 2018; Hussain et al., 2021); however, they are also widely applied in the evaluation of compound physicochemical and ADMET properties (Göller et al., 2020; Göller et al., 2022; Jia and Gao, 2022). The ML role in computer-aided drug design is not limited to the assessment of compound libraries, but a number of generative approaches is used to enumerate new sets of potentially active compounds (Baskin, 2020). Moreover, ML can help in the compound optimization and indication of features, which are important for a particular type of activity, thanks to the wide range of interpretability tools (Hudson, 2021). ML methods also support structure-based path of virtual screening tasks – they assist in the detection of ligand-protein interaction patterns characteristic for considered activity profiles (Khamis et al., 2015; Khamis and Gomaa, 2015; Khamis et al., 2016), as well as in the detection of complex relationships between ligand-protein interaction schemes occurring during MD simulations (Podlowska et al., 2020; Kucwaj-Brysz et al., 2021).

In this review, we comprehensively summarize existing approaches to automatic handling of MD simulations’ outputs. We will describe approaches available within the MD software, but our main focus is on the automatic statistical and ML-based post-processing tools.

TOOLS AVAILABLE WITHIN THE MD SOFTWARE OR PACKAGES DEDICATED TO MD OUTPUT ANALYSIS

Numerous software packages are able to perform MD simulations. The list of the most popular programs includes GROMACS (Abraham et al., 2015), HyperChem (Laxmi and Priyadarshy, 2002), AMBER (Case et al., 2005), LAMMPS (Thompson et al., 2021), CHARMM (Brooks et al., 2009), DL_POLY (Todorov et al., 2006), HOOMD (Glaser et al., 2015), TINKER (Lagardère et al., 2018), NAMD (Phillips et al., 2005), and Desmond (Bowers et al., 2006). The resulting simulation trajectory can then be analyzed at different levels – from the qualitative visualization of changes occurring in the modeled system to detailed investigation of variations in atom positions and ligand-protein interactions. Due to the high

amount of data produced during MD simulations (of up to several terabytes size), programs for MD analysis should also be able to efficiently deal with such data volumes.

The list of the most known packages for MD simulations analysis opens VMD [Visual Molecular Dynamics (Humphrey et al., 1996)], developed by the Theoretical and Computational Biophysics Group at the University of Illinois at Urbana-Champaign. VMD is a program designed for interactive visualization and analysis of biomolecular systems including processing of very large systems (composed of up to billion particles). The software is written in C and C++ (source code available) and is distributed free of charge. Convenient graphical interface supports performing various types of coordinate analysis on Unix, MacOS, and Windows operating system, along with NVIDIA OptiX and CUDA support. In addition to the built-in analysis tools applicable to trajectories processing, VMD has a broad collection of plugins and scripts (VMD Plugin Library, 2021, n. d.; VMD Script Library (2021), n. d.).

Execution of Tcl and Python scripts and implementation of developed plugins enables adjustment of VMD capabilities to users' needs without recompiling the source code. Both types of tools are distributed under an open-source license, unless otherwise stated. Moreover, researchers are encouraged to develop and share new utilities in order to support the growth of the VMD community and development of the software. VMD plugins are divided into the "molfile" plugins, which enable working with multiple file formats of molecular data, and scripting extensions used to perform requested tasks. Plugins dedicated to data analysis allow performing various calculations: from RMSD (*RMSD Tool*, *RMSD Trajectory Tool*) to electrostatic potentials (*APBSRun*, *Delphi Force*) and IR spectral density (*IRSpecGUI*). Resulting outcomes can be visualized through generated plots—*GofrGUI*, *NAMD Plot*, *RamaPlot*, *Timeline*—or as maps—*Contact Map*, *VolMap*, *HeatMapper*, *PMEpot*. There are also plugins capable of analysing free-energy perturbation calculations (*AlaScan*, *ParseFEP*) and obtaining data on proteins—*Intervor* (extracts and displays protein-protein interface), *SurfVol* (measures surface area and volume of proteins), and *NetworkView* (shows protein interaction networks). Developed statistical tools visualize clusters of structure conformations (*Clustering Tool*) or perform normal mode visualization and comparative analysis (*NMWiz*). VMD has constantly been developed: the latest version (1.9.3) includes introduction of the following major features: introduction of new QwikMD plugin connecting VMD with MD program NAMD, enabling quick preparation of common molecular simulations; the TopoTools plugin used for automated topology conversion from CHARMM to GROMACS; the new TachyonL-OSpray ray tracing engine for generating high quality renderings of molecular systems containing hundreds millions of particles; and OpenGL rendering for parallel visualization runs on "headless" clouds and petascale computers.

PTRAJ (Process TRAJectory) is another example of a tool enabling post-processing of MD data (Roe et al., 2013). It was dedicated for the analysis of the AMBER output. Its successor, CPPTRAJ, emerged as a response to the growing trajectory sizes, offering a wider range of functionalities and more efficient data processing. In contrast to PTRAJ (written primarily in C), CPPTRAJ code is based on C++ and the whole program

structure was reorganized to facilitate the addition of new functionalities. The programs and their source code are freely available under the GNU General Public License version 3 and are distributed within the AmberTools21. The strong point of CPPTRAJ is batch-processing, which allows the use of remote sites for analysis and possibility of combining various types of commands, trajectories, and topologies in the same run. Other important features of CPPTRAJ are: the availability of MPI, OpenMP, and CUDA parallelization, support for implementation of variables and loops, and possibility to apply atom masking to specify which part of the system should be analyzed. The number of developed commands applicable for MD data analysis is great, including simple calculations, such as estimation of the number of hydrogen bonds (*hbond*), and multiple examples of more complex tools, such as performing non-linear curve fitting (*curvefit*, *multicurve*) and linear regression (*regress*), matrix based calculations (*crosscorr*, *diagmatrix*, *hausdorff*, *modes*), estimating auto-/cross-correlation (*autocorr*, *correlationcoe*, *timecorr*), creating histograms (*hist*, *kde*, *multihist*), and many more (Case et al., 2021). CPPTRAJ development has resulted in new features, among which are: rewritten code expanding clustering capabilities, ability to RMS-fit grids onto coordinates, automatic calculation of multiple puckers, speeding up the non-bonded energy calculation, enhancing the performance of the *permutedihedrals* and *randomizeions* commands, and automation of downloading and building external libraries in CPPTRAJ (2021).

MDAnalysis is an object-oriented library developed for the analysis of MD trajectories and protein structures (Michaud-Agrawal et al., 2011). The package is written in Python and Cython and uses NumPy arrays to expand its functionality. MDAnalysis is available under the GNU General Public License version 2.0 (<https://github.com/MDAnalysis/mdanalysis>). The analysis modules are capable of assessing distances and contacts (e.g., calculating path similarity, which reveals geometric similarity of trajectories useful for identification of patterns in trajectory), performing dimensionality reduction and carrying out volumetric analysis (e.g., linear density estimation). Other modules analyze the structure of macromolecules (such as HELANAL (Sugeta and Miyazawa, 1967; Bansal et al., 2000)—a tool for the analysis of protein helices), polymers (including determination of the polymer persistence length), nucleic acids and, finally, membrane and membrane proteins (namely, HOLE (Stelzl et al., 2014), a suite of tools used to assess pore dimensions of the holes as a function of time). Recently MDAnalysis announced the introduction of a command-line interface in answer to user needs, and a number of supported analysis modules is provided in the documentation.

MDTraj (McGibbon et al., 2015) is a Python library applied for MD trajectory manipulation and analysis, whose goal is to provide interface between MD data and modern tools and programs for statistical analysis and visualization based on Python. MDTraj is licensed under the Lesser GNU General Public License (LGPL v2.1+) on GitHub (<https://github.com/mdtraj/mdtraj>). MDTraj works with every possible MD data format, focusing on speed and efficient performance and

providing multiple analysis possibilities. Available functions identify hydrogen bonds, compute distances to create residue-residue contact maps, assess secondary structure of the protein and assign code according to the implemented Dictionary (Kabsch and Sander, 1983), calculate solvent-accessible surface area (SASA) and NMR scalar coupling, as well as determine nematic order parameters, which describe the orientational order of a system from 0 to 1. Another special feature is the particularly fast RMSD computations due to performance optimization based on Haque et al. (2014) along with C/C++ code implementation. Moreover, MDTraj documentation gives access to 14 notebooks containing analysis examples with executable code—e.g., PCA with scikit-learn ML library followed by plotting data using Matplotlib.

LOOS (Lightweight Object-Oriented Structure-analysis) (Romo et al., 2014; Grossfield and Romo, 2021) aims at enabling rapid development and testing of new tools for MD analysis. Additionally, the program includes a number of easy-to-use prebuilt applications. As LOOS is a C++ library, its combination with Python interface (PyLOOS) resulted in high performance and simplicity of use and further development. Moreover, the C++ layers could be used independently for even more efficient utilization of resources. LOOS is freely distributed under the GPLv3 license and is available via GitHub (<https://github.com/GrossfieldLab/loos>). In LOOS, 140 prebuilt tools are grouped into the following categories: macromolecule tools (e.g., computation of the radial distribution function), hydrogen bonding handling, principal component analysis (PCA), elastic network models (ENM), clustering, assessment of statistical error (e.g., block-averaged standard error calculations), and convergence. The tools included in the "membrane systems" category are dedicated for analyzing lipid bilayers and associated systems (e.g., calculation of molecular order parameters). Furthermore, 2D Voronoi decomposition tools are used to obtain data within a particular membrane slice. 3D density distributions tools generate 3D histograms from MD trajectories. They were originally created for visualization of water distribution; however, they are able to estimate membrane lipid density as well.

Pteros (Yesylevskyy, 2012; Yesylevskyy, 2015) is a high-performance molecular modeling library available for C++ and Python. It lets users analyze MD data and develop new analysis tools with the assistance of the easy-to-use APIs in both of the above-mentioned programming languages. In order to accelerate the analysis process, Pteros asynchronously reads files with MD trajectories and performs analysis tasks in parallel. Analysis plugins are completely independent and, besides typical calculations, provide more specific manipulations. For example, they enable assessing properties related to curvature with the Curvature plugin, which computes mean and Gaussian curvatures of various lipid aggregates, smooths membrane surfaces, and calculates other properties of molecules embedded into the lipid membrane. While the above-mentioned plugin is not open-source, Pteros is a free software distributed under Artistic License and available at GitHub (<https://github.com/yesint/pteros>).

Till now, we have described exclusively open source software and libraries, which serve as powerful and freely available tools for MD output analysis. Nevertheless, some commercial software is also worth mentioning, e.g., Molecular Operating Environment (MOE)

[Molecular Operating Environment (MOE), 2019], Desmond (Schrödinger Release 2021–4: Desmond Molecular Dynamics System, 2021), and CHARMM (Brooks et al., 2009). MOE constitutes a platform for integrated computer-aided molecular design with vast capabilities: QSAR models generation, virtual screening, protein engineering, homology modeling, as well as carrying out MD simulations. However, MOE offers limited opportunities for MD analysis, as only Free Energy Calculations along with Torsion Scan and Analysis are mentioned at the official software webpage. Greater analysis possibilities are provided by Desmond—a commercial software available without cost for non-commercial use, developed by D. E. Shaw Research for high-speed MD simulations of biological systems. Desmond offers multiple panels for different post-processing operations, such as Trajectory Frame Clustering Panel, Simulation Quality Analysis Panel (enabling estimation of potential energy, temperature, pressure, etc.), Simulation Event Analysis Panel (enabling calculation of geometric and energy-based properties, e.g., RMSF, hydrogen bonds, Coulomb energy), and Radial Distribution Function Panel. What is more, Desmond provides distinct panels for metadynamics and replica exchange simulations analysis, and Python scripts applicable for PCA, density profile calculations, and others. The advantages of MD data analysis in Desmond are its detailed tutorials, intuitive GUI, and convenience of some tools, such as Simulation Interaction Diagram. Its output is saved as a pdf file, which contains results of protein-ligand system analysis in the form of colored plots, together with the short explanation of the meaning of each calculated property.

Plenty of other software and tools are useful in MD data analysis; among them are GROMACS (Abraham et al., 2015) and CHARMM (Brooks et al., 2009)—well-known MD programs capable of performing analysis tasks as well. Carma (Glykos, 2006) is a lightweight program written in C along with its graphical user interface grcarma (Koukos and Glykos, 2013) and Wordom (Seeber et al., 2007; Seeber et al., 2011) - a simple and fast command-line utility. MMTSB (Feig and Karanicolas, 2004) is a set of tools for enhanced sampling and multiscale molecular modeling approaches, while Simulaid (Mezei, 2010) is a program for carrying out analysis tasks of multiple types and MD trajectory data manipulation. MMTK (Hinsen, 2000), the Molecular Modeling Toolkit, contains MD analysis scripts; both Bio3D package (Grant et al., 2006) written in R language, and Python toolkit. MD-Tracks (Verstraelen et al., 2008) provides statistical analysis of MD data, and ST-Analyzer (Jeong et al., 2014) is an intuitive and simple web-based GUI environment, with nine analysis modules for extraction of various parameters from MD output.

MACHINE LEARNING—CLASSES OF MODELS USED IN THE STRUCTURE-BASED DRUG DESIGN

ML methods have become an integral element of structure-based path of drug design, and they assist in the analysis of both docking and MD simulations (Dutta and Bose, 2021). The general task of ML is to detect relationships and complex patterns in large

datasets. As the amount of data produced in the structure-based path has recently grown enormously, the application of ML methods for MD outcome analysis is becoming more and more popular. Within ML methods, we can also distinguish deep learning (DL) algorithms with their main usage in computer-aided drug design to generate examples of new potential ligands via generative approaches.

The most popular classes of ML models applied in the broadly understood campaigns for searching for new drugs include:

- 1) Bayesian models—a collection of models based on the Bayes' theorem. It defines the probability of an event on the basis of prior knowledge of conditions, which might be influencing this event. The Bayes' theorem in its simplest form (taking into account only two events, A and B) can be described using the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(A|B)$ is a conditional probability of occurrence of event A, given that B is true; $P(B|A)$ is a conditional probability of occurrence of event B, given that A is true; and $P(A)$ and $P(B)$ are probabilities of occurrence A and B, respectively, without any conditions ($P(B) > 0$).

Bayes' theorem for a higher number of events adopts the following form:

If B, T_1, \dots, T_n are such events that:

$P(B) > 0$, $BCU_{i=1}^n T_i$ and $T_i \cap T_j = \emptyset (i \neq j)$, then:

$$P(T_j|B) = \frac{P(B|T_j)P(T_j)}{\sum_{i=1}^n P(B|T_i)P(T_i)}$$

In drug design approaches, Bayes' theorem is most often used within the Naïve Bayes algorithm. In such a case, Bayes' theorem is used together with an assumption of events (features) independence (Berrari, 2019).

Another concept using Bayes' theorem is Bayesian statistics, in which all observed and unobserved parameters of a statistical model are given a joint probability distribution (prior and data distribution). Bayesian statistics expresses probability as a *degree of belief*, and Bayes' theorem is used to assign a probability distribution to quantitatively describe this *degree of belief* in the form of a set of parameters (van de Schoot et al., 2021).

The Bayesian concept is also used in fuzzy clustering (Glenn et al., 2015).

- 2) K-nearest neighbors methods – based on the determination of distances between an evaluated sample and representatives of the training set. In its simplest form ($K = 1$), the evaluated sample is assigned to the class of its closest neighbor from the training set (or value of the considered parameter of the closest neighbor is returned in the case of regression). If a higher number of examples closest to the query is considered ($K > 1$), voting for the most frequent class label is carried out (classification) or values of evaluated parameters are averaged (regression)—**Figure 2 a** (Cover and Hart, 1967; Hall et al., 2008).

In MD studies, k-nearest neighbors algorithm is also used in clustering procedures aimed at the formation of groups of geometrically similar conformations (Keller et al., 2010).

- 3) Trees—tree-based algorithms are considered to be one of the most efficient and most broadly used types of ML models. Their important advantage is their simplicity and ease of interpretation, which play a role in drug design protocols (e.g., by the possibility of indication of features important for a particular compound activity). Predictions can be made using one decision tree or multiple trees (as it is in the case of Random Forest). Attributes for a root and subsequent nodes are selected on the basis of their discrimination power (at each level, a feature which provides the best distinguishment between considered classes is selected). Evaluation of new examples is carried out via checking values of features present in the subsequent nodes -**Figure 2B** (Breiman et al., 1984; Quinlan, 1986).
- 4) Neural networks—neural networks search for relationships in data in such a way that they mimic the processes occurring in the human brain. Their neurons are constituted by a mathematical function, which collects and classifies information. Such artificial neurons are interconnected (such connections reflect biological synapses, called edges) and they have the ability to communicate with each other. A neuron (node) receives a signal, processes it, and passes the respective information to the connected neurons. Typically, neurons are organized into layers, and the signal is passed from the input layer (the first one) to the output layer (the last one) (Hopfield, 1982).

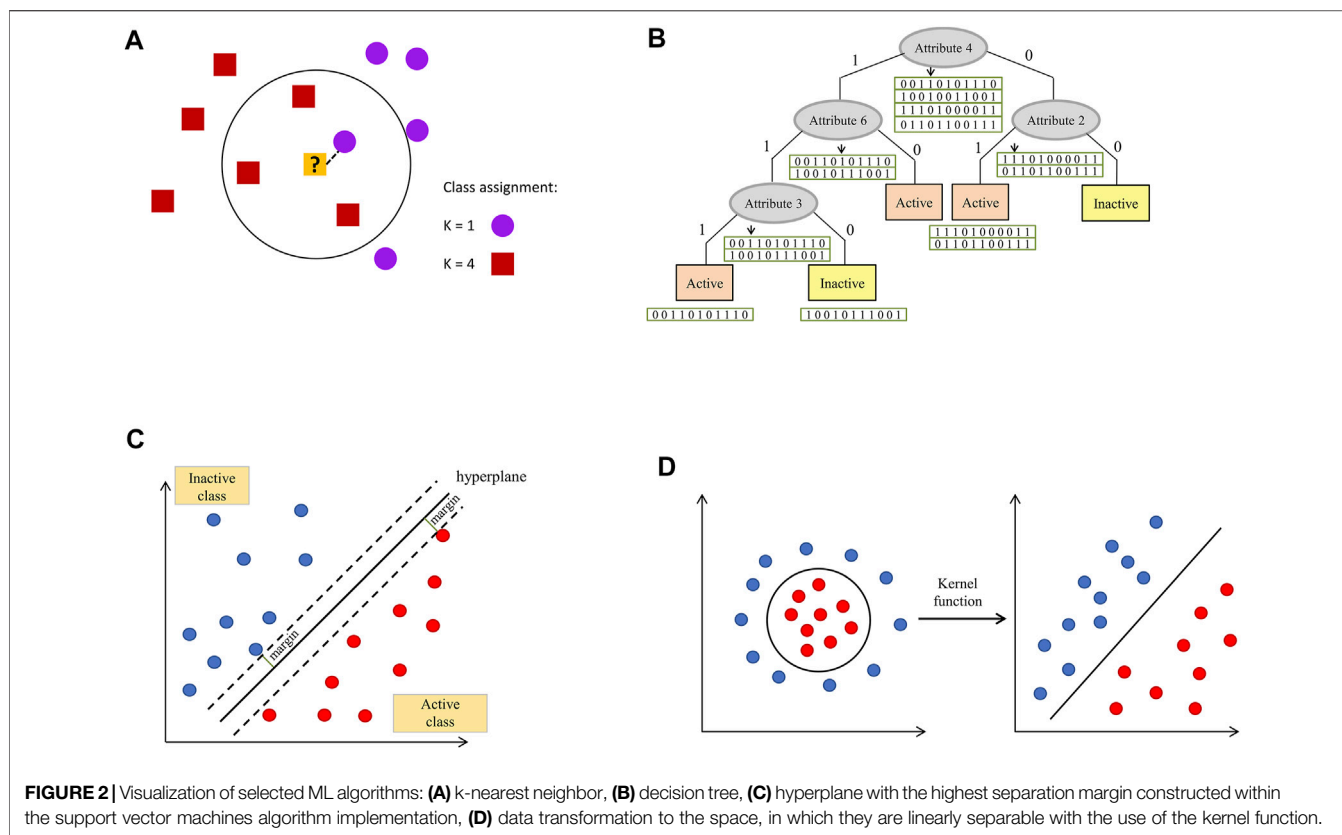
A special type of neural network that has recently gained enormous popularity is deep neural network (DNN) with “deep” referring to the application of multiple layers in the network (LeCun et al., 2015; Schmidhuber, 2015).

Neural networks concept is also applied in unsupervised approaches for MD data clustering, e.g., in the form of Self Organizing Maps (SOMs) (Hyvönen et al., 2001; Fraccalvieri et al., 2013; Mallet et al., 2021). In order not to lose the topological properties of the input space, a neighborhood function is used.

- 5) Support Vector Machines (SVM)—an algorithm according to which each data item constitutes a point in n-dimensional space (n is equal to the number of features), with coordinates defined by the particular feature value. The task of the model is to find a hyperplane, which discriminates example classes with the highest margin (**Figure 2C**). As linear discrimination is often not possible, a kernel function needs to be applied in order to transform the input into a space of higher dimension, so an inseparable problem is converted into a separable one—**Figure 2D** (Cortes and Vapnik, 1995).

CLUSTERING AND REDUCTION OF DATA DIMENSIONALITY

The most common approach to use the automatic post-processing of the MD simulations output is the reduction of



dimensionality and clustering (Amadei et al., 1993; Lange and Grubmüller, 2006).

Clustering

Clustering, from its assumptions, is an unsupervised technique of finding patterns and relationships in data. In contrast to the previously described techniques, clustering does not require the presence of the training set, as its aim is to form subgroups of similar objects. Clustering algorithms use various “distance” measures to evaluate object similarity. Two main groups of clustering approaches can be distinguished, namely partitional and hierarchical, both of which can be carried out in the bottom-up agglomerative way or using a top-down divisive approach (Kaufman and Rousseeuw, 1990). Another group of data grouping methods are density-based schemes, in which the clusters refer to the peaks of the probability distribution (or free energy minima) from which the data are collected (Sander, 2011; Glielmo et al., 2021). In MD simulations, such probability peaks typically correspond to metastable states of the system. An example application of density-based clustering to the analysis of MD data is density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996; Schubert et al., 2017), in which the clusters are defined as regions with density above the particular threshold. Such an approach was used to find representative structures from MD simulations and analyze MD trajectories (Wang et al., 2013). MD trajectories have also been analyzed by the density peak clustering.

The most popular partitional clustering technique is the K-means algorithm. Clustering in this approach starts from the random placement of K initial centroids. Then, K clusters are formed iteratively in such a way that a point which is closest to a particular centroid is added to the respective cluster, and a new centroid for each cluster is determined. When the cluster membership does not change (the convergence is obtained), the process is stopped. The drawback of K-means clustering is the dependence of the final outcome on the initial choice of the centroids. Problems might also occur when significant variations in the cluster sizes or densities appear, when data outliers are present, or when the ‘natural’ clusters have non-spherical shapes (Hartigan and Wong, 1979; Huang, 1998).

The starting point of agglomerative hierarchical clustering is a formation of singleton clusters from each object from the dataset. Then, iterative linkage of the nearest clusters is carried out, until the whole dataset constitutes one group. On the basis of the resulting dendrogram, the final division of data is produced. Hierarchical clustering is deterministic, but it requires high computational power and storage abilities, which limits its application to small datasets.

The most popular metric used to evaluate MD simulations’ output in terms of data proximity is Root Mean Square Deviation (RMSD). Despite the presence of some drawbacks [e.g., incidents of wrong conclusions when applied to equilibrium evaluation (Grossfield and Zuckerman, 2009)], it is still the most frequently used method for comparison of

conformation similarity. Several different solutions were also proposed, such as the application of Euclidean Distances Matrices (EDM) (de Souza et al., 2017); however, they have not gained such wide popularity as RMSD.

Evaluation of Clustering Approaches

The evaluation of clustering is not easy, as falling into the group of unsupervised approaches, clustering does not refer to true labels. One group of cluster assessment methods is the so-called “internal evaluation,” where clusters are evaluated on the basis of the clustered data. In general, in such an evaluation, the highest scores are assigned to the approaches which produce clusters of high similarity between particular cluster elements and low similarity between elements belonging to different clusters (Rand, 1971). An example of internal measure of clustering quality is Davies-Bouldin index (DB) (Davies and Bouldin, 1979):

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

with n being the number of clusters, c_i , c_j being centroids of clusters i and j , respectively; σ_i refers to the average distance of elements belonging to cluster i to its centroid c_i ; and $d(c_i, c_j)$ is the distance between centroids of clusters i and j . The lower the values of DB index, the better they are.

Another approach of the assessment of clustering quality is external evaluation, which refers to pieces of information that were not used during clustering. External evaluation can be based on the known class labels or on some benchmark datasets. However, if the true class labels are known, the clustering is actually not needed (de Souto et al., 2012).

Before the application of methods for clustering evaluation, the dataset should be examined in terms of the clustering tendency. If the dataset is composed of the uniformly distributed points (therefore, there is no clustering tendency present), then the identified clusters may be invalid. In order to verify the clustering tendency, the Hopkins test (Hopkins and Skellam, 1954) can be used (statistical test for spatial randomness of a variable).

Reduction of Data Dimensionality

Principal Component Analysis (PCA) is an approach for the reduction of the data dimensionality via transformation of a large set of variables into a smaller one, preserving as much information of the original set as possible (Ichiye and Karplus, 1991; Jolliffe, 2002; Jolliffe and Cadima, 2016). The goal is obtained via extraction of important information from the data table and its representation in the form of new orthogonal (linearly independent) variables (principal components). Then, the relationships between observations and variables can be displayed in the form of points in the maps. PCA is based on the assumption that the phenomena of interest can be explained by variances and covariances between original variables from the dataset. PCA is often applied before performing the clustering procedure. In MD-related applications, PCA is responsible for extracting the dominant modes in the molecule motion. It should be pointed out that, during the MD,

the Cartesian positions of all atoms of the simulated system (of a size of thousands or even millions of atoms) are recorded in every time step, which indicates the importance of application of post-processing methods. If the dimensionality reduction is carried out properly, all relevant information is preserved, and the analysis of the MD output is valid.

Another approach for reduction of data dimensionality is multidimensional scaling (MDS), which determines the data space of lower dimension with the best possible preservation of the pairwise distances between data points (Young and Householder, 1938; Torgerson, 1952). Its mode of action is closely related to PCA; however, for MDS it is sufficient to provide a pairwise distance between points (their exact positions are not necessary).

PCA and MDS are representatives of linear methods of data dimensionality reduction; however, there is also a number of non-linear approaches to this task, with such examples as isometric features mapping (Tenenbaum et al., 2000), kernel PCA (Schölkopf et al., 1998), diffusion map (Coifman et al., 2005; Coifman and Lafon, 2006), and t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008). Low-dimensional spaces to embed high-dimensional data are also more and more often determined using DL approaches. One of the most popular DL techniques for reduction of data dimensionality is autoencoder (Kramer, 1991). Autoencoder maps input configuration to representation of lower dimension and then maps it back to the original space *via* respective decoder. Low-dimensional representation is learned *via* minimization of error between the original data points and data points obtained by the application of the above-mentioned decoder. Another DL-based approach for reduction of data dimensionality falls into the group of generative neural networks. Its representatives include Variational Autoencoders (VAEs) (Lopez et al., 2018) and Generative Adversarial Networks (GANs) (Goodfellow, et al., 2014).

Examples of Clustering and Data Dimensionality Reduction for MD Output Analysis

Unsupervised procedures are widely applied in the MD outcome analysis, due to the above-mentioned problem of the vast amount of data produced during simulations: clustering data into groups gathering similar conformations obtained during MD, and reduction of data dimensionality which lowers the number of features considered. Both these approaches help in the analysis of MD output.

The problem of clustering MD data emerged quite early. The first reports of clustering MD output were released in the early 1990s (Gordon and Somorjai, 1992; Torda and van Gunstered, 1994). Various groups also compared effectiveness of various clustering algorithms (Shao et al., 2007; Keller et al., 2010; Abramyan et al., 2016). Nowadays, clustering of MD data has become a standard procedure applied in order to facilitate interpretation and analysis of MD trajectories (Bruno et al., 2011; De Paris et al., 2015a; De Paris et al., 2015b; Rudling et al., 2018; Takemura et al., 2018; Evangelista et al., 2019;

Yoshino et al., 2019; Bekker et al., 2020; Roither et al., 2020; Araki et al., 2021; Mallet et al., 2021; Wu et al., 2021) and new algorithms to improve this procedure are constantly developed.

Dimensionality reduction of MD data with the use of PCA was also first used in the early 90s (Ichiye and Karplus, 1991; Amadei et al., 1993) and since that time its application in MD output analysis has been constantly growing (Das and Mukhopadhyay, 2007; Chiappori et al., 2010; Kim et al., 2010; Casoni et al., 2013; Ng et al., 2013; Novikov et al., 2013; Bhakat et al., 2014; Sittel et al., 2014; Ernst et al., 2015; Chaturvedi et al., 2017; Cossio-Pérez et al., 2017; Fakhar et al., 2017; Chen, 2018; Cholko et al., 2018; An et al., 2019; Barletta et al., 2019; Girdhar et al., 2019; Karnati and Wang, 2019; Lipiński et al., 2019; Martínez-Archundia et al., 2019; Wu et al., 2019; Magudeeswaran and Poomani, 2020; David et al., 2021; Majumder and Giri, 2021). Although PCA is the most popular approach applied to handle MD trajectories, other data dimensionality reduction methods are also used in the MD field. Pisani et al. used MDS to examine conformational landscapes of CDK2 (Pisani et al., 2016) and Bécavin et al. improved the application of MDS for MD data by using singular value decomposition. MDS in the context of MD was also described by Troyer and Cohen (1995), Andrecut (2009), Tribello and Gasparotto (2019), and Srivastava et al. (2020). There are also examples of the application of other approaches: isometric feature mapping (Stamati et al., 2010), kernel PCA (Antonioni and Schwartz, 2011), diffusion map (Rohrdanz et al., 2011; Zheng et al., 2011; Zheng et al., 2013a; Zheng et al., 2013b; Preto and Clementi, 2014), t-SNE (Zhou et al., 2018; Zhou et al., 2019; Spiwok and Kříž, 2020), and VAE (Hernández et al., 2018; Shamsi et al., 2018; Moritsugu, 2021; Tian et al., 2021).

MARKOV STATE MODELING

Markov state modeling (MSM) (Pande et al., 2010; Husic and Pande, 2018) is another approach widely applied in the MD-based studies. MSM can be used to characterize events that occur at longer timescales than available computational power to perform such long simulation. Such MDs are simulated as transitions between a set of discrete stable states. The MSM parametrization can be performed *via* running several short MDs, which can be computed in parallel. The main difficulty in the MSM application is definition of the above-mentioned stable states (Abella et al., 2020). In general, MSM is an approach for modeling random processes with the use of the Markov assumption, which is when the present state is given, all following states are independent of all past states. MSMs describe the stochastic dynamics of a biomolecular system using two objects: a discretization of the high-dimensional molecular state space into *n* disjoint conformational sets and a model of the stochastic transitions between these states [usually described by a matrix of conditional transition probabilities (Chodera and Noé, 2014)].

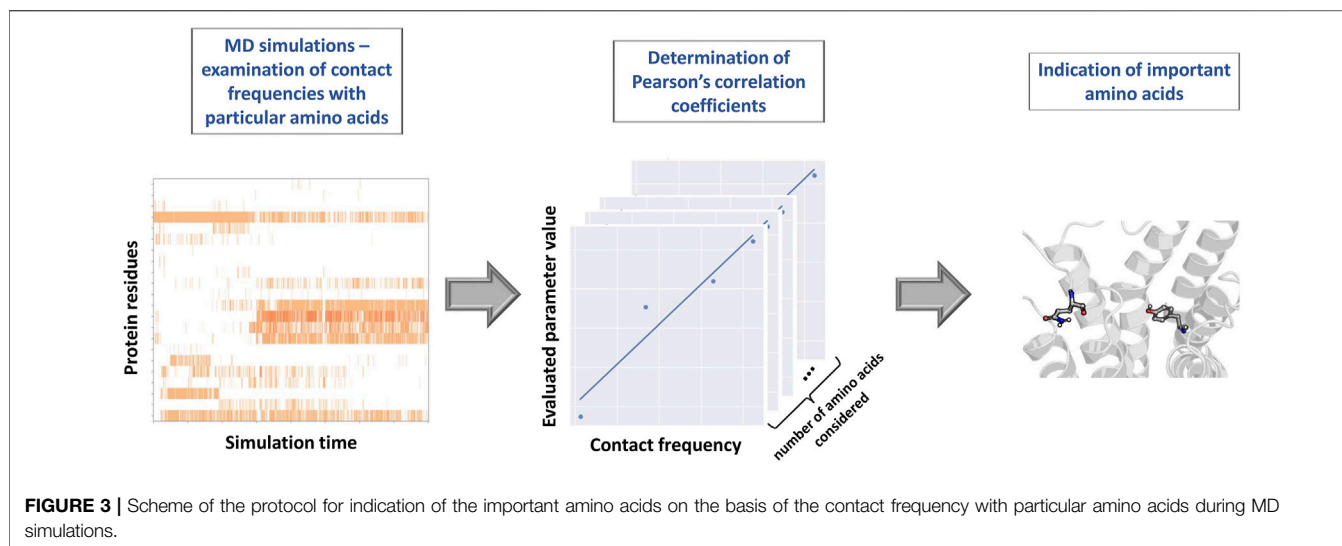
Examples of MSM applications in drug design include: examination of the binding kinetics of the trypsin inhibitor benzamidine (Buch et al., 2011), description of the multiple unbinding pathways of ligands dissociating from FKBP

(Huang and Caflisch, 2011), examination of substrate binding mechanism of HIV-1 protease (Pietrucci et al., 2009), analysis of binding pathways of opiates to μ -opioid receptors (Barati et al., 2018), reconstruction of binding process of alprenolol to the beta2-adrenergic receptor (Bernetti et al., 2019), membrane-mediated ligand unbinding of the PK-11195 ligand from the translocator protein (TSPO) (Dixon et al., 2021), study of the two bromodomain-inhibitor systems using multiple docked starting poses (Dickson, 2018), examination of the unbinding kinetics of a p38 MAP kinase type II inhibitor (Casasnovas et al., 2017), examination of ligand-induced active-inactive conformation change of beta-2 adrenergic receptor (Bai et al., 2014), and investigation of the interplay of conformational change and ligand-binding kinetics for the serine protease trypsin and its competitive inhibitor benzamidine (Plattner and Noé, 2015).

EXAMPLES OF ML-BASED ANALYSIS OF MD

The proper representation of MD outcome opens the door to the wide range of possibilities in terms of the post-processing approaches. Podlewska et al. (2020) and Kucwaj-Brysz et al. (2021) analyzed ligand-receptor contact patterns occurring during MD simulations and examined them with reference to the modeled property. Via the calculation of the Pearson's correlation coefficient between the contact frequencies and values of examined parameters, the highest correlated residues (considered as the most important for the modeled property) were detected. Scheme of the above-described protocol is presented in **Figure 3**. At first, each simulation frame was represented with the use of the Structural Interaction Fingerprints (Singh et al., 2006). Then, for each amino acid, the contact frequency during simulation was calculated. Finally, for each protein residue, the Pearson's correlation coefficients between the respective contact frequency and values of the evaluated compound parameters were determined. The highest correlated positions were indicated as those which should be considered in detail during the further design of compounds of particular activity profile.

Riniker (2017) developed a molecular dynamics fingerprint (MDFP) to combine MD approach with ML methods. MDFPs were obtained *via* the extraction of three properties from MD trajectories: intramolecular and total potential energy of the solute, radius of gyration, and solvent-accessible surface area resulting in a vector of floats. The fingerprint also contained information on the distribution of each property, characterized by its average, standard deviation, and median values. In addition, MDFP was enriched with standard 2D fingerprints: Morgan fingerprints and 2D-counts fingerprints from RDKit (number of heavy atoms, number of rotatable bonds, number of N, O, F, P, S, Cl, Br, and I atoms in the compound). Such representation constituted an input for ML models, which were trained to predict solvation free energies in five different solvents (water, octanol, chloroform, hexadecane, and cyclohexane) and partition coefficient in octanol/water, hexadecane/water, and cyclohexane/water.



MDFP was also used by Gebhardt et al. (2020). In this approach, ML was combined with the atomistic MD simulations encoded with MDFPs enabling the large-scale free-energy calculations. The so-called ML/MDFP method overcomes limitations related to free-energy estimation with MD – high computational expense and imperfections of force-fields. ML models are able to detect systematic force field errors caused by specific chemical groups and, afterwards, decrease their influence on final prediction. Moreover, ML models provide efficient and fast calculations when working with fingerprints databases; as an example, Gebhardt et al. utilized the distributions of potential energy of the solute, radius of gyration, and SASA, which were generated from MD data. The outcomes proved that ML/MDFP approach predicted free-energy not worse or even slightly better than rigorous free-energy simulations and two models, namely quantum chemistry-based COSMO-RS. When two models for free energy predictions (COSMO-RS and UNIFAC) were compared with the support vector regression (SVR), it appeared that the latter one demonstrated the best results. The other application of fingerprints extracted from MD could be distinguishing active compounds, as Jamal et al. (2019) proved on the example of caspase-8 ligands. MD descriptors determined in this work were analogous to those obtained by Gebhardt et al. Moreover, fingerprints of different types were also calculated for reference. Multiple combinations of 2D, 3D, and MD descriptors were used to train two ML models: artificial neural networks and Random Forest. MD descriptors used individually showed better performance than being combined with other 2D/3D descriptors, which proved applicability of MD descriptors for lead prioritization and optimization of caspase-8 ligands.

Ash and Fourches (2017) made benefits of combination of MD and chemical descriptors to generate innovative QSAR models based on MD data, resulting in the construction of the so-called hyperpredictive MDQSAR models. The researchers in their work hypothesized that exploring dynamic noncovalent protein-ligand interactions would help to distinguish active compounds from

non-active. A set of ERK2 inhibitors served as a case study, after previous unsuccessful attempts to rank them using conventional QSAR and sophisticated molecular docking techniques. Each ligand was docked in the ERK2 binding site using Glide, then 20 ns simulations of obtained ligand-protein complexes were performed in Desmond. MDs were followed by the extraction of descriptors on the basis of MD data with KNIME, such as traditional 1D-MACCS fingerprints, as well as 2D RDKit, 3D-D Moments and 3D-WHIM descriptors. The results indicate that MD descriptors successfully tackled the primary challenge and clearly pointed out the most active ligands. The hierarchical clustering highlighted similarities between MD descriptors and activities; furthermore, MD descriptors turned out to be useful in the identification of activity cliffs in all descriptor spaces. The research underlines the importance of further investigation of the MD descriptors usage, which could lead to implementation of new highly effective MDQSAR models in the future computer-aided drug design workflows.

MD data were also used by Vitek et al. (2013) to develop Support Vector Regression (SVR) model for water molecule energy estimation and by Jamroz et al. (2012) to examine fluctuations of protein residues during simulation.

Exploring protein conformations is extremely useful in understanding protein structure and function. However, to capture conformational changes we would need to perform long-time simulations and overcome multiple high energy barriers between local energy minima, which is related to the consumption of significant amounts of computational resources. Traditionally, enhanced sampling methods are exploited to solve these problems; however, their efficiency requires improvement (Yang et al., 2019). Fortunately, owing to technology advances, numerous novel efficient techniques have been developed. For example, a number of DL-based, approaches have already been proposed, such as variational autoencoders (VAEs), which significantly increases sampling “power”, if combined with MD potential. Tian et al. (2021) demonstrated successful protein sampling with VAEs on the example of adenosine

kinase (ADK) conformational change from its closed state to the open one. Decoded conformations were similar to the training ones. Additionally, the latent space provided by VAEs could serve as a starting point for new simulations and studying of unexplored conformational spaces. VAEs application allows to perform short simulations of 20 ns and reach sampling efficiency comparable to a single long MD simulation. Another example of analysis of MD trajectories of proteins applies the Bayesian interference method to perform structural fitting for removing time-dependent translational and rotational movements (Miyashita and Yonezawa, 2017). On the other hand, Perez et al. (2015) combined MD with Bayesian interference to speed up simulation. The combination of Bayesian interference with MD simulations was also used by Shevchuk and Hub (2017) to refine structures and ensembles against small-angle X-ray scattering (SAXS) data.

Proteins change their conformations upon the influence of many factors, such as temperature, pH, and more importantly as a consequence of molecular recognition due to ligand binding (Doms et al., 1985; Takeda et al., 1989; Andersen et al., 1990). What is more, the ligand-protein complex is formed by the induced fit of both molecules, and the resulting protein conformations depend on the structure of the ligand (Bosshard, 2001). Conformational dynamics of proteins have a profound effect on cell functioning, such as in the case of G-protein coupled receptors (GPCRs), which transduce external signals into cells by activation of specific cellular pathways. The binding of different ligands stabilizes certain conformational state, which results in the elicitation of distinct signalling—a phenomenon called functional signalling, or biased agonism (Hilger et al., 2018; Wooten et al., 2018). An essential role of GPCRs in signal transmission highlights the importance of understanding how ligand binding alters protein conformations, in order to design new GPCR ligands, which would target desired pathways and avoid others, potentially causing side effects. MD is perfectly suited for perceiving ligand-protein conformational change; however, the difficulty lies in the necessity to analyze long-scale MD simulations, which are required to capture tiny structural changes, responsible for functional signalling. Plante et al. (2019) successfully applied deep neural networks (DNNs), to analyze MD data. MD output was transformed into the pixel representation, which is interpretable by the state-of-art DL object-recognition technology. When the method was applied to the pharmacological classification of 5-HT_{2A} and D₂ receptors ligands, among which were full, partial, and inverse agonists, DNN achieved near-perfect accuracy, classifying correctly >99% frames. Moreover, the sensitivity analysis identified the molecular determinants, which were considered by the model as the most important for the correct prediction. Even if the study has limited scope, including only eight ligands and two receptors, it gives hope for the highly accurate and efficient estimation of ligand-protein functional selectivity with the help of DNN.

Allostery is called the second secret of life (Fenton, 2008), as it is crucial for the adaptation of living organisms to changing environmental conditions by altering multiple cell functions, like enzyme catalysis, cell signalling, gene transcription, and others (Goodey and Benkovic, 2008; Nussinov et al., 2014). Designing

allosteric drugs is a challenging task for multiple reasons. First of all, classical docking alone is unable to predict how orthosteric binding sites would adjust to allosteric modulation, and, importantly, which functional effect ligands would exert on protein's function (Nussinov and Tsai, 2013; Lu et al., 2019; Sheik et al., 2020). Luckily, MD simulations give insight into the nature of allosteric perturbations; moreover, the application of ML algorithms to MD data expands possibilities to extract valuable information from long-scale simulations. Recently conducted research proved that such a combined MD-ML approach is able to efficiently determine ligand's functional activity and models explaining ligand efficacy can be constructed. Marchetti et al. (2021) brought together the benefits of ensemble docking, MD and ML, in order to predict whether a set of ligands would inhibit or activate molecular chaperone Hsp90. MD of Hsp90 with several ligands was followed by cluster analysis of the obtained metatrayjectory, subsequently, representative protein conformations were chosen for ensemble docking. The features obtained from docking, notably docking score, RMS, and RMSD, were used for training a supervised model, which served as a classification tool. Among three popular algorithms—logistic regression, SVM, and Random Forest - SVM reached the highest accuracy (0.9), as well as showed the best performance. On the other hand, attempts to classify ligands on the basis of separate features or chemometrics properties (here, molecular fingerprints) were far less efficient. In contrast, Ferraro et al. (2021) aimed to predict allosteric ligand functionality quantitatively. A computational experiment was performed on the allosteric modulators of the molecular chaperone TRAP1, which had similar affinities, but inhibited ATPase function with different efficacy. Two ML algorithms—Naïve Bayes and SVM—were applied to extract the local dynamic patterns responsible for the allosteric perturbation. The models were trained and validated on MD simulations of the perturbed and unperturbed systems. Whereas the discriminative SVM models qualitatively assessed the disparities between the perturbed and unperturbed ensembles, the implementation of the generative Naïve Bayes model produced a linear regression model with a 0.71 correlation between predicted states in the inhibitor-bound trajectories (TPR percentage) and the TRAP1 inhibition percentage. Additionally, Naïve Bayes could estimate the weight of ligand effects on each feature, which would support the identification of the features crucial for the allosteric propagation. Therefore, ML expands the possibilities of computer-aided drug design of allosteric modulators and could bring drug design to a new level with limited experimental testing.

The number of proteins with unknown functions is increasing due to the advances in bioinformatics, especially in the field of structural genomics. Identification of binding pockets could potentially be the key to understanding which functions specific proteins carry out. The FEATURE (Wei and Altman, 1998) is an ML-based algorithm for the identification of Ca²⁺-binding sites, utilizing the Bayesian scoring scheme. The FEATURE prediction does not depend on the sequence or structure, as the models examine local 3D physicochemical environment and that is why they are able to recognize

diverse binding sites. However, the applications of the algorithm were limited to static structures, until Glazer et al. (2008) applied MD to improve the FEATURE detection ability by increasing structural diversity. The hypothesis was tested on parvalbumin β – an EF-hand Ca^{2+} -binding protein, which has two Ca^{2+} -binding sites – and MD-assisted calcium-binding pockets recognition. Moreover, relatively small time steps were characterized by significant change in the FEATURE scores, meaning that the FEATURE is very sensitive to small conformational changes, which might have an impact on calcium binding. These promising results could help to implement MD methodology in the exploration of protein functions.

Researchers' efforts and technological advancement resulted in the development of a framework designed to support performing of MD simulations by means of ML algorithms – TorchMD (Doerr et al., 2021). Since the toolset is written in PyTorch (Paszke et al., 2019), it can be easily integrated with other models from this ML library. Among essential features of the framework is TorchMD-Net, which takes advantage of training neural network potential in order to improve force-field development. Furthermore, TorchMD enables running simulations with end-to-end differentiability of parameters, beneficial for the performance of steered and highly constrained MD simulations, sensitivity analysis, and others. Additionally, TorchMD with implemented neural network potential is used for coarse-grained MD simulations, which are helpful in studying protein folding and exploring conformational space. Code, step-by-step tutorials, and data are available at GitHub (<https://www.github.com/torchmd>).

CONCLUSION

Both intense growth in the amount of data, as well as increasing capabilities of various algorithms to detect patterns and relationships in various sets of information,

dramatically increased the popularity of automatic approaches for MD outcome analysis. The output of such experiments consists of billions of timesteps, and recorded positions and velocities of thousands of atoms. Therefore, extracting important information from such a data package can be very challenging, and so the application of various post-processing approaches is needed. The post-processing protocols can help in the finding of non-obvious ligand-protein interaction patterns, detection of rare conformational states, or examining dependence of conformational changes of the examined system in time. Moreover, thanks to the post-processing approaches, the prediction of the system behavior in longer time scales than modeled can be made.

However, given all the advantages of ML approaches, we should still be aware of their limitations and pay attention to data used for models training, as it will substantially define the quality of the outcome. Importantly, ML models could have limited transferability and must be applied to other types of data carefully. Nevertheless, application of ML to MD data is undoubtedly the future, which makes the potential of MD applications almost unlimited.

AUTHOR CONTRIBUTIONS

HB: literature search; preparation of the manuscript draft, review, editing, figures preparation; SP: literature search, preparation of the manuscript, review, editing, figures preparation, supervision.

FUNDING

The study was supported by the grant OPUS 2018/31/B/NZ2/00165 financed by the National Science Centre, Poland (www.ncn.gov.pl).

REFERENCES

- Abella, J. R., Antunes, D., Jackson, K., Lizée, G., Clementi, C., and Kavraki, L. E. (2020). Markov State Modeling Reveals Alternative Unbinding Pathways for Peptide-MHC Complexes. *Proc. Natl. Acad. Sci. U S A* 117 (48), 30610–30618. doi:10.1073/pnas.2007246117
- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* 1–2, 19–25. doi:10.1016/j.softx.2015.06.001
- Abramyan, T. M., Snyder, J. A., Thyparambil, A. A., Stuart, S. J., and Latour, R. A. (2016). Cluster Analysis of Molecular Simulation Trajectories for Systems where Both Conformation and Orientation of the Sampled States Are Important. *J. Comput. Chem.* 37 (21), 1973–1982. doi:10.1002/jcc.24416
- Acharya, A., Agarwal, R., Baker, M., Baudry, J., Bhowmik, D., Boehm, S., et al. (2020). Supercomputer-based Ensemble Docking Drug Discovery Pipeline with Application to COVID-19. *ChemRxiv* 60 (12), 5832–5352. doi:10.26434/chemrxiv.12725465.v1
- Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993). Essential Dynamics of Proteins. *Proteins* 17 (4), 412–425. doi:10.1002/prot.340170408
- Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J. A., Miao, Y., et al. (2018). Ensemble Docking in Drug Discovery. *Biophys. J.* 114 (10), 2271–2278. doi:10.1016/j.bpj.2018.02.038
- An, Y., Jessen, H. J., Wang, H., Shears, S. B., and Kireev, D. (2019). Dynamics of Substrate Processing by PPIP5K2, a Versatile Catalytic Machine. *Structure* 27 (6), 1022–e2. doi:10.1016/j.str.2019.03.007
- Andersen, B. F., Baker, H. M., Morris, G. E., Rumball, S. V., and Baker, E. N. (1990). Apolactoferrin Structure Demonstrates Ligand-Induced Conformational Change in Transferrins. *Nature* 344 (6268), 784–787. doi:10.1038/344784a0
- Andreucot, M. (2009). Molecular Dynamics Multidimensional Scaling. *Phys. Lett. A* 373 (23–24), 2001–2006. doi:10.1016/j.physleta.2009.04.007
- Antoniou, D., and Schwartz, S. D. (2011). Response to Comment on “Towards Identification of the Reaction Coordinate Directly from the Transition State Ensemble Using the Kernel PCA Method” by D. Antoniou and S. Schwartz, *J. Phys. Chem. B* 115, 2465–2469 (2011). *J. Phys. Chem. B* 115 (10), 12674–12675. doi:10.1021/jp207463g
- Araki, M., Matsumoto, S., Bekker, G. J., Isaka, Y., Sagae, Y., Kamiya, N., et al. (2021). Exploring Ligand Binding Pathways on Proteins Using Hypersound-Accelerated Molecular Dynamics. *Nat. Commun.* 12 (1), 2793. doi:10.1038/s41467-021-23157-1
- Ash, J., and Fourches, D. (2017). Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics

- Trajectories. *J. Chem. Inf. Model.* 57 (6), 1286–1299. doi:10.1021/acs.jcim.7b00048
- Bai, Q., Pérez-Sánchez, H., Zhang, Y., Shao, Y., Shi, D., Liu, H., et al. (2014). Ligand Induced Change of $\beta 2$ Adrenergic Receptor from Active to Inactive Conformation and its Implication for the Closed/open State of the Water Channel: Insight from Molecular Dynamics Simulation, Free Energy Calculation and Markov State Model Analysis. *Phys. Chem. Chem. Phys.* 16 (30), 15874–15885. doi:10.1039/c4cp01185f
- Ballester, P. J. (2019). Machine Learning for Molecular Modelling in Drug Design. *Biomolecules* 9 (6), 216. doi:10.3390/biom9060216
- Bansal, M., Kumar, S., and Velavan, R. (2000). HELANAL: a Program to Characterize helix Geometry in Proteins. *J. Biomol. Struct. Dyn.* 17 (5), 811–819. doi:10.1080/07391102.2000.10506570
- Barati Farimani, A., Feinberg, E., and Pande, V. (2018). Binding Pathway of Opiates to μ -Opioid Receptors Revealed by Machine Learning. *Biophysical J.* 114 (3), 62a–63a. doi:10.1016/j.bpj.2017.11.390
- Barletta, G. P., Franchini, G., Corsico, B., and Fernandez-Alberti, S. (2019). Fatty Acid and Retinol-Binding Protein: Unusual Protein Conformational and Cavity Changes Dictated by Ligand Fluctuations. *J. Chem. Inf. Model.* 59 (8), 3545–3555. doi:10.1021/acs.jcim.9b00364
- Baskin, I. I. (2020). The Power of Deep Learning to Ligand-Based Novel Drug Discovery. *Expert Opin. Drug Discov.* 15 (7), 755–764. doi:10.1080/17460441.2020.1745183
- Bekker, G. J., Araki, M., Oshima, K., Okuno, Y., and Kamiya, N. (2020). Exhaustive Search of the Configurational Space of Heat-Shock Protein 90 with its Inhibitor by Multicanonical Molecular Dynamics Based Dynamic Docking. *J. Comput. Chem.* 41 (17), 1606–1615. doi:10.1002/jcc.26203
- Bernetti, M., Masetti, M., Recanatini, M., Amaro, R. E., and Cavalli, A. (2019). An Integrated Markov State Model and Path Metadynamics Approach to Characterize Drug Binding Processes. *J. Chem. Theor. Comput.* 15 (10), 5689–5702. doi:10.1021/acs.jctc.9b00450
- Berrai, D. (2019). “Bayes’ Theorem and Naive Bayes Classifier,” in *Encyclopedia of Bioinformatics and Computational Biology*. Editors S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Wadern: Academic Press), 403–412. doi:10.1016/b978-0-12-809633-8.20473-1
- Bhakat, S., Martin, A. J., and Soliman, M. E. (2014). An Integrated Molecular Dynamics, Principal Component Analysis and Residue Interaction Network Approach Reveals the Impact of M184V Mutation on HIV Reverse Transcriptase Resistance to Lamivudine. *Mol. Biosyst.* 10 (8), 2215–2228. doi:10.1039/c4mb00253a
- Bhattacharai, A., Wang, J., and Miao, Y. (2020). Retrospective Ensemble Docking of Allsteric Modulators in an Adenosine G-Protein-Coupled Receptor. *Biochim. Biophys. Acta Gen. Subj.* 1864 (8), 129615. doi:10.1016/j.bbagen.2020.129615
- Binder, K., Horbach, J., Kob, W., Paul, W., and Varnik, F. (2004). Molecular Dynamics Simulations. *J. Phys. Condens. Matter* 16 (5), S429–S453. doi:10.1088/0953-8984/16/5/006
- Bosshard, H. R. (2001). Molecular Recognition by Induced Fit: How Fit Is the Concept. *News Physiol. Sci.* 16 (4), 171–173. doi:10.1152/physiologyonline.2001.16.4.171
- Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). “Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters,” in SC ’06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, 43. doi:10.1109/SC.2006.54
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. 1st ed. New York: Routledge.
- Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr, Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). CHARMM: the Biomolecular Simulation Program. *J. Comput. Chem.* 30 (10), 1545–1614. doi:10.1002/jcc.21287
- Bruno, A., Beato, C., and Costantino, G. (2011). Molecular Dynamics Simulations and Docking Studies on 3D Models of the Heterodimeric and Homodimeric 5-HT_{2A} Receptor Subtype. *Future Med. Chem.* 3 (6), 665–681. doi:10.4155/fmc.11.27
- Buch, I., Giorino, T., and de Fabritiis, G. (2011). Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U S A.* 108, 10184–10189. doi:10.1073/pnas.1103547108
- Carpenter, K. A., and Huang, X., (2018). Machine Learning-Based Virtual Screening and its Applications to Alzheimer’s Drug Discovery: A Review. *Curr. Pharm. Des.* 24 (28), 3347–3358. doi:10.2174/1381612824666180607124038
- Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P., and Parrinello, M. (2017). Unbinding Kinetics of a P38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *J. Am. Chem. Soc.* 139 (13), 4780–4788. doi:10.1021/jacs.6b12950
- Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., and Merz, K. M., Jr. (2005). The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* 26(16), 1668. doi:10.1002/jcc.20290
- Case, D. A., Aktulga, H. M., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., et al. (2021). *Amber 2021*. San Francisco: University of California.
- Casoni, A., Clerici, F., and Contini, A. (2013). Molecular Dynamic Simulation of mGluR5 Amino Terminal Domain: Essential Dynamics Analysis Captures the Agonist or Antagonist Behaviour of Ligands. *J. Mol. Graph. Model.* 41, 72–78. doi:10.1016/j.jmgm.2013.02.002
- Chaturvedi, N., Yadav, B. S., Pandey, P. N., and Tripathi, V. (2017). The Effect of β -glucan and its Potential Analog on the Structure of Dectin-1 Receptor. *J. Mol. Graph. Model.* 74, 315–325. doi:10.1016/j.jmgm.2017.04.014
- Chen, J. (2018). Functional Roles of Magnesium Binding to Extracellular Signal-Regulated Kinase 2 Explored by Molecular Dynamics Simulations and Principal Component Analysis. *J. Biomol. Struct. Dyn.* 36 (2), 351–361. doi:10.1080/07391102.2016.1277783
- Chiappori, F., Merelli, I., Milanese, L., and Rovida, E. (2010). Exploring the Role of the Phospholipid Ligand in Endothelial Protein C Receptor: a Molecular Dynamics Study. *Proteins* 78 (12), 2679–2690. doi:10.1002/prot.22782
- Chodera, J. D., and Noé, F. (2014). Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Struct. Biol.* 25, 135–144. doi:10.1016/j.sbi.2014.04.002
- Cholko, T., Chen, W., Tang, Z., and Chang, C. A. (2018). A Molecular Dynamics Investigation of CDK8/CycC and Ligand Binding: Conformational Flexibility and Implication in Drug Discovery. *J. Comput. Aided Mol. Des.* 32 (6), 671–685. doi:10.1007/s10822-018-0120-3
- Coifman, R. R., and Lafon, S. (2006). Diffusion Maps. *Appl. Comput. Harmon. Anal.* 21 (1), 5–30. doi:10.1016/j.acha.2006.04.006
- Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., et al. (2005). Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps. *Proc. Natl. Acad. Sci. U. S. A.* 102 (21), 7426–7431. doi:10.1073/pnas.0500334102
- Cortes, C., and Vapnik, V. N. (1995). Support-vector Networks. *Mach. Learn.* 20 (3), 273–297. doi:10.1007/BF00994018
- Cossio-Pérez, R., Palma, J., and Pierdominici-Sottile, G. (2017). Consistent Principal Component Modes from Molecular Dynamics Simulations of Proteins. *J. Chem. Inf. Model.* 57 (4), 826–834. doi:10.1021/acs.jcim.6b00646
- Cover, T. M., and Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theor.* 13 (1), 21–27. doi:10.1109/TIT.1967.1053964
- CPPTRAJ (2021). CPPTRAJ Wiki. Available at: <https://github.com/Amber-MD/cpptraj/wiki> (Accessed December 22, 2021).
- Das, A., and Mukhopadhyay, C. (2007). Application of Principal Component Analysis in Protein Unfolding: an All-Atom Molecular Dynamics Simulation Study. *J. Chem. Phys.* 127 (16), 165103. doi:10.1063/1.2796165
- David, C. C., Avery, C. S., and Jacobs, D. J. (2021). JEDI: Java Essential Dynamics Inspector - a Molecular Trajectory Analysis Toolkit. *BMC Bioinform* 22 (1), 226. doi:10.1186/s12859-021-04140-5
- Davies, D. L., and Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-1 (2), 224–227. doi:10.1109/TPAMI.1979.4766909
- De Paris, R., Quevedo, C. V., Ruiz, D. D. A., and Norberto de Souza, O. (2015a). An Effective Approach for Clustering InhA Molecular Dynamics Trajectory Using Substrate-Binding Cavity Features. *PLoS ONE* 10 (7), e0133172. doi:10.1371/journal.pone.0133172
- De Paris, R., Quevedo, C. V., Ruiz, D. D., Norberto de Souza, O., and Barros, R. C. (2015b). Clustering Molecular Dynamics Trajectories for Optimizing Docking Experiments. *Comput. Intell. Neurosci.* 2015, 916240. doi:10.1155/2015/916240
- de Souto, M. C. P., Coelho, A. L. V., Faceli, K., Sakata, T. C., Bonadia, V., and Costa, I. G. (2012). A Comparison of External Clustering Evaluation Indices in the Context of Imbalanced Data Sets. *Braz. Symp. Neural Networks* 2012, 49–54. doi:10.1109/SBRN.2012.25
- de Souza, V. C., Golliat, L., and Goliati, P. V. Z. C. (2017). “Clustering Algorithms Applied on Analysis of Protein Molecular Dynamics,” in Proceedings of the 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 1–6.

- Dickson, A. (2018). Mapping the Ligand Binding Landscape. *Biophys. J.* 115 (9), 1707–1719. doi:10.1016/j.bpj.2018.09.021
- Dixon, T., Uyar, A., Ferguson-Miller, S., and Dickson, A. (2021). Membrane-Mediated Ligand Unbinding of the PK-11195 Ligand from TSPO. *Biophys. J.* 120 (1), 158–167. doi:10.1016/j.bpj.2020.11.015
- Desmond Molecular Dynamics System (2021). Maestro-Desmond Interoperability Tools. New York, NY: Schrödinger.
- Doerr, S., Majewski, M., Pérez, A., Krämer, A., Clementi, C., Noe, F., et al. (2021). TorchMD: A Deep Learning Framework for Molecular Simulations. *J. Chem. Theor. Comput.* 17 (4), 2355–2363. doi:10.1021/acs.jctc.0c01343
- Doms, R. W., Helenius, A., and White, J. (1985). Membrane Fusion Activity of the Influenza Virus Hemagglutinin. The Low pH-Induced Conformational Change. *J. Biol. Chem.* 260 (5), 2973–2981. doi:10.1016/s0021-9258(18)89461-3
- Dutta, S., and Bose, K. (2021). Remodelling Structure-Based Drug Design Using Machine Learning. *Emerg. Top. Life Sci.* 5 (1), 13–27. doi:10.1042/ETLS20200253
- Ellingson, S. R., Miao, Y., Baudry, J., and Smith, J. C. (2015). Multi-conformer Ensemble Docking to Difficult Protein Targets. *J. Phys. Chem. B.* 119 (3), 1026–1034. doi:10.1021/jp506511p
- Ernst, M., Sittel, F., and Stock, G. (2015). Contact- and Distance-Based Principal Component Analysis of Protein Dynamics. *J. Chem. Phys.* 143 (24), 244114. doi:10.1063/1.4938249
- Ester, M., Kriegl, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Kdd-96 Proc.* 96, 226–231. doi:10.5555/3001460.3001507
- Evangelista, F. W., Ellingson, S. R., Smith, J. C., and Baudry, J. (2019). Ensemble Docking in Drug Discovery: How Many Protein Configurations from Molecular Dynamics Simulations Are Needed to Reproduce Known Ligand Binding. *J. Phys. Chem. B.* 123 (25), 5189–5195. doi:10.1021/acs.jpcc.8b11491
- Fakhar, Z., Govender, T., Maguire, G. E. M., Lamichane, G., Walker, R. C., Kruger, H. G., et al. (2017). Differential Flap Dynamics in Ld-Transpeptidase2 from mycobacterium Tuberculosis Revealed by Molecular Dynamics. *Mol. Biosyst.* 13 (6), 1223–1234. doi:10.1039/c7mb00110j
- Feig, M., Karanicolas, J., and Brooks, C. L., 3rd (2004). MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graph. Model.* 22 (5), 377–395. doi:10.1016/j.jmgl.2003.12.005
- Fenton, A. W. (2008). Allostery: an Illustrated Definition for the ‘second Secret of Life. *Trends Biochem. Sci.* 33 (9), 420–425. doi:10.1016/j.tibs.2008.05.009
- Ferraro, M., Moroni, E., Ippoliti, E., Rinaldi, S., Sanchez-Martin, C., Rasola, A., et al. (2021). Machine Learning of Allosteric Effects: the Analysis of Ligand-Induced Dynamics to Predict Functional Effects in TRAP1. *J. Phys. Chem. B.* 125 (1), 101–114. doi:10.1021/acs.jpcc.0c09742
- Ferreira, L. G., Dos Santos, R. N., Oliva, G., and Andricopulo, A. D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* 20 (7), 13384–13421. doi:10.3390/molecules200713384
- Fraccalvieri, D., Bonati, L., and Stella, F. (2013). “Self Organizing Maps to Efficiently Cluster and Functionally Interpret Protein Conformational Ensembles,” in In Proceedings Wivace. arXiv:1309.7122. doi:10.4204/eptcs.130.13
- Fu, G., Sivaprakasam, P., Dale, O. R., Manly, S. P., Cutler, S. J., and Doerksen, R. J. (2014). Pharmacophore Modeling, Ensemble Docking, Virtual Screening, and Biological Evaluation on Glycogen Synthase Kinase-3 β . *Mol. Inform.* 33 (9), 610–626. doi:10.1002/minf.201400044
- Gebhardt, J., Kiesel, M., Riniker, S., and Hansen, N. (2020). Combining Molecular Dynamics and Machine Learning to Predict Self-Solvation Free Energies and Limiting Activity Coefficients. *J. Chem. Inf. Model.* 60 (11), 5319–5330. doi:10.1021/acs.jcim.0c00479
- Girdhar, K., Dehury, B., Kumar, S. M., Daniel, V. P., Choubey, A., Dogra, S., et al. (2019). Novel Insights into the Dynamics Behavior of Glucagon-like Peptide-1 Receptor with its Small Molecule Agonists. *J. Biomol. Struct. Dyn.* 37 (15), 3976–3986. doi:10.1080/07391102.2018.1532818
- Glaser, J., Nguyen, T. D., Anderson, J. A., Lui, P., Spiga, F., Millan, J. A., et al. (2015). Strong Scaling of General-Purpose Molecular Dynamics Simulations on GPUs. *Comput. Phys. Commun.* 192, 97–107. doi:10.1016/j.cpc.2015.02.028
- Glazer, D. S., Radmer, R. J., and Altman, R. B. (2008). Combining Molecular Dynamics and Machine Learning to Improve Protein Function Recognition. *Pac. Symp. Biocomput.* 13, 332–343. doi:10.1249/jsr.0b013e31818f03c5
- Glenn, T. C., Zare, A., and Gader, P. D. (2015). Bayesian Fuzzy Clustering. *IEEE Trans. Fuzzy Syst.* 23 (5), 1545–1561. doi:10.1109/TFUZZ.2014.2370676
- Glielmo, A., Husic, B. E., Rodriguez, A., Clementi, C., Noé, F., and Laio, A. (2021). Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* 121 (16), 9722–9758. doi:10.1021/acs.chemrev.0c01195
- Glykos, N. M. (2006). Software News and Updates Carma: A Molecular Dynamics Analysis Program. *J. Comput. Chem.* 27 (14), 1765–1768. doi:10.1002/jcc.20482
- Göller, A. H., Kuhnke, L., Montanari, F., Bonin, A., Schneckener, S., Ter Laak, A., et al. (2020). Bayer's In Silico ADMET Platform: a Journey of Machine Learning over the Past Two Decades. *Drug Discov. Today* 25 (9), 1702–1709. doi:10.1016/j.drudis.2020.07.001
- Göller, A. H., Kuhnke, L., Ter Laak, A., Meier, K., and Hillisch, A. (2022). Machine Learning Applied to the Modeling of Pharmacological and ADMET Endpoints. *Methods Mol. Biol.* 2390, 61–101. doi:10.1007/978-1-0716-1787-8_2
- Goodey, N. M., and Benkovic, S. J. (2008). Allosteric Regulation and Catalysis Emerge via a Common Route. *Nat. Chem. Biol.* 4 (8), 474–482. doi:10.1038/nchembio.98
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* 27, 2672–2680. doi:10.3156/jsoft.29.5_177_2
- Gordon, H. L., and Somorjai, R. L. (1992). Fuzzy Cluster Analysis of Molecular Dynamics Trajectories. *Proteins* 14 (2), 249–264. doi:10.1002/prot.340140211
- Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006). Bio3d: an R Package for the Comparative Analysis of Protein Structures. *Bioinformatics* 22 (21), 2695–2696. doi:10.1093/bioinformatics/btl461
- Grossfield, A., and Romo, T. D. (2021). Loos, a Better Tool to Analyze Molecular Dynamics Simulations. *Biophys. J.* 120 (3), 178a. doi:10.1016/j.bpj.2020.11.1245
- Grossfield, A., and Zuckerman, D. M. (2009). Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. *Annu. Rep. Comput. Chem.* 5, 23–48. doi:10.1016/S1574-1400(09)00502-7
- Guedes, I. A., de Magalhães, C. S., and Dardenne, L. E. (2014). Receptor–ligand Molecular Docking. *Biophys. Rev.* 6 (1), 75–87. doi:10.1007/s12551-013-0130-2
- Hall, P., Park, B. U., and Samworth, R. J. (2008). Choice of Neighbor Order in Nearest-Neighbor Classification. *Ann. Stat.* 36 (5), 2135–2152. doi:10.1214/07-AOS537
- Haque, I. S., Beauchamp, K. A., and Pande, V. S. (2014). A Fast $3 \times N$ Matrix Multiply Routine for Calculation of Protein RMSD. *Biorxiv* 8631. doi:10.1101/008631
- Hartigan, A., and Wong, M. A. (1979). A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* 28 (1), 100–108. doi:10.2307/2346830
- Hernández, C. X., Wayment-Steele, H. K., Sultan, M. M., Husic, B. E., and Pande, V. S. (2018). Variational Encoding of Complex Dynamics. *Phys. Rev. E* 97 (6-1), 062412. doi:10.1103/PhysRevE.97.062412
- Hilger, D., Masureel, M., and Kobilka, B. K. (2018). Structure and Dynamics of GPCR Signaling Complexes. *Nat. Struct. Mol. Biol.* 25 (1), 4–12. doi:10.1038/s41594-017-0011-7
- Hinsen, K. (2000). The Molecular Modeling Toolkit: a New Approach to Molecular Simulations. *J. Comput. Chem.* 21 (2), 79–85. doi:10.1002/(SICI)1096-987X(20000130)21:2<79::AID-JCC1>3.0.CO;2-B
- Hollingsworth, S. A., and Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron* 99 (6), 1129–1143. doi:10.1016/j.neuron.2018.08.011
- Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79 (8), 2554–2558. doi:10.1073/pnas.79.8.2554
- Hopkins, B., and Skellam, J. G. (1954). A New Method for Determining the Type of Distribution of Plant Individuals. *Ann. Bot.* 18 (2), 213–227. doi:10.1093/oxfordjournals.aob.a083391
- Huang, D., and Cafisch, A. (2011). The Free Energy Landscape of Small Molecule Unbinding. *Plos Comput. Biol.* 7 (2), e1002002. doi:10.1371/journal.pcbi.1002002
- Huang, Z. (1998). Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.* 2 (3), 283–304. doi:10.1023/A:1009769707641
- Hudson, I. L. (2021). Data Integration Using Advances in Machine Learning in Drug Discovery and Molecular Biology. *Methods Mol. Biol.* 2190, 167–184. doi:10.1007/978-1-0716-0826-5_7

- Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graph.* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5
- Husic, B. E., and Pande, V. S. (2018). Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* 140 (7), 2386–2396. doi:10.1021/jacs.7b12191
- Hussain, W., Rasool, N., and Khan, Y. D. (2021). Insights into Machine Learning-Based Approaches for Virtual Screening in Drug Discovery: Existing Strategies and Streamlining through FP-CADD. *Curr. Drug Discov. Technol.* 18 (4), 463–472. doi:10.2174/1570163817666200806165934
- Hyvönen, M. T., Hiltunen, Y., El-Dereby, W., Ojala, T., Vaara, J., Kovanen, P. T., et al. (2001). Application of Self-Organizing Maps in Conformational Analysis of Lipids. *J. Am. Chem. Soc.* 123 (5), 810–816. doi:10.1021/ja0025853
- Ichiye, T., and Karplus, M. (1991). Collective Motions in Proteins: a Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and normal Mode Simulations. *Proteins* 11 (3), 205–217. doi:10.1002/prot.340110305
- Jamal, S., Grover, A., and Grover, S. (2019). Machine Learning from Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors against Alzheimer's Disease. *Front. Pharmacol.* 10, 780. doi:10.3389/fphar.2019.00780
- Jeong, J. C., Jo, S., Wu, E. L., Qi, Y., Monje-Galvan, V., Yeom, M. S., et al. (2014). ST-analyzer: A Web-based User Interface for Simulation Trajectory Analysis. *J. Comput. Chem.* 35 (12), 957–963. doi:10.1002/jcc.23584
- Jia, L., and Gao, H. (2022). Machine Learning for In Silico ADMET Prediction. *Methods Mol. Biol.* 2390, 447–460. doi:10.1007/978-1-0716-1787-8_20
- Jolliffe, I. T., and Cadima, J. (2016). Principal Component Analysis: a Review and Recent Developments. *Phil. Trans. R. Soc. A* 374 (2065), 20150202. doi:10.1098/rsta.2015.0202
- Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer-Verlag.
- Kabsch, W., and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. *Biopolym. Orig. Res. Biomol.* 22 (12), 2577–2637. doi:10.1002/bip.360221211
- Karnati, K. R., and Wang, Y. (2019). Structural and Binding Insights into HIV-1 Protease and P2-Ligand Interactions through Molecular Dynamics Simulations, Binding Free Energy and Principal Component Analysis. *J. Mol. Graph. Model.* 92, 112–122. doi:10.1016/j.jmgm.2019.07.008
- Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons.
- Keller, B., Daura, X., and van Gunsteren, W. F. (2010). Comparing Geometric and Kinetic Cluster Algorithms for Molecular Simulation Data. *J. Chem. Phys.* 132 (7), 074110. doi:10.1063/1.3301140
- Khamis, M. A., Gomaa, W., and Ahmed, W. F. (2015). Machine Learning in Computational Docking. *Artif. Intell. Med.* 63 (3), 135–152. doi:10.1016/j.artmed.2015.02.002
- Khamis, M. A., and Gomaa, W. (2015). Comparative Assessment of Machine-Learning Scoring Functions on PDBbind 2013. *Eng. Appl. Artif. Intell.* 45 (C), 136–151. doi:10.1016/j.engappai.2015.06.021
- Khamis, M., Gomaa, W., and Galal, B. (2016). *Deep Learning Is Competing Random forest in Computational Docking*. New Borg El-Arab City, Egypt: arXiv: 1608.06665.
- Kim, H. J., Choi, M. Y., Kim, H. J., and Llinás, M. (2010). Conformational Dynamics and Ligand Binding in the Multi-Domain Protein PDC109. *PLoS One* 5 (2), e9180. doi:10.1371/journal.pone.0009180
- Koukos, P. I., and Glykos, N. M. (2013). Grcarma: a Fully Automated Task-oriented Interface for the Analysis of Molecular Dynamics Trajectories. *J. Comput. Chem.* 34 (26), 2310–2312. doi:10.1002/jcc.23381
- Kramer, M. A. (1991). Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *Aiche J.* 37, 233–243. doi:10.1002/aic.690370209
- Kucwaj-Brysz, K., Dela, A., Podlowska, S., Bednarski, M., Siwek, A., Satała, G., et al. (2021). The Structural Determinants for α 1-adrenergic/serotonin Receptors Activity Among Phenylpiperazine-Hydantoin Derivatives. *Molecules* 26 (22), 7025. doi:10.3390/molecules26227025
- Lagardère, L., Jolly, L.-H., Lipparini, F., Aviat, F., Stamm, B., Jing, Z. F., et al. (2018). Tinker-HP: A Massively Parallel Molecular Dynamics Package for Multiscale Simulations of Large Complex Systems with Advanced Point Dipole Polarizable Force Fields. *Chem. Sci.* 9 (4), 956–972. doi:10.1039/c7sc04531j
- Lange, O. F., and Grubmüller, H. (2006). Can Principal Components Yield a Dimension Reduced Description of Protein Dynamics on Long Time Scales? *J. Phys. Chem. B* 110 (45), 22842–22852. doi:10.1021/jp062548j
- Laxmi, D., and Priyadarshy, S. (2002). HyperChem 6.03. *Biotech. Softw. Internet Rep.* 3 (1), 5–9. doi:10.1089/152791602317250351
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539
- Leimkuhler, B., and Matthews, C. (2016). *Molecular Dynamics*. Switzerland: Springer.
- Lindahl, E. R. (2008). Molecular Dynamics Simulations. *Methods Mol. Biol.* 443, 3–23. doi:10.1007/978-1-59745-177-2_1
- Lipiński, P. F. J., Jarończyk, M., Dobrowolski, J. C., and Sadlej, J. (2019). Molecular Dynamics of Fentanyl Bound to μ -opioid Receptor. *J. Mol. Model.* 25 (5), 144. doi:10.1007/s00894-019-3999-2
- Lopez, R., Regier, J., Jordan, M. I., and Yosef, N. (2018). Information Constraints on Auto-Encoding Variational Bayes. *Adv. Neural Inf. Process. Syst.* 31, 6114–6125.
- Lu, S., Shen, Q., and Zhang, J. (2019). Allosteric Methods and Their Applications: Facilitating the Discovery of Allosteric Drugs and the Investigation of Allosteric Mechanisms. *Acc. Chem. Res.* 52 (2), 492–500. doi:10.1021/acs.accounts.8b00570
- Magudeeswaran, S., and Poomani, K. (2020). Binding Mechanism of Spinosine and Venenatine Molecules with P300 HAT Enzyme: Molecular Screening, Molecular Dynamics and Free-Energy Analysis. *J. Cel. Biochem.* 121 (2), 1759–1777. doi:10.1002/jcb.29412
- Majumder, S., and Giri, K. (2021). An Insight into the Binding Mechanism of Viprinin and its Morpholine and Piperidine Derivatives with HIV-1 VPR: Molecular Dynamics Simulation, Principal Component Analysis and Binding Free Energy Calculation Study. *J. Biomol. Struct. Dyn.*, 1–13. doi:10.1080/07391102.2021.1954553
- Mallet, V., Nilges, M., and Bouvier, G. (2021). Quicksom: Self-Organizing Maps on GPUs for Clustering of Molecular Dynamics Trajectories. *Bioinformatics* 37 (14), 2064–2065. doi:10.1093/bioinformatics/btaa925
- Marchetti, F., Moroni, E., Pandini, A., and Colombo, G. (2021). Machine Learning Prediction of Allosteric Drug Activity from Molecular Dynamics. *J. Phys. Chem. Lett.* 12 (15), 3724–3732. doi:10.1021/acs.jpclett.1c00045
- Martínez-Archundia, M., Correa-Basurto, J., Montaña, S., and Rosas-Trigueros, J. L. (2019). Studying the Collective Motions of the Adenosine A2A Receptor as a Result of Ligand Binding Using Principal Component Analysis. *J. Biomol. Struct. Dyn.* 37 (18), 4685–4700. doi:10.1080/07391102.2018.1564700
- McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., et al. (2015). MDTraj: a Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* 109 (8), 1528–1532. doi:10.1016/j.bpj.2015.08.015
- Melville, J. L., Burke, E. K., and Hirst, J. D. (2009). Machine Learning in Virtual Screening. *Comb. Chem. High Throughput Screen.* 12 (4), 332–343. doi:10.2174/138620709788167980
- Mezei, M. (2010). Simulaid: a Simulation Facilitator and Analysis Program. *J. Comput. Chem.* 31 (14), 2658–2668. doi:10.1002/jcc.21551
- Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDAnalysis: a Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* 32 (10), 2319–2327. doi:10.1002/jcc.21787
- Miyashita, N., and Yonezawa, Y. (2017). On-the-fly Analysis of Molecular Dynamics Simulation Trajectories of Proteins Using the Bayesian Inference Method. *J. Chem. Phys.* 147 (12), 124108. doi:10.1063/1.4997099
- Molecular Operating Environment (MOE) (2020). Chemical Computing Group ULC. Montreal, QC, Canada.
- Moritsugu, K. (2021). Multiscale Enhanced Sampling Using Machine Learning. *Life (Basel)* 11 (10), 1076. doi:10.3390/life11101076
- Morris, G. M., and Lim-Wilby, M. (2008). Molecular Docking. *Methods Mol. Biol.* 443, 365–382. doi:10.1007/978-1-59745-177-2_19
- Ng, H. W., Laughton, C. A., and Doughty, S. W. (2013). Molecular Dynamics Simulations of the Adenosine A2a Receptor: Structural Stability, Sampling, and Convergence. *J. Chem. Inf. Model.* 53 (5), 1168–1178. doi:10.1021/ci300610w
- Novikov, G. V., Sivozhelezov, V. S., and Shaitan, K. V. (2013). Study of Structural Dynamics of Ligand-Activated Membrane Receptors by Means of Principal Component Analysis. *Biochemistry (Mosc)* 78 (4), 403–411. doi:10.1134/S0006297913040093
- Nussinov, R., and Tsai, C.-J. (2013). Allostery in Disease and in Drug Discovery. *Cell* 153 (2), 293–305. doi:10.1016/j.cell.2013.03.034
- Nussinov, R., Tsai, C.-J., and Liu, J. (2014). Principles of Allosteric Interactions in Cell Signaling. *J. Am. Chem. Soc.* 136 (51), 17692–17701. doi:10.1021/ja510028c
- Pande, V. S., Beauchamp, K., and Bowman, G. R. (2010). Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* 52 (1), 99–105. doi:10.1016/j.ymeth.2010.06.002

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*. Editors H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver: Curran Associates, Inc.), 8024–8035. 32.
- Patel, L., Shukla, T., Huang, X., Ussery, D. W., and Wang, S. (2020). Machine Learning Methods in Drug Discovery. *Molecules* 25 (22), 5277. doi:10.3390/molecules25225277
- Perez, A., MacCallum, J. L., and Dill, K. A. (2015). Accelerating Molecular Simulations of Proteins Using Bayesian Inference on Weak Information. *PNAS* 112 (38), 11846–11851. doi:10.1073/pnas.1515561112
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* 26 (16), 1781–1802. doi:10.1002/jcc.20289
- Pietrucci, F., Marinelli, F., Carloni, P., and Laio, A. (2009). Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J. Am. Chem. Soc.* 131 (33), 11811–11818. doi:10.1021/ja903045y
- Pisani, P., Caporuscio, F., Carlini, L., and Rastelli, G. (2016). Molecular Dynamics Simulations and Classical Multidimensional Scaling Unveil New Metastable States in the Conformational Landscape of CDK2. *PLoS One* 11 (4), e0154066. doi:10.1371/journal.pone.0154066
- Plante, A., Shore, D. M., Morra, G., Khelashvili, G., and Weinstein, H. (2019). A Machine Learning Approach for the Discovery of Ligand-Specific Functional Mechanisms of GPCRs. *Molecules* 24 (11), 2097. doi:10.3390/molecules24112097
- Plattner, N., and Noé, F. (2015). Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* 6, 7653. doi:10.1038/ncomms8653
- Podlowska, S., Latacz, G., Łażewska, D., Kieć-Kononowicz, K., and Handzlik, J. (2020). In Silico and In Vitro Studies on Interaction of Novel Non-Imidazole Histamine H3R Antagonists with CYP3A4. *Bioorg. Med. Chem. Lett.* 30 (11), 127147. doi:10.1016/j.bmcl.2020.127147
- Preto, J., and Clementi, C. (2014). Fast Recovery of Free Energy Landscapes via Diffusion-Map-Directed Molecular Dynamics. *Phys. Chem. Chem. Phys.* 16 (36), 19181–19191. doi:10.1039/c3cp54520b
- Quinlan, J. R. (1986). Induction of Decision Trees. *Mach. Learn.* 1, 81–106. doi:10.1007/BF00116251
- Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* 66 (336), 846–850. doi:10.1080/01621459.1971.10482356
- Riniker, S. (2017). Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data to Predict Free-Energy Differences. *J. Chem. Inf. Model.* 57 (4), 726–741. doi:10.1021/acs.jcim.6b00778
- Roe, D. R., and Cheatham, T. E., 3rd (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theor.* 9 (7), 3084–3095. doi:10.1021/ct400341p
- Rohrdanz, M. A., Zheng, W., Maggioni, M., and Clementi, C. (2011). Determination of Reaction Coordinates via Locally Scaled Diffusion Map. *J. Chem. Phys.* 134 (12), 124116. doi:10.1063/1.3569857
- Roither, B., Oostenbrink, C., and Schreiner, W. (2020). Molecular Dynamics of the Immune Checkpoint Programmed Cell Death Protein 1, PD-1: Conformational Changes of the BC-Loop upon Binding of the Ligand PD-L1 and the Monoclonal Antibody Nivolumab. *BMC Bioinform.* 21 (Suppl. 17), 557. doi:10.1186/s12859-020-03904-9
- Romo, T. D., Leioatts, N., and Grossfield, A. (2014). Lightweight Object Oriented Structure Analysis: Tools for Building Tools to Analyze Molecular Dynamics Simulations. *J. Comput. Chem.* 35 (32), 2305–2318. doi:10.1002/jcc.23753
- Rudling, A., Orro, A., and Carlsson, J. (2018). Prediction of Ordered Water Molecules in Protein Binding Sites from Molecular Dynamics Simulations: The Impact of Ligand Binding on Hydration Networks. *J. Chem. Inf. Model.* 58 (2), 350–361. doi:10.1021/acs.jcim.7b00520
- Sander, J. (2011). in *Density-Based Clustering* in *Encyclopedia of Machine Learning*. Editors C. Sammut and G. I. Webb (Boston: Springer).
- Santos, L. H. S., Ferreira, R. S., and Caffarena, E. R. (2019). Integrating Molecular Docking and Molecular Dynamics Simulations. *Methods Mol. Biol.* 2053, 13–34. doi:10.1007/978-1-4939-9752-7_2
- Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Netw.* 61, 85–117. doi:10.1016/j.neunet.2014.09.003
- Schölkopf, B., Smola, A., and Müller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* 10 (5), 1299–1319. doi:10.1162/089976698300017467
- Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* 42 (3), 19. doi:10.1145/3068335
- Seeber, M., Cecchini, M., Rao, F., Settanni, G., and Cafilisch, A. (2007). Wordom: A Program for Efficient Analysis of Molecular Dynamics Simulations. *Bioinformatics* 23 (19), 2625–2627. doi:10.1093/bioinformatics/btm378
- Seeber, M., Felline, A., Raimondi, F., Muff, S., Friedman, R., Rao, F., et al. (2011). Wordom: A User-friendly Program for the Analysis of Molecular Structures, Trajectories, and Free Energy Surfaces. *J. Comput. Chem.* 32 (6), 1183–1194. doi:10.1002/jcc.21688
- Shamsi, Z., Cheng, K. J., and Shukla, D. (2018). Reinforcement Learning Based Adaptive Sampling: REAPing Rewards by Exploring Protein Conformational Landscapes. *J. Phys. Chem. B* 122 (35), 8386–8395. doi:10.1021/acs.jpcc.8b06521
- Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E. (2007). Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theor. Comput.* 3 (6), 2312–2334. doi:10.1021/ct700119m
- Sheik, A. O., Veldman, W., Manyumwa, C., Khairallah, A., Agajanian, S., Oluyemi, O., et al. (2020). Integrated Computational Approaches and Tools for Allosteric Drug Discovery. *Int. J. Mol. Sci.* 21 (3), 847. doi:10.3390/ijms21030847
- Shevchuk, R., and Hub, J. S. (2017). Bayesian Refinement of Protein Structures and Ensembles against SAXS Data Using Molecular Dynamics. *PLOS Comput. Biol.* 13 (10), e1005800. doi:10.1371/journal.pcbi.1005800
- Singh, J., Deng, Z., Narale, G., and Chuaqui, C. (2006). Structural Interaction Fingerprints: a New Approach to Organizing, Mining, Analyzing, and Designing Protein-Small Molecule Complexes. *Chem. Biol. Drug Des.* 67 (1), 5–12. doi:10.1111/j.1747-0285.2005.00323.x
- Sittel, F., Jain, A., and Stock, G. (2014). Principal Component Analysis of Molecular Dynamics: on the Use of Cartesian vs. Internal Coordinates. *J. Chem. Phys.* 141 (1), 014111. doi:10.1063/1.4885338
- Spiwok, V., and Kříž, P. (2020). Time-Lagged T-Distributed Stochastic Neighbor Embedding (T-SNE) of Molecular Simulation Trajectories. *Front. Mol. Biosci.* 7, 132. doi:10.3389/fmolb.2020.00132
- Srivastava, A., Bala, S., Motomura, H., Kohda, D., Tama, F., and Miyashita, O. (2020). Conformational Ensemble of an Intrinsically Flexible Loop in Mitochondrial Import Protein Tim21 Studied by Modeling and Molecular Dynamics Simulations. *Biochim. Biophys. Acta Gen. Subj.* 1864 (2), 129417. doi:10.1016/j.bbagen.2019.129417
- Stamati, H., Clementi, C., and Kavraki, L. E. (2010). Application of Nonlinear Dimensionality Reduction to Characterize the Conformational Landscape of Small Peptides. *Proteins* 78 (2), 223–235. doi:10.1002/prot.22526
- Stelzl, L. S., Fowler, P. W., Sansom, M. S. P., and Beckstein, O. (2014). Flexible gates Generate Occluded Intermediates in the Transport Cycle of LacY. *J. Mol. Biol.* 426 (3), 735–751. doi:10.1016/j.jmb.2013.10.024
- Sugeta, H., and Miyazawa, T. (1967). General Method for Calculating Helical Parameters of Polymer Chains from Bond Lengths, Bond Angles, and Internal-Rotation Angles. *Biopolym. Orig. Res. Biomol.* 5 (7), 673–679. doi:10.1002/BIP.1967.360050708
- Sutmann, G. (2002). "Classical Molecular Dynamics," in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*. Editors J. Grotendorst, D. Marx, and A. Muramatsu (Jülich: John von Neumann Institute for Computing), 211–254.
- Takeda, K., Wada, A., Yamamoto, K., Moriyama, Y., and Aoki, K. (1989). Conformational Change of Bovine Serum Albumin by Heat Treatment. *J. Protein Chem.* 8 (5), 653–659. doi:10.1007/BF01025605
- Takemura, K., Sato, C., and Kitao, A. (2018). ColDock: Concentrated Ligand Docking with All-Atom Molecular Dynamics Simulation. *J. Phys. Chem. B* 122 (29), 7191–7200. doi:10.1021/acs.jpcc.8b02756
- Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 2319–2323. doi:10.1126/science.290.5500.2319
- Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., et al. (2021). LAMMPS-A Flexible Simulation Tool for

- Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Comput. Phys. Commun.* 271, 108171. doi:10.1016/j.cpc.2021.108171
- Tian, H., Jiang, X., Trozzi, F., Xiao, S., Larson, E. C., and Tao, P. (2021). Explore Protein Conformational Space with Variational Autoencoder. *Front. Mol. Biosci.* 8, 781635. doi:10.3389/fmolb.2021.781635
- Todorov, I. T., Smith, W., Trachenko, K., and Dove, M. T. (2006). DL_POLY_3: New Dimensions in Molecular Dynamics Simulations via Massive Parallelism. *J. Mater. Chem.* 16 (20), 1911–1918. doi:10.1039/B517931A
- Torda, A. E., and van Gunsteren, W. F. (1994). Algorithms for Clustering Molecular Dynamics Configurations. *J. Comput. Chem.* 15 (12), 1331–1340. doi:10.1002/jcc.540151203
- Torgerson, W. S. (1952). Multidimensional Scaling: I. Theory and Method. *Psychometrika* 17, 401–419. doi:10.1007/BF02288916
- Tribello, G. A., and Gasparotto, P. (2019). Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. Biosci.* 6, 46. doi:10.3389/fmolb.2019.00046
- Troyer, J. M., and Cohen, F. E. (1995). Protein Conformational Landscapes: Energy Minimization and Clustering of a Long Molecular Dynamics Trajectory. *Proteins Struct. Funct. Bioinform.* 23, 97–110. doi:10.1002/prot.340230111
- Uehara, S., and Tanaka, S. (2017). Cosolvent-based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Druggable Protein Conformations. *J. Chem. Inf. Model.* 57 (4), 742–756. doi:10.1021/acs.jcim.6b00791
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* 18 (6), 463–477. doi:10.1038/s41573-019-0024-5
- van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., and Tadesse, M. G. (2021). Bayesian Statistics and Modelling. *Nat. Rev. Methods Primers* 1, 1–26. doi:10.1038/s43586-020-00001-2
- van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* 9 (86), 2579–2605.
- Verstraelen, T., Van Houteghem, M., Van Speybroeck, V., and Waroquier, M. (2008). Md-tracks: a Productive Solution for the Advanced Analysis of Molecular Dynamics and Monte Carlo Simulations. *J. Chem. Inf. Model.* 48 (12), 2414–2424. doi:10.1021/ci800233y
- Vitek, A., Stachon, M., Krömer, P., and Šnáel, V. (2013). “Towards the Modeling of Atomic and Molecular Clusters Energy by Support Vector Regression,” in 2013 5th International Conference on Intelligent Networking and Collaborative Systems, 121–126. doi:10.1109/INCoS.2013.26
- VMD Plugin Library (2021). Theoretical and Computational Biophysics Group. Available at: <https://www.ks.uiuc.edu/Research/vmd/plugins/> (Accessed December 22, 2021).
- VMD Script Library (2021). Theoretical and Computational Biophysics Group. Available at: https://www.ks.uiuc.edu/Research/vmd/script_library (Accessed December 22, 2021).
- Wang, K., Chodera, J. D., Yang, Y., and Shirts, M. R. (2013). Identifying Ligand Binding Sites and Poses Using GPU-Accelerated Hamiltonian Replica Exchange Molecular Dynamics. *J. Comput.-Aided Mol. Des.* 27 (12), 989–1007. doi:10.1007/s10822-013-9689-8
- Wang, W., Gan, N., Sun, Q., Wu, D., Gan, R., Zhang, M., et al. (2019). Study on the Interaction of Ertugliflozin with Human Serum Albumin *In Vitro* by Multispectroscopic Methods, Molecular Docking, and Molecular Dynamics Simulation. *Spectrochim. Acta A. Mol. Biomol. Spectrosc.* 219, 83–90. doi:10.1016/j.saa.2019.04.047
- Wang, X., Song, K., Li, L., and Chen, L. (2018). Structure-Based Drug Design Strategies and Challenges. *Curr. Top. Med. Chem.* 18 (12), 998–1006. doi:10.2174/1568026618666180813152921
- Wei, L., and Altman, R. B. (1998). Recognizing Protein Binding Sites Using Statistical Descriptions of Their 3D Environments. *Pac. Symp. Biocomput.*, 497–508.
- Wooten, D., Christopoulos, A., Marti-Solano, M., Babu, M. M., and Sexton, P. M. (2018). Mechanisms of Signalling and Biased Agonism in G Protein-Coupled Receptors. *Nat. Rev. Mol. Cell Biol.* 19 (10), 638–653. doi:10.1038/s41580-018-0049-3
- Wu, W., Han, L., Wang, C., Wen, X., Sun, H., and Yuan, H. (2019). Structural Insights into Ligand Binding Features of Dual FABP4/5 Inhibitors by Molecular Dynamics Simulations. *J. Biomol. Struct. Dyn.* 37 (18), 4790–4800. doi:10.1080/07391102.2018.1561328
- Wu, X., Zheng, Z., Guo, T., Wang, K., and Zhang, Y. (2021). Molecular Dynamics Simulation of Lentinan and its Interaction with the Innate Receptor Dectin-1. *Int. J. Biol. Macromol.* 171, 527–538. doi:10.1016/j.ijbiomac.2021.01.032
- Yang, G. F. (2014). Structure-based Drug Design: Strategies and Challenges. *Curr. Pharm. Des.* 20 (5), 685–686. doi:10.2174/138161282005140214161643
- Yang, Y. L., Shao, Q., Zhang, J., Yang, L., and Gao, Y. Q. (2019). Enhanced Sampling in Molecular Dynamics. *J. Chem. Phys.* 151 (7), 70902. doi:10.1063/1.5109531
- Yesylevskyy, S. O. (2015). Pteros 2.0: Evolution of the Fast Parallel Molecular Analysis Library for C++ and Python. *J. Comput. Chem.* 36 (19), 1480–1488. doi:10.1002/jcc.23943
- Yesylevskyy, S. O. (2012). Pteros: Fast and Easy to Use Open-source C++ Library for Molecular Analysis. *J. Comput. Chem.* 33 (19), 1632–1636. doi:10.1002/jcc.22989
- Yoshino, R., Yasuo, N., and Sekijima, M. (2019). Molecular Dynamics Simulation Reveals the Mechanism by Which the Influenza Cap-Dependent Endonuclease Acquires Resistance Against Baloxavir Marboxil. *Sci. Rep.* 9 (1), 17464. doi:10.1038/s41598-019-53945-1
- Young, G., and Householder, A. S. (1938). Discussion of a Set of Points in Terms of Their Mutual Distances. *Psychometrika* 3, 19–22. doi:10.1007/BF02287916
- Zheng, W., Qi, B., Rohrdanz, M. A., Caffisch, A., Dinner, A. R., and Clementi, C. (2011). Delineation of Folding Pathways of a β -sheet Miniprotein. *J. Phys. Chem. B* 115 (44), 3065–3074. doi:10.1021/jp2076935
- Zheng, W., Rohrdanz, M. A., and Clementi, C. (2013a). Rapid Exploration of Configuration Space with Diffusion-Map-Directed Molecular Dynamics. *J. Phys. Chem. B* 117 (42), 12769–12776. doi:10.1021/jp401911h
- Zheng, W., Vargiu, A. V., Rohrdanz, M. A., Carloni, P., and Clementi, C. (2013b). Molecular Recognition of DNA by Ligands: Roughness and Complexity of the Free Energy Profile. *J. Chem. Phys.* 139 (14), 145102. doi:10.1063/1.4824106
- Zhou, H., Wang, F., Bennett, D. I., and Tao, P. (2019). Directed Kinetic Transition Network Model. *J. Chem. Phys.* 151 (14), 144112. doi:10.1063/1.5110896
- Zhou, H., Wang, F., and Tao, P. (2018). t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. *J. Chem. Theor. Comput.* 14 (11), 5499–5510. doi:10.1021/acs.jctc.8b00652

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Baltrukevich and Podlowska. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Pocket2Drug: An Encoder-Decoder Deep Neural Network for the Target-Based Drug Design

Wentao Shi¹, Manali Singha², Gopal Srivastava², Limeng Pu³, J. Ramanujam^{1,3} and Michal Brylinski^{2,3*}

¹Division of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA, United States, ²Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, United States, ³Center for Computation and Technology, Louisiana State University, Baton Rouge, LA, United States

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Jahan B. Ghasemi,
University of Tehran, Iran
Guang Hu,
Soochow University, China

*Correspondence:

Michal Brylinski
michal@brylinski.org

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 17 December 2021

Accepted: 10 February 2022

Published: 11 March 2022

Citation:

Shi W, Singha M, Srivastava G, Pu L,
Ramanujam J and Brylinski M (2022)
Pocket2Drug: An Encoder-Decoder
Deep Neural Network for the Target-
Based Drug Design.
Front. Pharmacol. 13:837715.
doi: 10.3389/fphar.2022.837715

Computational modeling is an essential component of modern drug discovery. One of its most important applications is to select promising drug candidates for pharmacologically relevant target proteins. Because of continuing advances in structural biology, putative binding sites for small organic molecules are being discovered in numerous proteins linked to various diseases. These valuable data offer new opportunities to build efficient computational models predicting binding molecules for target sites through the application of data mining and machine learning. In particular, deep neural networks are powerful techniques capable of learning from complex data in order to make informed drug binding predictions. In this communication, we describe Pocket2Drug, a deep graph neural network model to predict binding molecules for a given a ligand binding site. This approach first learns the conditional probability distribution of small molecules from a large dataset of pocket structures with supervised training, followed by the sampling of drug candidates from the trained model. Comprehensive benchmarking simulations show that using Pocket2Drug significantly improves the chances of finding molecules binding to target pockets compared to traditional drug selection procedures. Specifically, known binders are generated for as many as 80.5% of targets present in the testing set consisting of dissimilar data from that used to train the deep graph neural network model. Overall, Pocket2Drug is a promising computational approach to inform the discovery of novel biopharmaceuticals.

Keywords: ligand binding sites, drug discovery and development, in silico drug design, deep learning, graph neural network, recurrent neural network, generative model, machine learning

INTRODUCTION

Recent developments in genomics revealed novel disease-related molecular targets, many of which are yet to be characterized with respect to the possibility of modulating their functions with pharmaceutical agents. Another challenge in pharmacotherapy arises from resistance effects to existing drugs complicating the treatment of particularly infectious diseases (Trebbosc et al., 2019) and cancer (Shou et al., 2004). Therefore, many drug development projects are focused on the discovery of small molecule therapeutics with new mode of action (Gerry and Schreiber, 2018). Generating novel small molecules is a difficult endeavor due to the high complexity of biological systems and the enormous size of chemical space of organic compounds. Traditional experimental techniques can be

used to identify drug-like molecules performing specific biochemical tasks by binding to macromolecular targets with a high specificity in order to modulate their cellular functions. Nonetheless, even advanced high-throughput screening methods have notable limitations due to the long time and high costs of screening a large number of drug candidates.

To make the drug discovery process more efficient, modern approaches incorporate miscellaneous computational components. Virtual screening (VS) is perhaps the most widely used strategy to help identify potentially bioactive molecules from large collections of commercially available as well as virtual compounds (Segler et al., 2018). Despite its utility, this technology has certain drawbacks such as high false-positive rates, the requirement of predefined ligand libraries for structure-based VS, oversimplified scoring functions, and protein structure frameworks absent in ligand-based VS (Wu et al., 2019). More recently, machine learning (ML) methods addressing many of these issues have become available for drug discovery. New ML techniques include a quantitative structure-activity relationship model to predict the target affinity, toxicity, and side effects (Mouchlis et al., 2021) and an approach to model polypharmacy side effects with graph convolutional networks (GCN) (Zitnik et al., 2018).

Deep learning (DL) is a family of modern machine learning models utilizing deep neural networks (DNNs). DL models have been demonstrated to be powerful feature extractors for ligand binding site classifiers (Jiménez et al., 2017; Pu et al., 2019; Shi et al., 2020) and metric learning models for binding sites in proteins (Simonovsky and Meyers, 2020). Recurrent neural networks (RNNs) are iterative DL models that generate sequences through multiple iterations. In each iteration, the RNN model generates an output of time t taking the output of iteration $t - 1$ as the input. According to the probabilistic language model (Graves, 2013), the probability of input token x_{t+1} is modeled as $P(x_{t+1}|y_t)$, which is the probability of x_{t+1} conditioned on the output token y_t from the previous iteration. This powerful methodology was applied to *de novo* drug discovery, where RNNs were trained to model the probability distribution of a drug dataset (Ertl et al., 2017; Segler et al., 2018; Gupta et al., 2018; Yasonik, 2020). These methods treat a drug dataset as a set of languages and employ an RNN to learn the corresponding language models. After the training stage is completed, the RNN learns the probability distribution $P(\text{molecule})$ of the drug dataset, from which molecules can be sampled. RNN-based approaches often represent molecules using a simplified molecular-input line-entry system (SMILES) (Weininger, 1988), where individual string characters represent tokens of time steps. Although using RNNs to learn the distributions of drug datasets offers new opportunities to find drugs, these techniques still employ a random search of the chemical space leading to long virtual screening times. From a computational standpoint, when the aim is to identify promising lead molecules against a target binding site, it is certainly advantageous to have the search space significantly reduced.

In order to achieve this goal, we developed Pocket2Drug, a new deep generative model with the encoder-decoder architecture. Inspired by the framework of image captioning

models taking images as the input to generate corresponding captions (Vinyals et al., 2015; Xu et al., 2015), the basic idea is to provide RNN with the prior information on ligand binding pockets to improve the chances of finding bioactive molecules. A typical image captioning model consists of two parts, an encoder/feature extractor and a decoder. A convolutional neural network (CNN) is often used as the encoder extracting fixed-size latent feature vectors from the input images containing the prior information that can subsequently be decoded by an RNN to generate image captions. Formally, image captioning models learn the probability of sequences conditioned on prior information, i.e., $P(\text{caption}|\text{image})$.

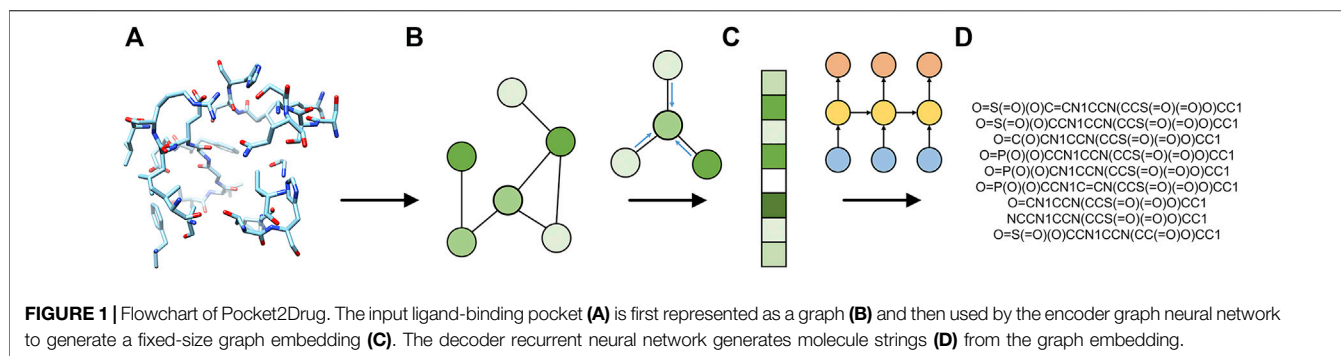
Pocket2Drug has a similar encoder-decoder architecture consisting of an encoder to extract features and a decoder to generate molecules. Nonetheless, Pocket2Drug differs from typical image captioning models in that it employs a graph representation of drug binding sites instead of images. Consequently, a GNN is employed as the encoder to extract the prior information from input pockets followed by an RNN decoder to generate molecule strings, which are the equivalents of image captions. In comprehensive benchmarking simulations against ligand-bound, ligand-free, and low-homology datasets of binding sites, we show that Pocket2Drug employing the encoder-decoder DNN effectively predicts binding drugs for input pocket structures.

MATERIALS AND METHODS

Datasets

Datasets used in this study were compiled from a non-redundant library of 51,677 pockets with bound ligands constructed for binding site prediction with eFindSite (Brylinski and Feinstein, 2013). The redundancy in the original library was already removed by excluding proteins with the template modeling (TM)-score, measuring the structure similarity (Zhang and Skolnick, 2004), of ≥ 0.4 and the 3D Tanimoto coefficient (TC), measuring the ligand similarity (Kawabata, 2011), of ≥ 0.7 . We further filtered the dataset based on the synthetic accessibility (SA) score (Ertl and Schuffenhauer, 2009) removing low- and high-complexity compounds whose SA scores are ≤ 1 and ≥ 6 , respectively. This procedure resulted in a high-quality dataset of 48,365 pockets binding small organic compounds, which were randomly split into training (90%) and testing (10%) subsets. The training subset of 43,529 pockets is referred to as the Pocket2Drug-train dataset while the remaining 4,836 (testing) pockets are called the Pocket2Drug-holo dataset.

Next, 433 pockets having a protein sequence identity of ≤ 0.5 with pockets in the training subset were selected from the Pocket2Drug-holo dataset creating the Pocket2Drug-lowhomol dataset to evaluate the ability to generalize to unseen data. Finally, the basic local alignment search tool (BLAST) (Altschul et al., 1990) was used with a sequence identity threshold of 95% to identify the apo structures of Pocket2Drug-holo proteins in the Protein Data Bank (PDB) (Berman et al., 2002). Ligand-free structures were then aligned on the corresponding holo-



proteins with TM-align (Zhang and Skolnick, 2005) and those producing significant alignments with a TM-score of ≥ 0.5 (Xu and Zhang, 2010) were retained. This procedure resulted in 828 ligand-free pockets referred to as the Pocket2Drug-apo dataset.

Graph Representation of Pockets

Binding pockets are represented as graphs, in which nodes are non-hydrogen atoms and edges connect pairs of atoms spatially located within 4.5 Å from one another (Shi et al., 2021). Node features include the hydrophobicity (Mahn et al., 2009), the charge, the binding probability (Jian et al., 2016), the solvent accessible surface area (Ali et al., 2014), and the sequence entropy (Liao et al., 2005), whereas the edge attribute is the bond multiplicity for covalently bonded atoms and 0 for atoms interacting non-covalently. Pockets are centered at the origin with principal axes aligned to Cartesian axes. The coordinates of individual atoms are also used as node features in order to provide the additional 3D information on binding pockets. This graph representation of ligand binding sites was used to accurately classify pockets in protein structures with GraphSite (Shi et al., 2021).

Encoder-Decoder Architecture

Pocket2Drug is implemented in PyTorch v1.7.1 (Paszke et al., 2019) and employs a DNN with the encoder-decoder architecture. The model learns the probability distribution of molecules conditioned on ligand binding pockets, $P(\text{molecule}|\text{pocket})$, which is then used to sample molecules for a given pocket as the prior condition. The pipeline implemented in Pocket2Drug is illustrated in Figure 1. For the input binding site (Figure 1A), a graph representation is generated by GraphSite (Shi et al., 2021) (Figure 1B) and the resulting graph is processed by an encoder to generate a fixed-size graph embedding (Figure 1C). As the encoder, we use a GNN constructed by removing the fully connected layers of the GraphSite classifier with parameters pretrained on binding site classification tasks (Shi et al., 2021). Subsequently, an RNN decoder takes the generated embedding vector as the input to compute SMILES sequences representing binding drugs (Figure 1D). Pocket2Drug is trained in an end-to-end fashion meaning that the parameters of both encoder and decoder are updated during backpropagation.

Graph Neural Network Encoder

The GNN encoder extracts latent features from the input pocket graphs. We use the embedding network implemented in the GraphSite classifier as the feature extractor with the last fully connected layer removed and the remaining parts of the classifier employed as the feature extractor. The message passing function utilizes weighted neighbor node features, in which weights are generated by a two-layer, fully connected neural network taking edge features as the input. Updated node features in k -th layer of node $x_i^{(k)}$, defined as

$$x_i^{(k)} = h_{\theta} \left(\text{concat}_{c \in \text{Channels}} \left((1 + \epsilon_c) \cdot x_i^{(k-1)} + \sum_{j \in N(i)} h_{wc}(e_{ij}) \cdot x_j^{(k-1)} \right) \right) \quad (1)$$

are first computed as a weighted sum of the first-order neighbors. The features of $x_i^{(k-1)}$ are weighted by $(1 + \epsilon_c)$, where ϵ_c is a trainable parameter. The weights of the first-order neighbors are generated by a neural network h_{wc} taking the edge feature, e_{ij} , as the input. Then, multiple channels of the weighted sum of the node features are concatenated and updated by another neural network h_{θ} . Finally, the output of each layer is connected by the jumping knowledge (JK)-network (Xu et al., 2018). The JK-network enables an automatic selection of the number of layers for individual nodes. Finally, the initial node embeddings are processed by the Set2Set graph read-out layer (Vinyals et al., 2016) to construct final, fixed-size graph embeddings.

Recurrent Neural Network Decoder

As a decoder, we use the gated recurrent unit (GRU), which is a variation of the vanilla RNN (Cho et al., 2014). The decoder network models a conditional probability of the output sequence based on the prior information on a ligand binding pocket:

$$P(\text{molecule}|\text{pocket}) = P(s_0|\text{pocket}) \prod_{t=1}^n P(s_t|\text{pocket}, s_0, \dots, s_{t-1}) \quad (2)$$

where s_t is the token of a molecule string at iteration t , and n is the length of the output string. Note that s_n represents the “end of string”, or *eos*, token. Figure 2 shows that the GRU network works differently during training and inference stages. During training, the graph embedding is taken by the GRU as the prior information to model the

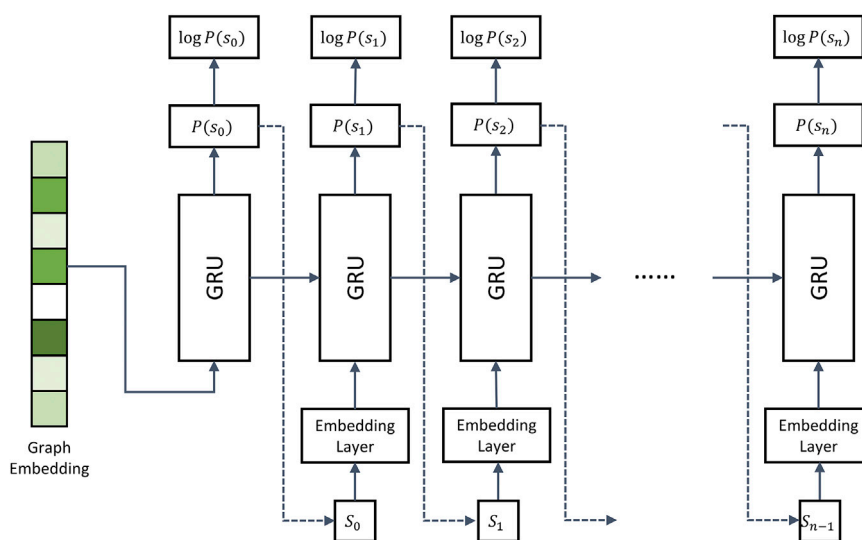


FIGURE 2 | Architecture of the recurrent neural network decoder. The decoder employs multiple gated recurrent units (GRUs). During model training, the molecule strings of binding drugs are used as the input. Dashed arrows represent the inference stage, in which the token sampled from $P(s_{t-1})$ is used as the input at iteration t .

probability distribution of all tokens, where the probability of a token s_0 is $P(s_0)$. In the remaining iterations, input tokens s_t of the binding drug string are mapped to vectors by the embedding layer and passed to the GRU as the input. The GRU then predicts the next token by generating another probability distribution $P(s_{t+1})$. The negative log likelihood of the binding drug is used as the loss function:

$$L = -\sum_{t=0}^n \log P(s_t) \quad (3)$$

Dashed arrows in **Figure 2** represent the inference stage. Here, the first iteration is the same as during training, i.e., the encoder generates graph embeddings used as the input in the first iteration. However, in the subsequent iterations, the RNN model takes the token s_t , sampled from the distribution of the previous step, to generate the distribution s_{t+1} . The inference stops when the *eos* token is reached.

Tokenization Scheme

Molecules can be represented by strings encoded by different tokenization schemes. Although SMILES is a widely used molecular string system, it was not designed for ML applications. Because of a strict syntax of SMILES, a significant portion of molecules generated by machine learning models are invalid. In addition, parentheses and ring indicators may be separated by long distances in SMILES strings causing problems for RNNs that have difficulty learning long-term dependencies (Öztürk et al., 2020). This issue can be addressed by improving either the RNN model or the tokenization scheme. For instance, RNN variants implementing “shortcuts” were developed to model long-term dependencies (Hochreiter and Schmidhuber, 1997). A long short-term memory (LSTM) model can also be used instead of a vanilla RNN in *de novo* drug design applications to learn the distribution of a drug dataset (Ertl et al., 2017). Another workaround is to improve the

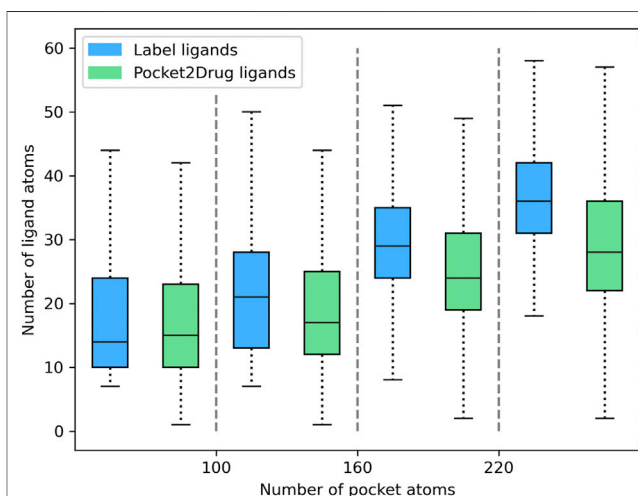


FIGURE 3 | Relationship between the ligand size and the size of binding pockets. The size of ligands and pockets is quantified by the number of non-hydrogen atoms. Binding pockets are assigned to four size groups: <100, 100–160, 161–220, and >220 atoms. For each pocket group, quartiles and the interquartile range are calculated for the size of label ligands (blue bars) and those molecules generated by Pocket2Drug (green bars).

tokenization scheme to make the string representation of molecules more suitable for ML applications. An example is DeepSMILES developed to enhance DL-based models taking SMILES as the input (O’Boyle and Dalke, 2018).

Pocket2Drug employs SELF-referencing Embedding Strings (SELFIES), another molecule tokenization scheme designed for machine learning applications (Krenn et al., 2020). The SELFIES method was selected because of several important properties. Not only any molecule can be represented by a SELFIES string, but also all virtual molecules generated by an ML model are valid.

TABLE 1 | Hit rates for the Pocket2Drug-holo dataset.

| Method | Sample size of 20,480 | | | | Sample size of 81,920 | | | |
|-------------|-----------------------|-------------------|-------------------|-------------|-----------------------|-------------------|-------------------|-------------|
| | TC ≥ 0.7 (%) | TC ≥ 0.8 (%) | TC ≥ 0.9 (%) | TC =1.0 (%) | TC ≥ 0.7 (%) | TC ≥ 0.8 (%) | TC ≥ 0.9 (%) | TC =1.0 (%) |
| Pocket2Drug | 95.9 | 79.9 | 64.8 | 52.5 | 98.4 | 86.8 | 69.7 | 56.4 |
| ZINC | 58.9 | 23.8 | 3.3 | 0.4 | 73.6 | 40.5 | 8.4 | 1.2 |
| Vanilla RNN | 57.1 | 19.7 | 1.6 | 0.1 | 70.9 | 35.3 | 4.7 | 0.3 |

Importantly, the information on rings and branches in SELFIES is localized by storing the branch size and ring size together with their identifiers. This tokenization scheme makes it easier for RNNs to learn from the “past” information compared to, e.g., SMILES that require RNNs to infer ring/branch indicators based on non-localized information.

EVALUATION AND RESULTS

Pocket2Drug was trained on the Pocket2Drug-train dataset and validated against Pocket2Drug-holo, -apo, and -lowhomol datasets. We first analyze the size of molecules generated for the Pocket2Drug-holo dataset. **Figure 3** shows that there is a notable correlation between the size of pockets and the size of binding molecules, referred to as label ligands, across experimental complex structures (blue bars). Encouragingly, the size of ligands constructed by Pocket2Drug is also correlated with the pocket size, although these molecules tend to be somewhat smaller than the corresponding label ligands (green bars). This result can be attributed to the fact that capturing longer dependencies in molecular strings is more difficult for the RNN trained to minimize the sum of cross-entropy loss function. In other words, the model makes fewer mistakes by generating smaller molecules.

Next, the quality of molecules generated for the Pocket2Drug-holo dataset is evaluated using two complementing protocols, one based on the chemical similarity of binding molecules (Baldi and Nasr, 2010) and another utilizing the structure alignments of protein pockets (Yeturu and Chandra, 2011). Pocket2Drug is compared to two baselines. The first method randomly selects drug candidates from the ZINC database, a curated collection of commercially available chemical compounds prepared specifically for virtual screening (Irwin and Shoichet, 2005). The second baseline method selects drug candidates from the output of a vanilla RNN (Segler et al., 2018) representing a typical DL-based approach for *de novo* drug design.

Evaluation by Ligand Chemical Similarity

The performance of Pocket2Drug, ZINC, and vanilla RNN are evaluated with the TC between the generated molecules and label ligands. For each pocket in the Pocket2Drug-holo dataset, TC values are calculated for a specified number of molecules sampled from the model output and the highest TC is selected as the final score. **Table 1** reports the percentage of Pocket2Drug-holo pockets with the corresponding score greater than or equal to a TC threshold ranging from 0.7 to 1.0. Encouragingly, using Pocket2Drug significantly improves chances to find binding

molecules compared to ZINC and vanilla RNN. For a sample size of 20,480 (10 batches of 2,048 molecules each to maximize the GPU utilization), Pocket2Drug generates at least one molecule which a TC of ≥ 0.7 to the label ligand for as many as 95.9% pockets. Note that two molecules sharing chemical similarity with a TC of ≥ 0.7 tend to have a similar bioactivity (Kumar, 2011; Ben Lo, 2016). For the majority of pockets (52.5%), Pocket2Drug selects the label ligand itself (a TC of 1.0). This performance is significantly higher than that of ZINC/vanilla RNN that selects ligands with a TC of ≥ 0.7 for 58.9%/57.1% of pockets and label ligands for merely 0.4%/0.1% of pockets. Increasing the sample size to 81,920 slightly improves the performance because four times more molecules are used to select that with the highest TC value. A significantly improved performance of Pocket2Drug over vanilla RNN can be attributed to the effective utilization of the prior information on ligand binding pockets learned by the ML model.

Next, the performance of Pocket2Drug is assessed against the Pocket2Drug-apo dataset. The mean root-mean-square deviation (RMSD) (Kabsch, 1976) of ligand-free structures against ligand-bound conformations is $1.2 \text{ \AA} \pm 0.9$. This low RMSD is expected because, with a few exceptions, the structures of apo- and holo-proteins tend to be highly similar (Brylinski and Skolnick, 2008). **Table 2** reports hit rates for molecules generated by Pocket2Drug using ligand-free and the corresponding ligand-bound pockets in the Pocket2Drug-holo dataset. Encouragingly, the performance of Pocket2Drug is independent on the ligand binding state of target proteins, therefore, the model does not require input proteins to be co-crystallized with ligands in order to successfully generate binding molecules.

We also evaluate the ability of Pocket2Drug to generalize to unseen data by measuring its performance against the Pocket2Drug-lowhomol dataset. As reported in **Table 3**, label ligands (a TC of 1.0) are generated by Pocket2Drug in 77.1%/80.5% of the cases when the sample size is 20,480/81,920. This performance represents a notable improvement over ZINC and vanilla RNN selecting a very few label ligands. Pocket2Drug also achieves the highest performance for other TC thresholds ranging from 0.7 to 0.9. These results show that Pocket2Drug not only performs exceptionally well against Pocket2Drug-holo and -apo datasets, but also against the Pocket2Drug-lowhomol dataset comprising proteins with a low sequence homology to the training subset demonstrating that it generalizes well to unseen data.

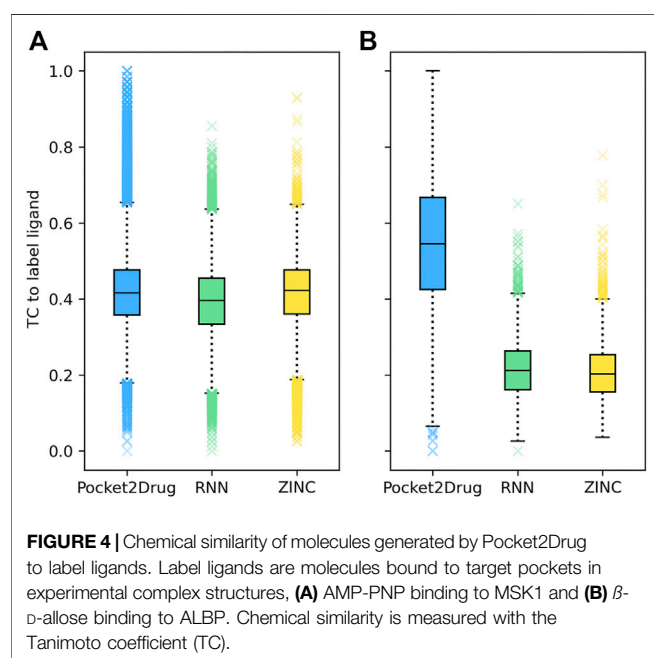
Two representative examples of pockets in the Pocket2Drug-lowhomol dataset are discussed in detail, a nucleotide binding site in the human mitogen and stress activated protein kinase 1 (MSK1) and a sugar binding site in D-allose binding protein (ALBP) from *E. coli*. MSK1 is involved in the regulation of

TABLE 2 | Hit rates for the Pocket2Drug-apo dataset. For each ligand-free structure, the corresponding ligand-bound pocket is selected from the Pocket2Drug-holo dataset for the apples-to-apples comparison.

| Binding state | Sample size of 20,480 | | | | Sample size of 81,920 | | | |
|---------------|-----------------------|-------------------|-------------------|-------------|-----------------------|-------------------|-------------------|-------------|
| | TC ≥ 0.7 (%) | TC ≥ 0.8 (%) | TC ≥ 0.9 (%) | TC =1.0 (%) | TC ≥ 0.7 (%) | TC ≥ 0.8 (%) | TC ≥ 0.9 (%) | TC =1.0 (%) |
| Ligand-free | 95.3 | 72.7 | 53.3 | 37.4 | 98.2 | 82.2 | 57.2 | 40.5 |
| Ligand-bound | 95.3 | 72.2 | 52.3 | 37.0 | 98.2 | 81.6 | 58.1 | 41.2 |

TABLE 3 | Hit rates for the Pocket2Drug-lowhomol dataset.

| Method | Sample size of 20,480 | | | | Sample size of 81,920 | | | |
|-------------|-----------------------|-------------------|-------------------|-------------|-----------------------|-------------------|-------------------|-------------|
| | TC ≥ 0.7 (%) | TC ≥ 0.8 (%) | TC ≥ 0.9 (%) | TC =1.0 (%) | TC ≥ 0.7 (%) | TC ≥ 0.8 (%) | TC ≥ 0.9 (%) | TC =1.0 (%) |
| Pocket2Drug | 98.2 | 95.2 | 87.5 | 77.1 | 98.9 | 96.8 | 90.0 | 80.5 |
| ZINC | 49.2 | 18.4 | 2.7 | 0.2 | 66.7 | 36.3 | 10.4 | 2.3 |
| Vanilla RNN | 50.8 | 16.1 | 0.9 | 0.0 | 62.8 | 28.8 | 5.7 | 0.9 |

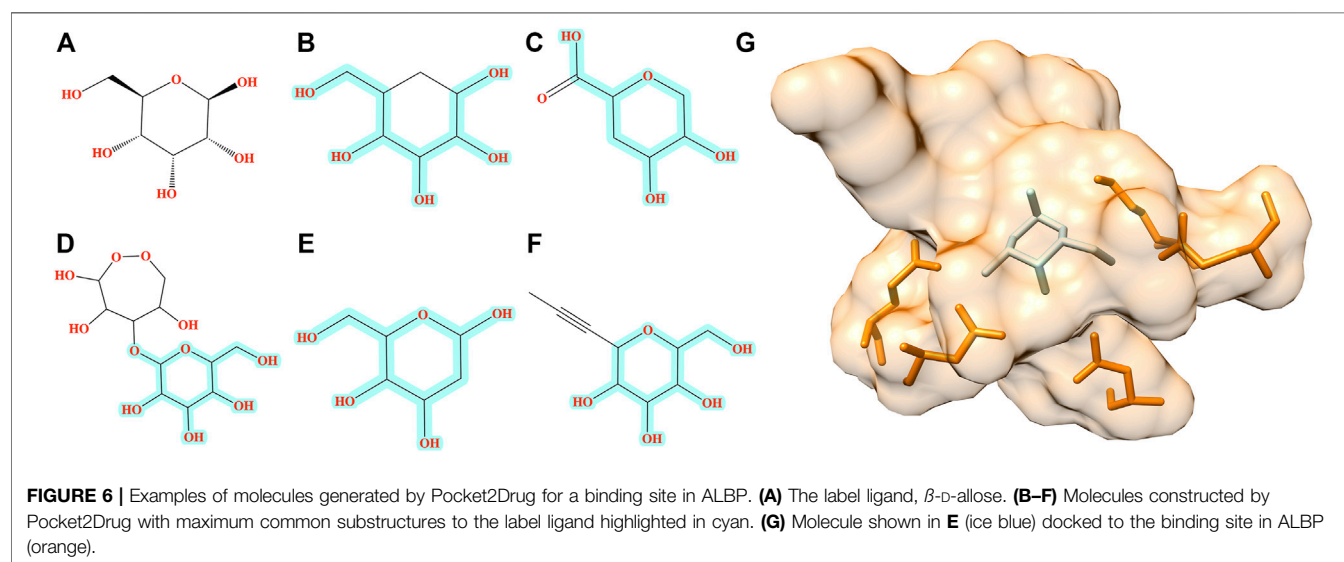
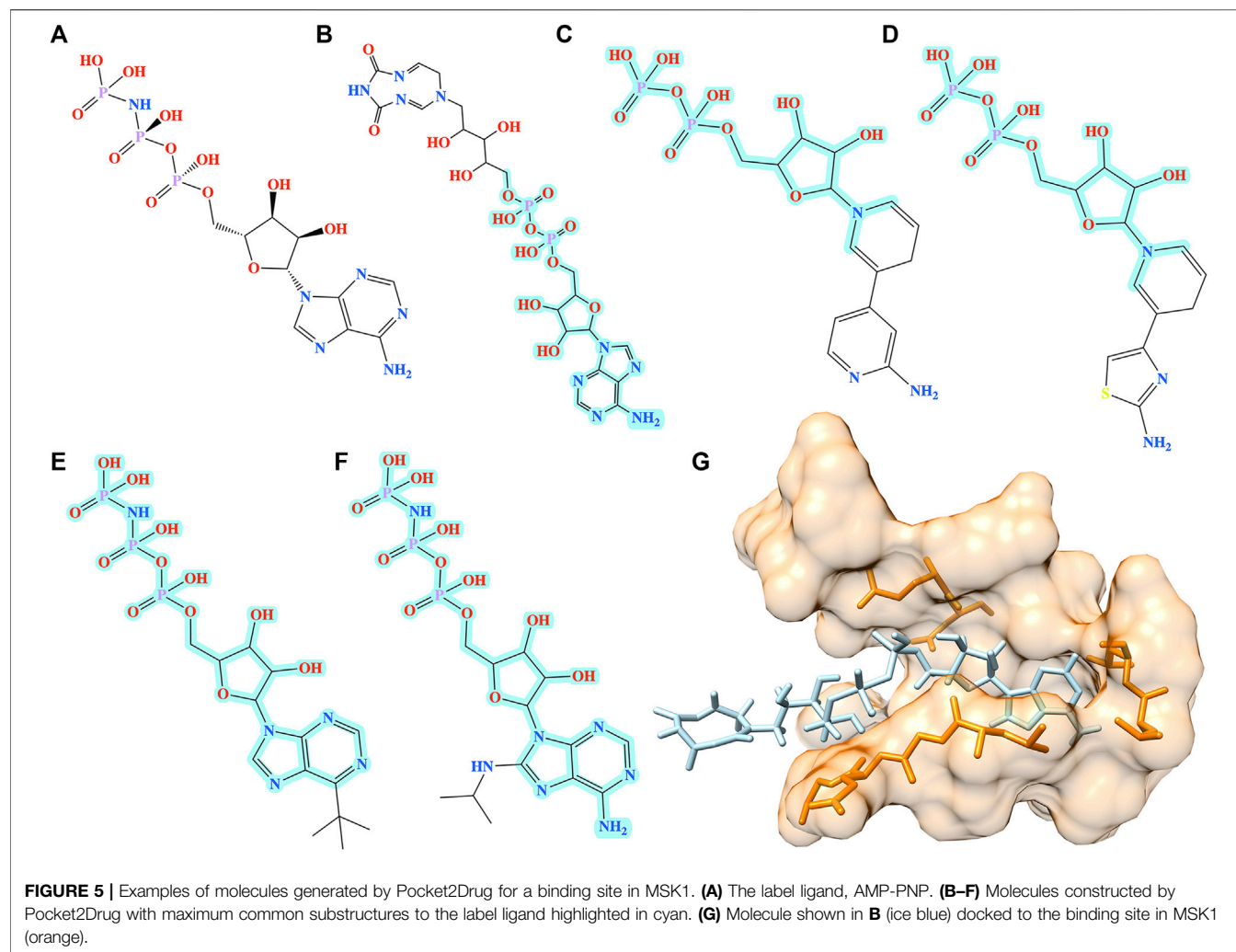


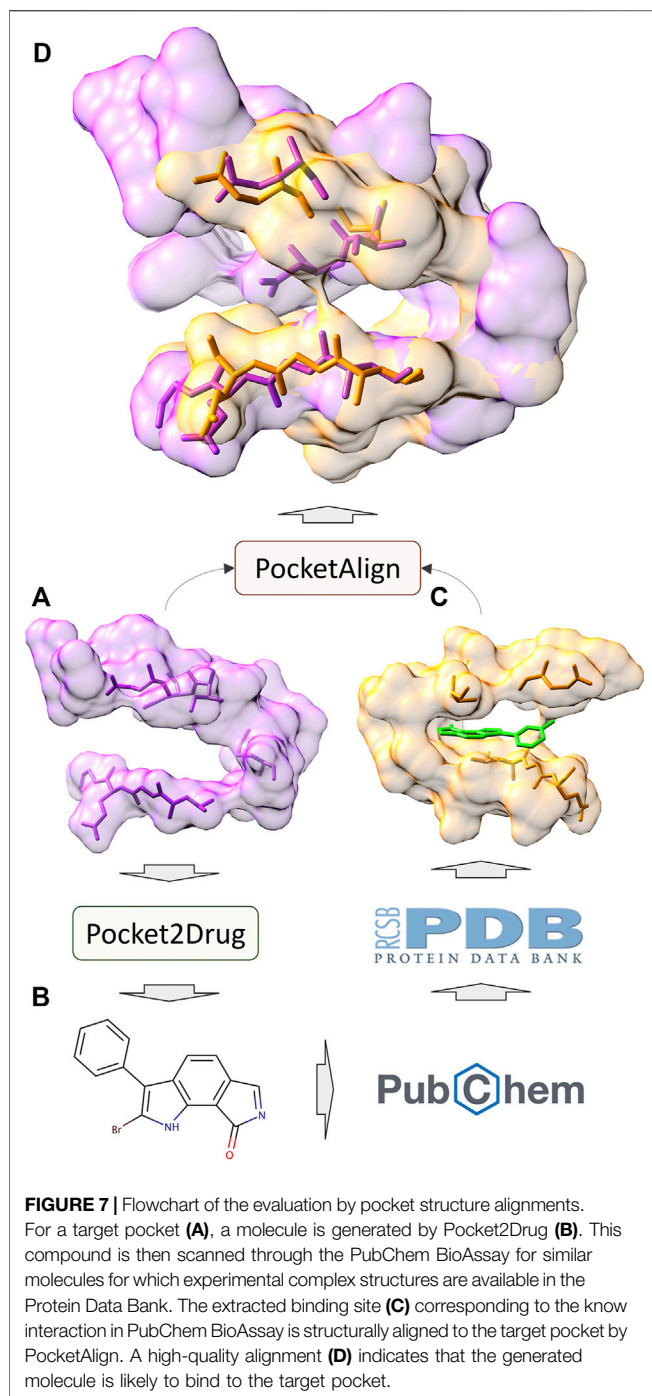
mitogen activated kinases and it is required by the tumor-promoter-induced neoplastic cell transformation (Malakhova et al., 2010). The complex structure of MSK1 and the phospho-amino-phosphonic acid-adenylate ester (AMP-PNP) (Malakhova et al., 2010) was chosen as the target. AMP-PNP is a competitive ATPase inhibitor blocking the ATP-dependent oxidative phosphorylation (Lardy et al., 1975). **Figure 4A** shows the distribution of TC similarities between the label ligand, AMP-PNP, and molecules generated by Pocket2Drug and two baseline methods. Although most virtual molecules have relatively low TC similarities to AMP-PNP, more molecules with high TC values are sampled from the Pocket2Drug model compared to ZINC and vanilla RNN. According to the Fisher-Pitman permutation test (Neuhäuser and Manly, 2004), the difference between

Pocket2Drug and vanilla RNN is statistically significant with a p -value close to 0 and that between Pocket2Drug and ZINC is insignificant with a p -value of 0.1.

To better understand the biological relevance of molecules generated by Pocket2Drug, five representative compounds with TC similarities against AMP-PNP ranging from 1.0 to 0.8 are presented in **Figure 5**. **Figure 5A** shows AMP-PNP, which is a nonhydrolyzable ATP analogue forming hydrogen bonds with MSK1 pocket residues through several moieties, NH_2 in adenine, $3'$ OH in pentose sugar, OH in β -phosphate, NH linking β - and γ -phosphates and OH in γ -phosphate in the complex crystal structure (Malakhova et al., 2010). Interestingly, several molecules generated by Pocket2Drug have common substructures with either substitutions in the adenine moiety (**Figures 5E,F**) and the terminal phosphate group (**Figure 5B**) or sharing the PNP subunit (**Figures 5C,D**). These virtual molecules contain groups forming important hydrogen bonds with MSK1 pocket residues. To further evaluate the possibility of binding, all molecules were docked into the AMP-PNP pocket of MSK1 with fcombu (Kawabata and Nakamura, 2014). The docking scores of the generated molecules are 12.5, 18.1, 21.8, 17.6, and 13.0 (**Figures 5B–F**, respectively). These results indicate that molecules generated by Pocket2Drug dock favorably to the target pocket with the compound shown in **Figures 5B,G** having the best docking score due to the substitution in β -phosphate group.

The improvement of Pocket2Drug over baseline methods is even more perceptible for ALBP where the distribution of TC similarities to the label ligand is shifted toward much higher values for molecules sampled from the Pocket2Drug model (**Figure 4B**). Differences between Pocket2Drug and both baseline methods are statistically significant with p -values close to 0. ALBP is a member of the ATP-binding cassette (ABC) transporter family facilitating the import and export of various molecules across the cell membrane (Fath and Kolter, 1993). ALBP binds β -D-allose, shown in **Figure 6A**, with a K_d of 0.33 μM (Chaudhuri et al., 1999). In the crystal complex structure, β -D-allose forms multiple interactions with the pocket residues of





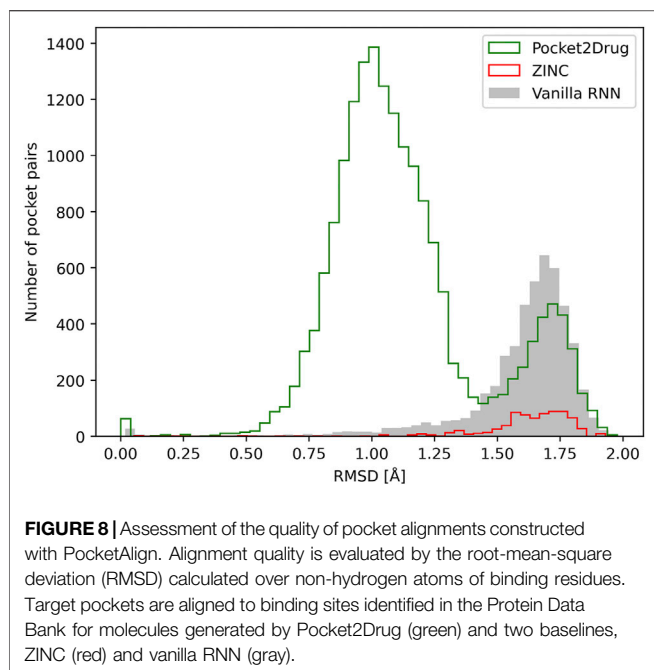
ALBP through the ring oxygen and five hydroxyl moieties (Chaudhuri et al., 1999). Selected compounds generated by Pocket2Drug are presented in Figures 6B–F. In addition to a substituted cyclohexane (Figure 6B), several substituted allose molecules (Figures 6C–F) sharing a high chemical similarity with the label ligand, β -D-allose (Figure 6A), were constructed. Most of these molecules dock well to ALBP pocket with docking scores of 4.1, 3.7, 20.9, 3.5, and 9.8 for compounds shown in Figures

6B–F, respectively. Interestingly, a substituted cyclohexane in the molecule shown in Figure 6B adopts the chair conformation similarly to β -D-allose bound to ALBP in the experimental complex structure. A compound shown in Figures 6E,G has the best docking score, whereas that shown in Figure 6D has less favorable docking score than those ligands having a comparable size to β -D-allose because of the large substitution at 5' position that does not fit in the binding pocket of ALBP. Docking results suggest that molecules generated by Pocket2Drug are capable of forming favorable interactions with the target pocket.

Evaluation by Pocket Structure Alignments

In addition to the assessment by ligand chemical similarity described above, the performance of Pocket2Drug is also evaluated with pocket structure alignments. This approach is based on an assumption that a molecule generated for the target pocket is a hit if a similar molecule binds to a site that is structurally similar to the target pocket (Govindaraj and Brylinski, 2018; Gaieb et al., 2019). A flowchart of the evaluation procedure is shown in Figure 7. For a target pocket in the testing set (Figure 7A), molecules generated by Pocket2Drug are ranked according to their frequencies and 100 of the most frequent molecules are selected. For each drug candidate (Figure 7B), chemically similar ligands with a TC of ≥ 0.7 are identified in the PubChem BioAssay dataset comprising 73,021 active interactions involving 919 unique proteins and 17,367 unique compounds (Wang et al., 2012). Next, the experimental complex structures of these ligands bound to similar proteins with a sequence identity of $\geq 70\%$ to PubChem BioAssay targets are retrieved from the PDB. The extracted binding sites (Figure 7C) are finally structurally aligned to the initial target pocket with PocketAlign, an accurate method to superpose ligand binding sites in a sequence order-independent manner (Yeturu and Chandra, 2011). Essentially, this procedure validates molecules generated for target pockets by finding similar interactions that have already been determined experimentally through binding assays and protein crystallography.

Similar to the evaluation protocol by ligand chemical similarity, Pocket2Drug is compared to ZINC and vanilla RNN. For each target pocket, 100 molecules from the ZINC database and 100 molecules generated by vanilla RNN are selected so that their molecular weight distributions match those calculated for compounds selected by Pocket2Drug. In terms of statistics, the number of pocket pairs used as input for structure alignments is 17,620 for Pocket2Drug, 6,307 for ZINC, and 6,694 for vanilla RNN. The number of valid pocket alignments constructed by PocketAlign (Yeturu and Chandra, 2011) are 16,987 (Pocket2Drug), 741 (ZINC), and 4,902 (vanilla RNN). A valid pocket alignment has the RMSD of ≤ 2 Å; higher RMSD values indicate that two pockets are structurally dissimilar. According to this criterion, as many as 96.4% of validation pairs of pockets identified using output molecules generated by Pocket2Drug produce valid structure alignments, while these percentages are notably lower for ZINC (11.7%) and vanilla RNN (73.2%). The distribution of the RMSD scores of pocket alignments for all tested methods is

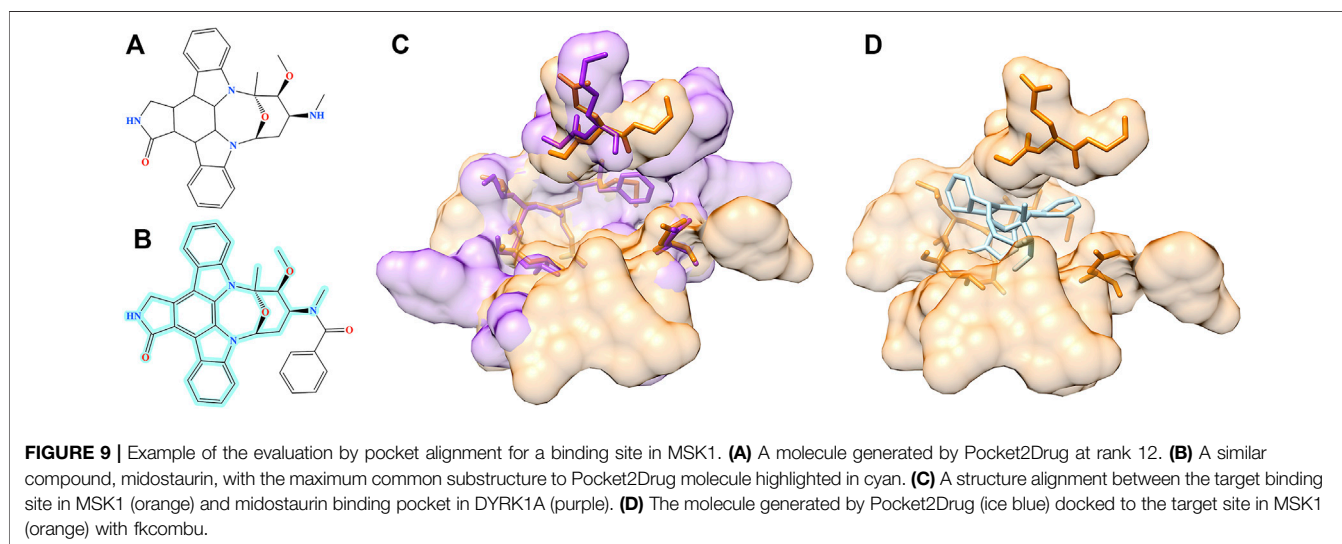


presented in **Figure 8**. Not only using molecules selected by Pocket2Drug results in the highest percentage of valid structure alignments, but also RMSD values for these superpositions are generally much lower compared to ZINC and vanilla RNN. The mean RMSD scores for pocket2Drug, ZINC, and vanilla RNN are 1.1 Å, 1.6 Å, and 1.6 Å, respectively.

Structure alignment results demonstrate that for a large number of molecules generated by Pocket2Drug for target pockets, there are experimentally determined interactions between chemically similar ligands binding to structurally similar pockets. Two representative cases are selected to exemplify the evaluation by pocket structure alignments. The first target pocket is a nucleotide binding site in

MSK1 used in the previous section to illustrate the results of the evaluation by ligand chemical similarity. Among molecules generated by Pocket2Drug, a compound ranked 12 with the frequency of 21 (**Figure 9A**) is chemically similar to midostaurin (PubChem-CID: 9829523, **Figure 9B**), a protein kinase C (PKC) inhibitor (Eder et al., 2004) used to treat systemic mastocytosis, acute myeloid leukemia, and mast cell leukemia (National Cancer Institute Dictionary, 2021). According to the bioassay data (PubChem-BAID: 208295368), midostaurin inhibits PKC- α isoform with the half-maximal inhibitory concentration (IC_{50}) of 22 nM (Millward et al., 2006). Midostaurin has been co-crystallized with the human dual specificity tyrosine-phosphorylation-regulated kinase 1A (DYRK1A, 25% sequence identity with PKC- α) with the equilibrium dissociation constant (K_d) of 100 nM (PDB-ID: 4nct) (Alexeeva et al., 2015). **Figure 9C** shows the structure alignment constructed by PocketAlign between AMP-PNP binding pocket in MSK1 and midostaurin binding pocket in DYRK1A. Despite a low global sequence identity between these proteins of only 26%, their binding pockets are structurally highly similar with the RMSD of 0.90 Å. The compound generated by Pocket2Drug docks to the AMP-PNP binding pocket in MSK1 with a score of 58.5 (**Figure 9D**).

The second example is the human angiotensin-1 receptor (Tie-2), an enzyme involved in vessel remodeling, branching, stability, and maturation (Yu, 2005). Using the binding site of Tie-2 as the input, Pocket2Drug generated a molecule shown in **Figure 10A** at rank 9 with a frequency of 5. This compound is chemically similar to doramapimod (PubChem-CID: 156422, **Figure 10B**), an inhibitor of ephrin type-A receptor 2 (EphA2) with a TC of 0.73. According to the bioassay data (PubChem-BAID: 40394839), doramapimod binds to EphA2 with a K_d of 0.37 nM and has been tested for its anti-proliferative activity in the SF-268 cell line. It inhibits the viability of EphA2 growth dependent glioblastoma cells with a half-maximal effective concentration (EC_{50}) of 5 μ M (Heinzlmeir et al., 2017). Despite a low global sequence identity of 37%, the structure alignment of binding sites in Tie-2 (PDB-ID:



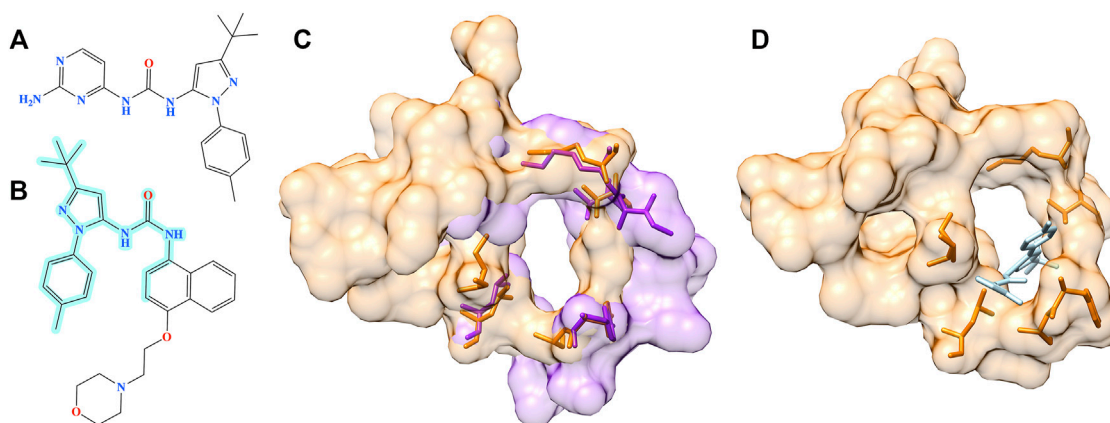


FIGURE 10 | Example of the evaluation by pocket alignment for a binding site in Tie-2. **(A)** A molecule generated by Pocket2Drug at rank 9. **(B)** A similar compound, doramapimod, with the maximum common substructure to Pocket2Drug molecule highlighted in cyan. **(C)** A structure alignment between the target binding site in Tie-2 (orange) and doramapimod binding pocket in EphA2 (purple). **(D)** The molecule generated by Pocket2Drug (ice blue) docked to the target site in Tie-2 (orange) with fkcombu.

2008) and EphA2 (PDB-ID: 5nkd) yields an RMSD of 0.95 Å (**Figure 10C**). Docking simulations with fkcombu confirmed that the molecule generated by Pocket2Drug fits well into the binding site of Tie-2 with a score of 24.3 (**Figure 10D**).

DISCUSSION

In this communication, we describe Pocket2Drug, a novel deep learning model employing an encoder-decoder architecture to predict binding molecules for a ligand binding site. Pocket2Drug was trained in an end-to-end supervised manner against a large collection of ligand-pocket pairs. The analysis of molecules generated by Pocket2Drug using two evaluation protocols based on ligand chemical similarity and pocket structure alignments revealed that this algorithm significantly improves the chances of finding binding ligands compared to traditional techniques. Pocket2Drug not only yields a high accuracy against ligand-free structures, but it also generalizes well to unseen data, *viz.* those pockets extracted from proteins that are different from training instances. These findings are particularly important in drug discovery against novel protein structures, where it can help significantly reduce the search space of drug candidates. In contrast to traditional virtual screening typically employing a library of 200,000 to over 1,000,000 molecules (Hughes et al., 2011), Pocket2Drug generates molecules that have high chances to bind to target pockets within a smaller sample of 81,920 compounds. Therefore, it can potentially decrease the number of molecules to be subjected to structure-based virtual screening from millions to tens of thousands.

Pocket2Drug can be improved by incorporating reinforcement learning imposing additional restraints on the synthetic accessibility, solubility, and toxicity of generated molecules, depending on a specific application. Additional improvements can also be achieved by applying a framework similar to the conditional recurrent neural network (cRNN), utilizing the RNN with the

prior information (Xu et al., 2021), to the heterogeneous input data. In contrast to cRNN, in which the pre-computed information is used as the prior condition for RNN, Pocket2Drug is an end-to-end DNN, therefore the encoder is updated during training. Another difference is the data representation; cRNN uses a voxel representation as the prior information, whereas Pocket2Drug employs a computationally more efficient graph representation. Nonetheless, the heterogeneous pocket data can be combined by concatenating embedding vectors generated by different feature extractors in order to provide the prior information on ligand binding sites.

An attention mechanism was shown to significantly improve the performance of image captioning because it helps the model capture more semantically meaningful parts of images (Xu et al., 2015). We expect that the same methodology can be implemented in Pocket2Drug since pocket residues contribute differently to the formation of molecular interactions with binding ligands. These are examples of future research directions that will be explored to further improve the performance of Pocket2Drug in the discovery of novel biopharmaceuticals.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://github.com/shiwentao00/Pocket2Drug>, <https://osf.io/qacwj/>.

AUTHOR CONTRIBUTIONS

Conceptualization: WS; Methods: WS and LP; Dataset: MB; Evaluation and case studies: WS, MS, and GS; Supervision: MB; Funding requisition: JR and MB; Manuscript draft: WS, LP, MS, and GS; Final manuscript: MB.

FUNDING

This work has been supported in part by the National Institute of General Medical Sciences of the National Institutes of Health

REFERENCES

- Alexeeva, M., Åberg, E., Engh, R. A., and Rothweiler, U. (2015). The Structure of a Dual-Specificity Tyrosine Phosphorylation-Regulated Kinase 1A-Pkc412 Complex Reveals Disulfide-Bridge Formation with the Anomalous Catalytic Loop HRD(HCD) Cysteine. *Acta Crystallogr. D Biol. Crystallogr.* 71 (Pt 5), 1207–1215. doi:10.1107/S1399004715005106
- Ali, S. A., Hassan, M. I., Islam, A., and Ahmad, F. (2014). A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. *Curr. Protein Pept. Sci.* 15 (5), 456–476. doi:10.2174/1389203715666140327114232
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/S0022-2836(05)80360-2
- Baldi, P., and Nasr, R. (2010). When Is Chemical Similarity Significant? the Statistical Distribution of Chemical Similarity Scores and its Extreme Values. *J. Chem. Inf. Model.* 50 (7), 1205–1222. doi:10.1021/ci100010v
- Ben Lo, J. Z. T. (2016). “Chemical Similarity Networks for Drug Discovery,” in *Special Topics in Drug Discovery* (Intech), 53–72. doi:10.5772/65106
- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., et al. (2002). The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* 58 (Pt 6 No 1), 899–907. doi:10.1107/s0907444902003451
- Brylinski, M., and Feinstein, W. P. (2013). eFindSite: Improved Prediction of Ligand Binding Sites in Protein Models Using Meta-Threading, Machine Learning and Auxiliary Ligands. *J. Comput. Aided Mol. Des.* 27 (6), 551–567. doi:10.1007/s10822-013-9663-5
- Brylinski, M., and Skolnick, J. (2008). What Is the Relationship between the Global Structures of Apo and Holo Proteins. *Proteins* 70 (2), 363–377. doi:10.1002/prot.21510
- Chaudhuri, B. N., Ko, J., Park, C., Jones, T. A., and Mowbray, S. L. (1999). Structure of D-Allose Binding Protein from *Escherichia coli* Bound to D-Allose at 1.8 Å Resolution. *J. Mol. Biol.* 286 (5), 1519–1531. doi:10.1006/jmbi.1999.2571
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25–29, 2014.
- Eder, J. P., Jr., Garcia-Carbonero, R., Clark, J. W., Supko, J. G., Puchalski, T. A., Ryan, D. P., et al. (2004). A Phase I Trial of Daily Oral 4'-N-Benzoyl-Staurosporine in Combination with Protracted Continuous Infusion 5-fluorouracil in Patients with Advanced Solid Malignancies. *Invest. New Drugs* 22 (2), 139–150. doi:10.1023/B:DRUG.0000011790.31292.ef
- Ertl, P., and Schuffenhauer, A. (2009). Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform* 1 (1), 8. doi:10.1186/1758-2946-1-8
- Ertl, P., Lewis, R., Martin, E., and Polyakov, V. (2017). *In Silico Generation of Novel, Drug-like Chemical Matter Using the LSTM Neural Network*. arXiv preprint arXiv:1712.07449.
- Fath, M. J., and Kolter, R. (1993). ABC Transporters: Bacterial Exporters. *Microbiol. Rev.* 57 (4), 995–1017. doi:10.1128/mr.57.4.995-1017.1993
- Gaieb, Z., Parks, C., and Amaro, R. (2019). *Evaluation of Binding Site Comparison Algorithms and Proteomic Machine Learning Models in the Detection of Protein Pockets Capable of Binding the Same Ligand*. ChemRxiv preprint ChemRxiv:9178136.
- Gerry, C. J., and Schreiber, S. L. (2018). Chemical Probes and Drug Leads from Advances in Synthetic Planning and Methodology. *Nat. Rev. Drug Discov.* 17 (5), 333–352. doi:10.1038/nrd.2018.53
- Govindaraj, R. G., and Brylinski, M. (2018). Comparative Assessment of Strategies to Identify Similar Ligand-Binding Pockets in Proteins. *BMC Bioinformatics* 19 (1), 91. doi:10.1186/s12859-018-2109-2
- award R35GM119524, the US National Science Foundation award CCF1619303, the Louisiana Board of Regents contract LEQSF(2016-19)-RD-B03 and by the Center for Computation and Technology, Louisiana State University.
- Graves, A. (2013). *Generating Sequences with Recurrent Neural Networks*. arXiv preprint arXiv:1308.0850.
- Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., and Schneider, G. (2018). Generative Recurrent Networks for De Novo Drug Design. *Mol. Inform.* 37 (1–2), 1700111. doi:10.1002/minf.201700111
- Heinzlmeir, S., Lohse, J., Treiber, T., Kudlinzki, D., Linhard, V., Gande, S. L., et al. (2017). Chemoproteomics-Aided Medicinal Chemistry for the Discovery of EPHA2 Inhibitors. *ChemMedChem* 12 (12), 999–1011. doi:10.1002/cmdc.201700217
- Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735
- Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of Early Drug Discovery. *Br. J. Pharmacol.* 162 (6), 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x
- Irwin, J. J., and Shoichet, B. K. (2005). ZINC—a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* 45 (1), 177–182. doi:10.1021/ci049714+
- Jian, J. W., Elumalai, P., Pitti, T., Wu, C. Y., Tsai, K. C., Chang, J. Y., et al. (2016). Predicting Ligand Binding Sites on Protein Surfaces by 3-dimensional Probability Density Distributions of Interacting Atoms. *PloS one* 11 (8), e0160315. doi:10.1371/journal.pone.0160315
- Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., and De Fabritiis, G. (2017). DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* 33 (19), 3036–3042. doi:10.1093/bioinformatics/btx350
- Kabsch, W. (1976). A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Cryst. Sect. A* 32 (5), 922–923. doi:10.1107/s0567739476001873
- Kawabata, T. (2011). Build-up Algorithm for Atomic Correspondence between Chemical Structures. *J. Chem. Inf. Model.* 51 (8), 1775–1787. doi:10.1021/ci2001023
- Kawabata, T., and Nakamura, H. (2014). 3D Flexible Alignment Using 2D Maximum Common Substructure: Dependence of Prediction Accuracy on Target-Reference Chemical Similarity. *J. Chem. Inf. Model.* 54 (7), 1850–1863. doi:10.1021/ci500006d
- Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Machine Learn. Sci. Techn.* 1 (4), 045024. doi:10.1088/2632-2153/aba947
- Kumar, A. (2011). Chemical Similarity Methods : A Tutorial Review. *The Chem. educator* 16, 46–50. doi:10.1333/s00897112344a
- Lardy, H. A., Schuster, S. M., and Ebel, R. E. (1975). Exploring Sites on Mitochondrial ATPase for Catalysis, Regulation, and Inhibition. *J. Supramol. Struct.* 3 (3), 214–221. doi:10.1002/jss.400030303
- Liao, H., Yeh, W., Chiang, D., Jernigan, R. L., and Lustig, B. (2005). Protein Sequence Entropy Is Closely Related to Packing Density and Hydrophobicity. *Protein Eng. Des. Sel* 18 (2), 59–64. doi:10.1093/protein/gzi009
- Mahn, A., Lienqueo, M. E., and Salgado, J. C. (2009). Methods of Calculating Protein Hydrophobicity and Their Application in Developing Correlations to Predict Hydrophobic Interaction Chromatography Retention. *J. Chromatogr. A* 1216 (10), 1838–1844. doi:10.1016/j.chroma.2008.11.089
- Malakhova, M., D'Angelo, I., Kim, H. G., Kurinov, I., Bode, A. M., and Dong, Z. (2010). The crystal Structure of the Active Form of the C-Terminal Kinase Domain of Mitogen- and Stress-Activated Protein Kinase 1. *J. Mol. Biol.* 399 (1), 41–52. doi:10.1016/j.jmb.2010.03.064
- Millward, M. J., House, C., Bowtell, D., Webster, L., Olver, I. N., Gore, M., et al. (2006). The Multikinase Inhibitor Midostaurin (PKC412A) Lacks Activity in Metastatic Melanoma: a Phase IIA Clinical and Biologic Study. *Br. J. Cancer* 95 (7), 829–834. doi:10.1038/sj.bjc.6603331
- Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., et al. (2021). Advances in De Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* 22 (4), 1676. doi:10.3390/ijms22041676

- National Cancer Institute Dictionary (2021). Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/n-benzoyl-staurosporine> (Accessed December 4, 2021).
- Neuhäuser, M., and Manly, B. F. (2004). The Fisher-Pitman Permutation Test when Testing for Differences in Mean and Variance. *Psychol. Rep.* 94 (1), 189–194. doi:10.2466/pr0.94.1.189-194
- O'Boyle, N., and Dalke, A. (2018). DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv*. doi:10.26434/chemrxiv.7097960.v1
- Öztürk, H., Özgür, A., Schwaller, P., Laino, T., and Ozkirimli, E. (2020). Exploring Chemical Space Using Natural Language Processing Methodologies for Drug Discovery. *Drug Discov. Today* 25 (4), 689–705. doi:10.1016/j.drudis.2020.01.020
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Proceedings of the Thirty-third Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, December 8–14, 2019.
- Pu, L., Govindaraj, R. G., Lemoine, J. M., Wu, H. C., and Brylinski, M. (2019). DeepDrug3D: Classification of Ligand-Binding Pockets in Proteins with a Convolutional Neural Network. *Plos Comput. Biol.* 15 (2), e1006718. doi:10.1371/journal.pcbi.1006718
- Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* 4 (1), 120–131. doi:10.1021/acscentsci.7b00512
- Shi, W., Lemoine, J. M., Shawky, A. A., Singha, M., Pu, L., Yang, S., et al. (2020). BionoiNet: Ligand-Binding Site Classification with Off-The-Shelf Deep Neural Network. *Bioinformatics* 36 (10), 3077–3083. doi:10.1093/bioinformatics/btaa094
- Shi, W., Singha, M., Pu, L., Ramanujam, J., and Brylinski, M. (2021). Graphsite: Ligand-Binding Site Classification Using Deep Graph Neural Network. *bioRxiv*, 2021.12.06.471420.
- Shou, J., Massarweh, S., Osborne, C. K., Wakeling, A. E., Ali, S., Weiss, H., et al. (2004). Mechanisms of Tamoxifen Resistance: Increased Estrogen Receptor-HER2/neu Cross-Talk in ER/HER2-positive Breast Cancer. *J. Natl. Cancer Inst.* 96 (12), 926–935. doi:10.1093/jnci/djh166
- Simonovsky, M., and Meyers, J. (2020). DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J. Chem. Inf. Model.* 60 (4), 2356–2366. doi:10.1021/acs.jcim.9b00554
- Trebosc, V., Gartenmann, S., Tötzel, M., Lucchini, V., Schellhorn, B., Pieren, M., et al. (2019). Dissecting Colistin Resistance Mechanisms in Extensively Drug-Resistant *Acinetobacter Baumannii* Clinical Isolates. *mBio* 10 (4), e01083. doi:10.1128/mBio.01083-19
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). "Show and Tell: A Neural Image Caption Generator," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, June 8–10, 2015, 3156–3164. doi:10.1109/cvpr.2015.7298935
- Vinyals, O., Bengio, S., and Kudlur, M. (2016). "Order Matters: Sequence to Sequence for Sets," in Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, May 2–4, 2016, 3156–3164. doi:10.1109/cvpr.2015.7298935 arXiv preprint arXiv:1511.06391.
- Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., et al. (2012). PubChem's BioAssay Database. *Nucleic Acids Res.* 40 (Database issue), D400–D412. doi:10.1093/nar/gkr1132
- Weininger, D. (1988). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* 28 (1), 31–36. doi:10.1021/ci00057a005
- Wu, K. J., Lei, P. M., Liu, H., Wu, C., Leung, C. H., and Ma, D. L. (2019). Mimicking Strategy for Protein-Protein Interaction Inhibitor Discovery by Virtual Screening. *Molecules* 24 (24), 4428. doi:10.3390/molecules24244428
- Xu, J., and Zhang, Y. (2010). How Significant Is a Protein Structure Similarity with TM-Score = 0.5? *Bioinformatics* 26 (7), 889–895. doi:10.1093/bioinformatics/btq066
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in International conference on machine learning (Lille, France: PMLR).
- Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K., and Jegelka, S. (2018). "Representation Learning on Graphs with Jumping Knowledge Networks," in International Conference on Machine Learning (Stockholm, Sweden: PMLR).
- Xu, M., Ran, T., and Chen, H. (2021). De Novo molecule Design through the Molecular Generative Model Conditioned by 3D Information of Protein Binding Sites. *J. Chem. Inf. Model.* 61 (7), 3240–3254. doi:10.1021/acs.jcim.0c01494
- Yasonik, J. (2020). Multiobjective De Novo Drug Design with Recurrent Neural Networks and Nondominated Sorting. *J. Cheminform* 12 (1), 14–19. doi:10.1186/s13321-020-00419-6
- Yeturu, K., and Chandra, N. (2011). PocketAlign a Novel Algorithm for Aligning Binding Sites in Protein Structures. *J. Chem. Inf. Model.* 51 (7), 1725–1736. doi:10.1021/ci200132z
- Yu, Q. (2005). The Dynamic Roles of Angiopoietins in Tumor Angiogenesis. *Future Oncol.* 1 (4), 475–484. doi:10.2217/14796694.1.4.475
- Zhang, Y., and Skolnick, J. (2004). Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins* 57 (4), 702–710. doi:10.1002/prot.20264
- Zhang, Y., and Skolnick, J. (2005). TM-align: a Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* 33 (7), 2302–2309. doi:10.1093/nar/gki524
- Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *Bioinformatics* 34 (13), i457–i466. doi:10.1093/bioinformatics/bty294

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Shi, Singha, Srivastava, Pu, Ramanujam and Brylinski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Set of Experimentally Validated Decoys for the Human CC Chemokine Receptor 7 (CCR7) Obtained by Virtual Screening

Matic Proj¹, Steven De Jonghe², Tom Van Loy², Marko Jukič^{3,4}, Anže Meden¹, Luka Ciber⁵, Črtomir Podlipnik⁵, Uroš Grošelj⁵, Janez Konc⁶, Dominique Schols² and Stanislav Gobec^{1*}

¹Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Ljubljana, Ljubljana, Slovenia, ²Laboratory of Virology and Chemotherapy, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, KU Leuven, Leuven, Belgium, ³Faculty of Chemistry and Chemical Engineering, Laboratory of Physical Chemistry and Chemical Thermodynamics, University of Maribor, Maribor, Slovenia, ⁴Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Koper, Slovenia, ⁵Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana, Slovenia, ⁶National Institute of Chemistry, Ljubljana, Slovenia

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Martin Gustavsson,
University of Copenhagen, Denmark
James Edward Pease,
Imperial College London,
United Kingdom
Maha Abedalwahab Habash,
Aqaba University of Technology,
Jordan

*Correspondence:

Stanislav Gobec
stanislav.gobec@ffa.uni-lj.si

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 15 January 2022

Accepted: 28 February 2022

Published: 18 March 2022

Citation:

Proj M, De Jonghe S, Van Loy T,
Jukič M, Meden A, Ciber L,
Podlipnik Č, Grošelj U, Konc J,
Schols D and Gobec S (2022) A Set of
Experimentally Validated Decoys for
the Human CC Chemokine Receptor 7
(CCR7) Obtained by Virtual Screening.
Front. Pharmacol. 13:855653.
doi: 10.3389/fphar.2022.855653

We present a state-of-the-art virtual screening workflow aiming at the identification of novel CC chemokine receptor 7 (CCR7) antagonists. Although CCR7 is associated with a variety of human diseases, such as immunological disorders, inflammatory diseases, and cancer, this target is underexplored in drug discovery and there are no potent and selective CCR7 small molecule antagonists available today. Therefore, computer-aided ligand-based, structure-based, and joint virtual screening campaigns were performed. Hits from these virtual screenings were tested in a CCL19-induced calcium signaling assay. After careful evaluation, none of the *in silico* hits were confirmed to have an antagonistic effect on CCR7. Hence, we report here a valuable set of 287 inactive compounds that can be used as experimentally validated decoys.

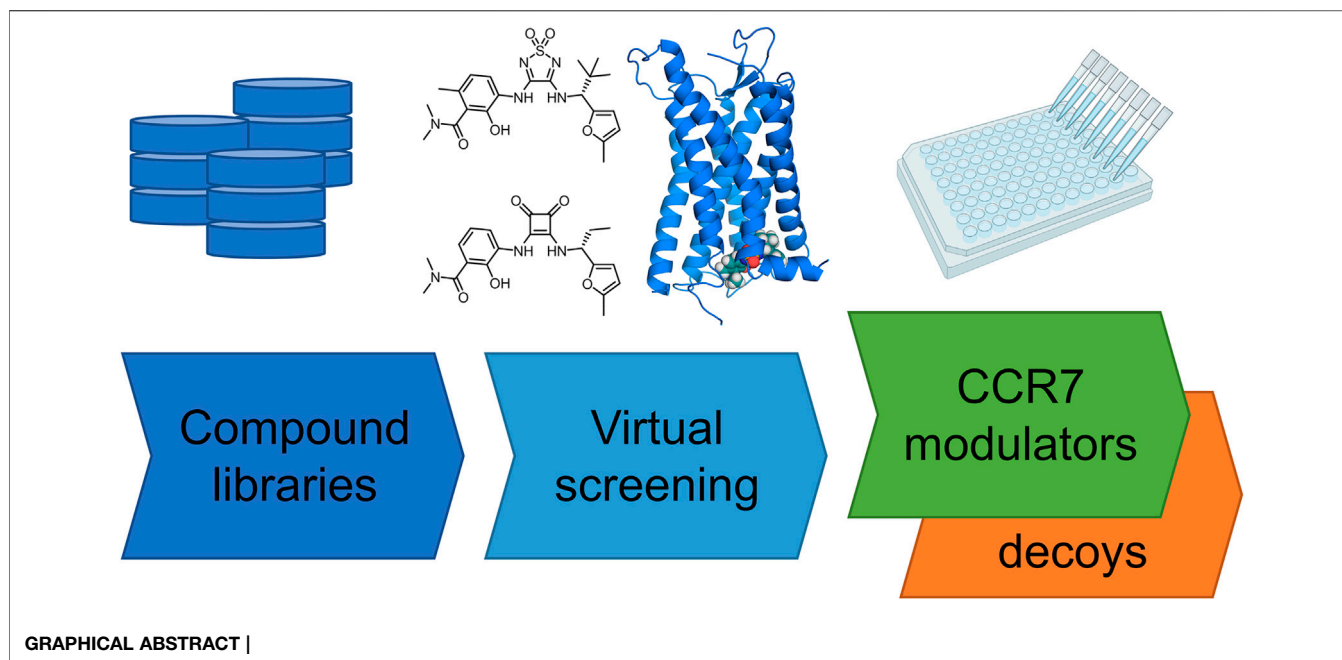
Keywords: virtual screening, decoys, chemical library, computer-aided drug design, CCR7, GPCR

1 INTRODUCTION

Chemokines (chemoattractant cytokines) are small, secreted proteins (~10 kDa) that were first described as essential mediators of immune cell migration throughout the human body. They are characterized by conserved N-terminal cysteine (C) residues (i.e., C, CC, CXC, and CX3C chemokines, where X is a variable amino acid) and exert their function through activation of G protein-coupled receptors (GPCRs). About 20 human chemokine receptors and approximately 50 different human chemokines are known. A given chemokine receptor can sometimes be activated by more than one chemokine and, at the same time, a particular chemokine can signal via multiple receptors (Griffith et al., 2014).

The CC chemokine receptor 7 (CCR7) is crucial for lymphoid organogenesis and the recruitment of naïve T lymphocytes and activated dendritic cells towards the lymph nodes, where they initiate the

Abbreviations: ATP, adenosine triphosphate; CCR7, CC Chemokine Receptor 7; GPCRs, G protein-coupled receptors; LBVS, ligand-based virtual screening; LiSiCA, ligand similarity using clique algorithm; ProBiS-Dock, protein binding sites docking; ROC, receiver-operating characteristic; ROCS, rapid overlay of chemical structures; RMSD, root-mean-square deviation; SBVS, structure-based virtual screening.



immune response (Zlotnik et al., 2011). CCR7 can be activated by two receptor ligands, the chemokines CCL19 and CCL21 that bind with high affinity to CCR7 (Sullivan et al., 1999). Unlike CCL19, CCL21 harbors an extended and highly positively charged C-terminal tail that mediates strong binding to glycosaminoglycans (GAGs) expressed at the cell surface (Barmore et al., 2016). Several studies revealed the biased signaling properties of CCL19 and CCL21 and indicate that both chemokines differentially target CCR7 in terms of G protein activation, β -arrestin recruitment and receptor internalization (Kohout et al., 2004; Corbisier et al., 2015; Hjortø et al., 2016). CCR7 signaling can contribute to the progression of severe human diseases. Tumor cells of diverse origins can hijack CCR7-mediated migration to metastasize, primarily to the lymph nodes (Zlotnik et al., 2011; Jørgensen et al., 2018). Recruitment of leukaemic T cells to the central nervous system is also dependent on CCR7 (Buonamici et al., 2009). Other human diseases associated with CCR7 signaling include chronic inflammatory diseases (e.g., rheumatoid arthritis) (Moschovakis and Förster, 2012). Hence, CCR7 has emerged as a promising therapeutic target, but remains understudied from a drug discovery perspective.

Even though CCR7 is implicated in various human diseases, to the best of our knowledge, no selective and potent small molecule antagonists for CCR7 have been developed so far. Recently, a high-throughput screening of 150,000 compounds using Chinese hamster ovary (CHO)-K1 cells expressing human or murine CCR7 in a β -arrestin recruitment assay was described (Hull-Ryde et al., 2018). The most potent CCR7 antagonist that emerged from this campaign was **cosalane** (Figure 1) with a half maximal inhibitory concentration (IC_{50}) value of 0.2 μ M (when CCL19 was used as the natural CCR7 ligand) and 2.7 μ M (when CCL21 was used as the agonist). In

addition, **cosalane** exhibited nearly identical activity against the human and murine CCR7 orthologues. However, the high lipophilicity of **cosalane** and its complex chemical structure make it unattractive as lead structure for further chemical optimization. Recently, the X-ray co-crystal structure of CCR7, complexed with **cmp2105** (Figure 1), was solved (Jaeger et al., 2019). This compound was shown to bind to a conserved allosteric G_i protein binding pocket at the intracellular side of the receptor. Validation of its CCR7 binding was performed in a membrane-based competition binding experiment with radiolabeled CCL19, in which an IC_{50} value of 35 nM was determined for **cmp2105**. Furthermore, CCR7 antagonism of **cmp2105** was confirmed in a cell-based β -arrestin recruitment assay, which yielded an IC_{50} value of 7.3 μ M (Jaeger et al., 2019). **Cmp2105** was initially discovered by screening in a CCR7 thermal-shift assay. **Navarixin** (Figure 1) also displayed a thermostabilizing effect in this assay and subsequently an IC_{50} value of 33.9 μ M was determined in the β -arrestin recruitment assay (Jaeger et al., 2019). Other analogues (i.e., **CS-1**, **CS-2**, and **CS-3**, Figure 1) also proved to be hits in the thermofluor stability assay, albeit less potent than **cmp2105** and **navarixin**, and were not further pharmacologically validated (Jaeger et al., 2019). Other known chemokine receptor ligands, such as **verciron** (a CCR9 antagonist) and **maraviroc** (a CCR5 antagonist) completely lacked the ability to thermally stabilize CCR7 (Jaeger et al., 2019).

To improve our understanding of the role of CCR7 in various pathologies, there is a clear need for potent, drug-like, and selective CCR7 antagonists that can be used as chemical probes to validate CCR7 as a drug target. In addition, these chemical tools can be used as starting points for medicinal

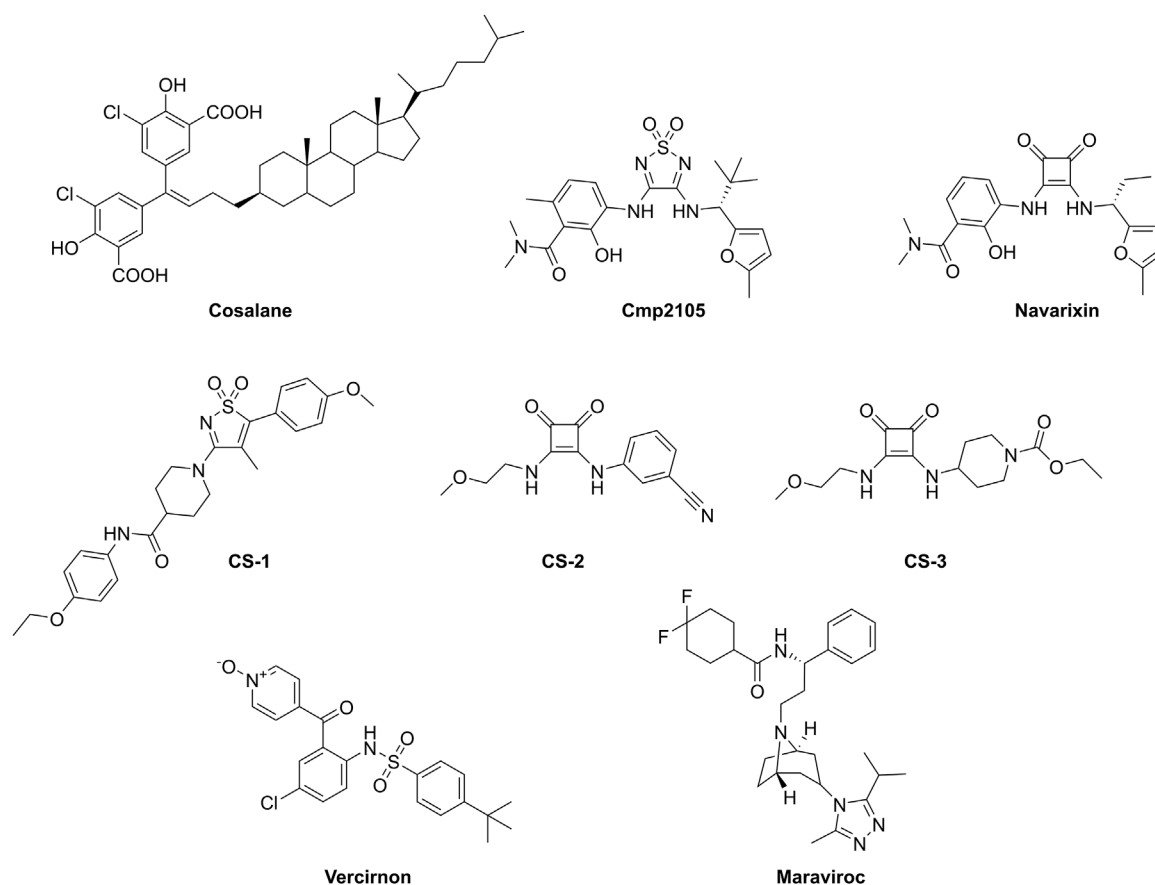


FIGURE 1 | Compounds studied as CCR7 antagonists.

chemistry-based optimization campaigns. In this study, we describe a virtual screening workflow, followed by experimental validation, in search for novel CCR7 small molecule antagonists. Known CCR7 ligands from the patent and scientific literature, whose CCR7 antagonism was independently confirmed, were used as starting points for ligand-based virtual screening (LBVS) protocols. In addition, the recently published crystal structure of CCR7 was used to perform molecular docking and to generate a pharmacophore model.

2 MATERIALS AND METHODS

2.1 GPCR Assays

2.1.1 CCR7 Competition Binding Assay

Human U87 glioblastoma cells that stably express CD4 and the human CCR7 receptor (U87.CD4.CCR7) were used to determine CCR7 binding affinity, essentially by adopting a previously published protocol used to study binding to another chemokine GPCR, CXCR4 (Schoofs et al., 2018). U87.CD4 cells that do not overexpress CCR7 were used as control cells

to evaluate the level of non-specific cell binding of the fluorescently labeled ligand (Alexa-Fluor647 labeled CCL19, CCL19^{AF647}, Almac, United Kingdom). In brief, U87.CD4.CCR7 cells were pre-incubated with compound (at different concentrations) in 150 μ L assay buffer [Hank's balanced salt solution (HBSS), 20 mM HEPES and 0.5% Fetal Calf Serum] for 15 min at room temperature (RT) in the dark. Afterwards, 50 μ L of CCL19^{AF647} was added (25 ng/mL final concentration) and samples were incubated for another 30 min at RT, protected from light. Then, cells were washed twice with assay buffer and fixed in 1% paraformaldehyde in Dulbecco's phosphate-buffered saline (DPBS).

Samples were immediately analyzed by flow cytometry (FACS Canto II, BD). Data were analysed using Flowjo. The percentage inhibition of CCL19^{AF647} binding was calculated according to formula $\{1 - [(MFI_X - MFI_{NC}) / (MFI_{PC} - MFI_{NC})]\} \times 100$, where MFI_X is the mean fluorescence intensity (MFI) of the compound-treated sample, MFI_{NC} the MFI of the negative control (i.e., autofluorescence of untreated and unlabeled cells) and MFI_{PC} the MFI of the positive control (i.e., cells exposed to CCL19^{AF647} only). IC_{50} values (i.e., the compound concentration that inhibits CCL19^{AF647} binding by 50%) were calculated using

four parameter non-linear curve fitting in GraphPad Prism 9.0.2. For each experiment the stain index (SI) was calculated as the ratio of the separation between MFI_{PC} and MFI_{NC} , divided by two times the standard deviation of MFI_{NC} .

2.1.2 Calcium Mobilization Assays

U87.CD4 cells that stably express either human CCR7, CXCR2, CCR5 or CXCR4 were seeded (20,000 cells/well) in gelatin-coated black-walled polystyrene 96-well plates with clear bottom and incubated overnight at 37°C and 5% CO₂. The next day, a fluorescent Ca²⁺-sensitive dye solution (Fluo-2 AM) was prepared as described before (Claes et al., 2018). Culture medium was removed, and cells were incubated for 45 min at room temperature in the dark. Meanwhile, 96-well polypropylene plates containing 5-fold concentrated compound dilutions and 5-fold concentrated solution of chemokine ligands (CCL19, CXCL8, LD78-β, CXCL12, respectively; all purchased from PeproTech) were prepared for use with the FLIPR Tetra device (Molecular Devices) as described before (Claes et al., 2018). The antagonistic properties of the compounds were calculated based on their capacity to inhibit the Ca²⁺ release induced by a fixed concentration of chemokine (i.e. 50 ng/mL final concentration for CCL19, CXCL8 and CXCL12 and 100 ng/mL for LD78-β), as described (Claes et al., 2018). Exactly the same protocol was used to record calcium responses in Chinese hamster ovary (CHO)-K1 cells, upon stimulation with adenosine triphosphate (ATP, purchased from Sigma).

2.2 Preparation of Chemical Libraries

2.2.1 Active Compounds and Generated Decoys

The survey of patent and scientific literature revealed the existence of eight CCR7 antagonists (**Supplementary Figure S1**). Based on this, a set of 600 decoy molecules (see Supporting Excel file) was generated using DUD-E server (Mysinger et al., 2012). The generated decoys have similar physicochemical properties (molecular weight, estimated water-octanol partition coefficient (miLogP), rotatable bonds, hydrogen bond acceptors, hydrogen bond donors, and net charge), but have a different 2D topology when compared to the active compounds. Decoys can be used as alternatives to experimentally confirmed inactive compounds for the purpose of model validation.

Primary literature search identified an additional 104 compounds active on multiple chemokine receptors, namely CCR1, CCR2, CCR3, CCR4, CCR5, CCR7, CCR8, CCR9, CCR10, CXCR1, CXCR2, CXCR3, CXCR4, and CXCR7 (**Supplementary Table S1**). Those compounds were used to construct a focused chemokine receptor targeted compound library, as described below.

2.2.2 FKKTlib Academic Compound Library

The FKKTlib academic compound library currently contains 3,428 unique synthesized compounds resulting from many years of research across various projects at the University of Ljubljana, Faculty of Chemistry and Chemical Technology. Most of the compounds in this library are heterocycles that are

documented in the scientific literature. Most of the samples are available as solids and are stored in cryogenic vials labelled with a QR code that allows for quick retrieval of the samples. To ensure the stability of the samples, they are stored under argon at -25°C. Information about the compounds in the library is stored in a web-based, fully retrievable molecular structure database based on the open-source solution MolDB6, developed by Prof. Norbert Haider from the University of Vienna (Haider, 2010). The system uses MySQL as a database engine, and the molecular structures with their corresponding data are stored in MySQL tables. The check/matchmol programme is used for structure or substructure searches, which is performed in a two-step procedure: pre-selection by fingerprint matching, followed by a complete atom-by-atom comparison of the remaining candidates. Structures and data can be added via the web interface or by importing from an MDL SD file using a Perl script on the server. The library is freely accessible at: <https://knjiznica-spojini.fkkt.uni-lj.si/fkktlib/>.

2.2.3 ZINC Library

The ZINC in-stock subset (Sterling and Irwin, 2015), containing 13.7 million drug-like compounds, was used for the virtual screening using the Ligand Similarity Using Clique Algorithm (LiSiCA) software (Lešnik et al., 2015). The ZINC subset was first filtered using the FILTER 3.1.2.2 software (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com), eliminating known or predicted aggregators, compounds containing metals, and compounds with reactive functional groups, and retaining only compounds with appropriate molecular weights (200–800 Da) and partition coefficients (-4.0–6.9) (see Supplementary Material for FILTER configuration file). The filtered ZINC library contained 8.9 million compounds. Finally, the stereoisomer and conformational model generator OMEGA 3.1.2.2 (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com) was used to enumerate stereocenters and to generate up to 30 conformers per compound.

2.2.4 Chemokine Receptor Targeted Compound Library

The ZINC in-stock subset (Sterling and Irwin, 2015) was also used for the construction of a library covering compounds targeting chemokine receptors. FP2 molecular fingerprints were calculated for the 104 compounds targeting various chemokine receptors (details in Supplementary Material) as well as for the complete ZINC subset. Using OpenBabel (v2.3.0), a similarity search in ZINC was carried out with 104 queries and a Tanimoto index of ≥0.5 to obtain a similarity library of 951,471 unique structures. The similarity library was then filtered using the FILTER software (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com), as described beforehand, to obtain a focused chemokine receptor library of 539,814 compounds. Finally, OMEGA (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com) was used to enumerate all possible stereocenters and generate up to 10,000 conformers per compound (RMS of 0.3).

2.2.5 MolPort Library

The second library was prepared from the MolPort database of in-stock compounds (7.5 million). It was used for core motif substructure searches, virtual screening with ROCS, docking with FRED, Glide, and ProBiS-Dock. Duplicates were removed using OpenBabel 2.4.1 (O'Boyle et al., 2011), and the database was processed using the FILTER 3.1.2.2 software (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com), eliminating known or predicted aggregators, compounds containing metals, and compounds with reactive functional groups, retaining only compounds with appropriate molecular weights (200–800 Da) and partition coefficients (−4.0–6.9) (see Supplementary Material for FILTER configuration file). Compounds known to cause interference in assay systems (Dahlin et al., 2015) were removed using the RDKit Molecule Catalog Filter node (catalog PAINS A) (RDKit: Open-source cheminformatics, 2021) as implemented in the KNIME platform (Berthold et al., 2007). Compounds with reactive functional groups (Brenk et al., 2008) were also removed. Protonation states at pH 7.4 were generated using OpenBabel 2.4.1 (O'Boyle et al., 2011). The final library contained 3.5 million compounds. Finally, the stereoisomer and conformational model generator OMEGA 3.1.2.2 (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com) was used to enumerate stereocenters and generate up to 200 conformers per compound.

2.2.6 Diversity Set of Compounds Available From Trusted Commercial Vendors

The third library used for pharmacophore-based screening contained 1.1 million compounds based on curated diversity sets from Asinex, ChemBridge, ChemDiv, Enamine, KeyOrganics, and Pharmeks. The libraries were downloaded in SDF format, merged, and duplicates removed using the LigandScout database Merger and Duplicates Remover nodes implemented in the Inte:Ligand Expert KNIME Extensions. Protonation states at pH 7.4 were generated using OpenBabel 2.4.1 (O'Boyle et al., 2011). Finally, a maximum of 200 conformations were generated for each molecule using the iCon algorithm of LigandScout (Poli et al., 2018) with default “BEST” settings and saved in LDB (LigandScout database format) using the idbgen algorithm.

2.3 Core Motif Substructure Search

Core motif substructure searches were performed using SMILES filters applied to the MolPort library and the FKKTlib. The core motifs of cyclobutenedione (with an additional nitrogen atom) and thiourea were defined by SMILES expressions O=C1C=C(N)C1=O and NC(N)=S, respectively. First, MolPort library filtering was performed using the RDKit substructure filter node (RDKit: Open-source cheminformatics, 2021) as implemented in the KNIME analytics platform (Berthold et al., 2007). 2,452 cyclobutenediones were extracted from the MolPort database and docked to the prepared CCR7 receptor using Glide XP (Schrödinger Suite 2020-2, Schrödinger, LLC, New York, NY, 2020) (Friesner et al., 2006) as described below. Of the 100 highest scoring compounds, 16 diverse compounds were selected for

purchase. Second, the MolPort database was searched for the thiourea core motif, which yielded more than 90k available compounds. Duplicates, PAINS (Dahlin et al., 2015), and compounds with reactive functional groups were removed to yield 63k compounds. Docking with Glide SP (Schrödinger Suite 2020-2, Schrödinger, LLC, New York, NY, 2020) was performed as described below, and after clustering, from the top 500 hits, 13 diverse compounds were purchased. Second, the FKKTlib was filtered and all 9 and 13 compounds available in solid form with cyclobutenedione and thiourea core motifs, respectively, were experimentally evaluated.

2.4 LBVS With LiSiCA Software

Ligand-based virtual screening (LBVS) of the ZINC database of purchasable compounds using LiSiCA software (Lešnik et al., 2015) was performed with the bioactive 3D conformation of **cmp2105** as the reference compound (PDB ID: 6QZH, ligand **JLW**) (Jaeger et al., 2019). The double bonds of the thiadiazole-dioxide of the **JLW** ligand were correctly assigned, since they are missing in the PDB structure. Both 2D and 3D options of the LiSiCA were used with all other settings set to default values. From the 200 compounds most similar to the reference **cmp2105** according to the Tanimoto score, 27 diverse compounds were purchased—12 of them arising from the 2D method and 15 were discovered with the 3D method.

2.5 LBVS With ROCS Software

The MolPort library was screened using ROCS 3.3.2.2 software (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com) (Hawkins et al., 2007). For model A, a 3D pose for the first query, **navarixin**, was obtained by docking with Glide XP to the prepared CCR7 receptor, as described below. The bioactive 3D conformation of **cmp2105** (PDB ID: 6QZH, ligand **JLW**) (Jaeger et al., 2019) was used to create models B and C. All models were validated with the set of active compounds and generated decoys. The default settings of ROCS were used for virtual screening of all three queries. Virtual hits were prioritized based on the ComboScore, which considers similarity of 3D shape (“ShapeTanimoto”) and chemical pattern (“ColorScore”). For each query, 27–30 top scoring compounds were purchased from the clustered list of top 100 scoring hits.

2.6 Homology Modelling

Before the release of the CCR7 X-ray crystal, a homology model of CCR7 was built using the structure of human CCR9 (PDB ID: 5LWE; B chain). The template was identified by running 10 PSI Blast iterations on the starting CCR7 (UniProt ID: P32248) sequence to identify five top scoring templates (PDB IDs: 5LWE, 5UIW, 4YAY, 5WB2 and 5UNF) (Müller et al., 1999). The alignment and template was used to build the homology model using YASARA Twinset software (Krieger et al., 2002; Krieger and Vriend, 2015) using the following parameters: speed: Slow, EValue Max: 0.5, Templates Total: 5, Templates SameSeq: OligoState: 4, Alignments: 5, LoopSamples: 50 and TermExtension: 10.18 models were built and each model subjected to an unrestrained energy

minimization with explicit water molecules by simulated annealing employing the YASARA2 force field (Krieger et al., 2009). The models were rated according to a quality Z-score and the best scoring model was used. The latter contained 276 of 378 target residues (73.0%) aligned to template residues. The sequence identity was 46.0% and the sequence similarity 68.1% (BLOSUM62 > 0). The monomer homology model after full unrestrained simulated annealing minimization was rated as optimal by YASARA with internal quality Z-score of 0.110, comprised amino acids 47–352, and was further checked with WHAT-IF test set.

2.7 Receptor Preparation

The CCR7 receptor was prepared from the X-ray crystal structure (PDB ID: 6QZH) (Jaeger et al., 2019) using Protein Preparation Wizard (Schrödinger Suite 2020-2, Schrödinger, LLC, New York, NY, 2020) (Madhavi Sastry et al., 2013). Briefly, missing side chains and missing loop 255–261 were modelled with Prime (Jacobson et al., 2004), hydrogen atoms were added, residues were protonated at pH 7.0, the hydrogen bonding network was refined, waters beyond 3.0 Å from other heteroatoms were removed, and restrained minimization was performed. The double bonds of the thiadiazole-dioxide core in the co-crystallized ligand **cmp2105** (PDB ID: 6QZH, ligand **JLW**) were assigned (Jaeger et al., 2019). Only the allosteric binding site was considered for pharmacophore-based screening and molecular docking.

2.8 Pharmacophore-Based Screening

The prepared CCR7 receptor was used to generate a structure-based pharmacophore model using LigandScout 4.4 (Inte:Ligand GmbH) (Wolber and Langer, 2005). Exclusion volumes defining regions based on the shape of the binding site residues were generated, and all features were converted to vectors. One hydrogen bond donor and one hydrophobic feature were marked as optional. This model was validated with the set of active compounds and generated decoys. Default settings in LigandScout were used. Virtual screening of the diversity set of compounds available from trusted commercial vendors yielded 78 virtual screening hits, which were then visually inspected, clustered according to Morgan fingerprints, and 23 diverse compounds were purchased.

2.9 Molecular Docking With FRED and Glide Software

Molecular docking was performed sequentially with FRED 3.4.0.2 (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com) and Glide software (Schrödinger Suite 2020-2: Glide, Schrödinger, LLC, New York, NY, 2020). First, Make Receptor 3.4.0.2 (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com) was used to define grid box of the allosteric binding site of the prepared receptor. The volume of the box was 6,725 Å³ (17.75 Å × 21.15 Å × 17.92 Å) and size of the outer contour was reduced to 1,139 Å³. Re-docking of the co-crystallized ligand **cmp2105** using FRED and Glide SP resulted in a root-mean-square deviation (RMSD) of 0.77 Å and 0.

34 Å, respectively, confirming the validity of the pose prediction during docking. In the same manner, docking with Glide XP (Schrödinger Suite 2020-2, Schrödinger, LLC, New York, NY, 2020) (Friesner et al., 2006) was used to obtain the bioactive 3D conformation of **navarixin**. The MolPort library was docked with FRED and 100,000 highest scoring hits were used for sequential docking with Glide. A 3D structure of one stereoisomer was generated using LigPrep (Schrödinger Suite 2020-2, Schrödinger, LLC, New York, NY, 2020). The prepared receptor's grid box was centered on the co-crystallized ligand and docking was performed using Glide SP (Schrödinger Suite 2020-2, Schrödinger, LLC, New York, NY, 2020) (Friesner et al., 2004). The 100 highest scoring virtual hits were clustered according to Morgan fingerprints, and 46 diverse compounds were purchased.

2.10 Molecular Docking With ProBiS-Dock Algorithm

Molecular docking with ProBiS-Dock algorithm (Konc et al., 2022) was performed with the prepared CCR7 receptor. Similar receptors with allosterically bound ligands that were known at the time of screening, i.e., CCR9 (PDB ID: 5LWE, ligand **79K** [**vercirnon**]) (Oswald et al., 2016) and CCR2 (PDB ID: 5T1A, ligand **VT5**) (Zheng et al., 2016), were aligned to the prepared CCR7 receptor. All three ligands, namely **cmp2105**, **vercirnon** and **VT5**, were extracted and used as template ligands that are required for molecular docking with ProBiS-Dock. Re-docking of the co-crystallized ligand **cmp2105** with an RMSD of 0.77 Å was performed for validation. From the MolPort library, one million compounds were randomly selected and used for virtual screening. The top 450 virtual hits were clustered according to Morgan fingerprints, and 40 diverse compounds were purchased.

2.11 Coupled Virtual Screening Approach

For the LBVS coupled to SBVS approach, we examined similar protein complexes published in the PDB database using ProBiS server (<https://probis.nih.gov/>) to identify all possible binding sites and small-molecule binding modes on the CCR7. Therefore, the built CCR7 homology model was used as an input for ProBiS calculation and one binding site identified (*binding site one in ProBiS; proximity of ligand vercirnon from CCR9 PDB ID: 5LWE*) (Konc and Janežič, 2012). The postulated binding site comprised of two pockets was defined by residues: Thr93, Leu147, Ile150, Val264, Ile265, Val268 and Val79, Val80, Thr82, Tyr83, Phe86, Asp94, Thr95, Leu97, Leu98, Leu100, Asp110, Asp336. Model was later validated with an all-atom RMSD of 1.942 Å towards PDB 6QZH crystal with allosteric site correctly identified relative to crystal **JLW** ligand. With the binding site defined, the receptor structure was generated using OEDocking 3.2.0.2 software package (OpenEye Scientific Software, Inc, Santa Fe, NM, United States; www.eyesopen.com) with MakeReceptor. A box with the volume of 5,805 Å³ (18,33 × 19,00 × 16,67 Å) was defined around reference ligand **vercirnon**. A balanced site shape potential was calculated where docking volume was 587 Å³. No

constraints were used. Docking of the chemokine receptor targeted library (539,814 compounds) to the prepared receptor was performed using FRED from OpenEye as described above with dock_resolution parameter set to *High*. Top 100 scoring compounds according to Chemgauss4 score were collected, clustered according to Morgan fingerprints (20 clusters, average RMSD linkage) and best scoring representatives selected for purchase and testing.

3 RESULTS AND DISCUSSION

3.1 Experimental Hit Validation

Before we initiated an extensive *in silico*-based screening program, several compounds described in the literature were either resynthesized (see Supplementary material) or purchased from commercial vendors to confirm their CCR7 antagonism in various pharmacological assays. Compounds previously shown to be CCR7 antagonists (i.e., **cmp2105**, **navarixin**, **CS-2**, and **CS-3**), as well as inactive control compounds (**maraviroc** and **vercirnon**), were included in this study (Figure 1; Table 1). A competition binding assay was established based on the specific interaction of fluorescently labeled CCL19 with CCR7 overexpressed on whole living cells (Figure 2A). Using CCL19^{AF647} as a tracer, the binding affinity for **cmp2105**, **navarixin**, **CS-2** and **CS-3** was evaluated (Figure 2B; Table 1). Whereas dose-dependent inhibition of CCL19^{AF647} binding was confirmed for **cmp2105** and **navarixin**, **CS-2** and **CS-3** did not show any CCR7 affinity. The previously observed stabilizing effect of **CS-2** and **CS-3** in thermal stability experiments was much smaller than for **cmp2105** and **navarixin** (Jaeger et al., 2019), suggesting a very low binding affinity for CCR7. It should further be noted that for **cmp2105** an IC₅₀ value of 35 nM was previously reported when this compound was assessed in membrane-based competition experiments using radioactively labeled CCL19 (Jaeger et al., 2019). The fact that in our assay whole cells are used instead of membrane preparation, which requires the compound to first enter the cell before reaching its

intracellular binding pocket, may therefore partly explain the increased apparent IC₅₀ value observed here.

Reference compounds were also evaluated in a CCR7 kinetic, fluorescence-based calcium mobilization assay. **cmp2105** and **navarixin** showed IC₅₀ values in the 5–15 μM range for antagonizing the CCL19-induced calcium response (Figure 3; Table 1) in line with their CCR7 antagonistic activity previously determined in a β-arrestin recruitment assay (Jaeger et al., 2019). In agreement with the lack of observed binding affinity, **CS-2** and **CS-3** were also inactive in this CCR7 calcium mobilization assay (Table 1). Furthermore, the absence of activity of **vercirnon** and **maraviroc** in the calcium mobilization assay is in agreement with their lack of activity in the thermal shift assay (Jaeger et al., 2019).

3.2 Virtual Screening Campaign

To expand the current set of potent CCR7 modulators, we launched a virtual screening campaign. Based on known CCR7 ligands, in particular **cmp2105** and **navarixin**, an LBVS was performed. Furthermore, a recently published crystal structure of the receptor (Jaeger et al., 2019) was used for structure-based virtual screening (SBVS). In addition to libraries of commercially available compounds, we also used the FKKTlib academic library for screening (Figure 4).

LBVS is commonly used in drug discovery and is based on the assumption that structurally similar compounds have similar biological properties. Various metrics are used to express similarity between compounds (Maggiora et al., 2014). We started with a simple substructure search for core motifs that are typical for small molecule ligands targeting the intracellular binding sites of various chemokine receptors. Thioureas bind to an intracellular binding pocket of CXCR2 (Nicholls et al., 2008) and cyclobutenediones have been shown to bind intracellularly to CXCR2 (Liu et al., 2021, 2) and CCR7 (Jaeger et al., 2019). This approach was applied to both the FKKTlib academic library and a library of commercially available compounds.

Second, we used ligand-based virtual screening software LiSiCA (Lešnik et al., 2015) to find compounds with different scaffolds and core motifs than those of the reference compound **cmp2105**. LiSiCA

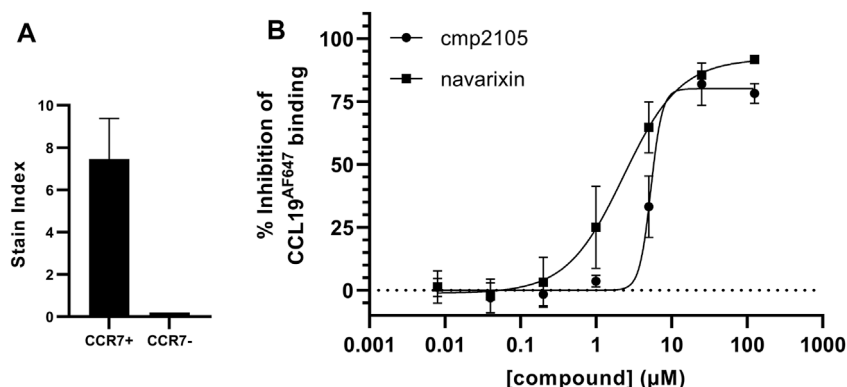


FIGURE 2 | Inhibition of CCL19^{AF647} binding by **cmp2105** and **navarixin** (A) Incubation of U87 cells that overexpress CCR7 (CCR7+) with CCL19^{AF647} generates a strong fluorescent binding signal, which is not present when CCL19^{AF647} is incubated with cells that do not overexpress CCR7 (CCR7-) (Mean stain index ±SD of two (CCR7-) or four (CCR7+) independent experiments) (B) Dose dependent inhibition of CCL19^{AF647} binding by **cmp2105** and **navarixin**.

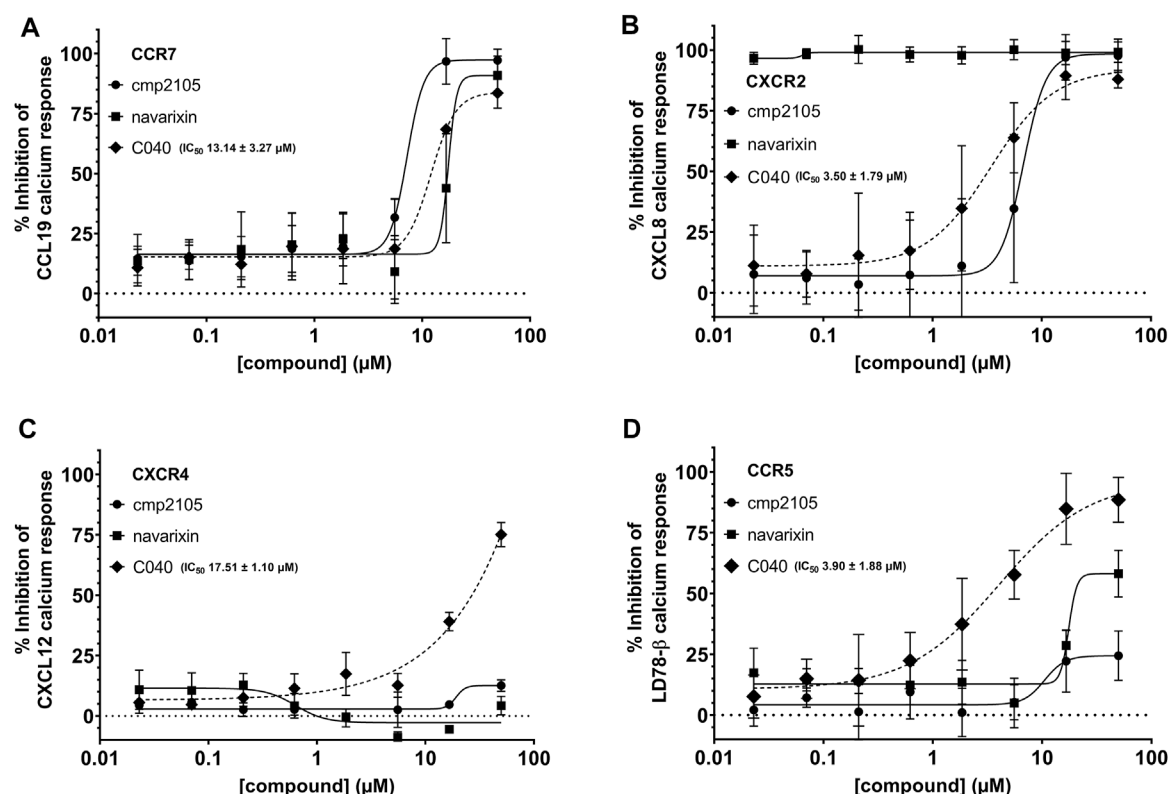


FIGURE 3 | Inhibition of the intracellular Ca^{2+} release. The ability of **cmp2105**, **navarixin**, and **C040** to inhibit the Ca^{2+} response induced by (A) CCL19-CCR7 (B) CXCL8-CXCR2 (C) CXCL12-CXCR4, and (D) LD78- β -CCR5 was evaluated. Mean \pm SD of at least three independent experiments is shown.

is based on a graph-theoretical representation of molecules and uses a fast maximum clique algorithm (Konc and Janežič, 2007) to search for 2D or 3D similarities between a reference compound and a database of target compounds. The similarities found are expressed by the Tanimoto coefficients. A library of commercially available compounds was compared to the reference based on both 2D and 3D molecular representations.

TABLE 1 | CCL19 competition binding, CCR7 calcium mobilization and CCR7 β -arrestin data of reference compounds.

| Compound | β -arrestin IC_{50} (μM) ^a | Calcium assay IC_{50} (μM) ^b | Binding assay IC_{50} (μM) ^c |
|-----------|--|--|--|
| Cmp2105 | 7.3 | 7.30 ± 1.66 | 6.12 ± 2.36 |
| Navarixin | 33.9 | 17.39 ± 1.12 | 2.43 ± 0.98 |
| CS-2 | NA ^d | >50 μM | >50 μM |
| CS-3 | NA ^d | >50 μM | >50 μM |
| Vercimon | NA ^d | >50 μM | ND ^e |
| Maraviroc | NA ^d | >50 μM | ND ^e |

^a IC_{50} : compound concentration inhibiting β -arrestin recruitment in CHO-K1 cells by 50%. Data from (Jaeger et al., 2019).

^b IC_{50} : compound concentration inhibiting CCL19 induced intracellular calcium flux by 50% (Mean \pm SD of at least three independent experiments).

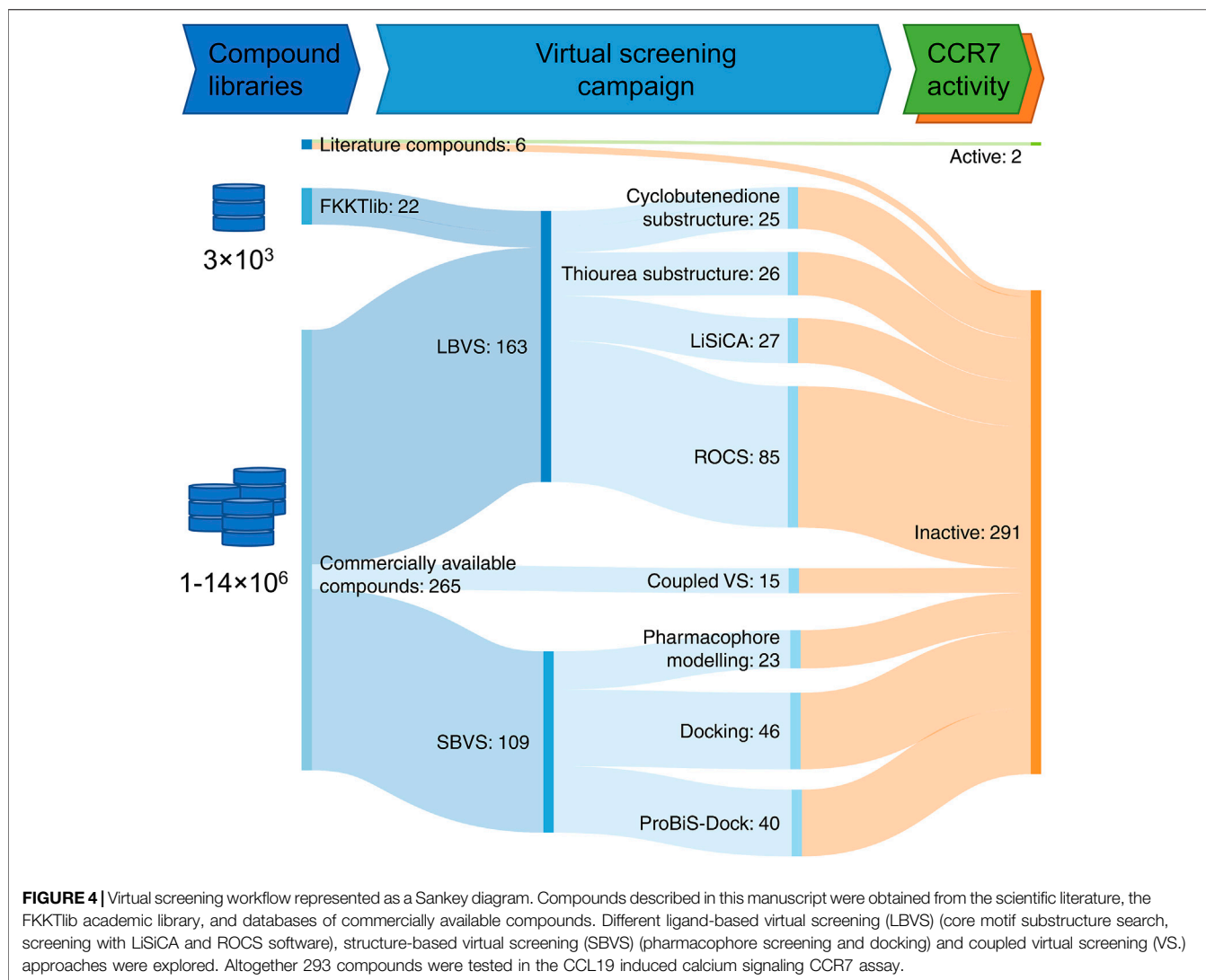
^c IC_{50} : compound concentration inhibiting CCL19^{AF647} binding by 50% (Mean \pm SD of four independent experiments).

^dNA: not available.

^eND: not determined.

Third, 3D shape-based virtual screening was performed by rapid overlay of chemical structures (ROCS) (Hawkins et al., 2007). This method is based on the concept that compounds have a similar shape if their volumes, described by a Gaussian function, overlap well. In addition to molecular volume, a color force field is used to describe other molecular features, such as hydrogen bond donors and acceptors, anions, cations, hydrophobes and rings (Kirchmair et al., 2009). As a starting point for the ROCS search, we modeled a 3D conformation of **navarixin** by docking with Glide XP and used it to generate ROCS model A (Figure 5A). The bioactive conformation of **cmp2105** was extracted from the co-crystal structure (PDB ID: 6QZH) and used directly to create models B and C, which differed in the selection of color features. Only relevant hydrogen bond donors and acceptors based on the distances in the crystal structure were used for the model B (Figure 5B). For model C, only color features in the inner part of the binding pocket were selected, leaving more degrees of freedom for the part of the molecule that extends toward the solvent (Figure 5C). All three models performed well in screening a set of active compounds and generated decoys. The results are presented in the form of receiver-operating characteristic (ROC) curves (Figure 5). Subsequently, the models were used to screen a library of commercially available compounds.

A structure-based pharmacophore model was constructed from the crystal structure of **cmp2105** (Figure 6A). The model consisted

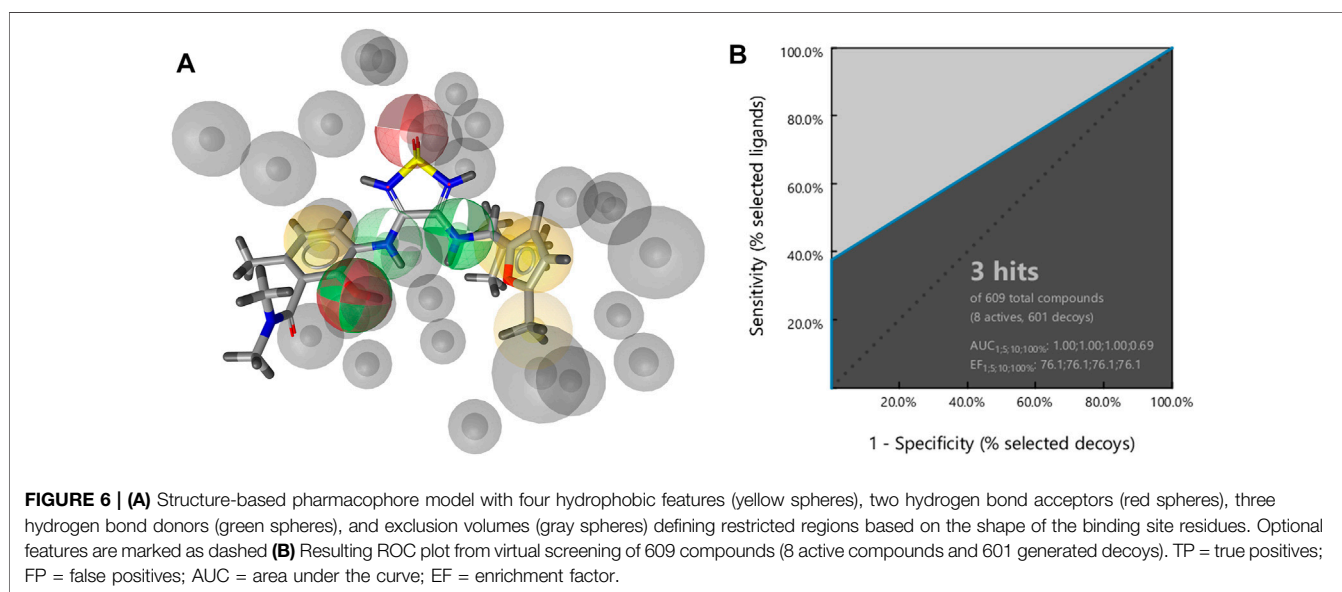
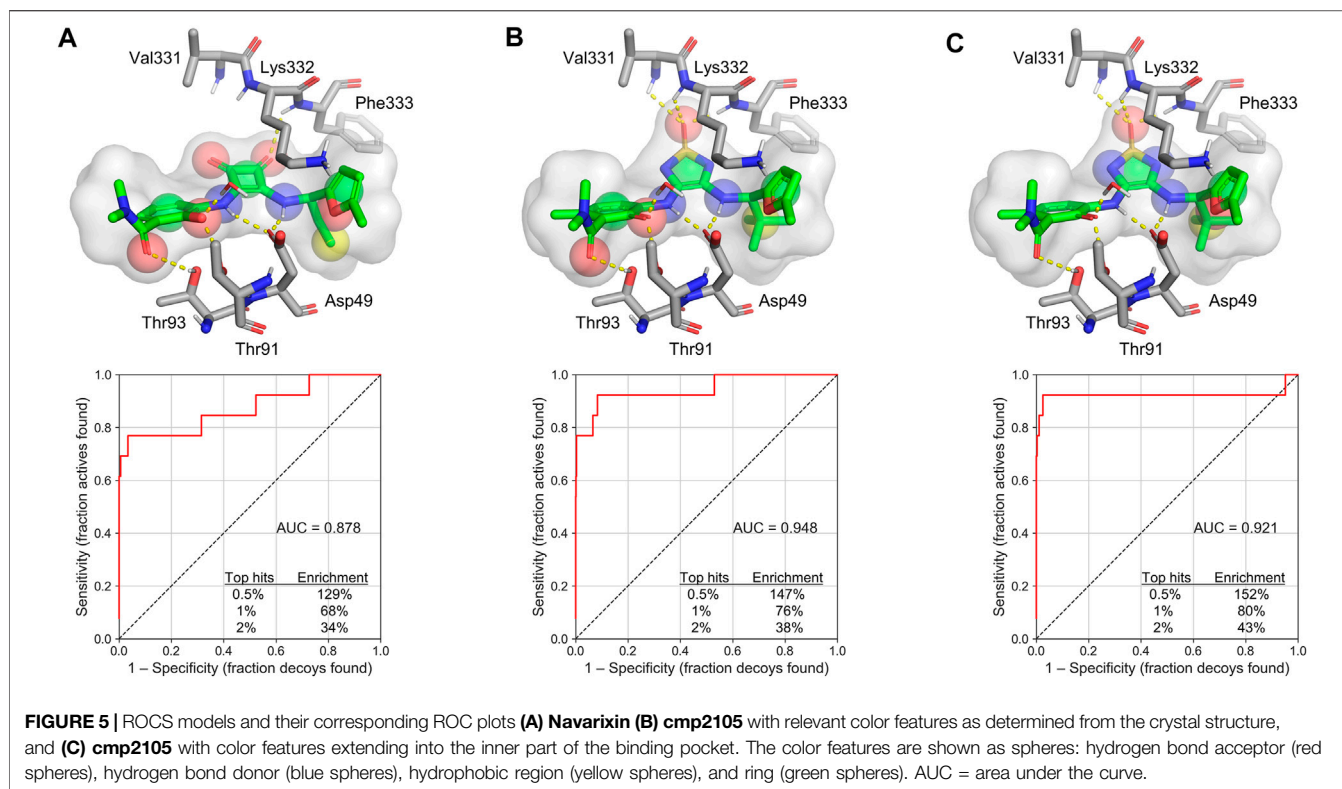


of a hydrogen bond acceptor for the sulfonyl moiety, two hydrogen bond donors for the secondary amines (one labeled as optional), a hydrogen bond donor and acceptor for the phenol moiety, four hydrophobic features, and exclusion spheres. A hydrophobic feature for the methyl moiety on furan ring was also labeled as optional. The results of screening a set of active compounds and generated decoys were visualized by ROC plot, with the rate of active compounds on the y-axis and the rate of decoys on the x-axis (**Figure 6B**). Three of eight active compounds were detected by this model, which was in turn used to screen a library of commercially available compounds.

In the next SBVS approach, molecular docking into the allosteric binding site of CCR7 was employed (**Figure 7A**). First, we used FRED software (McGann, 2012), which was capable of high-throughput docking of a prepared library containing more than 3.5 million commercially available compounds. Then, only 100,000 highest-scoring compounds were used for subsequent docking with Glide software (Friesner et al., 2004), which was expected to be more

successful in enriching a virtual hit list but is also more computationally intensive (Kellenberger et al., 2004; McGaughey et al., 2007). Besides, Glide performed better in a re-docking experiment with **cmp2105**, achieving an RMSD of 0.34 Å, compared to FRED with an RMSD of 0.77 Å.

A SBVS approach was also explored using the ProBiS-Dock algorithm (Konc et al., 2022). Allosteric binding sites of other chemokine receptor crystal structures available at the time of our study were aligned and compared. Accordingly, three template ligands were selected: **cmp2105** (CCR7), **verciron** (CCR9), and **VT5** (CCR2). The template ligands were used together with the CCR7 crystal structure (PDB ID: 6QZH) (Jaeger et al., 2019) as input to the ProBiS-Dock algorithm (**Figure 7B**). When docking the library of commercially available compounds, both the docked compound and the receptor were treated as fully flexible to account for the induced fit of ligand binding. The obtained poses were scored using a combination of a site-specific and a generalized statistical scoring function. A site-specific



scoring function scores the docked compounds based on their overlap with the template ligands, while a generalized statistical scoring function scores the compounds based on their interactions with the receptor.

Finally, a coupled virtual screening approach was explored, in which a chemokine receptor targeted compound library

containing 539,814 compounds was docked using FRED software. This library covered similar compounds to the literature actives on all chemokine receptors. The allosteric binding site of a CCR7 homology model, 5 Å around the **vercirnon** ligand from the template structure of CCR9, was used for docking.

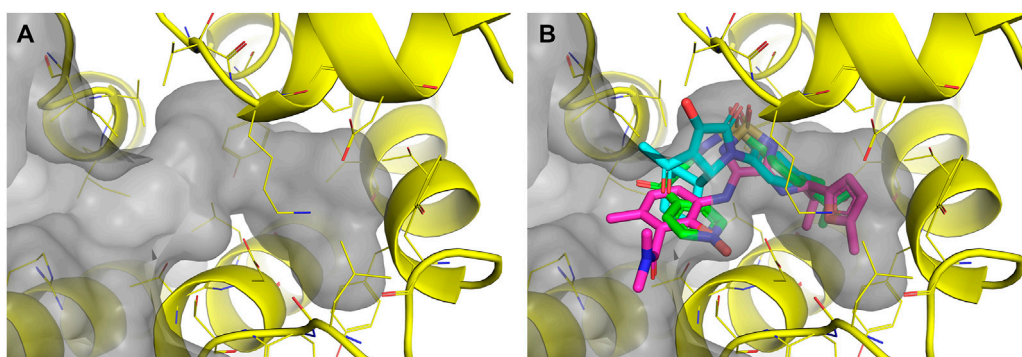


FIGURE 7 | (A) The prepared CCR7 receptor (PDB ID: 6QZH, yellow) was used for docking with FRED and subsequently with Glide. The surface of the allosteric pocket is shown in gray **(B)** The input for the ProBiS-Dock algorithm consisted of the prepared CCR7 receptor (PDB ID: 6QZH, yellow) and three template ligands, **cmp2105** (magenta), **vercirnon** (green), **VT5** (cyan).

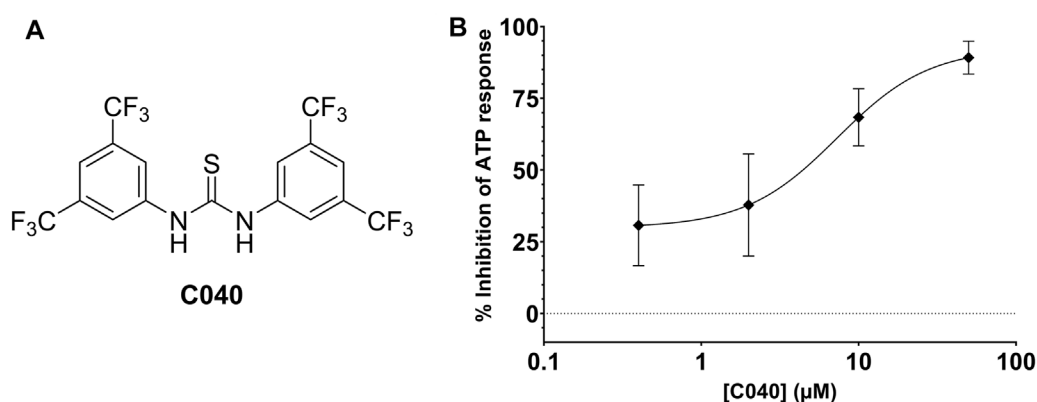


FIGURE 8 | (A) Chemical structure of **C040** **(B)** Inhibition of the ATP-induced Ca^{2+} release by **C040**.

In total, 287 virtual screening hits were selected from the various approaches and experimentally evaluated as potential CCR7 antagonists in the calcium mobilization assay (Supplementary Excel File). One compound (**C040**, **Figure 8A**) showed dose-dependent inhibition of the CCR7-mediated calcium response, affording an EC_{50} value of $13.14 \mu\text{M}$ (**Figure 3A**). Given the conserved nature of the intracellular binding pocket targeted by **cmp2105** and **navarixin** (Jaeger et al., 2019), the inhibiting effect of **C040**, alongside the reference compounds **cmp2105** and **navarixin**, on the intracellular calcium mobilization mediated by several other chemokine receptors was evaluated (**Figure 3B–D**). **Cmp2105** and **navarixin** inhibited the CCR7 and CXCR2 mediated calcium mobilization, in line with literature data (Gonsiorek et al., 2007; Jaeger et al., 2019), but had no (or only very limited) effect on CXCR4 and CCR5 mediated responses. In contrast, **C040** completely lacked receptor specificity as it inhibited the calcium mobilization downstream all tested chemokine receptors, with similar potencies (IC_{50} values in the $3\text{--}18 \mu\text{M}$ range). These

data suggested that the chemokine receptor antagonistic activity of **C040** may, at least partially, be due to interference with the fluorescent assay readout. To further explore this hypothesis, the ability of **C040** to inhibit calcium responses not mediated by human chemokine receptors was investigated. It is known that stimulation of CHO-K1 cells with adenosine triphosphate (ATP) leads to a rapid release of Ca^{2+} from intracellular stores (Iredale and Hill, 1993). Using CHO-K1 cells, exactly the same experimental set-up as for the chemokine receptor expressing U87 cells was applied, essentially including the same fluorescent calcium dye (Fluo-2) for cell loading. Also in this experimental setup **C040** was able to dose-dependently inhibit the measured calcium response induced by ATP ($10 \mu\text{M}$ final concentration) (**Figure 8B**), confirming its interference with this particular fluorescent readout. Furthermore, when **C040** was assessed in the CCR7 competition binding assay described above, it was inactive at the highest concentration tested ($25 \mu\text{M}$). Altogether, these data indicate that **C040**, despite showing activity in the CCR7 calcium assay, should not be selected as a hit compound, for a medicinal chemistry-based optimization campaign.

4 CONCLUSION

The CCR7 antagonistic activity of previously reported ligands (**cmp2105** and **navarixin**) was confirmed in two independent assays, namely a kinetic, fluorescence-based calcium mobilization CCR7 assay and a CCR7 competition binding assay. Starting from this, an *in silico* virtual screening campaign for the identification of novel CCR7 antagonists was carried out using several strategies. A library of commercially available compounds and an academic library FKKTlib (available at: <https://knjiznica-spojini.fkktlib.uni-lj.si/fkktlib/>) were used to prepare the input libraries. LBVS, SBVS, and coupled virtual screenings were followed by experimental validation. A selection of 287 *in silico* hits was experimentally investigated for CCR7 antagonism. Initial data revealed that one analogue (**C040**) showed promising CCR7 antagonistic activity in the calcium mobilization assay. Unfortunately, **C040** was equally active against other chemokine receptors tested and was completely devoid of activity in a CCR7 binding assay. Since **C040** also behaves as an antagonist of a purinergic receptor, it strongly suggests that **C040** interferes with the assay read-out, rather than being a bona fide chemokine receptor antagonist. This study highlights the importance of experimental validation of virtual hits, using an array of orthogonal assays to confirm activity before nominating any hits. Since none of the compounds disclosed in this manuscript showed any CCR7 antagonistic activity, we report them as a large set of inactive compounds that can be used by the medicinal chemistry community as a set of experimentally validated decoys. We believe this will facilitate the identification and computational design of new CCR7 ligands in the future.

REFERENCES

- Barmore, A. J., Castex, S. M., Gouletas, B. A., Griffith, A. J., Metz, S. W., Muelder, N. G., et al. (2016). Transferring the C-Terminus of the Chemokine CCL21 to CCL19 Confers Enhanced Heparin Binding. *Biochem. Biophys. Res. Commun.* 477, 602–606. doi:10.1016/j.bbrc.2016.06.098
- Berthold, M. R., Cebon, N., Dill, F., Gabriel, T. R., Kötter, T., Meinel, T., et al. (2007). KNIME: The Konstanz Information Miner. In *Studies In Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.
- Brenk, R., Schipani, A., James, D., Krasowski, A., Gilbert, I. H., Frearson, J., et al. (2008). Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* 3, 435–444. doi:10.1002/cmdc.200700139
- Buonamici, S., Trimarchi, T., Ruocco, M. G., Reavie, L., Cathelin, S., Mar, B. G., et al. (2009). CCR7 Signalling as an Essential Regulator of CNS Infiltration in T-Cell Leukaemia. *Nature* 459, 1000–1004. doi:10.1038/nature08020
- Claes, S., D'huyts, T., Van Hout, A., Schols, D., and Van Loy, T. (2018). A Kinetic Fluorescence-Based Ca²⁺ Mobilization Assay to Identify G Protein-Coupled Receptor Agonists, Antagonists, and Allosteric Modulators. *JoVE* (132), e56780. doi:10.3791/56780
- Corbisier, J., Galès, C., Huszagh, A., Parmentier, M., and Springael, J.-Y. (2015). Biased Signaling at Chemokine Receptors. *J. Biol. Chem.* 290, 9542–9554. doi:10.1074/jbc.M114.596098
- Dahlin, J. L., Nissink, J. W., Strasser, J. M., Francis, S., Higgins, L., Zhou, H., et al. (2015). PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *J. Med. Chem.* 58, 2091–2113. doi:10.1021/jm5019093
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

MP, MJ, and JK. performed computational experiments. SDJ and TVL performed CCR7 assays and analyzed data. AM and LC performed the synthesis and compound analytics. UG, ČP, MJ and LC established and maintained the academic compound library FKKTlib. MP and SDJ wrote the manuscript with contributions from all co-authors. SG and SDJ supervised the project. SG and DS secured funding. All authors have read and approved the final article.

FUNDING

Slovenian Research Agency, research core funding No. P1-0208 and CELSA project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.855653/full#supplementary-material>

- Method and Assessment of Docking Accuracy. *J. Med. Chem.* 47, 1739–1749. doi:10.1021/jm0306430
- Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra Precision glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* 49, 6177–6196. doi:10.1021/jm051256o
- Gonsiorek, W., Fan, X., Hesk, D., Fossetta, J., Qiu, H., Jakway, J., et al. (2007). Pharmacological Characterization of Sch527123, a Potent Allosteric CXCR1/CXCR2 Antagonist. *J. Pharmacol. Exp. Ther.* 322, 477–485. doi:10.1124/jpet.106.118927
- Griffith, J. W., Sokol, C. L., and Luster, A. D. (2014). Chemokines and Chemokine Receptors: Positioning Cells for Host Defense and Immunity. *Annu. Rev. Immunol.* 32, 659–702. doi:10.1146/annurev-immunol-032713-120145
- Haider, N. (2010). Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: an Open-Source Approach. *Molecules* 15, 5079–5092. doi:10.3390/molecules15085079
- Hawkins, P. C., Skillman, A. G., and Nicholls, A. (2007). Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* 50, 74–82. doi:10.1021/jm0603365
- Hjort, G. M., Larsen, O., Steen, A., Daugvilaite, V., Berg, C., Fares, S., et al. (2016). Differential CCR7 Targeting in Dendritic Cells by Three Naturally Occurring CC-Chemokines. *Front. Immunol.* 7, 568. doi:10.3389/fimmu.2016.00568
- Hull-Ryde, E. A., Porter, M. A., Fowler, K. A., Kireev, D., Li, K., Simpson, C. D., et al. (2018). Identification of Cosalane as an Inhibitor of Human and Murine CC-Chemokine Receptor 7 Signaling via a High-Throughput Screen. *SLAS Discov.* 23, 1083–1091. doi:10.1177/2472555218780917
- Iredale, P. A., and Hill, S. J. (1993). Increases in Intracellular Calcium via Activation of an Endogenous P2-Purinoceptor in Cultured CHO-K1 Cells. *Br. J. Pharmacol.* 110, 1305–1310. Available at: <https://www.ncbi.nlm.nih.gov/>

- pmc/articles/PMC2175888/(Accessed February 22, 2022). doi:10.1111/j.1476-5381.1993.tb13960.x
- Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., et al. (2004). A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins* 55, 351–367. doi:10.1002/prot.10613
- Jaeger, K., Bruenle, S., Weinert, T., Guba, W., Muehle, J., Miyazaki, T., et al. (2019). Structural Basis for Allosteric Ligand Recognition in the Human CC Chemokine Receptor 7. *Cell* 178, 1222–e10. doi:10.1016/j.cell.2019.07.028
- Jørgensen, A. S., Rosenkilde, M. M., and Hjortø, G. M. (2018). Biased Signaling of G Protein-Coupled Receptors - from a Chemokine Receptor CCR7 Perspective. *Gen. Comp. Endocrinol.* 258, 4–14. doi:10.1016/j.ygcen.2017.07.004
- Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. (2004). Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* 57, 225–242. doi:10.1002/prot.20149
- Kirschmair, J., Distinto, S., Markt, P., Schuster, D., Spitzer, G. M., Liedl, K. R., et al. (2009). How to Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* 49, 678–692. doi:10.1021/ci8004226
- Kohout, T. A., Nicholas, S. L., Perry, S. J., Reinhart, G., Junger, S., and Struthers, R. S. (2004). Differential Desensitization, Receptor Phosphorylation, Beta-Arrestin Recruitment, and ERK1/2 Activation by the Two Endogenous Ligands for the CC Chemokine Receptor 7. *J. Biol. Chem.* 279, 23214–23222. doi:10.1074/jbc.M402125200
- Konc, J., and Janežič, D. (2012). ProBiS-2012: Web Server and Web Services for Detection of Structurally Similar Binding Sites in Proteins. *Nucleic Acids Res.* 40, W214–W221. doi:10.1093/nar/gks435
- Konc, J., and Janežič, D. (2007). An Improved branch and Bound Algorithm for the Maximum Clique Problem. *MATCH Commun. Math. Comp. Chem.* 58, 569–590.
- Konc, J., Lešnik, S., Škrlj, B., Sova, M., Proj, M., and Knez, D. (2022). ProBiS-Dock: A Hybrid Multi-Template Homology Flexible Docking Algorithm Enabled by Protein Binding Site Comparison. *J. Chem. Inf. Model.* In press
- Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., et al. (2009). Improving Physical Realism, Stereochemistry, and Side-Chain Accuracy in Homology Modeling: Four Approaches that Performed Well in CASP8. *Proteins* 77 Suppl 9, 114–122. doi:10.1002/prot.22570
- Krieger, E., Koraimann, G., and Vriend, G. (2002). Increasing the Precision of Comparative Models with YASARA NOVA--a Self-Parameterizing Force Field. *Proteins* 47, 393–402. doi:10.1002/prot.10104
- Krieger, E., and Vriend, G. (2015). New Ways to Boost Molecular Dynamics Simulations. *J. Comput. Chem.* 36, 996–1007. doi:10.1002/jcc.23899
- Lešnik, S., Štular, T., Brus, B., Knez, D., Gobec, S., Janežič, D., et al. (2015). LiSiCA: A Software for Ligand-Based Virtual Screening and its Application for the Discovery of Butyrylcholinesterase Inhibitors. *J. Chem. Inf. Model.* 55, 1521–1528. doi:10.1021/acs.jcim.5b00136
- Liu, K., Shen, L., Wu, M., Liu, Z. J., and Hua, T. (2021). Structural Insights into the Activation of Chemokine Receptor CXCR2. *FEBS J.* 289, 386–393. doi:10.1111/febs.15865
- Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* 57, 3186–3204. doi:10.1021/jm401411z
- McGann, M. (2012). FRED and HYBRID Docking Performance on Standardized Datasets. *J. Comput. Aided Mol. Des.* 26, 897–906. doi:10.1007/s10822-012-9584-8
- McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., et al. (2007). Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* 47, 1504–1519. doi:10.1021/ci700052x
- Moschovakis, G. L., and Förster, R. (2012). Multifaceted Activities of CCR7 Regulate T-Cell Homeostasis in Health and Disease. *Eur. J. Immunol.* 42, 1949–1955. doi:10.1002/eji.201242614
- Müller, A., MacCallum, R. M., and Sternberg, M. J. (1999). Benchmarking PSI-BLAST in Genome Annotation. *J. Mol. Biol.* 293, 1257–1271. doi:10.1006/jmbi.1999.3233
- Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* 55, 6582–6594. doi:10.1021/jm300687e
- Nicholls, D. J., Tomkinson, N. P., Wiley, K. E., Brammall, A., Bowers, L., Grahames, C., et al. (2008). Identification of a Putative Intracellular Allosteric Antagonist Binding-Site in the CXC Chemokine Receptors 1 and 2. *Mol. Pharmacol.* 74, 1193–1202. doi:10.1124/mol.107.044610
- O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An Open Chemical Toolbox. *J. Cheminform* 3, 33. doi:10.1186/1758-2946-3-33
- Oswald, C., Rappas, M., Kean, J., Doré, A. S., Errey, J. C., Bennett, K., et al. (2016). Intracellular Allosteric Antagonism of the CCR9 Receptor. *Nature* 540, 462–465. doi:10.1038/nature20606
- Poli, G., Seidel, T., and Langer, T. (2018). Conformational Sampling of Small Molecules with iCon: Performance Assessment in Comparison with OMEGA. *Front. Chem.* 6, 229. doi:10.3389/fchem.2018.00229
- RDKit: Open-source cheminformatics (2021). Available at: <https://www.rdkit.org/> (Accessed January 8, 2021).
- Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided Mol. Des.* 27, 221–234. doi:10.1007/s10822-013-9644-8
- Schoofs, G., Van Hout, A., D'huys, T., Schols, D., and Van Loy, T. (2018). A Flow Cytometry-Based Assay to Identify Compounds that Disrupt Binding of Fluorescently-Labeled CXC Chemokine Ligand 12 to CXC Chemokine Receptor 4. *JoVE* (133), e57271. doi:10.3791/57271
- Sterling, T., and Irwin, J. J. (2015). ZINC 15--Ligand Discovery for Everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi:10.1021/acs.jcim.5b00559
- Sullivan, S. K., McGrath, D. A., Grigoriadis, D., and Bacon, K. B. (1999). Pharmacological and Signaling Analysis of Human Chemokine Receptor CCR-7 Stably Expressed in HEK-293 Cells: High-Affinity Binding of Recombinant Ligands MIP-3beta and SLC Stimulates Multiple Signaling Cascades. *Biochem. Biophys. Res. Commun.* 263, 685–690. doi:10.1006/bbrc.1999.1442
- Wolber, G., and Langer, T. (2005). LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* 45, 160–169. doi:10.1021/ci049885e
- Zheng, Y., Qin, L., Zacarias, N. V., de Vries, H., Han, G. W., Gustavsson, M., et al. (2016). Structure of CC Chemokine Receptor 2 with Orthosteric and Allosteric Antagonists. *Nature* 540, 458–461. doi:10.1038/nature20605
- Zlotnik, A., Burkhardt, A. M., and Homey, B. (2011). Homeostatic Chemokine Receptors and Organ-specific Metastasis. *Nat. Rev. Immunol.* 11, 597–606. doi:10.1038/nri3049

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Proj, De Jonghe, Van Loy, Jukić, Meden, Ciber, Podlipnik, Grošelj, Konc, Schols and Gobec. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A New Strategy for Multitarget Drug Discovery/Repositioning Through the Identification of Similar 3D Amino Acid Patterns Among Proteins Structures: The Case of Tafluprost and its Effects on Cardiac Ion Channels

OPEN ACCESS

Edited by:

Leonardo L. G. Ferreira,
University of São Paulo, Brazil

Reviewed by:

Marcus Scotti,
Federal University of Paraíba, Brazil
Simon Wang,
Howard University, United States

*Correspondence:

Wendy González
wgonzalez@utalca.cl
Miguel Reyes-Parada
miguel.reyes@usach.cl
Gabriel Núñez-Vivanco
gabriel.nunez@uaysen.cl

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 16 January 2022

Accepted: 21 February 2022

Published: 18 March 2022

Citation:

Valdés-Jiménez A, Jiménez-González D, Kiper AK, Rinné S, Decher N, González W, Reyes-Parada M and Núñez-Vivanco G (2022) A New Strategy for Multitarget Drug Discovery/Repositioning Through the Identification of Similar 3D Amino Acid Patterns Among Proteins Structures: The Case of Tafluprost and its Effects on Cardiac Ion Channels. *Front. Pharmacol.* 13:855792. doi: 10.3389/fphar.2022.855792

Alejandro Valdés-Jiménez^{1,2}, Daniel Jiménez-González^{2,3}, Aytug K. Kiper⁴, Susanne Rinné⁴, Niels Decher⁴, Wendy González^{1,5*}, Miguel Reyes-Parada^{6,7*} and Gabriel Núñez-Vivanco^{8*}

¹Center for Bioinformatics, Simulations and Modelling, Faculty of Engineering, University of Talca, Talca, Chile, ²Computer Architecture Department, Universitat Politècnica de Catalunya, Barcelona, Spain, ³Barcelona Supercomputing Center, Barcelona, Spain, ⁴Institute for Physiology and Pathophysiology, Philipps-University Marburg, Marburg, Germany, ⁵Millennium Nucleus of Ion Channels-Associated Diseases (MiNICAD), Universidad de Talca, Talca, Chile, ⁶Centro de Investigación Biomédica y Aplicada (CIBAP), Escuela de Medicina, Facultad de Ciencias Médicas, Universidad de Santiago de Chile, Santiago, Chile, ⁷Facultad de Ciencias de la Salud, Universidad Autónoma de Chile, Talca, Chile, ⁸Departamento de Ciencias Naturales y Tecnología, Universidad de Aysén, Coyhaique, Chile

The identification of similar three-dimensional (3D) amino acid patterns among different proteins might be helpful to explain the polypharmacological profile of many currently used drugs. Also, it would be a reasonable first step for the design of novel multitarget compounds. Most of the current computational tools employed for this aim are limited to the comparisons among known binding sites, and do not consider several additional important 3D patterns such as allosteric sites or other conserved motifs. In the present work, we introduce Geomfinder2.0, which is a new and improved version of our previously described algorithm for the deep exploration and discovery of similar and druggable 3D patterns. As compared with the original version, substantial improvements that have been incorporated to our software allow: (i) to compare quaternary structures, (ii) to deal with a list of pairs of structures, (iii) to know how druggable is the zone where similar 3D patterns are detected and (iv) to significantly reduce the execution time. Thus, the new algorithm achieves up to 353x speedup as compared to the previous sequential version, allowing the exploration of a significant number of quaternary structures in a reasonable time. In order to illustrate the potential of the updated Geomfinder version, we show a case of use in which similar 3D patterns were detected in the cardiac ions channels NaV1.5 and TASK-1. These channels are quite different in terms of structure, sequence and function and both have been regarded as important targets for drugs aimed at treating atrial fibrillation. Finally, we describe the *in vitro* effects of tafluprost (a drug currently used to treat glaucoma, which was identified as a novel putative ligand of NaV1.5 and TASK-1) upon both ion channels' activity and discuss its possible repositioning as a novel antiarrhythmic drug.

Keywords: polypharmacology, binding site similarity, cardiac ion channels, tafluprost, binding site comparisons

INTRODUCTION

Although most novel drugs are still developed using the “magic bullet” paradigm, which involves highly selective profiles, chemical compounds are naturally promiscuous in practice. Indeed, most therapeutically beneficial agents interact with more than one molecular target (Feldmann and Bajorath, 2020). Interestingly, this promiscuity is now, in some cases, considered an advantageous feature and is proactively pursued. Thus, many drug development initiatives are focused on the design of multitarget compounds with a polypharmacological profile that may improve drug-based treatments’ efficacy and/or safety (Ramsay et al., 2018; Proschak et al., 2019). Unfortunately, the design of drugs with multiple activities on a selected handful of different protein structures remains a significant experimental and computational challenge (Konc, 2019). Recent reports have proposed several strategies to identify multitarget drugs, such as clinical observations and target combinations based on phenotypic screening. Bioinformatics is also a useful tool to address these challenges through molecular modelling techniques for detecting similar targets, machine learning methods to find disease-related targets, target-fishing using molecular docking, ligand-based pharmacophore searching, virtual screening simulations, and the search of binding sites similarities (Ma et al., 2010; Ren et al., 2021; Stepnicki et al., 2021). Finding compounds with multitarget action on related proteins which share a similar function, folding or binding sites, is currently an accessible task. Unfortunately, complex diseases often comprise a wide range of evolutionary distant and structurally different proteins where current methods are not entirely precise. For example, neuropsychiatric, cardiac or autoimmune disorders (among others), including atrial fibrillation or major depression, are complex diseases that often encompass dysfunctions in a wide range of types of proteins such as ion channels, enzymes, transporters, globular proteins, etc. (Bolognesi, 2019; Konc, 2019). Thus, the identification of conserved/similar sites (or more broadly three-dimensional (3D) patterns, defined as a local structural arrangement of amino acids) among a set of proteins (related or not between them), can be useful for the rational design/repositioning of polypharmacological drugs (Konc, 2019; Adasme et al., 2020; Li et al., 2021). In this context, tools such as G-LoSA (Lee and Im, 2016), Geomfinder (Núñez-Vivanco et al., 2016), 3D-PP (Valdés-Jiménez et al., 2019) and others (Ehrt et al., 2016; Ehrt et al., 2018), which work regardless of information about known ligands, binding sites, sequence similarity, or structural folding of the proteins, improve the chances of finding similar 3D patterns among very different or unrelated targets. Although some of these similarities may appear by chance, others might represent distant evolutionary relationships and correspond, for instance, to secondary binding sites. These sites have recently gained attention for the rational design of polypharmacological allosteric modulators (Abdel-Magid, 2015; Meysman et al., 2015; Reyes-Parada and Iturriaga-Vasquez, 2016; Wakefield et al., 2019). Indeed, by using our in-house algorithm Geomfinder (Núñez-Vivanco et al., 2016), we have reported the finding of some similar 3D

patterns among very different protein structures, which cannot be observed through other structural tools or with sequence-based methods. Noteworthy, in the original version of Geomfinder, the residues of each detected 3D pattern could only be part of one specific protein chain. However, it is well known that several important 3D patterns (e.g. binding sites, catalytic sites) are located either at the interface between different subunits of a single target (Lee et al., 2017) or at the oligomer interface in multimeric proteins (Baek et al., 2017).

In the present work, we introduce Geomfinder2.0, which is a new and improved version of our previous tool for the deep exploration and discovery of similar and druggable 3D patterns. Thus, the possibility of exploring 3D patterns formed by different chains of a quaternary protein structure is one of the major novel features of this version. Also, the updated algorithm includes a function that predicts the zone where similar 3D patterns may be druggable (Le Guilloux et al., 2009). Thus, a 3D pattern with a high level of druggability found in several protein structures might be used as the input for the design of multitarget compounds using approaches such as Pocket-Based Drug Design (Zheng et al., 2013). From a computational perspective, the currently available version of Geomfinder has been fully migrated from Python 2.7 to C++ language and parallelized using the shared memory programming model OpenMP (Dagum and Menon, 1998). These improvements allowed a speedup of up to 353x. Furthermore, as a functional enhancement, Geomfinder can now compute several pairs of comparisons simultaneously.

In addition, in order to illustrate the potential of the updated Geomfinder version, we show a case of use in which similar 3D patterns were detected in two cardiac ion channels, specifically NaV1.5 and TASK-1, which are selective for sodium or potassium, respectively. These channels have been regarded as important targets for drugs aimed at treating atrial fibrillation (Sossalla et al., 2010; Wiedmann et al., 2021). Finally, we describe the *in vitro* effects of tafluprost (a putative ligand identified after a receptor-based drug search approach) upon both ion channels’ activity and discuss its possible repositioning as a novel antiarrhythmic drug.

MATERIALS AND METHODS

Computational Methods

Software Improvements

The new version of Geomfinder includes all geometrical characteristics and processes described previously (Núñez-Vivanco et al., 2016) and incorporates new features and substantial improvements. In the **Supplementary Figure S1**, the implemented architecture and essential components and services of Geomfinder are shown. Remarkably, the input can be a list of pairs of structures (always as PDB files) in this new version. This new feature makes it possible to find similar 3D patterns simultaneously in several pairs of protein structures submitted by the user (**Figure 1**). Another essential feature of the new version is the possibility of searching for 3D patterns in the interface between two or more protein chains/subunits. To

Pair of proteins to compare.

☒ Protein A code*

6uz3

☐ or upload pdb file A ([Download sample PDB file A](#))

Choose File No file chosen

☒ Protein B code*

6rv3

☐ or upload pdb file B ([Download sample PDB file B](#))

Choose File No file chosen

*Four characters code (Example: 1m9o). It will be downloaded from <https://www.rcsb.org/>

Next

List of pairs of proteins to compare (Max 10 lines).

Upload a text file containing list of pairs of proteins. Example:

 1tf6:*-1exp:A (all chains in 1tf6 against chain A in 1exp)

 1tf6:A,B,D-1exp:A (chains A,B,D in 1tf6 against chain A in 1exp)

 HOM1:A-1exp:A (chain A in HOM1 against chain A in 1exp)

 HOM2:A-1aay:A (chain A in HOM2 against chain A in 1aay)

 ...

 Download text file: [\[Sample basic\]](#) [\[Sample selecting the chains\]](#) [\[Sample with models by homology\]](#)

Choose File No file chosen

Structures will be downloaded from <https://www.rcsb.org/>

Homology modeled structures must be uploaded in the next page (lines starting with HOM).

Next

FIGURE 1 | The user may request the measures between two protein structures or within a list of pairs of proteins. Also, the protein structures can be both experimentally solved (downloaded from Protein Data Bank) or homology models (uploaded by the user).

Chains. If more of one chain is selected, then a new PDB file will be created with these chains concatenated (by default in chain A). If only one chain is selected (and there is more than one), a new PDB file will be created that will include only that chain.

Protein 6uz3 ⓘ

☒ A ☐ B ☐ C ☐ D

Details:

Type: protein [A] Sodium channel protein type 5 subunit alpha, Green fluorescent protein [Aequorea victoria]

Protein 6rv3 ⓘ

☒ A ☒ B ☐ C ☐ D

Details:

Type: protein [A B C D] Potassium channel subfamily K member 3 [Homo sapiens]

FIGURE 2 | The user can concatenate different protein chains. The information concerning chains, names, and the source organism of each protein is obtained from the Protein Data Bank or from the same PDB file (in the case of homology models).

Parameters

☒ Apply fpocket

Uses fpocket to find pockets.

Minimum druggability percentage (%)

50

FIGURE 3 | The user can select the application of Fpocket on the exploration of pockets. In addition, if this option is activated, a minimum druggability percentage must be included.

Frontiers in Pharmacology | www.frontiersin.org

238

March 2022 | Volume 13 | Article 855792

this end, in the submission process, the user can select one particular chain of each or concatenate either all or some of the structure's chains (**Figure 2**). Also, it is now possible to search 3D patterns only in those zones of the proteins where the cavities detected achieve the user-defined druggability score threshold, which is interpreted as the probability of the cavity to bind a drug and alter its normal activity (Pérot et al., 2010; Schmidtke and Barril, 2010). This preprocessing step is an optional parameter (**Figure 3**) calculated by the Fpocket algorithm (Le Guilloux et al., 2009), considering features such as the size, the hydrophobicity, and the normalized polarity of the residues lining the cavity. Technically, several sections of the algorithm were also optimized. The original version of Geomfinder (PythonThreading) was developed in Python 2.7 using threads as a parallel strategy. Searching for a better performance, three new versions were built: PythonMultiprocessing using Python 2.7 with multiprocessing, C++/Pthreads using C++ and POSIX threads, and C++/OpenMP using C++ with OpenMP annotations. These versions were analyzed (using a 32-CPU machine with hyperthreading activated) against a sequential version of Geomfinder implemented with benchmarking purposes.

Detection of Similar 3D Amino Acid Patterns Using Geomfinder

Searching for similar 3D amino acid patterns begins with the creation of a virtual grid of coordinates on each protein structure. Then, using all geometrical centers of all side chains of the residues, Geomfinder performs a residues grouping step depending on the distance between each virtual coordinate. With this, hundreds of 3D amino acid patterns are defined in each protein structure. After that, for each 3D pattern, four descriptors are measured: a) The list of distances between the geometric centers of the side chains of all the residues forming the 3D pattern; b) the sum of the short and medium-range of non-bonded energy of each residue forming the 3D pattern; c) the list of the residues forming a 3D pattern; and d) the list of the distances constituting the shortest pathway necessary to go over all the residues lining the 3D pattern. Finally, all pairs of 3D patterns identified in the two tested proteins are compared using an all-versus-all approach. Thus, at the end of the analysis, each pair of the 3D pattern has a final similarity score named GScore. The GScore is defined as a combination of the similarities (S) of the four descriptors:

$$GScore = S_{Dist} * D_p + S_{NbE} * C_p + S_{Tsp} * T_p + S_{Sc} * S_p$$

which is calculated as the relative changes on each pair of the 3D pattern as follows:

$$S_{Dist} = \frac{|Dist_A \cap Dist_B|}{\max(|Dist_A|, |Dist_B|)}$$

$$S_{NbE} = \frac{\min(|NbE_A|, |NbE_B|)}{\max(|NbE_A|, |NbE_B|)}$$

$$S_{Tsp} = \frac{|Tsp_A \cap Tsp_B|}{\max(|Tsp_A|, |Tsp_B|)} \quad S_{Sc} = \frac{|Sc_A \cap Sc_B|}{\max(|Sc_A|, |Sc_B|)}$$

S_{Dist}, S_{NbE}, S_{Tsp}, and S_{Sc}, are the partial scores of similarity of the distances, the non-bonded energies, the perimeter, and the sequence components, between any two 3D patterns.

Receptophore Determination

As we have previously proposed (Núñez-Vivanco et al., 2018), a "receptophore" can be defined -by analogy with the pharmacophore concept-as a 3D ensemble (present in two or more receptors), of molecular, steric, and electronic features that ensure the optimal molecular interactions with a common promiscuous ligand. Therefore, here we describe how we determine it from the local similarities identified with Geomfinder. The method consists basically of the structural alignment of the similar 3D patterns identified between the protein structures. This process was performed using the external computational methods PocketAlign and MultiBind (Shulman-Peleg et al., 2008; Yeturu and Chandra, 2011). This approach finds the best match of physicochemical properties among the residues forming each site. PocketAlign carries out multiple alignments between the similar 3D patterns detected to recognize similar amino acids matches. Then, numerous structural rearrangements of superimposed binding sites are applied to find the best structural fit. Briefly, this method consists of two main processes: a) the preprocessing of the features of each 3D pattern and hashing them into a table; and b) the recognition of the similar features in the objects of the hash table. In the preprocessing, each amino acid is denoted by pseudocenters (X, Y, and Z coordinates), which provides a unique physicochemical property to the binding site: hydrogen-bond donor, hydrogen-bond acceptor, mixed donor/acceptor, hydrophobic aliphatic or aromatic contacts. Finally, MultiBind performs a combination of multiple superimposed binding site conformations to find common patterns. Because 3D patterns that originate the receptophore are not necessarily identical, different residues might be partially aligned if the overall structural alignment score turns out to be maximized with that arrangement. After that, to generate the final receptophore, all equivalent amino acids between both 3D patterns (same physicochemical group: polar, non-polar, positively or negatively charged) appearing aligned or superimposed, are manually merged in the resultant PDB file. In contrast, all non-equivalent amino acids are preserved in the final receptophore.

Receptor/Pocket Based Virtual Screening

The previously defined receptophore was used to perform a virtual screening analysis through the computation tool e-LEA3D (Douguet, 2010). This program can create new molecules with a fragment-based approach or evaluate a user-defined data set of compounds. In our case, we used the "FDA-approved" data set to determine if some of these compounds exhibit affinity for the receptophore. Thus, each molecule's fitness was evaluated via a function that inputs the molecular structure and returns a numeric score. The evaluation can integrate a selected number of molecular properties and/or a protein-ligand docking score calculated by the program PLANTS (Korb et al., 2009). After that, a list of FDA-approved drugs, ranked by their affinity for the receptophore, was obtained and analyzed for further experiments.

Protein Structures

For the case of use, we employed the crystallographic protein structures (obtained from the Protein Data Bank) of the sodium channel NaV1.5 (PDBid: 6UZ3, resolution 3.50 Å) and the two-pore domain potassium ion channel TASK-1 (PDBid: 6RV3, resolution 2.90 Å). Only chain A was selected in the case of NaV1.5, whereas for TASK-1, chains A and B were concatenated. All details of input parameters and results can be found at <https://geomfinder2.appsbio.utalca.cl/result/11598984329232/>

Pharmacological Methods

Drugs and IC50 Values

Tafluprost was purchased from Merck. All other reagents used were of analytical grade. Tafluprost was dissolved in DMSO and added to the external solution just before the recordings. The IC50 (or EC50) were determined from Hill plots using up to five concentrations for each construct and are expressed as mean \pm SD coming from the different replicates measurements ($n = 3$ –9 replicates).

Oocyte Preparation and cRNA Injection

All procedures performed in this study involving animals were carried out in accordance with the EU Directive 2010/63/EU for animal experiments. The work with *Xenopus laevis* at the University of Marburg with all experimental protocols were approved by the Regierungspräsidium Gießen, Germany (V54-19c20 15 h 02 MR 20/28 Nr.A 4/2013).

Oocytes were obtained from anesthetized *Xenopus laevis* frogs and incubated in OR2 solution containing in mM: 82.5 NaCl, 2 KCl, 1 MgCl₂, 5 HEPES (pH 7.5) substituted with 2 mg/ml collagenase II (Sigma) to remove residual connective tissue. Then the oocytes were stored at 18°C in ND96 supplemented with 50 mg/L gentamycin, 274 mg/L sodium pyruvate, and 88 mg/L theophylline.

Oocytes were each injected with either 50 nL of cRNA of human TASK-1 (KCNC3, NM_002246) or 10 ng of cRNA of human Nav1.5 (hH1, M77235), as previously described (Ortiz-Bonnin et al., 2016; Rinné et al., 2019).

Two-Electrode Voltage Clamp Recordings

Two-electrode voltage clamp recordings were performed at room temperature (20–22°C) with a TurboTEC 10CD (npi) amplifier and a Digidata1200 Series (Axon Instruments) as analog/digital converter, as previously described (Ortiz-Bonnin et al., 2016; Rinné et al., 2019). Briefly, micropipettes were made from borosilicate glass capillaries GB150TF-8P (Science Products) and pulled with a DMZ Universal Puller (Zeitz). Recording pipettes had a resistance of 0.5–1.5 M Ω when filled with 3M KCl solution.

For both, TASK-1 or NaV1.5 channel measurements, recording solution ND96 contained in mM: 96 NaCl, 2 KCl, 1.8CaCl₂, 1 MgCl₂, 5 HEPES (pH 7.5). In the case of TASK-1 channel, block was analyzed with voltage steps from a holding potential of –80 mV. A first test pulse to 0 mV of 1 s duration was followed by a repolarizing step to –80 mV for 1 s directly followed by another 1 s test pulse to +40 mV. The sweep time interval was 10 s. For NaV1.5 channel, block was analyzed with voltage steps

from a holding potential of –120 mV. A first depolarizing pulse to –10 mV of 20 ms duration, then, after holding at –40 mV for 4s, a 20 ms step at –120 mV was carried out. The sweep time interval was 10 s.

Tafluprost at different concentrations was evaluated on channel currents. Stability in recordings was monitored prior to the addition of compounds, which were removed from the bath to show recovery.

Data were acquired with Clampex 10 (Molecular Devices) and analyzed with Clampfit 10 (MolecularDevices) and Origin 7 (OriginLab Corp.).

RESULTS

Software Optimization

All the new versions of Geomfinder, implemented with benchmarking purposes, incorporated several optimizations such as more efficient data structures and compilation flags for the machines utilized on the webserver. As denoted by the green line in **Figure 4**, the version C++/OpenMP always showed a better performance than the other implementations, achieving a maximal 353x speedup compared to the original Sequential version (**Supplementary Figure S6**). On this basis, the C++/OpenMP version was selected for the implementation of the new server of Geomfinder.

Case of Use

It has been shown that some local anesthetics are multi-channel blocking drugs, which interact with cardiac ion channels such as NaV1.5 and TASK-1 (Tikhonov and Zhorov, 2017; Rinné et al., 2019). Considering that this polypharmacological profile likely underlies their antiarrhythmic effect, we searched similar and druggable 3D patterns between these channels. Although NaV1.5 and TASK-1 have different sequences, structures, and topologies (**Figure 5**), Geomfinder required a few seconds to detect several similar 3D patterns (**Supplementary Figure S2**).

Interestingly, one of the patterns found contains two key residues of the local anesthetics binding site in NaV1.5 (Phe1762, Ile1468; (Nguyen et al., 2019)) and one key residue in TASK-1 (Phe238; (Rinné et al., 2019)) located at the fenestrations. Considering that these 3D patterns are similar (GScore = 66.5%) and are located in zones with high druggability values (Phe1762 and Ile1468 are located in pockets with 99.9% druggability in NaV1.5; and Phe238 in a pocket with 100% druggability in TASK-1; result #25 in **Supplementary Figure S3**), it is enticing to state that the use of these 3D patterns is a promising starting point to understand the polypharmacological profile of local anesthetics and to design/search new multitarget compounds aimed to these cardiac ion channels. The fact that the residues Phe1762 in NaV1.5 and Phe238 in TASK-1 were found by Geomfinder in similar tridimensional orientations into each 3D pattern (**Figure 6**), supports the idea that the aromatic ring of local anesthetics interacts with the ion channels establishing a π – π interaction (González et al., 2001).

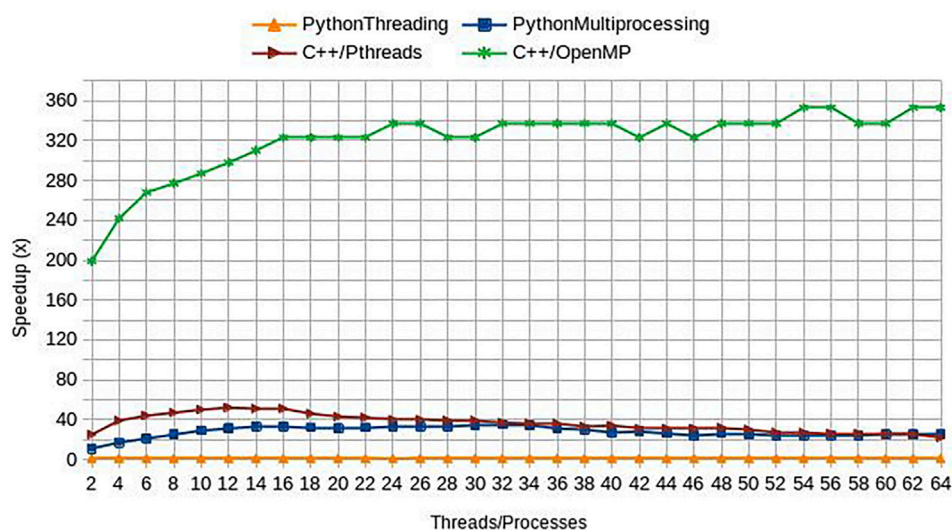


FIGURE 4 | Performance of new Geomfinder implementations. The original (PythonThreading) and three new versions, including parallel programming paradigm and code optimization, were compared against a sequential Python version (Sequential) developed with benchmarking purposes. These versions were built using Python 2.7 with multiprocessing (PythonMultiprocessing), C++ with POSIX threads (C++/Pthreads), and C++ with OpenMP annotations (C++/OpenMP). As is denoted by the green line in the graph, the version C++/OpenMP always showed better performance than the other implementations, reaching their best acceleration using 54 threads: 353x speedup compared to Sequential version. Thus, the version C++/OpenMP was selected for the implementation of the new server of Geomfinder. The experimental setup used in the comparisons is a 32 Intel Xeon CPU E5-2683 (2.10 GHz) SMP system with hyperthreading enabled (64 virtual cores/threads), 252 GB RAM, 40 MB Intel Smart Cache.

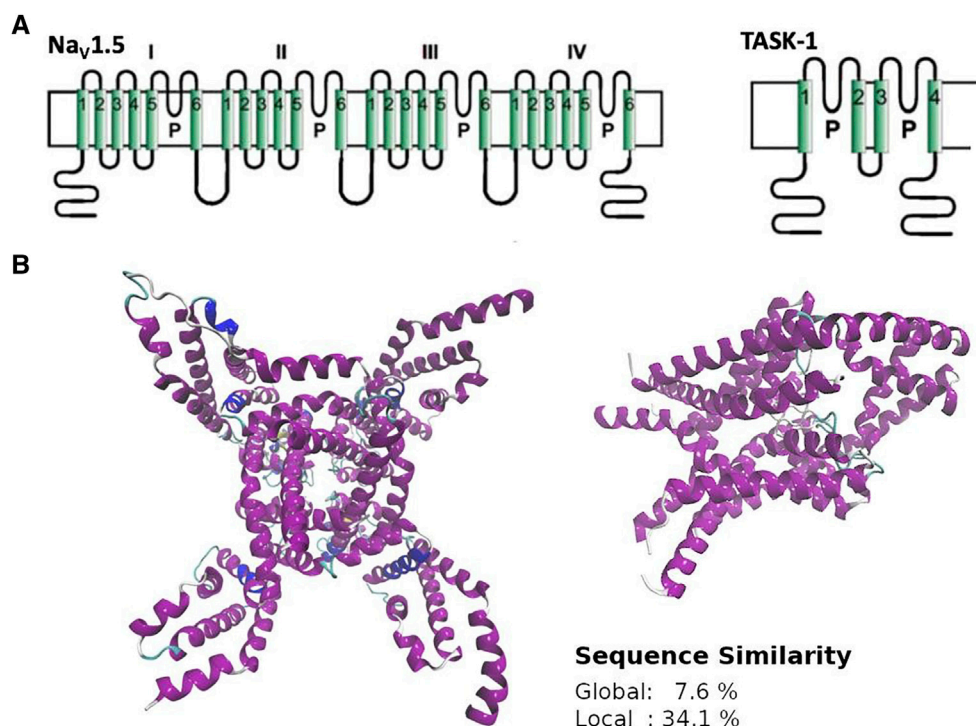
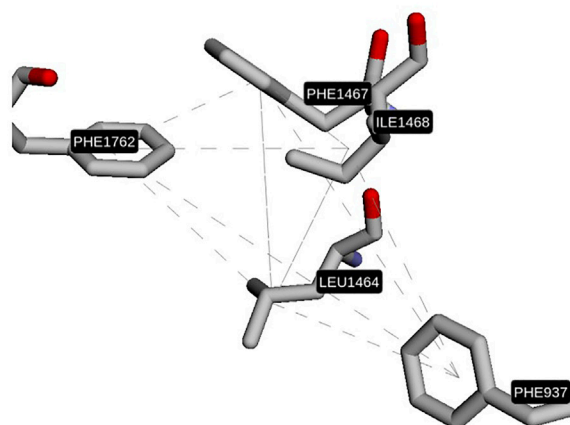


FIGURE 5 | General overview of structure (A), topology (B) and sequence (inset) similarity of NaV1.5 and TASK-1. NaV1.5 corresponds to the PDBid:6UZ3 and TASK-1 to the PDBid:6RV3. NaV1.5 is a monomer with four domains, each containing a pore sequence and TASK-1 is a dimer with four transmembrane segments and two pore sequences each monomer.

NaV1.5



TASK-1

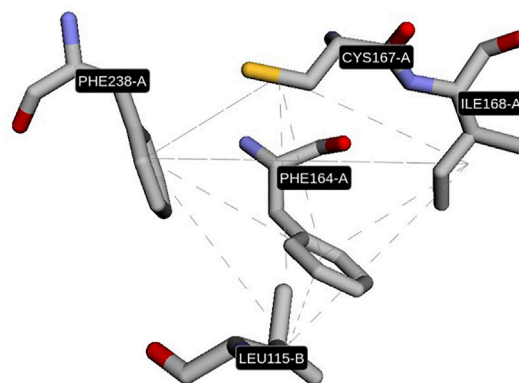


FIGURE 6 | Similar 3D pattern found in NaV1.5 and TASK-1. In both sites, the residues PHE1762 and PHE238 show similar orientation.

| Ligand 1 : BU_0(Chain:A,B) | | | Ligand 2 : BU_0(Chain:A) | | |
|--|-------|------|-----------------------------|-------|------|
| Rigid transformation of binding site 1 : 1.44238 0.187857 -2.29709 1 | | | | | |
| Site 1: 6rv3-TASK-1-BU_.pdb | | | Site 2: 6uz3-NAV1.5-BU_.pdb | | |
| Chain.ID | A. A. | Type | Chain.ID | A. A. | Type |
| A.160 | Val | DON | A.1468 | Ile | DON |
| A.160 | Val | ALI | A.1468 | Ile | ALI |
| A.161 | Leu | ALI | A.1472 | Ile | ALI |
| A.238 | Phe | PII | A.1762 | Phe | PII |
| A.242 | Val | ALI | A.1464 | Leu | ALI |
| B.115 | Leu | ALI | A.1765 | Val | ALI |

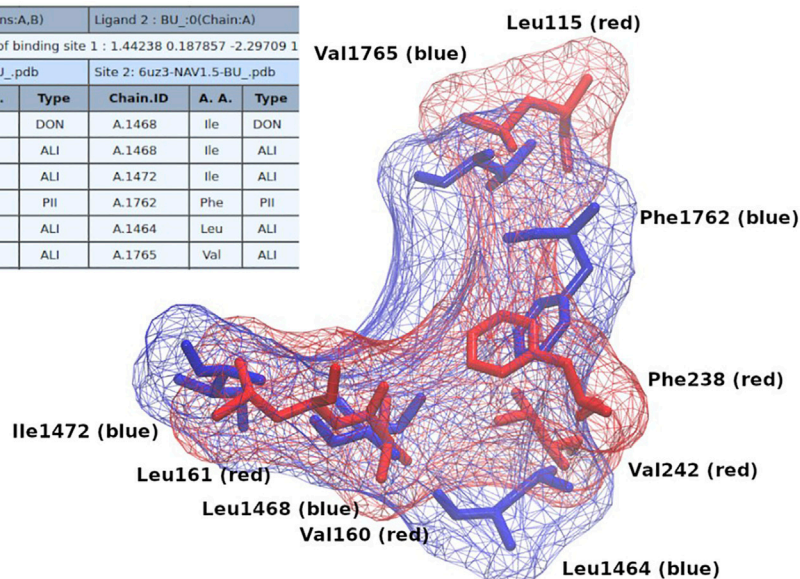


FIGURE 7 | Common binding site of bupivacaine on TASK-1 and NaV1.5. The table (from Multibind) indicates the equivalent physicochemical properties matched. The colored residues depicted in licorice show the alignment (from PocketAlign) of the bupivacaine binding sites on both proteins (Red: NaV1.5; Blue: TASK-1).

In order to show how these results could be useful for the design/search of new multitarget compounds, we initially performed docking molecular simulations of bupivacaine (a common local anesthetic) on NaV1.5 and TASK-1 using the coordinates of the similar pockets detected with Geomfinder as grid centers. As shown in **Supplementary Figures S4, S5**, the aromatic ring of the bupivacaine seems to be establishing π - π interactions with Phe238 in TASK-1 and with Phe1762 in NaV1.5. After selecting the best conformers of bupivacaine (those with the lowest estimation of free energy of binding) in both proteins, we extracted the residues located at 4 Å of the

ligands, and constructed a common binding site for TASK-1 and NaV1.5, as we have previously described (Möller-Acuña et al., 2015; Núñez-Vivanco et al., 2018). Briefly, this common binding site was constructed using external computational tools such as PocketAlign (Yeturu and Chandra, 2011) and MultiBind (Shulman-Peleg et al., 2008), which perform several structural alignments and tridimensional rearrangements, looking for the best match of the physicochemical properties of the selected binding sites. As shown in **Figure 7**, it was possible to find a tridimensional fit for adjusting the physicochemical properties and the structural alignment of the residues of both binding sites,

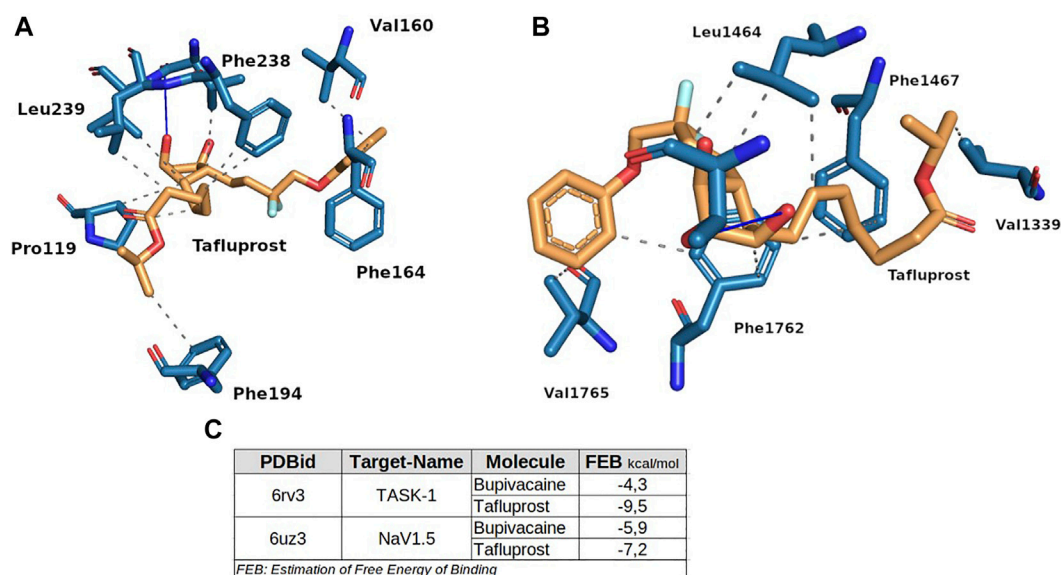


FIGURE 8 | Binding modes of tafluprost on TASK-1 (A) and NaV1.5 (B). The box at (C) shows the estimation of Free Energy of Binding of bupivacaine and tafluprost for both proteins. These experiments were performed at the same conditions.

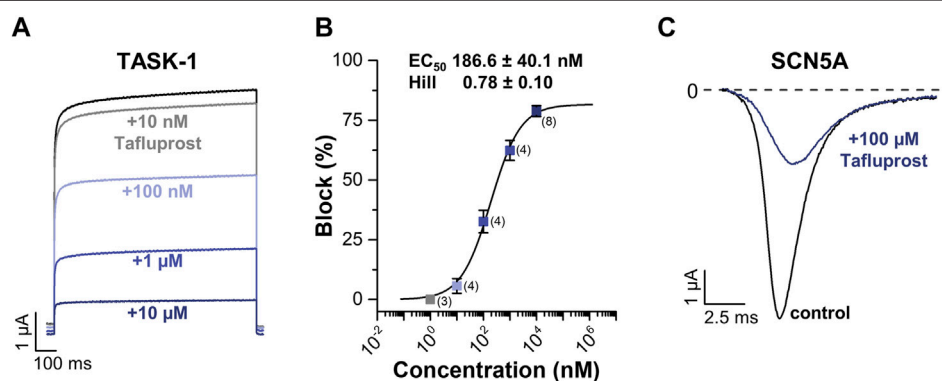


FIGURE 9 | (A) Representative current traces of TASK-1 in the absence (black) or the presence of increasing concentrations of tafluprost (colored). (B) Percentage of block of outward currents in TASK-1 channel by tafluprost. Each point is the mean \pm SD of 3-8 determinations. (C) Representative current traces of NaV1.5 (also referred as to SCN5A) before (black) and after (blue) application of 100 μ M tafluprost.

which we call the common “receptophore” (Núñez-Vivanco et al., 2018). After that, we used this “receptophore” to perform a structure-based virtual screening with the e-LEA3D tool (Douguet, 2010). This software requires as input a PDB file of one pocket structure, which can then be used to make either a *de novo* drug design or a virtual screening into a collection of known compounds, which in both cases should lead to find molecules with high potential affinity for the pocket. In our case, we merged both aligned binding sites into one unique PDB file and after run e-LEA3D in the virtual screening mode (using the “FDA approved drugs” data set available at the e-LEA3D web server), a list of molecules ranked by their theoretical affinities for the pocket submitted was obtained. The top ranked molecule was tafluprost (Drugbank_ID: DB08819), a prostaglandin $F_{2\alpha}$

(PGF $_{2\alpha}$)-type agonist currently used as a treatment for glaucoma and ocular hypertension (Papadia et al., 2011; Klimko and Sharif, 2019), which in theory should be, a new putative multitarget ligand of NaV1.5 and TASK-1.

To test this idea, we performed new docking molecular simulations with tafluprost on the structures of NaV1.5 and TASK-1, setting the same parameters that were used for the experiments with bupivacaine. As stated in Figure 8, tafluprost showed better affinities than those obtained for bupivacaine at both protein structures. The stabilization of the protein-ligand complexes seems to be determined by several hydrophobic interactions of residues which have been previously reported as key residues for the local anesthetics (e.g. Phe238, Leu1464 and Phe1762).

Experimental Evaluation of Tafluprost

As shown in **Figure 9**, tafluprost potently blocked TASK-1 outward currents in a concentration-dependent manner ($EC_{50} = 186 \pm 40$ nM; **Figures 9A,B**). It should be noted that since the highest tafluprost concentration used in these experiments was 10 μ M (**Figure 9B**), here we report EC_{50} instead of IC_{50} , as the hill fit determines the concentration where tafluprost reaches 50% of its maximal effect but not 50% inhibition. As we reached 82% maximal inhibition, EC_{50} concentration corresponds to about 40% inhibition (not 50% inhibition). On the other hand, tafluprost also blocked NaV1.5 inward currents, although with much less potency than that observed at TASK-1 (estimated IC_{50} of about 76 μ M **Figure 9B**).

DISCUSSION

The search of multitarget compounds might be a difficult task, particularly when the drugs are aimed to act at receptors with highly diverse structure and function. Based on the idea that a given compound could simultaneously interact with two (or more) relevant targets if they have similar binding sites (Jalencas and Mestres, 2013; Salentin et al., 2014; Ehrt et al., 2016; Konc, 2019; Naderi et al., 2019), one reasonable approach to find promiscuous drugs under these circumstances is to look for similar binding sites at the addressed targets. In this context, Geomfinder2.0 appears as a valuable tool since it is a fast web server for the discovery of similar and druggable 3D patterns between any pair of protein structures. This new version has significantly improved its usefulness and performance as compared with the original version, with up to 353x speedup for the input data set analyzed and the available machine, and allowing to compare a list of pairs of structures. It also identifies 3D patterns formed by different protein chains and characterizes how druggable is the zone where the 3D patterns were detected. It is important to note that, beyond these functional and performance improvements, the core method for searching and comparison of 3D patterns is the same as in the original version (Núñez-Vivanco et al., 2016). The accuracy and precision of this core method has already been compared with those of computational tools such as PocketMatch and ClickTopology (Núñez-Vivanco et al., 2016). In addition, in the present work we confirmed the reliability of our algorithm with a case of use, in which the theoretical predictions were experimentally confirmed.

Interestingly, when analyzing TASK-1 and NaV1.5, two cardiac channels with highly different structures and functions, Geomfinder2.0 was able to find (in a few seconds) several similar 3D patterns, a pair of which seemed remarkably attractive since they included some key residues involved in the binding of local anesthetics at both types of ion channels (Nguyen et al., 2019; Rinné et al., 2019). On this basis, we dissected these 3D patterns and constructed a common binding site which was used to search for possible novel multitarget ligands. As the development of novel polypharmacological agents can be a difficult, time-consuming and expensive task, drug repurposing (i.e. the establishment of new indications of existing drugs), has been proposed as an efficient alternative over the *de novo* drug

development approach. Thus, using a “FDA-approved” data set we searched for known compounds that might show an unanticipated ability to interact with TASK-1 and NaV1.5. Tafluprost, a prostaglandin analogue (PGF_{2 α} agonist) currently used for the treatment of open-angle glaucoma (Papadia et al., 2011; Klimko and Sharif, 2019), was the most promising drug arising from this analysis. Remarkably, when experimentally tested tafluprost blocked the corresponding currents at both NaV1.5 and TASK-1, although with a higher potency for the latter. The differential activity of tafluprost upon both ion channels (as well as its higher potency as compared with the local anesthetic bupivacaine (Stoetzer et al., 2016)) roughly agrees with the theoretical binding energies predicted by the docking simulations. Noteworthy, although tafluprost was clearly less active in NaV1.5 than in TASK-1, it still shows a potency in a similar range as that shown by the well-known sodium channel blocker bupivacaine (Zhang et al., 2014). Even though this is the first time that an activity of tafluprost on cardiac channels is described, it should be noted that it had been reported that the drug is able to induce a relaxation of rabbit ciliary arteries precontracted with a high-potassium solution (Dong et al., 2008). Accordingly, this effect might be related with its potent TASK-1 blocking properties.

Both NaV1.5 and TASK-1 are attractive drug targets in particular for the development of treatments of atrial fibrillation, the most common cardiac arrhythmia (Sossalla et al., 2010; Wiedmann et al., 2021). In addition, multichannel blockers such as amiodarone or dronedarone have been used in clinical settings (Kraft et al., 2021). In this context, the pharmacological activity predicted and demonstrated here for tafluprost, suggests that this compound could be a good candidate to evaluate its properties and repurpose it as a novel antiarrhythmic drug.

In summary, despite mounting evidence indicating that polypharmacological drugs might show better efficacy and less side effects than more selective compounds, early work in this area had already recognized that the rational search of multitarget drugs faces at least two major challenges, including a) the need to identify a combination of nodes in a biological network whose perturbation results in a desired therapeutic outcome, and b) to develop drugs whose polypharmacological profile allows those nodes to be perturbed specifically (Hopkins, 2008). Therefore, computational tools such as Geomfinder can be helpful to discover similar and druggable 3D patterns among proteins which have been tagged as targets in a multi-factorial disease, which appears as an important first step to either rationally design or -as in this case-purpose novel indications for compounds already in use.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The animal study was reviewed and approved by Regierungspräsidium Giesen, Germany (V54-19c20 15 h 02 MR 20/28 Nr.A 4/2013) (Work with *Xenopus laevis* at the University of Marburg).

AUTHOR CONTRIBUTIONS

AV-J, DJ-G, AK, SR, ND, WG, MR-P, and GN-V wrote the manuscript. WG, MR-P, ND, and GN-V designed the experiments and together with AV-J and DJ-G interpreted the results. AV-J, DJ-G, and GN-V designed, built and implemented computational development. SR, AK, and ND performed and interpreted electrophysiological experiments. Global conceptualization by AV-J, WG, MR-P, and GN-V.

REFERENCES

- Abdel-Magid, A. F. (2015). Allosteric Modulators: An Emerging Concept in Drug Discovery. *ACS Med. Chem. Lett.* 6, 104–107. doi:10.1021/ml5005365
- Adasme, M. F., Parisi, D., Van Belle, K., Salentin, S., Haupt, V. J., Jennings, G. S., et al. (2020). Structure-based Drug Repositioning Explains Ibrutinib as VEGFR2 Inhibitor. *PLoS One* 15, e0233089. doi:10.1371/journal.pone.0233089
- Baek, M., Park, T., Heo, L., Park, C., and Seok, C. (2017). GalaxyHomomer: a Web Server for Protein Homo-Oligomer Structure Prediction from a Monomer Sequence or Structure. *Nucleic Acids Res.* 45, W320–W324. doi:10.1093/nar/gkx246
- Bolognesi, M. L. (2019). Harnessing Polypharmacology with Medicinal Chemistry. *ACS Med. Chem. Lett.* 10, 273–275. doi:10.1021/acsmchemlett.9b00039
- Dagum, L., and Menon, R. (1998). OpenMP: an Industry Standard API for Shared-Memory Programming. *IEEE Comput. Sci. Eng.* 5, 46–55. doi:10.1109/99.660313
- Dong, Y., Watabe, H., Su, G., Ishikawa, H., Sato, N., and Yoshitomi, T. (2008). Relaxing Effect and Mechanism of Tafluprost on Isolated Rabbit Ciliary Arteries. *Exp. Eye Res.* 87, 251–256. doi:10.1016/j.exer.2008.06.005
- Douguet, D. (2010). e-LEA3D: a Computational-Aided Drug Design Web Server. *Nucleic Acids Res.* 38, W615. doi:10.1093/nar/gkq322
- Ehrt, C., Brinkjost, T., and Koch, O. (2016). Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.* 59, 4121–4151. doi:10.1021/acs.jmedchem.6b00078
- Ehrt, C., Brinkjost, T., and Koch, O. (2018). A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (ProSPECCTs). *Plos Comput. Biol.* 14, e1006483. doi:10.1371/journal.pcbi.1006483
- Feldmann, C., and Bajorath, J. (2020). Biological Activity Profiles of Multitarget Ligands from X-ray Structures. *Molecules* 25, 794. doi:10.3390/molecules25040794
- González, T., Longobardo, M., Caballero, R., Delpón, E., Sinisterra, J. V., Tamargo, J., et al. (2001). Stereoselective Effects of the Enantiomers of a New Local Anaesthetic, IQB-9302, on a Human Cardiac Potassium Channel (Kv1.5). *Br. J. Pharmacol.* 132, 385–392. doi:10.1038/sj.bjp.0703844
- Hopkins, A. L. (2008). Network Pharmacology: the Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* 4, 682–690. doi:10.1038/nchembio.118
- Jalencas, X., and Mestres, J. (2013). Identification of Similar Binding Sites to Detect Distant Polypharmacology. *Mol. Inform.* 32, 976–990. doi:10.1002/minf.201300082
- Klimko, P. G., and Sharif, N. A. (2019). Discovery, Characterization and Clinical Utility of Prostaglandin Agonists for the Treatment of Glaucoma. *Br. J. Pharmacol.* 176, 1051–1058. doi:10.1111/bph.14327

FUNDING

This research was funded by the Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) grants numbers 1191133, 1170662 and from Spanish Ministry of Economy and Competitiveness (projects SEV-2015-0493 and TIN2015-65316-P, grant BES-2016-078046), and from Generalitat de Catalunya (contracts 2017-SGR-1414 and 2017-SGR-1328). The financial support by DICYT-USACH grant 5392102RP-ACDicyt is also acknowledged. The web-server is hosted in the cluster obtained with the grant CONICYT-FONDEQUIP-EQM160063.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.855792/full#supplementary-material>

- Konc, J. (2019). Binding Site Comparisons for Target-Centered Drug Discovery. *Expert Opin. Drug Discov.* 14, 445–454. doi:10.1080/17460441.2019.1588883
- Korb, O., Stützle, T., and Exner, T. E. (2009). Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* 49, 84–96. doi:10.1021/ci800298z
- Kraft, M., Büscher, A., Wiedmann, F., L'hoste, Y., Haefeli, W. E., Frey, N., et al. (2021). Current Drug Treatment Strategies for Atrial Fibrillation and TASK-1 Inhibition as an Emerging Novel Therapy Option. *Front. Pharmacol.* 12, 638445. doi:10.3389/fphar.2021.638445
- Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* 10, 168. doi:10.1186/1471-2105-10-168
- Lee, H. S., and Im, W. (2016). G-LoSA: An Efficient Computational Tool for Local Structure-Centric Biological Studies and Drug Design. *Protein Sci.* 25, 865–876. doi:10.1002/pro.2890
- Lee, J., Konc, J., Janežič, D., and Brooks, B. R. (2017). Global Organization of a Binding Site Network Gives Insight into Evolution and Structure-Function Relationships of Proteins. *Sci. Rep.* 7, 11652. doi:10.1038/s41598-017-10412-z
- Li, S., Cai, C., Gong, J., Liu, X., and Li, H. (2021). A Fast Protein Binding Site Comparison Algorithm for Proteome-wide Protein Function Prediction and Drug Repurposing. *Proteins* 89, 1541–1556. doi:10.1002/prot.26176
- Ma, X. H., Shi, Z., Tan, C., Jiang, Y., Go, M. L., Low, B. C., et al. (2010). In-silico Approaches to Multi-Target Drug Discovery: Computer Aided Multi-Target Drug Design, Multi-Target Virtual Screening. *Pharm. Res.* 27, 739–749. doi:10.1007/s11095-010-0065-2
- Meysman, P., Zhou, C., Cule, B., Goethals, B., and Laukens, K. (2015). Mining the Entire Protein DataBank for Frequent Spatially Cohesive Amino Acid Patterns. *BioData Min.* 8, 4. doi:10.1186/s13040-015-0038-4
- Möller-Acuña, P., Contreras-Riquelme, J. S., Rojas-Fuentes, C., Nuñez-Vivanco, G., Alzate-Morales, J., Iturriaga-Vásquez, P., et al. (2015). Similarities between the Binding Sites of SB-206553 at Serotonin Type 2 and Alpha7 Acetylcholine Nicotinic Receptors: Rationale for its Polypharmacological Profile. *PLoS one* 10, e0134444. doi:10.1371/journal.pone.0134444
- Naderi, M., Lemoine, J. M., Govindaraj, R. G., Kana, O. Z., Feinstein, W. P., and Brylinski, M. (2019). Binding Site Matching in Rational Drug Design: Algorithms and Applications. *Brief. Bioinform.* 20, 2167–2184. doi:10.1093/bib/bby078
- Nguyen, P. T., DeMarco, K. R., Vorobyov, I., Clancy, C. E., and Yarov-Yarovoy, V. (2019). Structural Basis for Antiarrhythmic Drug Interactions with the Human Cardiac Sodium Channel. *Proc. Natl. Acad. Sci. U S A.* 116, 2945–2954. doi:10.1073/pnas.1817446116
- Núñez-Vivanco, G., Valdés-Jiménez, A., Besoain, F., and Reyes-Parada, M. (2016). Geomfinder: a Multi-Feature Identifier of Similar Three-Dimensional Protein Patterns: a Ligand-independent Approach. *J. Cheminform.* 8, 19. doi:10.1186/s13321-016-0131-9

- Núñez-Vivanco, G., Fierro, A., Moya, P., Iturriaga-Vásquez, P., and Reyes-Parada, M. (2018). 3D Similarities between the Binding Sites of Monoaminergic Target Proteins. *PLoS one* 13, e0200637. doi:10.1371/journal.pone.0200637
- Ortiz-Bonnin, B., Rinné, S., Moss, R., Streit, A. K., Scharf, M., Richter, K., et al. (2016). Electrophysiological Characterization of a Large Set of Novel Variants in the SCN5A-Gene: Identification of Novel LQTS3 and BrS Mutations. *Pflugers Arch.* 468, 1375–1387. doi:10.1007/s00424016-1844-310.1007/s00424-016-1844-3
- Papadia, M., Bagnis, A., Scotto, R., and Traverso, C. E. (2011). Tafluprost for Glaucoma. *Expert Opin. Pharmacother.* 12, 2393–2401. doi:10.1517/14656566.2011.606810
- Pérot, S., Sperandio, O., Miteva, M. A., Camproux, A. C., and Villoutreix, B. O. (2010). Druggable Pockets and Binding Site Centric Chemical Space: a Paradigm Shift in Drug Discovery. *Drug Discov. Today* 15, 656–667. doi:10.1016/j.drudis.2010.05.015
- Proschak, E., Stark, H., and Merk, D. (2019). Polypharmacology by Design: A Medicinal Chemist's Perspective on Multitargeting Compounds. *J. Med. Chem.* 62, 420–444. doi:10.1021/acs.jmedchem.8b00760
- Ramsay, R. R., Popovic-Nikolic, M. R., Nikolic, K., Uliassi, E., and Bolognesi, M. L. (2018). A Perspective on Multi-Target Drug Discovery and Design for Complex Diseases. *Clin. Transl. Med.* 7, 3. doi:10.1186/s40169-017-0181-2
- Ren, P.-x., Shang, W.-j., Yin, W.-c., Ge, H., Wang, L., Zhang, X.-l., et al. (2021). A Multi-Targeting Drug Design Strategy for Identifying Potent Anti-SARS-CoV-2 Inhibitors. *Acta Pharmacol. Sin.* 43, 483–493. doi:10.1038/s41401-021-00668-7
- Reyes-Parada, M., and Iturriaga-Vasquez, P. (2016). The Development of Novel Polypharmacological Agents Targeting the Multiple Binding Sites of Nicotinic Acetylcholine Receptors. *Expert Opin. Drug Discov.* 11, 969–981. doi:10.1080/17460441.2016.1227317
- Rinné, S., Kiper, A. K., Vowinkel, K. S., Ramírez, D., Schewe, M., Bedoya, M., et al. (2019). The Molecular Basis for an Allosteric Inhibition of K⁺-flux Gating in K2P Channels. *eLife* 8, e39476. doi:10.7554/eLife.39476
- Salentin, S., Haupt, V. J., Daminelli, S., and Schroeder, M. (2014). Polypharmacology Rescored: Protein-Ligand Interaction Profiles for Remote Binding Site Similarity Assessment. *Prog. Biophys. Mol. Biol.* 116, 174–186. doi:10.1016/j.pbiomolbio.2014.05.006
- Schmidtke, P., and Barril, X. (2010). Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J. Med. Chem.* 53, 5858–5867. doi:10.1021/jm100574m
- Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. J. (2008). MultiBind and MAPPIS: Webservers for Multiple Alignment of Protein 3D-Binding Sites and Their Interactions. *Nucleic Acids Res.* 36, W260. doi:10.1093/nar/gkn185
- Sossalla, S., Kallmeyer, B., Wagner, S., Mazur, M., Maurer, U., Toischer, K., et al. (2010). Altered Na⁺ Currents in Atrial Fibrillation Effects of Ranolazine on Arrhythmias and Contractility in Human Atrial Myocardium. *J. Am. Coll. Cardiol.* 55, 2330–2342. doi:10.1016/j.jacc.2009.12.055
- Stepnicki, P., Kondej, M., Koszła, O., Żuk, J., and Kaczor, A. A. (2021). Multi-targeted Drug Design Strategies for the Treatment of Schizophrenia. *Expert Opin. Drug Discov.* 16, 101–114. doi:10.1080/17460441.2020.1816962
- Stoetzer, C., Doll, T., Stueber, T., Herzog, C., Echtermeyer, F., Greulich, F., et al. (2016). Tetrodotoxin-sensitive α -subunits of Voltage-Gated Sodium Channels Are Relevant for Inhibition of Cardiac Sodium Currents by Local Anesthetics. *Naunyn Schmiedebergs Arch. Pharmacol.* 389, 625–636. doi:10.1007/s00210-016-1231-9
- Tikhonov, D. B., and Zhorov, B. S. (2017). Mechanism of Sodium Channel Block by Local Anesthetics, Antiarrhythmics, and Anticonvulsants. *J. Gen. Physiol.* 149, 465–481. doi:10.1085/jgp.201611668
- Valdés-Jiménez, A., Larriba-Pey, J. L., Núñez-Vivanco, G., and Reyes-Parada, M. (2019). 3D-PP: A Tool for Discovering Conserved Three-Dimensional Protein Patterns. *Int. J. Mol. Sci.* 20, 3174. doi:10.3390/ijms20133174
- Wakefield, A. E., Mason, J. S., Vajda, S., and Keserű, G. M. (2019). Analysis of Tractable Allosteric Sites in G Protein-Coupled Receptors. *Sci. Rep.* 9, 6180. doi:10.1038/s41598-019-42618-8
- Wiedmann, F., Beyersdorf, C., Zhou, X. B., Kraft, M., Paasche, A., Javorszky, N., et al. (2021). Treatment of Atrial Fibrillation with Doxapram: TASK-1 Potassium Channel Inhibition as a Novel Pharmacological Strategy. *Cardiovasc. Res.* cvab177. doi:10.1093/cvr/cvab177
- Yeturu, K., and Chandra, N. (2011). PocketAlign a Novel Algorithm for Aligning Binding Sites in Protein Structures. *J. Chem. Inf. Model.* 51, 1725–1736. doi:10.1021/ci200132z
- Zhang, H., Ji, H., Liu, Z., Ji, Y., You, X., Ding, G., et al. (2014). Voltage-dependent Blockade by Bupivacaine of Cardiac Sodium Channels Expressed in Xenopus Oocytes. *Neurosci. Bull.* 30, 697–710. doi:10.1007/s12264-013-1449-1
- Zheng, X., Gan, L., Wang, E., and Wang, J. (2013). Pocket-based Drug Design: Exploring Pocket Space. *AAPS J.* 15, 228–241. doi:10.1208/s12248-012-9426-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Valdés-Jiménez, Jiménez-González, Kiper, Rinné, Decher, González, Reyes-Parada and Núñez-Vivanco. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Identification of Potent and Selective JAK1 Lead Compounds Through Ligand-Based Drug Design Approaches

Sathya Babu¹, Santhosh Kumar Nagarajan¹, Sruthy Sathish¹, Vir Singh Negi², Honglae Sohn^{3*} and Thirumurthy Madhavan^{1*}

¹Computational Biology Lab, Department of Genetic Engineering, School of Bioengineering, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, India, ²Department of Clinical Immunology, Jawaharlal Institute of Post-Graduate Medical Education and Research, Pondicherry, India, ³Department of Chemistry and Department of Carbon Materials, Chosun University, Gwangju, South Korea

OPEN ACCESS

Edited by:

Leonardo L. G. Ferreira,
University of São Paulo, Brazil

Reviewed by:

Prasanth Kumar,
Gujarat University, India
Mariana Laureano De Souza,
Universidade de São Paulo São
Carlos, Brazil

*Correspondence:

Thirumurthy Madhavan
thiru.murthyunom@gmail.com
Honglae Sohn
hsohn@chosun.ac.kr

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 16 December 2021

Accepted: 07 March 2022

Published: 21 April 2022

Citation:

Babu S, Nagarajan SK, Sathish S,
Negi VS, Sohn H and Madhavan T
(2022) Identification of Potent and
Selective JAK1 Lead Compounds
Through Ligand-Based Drug
Design Approaches.
Front. Pharmacol. 13:837369.
doi: 10.3389/fphar.2022.837369

JAK1 plays a significant role in the intracellular signaling by interacting with cytokine receptors in different types of cells and is linked to the pathogenesis of various cancers and in the pathology of the immune system. In this study, ligand-based pharmacophore modeling combined with virtual screening and molecular docking methods was incorporated to identify the potent and selective lead compounds for JAK1. Initially, the ligand-based pharmacophore models were generated using a set of 52 JAK1 inhibitors named C-2 methyl/hydroxyethyl imidazopyrrolopyridines derivatives. Twenty-seven pharmacophore models with five and six pharmacophore features were generated and validated using potency and selectivity validation methods. During potency validation, the Guner-Henry score was calculated to check the accuracy of the generated models, whereas in selectivity validation, the pharmacophore models that are capable of identifying selective JAK1 inhibitors were evaluated. Based on the validation results, the best pharmacophore models ADHRRR, DDHRRR, DRRR, DPRR, DHRRR, ADRRR, DDHRR, and ADPRR were selected and taken for virtual screening against the Maybridge, Asinex, Chemdiv, Enamine, Lifechemicals, and Zinc database to identify the new molecules with novel scaffold that can bind to JAK1. A total of 4,265 hits were identified from screening and checked for acceptable drug-like properties. A total of 2,856 hits were selected after ADME predictions and taken for Glide molecular docking to assess the accurate binding modes of the lead candidates. Ninety molecules were shortlisted based on binding energy and H-bond interactions with the important residues of JAK1. The docking results were authenticated by calculating binding free energy for protein-ligand complexes using the MM-GBSA calculation and induced fit docking methods. Subsequently, the cross-docking approach was carried out to recognize the selective JAK1 lead compounds. Finally, top five lead compounds that were potent and selective against JAK1 were selected and validated using molecular dynamics simulation. Besides, the density functional theory study was also carried out for the selected leads. Through various computational studies, we observed good potency and selectivity of these lead compounds when compared with the drug ruxolitinib.

Compounds such as T5923555 and T5923531 were found to be the best and can be further validated using *in vitro* and *in vivo* methods.

Keywords: JAK1, pharmacophore modeling, virtual screening, molecular docking, molecular dynamics simulation, density function theory

INTRODUCTION

Janus kinase 1 (JAK1) is the most widely employed JAK, according to biochemical and genetic research, since it is involved in the signaling of the gamma common (γ c), beta common (β c), gp130, type I and type II interferon, IL-6, and IL-10 cytokine subfamilies (Kulagowski et al., 2012). JAK1 comprises seven homology domains (JH1–JH7) (Harpur et al., 1992). The C-terminal kinase module (JH1) is the protein's physiologically active catalytic domain. The JH2 domain is a catalytically inactive pseudokinase domain that has been found to interact with the JH1 domain and control its activity (Saharinen and Silvennoinen 2002). Two Src homology 2 (SH2) domains (JH3 and JH4) precede the FERM domain (JH5–JH7) at the N-terminus. The JH1 domain has an ATP-binding site, which has been the target of a number of small-molecule inhibitors. All four members of JAK have a highly conserved kinase domain, particularly at the ATP-binding region, which complicates the development of particular inhibitors (Caspers et al., 2016). The active sites of JAKs comprise multiple subdomains that include the β -glycine loop, the catalytic loop, and activation loops (Taldaev et al., 2022). The amino acid present in and around the hinge region serves critical functions in the integrity of kinase activity control. Furthermore, since this area is adjacent to the ATP-binding site in the catalytic cleft, it is reasonable to believe that the mutations in this region might promote constitutive activation of the kinase (Gorden et al., 2010; Haan et al., 2010).

All STAT proteins (STAT1–STAT6) that are ubiquitously expressed in all the tissues may be phosphorylated by JAK1 enzyme (Gruber et al., 2020). JAK1 has been shown in mouse knockout experiments to have a critical function in signal transduction (Itteboina et al., 2017). According to earlier research, JAK1 is ascendant over JAK3, and in the absence of JAK1, JAK3 is unable to activate STATs (Haan et al., 2011). Furthermore, recent studies have shown that JAK1 rather than JAK3 kinase is the primary driver of the immune-relevant cytokine activity (Menet et al., 2015). JAK1 is involved in various types of cancer. Activation of JAK1 kinase by IL-6 family cytokines appeared to be the mechanism for constitutive STAT3 activation in human ovarian cancer cells (Wen et al., 2014). In gastric cancer, by activating the JAK1/STAT3 pathway, the upregulation of HOXA10 gene increased cell proliferation, cloning formation, and tumorigenesis and lowered cell apoptosis (Chen et al., 2019). In lung adenocarcinoma patients, the overall survival time was substantially reduced in patients with EGFR-amplified tumors expressing greater levels of phosphorylated JAK1 compared with individuals with tumors without one or both of these traits. Additionally, JAK inhibition was demonstrated to limit the development of human lung adenocarcinoma with a K-RAS mutation (Xie et al., 2021).

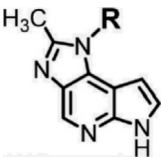
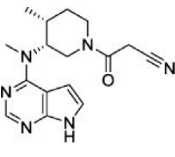
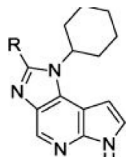
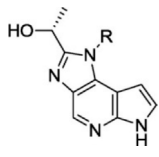
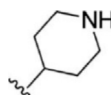
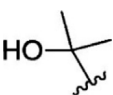
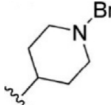
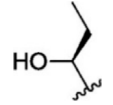
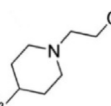
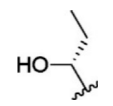

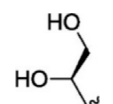
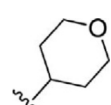
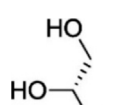
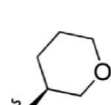
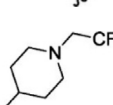
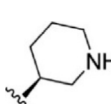
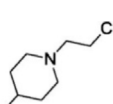
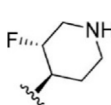
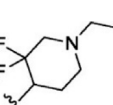
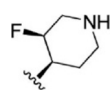
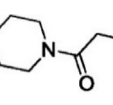
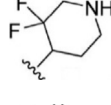
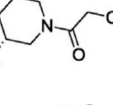
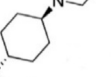
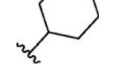
AML and breast cancer patients have exhibited several STAT5-activating JAK1 mutations (Hornakova et al., 2011). Moreover, in ER-negative breast cancer cell lines, the upregulation of phosphorylated JAK1 expression was observed (Yeh et al., 2007).

According to clinical and experimental investigations, rheumatoid arthritis synovial response may be influenced by the JAK1-mediated cytokine (IFN and IL-6) signaling. As a result, inhibiting JAK1 is regarded as a significant therapeutic strategy for the successful treatment of rheumatoid arthritis (Keretsu et al., 2021b). Recently, it has been discovered that inhibiting JAK1 selectively may be an effective therapy option for patients suffering from autoimmune and hematological illnesses because of the role that altered JAK1 signaling plays in these conditions (Kleppe et al., 2017). Moreover, JAK1 expression in cancer cells allows individual cells to contract, perhaps enabling them to transcend their tumor and spread to other areas of the body (Nordqvist 2011). Mutations in JAK1 are less common than in T-ALL patients with B-ALL or leukemia of the myeloid origin. In two AML patients, a JAK1 mutation V623A was found, emphasizing the capacity of constitutively active JAK1 to induce a variety of leukemias (Xiang et al., 2008; Raivola et al., 2021).

JAK inhibitors, which have been authorized for the treatment of cancer and autoimmune illnesses, have provided the first insight on the importance of JAK1 in NK cell biology (Schwartz et al., 2017). Ruxolitinib, JAK1/JAK2 inhibitor, has lowered the number of NK cells and hampered maturation and function in both mice and human patients (Schönberg et al., 2015; Bottos et al., 2016). Ruxolitinib's influence on NK cell development has been linked to JAK2 as well; therefore, it is not clear which of the two kinases is accountable for the reported results (Bottos et al., 2016; Kim et al., 2017). The fascinating finding by Sohn et al. (2013) has highlighted the importance of JAK1 inhibitor on the IL-6, IL-22, and INF-pathways. JAK1 inhibitors including ruxolitinib, tofacitinib, filgotinib, peficitinib, and numerous additional second-generation inhibitors are now under investigation for the treatment of inflammatory and autoimmune illnesses. Because of limited potency, non-targeting, and off-target effects (Keretsu et al., 2021a), new JAK1 inhibitors with high potency and selectivity are urgently needed.

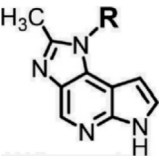
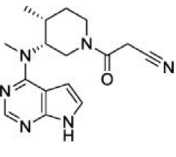
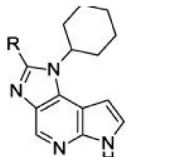
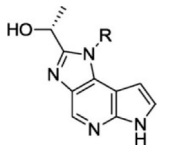
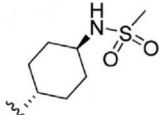

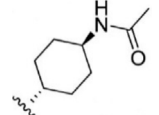

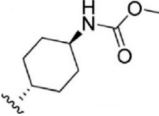
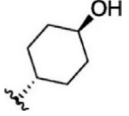
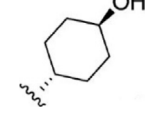
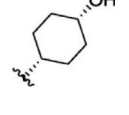
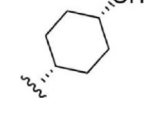
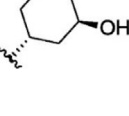
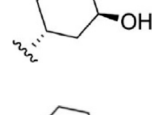
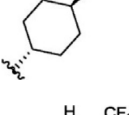
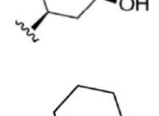
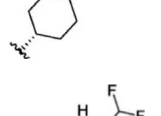
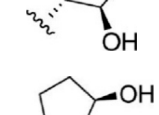
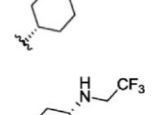
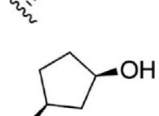
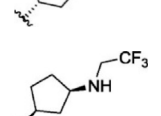

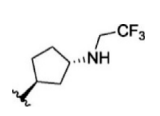

Pharmacophore models are widely employed to quantitatively explore common chemical characteristics among a considerable number of structures with great diversity (Taha et al., 2008; Xie et al., 2009). It is one of the widely used approaches to search for chemical databases and identify novel scaffolds for various targets (Wang H. et al., 2008; Lu et al., 2007). To discover the potent hits, the ligand-based and structure-based pharmacophore models can be used. In this study, the ligand-based pharmacophore models

TABLE 1 | The chemical structures and the biological activity of JAK1 inhibitors.

| Compound 1–21 | | | Compound 22 | | Compound 23–32 | | Compound 33–53 | |
|---|---|---------------------|---|------|---|---------------------|---|--|
|  | | |  | |  | |  | |
| S. no. | R | K _i (nM) | pK _i | S.no | R | K _i (nM) | pK _i | |
| 1 |  | 10 | 8.000 | 27 |  | 13 | 7.886 | |
| 2 |  | 1.3 | 8.886 | 28 |  | 16 | 7.796 | |
| 3 |  | 0.9 | 9.046 | 29 |  | 1.8 | 8.745 | |
| 4 |  | 1.3 | 8.886 | 30 |  | 7.2 | 8.143 | |
| 5 |  | 18 | 7.745 | 31 |  | 1.6 | 8.796 | |
| 6 |  | 2.8 | 8.553 | 32 |  | 2.6 | 8.585 | |
| 7 |  | 150 | 6.824 | 33 |  | 3.4 | 8.469 | |
| 8 |  | 1.2 | 8.921 | 34 |  | 2 | 8.699 | |
| 9 |  | 9.3 | 8.032 | 35 |  | 5.2 | 8.284 | |
| 10 |  | 2 | 8.699 | 36 |  | 43 | 7.367 | |
| 11 |  | 1.5 | 8.824 | 37 |  | 31 | 7.509 | |

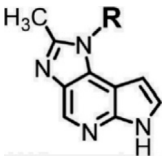
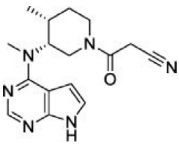
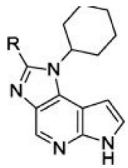
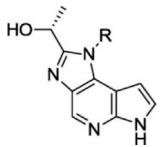

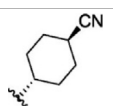

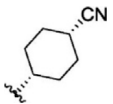

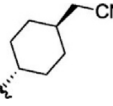

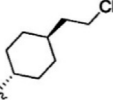
(Continued on following page)

TABLE 1 | (Continued) The chemical structures and the biological activity of JAK1 inhibitors.

| Compound 1–21 | | Compound 22 | | Compound 23–32 | | Compound 33–53 | |
|---|---|---|-------|--|---|---|-------|
|  | |  | |  | |  | |
| 12 |  | 4.5 | 8.347 | 38 |  | 68 | 7.167 |
| 13 |  | 6.1 | 8.215 | 39 |  | 2.8 | 8.553 |
| 14 |  | 2.6 | 8.585 | 40 |  | 5.4 | 8.268 |
| 15 |  | 5.8 | 8.237 | 41 |  | 12 | 7.921 |
| 16 |  | 6.7 | 8.174 | 42 |  | 2.7 | 8.569 |
| 17 |  | 1.8 | 8.745 | 43 |  | 4.9 | 8.310 |
| 18 |  | 7.3 | 8.137 | 44 |  | 0.8 | 9.097 |
| 19 |  | 90 | 7.046 | 45 |  | 1.1 | 8.959 |
| 20 |  | 4.8 | 8.319 | 46 |  | 53 | 7.276 |
| 21 |  | 5.4 | 8.268 | 47 |  | 2 | 8.699 |
| 22 | - | 0.7 | 9.155 | 48 |  | 1.2 | 8.921 |

(Continued on following page)

TABLE 1 | (Continued) The chemical structures and the biological activity of JAK1 inhibitors.

| Compound 1–21 | | Compound 22 | | Compound 23–32 | | Compound 33–53 | |
|---|---|---|-------|--|---|---|--------|
|  | |  | |  | |  | |
| 23 |  | 1.1 | 8.959 | 49 |  | 1.9 | 8.721 |
| 24 |  | 3.5 | 8.456 | 50 |  | 0.9 | 9.046 |
| 25 |  | 10 | 8.000 | 51 |  | 0.1 | 10.000 |
| 26 |  | 0.8 | 9.097 | 52 |  | 0.3 | 9.523 |

were generated using the 52 JAK1 inhibitors reported by Zak et al. It elucidates the spatial arrangement of structural features of various potent and structurally diverse inhibitors crucial for biological recognition. One efficacious approach toward the discovery and development of the drugs is the virtual screening of molecular libraries (Stahl et al., 2006). Virtual screening helps to identify the potential lead molecules and reduces the time and cost of the drug discovery process (Reddy et al., 2007). Thus, pharmacophore-based virtual screening was implemented. In many research works, it was proposed that the combination of pharmacophore modeling and molecular docking is a successful method to discover the novel and potent lead compounds (Sakkiah et al., 2009; Sakkiah et al., 2010; Sakkiah et al., 2011). Hence, the results of pharmacophore-based virtual screening were taken for molecular docking.

Docking results were used to predict the binding orientations of the hits as well as the filter to select the hits. The molecular docking results were validated by calculating the free energy of binding using the molecular mechanics-generalized born surface area (MM-GBSA) method for the protein–ligand complexes (Friesner et al., 2006). Furthermore, induced fit docking (IFD) was carried out to get additional understanding about the structure and flexibility of these hits into the binding site since IFD has been reported to be a powerful method to account for both receptor and ligand flexibility (Zhong et al., 2009). Subsequently, the cross-docking method was used to identify the selective hits by docking every hit to every receptor. By examining the results, the top five hits were selected and taken for

molecular dynamic simulation and density functional theory (DFT) study. To identify the potency and selectivity of the leads, a drug molecule named ruxolitinib was included in the study. The results of selected lead compounds and the drug were compared and analyzed.

MATERIALS AND METHODS

Dataset Selection

For ligand-based pharmacophore modeling, a set of 52 JAK1 inhibitors (C-2 methyl/hydroxyethyl imidazopyrrolopyridines derivatives) reported by Zak et al. (2012) and Zak et al. (2013) were selected because of their diverse biological activity. The K_i values of these inhibitors (0.1–150 nM) were derived using biochemical and cell-based assays. These inhibitors have shown higher selectivity toward JAK1 over JAK2. The experimental K_i values were converted into pK_i values that are simply the negative log of the K_i value. The chemical structures and biological activities of all molecules are given in **Table 1**.

Pharmacophore Model Generation

Phase 4.3, a high-performance program module of Schrödinger 2015, was used to generate the ligand-based pharmacophore models (Dixon et al., 2006). It uses a fine-grained conformational sampling method to predict a hypothesis consisting of the common pharmacophore features. The Ligprep module was used to clean, minimize, and generate

TABLE 2 | The summary of statistical data obtained for the pharmacophore hypotheses.

| S. no. | Hypothesis | Survival score | Survival inactive | Post hoc | Site | Vector | Volume |
|--------|---------------|----------------|-------------------|--------------|-------------|--------------|--------------|
| 1 | ADHRRR | 3.514 | 1.420 | 3.514 | 0.72 | 0.992 | 0.803 |
| 2 | AADHRR | 3.513 | 1.396 | 3.513 | 0.72 | 0.994 | 0.801 |
| 3 | AAADHR | 3.509 | 1.410 | 3.509 | 0.71 | 0.995 | 0.805 |
| 4 | DDHRRR | 3.424 | 1.371 | 3.424 | 0.77 | 0.953 | 0.705 |
| 5 | ADDHRR | 3.420 | 1.364 | 3.420 | 0.75 | 0.960 | 0.710 |
| 6 | AADDHR | 3.398 | 1.244 | 3.398 | 0.73 | 0.948 | 0.724 |
| 7 | AADRR | 4.378 | 2.106 | 3.607 | 0.85 | 0.992 | 0.761 |
| 8 | AAADR | 4.364 | 2.141 | 3.593 | 0.84 | 0.969 | 0.780 |
| 9 | ADRRR | 4.353 | 2.081 | 3.582 | 0.85 | 0.979 | 0.757 |
| 10 | ADRRR | 4.339 | 2.080 | 3.568 | 0.89 | 0.917 | 0.763 |
| 11 | AADDR | 4.330 | 1.877 | 3.559 | 0.88 | 0.920 | 0.755 |
| 12 | DDRRR | 4.292 | 1.841 | 3.521 | 0.87 | 0.901 | 0.747 |
| 13 | ADHRR | 4.259 | 1.747 | 3.487 | 0.80 | 0.979 | 0.709 |
| 14 | DHRRR | 4.257 | 1.883 | 3.486 | 0.81 | 0.969 | 0.710 |
| 15 | AADHR | 4.253 | 1.744 | 3.481 | 0.79 | 0.985 | 0.709 |
| 16 | AHRRR | 3.974 | 1.916 | 3.510 | 0.72 | 0.988 | 0.803 |
| 17 | AAHRR | 3.972 | 1.892 | 3.508 | 0.72 | 0.991 | 0.801 |
| 18 | AAHR | 3.965 | 1.907 | 3.501 | 0.70 | 0.993 | 0.806 |
| 19 | AAADH | 3.954 | 1.927 | 3.490 | 0.69 | 0.996 | 0.802 |
| 20 | DPRRR | 3.945 | 1.666 | 3.481 | 0.71 | 0.993 | 0.780 |
| 21 | ADPRR | 3.940 | 1.662 | 3.476 | 0.71 | 0.993 | 0.776 |
| 22 | AADPR | 3.929 | 1.631 | 3.465 | 0.69 | 0.993 | 0.778 |
| 23 | ADDHR | 3.896 | 1.788 | 3.432 | 0.73 | 0.959 | 0.741 |
| 24 | DDHRR | 3.896 | 1.794 | 3.432 | 0.78 | 0.951 | 0.701 |
| 25 | AADDH | 3.853 | 1.726 | 3.389 | 0.73 | 0.957 | 0.701 |
| 26 | DHHRR | 3.441 | 1.703 | 2.977 | 0.44 | 0.964 | 0.575 |
| 27 | ADHHR | 3.437 | 1.612 | 2.973 | 0.46 | 0.961 | 0.555 |

A- acceptor, D- donor, H- hydrophobic, R- aromatic ring, and P- positive group. The selected pharmacophore hypotheses are represented in bold.

conformations of all compounds. Based on the diversity of the chemical structure and its biological activity, the quantitative pharmacophore models were generated using the Develop Pharmacophore Model option. On the basis of biological activity distribution (pK_i values), the activity threshold value was set and the inhibitors were divided into actives, inactives, and moderately actives. In this study, both five and six featured pharmacophore hypotheses were generated by defining the minimum and maximum numbers of sites to five and six. The pharmacophore models were developed possessing different combinations of hydrogen-bond acceptor (A), hydrogen-bond donor (D), aromatic ring (R), hydrophobic group (H), positively ionizable (P), and negatively ionizable (N) groups. The resulting hypotheses were scored and ranked on the basis of scoring parameters. The scoring algorithm includes the alignment of site points and vectors, number of ligands matched, volume overlap, relative conformational energy, selectivity, and activity. The difference between the survival score and the survival inactive score notifies the ability of the hypotheses to correctly distinguish between actives and inactives.

Pharmacophore Model Validation

Since the pharmacophore model is just a theoretical model, it is necessary to analyze whether or not the generated model is able to predict the active compounds. Thus, two approaches, namely, potency validation and selectivity validation, were performed to measure the accuracy of pharmacophores in selecting the active compounds.

Potency Validation

Potency validation was carried out to test whether the pharmacophore model is good enough to pick a greater number of active molecules. This was achieved by screening the database consisting of both active molecules and decoys. Active molecules are the known inhibitors of JAK with higher biological activities, whereas decoys are the molecule that does not have any activity toward JAK and it was downloaded from DUD-E (a Database of Useful Decoys-Enhanced) database (Mysinger et al., 2012). DUD-E datasets were used only after removing the biasness through docking. Based on the number of actives and decoys retrieved by the pharmacophore models, statistical parameters such as Guner-Henry (GH) score, %A, %Y, and E score were calculated using the following formula:

$$\text{GH score} = \left(\frac{Ha (3A + Ht)}{4 * Ht * A} \right) \left(1 - \frac{Ht - Ha}{D - A} \right);$$

$$\%A = \frac{Ha}{A} * 100; \%Y = \frac{Ha}{Ht} * 100; E = \frac{Ha/Ht}{A/D},$$

where Ha is the number of actives in the hits list, Ht is the number of hits retrieved, A is the number of active compounds in the database, D is the number of compounds in the database, %A is the percentage of known active compounds obtained from the database, %Y is the percentage of known actives in the hits list, and E is the enrichment of the concentration of actives by the model relative to random screening without a pharmacophoric approach. GH score ranges from 0 to 1, which indicates a null model and an ideal model,

TABLE 3 | Pharmacophore validation results from potency validation.

| S. no. | Hypothesis | H _a (#40) | Decoys (#1000) | H _i | %A | %Y | E | GH score |
|--------|---------------|----------------------|----------------|----------------|--------------|---------------|--------------|--------------|
| 1 | ADHRRR | 11 | 0 | 11 | 27.50 | 100.00 | 26.00 | 0.819 |
| 2 | AADHRR | 11 | 10 | 21 | 27.50 | 52.38 | 13.62 | 0.457 |
| 3 | AAADHR | 11 | 22 | 33 | 27.50 | 33.33 | 8.67 | 0.312 |
| 4 | DDHRRR | 21 | 0 | 21 | 52.50 | 100.00 | 26.00 | 0.881 |
| 5 | ADDHRR | 24 | 13 | 37 | 60.00 | 64.86 | 16.86 | 0.628 |
| 6 | AADDHR | 22 | 27 | 49 | 55.00 | 44.90 | 11.67 | 0.461 |
| 7 | AADRR | 23 | 106 | 129 | 57.50 | 17.83 | 4.64 | 0.248 |
| 8 | AAADR | 21 | 242 | 263 | 52.50 | 7.98 | 2.08 | 0.145 |
| 9 | ADRRR | 23 | 9 | 32 | 57.50 | 71.88 | 18.69 | 0.677 |
| 10 | ADRRR | 26 | 39 | 65 | 65.00 | 40.00 | 10.40 | 0.444 |
| 11 | AADDR | 22 | 128 | 150 | 55.00 | 14.67 | 3.81 | 0.216 |
| 12 | DDRRR | 28 | 1 | 29 | 70.00 | 96.55 | 25.10 | 0.898 |
| 13 | ADHRR | 33 | 85 | 118 | 82.50 | 27.97 | 7.27 | 0.381 |
| 14 | DHRRR | 31 | 12 | 43 | 77.50 | 72.09 | 18.74 | 0.726 |
| 15 | AADHR | 32 | 252 | 284 | 80.00 | 11.27 | 2.93 | 0.213 |
| 16 | AHRRR | 11 | 3 | 14 | 27.50 | 78.57 | 20.43 | 0.656 |
| 17 | AAHRR | 11 | 78 | 89 | 27.50 | 12.36 | 3.21 | 0.149 |
| 18 | AAADR | 11 | 141 | 152 | 27.50 | 7.24 | 1.88 | 0.106 |
| 19 | AAADH | 11 | 54 | 65 | 27.50 | 16.92 | 4.40 | 0.185 |
| 20 | DPRRR | 11 | 0 | 11 | 27.50 | 100.00 | 26.00 | 0.819 |
| 21 | ADPRR | 11 | 4 | 15 | 27.50 | 73.33 | 19.07 | 0.616 |
| 22 | AADPR | 12 | 10 | 22 | 30.00 | 54.55 | 14.18 | 0.479 |
| 23 | ADDHR | 23 | 71 | 94 | 57.50 | 24.47 | 6.36 | 0.304 |
| 24 | DDHRR | 29 | 17 | 46 | 72.50 | 63.04 | 16.39 | 0.643 |
| 25 | AADDH | 19 | 132 | 151 | 47.50 | 12.58 | 3.27 | 0.185 |
| 26 | DHHRR | 23 | 53 | 76 | 57.50 | 30.26 | 7.87 | 0.351 |
| 27 | ADHHR | 19 | 164 | 183 | 47.50 | 10.38 | 2.70 | 0.164 |

The number of compounds used for the validation study is mentioned within parenthesis. The selected pharmacophore hypotheses are represented in bold.

respectively. GH score >0.6 indicates the acceptable quality of the pharmacophore model and is useful in differentiating the known active molecules from inactives and suitable for retrieving active JAK1 inhibitors (Sathe et al., 2014; Li et al., 2015).

Selectivity Validation

Selectivity validation was performed to check which pharmacophore models are more selective in choosing high number of JAK1 molecules. Selectivity validation was carried out in two ways. First, a database comprising 30 JAK1, 30 JAK2, 30 JAK3, and 30 TYK2 molecules (Yang et al., 2007; Wang et al., 2009; Pissot-Soldermann et al., 2010; Ioannidis et al., 2011; Kulagowski et al., 2012) was created and used for validation. The ability of pharmacophore models to differentiate the selective JAK inhibitors was evaluated using virtual screening workflow on a manually curated database. Second, to further confirm the selectivity of the selected models, the available 288 JAK1 (Kulagowski et al., 2012; Labadie et al., 2012; Hurley et al., 2013; Labadie et al., 2013), 627 JAK2 (Lucet et al., 2006; Wang et al., 2009; Pissot-Soldermann et al., 2010; Harikrishnan et al., 2011; Ioannidis et al., 2011; Schenkel et al., 2011; Dugan et al., 2012; Forsyth et al., 2012; Lynch et al., 2013; Vazquez et al., 2018), and 431 JAK3 (Chrencik et al., 2010; Thoma et al., 2011; Jaime-Figueroa et al., 2013; Lynch et al., 2013; Soth et al., 2013; De Vicente et al., 2014; Duan et al., 2014) inhibitors from diverse research papers that mention either IC₅₀ values or K_i values of these inhibitors were taken for validation.

Pharmacophore-Based Virtual Screening

Virtual screening is the process where the complete databases are used to identify the molecules in the database which are most likely to bind to a drug target (Vyas et al., 2008). In this study, pharmacophore-based virtual screening was carried out using the “find matches to hypothesis” option available in the phase module which efficiently search for pharmacophore matches from the database of fixed conformers. The pharmacophore-based virtual screening was performed against Maybridge (53,000) (www.maybridge.com), Lifechemicals (12, 92, 000) (<http://lifechemicals.com/>), Enamine (24,91,318) (<http://enamine.net/>), Chemdiv (15,00,000) (<http://www.chemdiv.com/>), Asinex (398,022) (<http://www.asinex.com/>), and Zinc chemical and Zinc natural databases (<http://zinc.docking.org/>) (44,92,226) (Irwin and Shoichet 2005; Irwin et al., 2012; Sterling and Irwin 2015) to identify the new molecules with novel scaffolds. After screening, fitness score that is a measure of how well the hypothesis matched to the aligned ligand conformers based on RMSD site matching, volume terms, and vector alignments was used to filter the molecules.

Absorption, Distribution, Metabolism, and Excretion Prediction

After virtual screening, the molecular descriptors and pharmaceutically applicable properties of the hits were calculated using Qikprop 4.4. Qikprop generates the

TABLE 4 | Pharmacophore validation results from selectivity validation.

| S. no. | Hypothesis | No. of inhibitors retrieved | | | | | | |
|--------|---------------|-----------------------------|------------|------------|------------|-------------|-------------|-------------|
| | | JAK1 (#30) | JAK2 (#30) | JAK3 (#30) | TYK2 (#30) | JAK1 (#288) | JAK2 (#627) | JAK3 (#431) |
| 1 | ADHRRR | 5 | - | 1 | - | 24 | 0 | 1 |
| 2 | AADHRR | - | 6 | 1 | - | - | - | - |
| 3 | AAADHR | - | 4 | - | - | - | - | - |
| 4 | DDHRRR | 8 | - | 1 | - | 54 | 0 | 1 |
| 5 | ADDHRR | 1 | 9 | 1 | - | - | - | - |
| 6 | AADDHR | 2 | 9 | - | - | - | - | - |
| 7 | AADRR | 7 | 16 | 3 | 3 | - | - | - |
| 8 | AAADR | 4 | 9 | 1 | - | - | - | - |
| 9 | ADRRR | 10 | - | 3 | - | 77 | 22 | 5 |
| 10 | ADRRR | 3 | - | 4 | - | - | - | - |
| 11 | AADRR | 2 | 9 | 1 | - | - | - | - |
| 12 | DDRRR | 18 | - | 2 | - | 86 | 1 | 10 |
| 13 | ADHRR | 23 | - | 3 | - | - | - | - |
| 14 | DHRRR | 26 | - | 1 | - | 142 | 56 | 20 |
| 15 | AADHR | 21 | 13 | 1 | 3 | - | - | - |
| 16 | AHRRR | - | - | 7 | - | - | - | - |
| 17 | AAHRR | - | 13 | 1 | - | - | - | - |
| 18 | AAHR | - | 8 | 1 | - | - | - | - |
| 19 | AAADH | - | 4 | - | - | - | - | - |
| 20 | DPRRR | 9 | - | - | - | 19 | 0 | 0 |
| 21 | ADPRR | 9 | 2 | - | - | 25 | 0 | 0 |
| 22 | AADPR | 9 | 1 | - | - | - | - | - |
| 23 | ADDHR | 4 | 9 | - | - | - | - | - |
| 24 | DDHRR | 17 | - | 1 | b | 60 | 6 | 37 |
| 25 | AADDH | - | 9 | - | - | - | - | - |
| 26 | DHHR | 17 | 13 | 3 | - | - | - | - |
| 27 | ADHHR | 12 | 23 | 4 | 3 | - | - | - |

The number of compounds used for the validation study is mentioned within parenthesis. The selected pharmacophore hypotheses are represented in bold.

physicochemical properties for a compound to find whether the compound follows drug likeliness properties. Lipinski's rule characterizes the important molecular properties of drug, including absorption, distribution, metabolism, and excretion (ADME) that is essential for a drug's pharmacokinetics in the human body (Jorgensen and Duffy 2002). Parameters that determine the ADME of the molecules were Molweight (Molecular weight), QPlogPo/w (partition coefficient), QPlogS (water solubility), percentage of human oral absorption, and intestinal absorption parameters such as Caco-2 and MDCK permeability. The compounds are expected to be active in humans only if the molecule passes through Lipinski's rule of five. Therefore, the compounds retrieved after filtration were subjected to ADME prediction and its physicochemical properties were analyzed.

Molecular Docking

Molecular docking predicts the binding mode and interaction of the small molecule to the protein. It distinguishes the behavior of small molecules in the binding site of target protein and explicates its fundamental biochemical processes (Gschwend et al., 1996; Lipinski 2000). The binding conformations of the hits inside the JAK1 ATP-binding site were investigated using Grid-based Ligand Docking with Energetics (Glide 6.7) module. The ATP-binding site of JAK1 comprises Leu881, Glu883, Val889, Ala906, Met956, Glu957, Phe958, Leu959, Gly962, Ser963, Glu966, Arg1007, Asn1008, Leu1010, Gly1020, and Asp1021 residues.

Before docking, protein preparation wizard was used to prepare protein structure (3EYG) (Williams et al., 2009) applying the default parameters that include adding hydrogens, filling missing atoms and residues using PRIME, assigning correct bond orders, and hydrogen-bond optimization and minimization. In the Receptor Grid Generation panel, the center of the grid box was defined on the centroid of the co-crystallized ligand (MI1), and the volume in the active-site region of the receptor was calculated by default settings (van der Waals radius scaling factor 1.0 and partial charge cutoff 0.25). Molecular docking was performed using both the Standard Precision (SP) and Extra Precision (XP) docking modes in which the receptor was held rigid and the ligand was free to move (Jain 2003; Halgren et al., 2004). Glide score is a combination of hydrophobic, hydrophilic, van der Waals energy, metal binding groups, freezing rotatable bonds, and polar interactions with the receptor. Comparing Glide SP and XP score, Glide SP score is a softer and more forgiving function whereas Glide XP score is a harder function and adept at reducing the false positives. Therefore, Glide XP score was considered for the selection of hits and further analysis.

MM-GBSA Calculations

The binding free-energy calculations procured via the MM-GBSA method are more precise and consistent than the glide XP score and improve the ranking of potential leads (Lyne et al., 2006; Das et al., 2009; Yang et al., 2009). Therefore, the binding free energy (ΔG_{bind}) of the protein-ligand complexes was calculated using

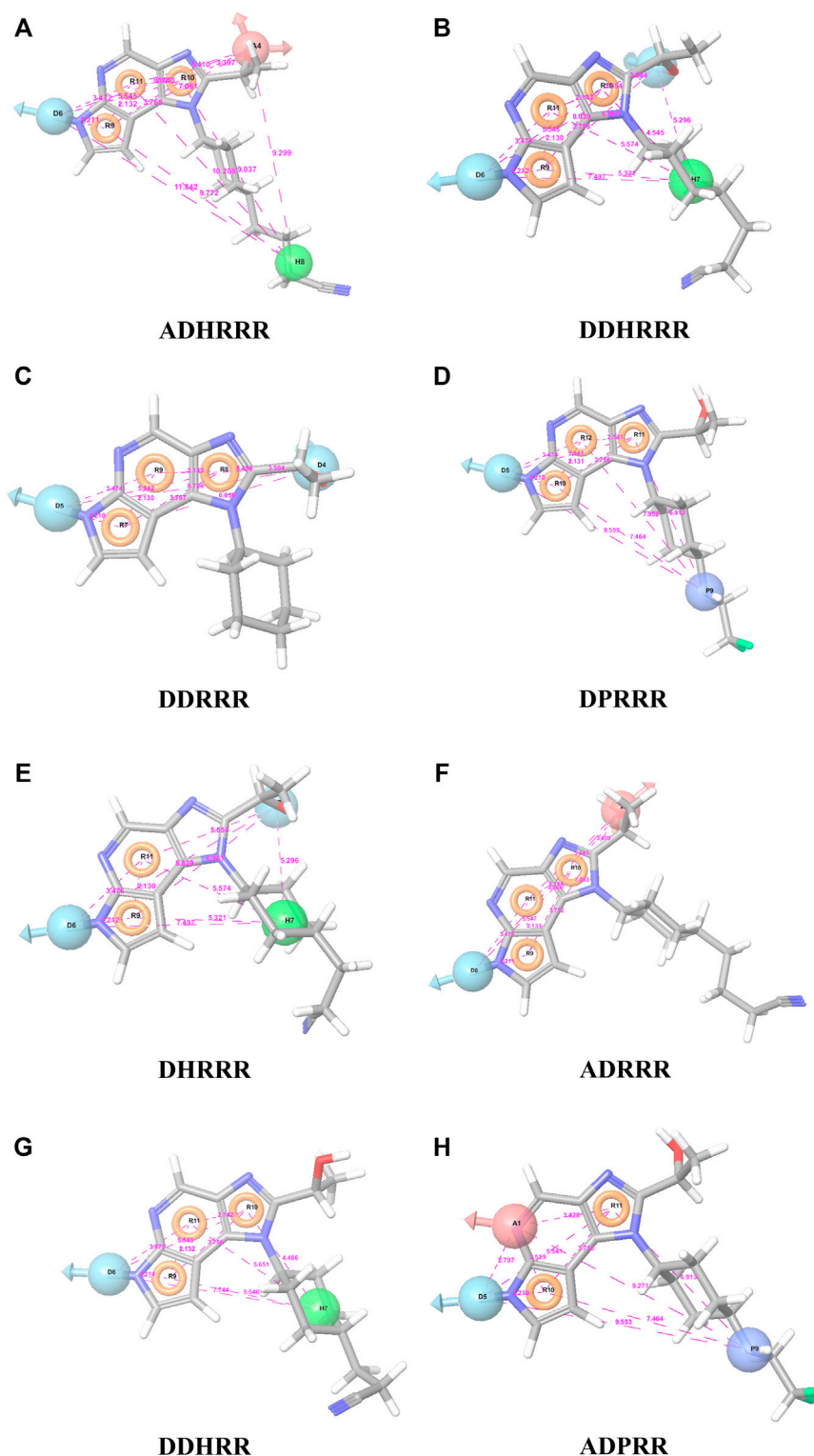


FIGURE 1 | The representation of selected pharmacophore models **(A)** ADHRRR, **(B)** DDHRRR, **(C)** DDRRR, **(D)** DPRRR, **(E)** DHRRR, **(F)** ADRRR, **(G)** DDHRR, and **(H)** ADPRR. Pharmacophore features are colored in light blue, brown, dark blue, brick red, and green contours representing the H-bond donor (D), H-bond acceptor (A), positives (P), aromatic ring (R), and hydrophobic (H) groups, respectively. The distances between the pharmacophore features (A) are given in pink dotted lines.

TABLE 5 | Number of hits obtained from pharmacophore-based virtual screening.

| S. no. | Hypothesis | Maybridge | Lifechemicals | Enamine | Asinex | Chemdiv | Zinc |
|--------|------------|-----------|---------------|---------|--------|---------|------|
| 1 | ADHRRR | 0 | 10 | 5 | 16 | 4 | 0 |
| 2 | DDHRRR | 0 | 6 | 9 | 0 | 1 | 0 |
| 3 | DDRRR | 3 | 117 | 167 | 18 | 16 | 16 |
| 4 | DPRRR | 1 | 57 | 68 | 9 | 40 | 14 |
| 5 | DHRRR | 9 | 250 | 321 | 161 | 151 | 155 |
| 6 | ADRRR | 3 | 134 | 224 | 335 | 17 | 113 |
| 7 | DDHRR | 9 | 151 | 558 | 44 | 88 | 213 |
| 8 | ADPRR | 1 | 74 | 436 | 30 | 9 | 202 |

the Prime MM-GBSA module implemented in Schrödinger 2015. The Prime MM-GBSA Module incorporates the OPLS3 force field and the VSGB dissolvable model to look through calculations (Li et al., 2011). The energy difference between the free and complex states of protein and ligand was calculated. The energy components such as covalent binding energy, van der Waals energy, generalized born electrostatic solvation energy, Coulomb energy, total energy, and H-bond correction were retrieved from the calculations.

Induced Fit Docking

In the docking protocol, to retain the flexibility of the receptor, a mixed molecular docking protocol called induced fit docking (IFD) developed by Schrödinger 2015 was employed (Wang H. Y. et al., 2008). IFD uses the refinement module in Prime to account for the receptor flexibility and Glide to account for the ligand flexibility (Jacobson et al., 2004). Protein preparation wizard and the Ligprep module were used for protein and ligand preparation, respectively. Grid was generated on the ATP-binding site amino acid residues based on the co-crystallized ligand. The ATP-binding site residues and their flexibility were considered for the IFD protocol. IFD was carried out with default parameters, and 20 conformational poses were calculated for each ligand. IFD scores (IFD score = 1.0 Glide_Gscore +0.05 Prime_Energy) were calculated based upon the total energy of the system and the protein–ligand interaction energy and used to rank the IFD poses (Luo et al., 2014). The electrostatic interactions formed between the receptor and the ligand were calculated by the docking scores under “Electro,” and hydrophilic interactions under “Lipophilic Evdw” mention the lipophilicity component acquired from the hydrophobic grid. IFD poses were ranked based on the scores, and the best pose was chosen for each hit.

Cross Docking

Cross docking is the process of taking a series of complexes of ligand–receptor pairs and docking every ligand to every receptor. This is used to study the specificity of the ligands and the receptors and, thus, yield valuable report regarding the effects of ligand upon binding. Protein preparation wizard and LigPrep were used to prepare proteins and the shortlisted hits, respectively. Grid was generated on the ATP-binding site residues. The hits shortlisted from the molecular docking study were docked against JAK1, JAK2, and JAK3 using the Glide XP module to identify the selective lead compounds.

Molecular Dynamics Simulation

Docking results could be the instantaneous state and were not considered decisive because binding of the inhibitor to a protein in an *in vivo* state is a dynamic process. For advanced studies, the stable binding mode of the ligand is more reliable. Hence, to explore the detailed binding modes and compare the stability and molecular interactions of the docked lead complexes, molecular dynamics simulation was carried out for 100ns using GROningen MAchine for Chemical Simulations (GROMACS version 2016.3 installed in Centos 7.3) software (Abraham et al., 2015). GROMACS works according to Newton’s laws of motions and simulates the behavior of bio-molecules such as nucleic acids, proteins, lipids, ligands, ions, and water. The coordinates for MD simulations have been achieved from the docking results. The PRODRG server (<http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrgr>) was used to calculate the ligand parameters in the framework of GROMOS96 54a7 force field. The SPC water model was used as a solvent during simulation. To achieve the stability of the simulated system, the potential energy, temperature, and pressure were monitored during the simulations. The temperature and pressure of the system were equilibrated (from ps to ns) till they reach 300 K and 1.05 bar, respectively. The stability of the secondary structure elements and conformational changes of the simulated complexes were evaluated by root mean square deviations (RMSDs), root mean square fluctuation (RMSF), radius of gyration (Rg), and solvent-accessible surface area (SASA) values obtained from MD trajectories. The molecular dynamics study was performed using High-Performance Computing server (Intel Xeon 14 core processor with 28 threads and 2.40 GHz processor speed).

MM-PBSA Calculation

The molecular mechanics energies combined with the Poisson–Boltzmann and surface area continuum solvation (MM/PBSA) method have been applied to predict binding free energies and to evaluate the relative stabilities of different bimolecular structures. The MM/PBSA calculations were performed for the simulated systems using g_mmpbsa, a GROMACS Tool for High-Throughput MM-PBSA Calculations (Kumari et al., 2014). Combined with molecular dynamics (MD) simulations, MM-PBSA can also incorporate conformational fluctuations and entropic contributions to the binding energy (Homeyer and Gohlke 2012).

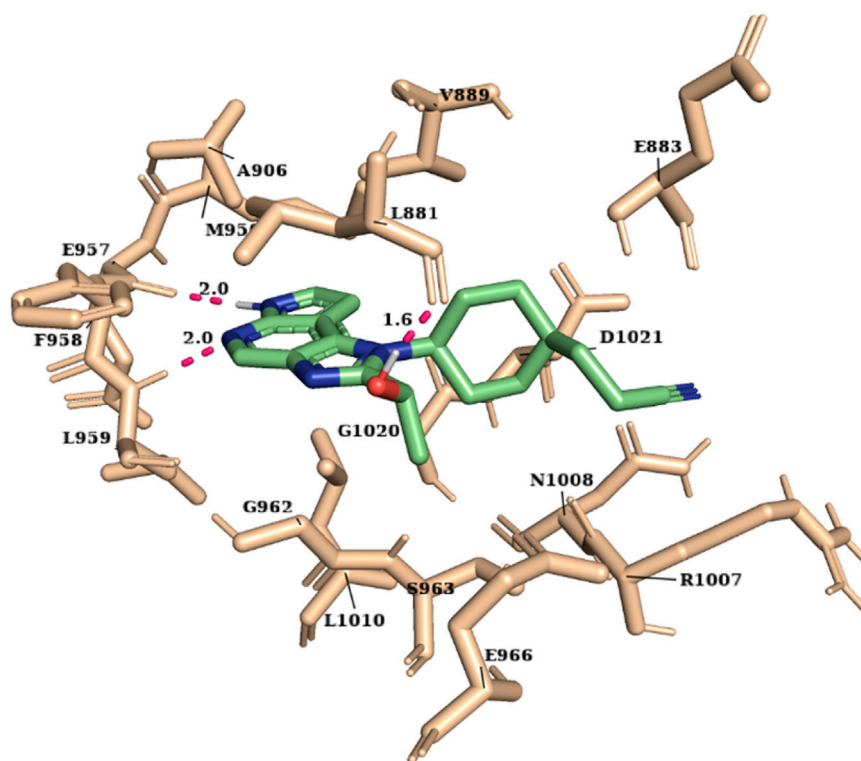


FIGURE 2 | The binding of highly active compound 51 into the ATP-binding site of JAK1.

Density Functional Theory Study

The density functional theory (DFT) study was carried out to observe the chemical behavior of the lead compounds using the electron density-relevant concepts (Zhao et al., 2011). Also, it provides a quantum-level understanding of the molecules and assists in building the relationship between the electronic properties and the biological activity of the molecule (Nagarajan et al., 2018). Molecular descriptors such as total energy, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), band energy gap (ΔE), molecular dipole moment, absolute hardness (η), global softness (σ), chemical potential (μ), electronegativity (χ), and electrophilicity index (ω) were studied for the selected lead compounds using Gaussian 16 software. Initially, the molecules were optimized using the B3LYP function with a 6-31G(d) basis set to calculate their molecular properties such as total energy and molecule dipole moment (Becke 1998). The dipole moment relates to the electro-chemical reactivity of the compounds. The electron donating and accepting ability of the molecules HOMO energy (E_{HOMO}) and LUMO energy (E_{LUMO}), respectively, were calculated.

RESULTS AND DISCUSSION

Pharmacophore Model Generation

In the phase module, ligand-based pharmacophore model generation was carried out utilizing 52 JAK1 inhibitors named

C-2 methyl/hydroxyethyl imidazopyrrolopyridine derivatives (Table 1) along with their activity values. Ten molecules whose $pK_i > 8.9$ were taken as actives, twelve molecules whose $pK_i < 8.1$ were taken as inactive, and the remaining thirty-one molecules were considered to be intermediates. Twenty-seven different pharmacophore hypotheses (six with six featured pharmacophores and twenty-one with five featured pharmacophores) were generated and put through the stringent scoring function. The generated pharmacophores were ranked by aligning them with the active ligands, and the statistical data obtained after scoring are tabulated in Table 2. Besides the survival active score, survival inactive score, and post-hoc score, fitness score was considered to measure the quality of the pharmacophores. The fitness score was calculated between the pharmacophores and the highly active (compound 51) and highly inactive (compound 7) compounds in the dataset. For all pharmacophores, the fitness score was higher with the highly active compound compared with the inactive compound. Subsequently, the pharmacophores were evaluated using different validation methods.

Pharmacophore Model Validation Potency Validation

For potency validation, a database containing 40 JAK1 actives and 1,000 decoys was created. The generated pharmacophore models were allowed to screen this database to calculate the GH score. It was observed that six featured hypotheses have picked less

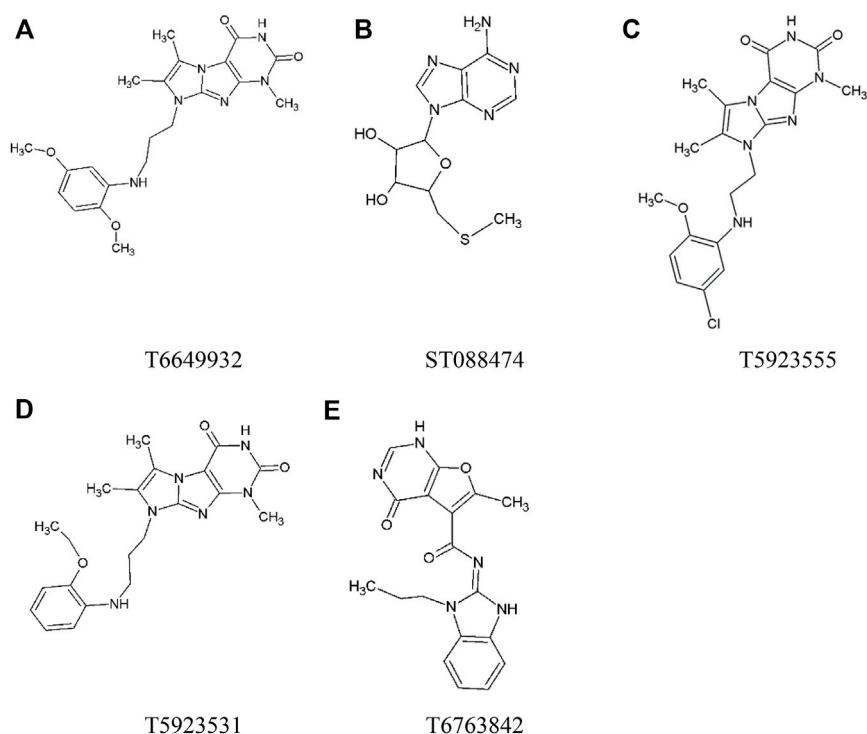


FIGURE 3 | The chemical structure of selected lead compounds. **(A)** T6649932, **(B)** ST088474, **(C)** T5923555, **(D)** T5923531, and **(E)** T6763842.

TABLE 6 | The drug likeliness properties of the selected lead compounds and the drug.

| S. no. | Molecule ID | molMW (130.0–725.0) | dHB (0.0–6.0) | aHB (2.0–20.0) | logPo/w (–2.0–6.5) | logS (–6.5–0.5) | logBB (–3.0–1.2) | PCaco (<25 poor>500 great) | PMDCK (<25 poor>500 great) |
|--------|-------------|------------------------|---------------|----------------|-----------------------|--------------------|---------------------|----------------------------------|----------------------------------|
| 1 | T6649932 | 426.5 | 2 | 8 | 3.5 | –5.3 | –1.1 | 756.168 | 365.722 |
| 2 | ST088474 | 297.3 | 4 | 10 | –0.4 | –2.1 | –1.3 | 132.825 | 95.256 |
| 3 | T5923555 | 416.9 | 2 | 7 | 3.3 | –5.1 | –0.8 | 582.955 | 674.629 |
| 4 | T5923531 | 410.5 | 2 | 7 | 3.4 | –5.4 | –1.1 | 679.576 | 325.851 |
| 5 | T6763842 | 351.4 | 1 | 6 | 2.7 | –4.4 | –1.2 | 385.229 | 176.425 |
| 6 | Ruxolitinib | 306.4 | 2 | 4.5 | 1.4 | –3.3 | –0.4 | 941.735 | 463.628 |

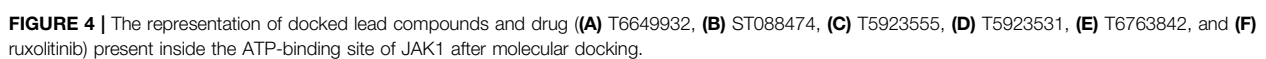
molMW, molecular weight; dHB, donor atoms; aHB, acceptor atoms; logPo/w, partition coefficient; logS, aqueous solubility; logBB, brain/blood partition coefficient; PCaco, predicted apparent; Caco-2, cell permeability in nm/sec; PMDCK, predicted apparent MDCK cell permeability in nm/sec. The qikprop recommended values are given inside the parenthesis.

TABLE 7 | Molecular docking results of the selected JAK1 lead compounds and the drug.

| S. no. | Molecule ID | XP score | Glide energy | Glide evdw | Glide ecoul | H-bond interaction |
|--------|-------------|----------|--------------|------------|-------------|--|
| 1 | T6649932 | –10.335 | –61.771 | –53.839 | –7.932 | Leu959, Glu957, Arg1007 |
| 2 | ST088474 | –10.653 | –50.800 | –34.600 | –16.200 | Leu959, Glu957, Leu881, Ser963, Glu966 |
| 3 | T5923555 | –10.015 | –57.500 | –49.348 | –8.151 | Leu959, Glu957, Arg1007 |
| 4 | T5923531 | –10.303 | –57.350 | –49.885 | –7.465 | Leu959, Glu957, Arg1007 |
| 5 | T6763842 | –10.671 | –51.703 | –46.166 | –5.536 | Leu959, Glu957 |
| 6 | Ruxolitinib | –9.282 | –57.553 | –43.488 | –14.065 | Leu959, Glu957 |

number of decoys compared with five featured hypotheses. The ADHRRR, DDHRRR, and DPRRR hypotheses were more potent because they do not pick any decoys. DDRRR, ADPRR, and

AHRRR have picked very less number of decoys. DDRRR, ADHRR, DHRRR, AADHR, DDHRR, and DDRRR have picked more active molecules. The results of potency



ADDHRR, DDRR, DPRR, DHRR, ADRR, AHRR, DDHRR, and ADPRR have obtained the GH score >0.6 indicating the goodness of these hypotheses.

TABLE 8 | MM-GBSA results of the selected JAK1 lead compounds and the drug.

| S. no. | Molecule ID | ΔG_{Bind} | $\Delta G_{\text{Bind_Coulomb}}$ | $\Delta G_{\text{Bind_Covalent}}$ | $\Delta G_{\text{Bind_Lipo}}$ | $\Delta G_{\text{Bind_vdW}}$ |
|--------|-------------|--------------------------|-----------------------------------|------------------------------------|--------------------------------|-------------------------------|
| 1 | T6649932 | -46.430 | -10.816 | 6.512 | -15.475 | -55.026 |
| 2 | ST088474 | -44.915 | -29.704 | 5.740 | -8.571 | -36.645 |
| 3 | T5923555 | -45.936 | -11.512 | 13.940 | -15.199 | -45.837 |
| 4 | T5923531 | -41.698 | -14.070 | 4.485 | -11.368 | -47.637 |
| 5 | T6763842 | -43.157 | -6.535 | 5.179 | -15.922 | -48.815 |
| 6 | Ruxolitinib | -46.184 | -18.883 | 1.186 | -13.853 | -38.185 |

TABLE 9 | Induced fit docking results of the selected JAK1 lead compounds and the drug.

| S. no. | Molecule ID | XP score | IFD score | Lipophilic EvdW | Electro | H-bond interaction |
|--------|-------------|----------|-----------|-----------------|---------|-------------------------|
| 1 | T6649932 | -8.557 | -598.316 | -5.175 | -0.353 | Leu959, Glu883 |
| 2 | ST088474 | -8.791 | -594.272 | -2.492 | -0.519 | Leu959, Glu957, Arg1007 |
| 3 | T5923555 | -9.343 | -598.889 | -4.142 | -0.638 | Leu959, Glu957, Ser963 |
| 4 | T5923531 | -9.941 | -598.697 | -4.919 | -0.701 | Leu959, Glu883 |
| 5 | T6763842 | -9.550 | -599.541 | -5.590 | -0.236 | Ser963 |
| 6 | Ruxolitinib | -9.725 | -595.395 | -3.654 | -1.055 | Leu959, Glu957 |

TABLE 10 | Cross-docking results of the selected JAK1 lead compounds and the drug.

| S. no. | Molecule ID | Glide XP Gscore | | | Glide energy (Kcal/mol) | | |
|--------|-------------|-----------------|--------|---------|-------------------------|---------|---------|
| | | JAK1 | JAK2 | JAK3 | JAK1 | JAK2 | JAK3 |
| 1 | T6649932 | -10.623 | -6.065 | -8.192 | -65.073 | -59.156 | -57.574 |
| 2 | ST088474 | -10.784 | -8.938 | -7.963 | -52.980 | -44.112 | -40.231 |
| 3 | T5923555 | -10.650 | -7.314 | -8.298 | -56.889 | -54.956 | -54.964 |
| 4 | T5923531 | -10.608 | -7.810 | -8.344 | -59.950 | -57.629 | -55.855 |
| 5 | T6763842 | -10.652 | -5.243 | -8.754 | -51.761 | -48.136 | -48.034 |
| 6 | Ruxolitinib | -9.178 | -9.091 | -10.209 | -49.475 | -48.884 | -48.434 |

Selectivity Validation

Initially, the selectivity validation was performed with a set of 30 JAK1, 30 JAK2, 30 JAK3, and 30 TYK2 molecules retrieved from different studies. DPRRR has picked only JAK1 molecules. The DHRRR, ADHRR, DDRRR, and DDHRR hypotheses have picked a high number of JAK1 molecules and very less JAK3 molecules. ADHRR and DDHRR have picked only few JAK1 and one JAK3 molecules. The results of selectivity validation are tabulated in **Table 4**. The fitness score for JAK1 inhibitors was greater than or equal to 1.5, whereas for other JAK inhibitors, the fitness score was <1.5 for most of the molecules indicating that the pharmacophore models were able to map well with the JAK1 inhibitors (Sathe et al., 2014; Babu et al., 2015).

Six feature pharmacophore hypotheses were more potent but not highly selective to JAK1. Based on potency and selectivity validation results, the DDHRRR, ADHRRR, DDRRR, DPRRR, DHRRR, ADRRR, DDHRR, and ADPRR hypotheses were selected because they were successful in retrieving active compounds from the database. The representation of the selected JAK1 pharmacophore models showing the distances between the pharmacophoric sites is shown in **Figures 1A–H**. On mapping the selected pharmacophore models with highly

active compound 51 and inactive compound 7, it was observed that the fitness score was >2.5 for the highly active compound mapping with all pharmacophore features whereas inactive compound 7 could map with either four or five pharmacophore features with low fitness score. The highest fitness score with compound 51 suggests screening using these models would pick the similar active compounds. From the results, we suggest the combination of two or three aromatic rings (R) and one or two donor atoms (D) with a hydrophobic (H) group is an important pharmacophoric feature for identifying the selective JAK1 inhibitors. The important pharmacophore features obtained were compared with the contribution maps obtained through the hologram-based fingerprint technique (**Supplementary Figure S1**). The contribution maps depict the imidazopyrrolopyridine ring which possesses one donor and three aromatic rings responsible for the intermediate contribution of the inhibitory activity. From the highly active compounds, we observed the cyano group attached to cyclohexanes (yellow) and the hydroxyethyl group attached to imidazopyrrolopyridines (green); a hydrophobic and an aromatic/donor group, respectively, are strongly responsible for the higher activity. Thus, these pharmacophore features are highly important for the inhibitory activity of JAK1.

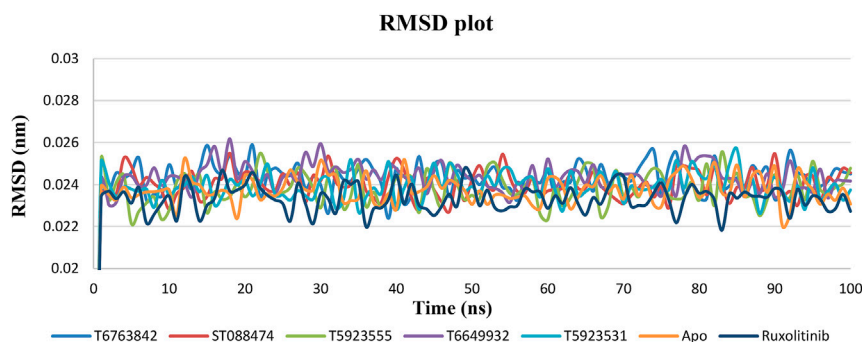


FIGURE 5 | The change in RMSD values of the backbone C α atoms of JAK1 systems over a period of 100 ns after binding with the lead compounds and drug.

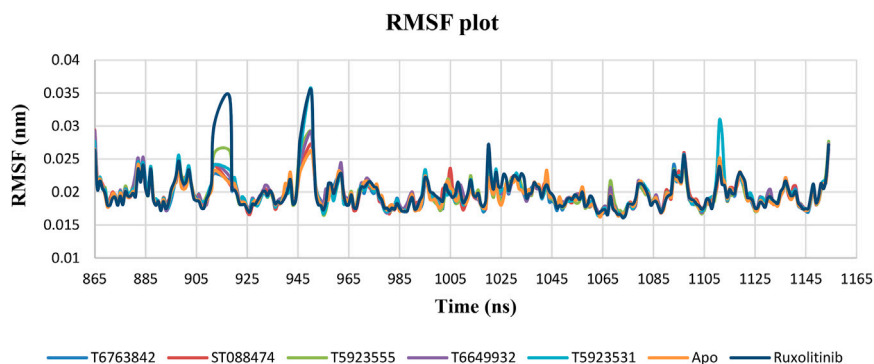


FIGURE 6 | The change in RMSF values of JAK1 residues over a period of 100 ns after binding with the lead compounds and drug.

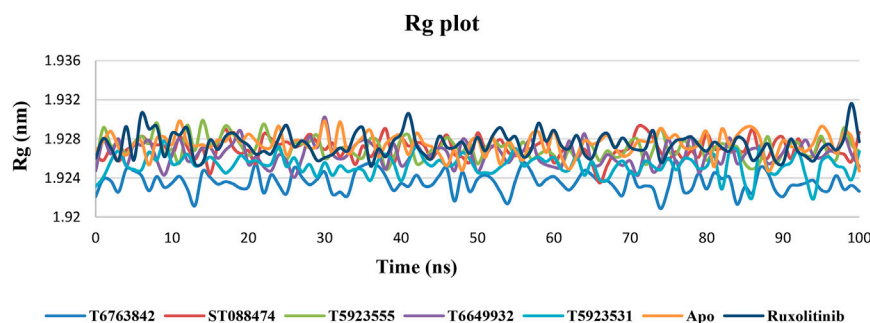


FIGURE 7 | The change in Rg values over a period of 100 ns after binding with the lead compounds and drug.

To confirm the selectivity of the selected pharmacophore models (DDHRRR, ADHRRR, DDRR, DPRR, DHRRR, ADRR, DDHRR, and ADPRR), the second round of selectivity validation was carried out with a set of 288 JAK1, 627 JAK2, and 431 JAK3 inhibitors with diverse activity. It was observed (**Table 4**) that all the selected pharmacophore models were able to pick more number of JAK1 inhibitors compared with its subtypes. Hence, the selected pharmacophore models were capable of discriminating the JAK1 inhibitors and appropriate for retrieving the novel and selective JAK1 inhibitors.

Pharmacophore-Based Virtual Screening

The selected pharmacophore models were screened against Maybridge, Lifechemicals, Enamine, Chemdiv, Asinex, and Zinc (chemical and natural) databases for the identification of new hits. The identified hits contain the structural features that overlap with the selected pharmacophore models. The hits obtained were ranked and filtered based on the fitness score. The fitness score was set to >1.5 for the Maybridge, Asinex, Chemdiv, Lifechemicals, and Enamine databases, whereas for the Zinc database, the fitness score was set to >2 because of high

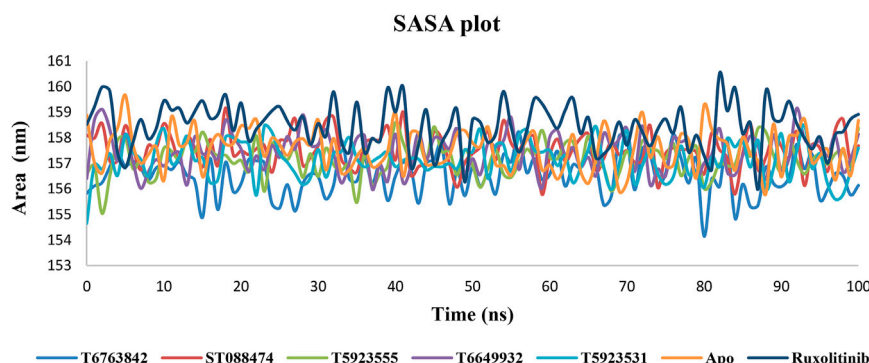


FIGURE 8 | The change in SASA values over a period of 100 ns after binding with the lead compounds and drug.

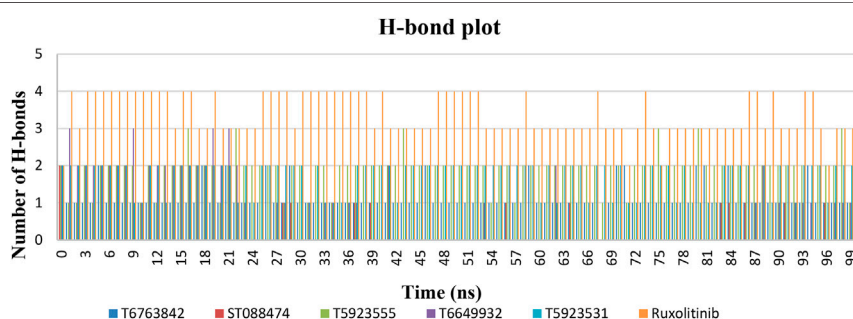


FIGURE 9 | The number of hydrogen bonds formed by lead compounds and drug over the simulation time.

number of molecules retrieved from the Zinc database. As a result of screening and filtration, 4,265 compounds were retrieved. The total numbers of hits retrieved from different databases are tabulated in **Table 5**.

The potentiality of the pharmacophore models was validated using receiver operating curves (ROCs) utilizing the screened molecules (Hevener et al., 2009); 10 compounds identified from the pharmacophore-based virtual screening were seeded with 500 decoys. Enrichment was estimated based on how well the compounds were fetched. After ranking the decoy set and

docked compounds by the Glide score, the enrichment was calculated using the ROC plot that provides the report on sensitivity and specificity. The ROC plot inferred that Glide XP ranked seven compounds in top 10% with the ROC value as 0.93 and the AUC value as 0.92. 80% of the true positives were fetched in top 20% which indicates its capability of retrieving the active compounds. The gentle increase in the ROC curve (**Supplementary Figure S2**) was noticed in the beginning, which implies that number of true positives was sacrificed to reduce the amount of false positives.

TABLE 11 | The protein–ligand interaction analysis of the selected JAK1 lead compounds and the drug before, during, and after MD simulation.

| S. no. | Molecule ID | H-bond interaction | | | | | |
|--------|-------------|--|------------------------|------------------------|------------------------|-------------------------|------------------|
| | | Before simulation | During simulation | | | | After simulation |
| | | | 25 ns | 50 ns | 75 ns | 99 ns | |
| 1 | T6649932 | Leu959, Glu957, Arg1007 | - | Leu959 | - | Leu959 | - |
| 2 | ST088474 | Leu959, Glu957, Leu881, Ser963, Glu966 | Asp1021 | Arg1007, Val1009 | Arg1007 | Asp1021 | Asp1021 |
| 3 | T5923555 | Leu959, Glu957, Arg1007 | Leu959, Glu957 | Leu959, Glu957 | Leu959, Glu957, Leu881 | Leu959, Glu957 | Leu959, Glu957 |
| 4 | T5923531 | Leu959, Glu957, Arg1007 | Leu959, Glu957 | Leu959, Glu957 | Leu959, Glu957 | Leu959, Glu957 | Leu959, Glu957 |
| 5 | T6763842 | Leu959, Glu957 | Ser963 | Ser963 | Ser963 | Ser963 | Ser963 |
| 6 | Ruxolitinib | Leu959, Glu957 | Leu959, Glu957, Leu881 | Leu959, Glu957, Leu881 | Leu959, Glu957 | Leu959, Glu957, Arg1007 | Leu959, Glu957 |

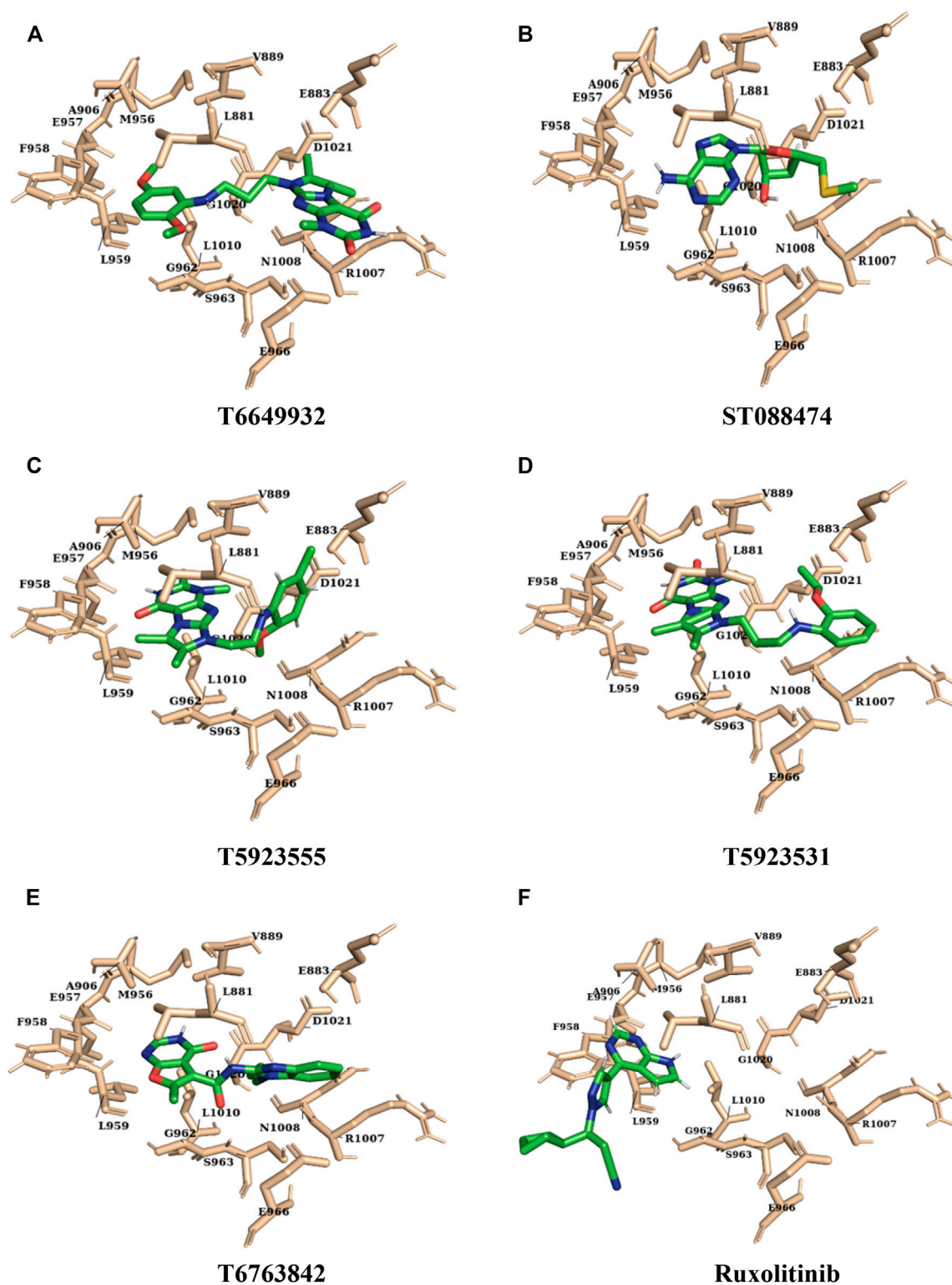


FIGURE 10 | The representation of final conformation of the docked lead compounds and drug ((**A**) T6649932, (**B**) ST088474, (**C**) T5923555, (**D**) T5923531, (**E**) T6763842, and (**F**) ruxolitinib) present inside the ATP-binding site of JAK after molecular dynamics simulation.

TABLE 12 | MM-PBSA results obtained from the molecular dynamics trajectory for the selected JAK1 lead compounds and the drug.

| S. no. | Molecule ID | van der Waals energy (kJ/mol) | Electrostatic energy (kJ/mol) | Polar solvation energy (kJ/mol) | SASA energy (kJ/mol) | Binding energy (kJ/mol) |
|--------|-------------|-------------------------------|-------------------------------|---------------------------------|----------------------|-------------------------|
| 1 | T6649932 | -194.226±16.777 | -74.970±32.984 | 203.123±32.793 | -18.207±1.508 | -24.281±30.279 |
| 2 | ST088474 | -135.299±6.176 | -19.370±4.198 | 97.066±9.529 | -14.916±0.839 | -22.519±9.921 |
| 3 | T5923555 | -212.143±8.554 | -2.460±7.922 | 256.402±9.826 | -19.217±0.649 | -42.581±11.158 |
| 4 | T5923531 | -214.550±9.829 | -13.931±5.705 | 219.798±14.381 | -20.106±0.900 | -38.790±15.145 |
| 5 | T6763842 | -163.284±11.976 | -89.322±11.300 | 230.810±17.465 | -15.839±0.839 | -27.636±15.186 |
| 6 | Ruxolitinib | -185.994±50.799 | -34.821±65.738 | 206.801±108.464 | -17.040±6.025 | -24.054±36.312 |

TABLE 13 | The statistical results of the DFT-based descriptors for the selected lead compounds and the drug.

| S. no. | Total energy (a.u.) | Energy of | | ΔE | Dipole moment (debye) | η | σ | χ | μ | ω |
|--------|---------------------|------------------|------------------|------|-----------------------|------|------|-------|------|------|
| | | εHOMO (Kcal/mol) | εLUMO (Kcal/mol) | | | | | | | |
| 1 | -1,195.61 | -0.20 | -0.04 | 4.48 | 3.06 | 2.24 | 0.22 | -3.20 | 3.20 | 2.29 |
| 2 | -1,325.80 | -0.21 | -0.01 | 5.43 | 3.48 | 2.71 | 0.18 | -3.08 | 3.08 | 1.75 |
| 3 | -1,675.02 | -0.29 | -0.07 | 6.09 | 5.59 | 3.04 | 0.16 | -4.83 | 4.83 | 3.84 |
| 4 | -1824.84 | -0.22 | -0.08 | 3.84 | 8.02 | 1.92 | 0.26 | -4.08 | 4.08 | 4.33 |
| 5 | -1,695.05 | -0.25 | -0.07 | 4.97 | 5.14 | 2.48 | 0.20 | -4.28 | 4.28 | 3.69 |
| R | -987.14 | -0.22 | -0.05 | 4.59 | 4.09 | 2.29 | 0.22 | -3.64 | 3.64 | 2.88 |

T6649932 (1), ST088474 (2), T5923555 (3), T5923531 (4), T6763842 (5), ruxolitinib (R). ΔE, band energy gap (εLUMO-εHOMO); η, absolute hardness; σ, global softness; χ, electronegativity; μ, chemical potential; ω, electrophilicity index.

Absorption, Distribution, Metabolism, and Excretion Prediction

Compounds that pass Lipinski's rule of five and other ADME properties of the drug are expected to be active in humans. Properties such as molecular weight, H-bond donors, H-bond acceptors, log p, van der Waals surface, aqueous solubility, blockage of HERG K⁺ channels, apparent Caco-2 cell permeability, apparent MDCK cell permeability, brain/blood partition coefficient, skin permeability, binding to human serum albumin, and human oral absorption of the hits were studied. Finally, 2,856 compounds whose drug-like properties were in the acceptable range (according to qikprop recommended range) were selected and subsequently exposed to glide SP and XP docking protocols to remove both the false-positive and false-negative hits.

Molecular Docking

The molecular docking study was carried out using the Glide SP and XP modes to explore the binding mode and interaction of hits on the ATP-binding site. The crystal structure of JAK1 protein 3EYG in complex with MI1 (Williams et al., 2009) was used to perform molecular docking. The grid was developed on the centroid of co-crystallized ligand MI1 surrounding the ATP-binding site residues (Leu881, Glu883, Val889, Ala906, Met956, Glu957, Phe958, Leu959, Gly962, Ser963, Glu966, Arg1007, Asn1008, Leu1010, Gly1020, and Asp1021) of JAK1. Initially, the docking of MI1 into the ATP-binding site was performed to check the accuracy and reproducibility of the docking program. Subsequently, the highly active compound 51 and 2,856 hits were

docked into the ATP-binding site. Considering the docking result of compound 51 (glide XP score -9.691), the glide XP threshold value was set to ≥ -9.60 to identify the novel hits. We observed that 90 molecules have exhibited glide score greater than the threshold and it was shortlisted (**Supplementary Table S1**). Among the JAK1 ATP-binding site residues, Leu959 and Glu957 that are present in the hinge region were found to be the most selective amino acid residues for the H-bond interaction and also crucial for selective inhibition of JAK1. Hence, the interactions with Leu959 and Glu957 were investigated for the hits. Compound 51 has shown H-bond interactions with Leu959, Glu957, and Leu881. The selected 90 hits have exhibited H-bond interaction with either Leu959 or Glu957 or both residues. Additionally, the Leu881, Glu883, Ser963, Glu966, and Arg1007 residues were involved in H-bond interaction with most of the hits. The hydrophobic interactions were formed mainly by the residues Leu881, Val889, Ala906, Val938, Met956, Phe958, Pro960, and Leu1010. The binding of compound 51 into the ATP-binding site is shown in **Figure 2**.

MM-GBSA Calculations

The highly ranked hits selected from glide docking were taken for MM-GBSA calculations to predict the binding energy of the protein-ligand complexes. The calculated free energy of binding (ΔG_{bind}) was lower than glide energy. It was observed that van der Waals (ΔG_{Bind_vdW}) energy contributes more for the ligand binding, whereas covalent interaction (ΔG_{Bind_Covalent}) and electrostatic salvation (ΔG_{Bind_Solv_GB}) energy terms disfavor for the inhibitor binding.

Induced Fit Docking

IFD was performed on the highly ranked hits using the crystal structure of JAK1 (3EYG). It was observed that IFD also produces good IFD score and XP score comparable to glide XP score. The IFD score of JAK1 hits was greater than or equal to -590 , and their corresponding XP score was greater than -8.00 which indicates the good binding ability of the hits. The observed H-bond and hydrophobic interaction with the IFD results was highly similar to glide results, indicating that these hits could bind and produce similar H-bond and hydrophobic interaction inside the binding site upon both receptor and ligand flexibility.

Cross Docking

Since an important objective of this work is to attain admissible levels of intra-family selectivity, the cross-docking approach was employed for the highly ranked hits. For cross docking, the crystal structure of JAK1 (3EYG), JAK2 (3KRR), and JAK3 (3ZEP) was used. Among 90 hits tested, the top five compounds (T6649932, ST088474, T5923555, T5923531, and T6763842) that have the highest docking score toward JAK1 (>-10.5) compared with JAK2 and JAK3 were selected and taken for further study.

Analysis of Selected Lead Compounds

The top five compounds that showed good potency and selectivity were selected and analyzed. To identify the potency and selectivity of the leads, a drug molecule named ruxolitinib was included in the study. Molecular docking, MM-GBSA calculations, IFD docking, and cross docking were performed for the drug and compared with the selected leads. Subsequently, the selected leads were taken for the molecular dynamics simulation study using the GROMACS and DFT calculations using Gaussian. The chemical structures of the selected lead compounds are shown in **Figures 3A–E**.

Absorption, Distribution, Metabolism, and Excretion Properties

ADME properties are the key determinants for the successful development of new drugs. All the analyzed pharmacokinetic parameters of these lead compounds were found to be within the permissible range. The percentage of the human oral absorption was found to be greater than 50%. The partition coefficient and water solubility that are important for the assessment of absorption and distribution of drugs within the body ranged between -0.4 and 3.5 and -5.4 and -2.1 , respectively. Compounds T6649932, T5923555, and T5923531 possessing good Caco-2 and MDCK permeability have good level of intestinal absorption. The drug-likeness properties of the selected leads and ruxolitinib are given in **Table 6**.

Glide XP Docking Analysis

For the selected leads, the glide XP score was greater than -10.015 , whereas for ruxolitinib, it was -9.282 . The highest docking score was observed for T6763842 (-10.671). The major contribution of vdW interactions was observed which indicate that the vdW interaction favors the protein-ligand complex. The glide XP score, glide energy, glide evdw, and glide ecout of the selected leads are given in **Table 7**. On

analyzing the interaction, it was observed that the lead compounds showed conserved H-bond interactions with both the selective residues of JAK1 (Leu959 and Glu957) similar to the drug indicating its remarkable selectivity. Compounds such as T6649932, T5923555, and T5923531 formed another H-bond with Arg1007. Additionally, hydrophobic interactions were formed with the ATP-binding site residues Leu881, Val889, Ala906, Val938, Met956, Phe958, Pro960, and Leu1010. **Figures 4A–F** show the docked pose of lead compounds and drug inside JAK1 ATP-binding site.

MM-GBSA Analysis

The binding free energy of the selected lead compounds ranges from -41.698 to -46.430 suggesting good binding affinity with JAK1. Many lead compounds have shown comparable free energy of binding with ruxolitinib. This provides the insight that these lead compounds have exhibited good specificity. Furthermore, the contribution of $\Delta G_{\text{Bind_vdW}}$ and $\Delta G_{\text{Bind_Lipo}}$ components to the binding free energy was compared. The high binding free energy was majorly contributed by $\Delta G_{\text{Bind_vdW}}$ than $\Delta G_{\text{Bind_Lipo}}$ component. The predicted binding free energy of the selected leads and drug is tabulated in **Table 8**.

Induced Fit Docking Analysis

The accuracy of glide scoring function in identifying the leads was checked using the IFD method. The IFD score was greater than -594 , and their corresponding docking score was greater than -8.5 . The IFD scores of four lead compounds were higher compared to ruxolitinib (-595.395), whereas the docking scores of four lead compounds were little lower compared with ruxolitinib (-9.725). The highest IFD score and XP score were observed for T6763842 and T5923531. The Electro and Lipophilic Evdw scores of the lead compounds and drugs showed higher lipophilicity compared with electrostatic interactions which implies the important role of lipophilicity in inhibitory activity. The glide XP score, IFD score, lipophilic evdw, electro, and H-bond interaction of the selected leads are given in **Table 9**. The representation of docked lead compounds and drug present inside the ATP-binding site of JAK1 after induced fit docking is shown in **Supplementary Figure S3**. The IFD results also confirmed that the selected lead compounds have occupied the ATP-binding site of JAK1 irrespective of receptor flexibility.

Cross-Docking Analysis

The cross-docking results of selected leads indicate that their docking score was greater than -10.00 with JAK1, whereas with respect to other JAKs, their docking score ranges from -5.2 to -8.9 . For ruxolitinib, docking scores were -9.178 with JAK1, -9.091 with JAK2, and -10.209 with JAK3. Therefore, the selected lead compounds have shown good selectivity in terms of docking score compared with ruxolitinib. The important components to determine the selectivity of the lead compounds were the electrostatic and hydrophilic components of the docking score. The drug and lead compounds showed higher lipophilicity compared with electrostatic interactions which implies that lipophilicity plays an important role in dictating the selectivity of these molecules. The cross-docking

results of the drug and selected leads are given in **Table 10**. The cross-docking results of ruxolitinib showed higher binding affinity with JAK3 compared with JAK1 and JAK2 indicating its lesser selectivity, whereas the selected lead compounds showed greater affinity and selectivity toward JAK1 compared with other JAK subtypes. Hence, the cross-docking results indicate the selected lead compounds are more selective than the drug.

Molecular Dynamics Simulation

RMSD Plot Analysis

RMSD relative to the respective initial conformations was monitored and analyzed to examine the stability and equilibration of all systems. The RMSD value for both the lead compounds and drug was calculated for the 100ns time scale using the apo form of JAK1 (3EYG) as reference. In **Figure 5**, it was observed that the RMSD values of both the drug and lead complexes were stable throughout the simulation. Furthermore, RMSD values for all protein backbone atoms attained convergence after 1 ns and maintained a plateau of 0.01 nm after the initial convergence. This suggests that all system simulations reached equilibrium and stabilization during the simulation. RMSD of both drug and lead complexes was found to be in the range of 0.02–0.03 nm indicating the similar stability. The observed smaller RMSD fluctuations for all compounds confirmed that the obtained binding conformations of these lead compounds and drug were highly reasonable.

RMSF Plot Analysis

RMSF of backbone atoms were monitored to identify the strong binding interactions and exemplify the pliability of these lead complexes in the ATP-binding site of JAK1. The RMSF plot shown in **Figure 6** indicates very minimal fluctuations were observed during the simulation except the terminal and loop regions of the protein. Most of the fluctuations were between 0.016 and 0.035 nm indicating the stability of the simulated system. Very minimal fluctuations were observed in the residues Pro912, His918, Glu946, Asn950, and Gly951 for all lead compounds and ruxolitinib. The JAK1 ATP-binding site residues that are crucial for binding and fixing the inhibitors have displayed insignificant fluctuations during the course of simulation. However, the most important and selective amino acid residues Leu959 and Glu957 that are important for inhibitor binding attained a quite stable behavior.

Rg Plot Analysis

The level of compactness in the structure of protein due to the presence or absence of ligands was calculated using radius of gyration (Rg) plot (Lobanov et al., 2008). It can be observed that all lead complexes and the drug showed consistently lower Rg values and exhibited a relatively similar nature of compactness in **Figure 7**. Thus, a relatively consistent Rg value indicates that a stably folded structure was observed throughout the MD simulation.

Solvent Accessible Surface Area Plot Analysis

The solvent accessible surface area (SASA) calculation of the protein–ligand complexes was used for predicting the extent of the conformational changes that occurred during the interaction. The

SASA plot shown in **Figure 8** indicates that no significant changes in the protein structure were caused by these lead compounds and drug during simulation. Hence, the protein–ligand complexes are relatively stable throughout the simulation.

Protein–Ligand Interaction Analysis

The most significant part in MD simulations is the analysis of protein–ligand interactions because it illustrates the changes in the binding mode of the ligands during simulations. **Figure 9** shows the number of H-bond formations over the trajectory for lead compounds and the drug. The H-bonds were the principal binding forces between protein and ligand. The drug ruxolitinib has produced 2–4 H-bonds, whereas lead compounds have produced 0–2 H-bonds throughout the simulation. T5923555, T5923531, and T6763842 have produced 1–3 H-bonds, whereas ST088474 produced one H-bond with JAK1 all through the simulation. Ruxolitinib, T6763842, and T5923555 had retained two H-bonds, and T5923531 had retained one H-bond, whereas T6649932 does not have an H-bond at the end of the simulation. A strong hydrogen bond network was formed mainly by the residues Glu957 and Leu959. T5923555 and T5923531 retained hydrogen bonds with Leu959 and Glu957 at the end of simulation. Moreover, the ATP-binding site residues were almost hydrophobic, which can form strong nonpolar interactions with lead compounds. The detailed protein–ligand interaction residues before and after molecular dynamics simulation (Saddala and Adi 2018) were studied and are given in **Table 11**. T5923555 and T5923531 were found to be more stable and reliable before and after simulation, and their important interaction (Glu957 and Leu959) remains unchanged throughout the simulation.

The binding mode of the drug and lead compounds after simulation is represented in **Figures 10A–F**. It was inferred that the initial docked conformation and the final conformation of the lead compounds and drug lie in the same binding pocket (**Supplementary Figure S4**). Hence, the conformation of the lead compounds was stable inside the binding pocket which, in turn, validates the reliability of the docking results. Furthermore, these absolute results suggest that the identified lead compounds are highly selective and potent and they can be taken for *in vitro* and *in vivo* studies.

MM-PBSA Calculation

The average binding energy of all the simulated complexes was calculated using the g_mmpbsa tool. The van der Waals energy, electrostatic energy, polar solvation energy, solvent-accessible surface area (SASA) energy, and binding energy were calculated and are tabulated in **Table 12**. T5923555 and T5923531 have shown good binding energy and van der Waals energy compared with other compounds.

Density Functional Theory Calculation

Molecular descriptors based on the electron density of the molecules were studied using Gaussian. Based on HOMO energy (E_{HOMO}) and LUMO energy (E_{LUMO}), descriptors such as ΔE , η , σ , μ , χ , and ω were calculated. The smaller energy gap (ΔE) for all lead compounds suggests that they can easily transit from HOMO to LUMO, which is important for the molecular

reactivity. Since the decrease in electronegativity (χ) value is proportional to the increase in inhibitive efficiency (Zhan et al., 2003), these leads would have higher inhibitory activity because of their lower electronegativity value. The statistical values of the calculated molecular descriptors are tabulated in **Table 13**. The smaller energy gap, lower electronegativity, and higher dipole moment that are vital for the inhibitory effect of a molecule were observed which validates the better inhibitory activity for the selected lead compounds.

CONCLUSION

Pharmacophore modeling, virtual screening, and molecular docking are the rational methods for the identification of novel leads with diverse chemical scaffold. Therefore, ligand-based pharmacophore modeling combined with virtual screening and docking was applied in this study to discover novel, potent, and selective virtual hits for JAK1 enzyme. Initially, the ligand-based pharmacophore models were generated and validated using the potency and selectivity validation methods. Eight pharmacophore models were selected and taken for pharmacophore-based virtual screening against six databases. The hits obtained from screening were filtered through ADME prediction and molecular docking. The binding free-energy calculation and induced fit docking methods were employed to validate the docking results. Subsequently, cross docking was carried out to identify the lead compounds that are more selective toward JAK1. Finally, the top five lead compounds were selected and taken for molecular dynamics and the DFT study. Among the five compounds, T5923555 and T5923531 were found to be the best leads and can be further validated using *in vitro* and *in vivo* methods.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

REFERENCES

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* 1–2, 19–25. doi:10.1016/j.softx.2015.06.001
- Babu, S., Kulkarni, S. A., Sohn, H., and Madhavan, T. (2015). Identification of Leads through In Silico Approaches Utilizing Benzylthio-1h-Benzo[d]imidazol-1-Yl Acetic Acid Derivatives: A Potent CRTh2 Antagonist. *J. Mol. Struct.* 1102, 25–41. doi:10.1016/j.molstruc.2015.08.031
- Becke, A. D. (1998). A New Inhomogeneity Parameter in Density-Functional Theory. *J. Chem. Phys.* 109 (6), 2092–2098. doi:10.1063/1.476722
- Bottos, A., Gotthardt, D., Gill, J. W., Gattelli, A., Frei, A., Tzankov, A., et al. (2016). Decreased NK-Cell Tumour Immunosurveillance Consequent to JAK Inhibition Enhances Metastasis in Breast Cancer Models. *Nat. Commun.* 7 (1), 12258–12312. doi:10.1038/ncomms12258
- Caspers, N. L., Han, S., Rajamohan, F., Hoth, L. R., Geoghegan, K. F., Subashi, T. A., et al. (2016). Development of a High-Throughput crystal Structure-

AUTHOR CONTRIBUTIONS

SB and TM have designed the experiment. SB has contributed to data acquisition, analysis, and interpretation. SB, SN, and SS validated the results. SB has drafted the manuscript. VN, HS, and TM contributed to the critical examination of the manuscript.

FUNDING

This research was supported by Start-Up Research Grant for Young Scientist (SB/YS/LS-128/2013), funded by the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India.

ACKNOWLEDGMENTS

This study was supported by the Computational Biology Lab, funded by the SERB Young Scientist grant (SB/YS/LS-128/2013). Author SB thanks CSIR, New Delhi, India, for providing Senior Research Fellowship (SRF). HS thanks the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. 2021R1F1A1062300) for the research support. This study was also supported by a research fund from Chosun University, 2021. The authors also express gratitude to StemOnc R&D Private Ltd., for the valuable revision of the manuscript. The authors thank the High-Performance Computing Centre, SRM Institute of Science and Technology, for providing the computational facility. This work is part of a PhD thesis submitted by SB to SRM Institute of Science and Technology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.837369/full#supplementary-material>

- Determination Platform for JAK1 Using a Novel Metal-Chelator Soaking System. *Acta Crystallogr. F Struct. Biol. Commun.* 72 (11), 840–845. doi:10.1107/S2053230X16016356
- Chen, W., Wu, G., Zhu, Y., Zhang, H., Zhou, Y., et al. (2019). HOXA10 Deteriorates Gastric Cancer through Activating JAK1/STAT3 Signaling Pathway. *Cancer Manag. Res.* 11, 6625–6635. doi:10.2147/CMAR.S201342
- Chrencik, J. E., Patny, A., Leung, I. K., Korniski, B., Emmons, T. L., Hall, T., et al. (2010). Structural and Thermodynamic Characterization of the TYK2 and JAK3 Kinase Domains in Complex with CP-690550 and CMP-6. *J. Mol. Biol.* 400 (3), 413–433. doi:10.1016/j.jmb.2010.05.020
- Das, D., Koh, Y., Tojo, Y., GhoshMitsuya, A. K. H., and Mitsuya, H. (2009). Prediction of Potency of Protease Inhibitors Using Free Energy Simulations with Polarizable Quantum Mechanics-Based Ligand Charges and a Hybrid Water Model. *J. Chem. Inf. Model.* 49 (12), 2851–2862. doi:10.1021/ci900320p
- De Vicente, J., Lemoine, R., Bartlett, M., Hermann, J. C., Hekmat-Nejad, M., Henningsen, R., et al. (2014). Scaffold Hopping towards Potent and Selective JAK3 Inhibitors: Discovery of Novel C-5 Substituted Pyrrolopyrazines. *Bioorg. Med. Chem. Lett.* 24 (21), 4969–4975. doi:10.1016/j.bmcl.2014.09.031

- Dixon, S. L., Smondyrev, A. M., Knoll, E. H., Rao, S. N., Shaw, D. E., and Friesner, R. A. (2006). PHASE: a New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput. Aided. Mol. Des.* 20 (10), 647–671. doi:10.1007/s10822-006-9087-6
- Duan, J. J., Lu, Z., Jiang, B., Yang, B. V., Doweiko, L. M., Nirschl, D. S., et al. (2014). Discovery of Pyrrolo[1,2-b]pyridazine-3-Carboxamides as Janus Kinase (JAK) Inhibitors. *Bioorg. Med. Chem. Lett.* 24 (24), 5721–5726. doi:10.1016/j.bmcl.2014.10.061
- Dugan, B. J., Gingrich, D. E., Mesaros, E. F., Milkiewicz, K. L., Curry, M. A., Zulli, A. L., et al. (2012). A Selective, Orally Bioavailable 1,2,4-Triazolo[1,5-a]pyridine-Based Inhibitor of Janus Kinase 2 for Use in Anticancer Therapy: Discovery of CEP-33779. *J. Med. Chem.* 55 (11), 5243–5254. doi:10.1021/jm300248q
- Forsyth, T., Kearney, P. C., Kim, B. G., Johnson, H. W., Aay, N., Arcalas, A., et al. (2012). SAR and *In Vivo* Evaluation of 4-Aryl-2-Aminoalkylpyrimidines as Potent and Selective Janus Kinase 2 (JAK2) Inhibitors. *Bioorg. Med. Chem. Lett.* 22 (24), 7653–7658. doi:10.1016/j.bmcl.2012.10.007
- Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra Precision glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* 49 (21), 6177–6196. doi:10.1021/jm051256o
- Gordon, G. M., Lambert, Q. T., Daniel, K. G., and Reuther, G. W. (2010). Transforming JAK1 Mutations Exhibit Differential Signalling, FERM Domain Requirements and Growth Responses to Interferon- γ . *Biochem. J.* 432 (2), 255–265. doi:10.1042/BJ20100774
- Gruber, C. N., Calis, J. J. A., Buta, S., Evrony, G., Martin, J. C., Uhl, S. A., et al. (2020). Complex Autoinflammatory Syndrome Unveils Fundamental Principles of JAK1 Kinase Transcriptional and Biochemical Function. *Immunity* 53 (3), 672–e11. doi:10.1016/j.immuni.2020.07.006
- Gschwend, D. A., Good, A. C., and Kuntz, I. D. (1996). Molecular Docking towards Drug Discovery. *J. Mol. Recognit.* 9 (2), 175–186. doi:10.1002/(sici)1099-1352(199603)9:2<175::aid-jmr260>3.0.co;2-d
- Haan, C., Behrmann, L., and Haan, S. (2010). Perspectives for the Use of Structural Information and Chemical Genetics to Develop Inhibitors of Janus Kinases. *J. Cel. Mol. Med.* 14, 504–527. doi:10.1111/j.1582-4934.2010.01018.x
- Haan, C., Rolvering, C., Raulf, F., Kapp, M., Drückes, P., Thoma, G., et al. (2011). Jak1 Has a Dominant Role over Jak3 in Signal Transduction through γ c-containing Cytokine Receptors. *Chem. Biol.* 18 (3), 314–323. doi:10.1016/j.chembiol.2011.01.012
- Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* 47 (7), 1750–1759. doi:10.1021/jm30644s
- Harikrishnan, L. S., Kamau, M. G., Wan, H., Inghrim, J. A., Zimmermann, K., Sang, X., et al. (2011). Pyrrolo[1,2-f]triazines as JAK2 Inhibitors: Achieving Potency and Selectivity for JAK2 over JAK3. *Bioorg. Med. Chem. Lett.* 21 (5), 1425–1428. doi:10.1016/j.bmcl.2011.01.022
- Harpur, A. G., Andres, A. C., Ziemiecki, A., Aston, R. R., and Wilks, A. F. (1992). JAK2, a Third Member of the JAK Family of Protein Tyrosine Kinases. *Oncogene* 7 (7), 1347–1353.
- Hevener, K. E., Zhao, W., Ball, D. M., Babaoglu, K., Qi, J., White, S. W., et al. (2009). Validation of Molecular Docking Programs for Virtual Screening against Dihydropterote Synthase. *J. Chem. Inf. Model.* 49 (2), 444–460. doi:10.1021/ci800293n
- Homeyer, N., and Gohlke, H. (2012). Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method. *Mol. Inform.* 31 (2), 114–122. doi:10.1002/minf.201100135
- Hornakova, T., Springuel, L., Devreux, J., Dusa, A., Constantinescu, S. N., Knoop, L., et al. (2011). Oncogenic JAK1 and JAK2-Activating Mutations Resistant to ATP-Competitive Inhibitors. *Haematologica* 96 (6), 845–853. doi:10.3324/haematol.2010.036350
- Hurley, C. A., Blair, W. S., Bull, R. J., Chang, C., Crackett, P. H., Deshmukh, G., et al. (2013). Novel Triazolo-Pyrrolopyridines as Inhibitors of Janus Kinase 1. *Bioorg. Med. Chem. Lett.* 23 (12), 3592–3598. doi:10.1016/j.bmcl.2013.04.018
- Ioannidis, S., Lamb, M. L., Wang, T., Almeida, L., Block, M. H., Davies, A. M., et al. (2011). Discovery of 5-Chloro-N2-[(1s)-1-(5-Fluoropyrimidin-2-Yl)ethyl]-N4-(5-Methyl-1h-Pyrazol-3-Yl)pyrimidine-2,4-Diamine (AZD1480) as a Novel Inhibitor of the Jak/Stat Pathway. *J. Med. Chem.* 54 (1), 262–276. doi:10.1021/jm1011319
- Irwin, J. J., and Shoichet, B. K. (2005). ZINC--a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* 45 (1), 177–182. doi:10.1021/ci049714+
- Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* 52 (7), 1757–1768. doi:10.1021/ci3001277
- Itteboina, R., Ballu, S., Sivan, S. K., and Manga, V. (2017). Molecular Modeling-Driven Approach for Identification of Janus Kinase 1 Inhibitors through 3D-QSAR, Docking and Molecular Dynamics Simulations. *J. Recept. Signal. Transduct. Res.* 37 (5), 453–469. doi:10.1080/10799893.2017.1328442
- Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., et al. (2004). A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins* 55 (2), 351–367. doi:10.1002/prot.10613
- Jaime-Figueroa, S., De Vicente, J., Hermann, J., Jahangir, A., Jin, S., Kuglstatler, A., et al. (2013). Discovery of a Series of Novel 5H-Pyrrolo[2,3-b]pyrazine-2-Phenyl Ethers, as Potent JAK3 Kinase Inhibitors. *Bioorg. Med. Chem. Lett.* 23 (9), 2522–2526. doi:10.1016/j.bmcl.2013.03.015
- Jain, A. N. (2003). Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* 46 (4), 499–511. doi:10.1021/jm020406h
- Jorgensen, W. L., and Duffy, E. M. (2002). Prediction of Drug Solubility from Structure. *Adv. Drug Deliv. Rev.* 54 (3), 355–366. doi:10.1016/s0169-409x(02)00008-x
- Keretsu, S., Bhujbal, S. P., and Cho, S. J. (2021b). Molecular Modeling Studies of Pyrrolo[2,3-d]pyrimidin-4-Amine Derivatives as JAK1 Inhibitors Based on 3D-QSAR, Molecular Docking, Molecular Dynamics (MD) and MM-PBSA Calculations. *J. Biomol. Struct. Dyn.* 39 (3), 753–765. doi:10.1080/07391102.2020.1714483
- Keretsu, S., Ghosh, S., and Cho, S. J. (2021a). Computer Aided Designing of Novel Pyrrolopyridine Derivatives as JAK1 Inhibitors. *Sci. Rep.* 11 (1), 23051–23112. doi:10.1038/s41598-021-0236410.1038/s41598-021-02364-2
- Kim, W. S., Kim, M. J., Kim, D. O., Byun, J. E., Huy, H., Song, H. Y., et al. (2017). Suppressor of Cytokine Signaling 2 Negatively Regulates NK Cell Differentiation by Inhibiting JAK2 Activity. *Sci. Rep.* 7 (1), 46153–46212. doi:10.1038/srep46153
- Kleppe, M., Spitzer, M. H., Li, S., Hill, C. E., Dong, L., Papalexi, E., et al. (2017). Jak1 Integrates Cytokine Sensing to Regulate Hematopoietic Stem Cell Function and Stress Hematopoiesis. *Cell stem cell* 21 (4), 489–e7. doi:10.1016/j.stem.2017.08.011
- Kulagowski, J. J., Blair, W., Bull, R. J., Chang, C., Deshmukh, G., Dyke, H. J., et al. (2012). Identification of Imidazo-Pyrrolopyridines as Novel and Potent JAK1 Inhibitors. *J. Med. Chem.* 55 (12), 5901–5921. doi:10.1021/jm300438j
- Kumar Nagarajan, S., Babu, S., Sohn, H., Devaraju, P., Madhavan, T., and Madhavan, T. (2018). Toward a Better Understanding of the Interaction between Somatostatin Receptor 2 and its Ligands: a Structural Characterization Study Using Molecular Dynamics and Conceptual Density Functional Theory. *J. Biomol. Struct. Dyn.* 37 (12), 3081–3102. doi:10.1080/07391102.2018.1508368
- Kumari, R., Kumar, R., and Lynn, A. (2014). g_mmpbsa-A GROMACS Tool for High-Throughput MM-PBSA Calculationsg_mmpbsa-Aa GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* 54 (7), 1951–1962. doi:10.1021/ci500020m
- Labadie, S., Barrett, K., Blair, W. S., Chang, C., Deshmukh, G., Eigenbrot, C., et al. (2013). Design and Evaluation of Novel 8-Oxo-Pyridopyrimidine Jak1/2 Inhibitors. *Bioorg. Med. Chem. Lett.* 23 (21), 5923–5930. doi:10.1016/j.bmcl.2013.08.082
- Labadie, S., Dragovich, P. S., Barrett, K., Blair, W. S., Bergeron, P., Chang, C., et al. (2012). Structure-based Discovery of C-2 Substituted Imidazo-Pyrrolopyridine JAK1 Inhibitors with Improved Selectivity over JAK2. *Bioorg. Med. Chem. Lett.* 22 (24), 7627–7633. doi:10.1016/j.bmcl.2012.10.008
- Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S., and Friesner, R. A. (2011). The VSGB 2.0 Model: a Next Generation Energy Model for High Resolution Protein Structure Modeling. *Proteins* 79 (10), 2794–2812. doi:10.1002/prot.23106
- Li, R. J., Wang, Y. L., Wang, Q. H., Wang, J., and Cheng, M. S. (2015). In Silico Design of Human IMPDH Inhibitors Using Pharmacophore Mapping and

- Molecular Docking Approaches. *Comput. Math. Methods Med.* 2015, 418767. doi:10.1155/2015/418767
- Lipinski, C. A. (2000). Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* 44 (1), 235–249. doi:10.1016/S1056-8719(00)00107-6
- Lobanov, M. Iu., Bogatyreva, N. S., and Galzitskaia, O. V. (2008). Radius of Gyration Is Indicator of Compactness of Protein Structure. *Mol. Biol. (Mosk)* 42 (4), 701–706. doi:10.1134/S0026893308040195
- Lu, A., Zhang, J., Yin, X., Luo, X., and Jiang, H. (2007). Farnesyltransferase Pharmacophore Model Derived from Diverse Classes of Inhibitors. *Bioorg. Med. Chem. Lett.* 17 (1), 243–249. doi:10.1016/j.bmcl.2006.09.055
- Lucet, I. S., Fantino, E., Styles, M., Bamert, R., Patel, O., Broughton, S. E., et al. (2006). The Structural Basis of Janus Kinase 2 Inhibition by a Potent and Specific Pan-Janus Kinase Inhibitor. *Blood* 107 (1), 176–183. doi:10.1182/blood-2005-06-2413
- Luo, H. J., Wang, J. Z., Huang, N. Y., Deng, W. Q., and Zou, K. (2014). Induced-fit Docking and Virtual Screening for 8-Hydroxy-3-Methoxy-5H-Pyrido [2,1-c] Pyrazin-5-One Derivatives as Inducible Nitric Oxide Synthase Inhibitors. *J. Chem. Pharm. Res.* 6 (3), 1187–1194.
- Lynch, S. M., DeVicente, J., Hermann, J. C., Jaime-Figueroa, S., Jin, S., Kuglstatler, A., et al. (2013). Strategic Use of Conformational Bias and Structure Based Design to Identify Potent JAK3 Inhibitors with Improved Selectivity against the JAK Family and the Kinome. *Bioorg. Med. Chem. Lett.* 23 (9), 2793–2800. doi:10.1016/j.bmcl.2013.02.012
- Lyne, P. D., Lamb, M. L., and Saeh, J. C. (2006). Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring. *J. Med. Chem.* 49 (16), 4805–4808. doi:10.1021/jm060522a
- Menet, C. J., Mammoliti, O., and López-Ramos, M. (2015). Progress toward JAK1-Selective Inhibitors. *Future Med. Chem.* 7 (2), 203–235. doi:10.4155/fmc.14.149
- Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* 55 (14), 6582–6594. doi:10.1021/jm300687e
- Nordqvist, C. (2011). *Protein JAK Makes Cancer Cells Contract, So They Can Squeeze Out of a Tumor*. Medical News Today.
- Pissot-Soldermann, C., Gerspacher, M., Furet, P., Gaul, C., Holzer, P., McCarthy, C., et al. (2010). Discovery and SAR of Potent, Orally Available 2,8-Diaryl-Quinoxalines as a New Class of JAK2 inhibitors. Discovery and SAR of Potent, Orally Available 2, 8-Diaryl-Quinoxalines as a New Class of JAK2 Inhibitors. *Bioorg. Med. Chem. Lett.* 20 (8), 2609–2613. doi:10.1016/j.bmcl.2010.02.056
- Raivola, J., Haikarainen, T., Abraham, B. G., and Silvennoinen, O. (2021). Janus Kinases in Leukemia. *Cancers* 13 (4), 800. doi:10.3390/cancers13040800
- Reddy, A. S., Pati, S. P., Kumar, P. P., Pradeep, H. N., and Sastry, G. N. (2007). Virtual Screening in Drug Discovery -- a Computational Perspective. *Curr. Protein Pept. Sci.* 8 (4), 329–351. doi:10.2174/138920307781369427
- Saddala, M. S., and Adi, P. J. (2018). Discovery of Small Molecules through Pharmacophore Modeling, Docking and Molecular Dynamics Simulation against Plasmodium Vivax Vivapain-3 (VP-3). *Heliyon* 4 (5), e00612. doi:10.1016/j.heliyon.2018.e00612
- Saharinen, P., and Silvennoinen, O. (2002). The Pseudokinase Domain Is Required for Suppression of Basal Activity of Jak2 and Jak3 Tyrosine Kinases and for Cytokine-Inducible Activation of Signal Transduction. *J. Biol. Chem.* 277 (49), 47954–47963. doi:10.1074/jbc.M205156200
- Sakkiah, S., Thangapandian, S., John, S., Kwon, Y. J., and Lee, K. W. (2010). 3D QSAR Pharmacophore Based Virtual Screening and Molecular Docking for Identification of Potential HSP90 Inhibitors. *Eur. J. Med. Chem.* 45 (6), 2132–2140. doi:10.1016/j.ejmech.2010.01.016
- Sakkiah, S., Krishnamoorthy, N., Gajendrarao, P., Thangapandian, S., Lee, Y. O., Kim, S. M., et al. (2009). Pharmacophore Mapping and Virtual Screening for SIRT1 Activators. *Bull. Korean Chem. Soc.* 30 (5), 1152–1156. doi:10.5012/bkcs.2009.30.5.1152
- Sakkiah, S., Thangapandian, S., John, S., and Lee, K. W. (2011). Identification of Critical Chemical Features for Aurora Kinase-B Inhibitors Using Hip-Hop, Virtual Screening and Molecular Docking. *J. Mol. Struct.* 985 (1), 14–26. doi:10.1016/j.molstruc.2010.08.050
- Sathe, R. Y., Kulkarni, S. A., Sella, R. N., and Madhavan, T. (2014). Computational Identification of JAK2 Inhibitors: a Combined Pharmacophore Mapping and Molecular Docking Approach. *Med. Chem. Res.* 24, 1449–1467. doi:10.1007/s00044-014-1223-6
- Schenkel, L. B., Huang, X., Cheng, A., Deak, H. L., Doherty, E., Emkey, R., et al. (2011). Discovery of Potent and Highly Selective Thienopyridine Janus Kinase 2 Inhibitors. *J. Med. Chem.* 54 (24), 8440–8450. doi:10.1021/jm200911r
- Schönberg, K., Rudolph, J., Vonnahme, M., Parampalli Jayanarayana, S., Cornez, I., Hejazi, M., et al. (2015). JAK Inhibition Impairs NK Cell Function in Myeloproliferative Neoplasms. *Cancer Res.* 75 (11), 2187–2199. doi:10.1158/0008-5472.CAN-14-3198
- Schwartz, D. M., Kanno, Y., Villarino, A., Ward, M., Gadina, M., and O'Shea, J. J. (2017). JAK Inhibition as a Therapeutic Strategy for Immune and Inflammatory Diseases. *Nat. Rev. Drug Discov.* 17 (12), 78–862. doi:10.1038/nrd.2017.20110.1038/nrd.2017.267
- Sohn, S. J., Barrett, K., Van Abbema, A., Chang, C., Kohli, P. B., Kanda, H., et al. (2013). A Restricted Role for TYK2 Catalytic Activity in Human Cytokine Responses Revealed by Novel TYK2-Selective Inhibitors. *J. Immunol.* 191 (5), 2205–2216. doi:10.4049/jimmunol.1202859
- Soth, M., Hermann, J. C., Yee, C., Alam, M., Barnett, J. W., Berry, P., et al. (2013). 3-Amido Pyrrolopyrazine JAK Kinase Inhibitors: Development of a JAK3 vs JAK1 Selective Inhibitor and Evaluation in Cellular and *In Vivo* Models. *J. Med. Chem.* 56 (1), 345–356. doi:10.1021/jm301646k
- Stahl, M., Guba, W., and Kansy, M. (2006). Integrating Molecular Design Resources within Modern Drug Discovery Research: the Roche Experience. *Drug Discov. Today* 11 (7–8), 326–333. doi:10.1016/j.drudis.2006.02.008
- Sterling, T., and Irwin, J. J. (2015). ZINC 15--Ligand Discovery for Everyone. *J. Chem. Inf. Model.* 55 (11), 2324–2337. doi:10.1021/acs.jcim.5b00559
- Taha, M. O., Atallah, N., Al-Bakri, A. G., Paradis-Bleau, C., Zalloum, H., Younis, K. S., et al. (2008). Discovery of New MurF Inhibitors via Pharmacophore Modeling and QSAR Analysis Followed by In-Silico Screening. *Bioorg. Med. Chem.* 16 (3), 1218–1235. doi:10.1016/j.bmc.2007.10.076
- Taldaev, A., Rudnev, V. R., Nikolsky, K. S., Kulikova, L. I., and Kaysheva, A. L. (2022). Molecular Modeling Insights into Upadacitinib Selectivity upon Binding to JAK Protein Family. *Pharmaceuticals* 15 (1), 30. doi:10.3390/ph15010030
- Thoma, G., Nuninger, F., Falchetto, R., Hermes, E., Tavares, G. A., VangrevelingheZerwes, E. H. G., et al. (2011). Identification of a Potent Janus Kinase 3 Inhibitor with High Selectivity within the Janus Kinase Family. *J. Med. Chem.* 54 (1), 284–288. doi:10.1021/jm101157q
- Vazquez, M. L., Kaila, N., Strohbach, J. W., Trzuppek, J. D., Brown, M. F., Flanagan, M. E., et al. (2018). Identification of N-[cis-3-[Methyl(7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino]cyclobutyl]propane-1-sulfonamide (PF-04965842): A Selective JAK1 Clinical Candidate for the Treatment of Autoimmune Diseases. *J. Med. Chem.* 61 (3), 1130–1152. doi:10.1021/acs.jmedchem.7b01598
- Vyas, V., Jain, A., Jain, A., and Gupta, A. (2008). Virtual Screening: a Fast Tool for Drug Design. *Sci. Pharm.* 76 (3), 333–360. doi:10.3797/sciparm.0803-03
- Wang, H., Aslanian, R., and Madison, V. S. (2008). Induced-fit Docking of Mometasone Furoate and Further Evidence for Glucocorticoid Receptor 17alpha Pocket Flexibility. *J. Mol. Graph. Model.* 27 (4), 512–521. doi:10.1016/j.jmgm.2008.09.002
- Wang, H. Y., Cao, Z. X., Li, L. L., Jiang, P. D., Zhao, Y. L., Luo, S. D., et al. (2008). Pharmacophore Modeling and Virtual Screening for Designing Potential PLK1 Inhibitors. *Bioorg. Med. Chem. Lett.* 18 (18), 4972–4977. doi:10.1016/j.bmcl.2008.08.033
- Wang, T., Duffy, J. P., Wang, J., Halas, S., Salituro, F. G., Pierce, A. C., et al. (2009). Janus Kinase 2 Inhibitors. Synthesis and Characterization of a Novel Polycyclic Azaindole. *J. Med. Chem.* 52 (24), 7938–7941. doi:10.1021/jm901383u
- Wen, W., Liang, W., Wu, J., Kowolik, C. M., Buettner, R., Scuto, A., et al. (2014). Targeting JAK1/STAT3 Signaling Suppresses Tumor Progression and Metastasis in a Peritoneal Model of Human Ovarian Cancer. *Mol. Cancer Ther.* 13 (12), 3037–3048. doi:10.1158/1535-7163.MCT-14-0077
- Williams, N. K., Bamert, R. S., Patel, O., Wang, C., Walden, P. M., Wilks, A. F., et al. (2009). Dissecting Specificity in the Janus Kinases: the Structures of JAK-specific Inhibitors Complexed to the JAK1 and JAK2 Protein Tyrosine Kinase Domains. *J. Mol. Biol.* 387 (1), 219–232. doi:10.1016/j.jmb.2009.01.041
- Xiang, Z., Zhao, Y., Mitaksov, V., Fremont, D. H., Kasai, Y., Molitoris, A., et al. (2008). Identification of Somatic JAK1 Mutations in Patients with Acute Myeloid Leukemia. *Blood* 111 (9), 4809–4812. doi:10.1182/blood-2007-05-090308

- Xie, H. Z., Li, L. L., Ren, J. X., Zou, J., Yang, L., Wei, Y. Q., et al. (2009). Pharmacophore Modeling Study Based on Known Spleen Tyrosine Kinase Inhibitors Together with Virtual Screening for Identifying Novel Inhibitors. *Bioorg. Med. Chem. Lett.* 19 (7), 1944–1949. doi:10.1016/j.bmcl.2009.02.049
- Xie, X., Wang, X., Shi, X., Zhang, Y., Laster, K. V., Liu, K., et al. (2021). Anwulignan Is a Novel JAK1 Inhibitor that Suppresses Non-small Cell Lung Cancer Growth. *J. Cel Mol Med* 25 (5), 2645–2654. doi:10.1111/jcmm.16289
- Yang, C. Y., Sun, H., Chen, J., Nikolovska-Coleska, Z., and Wang, S. (2009). Importance of Ligand Reorganization Free Energy in Protein-Ligand Binding-Affinity Prediction. *J. Am. Chem. Soc.* 131 (38), 13709–13721. doi:10.1021/ja9039373
- Yang, S. M., Malaviya, R., Wilson, L. J., Argentieri, R., Chen, X., Yang, C., et al. (2007). Simplified Staurosporine Analogs as Potent JAK3 Inhibitors. *Bioorg. Med. Chem. Lett.* 17 (2), 326–331. doi:10.1016/j.bmcl.2006.10.062
- Yeh, Y. T., Ou-Yang, F., Chen, I. F., Yang, S. F., Su, J. H., Hou, M. F., et al. (2007). Altered P-JAK1 Expression Is Associated with Estrogen Receptor Status in Breast Infiltrating Ductal Carcinoma. *Oncol. Rep.* 17 (1), 35–39. doi:10.3892/or.17.1.35
- Zak, M., Hurley, C. A., Ward, S. I., Bergeron, P., Barrett, K., Balazs, M., et al. (2013). Identification of C-2 Hydroxyethyl Imidazopyrrolopyridines as Potent JAK1 Inhibitors with Favorable Physicochemical Properties and High Selectivity over JAK2. *J. Med. Chem.* 56 (11), 4764–4785. doi:10.1021/jm4004895
- Zak, M., Mendonca, R., Balazs, M., Barrett, K., Bergeron, P., Blair, W. S., et al. (2012). Discovery and Optimization of C-2 Methyl Imidazopyrrolopyridines as Potent and Orally Bioavailable JAK1 Inhibitors with Selectivity over JAK2. *J. Med. Chem.* 55 (13), 6176–6193. doi:10.1021/jm300628c
- Zhan, C.-G., Nichols, J. A., and Dixon, D. A. (2003). Ionization Potential, Electron Affinity, Electronegativity, Hardness, and Electron Excitation Energy: Molecular Properties from Density Functional Theory Orbital Energies. *J. Phys. Chem. A* 107 (20), 4184–4195. doi:10.1021/jp0225774
- Zhao, X., Chen, M., Huang, B., Ji, H., and Yuan, M. (2011). Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) Studies on $\alpha(1A)$ -adrenergic Receptor Antagonists Based on Pharmacophore Molecular Alignment. *Int. J. Mol. Sci.* 12 (10), 7022–7037. doi:10.3390/ijms12107022
- Zhong, H., Tran, L. M., and Stang, J. L. (2009). Induced-fit Docking Studies of the Active and Inactive States of Protein Tyrosine Kinases. *J. Mol. Graph Model.* 28 (4), 336–346. doi:10.1016/j.jmgm.2009.08.012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Babu, Nagarajan, Sathish, Negi, Sohn and Madhavan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Drugsniffer: An Open Source Workflow for Virtually Screening Billions of Molecules for Binding Affinity to Protein Targets

Vishwesh Venkatraman^{1*†}, Thomas H. Colligan², George T. Lesica², Daniel R. Olson², Jeremiah Gaiser², Conner J. Copeland², Travis J. Wheeler^{2*†} and Amitava Roy^{2,3†}

¹Department of Chemistry, Norwegian University of Science and Technology, Trondheim, Norway, ²Department of Computer Science, University of Montana, Missoula, MT, United States, ³Rocky Mountain Laboratories, Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, United States

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Bruno Villoutreix,
Institut National de la Santé et de la
Recherche Médicale (INSERM), France
Aldo Oliveira,
Federal University of Santa Catarina,
Brazil

*Correspondence:

Vishwesh Venkatraman
vishwesh.venkatraman@ntnu.no
Travis J. Wheeler
travis.wheeler@umontana.edu

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 12 February 2022

Accepted: 04 April 2022

Published: 26 April 2022

Citation:

Venkatraman V, Colligan TH,
Lesica GT, Olson DR, Gaiser J,
Copeland CJ, Wheeler TJ and Roy A
(2022) Drugsniffer: An Open Source
Workflow for Virtually Screening Billions
of Molecules for Binding Affinity to
Protein Targets.
Front. Pharmacol. 13:874746.
doi: 10.3389/fphar.2022.874746

The SARS-CoV2 pandemic has highlighted the importance of efficient and effective methods for identification of therapeutic drugs, and in particular has laid bare the need for methods that allow exploration of the full diversity of synthesizable small molecules. While classical high-throughput screening methods may consider up to millions of molecules, virtual screening methods hold the promise of enabling appraisal of billions of candidate molecules, thus expanding the search space while concurrently reducing costs and speeding discovery. Here, we describe a new screening pipeline, called *drugsniffer*, that is capable of rapidly exploring drug candidates from a library of billions of molecules, and is designed to support distributed computation on cluster and cloud resources. As an example of performance, our pipeline required ~40,000 total compute hours to screen for potential drugs targeting three SARS-CoV2 proteins among a library of ~3.7 billion candidate molecules.

Keywords: virtual screening, machine learning, computer aided drug design, de novo design, SARS-CoV-2, protein-ligand docking

1 INTRODUCTION

The war against viruses is largely fought using vaccines and therapeutic drugs. As of December 2021, there are 55 FDA-approved vaccines against 19 human viruses (FDA, 2021), while only three viruses are targeted by approved antiviral drugs (FDA, 2020b). This disparity is particularly visible in the context of the ongoing SARS-CoV2 pandemic, in which vaccines were produced at a remarkable speed and with excellent effectiveness (FDA, 2020a; Wouters et al., 2021), while effective antiviral agents (Mahase, 2021; Jayk Bernal et al., 2022) only arrived 2 years into the pandemic, and with very limited availability. Despite vaccine success, there remains a vital need for development of effective antiviral drugs due to a combination of vaccine hesitancy, incomplete vaccine availability, breakthrough infection risk, and the continued emergence of viral variants (Kaplan and Milstein, 2021). Beyond SARS-CoV2, the cost and limited exploratory scope of current drug discovery pipelines will hamper efforts to quickly respond to future pandemic needs, and are an obstacle to development of antiviral drugs for viruses primarily afflicting relatively poor populations (Adamson et al., 2021).

TABLE 1 | Several open access software tools for virtual screening. In a number of the tools, such as dockECR and VirtualFlow, multiple docking programs are used to predict scores between a single target or multiple targets (merging and shrinking approach) and a library of compounds. The AMIDE software carries out large-scale chemical ligand docking over a large dataset of proteins with the aim of identifying potential side effects of new drugs. iDrug, Pharmit (for structure-based pharmacophore modeling), iStar, e-LEA3D, USR-VS (3D shape-based similarity), MTiOpenScreen and ChemicalToolbox are web-based platforms for computer-aided drug design. ChemicalToolbox allows for integration with other tools and workflows (molecular dynamics) that are part of the Galaxy software framework (<https://galaxyproject.org/>). e-LEA3D uses a *de novo* drug design strategy in which fragments or combination of fragments that fit a QSAR model or the binding site of a protein are identified. * iDrug uses a pocket structure to define the pharmacophore descriptors needed for LBVS. However, they do not explicitly calculate the interaction between a ligand and the pocket, such as docking. In our opinion, they are marginally SBVS.

| Software | LBVS | SBVS | ADMET |
|--|------|------|-------|
| dockECR Ochoa et al. (2021) | X | ✓ | X |
| MolAr Maia et al. (2020) | X | ✓ | X |
| iDrug Wang et al. (2014) | ✓ | ✓* | X |
| ChemicalToolbox Bray et al. (2020) | X | ✓ | ✓ |
| VirtualFlow Gorgulla et al. (2020), Gorgulla et al. (2021) | X | ✓ | ✓ |
| AMIDE Darne et al. (2021) | X | ✓ | X |
| VSPipe Álvarez-Carretero et al. (2018) | X | ✓ | X |
| DockBlaster Irwin et al. (2009) | X | ✓ | X |
| e-LEA3D Douguet (2010) | X | ✓ | X |
| Pharmit Sunseri and Koes (2016) | ✓ | X | X |
| iStar Li et al. (2014) | X | ✓ | X |
| USR-VS Li et al. (2016) | ✓ | X | X |
| MTiOpenScreen Labbé et al. (2015) | X | ✓ | X |
| DrugSniffer | ✓ | ✓ | ✓ |

Modern drug development efforts rely on high-throughput screening (HTS) analysis, which involves automated physical evaluation of activity across a library of thousands to millions of candidate small-molecule drugs (Berdigaliyev and Aljofan, 2020). HTS can be complemented by computer-aided drug design (CADD) and virtual screening (VS), in which interactions between small-molecules and a targets are estimated using computational models. In particular, computational analysis holds the promise of enabling expansion of the number of considered molecules from millions to billions.

VS strategies are traditionally divided into two categories: ligand-based (LBVS) and structure-based (SBVS) methods. In LBVS methods, a known active ligand is used as the basis for a search for chemically and structurally similar molecules (Ripphausen et al., 2011), with no consideration of the target protein. In SBVS approaches, small molecules are computationally docked into target binding sites to estimate their activities (Maia et al., 2020); this approach depends on availability of structural information, and is computationally intensive. The two methods can be integrated either by combining results (Wilson and Lill, 2011; Wang et al., 2020), or by using LBVS methods to quickly establish a set of candidates subjected to subsequent SBVS docking analysis (Drwal and Griffith, 2013).

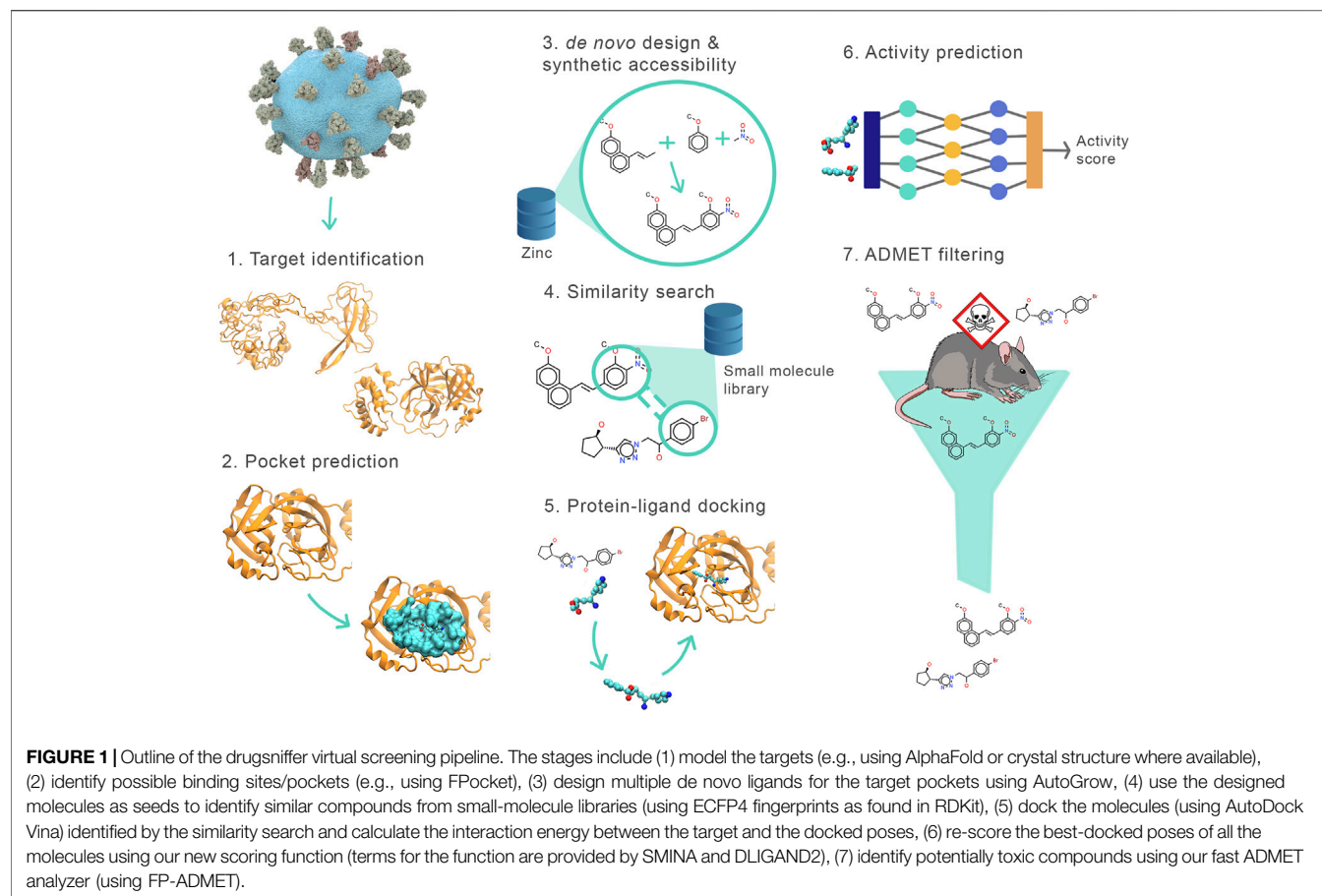
Table 1 provides a list of various open access VS tools. For large scale virtual screening of compound libraries, software pipelines such as VSPipe Álvarez-Carretero et al. (2018), VirtualFlow Gorgulla et al. (2020, 2021), AMIDE Darne et al. (2021) have been used. Many of these approaches make use of SBVS and facilitate the use of a variety of docking Bender et al. (2021) programs with significant emphasis on scaling the calculations. Recent GPU acceleration of docking (Santos-Martins et al., 2021) has

improved throughput, but resource requirements are still exceedingly high. For example, an effort to performing one billion docking assays was reported to require 664K GPU hours and 4.64M core hours for a single VS analysis (Acharya et al., 2020). With the aim of automating hit-selection protocols and minimizing human intervention, artificial intelligence-driven VS. pipeline have also been introduced Gentile et al. (2020), Gentile et al. (2021); Yaacoub et al. (2021).

Herein, we describe our development and release of an open source, massively-scalable LBVS-filtered SBVS pipeline, called *drugsniffer*, that is designed to achieve the goal of virtually screening bioactive drugs from datasets of billions of probably-synthesizable small molecules in a much-reduced time budget. *Drugsniffer* is easy to install and manages the distribution of computation across cluster or cloud resources. It reduces the computational burden to 10s of thousands of compute hours for search across a library of billions of candidate molecules, and provides a framework in which future methodological advances can be incorporated and evaluated. Using an early iteration of *drugsniffer*, we assessed ~3.7B molecules for binding potential against 3 SARS-CoV2 proteins (22 binding pockets), with total computational investment of ~40 K compute hours. The results of our analysis were accepted as a finalist in Joint European Disruptive Initiative (JEDI) “billion molecules against COVID19” challenge (Le et al., 2021).

2 METHODS

Drugsniffer consists of the following phases (see **Figure 1**): 1) select the protein target and determine its structure, 2) identify binding pockets, 3) design *de novo* ligands for each pocket, 4) use these as seeds to identify similar molecules in a large composite



database of synthesizable small molecules, 5) perform *in silico* docking assays on these candidates, 6) apply a new neural network model to predict and rank binding affinity based on features of the docked poses, 7) identify potential toxicity of compounds using a custom ADMET filter. In this section, we describe these stages in detail, then discuss our application of an early implementation of the pipeline to the JEDI COVID19 Grand Challenge.

2.1 Selecting Target Proteins and Determining Structure

The first step in the drug screening process is the selection of the target protein—the user must provide a structural model for the selected protein. *Drugsniffer* is agnostic about the source of the structural model, and will work with experimentally-validated or computationally-predicted structures. Though protein structures may be retrieved from a variety of sources, we have had good experiences with ChimeraX (Pettersen et al., 2021), which, for example, supports retrieval of structures from the Protein Data Bank (Berman et al., 2000) or prediction using AlphaFold2 (Jumper et al., 2021). AlphaFold2 achieved remarkable accuracy in the CASP14 competition; for example, in 92.5% of predictions, all side chain atoms are predicted with error ≤ 5 Å (Pereira et al., 2021). This accuracy is unprecedented for

computational models, and these models may provide insight into the diversity of conformations that extend beyond the single conformer of a crystal-based structure. Even so, a substantial fraction of the predicted atoms, primarily from the flexible parts of the proteins, may not be modeled correctly by AlphaFold2. We encourage users to evaluate the overall (IDDT) and residue-specific (pLDDT) scores to evaluate the predicted accuracy of the overall and pocket regions of an AlphaFold2 model.

2.2 Identifying Pockets

In addition to a target protein structure, *drugsniffer* must be provided with at least one pocket descriptor, as well as a preferred pocket box size. The most reliable way of detecting a ligand-binding pocket is a user's prior knowledge about the binding pocket from experience, experimental evidence, and literature search. Computational identification of a pocket-like region is challenging and an active area of research (Zhao et al., 2020). The *drugsniffer* pipeline includes a copy of the cavity detection software Fpocket (Le Guilloux et al., 2009) only because it is a stand-alone free program. We encourage users to use multiple pocket search algorithms, such as FTMAP Kozakov et al. (2015), POCASA Yu et al. (2010), and molecular dynamics simulations, and use their judgment to define a pocket-like region in the protein. The current implementation of the *drugsniffer* pipeline produces an FPOCKET output that includes all predicted

TABLE 2 | The small molecule databases searched as part of the VS protocol.

| Database | Number of ligands |
|-----------|-------------------|
| Sweetlead | ≈4,000 |
| Drugbank | ≈10,000 |
| MOLPROT | ≈7,600,000 |
| PUBCHEM | ≈103,000,000 |
| ZINC15 | ≈417,000,000 |
| GDB | ≈1,003,000,000 |
| SAVI | ≈1,009,000,000 |
| ENAMINE | ≈1,200,000,000 |
| Total | ≈3,700,000,000 |

<https://simtk.org/projects/sweetlead>

<https://www.drugbank.ca/releases/latest>

<https://www.molport.com/shop/libraries-collections>

<http://ftp.ncbi.nlm.nih.gov/pubchem/Compound/>

<http://files.docking.org/catalogs/>

<http://gdb.unibe.ch/downloads/>

https://cactus.nci.nih.gov/download/savi_download/

<https://enamine.net/library-synthesis/real-compounds/real-database>

pockets; the user is tasked with manually reviewing these and identifying the subset for which the downstream drug discovery stages should be performed, e.g., using ChimeraX or PyMol (Oliveira et al., 2014). Pocket descriptors identified outside of the *drugsniffer* pipeline may be provided as an alternative or supplementary source of predicted pockets. Box size must be determined for each pocket; we recommend basing this on the scheme proposed by (Feinstein and Brylinski, 2015).

2.3 De Novo Ligand Design

Following manual pocket identification, *drugsniffer* accepts as input the set of targeted pockets, and proceeds in an automatic fashion through the remaining stages. In the first stage, a large number of candidate ligand molecules are designed from scratch using the software AutoGrow4 (Spiegel and Durrant, 2020), which employs a genetic algorithm to evolve ligands from building blocks obtained from the ZINC library (Sterling and Irwin, 2015). AutoGrow4 utilizes a diversity score that acts as a secondary fitness metric and is used to select seed compounds that are structurally unique from previous generations. The molecules are subsequently docked into the pockets of the specified target protein using QuickVina (Alhossary et al., 2015) which is a faster version of Autodock Vina. Docked results are ranked based on the Vina docking score of the top ranking pose. A Lipinski RO5 filter is used to exclude candidate structures that do not satisfy drug-like criteria. The NIH filter (Jadhav et al., 2010) is also included to screen against compounds containing undesirable functional groups. AutoGrow4 performs *in silico* chemical reactions (Durrant and McCammon, 2012) derived from a set of robust organic reactions (Hartenfeller et al., 2011) to generate new child compounds from a parent molecule. These reaction-based structural transformations are used to increase the likelihood of the designed molecules being synthetically accessible. However, a drawback of using pre-defined reaction schemes is that they may match reaction handles and fail to consider the presence of competing functionalities that can compromise the reaction outcome (Ghiandoni et al., 2020; Meyers et al., 2021). By default, the pipeline runs AutoGrow4 for 10 generations, and

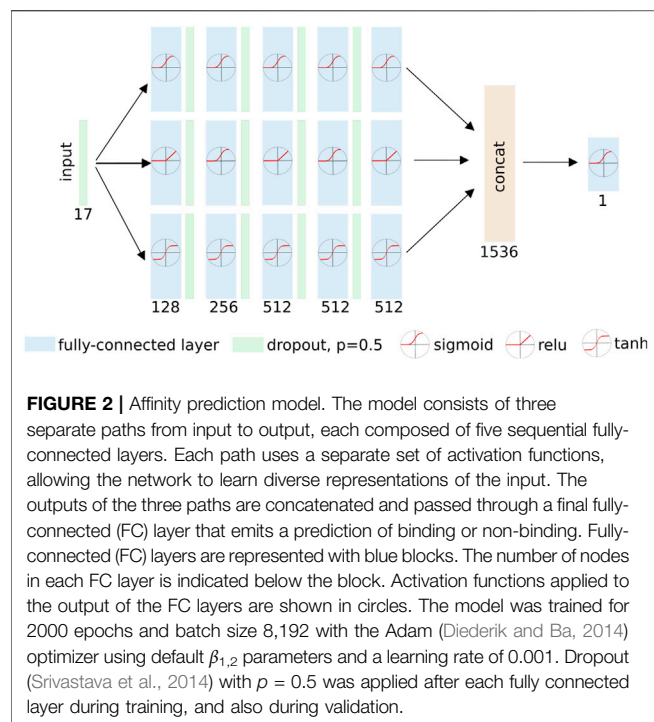
captures 150 *de novo* molecules from each of the final three generations. *Drugsniffer* can optionally forgo this AutoGrow4 step, and instead accept a collection of ligands provided by the user—these may be sourced from some prior *de novo* computation, or from a collection of co-crystallized protein-ligand complexes.

2.4 Molecular Similarity Search

The motivation for employing *de novo* ligand design is to produce drug-like compounds that can mimic known inhibitors or potentially active ligands with a diversity of chemical structures. While the molecules produced by AutoGrow4 are predicted to be synthesizable, factors such as establishing synthetic routes, material procurement, costs and time involved are difficult to predict. We therefore sought to build on the value of these designed molecules through an LBVS search strategy in which the *de novo* molecules serve as seeds in a search for similar compounds within a massive library of molecules.

We compiled a collection of molecules from various small-molecule libraries, with the aim of capturing a large diversity of molecules that either already exist, or are likely-synthesizable and can be made to order (see Table 2). The Enamine collection includes more than 1 billion compounds that comply with Lipinski's rule of five (RO5) criteria and are expected to be realized in 1–3 synthesis steps. The Synthetically Accessible Virtual Inventory (SAVI) (Patel et al., 2020) contains over 1 billion reliably-synthesizable compounds generated through expert-system rules. GDB-13 (Blum and Reymond, 2009) also contains over 1 billion compounds (containing up to 13 atoms of C, N, O, S, and Cl = , generated according to chemical stability and synthetic feasibility rules. PubChem (Kim et al., 2020), ZINC (Sterling and Irwin, 2015), and Molport are curated collections of commercially-available molecules. SweetLead (Novick et al., 2013) and DrugBank (Wishart et al., 2017) contain drugs that are in use or in clinical trials, and may therefore facilitate repurposing of established drugs. We removed molecules containing salts, because downstream docking methods fail in the face the apparent disjoint molecules. The full de-duplicated collection contains ~3.7 billion unique molecules.

To identify library-sourced compounds similar to the *de novo* seeds produced by AutoGrow4, 1024-bit ECFP4 fingerprints (O'Boyle and Sayle, 2016) are computed for all ~3.7 billion library compounds. The ECFP4 fingerprint is a 1024-element binary vector that encodes structural and chemical features. Though a multitude of fingerprint strategies exist, ECFP4 has been reported to effectively rank diverse structures by similarity (O'Boyle and Sayle, 2016). Future releases of *drugsniffer* will enable selection of other fingerprints, or related similarity measures. ECFP4 fingerprints are computed using RDKit (<https://www.rdkit.org>), then stored as a sequence of 1,024 bit vectors, so that a library of 3.7 billion molecules is represented by a ~475 Gbyte fingerprint database. Fingerprints are similarly computed for all seeds. A measure of similarity between two molecules is computed by comparing the 1024-bit fingerprints of each molecule, using the Tanimoto coefficient (aka Jaccard index): the ratio of the intersecting set (number of bits set to one in both fingerprints) to the union set (number of bits set to one in at least one of the two fingerprints) (Bajusz et al., 2015).



Similar (“neighbor”) molecules are identified by computing the Tanimoto coefficient for each seed against each molecule in the fingerprint database using SIMD vectorized bit-level comparison over 1,024 representative bits per molecule. By default, neighbors with Tanimoto similarity > 0.5 to at least one seed are captured for later docking estimates. This threshold is selected based on experience, with the aim of balancing stringency (reducing the computational burden of later stages) with permissiveness (expanding the pool of candidates that reach the next stage); it can be altered at run time.

2.5 Protein-Ligand Docking

For the seed-neighbor molecules identified by the similarity search, initial 3D coordinates are generated from the SMILES representations using OpenBabel (O’Boyle et al., 2011a). Diverse low-energy conformers for the molecules are generated using the Confab (O’Boyle et al., 2011b), then the lowest energy conformation is retained. These optimized structures of neighbors are docked into their respective targets using AutoDock Vina (Trott and Olson, 2010). The number of docking poses produced and the exhaustiveness parameter for the search for each ligand are parameterized by the user; the default values are 9 and 4, respectively.

2.6 Re-Scoring Docked Ligands, to Estimate Binding Affinity

AutoDock Vina reports a set of molecular poses within the pocket, along with a value representing a prediction of the quality of each docked pose. Because this prediction is only a loose estimate of actual binding affinity, a variety of post hoc re-scoring methods have been

devised [e.g., see (Koes et al., 2013a; Chen et al., 2019; McNutt et al., 2021)]. *Drugsniffer* can report either the Autodock Vina score, the SMINA (Koes et al., 2013b) rescoring value, or the result of a new neural network re-scoring strategy that we have produced for this workflow (*dock2bind*, which is the default). *Drugsniffer* supports retraining of this model with domain-specific binding affinity data, and also will accept an alternate re-scoring function that is injected by the user into the *drugsniffer* workflow by providing a Docker container meeting a simple documented API.

For each docked pose, our *dock2bind* receives 16 pose descriptors calculated by SMINA, along with the DFIRE estimate of protein–ligand potential (Chen et al., 2019), and computes a new affinity estimate for the pose. This estimate is a value between 0 and 1 and can be thought of as the model’s confidence that the molecule binds to the pocket, constrained by the specific pose. See **Figure 2** for model details. Ligand-protein pairs were taken from the DUD-E benchmark (Mysinger et al., 2012) and LIT-PCBA (Tran-Nguyen et al., 2020). To train the model, docked poses were generated for $\sim 14,000$ ligand-protein pairs from the DUD-E dataset, along with $\sim 800,000$ decoy ZINC-sourced compounds docked to the same protein partners. These were supplemented with an additional $\sim 4,000$ ligand-protein complexes from LIT-PCBA, and $\sim 121,000$ decoys docked to the same proteins. The active:decoy ratio is intended to reflect the large actual classification imbalance (most molecules are inactive for any specific target). For each target, 9 docked poses were produced, and the pose with the best SMINA score was provided to the *dock2bind* model for training.

2.7 ADMET Analysis

Drugsniffer includes a suite of models to predict properties tied to bioavailability and safety. Owing to their ease of computation, molecular fingerprints have been frequently used to predict these properties (Kim and Nam, 2017; Ai et al., 2018; Yang et al., 2019). Fingerprint-based classification models were trained on experimental data available [see (Venkatraman, 2021)] for solubility in dimethyl sulfoxide (DMSO), blood brain barrier permeability, human intestinal absorption (HIA), AMES mutagenicity, HERG cardiotoxicity, drug induced liver injury (DILI), Cytochrome p450 interaction (CYP3A4 and CYP2C9 isoforms), metabolic stability and acute LD₅₀ toxicity based on the criteria defined by the Environmental Protection Agency (EPA). For each property, various fingerprints (Hinselmann et al., 2011) (substructure and extended/functional connectivity fingerprints) were evaluated for their discriminant ability and the fingerprint model [using random forests (Breiman, 2001)] yielding the best balanced accuracy (Brodersen et al., 2010; Venkatraman, 2021) was retained. The *drugsniffer* pipeline applies these models to the list of candidates produced by previous stages, and appends the resultant vector of properties to the affinity prediction results. The models can be accessed at <https://gitlab.com/vishsoft/fpadmet>.

2.8 Software and Data

Drugsniffer is implemented as a Nextflow workflow (Di Tommaso et al., 2017) that orchestrates the activity of a curated set of open source tools, and supports analysis in cluster (SLURM) and cloud (AWS) environments. **Table 3**

TABLE 3 | Software used in the VS pipeline.

| Software | Comments |
|---------------------------|---|
| RDKit | Routines for ECFP4 fingerprint generation |
| Chemistry Development Kit | logP estimation routines |
| OpenBabel | interconvert chemical file formats |
| MGLTools | interconvert chemical file formats |
| AutoDock Vina | Protein-ligand docking |
| DLigand2 | statistical potential term for protein-ligand binding affinity prediction |
| SMINA | scoring terms for protein-ligand binding affinity prediction |
| AUTOGROW4 | <i>de novo</i> ligand design using docking |
| FP-ADMET | Prediction of ADMET properties |

<https://www.rdkit.org>

<https://cdk.github.io/>

http://openbabel.org/wiki/Main_Page

<https://ccsb.scripps.edu/mgltools/downloads/>

<https://github.com/ccsb-scripps/AutoDock-Vina>

<https://github.com/sysu-yanglab/DLIGAND2>

<https://github.com/mwojcikowski/smina>

<https://git.durrantlab.pitt.edu/jdurrant/autogrow4>

<https://gitlab.com/vishsoft/fpadmet>

lists the different software tools that are used in the workflow. The workflow depends on a collection of Docker containers and runner scripts wrapping each of our own tools as well as the external open source tools included in the analysis pipeline. This organizing principle makes it possible for the user to configure and run *drugsniffer* without concern for dependencies. Docker container files, NextFlow scripts, and tool code are all available via GitHub (<https://github.com/TravisWheelerLab/drug-sniffer>). Versioned Docker container images are published in the GitHub container registry, and the full library of ~3.7 billion molecules (with pre-computed fingerprints) is housed in a persistent OSF repository (Soderberg, 2018) and. Instructions for download and use are found at <http://drugsniffer.org>.

2.9 Application of *Drugsniffer* to JEDI COVID19 Grand Challenge

In May 2020, the Joint European Disruptive Initiative (JEDI) launched a “Grand Challenge” competition intended to motivate development of methods capable of searching a library of billions of molecules for those with potentially good binding affinity for target SARS CoV2 proteins. We developed *drugsniffer* to meet these goals, and submitted candidate molecules identified with an early version of the pipeline. Our submissions have reached the finalist stage, and are currently under experimental review. Here, we describe how our pipeline was used to prepare our submission, and document the differences between the version of the pipeline used for our JEDI submission and its current released form.

To begin, we selected three target proteins: RNA dependent RNA polymerase (Non-structural Protein 12, aka NSP12), 3C like protease (3CLPro), and Nucleocapsid protein (N). At the time of the analysis, no whole-protein experimental structure was available for any of the targets and AlphaFold2 was not yet released. We therefore downloaded models created by I-TASSER (Yang et al., 2015), and added hydrogen atoms with CHARMM (Brooks et al., 2009).

Candidate binding pockets for the three selected targets were identified using a combination of literature search and results

from the tools FTMAP (Kozakov et al., 2015) and POCASA (Yu et al., 2010) (*drugsniffer* incorporates Fpocket in lieu of these, because its license allows redistribution). Seven pocket-like regions were identified: 2 each for N and 3CLpro, and 3 for NSP12. Some of the pocket-like regions were too large to be occupied by a typical-sized ligand. Consequently, the larger pocket-like regions were subdivided into smaller pockets. A total of 22 pockets were finalized as targets: 8 each for N and NSP12 and 6 for 3CLPro. We searched the literature to identify any glycosylation sites for the three selected targets and did not find any. We also used N-GlyDe (Pitti et al., 2019) to identify any potential sites for N-linked glycans. Our predicted glycosylation sites are residue 269 of N and residues 767 and 911 of NSP12. As none of the glycosylation sites were near any of the predicted binding pockets, we did not consider glycosylation for our later docking exercises.

The next several pipeline stages were run as in the current release of the pipeline, including *de novo* ligand design, molecular similarity search, and protein-ligand docking. AutoGrow4 was run for 25 generations, over five independent runs. In total, 31,962 seed molecules were identified by AutoGrow4 (12,227 for nsp12 pockets, 14,334 for N pockets, and 5,401 for 3CLPro pockets). Molecular similarity search identified ~97,000 library compounds with Tanimoto similarity >0.6 to some seed, and another ~955,000 with Tanimoto similarities of 0.5–0.6. Among the 97,000 closest neighbours: ~43,000 were identified for nsp12, ~34,000 for N, ~20,000 for 3CLPro. For each pocket, all seed neighbor molecules were docked (AutoDock Vina) to the pocket, and poses were re-scored using dock2bind, using the top re-scored pose for each molecule as its predicted affinity. The top-scoring 30,000 candidates (10,000 per protein) were analyzed for ADMET and predicted synthetic complexity [SCSCORE (Coley et al., 2018)] of the target molecule. Candidates with no ADMET contraindications, and with an expected number of synthesis steps ≤5 were submitted to the JEDI challenge; 18 compounds passed JEDI criteria for the final evaluation, and are being synthesized and evaluated.

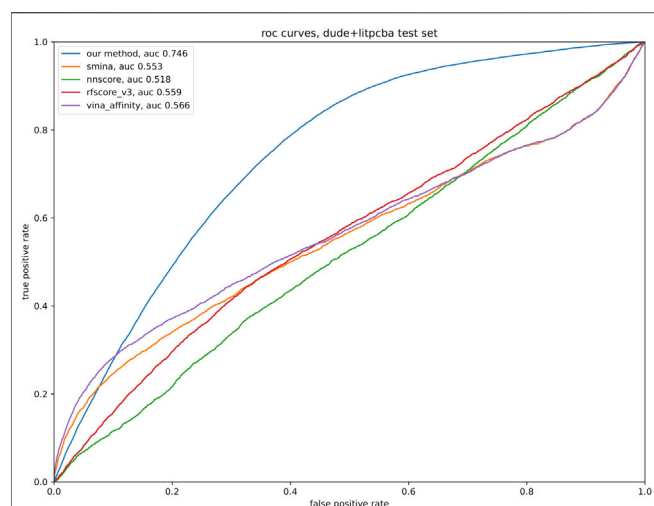


FIGURE 3 | Test data consisting of 3,900 ligand-protein pairs and 213,000 decoy-protein pairs was analyzed with the tools listed in the legend, with the relevant tool producing a binding affinity estimate for each pair. Default parameters were used for all tools; our model was trained as described in the text. A ROC curve was produced for each tool, based on the sorted list of predicted affinity.

3 RESULTS

Here, we have described the stages and availability of a new pipeline for exploring a pre-built library of billions of likely-synthesizable molecules for a small set of candidate molecules that are expected to show good binding affinity to a user-provided protein structure and pocket descriptor. As a proof of principle, we used a variant of this pipeline to identify drug candidates from our library of ~3.7 billion molecules, targeting 22 pockets in 3 proteins associated with SARS-CoV2, resulting in a list of ~30,000 candidate compounds. This collection was submitted for analysis to the JEDI “Grand Challenge,” and were advanced to “finalist” status; experimental review of a subset of these molecules is underway. Compute time for the total search for candidate molecules for all 22 pockets was ~40,000 CPU hours. By distributing workload across a cluster, the analysis required only a few days. In addition to these run time results, we explored the efficacy of our custom docking re-scoring model, as well as the outcomes of ADMET and synthesizability analysis.

3.1 Performance of the Deep Learning Re-Scoring Model

To quantitatively evaluate our model, a test set was developed from DUD-E and LIT-PCBA, consisting of complexes involving proteins not found in the training set. A total of ~3000 DUD-E ligand-protein pairs, ~186,000 decoys for DUD-E proteins, ~900 LIT-PCBA ligand-protein pairs, and ~27,000 decoys for LIT-PCBA. No hyperparameter tuning was performed on any of the models so a validation set was unnecessary. To test the

efficacy of our method of ranking potential binders, we compared our method to a variety of open-source implementations of affinity-predicting methods, including Vina’s default method, the SMINA default score, and the NNScore and RF-score (version 3) from the Open Drug Discovery Toolkit (Wójcikowski et al., 2015) (ODDT). **Figure 3** shows the performance of the model architecture trained on different subsets of the data.

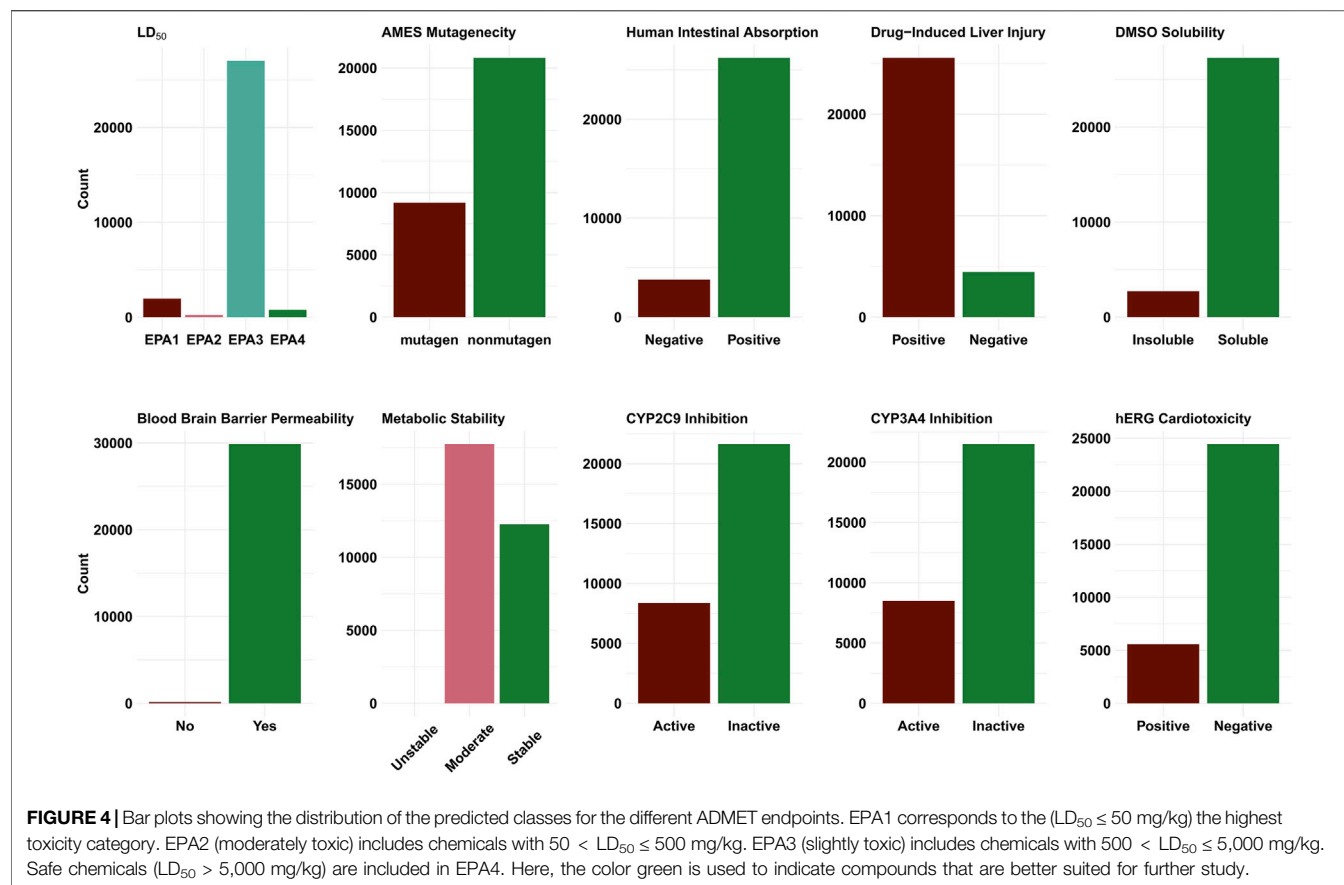
3.2 ADMET and Synthesizability Analysis

Figure 4 shows the distribution of the ADMET properties for the ~30,000 compounds that were submitted to the JEDI competition. For the most part, the shortlisted compounds were predicted to have favourable ADMET properties. Our ML model for DILI (Venkatraman, 2021) predicts a majority (~85%) of the compounds to be hepatotoxic. The DILI model however only provides a binary (yes/no) prediction and does not indicate the level of the underlying DILI severity. A strict application of the models (i.e., selecting only those compounds that are deemed to be favourable across all calculated properties) yielded a set of 1,635 compounds. Many ADMET properties are affected by the dosage, route and frequency. For better assessment of ADMET, knowledge of the underlying mechanisms is required. Given that it is far from trivial to prioritize one property over the other (leading to varying application of the filter), we have used the model predictions as a guide rather than a filter. With respect to synthesizability, ~79% of molecules identified by the pipeline were predicted to require three or fewer predicted reaction steps.

4 DISCUSSION

Virtual screening has seen a recent rise in prominence, supported by improved computational methods across the range of analyses represented in the *drugsniffer* pipeline. The ongoing pandemic has highlighted the need for improved speed and increased exploratory scope of virtual screening methods. Relatedly, the development of low-cost virtual screening methods holds the promise of improving opportunities for development of drugs targeting diseases prevalent in low-income regions, for which economic incentives discourage expensive high-throughput screening assays. We developed *drugsniffer* as a preliminary tool to meet this need, exploring billions of candidate molecules for a target protein pocket in a few thousand compute hours—relatively modest resources available to most HPC infrastructures. Even with its development, each of the stages of the *drugsniffer* pipeline will be well-served by methodological advances. We highlight a few such areas of opportunity here, and observe that *drugsniffer* can easily adapt to incorporate advances along these lines, due to its modular nature.

With the development and release of AlphaFold2 and similar structure prediction methods, structure prediction is perhaps no longer a general bottleneck in the drug discovery problem, though some protein types still suffer from relatively uncertain predictions. Pocket identification remains a



challenge, and most current techniques can detect pockets only with ~60% accuracy (Zhao et al., 2020). Advances in this field will reduce the dependency on expert manual analysis of structures and pockets.

4.1 Future Advances

Drugsniffer will also be improved by development of advances in *de novo* molecule production (where limitations include wall clock run time and molecule synthesizability and utility), molecular similarity search (where current molecule-centric approaches fail to account for pocket-specific interaction profiles), and docking-based affinity prediction (where re-scoring methods produce only modestly enrichment for actives vs. decoys (see **Figure 3**) and may not generalize well to structures that are not represented in the training set). *Drugsniffer* will be expanded by including molecular dynamics simulations to consider multiple conformations of a pocket region and refining binding energy estimation of shortlisted ligands. It should be emphasized that the scope of the *drugsniffer* pipeline is to identify possible ligands with high enrichment factors. Users should carry out such MD or QM studies on the possible ligands predicted by the *drugsniffer* for a more accurate prediction of binding affinity or to investigate the effect of protonation states in binding. Due to their approximate nature, docking forcefields are insensitive to such details.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <http://drugsniffer.org>.

AUTHOR CONTRIBUTIONS

VV, AR, and TW designed the pipeline, along with the study of application to SARS-CoV2 proteins; they also supervised efforts of others and collectively wrote the first draft of the manuscript. AR and VV developed approaches for identifying proteins, structures, pockets, and *de novo* seeds; they also collected the molecule library. TW, VV, and JG developed methods for molecule fingerprinting and rapid neighbour identification, and applied to SARS-Cov2 data. CC and GL incorporated docking into the pipeline. TC and DO developed the machine learning model for docking re-scoring. GL developed the NextFlow workflow, and all associated Docker images. All authors contributed to the manuscript edits.

FUNDING

VV acknowledges financial support from the Research Council of Norway (Grant No. 262152). AR acknowledges funding from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and

Human Services under BCBP Support Services Contract HHSN316201300006W/HHSN27200002 to MSC, Inc. The remainder of co-authors acknowledge support from the National Institute of General Medical Sciences (NIH NIGMS, R01GM132600) and the Genomic Science program (GSP) of the Office of Biological and Environmental Research in the Department of Energy (DE-SC0021216).

REFERENCES

- Acharya, A., Agarwal, R., Baker, M. B., Baudry, J., Bhowmik, D., Boehm, S., et al. (2020). Supercomputer-based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J. Chem. Inf. Model.* 60, 5832–5852. doi:10.1021/acs.jcim.0c01010.26434/chemrxiv.12725465
- Adamson, C. S., Chibale, K., Goss, R. J. M., Jaspars, M., Newman, D. J., and Dorrington, R. A. (2021). Antiviral Drug Discovery: Preparing for the Next Pandemic. *Chem. Soc. Rev.* 50, 3647–3655. doi:10.1039/d0cs01118e
- Ai, H., Chen, W., Zhang, L., Huang, L., Yin, Z., Hu, H., et al. (2018). Predicting Drug-Induced Liver Injury Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicol. Sci.* 165, 100–107. doi:10.1093/toxsci/kfy121
- Alhossary, A., Handoko, S. D., Mu, Y., and Kwok, C.-K. (2015). Fast, Accurate, and Reliable Molecular Docking with QuickVina 2. *Bioinformatics* 31, 2214–2216. doi:10.1093/bioinformatics/btv082
- Álvarez-Carretero, S., Pavlopoulou, N., Adams, J., Gilson, J., and Taberner, L. (2018). VSpi, an Integrated Resource for Virtual Screening and Hit Selection: Applications to Protein Tyrosine Phosphatase Inhibition. *Molecules* 23, 353. doi:10.3390/molecules23020353
- Bajusz, D., Rácz, A., and Héberger, K. (2015). Why Is Tanimoto index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* 7, 1–13. doi:10.1186/s13321-015-0069-3
- Bender, B. J., Gahbauer, S., Luttens, A., Lyu, J., Webb, C. M., Stein, R. M., et al. (2021). A Practical Guide to Large-Scale Docking. *Nat. Protoc.* 16, 4799–4832. doi:10.1038/s41596-021-00597-z
- Berdigaliyev, N., and Aljofan, M. (2020). An Overview of Drug Discovery and Development. *Future Med. Chem.* 12, 939–947. doi:10.4155/fmc-2019-0307
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235
- Blum, L. C., and Reymond, J.-L. (2009). 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* 131, 8732–8733. doi:10.1021/ja902302h
- Bray, S. A., Lucas, X., Kumar, A., and Grüning, B. A. (2020). The ChemicalToolbox: Reproducible, User-Friendly Cheminformatics Analysis on the Galaxy Platform. *J. Cheminform.* 12, 40. doi:10.1186/s13321-020-00442-7
- Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324
- Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). “The Balanced Accuracy and its Posterior Distribution,” in 2010 20th International Conference on Pattern Recognition, 3121–3124. doi:10.1109/icpr.2010.764
- Brooks, B. R., Brooks, C. L., III, Mackerell, A. D., Jr, Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). Charmm: the Biomolecular Simulation Program. *J. Comp. Chem.* 30, 1545–1614. doi:10.1002/jcc.21287
- Chen, P., Ke, Y., Lu, Y., Du, Y., Li, J., Yan, H., et al. (2019). DLIGAND2: an Improved Knowledge-Based Energy Function for Protein–Ligand Interactions Using the Distance-Scaled, Finite, Ideal-Gas Reference State. *J. Cheminf.* 11. doi:10.1186/s13321-019-0373-4
- Coley, C. W., Rogers, L., Green, W. H., and Jensen, K. F. (2018). SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* 58, 252–261. doi:10.1021/acs.jcim.7b00622
- Darme, P., Dauchez, M., Renard, A., Voutquenne-Nazabadioko, L., Aubert, D., Escotte-Binet, S., et al. (2021). AMIDE V2: High-Throughput Screening Based on AutoDock-GPU and Improved Workflow Leading to Better Performance and Reliability. *Int. J. Mol. Sci.* 22, 7489. doi:10.3390/ijms22147489
- Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* 35, 316–319. doi:10.1038/nbt.3820
- Diederik, K., and Ba, J. L. (2014). ADAM: A Method for Stochastic Optimization. *AIP Conf. Proc.* 1631, 58–62. doi:10.1063/1.4902458
- Douguet, D. (2010). e-LEA3D: a Computational-Aided Drug Design Web Server. *Nucleic Acids Res.* 38, W615–W621. doi:10.1093/nar/gkq322
- Drwal, M. N., and Griffith, R. (2013). Combination of Ligand-And Structure-Based Methods in Virtual Screening. *Drug Discov. Today Technol.* 10, e395–e401. doi:10.1016/j.ddtec.2013.02.002
- Durrant, J. D., and McCammon, J. A. (2012). Autoclickchem: Click Chemistry In Silico. *Plos Comput. Biol.* 8, 1–7. doi:10.1371/journal.pcbi.1002397
- FDA (2020a). Covid-19 Vaccines. [Dataset] (accessed May 04, 2021).
- FDA (2020b). Index to Drug-specific Information. [Dataset] (accessed May 04, 2021).
- FDA (2021). Vaccines Licensed for Use in the United States. [Dataset] (accessed May 04, 2021).
- Feinstein, W. P., and Brylinski, M. (2015). Calculating an Optimal Box Size for Ligand Docking and Virtual Screening against Experimental and Predicted Binding Pockets. *J. Cheminf.* 7. doi:10.1186/s13321-015-0067-5
- Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., et al. (2020). Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* 6, 939–949. doi:10.1021/acscentsci.0c00229
- Gentile, F., Fernandez, M., Ban, F., Ton, A.-T., Mslati, H., Perez, C. F., et al. (2021). Automated Discovery of Noncovalent Inhibitors of SARS-CoV-2 Main Protease by Consensus Deep Docking of 40 Billion Small Molecules. *Chem. Sci.* 12, 15960–15974. doi:10.1039/d1sc05579h
- Ghiandoni, G. M., Bodkin, M. J., Chen, B., Hristozov, D., Wallace, J. E. A., Webster, J., et al. (2020). Enhancing Reaction-Based De Novo Design Using a Multi-Label Reaction Class Recommender. *J. Comput. Aided Mol. Des.* 34, 783–803. doi:10.1007/s10822-020-00300-6
- Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Das, K. M. P., et al. (2020). An Open-Source Drug Discovery Platform Enables Ultra-large Virtual Screens. *Nature* 580, 663–668. doi:10.1038/s41586-020-2117-z
- Gorgulla, C., Çınaroğlu, S. S., Fischer, P. D., Fackeldey, K., Wagner, G., and Arthanari, H. (2021). VirtualFlow Ants-Ultra-Large Virtual Screenings with Artificial Intelligence Driven Docking Algorithm Based on Ant colony Optimization. *Int. J. Mol. Sci.* 22, 5807. doi:10.3390/ijms22115807
- Hartenfeller, M., Eberle, M., Meier, P., Nieto-Oberhuber, C., Altmann, K.-H., Schneider, G., et al. (2011). A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *J. Chem. Inf. Model.* 51, 3093–3098. doi:10.1021/ci200379p
- Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., and Zell, A. (2011). jCompoundMapper: An Open Source Java Library and Command-Line Tool for Chemical Fingerprints. *J. Cheminf.* 3. doi:10.1186/1758-2946-3-3
- Irwin, J. J., Shoichet, B. K., Mysinger, M. M., Huang, N., Colizzi, F., Wassam, P., et al. (2009). Automated Docking Screens: a Feasibility Study. *J. Med. Chem.* 52, 5712–5720. doi:10.1021/jm9006966
- Jadhav, A., Ferreira, R. S., Klumpp, C., Mott, B. T., Austin, C. P., Inglese, J., et al. (2010). Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* 53, 37–51. doi:10.1021/jm901070c
- Jayk Bernal, A., Gomes da Silva, M. M., Musungaie, D. B., Kovalchuk, E., Gonzalez, A., Delos Reyes, V., et al. (2022). Molnupiravir for Oral Treatment of Covid-19 in Nonhospitalized Patients. *N. Engl. J. Med.* 386, 509–520. doi:10.1056/NEJMoa2116044
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

ACKNOWLEDGMENTS

This work would not have been possible without the computational resources of the NIH HPC Biowulf cluster, and the University of Montana’s Griz Shared Computing Cluster (GSCC). We thank Rose Perry-Gottschalk, NIAID, RTB, NIH for help with the visual arts.

- Kaplan, R. M., and Milstein, A. (2021). Influence of a COVID-19 Vaccine's Effectiveness and Safety Profile on Vaccination Acceptance. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2021726118. doi:10.1073/pnas.2021726118
- Kim, E., and Nam, H. (2017). Prediction Models for Drug-Induced Hepatotoxicity by Using Weighted Molecular Fingerprints. *BMC Bioinform.* 18. doi:10.1186/s12859-017-1638-4
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2020). PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* 49, D1388–D1395. doi:10.1093/nar/gkaa971
- Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013a). Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* 53, 1893–1904. doi:10.1021/ci300604z
- Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013b). Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* 53, 1893–1904. doi:10.1021/ci300604z
- Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat. Protoc.* 10, 733–755. doi:10.1038/nprot.2015.043
- Labbé, C. M., Rey, J., Lagorce, D., Vavruša, M., Becot, J., Sperandio, O., et al. (2015). MTIOpenScreen: a Web Server for Structure-Based Virtual Screening. *Nucleic Acids Res.* 43, W448–W454. doi:10.1093/nar/gkv306
- Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an Open Source Platform for Ligand Pocket Detection. *BMC Bioinform.* 10, 1–11. doi:10.1186/1471-2105-10-168
- Le, T., Hempel, T., Winter, R., Olsson, S., Raich, L., Elez, K., et al. (2021). JEDI Billion Molecules against Covid-19: Compounds Synthesized. doi:10.6084/m9.figshare.14458896
- Li, H., Leung, K.-S., Ballester, P. J., and Wong, M.-H. (2014). Istar: A Web Platform for Large-Scale Protein-Ligand Docking. *PLoS One* 9, e85678. doi:10.1371/journal.pone.0085678
- Li, H., Leung, K.-S., Wong, M.-H., and Ballester, P. J. (2016). USR-VS: a Web Server for Large-Scale Prospective Virtual Screening Using Ultrafast Shape Recognition Techniques. *Nucleic Acids Res.* 44, W436–W441. doi:10.1093/nar/gkw320
- Mahase, E. (2021). Covid-19: Pfizer's Paxlovid Is 89% Effective in Patients at Risk of Serious Illness, Company Reports. *Br. Med. J.* 375, n2713. doi:10.1136/bmj.n2713
- Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M., and Taranto, A. G. (2020). Structure-based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* 8. doi:10.3389/fchem.2020.00343
- McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., et al. (2021). Gnina 1.0: Molecular Docking with Deep Learning. *J. Cheminf.* 13, 1–20. doi:10.1186/s13321-021-00522-2
- Meyers, J., Fabian, B., and Brown, N. (2021). De Novo molecular Design and Generative Models. *Drug Discov. Today* 26, 2707–2715. doi:10.1016/j.drudis.2021.05.019
- Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* 55, 6582–6594. doi:10.1021/jm300687e
- Novick, P. A., Ortiz, O. F., Poelman, J., Abdulhay, A. Y., and Pande, V. S. (2013). SWEETLEAD: an In Silico Database of Approved Drugs, Regulated Chemicals, and Herbal Isolates for Computer-Aided Drug Discovery. *PLoS ONE* 8, e79568. doi:10.1371/journal.pone.0079568
- O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011a). Open Babel: An Open Chemical Toolbox. *J. Cheminf.* 3, 1–14. doi:10.1186/1758-2946-3-33
- O'Boyle, N. M., and Sayle, R. A. (2016). Comparing Structural Fingerprints Using a Literature-Based Similarity Benchmark. *J. Cheminf.* 8. doi:10.1186/s13321-016-0148-0
- O'Boyle, N. M., Vandermeersch, T., Flynn, C. J., Maguire, A. R., and Hutchison, G. R. (2011b). Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminf.* 3. doi:10.1186/1758-2946-3-8
- Ochoa, R., Palacio-Rodriguez, K., Clemente, C. M., and Adler, N. S. (2021). dockECR: Open Consensus Docking and Ranking Protocol for Virtual Screening of Small Molecules. *J. Mol. Graph. Model.* 109, 108023. doi:10.1016/j.jmgm.2021.108023
- Oliveira, S. H., Ferraz, F. A., Honorato, R. V., Xavier-Neto, J., Sobreira, T. J., and de Oliveira, P. S. (2014). Kfinder: Steered Identification of Protein Cavities as a Pymol Plugin. *BMC Bioinform.* 15, 1–8. doi:10.1186/1471-2105-15-197
- Patel, H., Ihlenfeldt, W.-D., Judson, P. N., Moroz, Y. S., Pevzner, Y., Peach, M. L., et al. (2020). SAVI, In Silico Generation of Billions of Easily Synthesizable Compounds through Expert-System Type Rules. *Sci. Data* 7. doi:10.1038/s41597-020-00727-4
- Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., and Lupas, A. N. (2021). High-accuracy Protein Structure Prediction in Casp14. *Proteins: Struct. Funct. Bioinformatics* 89, 1687–1699. doi:10.1002/prot.26171
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., et al. (2021). UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers. *Protein Sci.* 30, 70–82. doi:10.1002/pro.3943
- Pitti, T., Chen, C.-T., Lin, H.-N., Choong, W.-K., Hsu, W.-L., and Sung, T.-Y. (2019). N-glyde: a Two-Stage N-Linked Glycosylation Site Prediction Incorporating Gapped Dipeptides and Pattern-Based Encoding. *Sci. Rep.* 9, 1–11. doi:10.1038/s41598-019-52341-z
- Ripphausen, P., Nisius, B., and Bajorath, J. (2011). State-of-the-art in Ligand-Based Virtual Screening. *Drug Discov. Today* 16, 372–376. doi:10.1016/j.drudis.2011.02.011
- Santos-Martins, D., Solis-Vasquez, L., Tillack, A. F., Sanner, M. F., Koch, A., and Forli, S. (2021). Accelerating AutoDock4 with GPUs and Gradient-Based Local Search. *J. Chem. Theor. Comput.* 17, 1060–1073. doi:10.1021/acs.jctc.0c01006
- Soderberg, C. K. (2018). Using OSF to Share Data: A Step-by-step Guide. *Adv. Methods Practices Psychol. Sci.* 1, 115–120. doi:10.1177/2515245918757689
- Spiegel, J. O., and Durrant, J. D. (2020). AutoGrow4: an Open-Source Genetic Algorithm for De Novo Drug Design and lead Optimization. *J. Cheminf.* 12. doi:10.1186/s13321-020-00429-4
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.
- Sterling, T., and Irwin, J. J. (2015). ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* 55, 2324–2337. doi:10.1021/acs.jcim.5b00559
- Sunseri, J., and Koes, D. R. (2016). Pharmit: Interactive Exploration of Chemical Space. *Nucleic Acids Res.* 44, W442–W448. doi:10.1093/nar/gkw287
- Tran-Nguyen, V.-K., Jacquemard, C., and Rognan, D. (2020). LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* 60, 4263–4273. doi:10.1021/acs.jcim.0c00155
- Trott, O., and Olson, A. J. (2010). Autodock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comp. Chem.* 31, 455–461. doi:10.1002/jcc.21334
- Venkatraman, V. (2021). FP-ADMET: a Compendium of Fingerprint-Based ADMET Prediction Models. *J. Cheminf.* 13. doi:10.1186/s13321-021-00557-5
- Wang, X., Chen, H., Yang, F., Gong, J., Li, S., Pei, J., et al. (2014). IDrug: a Web-Accessible and Interactive Drug Discovery and Design Platform. *J. Cheminform.* 6, 28. doi:10.1186/1758-2946-6-28
- Wang, Z., Sun, H., Shen, C., Hu, X., Gao, J., Li, D., et al. (2020). Combined Strategies in Structure-Based Virtual Screening. *Phys. Chem. Chem. Phys.* 22, 3149–3159. doi:10.1039/c9cp06303j
- Wilson, G. L., and Lill, M. A. (2011). Integrating Structure-Based and Ligand-Based Approaches for Computational Drug Design. *Future Med. Chem.* 3, 735–750. doi:10.4155/fmc.11.18
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2017). DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi:10.1093/nar/gkx1037
- Wójcikowski, M., Zielenkiewicz, P., and Siedlecki, P. (2015). Open Drug Discovery Toolkit (ODDT): a New Open-Source Player in the Drug Discovery Field. *J. Cheminform.* 7, 26. doi:10.1186/s13321-015-0078-2
- Wouters, O. J., Shadlen, K. C., Salcher-Konrad, M., Pollard, A. J., Larson, H. J., Teerawattananon, Y., et al. (2021). Challenges in Ensuring Global Access to COVID-19 Vaccines: Production, Affordability, Allocation, and Deployment. *The Lancet* 397, 1023–1034. doi:10.1016/s0140-6736(21)00306-8
- Yaacoub, J. C., Gleave, J., Gentile, F., Stern, A., and Cherkasov, A. (2021). DD-GUI: A Graphical User Interface for Deep Learning-Accelerated Virtual Screening of Large Chemical Libraries (Deep Docking). *Bioinformatics* 38, 1146–1148. doi:10.1093/bioinformatics/btab771
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-Tasser Suite: Protein Structure and Function Prediction. *Nat. Methods* 12, 7–8. doi:10.1038/nmeth.3213

- Yang, M., Tao, B., Chen, C., Jia, W., Sun, S., Zhang, T., et al. (2019). Machine Learning Models Based on Molecular Fingerprints and an Extreme Gradient Boosting Method lead to the Discovery of JAK2 Inhibitors. *J. Chem. Inf. Model.* 59, 5002–5012. doi:10.1021/acs.jcim.9b00798
- Yu, J., Zhou, Y., Tanaka, I., and Yao, M. (2010). Roll: a New Algorithm for the Detection of Protein Pockets and Cavities with a Rolling Probe Sphere. *Bioinformatics* 26, 46–52. doi:10.1093/bioinformatics/btp599
- Zhao, J., Cao, Y., and Zhang, L. (2020). Exploring the Computational Methods for Protein-Ligand Binding Site Prediction. *Comput. Struct. Biotechnol. J.* 18, 417–426. doi:10.1016/j.csbj.2020.02.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Venkatraman, Colligan, Lesica, Olson, Gaiser, Copeland, Wheeler and Roy. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Machine Learning in Antibacterial Drug Design

Marko Jukič^{1,2*} and Urban Bren^{1,2*}

¹Laboratory of Physical Chemistry and Chemical Thermodynamics, Faculty of Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia, ²Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Koper, Slovenia

OPEN ACCESS

Edited by:

Leonardo L. G. Ferreira,
University of São Paulo, Brazil

Reviewed by:

Tihomir Tomašič,
University of Ljubljana, Slovenia
Amit Kumar Banerjee,
Indian Institute of Chemical
Technology (CSIR), India

*Correspondence:

Marko Jukič
marko.jukic@um.si
Urban Bren
urban.bren@um.si

Specialty section:

This article was submitted to
Experimental Pharmacology and
Drug Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 28 January 2022

Accepted: 28 March 2022

Published: 03 May 2022

Citation:

Jukič M and Bren U (2022) Machine Learning in Antibacterial Drug Design. *Front. Pharmacol.* 13:864412. doi: 10.3389/fphar.2022.864412

Advances in computer hardware and the availability of high-performance supercomputing platforms and parallel computing, along with artificial intelligence methods are successfully complementing traditional approaches in medicinal chemistry. In particular, machine learning is gaining importance with the growth of the available data collections. One of the critical areas where this methodology can be successfully applied is in the development of new antibacterial agents. The latter is essential because of the high attrition rates in new drug discovery, both in industry and in academic research programs. Scientific involvement in this area is even more urgent as antibacterial drug resistance becomes a public health concern worldwide and pushes us increasingly into the post-antibiotic era. In this review, we focus on the latest machine learning approaches used in the discovery of new antibacterial agents and targets, covering both small molecules and antibacterial peptides. For the benefit of the reader, we summarize all applied machine learning approaches and available databases useful for the design of new antibacterial agents and address the current shortcomings.

Keywords: artificial intelligence, machine learning, computer-aided drug design (CADD), infectious diseases, antibacterial drug design, antibacterial, antibacterial target discovery, antibacterial drug resistance

INTRODUCTION

Modern antibacterial drug development currently notes a lack of novel antibacterial classes, an observation that is critical in the context of antibacterial drug resistance (Brown and Wright, 2016). Furthermore, not only single-drug resistance but also multiple-drug antibiotic resistance (MDR) has been observed in clinically relevant pathogens worldwide, rendering current established therapies ineffective (Laxminarayan et al., 2020; Vila et al., 2020). The annual number of deaths caused by infections with resistant pathogens alone is currently high and is expected to reach into millions by 2050, making high-quality data collection and reporting and antibacterial research essential (de Kraker et al., 2016; Matamoros-Recio et al., 2021). Recent advances in Computer-aided drug design (CADD) coupled with parallel and high-performance computing (HPC) platforms and new *in silico* methods represent a new paradigm for antibacterial drug discovery. In particular, machine learning methods have the potential to

Abbreviations: AI, artificial intelligence; ANN, artificial neural network; CADD, computer-assisted drug design; DT, decision tree; FSC, feedback system control; kNN, k-nearest neighbors; LOR, logistic regression; (M)LR, (multiple) linear regression; MDR, multidrug resistant; MIC, minimum inhibitory concentration; MRSA, methicillin-resistant *Staphylococcus aureus*; NB, naïve Bayes; RF, random forest; RiPPs, ribosomally synthesized and posttranslationally modified peptides; SCM, set covering machine; SVM, support vector machines.

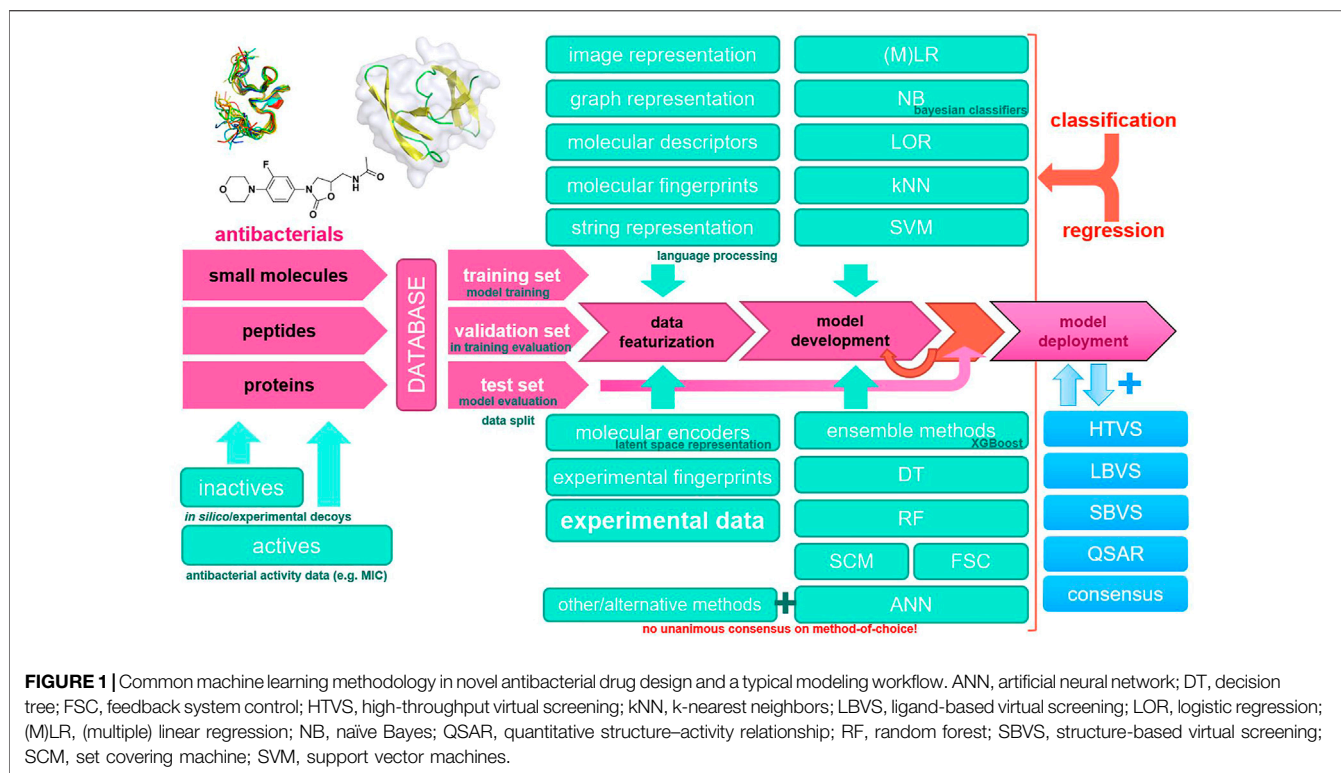


TABLE 1 | Currently available antibacterial compound and peptide databases suitable for *in silico* drug design.

| Database name | Type | Location | References |
|---|--|---|-------------------------------------|
| ChEMBL | Comprehensive bioactivity database and bioinformatics platform | https://www.ebi.ac.uk/chembl/ | Mendez et al. (2019) |
| Shared Platform for Antibiotic Research and Knowledge (SPARK) or CO-ADD | Community for open antimicrobial drug discovery | https://co-add.org/ | Thomas et al. (2018), Cooper (2015) |
| Antimicrobial Index | Microorganisms and antimicrobial agents | http://antibiotics.toku-e.com/ | Amirka and Qiubao, (2011) |
| MEGAres | Antibacterials and resistance determinants | https://megares.meglab.org/ | Doster et al. (2020) |
| Antimicrobial Combination Networks | Antibacterial combinations | http://www.sing-group.org/antimicrobialCombination/ | Jorge et al. (2016) |
| AntibioticDB | Antibacterial compounds | https://www.antibioticdb.com/ | Farrell et al. (2018) |
| The Drug Repurposing Hub | Compounds, targets, and indications | https://clue.io/repurposing/ | Corsello et al. (2017) |
| APD3 | Antibacterial peptides | https://aps.unmc.edu/ | Wang et al. (2016) |
| CAMP3 | Antibacterial peptides | http://www.camp3.bicnirrh.res.in/ | Waghu et al. (2016) |
| BAGEL4 | Bacteriocins and RiPPs | http://bagel4.molgenrug.nl/ | van Heel et al. (2018) |
| DBAASP v3 | Antibacterial peptides | https://dbaasp.org/ | Pirtskhalava et al. (2016) |
| Defensins knowledgebase | Defensins | http://defensins.bii.a-star.edu.sg/ | Seebah et al. (2007) |
| DRAMP | Antibacterial peptides | https://ngdc.cncb.ac.cn/ | Kang et al. (2019) |
| BaAMPs | Biofilm-active peptides | http://www.baamps.it/ | Di Luca et al. (2015) |
| dbAMP 2.0 | Antibacterial peptides | https://awi.cuhk.edu.cn/dbAMP/ | Jhong et al. (2022) |
| AECD | Antimicrobial enzyme combinations | https://www.ceb.uminho.pt/aecd/ | Jorge et al. (2019) |

increase the accuracy of high-throughput virtual screening using ligand-based, structure-based, or consensus-based approaches (Serafim et al., 2020). It should be noted that modern software implementations of machine learning algorithms efficiently utilize computer hardware and are ideal for the bioinformatics or chemoinformatics scenario; however, extreme care should be taken with input data

(Bzdok et al., 2017). Most importantly, the increasing availability of data makes machine learning methods even more important, either as a stand-alone method or in a consensus scenario where they can boost traditional medicinal chemistry approaches (He et al., 2021). In this review, we focus on machine learning approaches in CADD that have been reported in recent years and have been used in

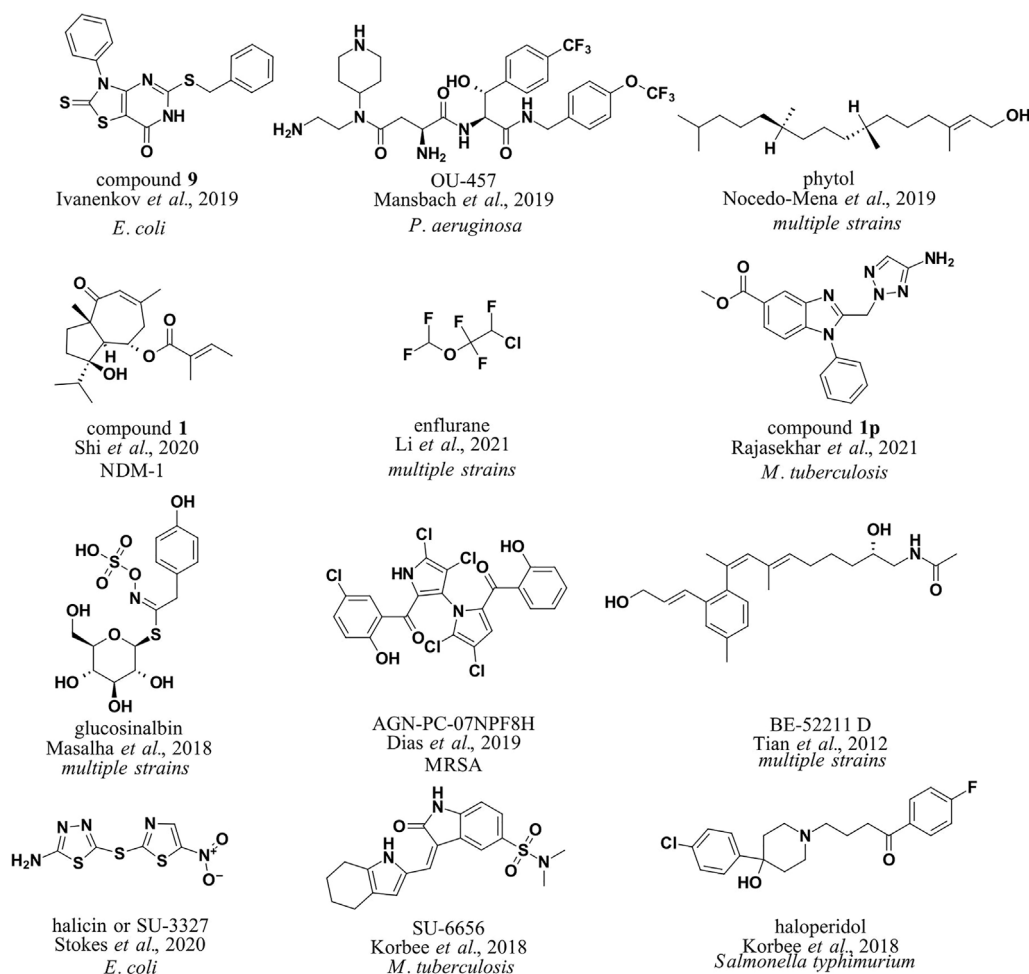


FIGURE 2 | Antibacterial compounds identified by machine learning boosted *in silico* methods in CADD.

the development of novel antibacterials. We summarize the relevant databases and consolidate the general workflow along with the methods used in **Figure 1**.

RELEVANT DATABASES FOR ANTIBACTERIAL DRUG DESIGN

The currently accessible libraries of antibacterial compounds are enlisted that include small molecules or peptides that can be used for the design of new antibacterial agents and model development (**Table 1**). The reader should also be aware of tailored or focused libraries and antibacterial libraries offered by commercial compound suppliers and complete online antibacterial drug discovery communities (CO-ADD; of special mention is that the industry also contributes to the CO-ADD community, or previously SPARK-database). The ChEMBL bioinformatics platform is by far the most comprehensive resource (especially considering small molecules), followed by CO-ADD (SPARK) and antimicrobial index. Databases supporting antibacterial peptides are far more common and offer quality data.

SMALL MOLECULES

To utilize machine learning approaches in the design of antibacterial small molecules and test different machine learning approaches, Yang *et al.* computed a simple set of molecular descriptors for small molecules with and without antibacterial properties and evaluated the decision tree, k-nearest neighbor, and support vector machine (SVM) classification models. The authors noted the good accuracy of the SVM approach and the applicability of the methodology for antibacterial drug design. Developed models produced the best prediction accuracies of 96.66 and 98.15% for antibacterial compounds and 99.50 and 98.02% for non-antibacterial compounds (Yang *et al.*, 2009). Ivanenkov *et al.* (2019) compiled a database of 145,000 small molecules, most of which came from a proprietary high-throughput screening campaign with *Escherichia coli* (*E. coli*; 1,786 active and 130,855 inactive compounds; all data points were obtained under the same experimental conditions). 1243 molecular descriptors were calculated using Dragon, ChemoSoft, MOE, and SmartMining software tools. Subsequently, self-organizing

maps (Kohonen maps) were used for classification and prediction of antibacterial activity with SmartMining software, and good results were obtained (predictive power of 75.5% on average). The developed models were deployed to identify new agents against *E. coli* (compound **9**, **Figure 2**). Maltarollo (2019) focused on *Staphylococcus aureus* (*S. aureus*), specifically FabI inhibitors. 166 literature compounds were collected and molecular descriptors and fingerprints were calculated using PaDEL software. Decision trees (DTs), random forests (RF), multilayer perceptron (MLP), k-nearest neighbors (kNN), Naive Bayes (NB), and support vector machine (SVM) models were trained for classification. RF models performed best in classifying known connections.

Shi et al. collected a database of New Delhi metallo beta-lactamase (NDM-1) inhibitors (511 compounds) from the literature (Shi et al., 2020). This was followed by the calculation of molecular descriptors (34 descriptors, MOE software) and the representation of SMILES strings padded with zeros up to a length of 550. Different methods were tested, such as RF, SVM, and linear discriminant analysis. Finally, it was decided to use the RF model, which performed much better than the classical virtual screening model (90.5 and 69.14%, respectively). The model was used to predict potential NDM-1 inhibitors from a natural product library that contained 2,172 compounds (compound **1**, **Figure 2**). The authors noted that the deep-learning method was not very powerful because of low data availability. Li et al. approached in a more general manner using more data points from the ChEMBL database (Li et al., 2021). The group collected a library of 2708 active antibacterial compounds (IC₅₀ cut-off of 10 μ M) and 78,620 inactive compounds and proceeded to calculate fingerprints (FP2, FP3, FP4, DLFP, MACCS, ECFP2, ECFP4, ECFP6, FCFP2, FCFP4, and FCFP6) and vector representations (mol2vec, SMILES2Vec, FP2VEC software; Jaeger et al., 2018; Öztürk et al., 2018; Jeon and Kim, 2019). Several machine learning methods were reviewed, and the FP2 database along with RF, SVM, and MLP methods was selected for screening (scikit-learn library; average accuracy of 0.85). The team then constructed a predictor for antibacterial agents based on all three models and applied it to the FDA-approved small-molecule database (DrugBank, Wishart et al., 2018). Of interest is the observed low FP2 similarity (<0.2) between the predicted and FDA-approved antibacterial agents. The group focused on the nine most different predicted compounds from the FDA antibacterials with the highest screening scores in all three models; however, it did not follow up with biological evaluation. The identified compounds belonged to the classes of anticancer drugs, ocular antihypertensives, and general anesthetics, with enflurane scoring the highest. Enflurane was previously demonstrated to possess antibacterial properties *in vitro* (enflurane, **Figure 2**).

The superiority of machine learning-assisted molecular docking was reported by de Avila et al. (2018). The group collected a database of 22 structurally supported 3-dehydroquinate dehydratase (DHQD) inhibitors with measured inhibition constants. They developed a new polynomial scoring function with selected energy terms from classical scoring functions. Using Sandres software (Lasso and

Ridge Regression), the newly developed scoring functions performed significantly better in the DHQD system test set supplemented by decoy compounds (the group did not further deploy the model).

Mansbach et al. focused on the permeation of Gram-negative bacteria and developed a fragment-based approach. They collected a database of compounds with MIC values in *Pseudomonas aeruginosa* (*P. aeruginosa*) and calculated fragment-based molecular representations for sparse regression and hierarchical clustering to identify the most relevant fragments thought to influence antibacterial activity (Mansbach et al., 2020). The method was used to predict new compounds with antibacterial properties and design “hybrid” molecules from multiple fragments (OU-457, **Figure 2**). Predicted molecules were experimentally evaluated.

Interestingly, an approach combining both antibacterial small molecules and antimicrobial peptides in a heterogeneous library was reported by Nava Lara et al. (2019). To identify compounds with antimicrobial activity in the intestinal flora, 1444 descriptors were calculated (PaDEL Descriptor software) and 52 different machine learning algorithms were tested (WEKA, AutoWEKA software) to finally select a random committee algorithm classifier with receiver operating characteristic (ROC) area under the curve (AUC) performance of 0.83 for the classification. The model was applied to the FDA-approved antimicrobial agents and found that almost half of them had potential broad-spectrum activity against intestinal bacteria; however, the predictions were not experimentally substantiated. Since antibacterial peptides make up a large proportion of antibacterial chemical substances, they are discussed in more detail in the section *Antibacterial Peptides*.

Mycobacteria infections are a significant public health problem worldwide. The development of novel antimycobacterial agents remains a challenge, especially in light of the increasing emergence of multidrug-resistant strains of mycobacteria. Several reviews have been published collecting the main therapeutic targets in this field and highlighting the importance of *in silico* methods, particularly promoted by machine learning approaches and focusing on cell-wall permeability studies (Aleksandrov and Myllykallio, 2019; Pushkaran et al., 2019; Ejalonibu et al., 2021). In this way, classical approaches of virtual screening against the mycobacterial target PrpR (Vina, Glide software), MMGBSA, and molecular dynamics (MD) studies on hit compounds were complemented by the MycoCSM method to identify novel benzimidazole derivatives as potential PrpR inhibitors (compound **1p**, **Figure 2**; Rajasekhar et al., 2021). MycoCSM is a graph-based DT model (scikit-learn library) based on 15,000 unique compounds (featurized with RDkit descriptors) with activity against bacteria of the genus *Mycobacterium* (MIC cut-off of 1 μ M), achieving correlation coefficients of up to 0.89 in predicting bioactivity in terms of minimum inhibitory concentration (Pires and Ascher, 2020).

Korbee et al. used predictive clustering trees (PCTs) to explore host-directed pathways toward antimycobacterial drug design (Clus software; <https://sourceforge.net/projects/clus/>). The group deployed their models on a library of pharmacologically

active compounds in a (LOPAC)-based drug-repurposing screen to identify experimentally validated compounds which target receptor tyrosine kinases (RTKs) and inhibit intracellular mycobacteria (SU-6656, **Figure 2**) and salmonellae (haloperidol, **Figure 2**; Korbbe et al., 2018).

NATURAL COMPOUNDS

Prediction of antibacterial activity while considering molecular structure and metabolic reaction networks was also attempted by Nocedo-Mena et al. (2019) (dataset: Jeong et al., 2000). The metabolic reaction network data were merged with compounds with MIC properties in ChEMBL, and machine learning modeling with multi-output perturbations was used to build predictive models. The models were deployed to identify natural antibacterial compounds from *C. incisa* (phytol, **Figure 2**).

The natural compounds were further explored by Masalha et al., (2018). The group assembled a library of 628 antibacterial compounds (Comprehensive Medicinal Chemistry Database) along with an inactive set of 2892 natural compounds (AnalytiCon Discovery GmbH database) and proceeded to calculate molecular descriptors (MOE software). An iterative indexing model based on stochastic elimination was created for discriminative filtering and antibacterial identification *via* the calculated molecular bioactivity index (Rayan et al., 2010). The model ROC AUC for antibacterial classification was 0.96, and the model was deployed for screening of the natural product database to identify 10 potential antibacterial hits, two of which were experimentally confirmed as active and others are still under research (glucosinabin, **Figure 2**). It is interesting to note that the authors found that comparable performance could not be achieved with either structure-based or ligand-based approaches due to non-efficient scoring or the number of false-positives.

Another report focused on marine natural sources to identify new compounds with activity against MRSA (Dias et al., 2019). Construction of a database of 6645 small molecules (ChEMBL, PubChem, ZINC; active molecules with MIC <5 μ M and inactive molecules with MIC \geq 5 μ M) was followed by a calculation of a comprehensive list of molecular descriptors and fingerprints (PaDEL and CDK Descriptor Software) to finally build a regression model using RF, SVM, Gaussian processes (GPs), and consensus approaches for pMIC determination against MRSA. The best consensus model (R^2 of 0.68) was deployed on the StreptomeDB database and resulted in 150 hits with 12 prioritized compounds, all with confirmed anti-MRSA experimental activity (AGN-PC-07NPF8H, **Figure 2**). The same group also reported a nuclear magnetic resonance (^1H and ^{13}C NMR)-based approach where compounds were featurized using experimental NMR-spectra assignment data. The compound library was a dataset of 155 samples that included 50 crude extracts, 55 fractions, and 50 pure compounds obtained from microbial actinobacteria isolated from marine sediments off the Madeira archipelago. RF, SVN, and convolutional neural network (CNN) models were generated

with an accuracy of 0.77 for the test set and were ready for further research and application.

Drug similarity identification was also attempted using molecular descriptors and fingerprints calculated using a database from the Current Medicinal Chemistry Database, MDL Drug Data Report, World Drug Index (drug-like molecules), and Available Chemicals Directory for non-drug-like molecules (180,000 compounds in total). Naive Bayesian classifiers and recursive partitioning models were developed and used for drug similarity prediction in the Traditional Chinese Medicine Compound Database (TCMD) (Tian et al., 2012). The research found that the classifiers can successfully provide valuable information in the early stages of drug design (drug-like compound identification accuracy of 0.86) and identify important drug-like scaffolds and even classify them by pharmacological activity, for example, label scaffolds of antibacterial compounds (BE-52211D, **Figure 2**).

Indeed, natural compounds represent an invaluable source of chemical diversity, and their drawbacks (availability, complexity, synergistic pharmacodynamics) in drug development could be mitigated by modern machine learning methods (Rodrigues et al., 2016). To this end, Zhang et al. have collected several machine learning protocols for activity prediction of natural products (Zhang et al., 2021).

ANTIBACTERIAL PEPTIDES

An important subfield of the discovery of new antibacterials is also the discovery of antibacterial peptides. The latter can serve as active agents, starting points for the design of peptidomimetics, or probes for further studies. The field and *in silico* tools have been reviewed previously (Lee et al., 2017; Cardoso et al., 2020; Wang et al., 2021), with the emphasis on machine learning-enabled antimicrobial peptide discovery and SVM for the discovery of membrane-active peptides (Lee et al., 2018). However, Frece reported a successful design of cationic antibacterial peptides derived from protegrin-1 as early as 2006 (Frece, 2006), and machine learning methodology contributed significantly to the design and discovery of novel peptides, as demonstrated by Fjell et al. To single out just one report, they reinforced the traditional QSAR approach with an artificial neural network model (ANN) that inferred a set of peptides with known antibacterial properties from computed descriptors (MOE software). After deploying the model in a screening scenario (*in silico* library with random peptides), short cationic peptides with MICs in the range of 0.3–10 μ M were identified (Fjell et al., 2009). The extended research group later reported an interesting approach for relational learning algorithms (RelF and WEKA software for regression) to explore patterns from the relational structures of the antibacterial peptides or an approximate attribute-value representation of the peptides (Szaboova et al., 2012). Feature vectors for peptide representation were also used using Chou's pseudo-amino acid composition (PseAAC), and the SVM was successfully used to classify antibacterial peptides (Khosravian et al., 2013).

The later approaches were also extended beyond antibacterial peptide identification to peptide target selectivity or prediction of Gram-positive or Gram-negative activities (Veltri et al., 2015). The group used an evolutionary feature construction and a fast correlation-based filter selection algorithm with logistic regression (WEKA) to successfully identify antibacterial peptides of up to 11 amino acids in length. The same group used APD3 database, converted peptide sequences into zero-padded numerical vectors of length 200, and trained a deep neural network (DNN; Keras, TensorFlow software) model to classify antimicrobial peptides (accuracy of 0.98 on APD3 data). Embedding vector visualization was also performed, and a reduced alphabet learnt from the DNN model was developed. Reduced sequence space retained good classification performance (Veltri et al., 2018). Müller et al. trained a recurrent neural network (RNN) with helical antimicrobial peptides (1554 peptides, APD). The sequences were padded according to the length of the longest sequence, N-terminal token added, and One-hot encoding employed (Müller et al., 2018). The resulting model was developed for *de novo* sequence generation, where 82% were predicted to be active antimicrobial peptides compared to 65% of randomly sampled sequences with the same amino acid distribution as the training set (CAMP AMP prediction tool; Waghu et al., 2014). Wu et al. used previous amino acid substitution data for antibacterial peptides and developed an amino acid activity contribution matrix (Wu et al., 2014). Using this methodology, the group developed a 12-mer DP7 peptide with antibacterial properties against multiple strains (Zhang et al., 2019). Similarly, Yoshida et al. used a natural antibacterial peptide Temporin-Ali (FFPIVGKLLSGLL-NH₂) and PSI BLAST to create a library of distantly related and functionally similar sequences, prepared the peptides, and evaluated their antibacterial activities *in vitro* on *E. coli* to construct a fitness matrix. The data were then used to train a model and deploy it to optimize peptide sequences. The group produced a peptide with 163-fold lower activity on *E. coli* bacteria (Yoshida et al., 2018). Another approach using rough set theory constructed quantitative structure–activity relationship rules for existing antibacterial peptides. New sequence development *via* a genetic algorithm and further *in vitro* testing resulted in a peptide being active against *Staphylococcus epidermidis* (*S. epidermidis*) (Boone et al., 2021).

Approaches were again extended by considering toxicity data in the development of novel antibacterial peptides intended for human drug development campaigns. Capecchi et al. used the Database of Antimicrobial Activity and Structure of Peptides (DBAASP; 4774 active peptides with an MIC threshold of 32 mg/ml) to train a recurrent neural network (RNN) generative model to develop nonhemolytic antibacterial peptides with activity against *P. aeruginosa*, *Acinetobacter baumannii* (*A. baumannii*), MRSA, and a broader range of MDR strains. To test the performance of machine learning models for antibacterial peptide design, Wani et al. trained models on a database of antibacterials (2638) and inactive peptides (3700) using RF, kNN, SVM, DT, NB, quadratic discriminant analysis (QDA), and ensemble learning. RF models were found to perform best in validation experiments. The group also highlighted three important peptide descriptors as essential for antibacterial

activity, namely, charge, polarity, and pseudo-amino acid composition (Wani et al., 2021). The field of *in silico* tools for designing antibacterial peptides using machine learning is also gaining traction, and targeted tools such as AMPGAN v2 are being developed (Van Oort et al., 2021). AMPGAN v2 is a bidirectional conditional generative adversarial network (BiCGAN) that targets *de novo* generation of antibacterial peptides. The group used training data by compiling the Database of Antimicrobial Activity and Structure of Peptides (DBAASP), Antiviral Peptide database (AVPdb), and UniProt databases (Apweiler et al., 2004; Gogoladze et al., 2014; Qureshi et al., 2014).

ANTIBACTERIAL DRUG RESISTANCE

Machine learning approaches are also being used to combat antibiotic resistance. Back in 2017, Macesic et al. published a review of antibacterial susceptibility testing using genotype–phenotype prediction, machine learning approaches to identify resistant strains, and the use of machine learning to improve treatment and optimize clinical approaches to MDR infections (Macesic et al., 2017). Interestingly, the authors lamented data abstraction and quality but pointed out that the methodology gains strength with the availability of quality data. A recent review article discusses several bioinformatics approaches involving machine learning that are useful for studying bacterial resistance, such as the use of modern bioinformatics approaches for the interpretation of data from increasing sequencing libraries; study of protein structures; *in silico* analysis of serovar, serogroup, and antigen markers; the development of *in silico* plasmid detection methods; *in silico* identification of resistance genes; antibacterial surveillance; and in turn, the prediction of the evolution of antibacterial drug resistance (Ndagi et al., 2020). In addition, machine learning approaches have been used beyond resistance prediction using genomic data to elucidate resistance mechanisms and for antibacterial stewardship applications. The latter are mainly concerned with patient data analysis, diagnosis, treatment, and prevention of resistance development in a clinical scenario (Anahtar et al., 2021). With the increasing use of antibiotics and the accompanying bacterial resistance, we cannot overemphasize the importance of these new approaches in translational research. Furthermore, the power of reported methods is increasing with the growth of quality data and availability of curated and resistance-focused libraries such as Plasmid ATLAS by Jesus et al., (2019), Ensembl Genomes (Bacteria) by Yates et al., (2022), BacDive by Reimer et al., (2019), Virulence Factor Database VFDB by Chen et al., (2005), Beta-Lactamase Database (BLDB) by Naas et al., (2017), Antibiotic Resistance Genes Database (ARDB, Liu and Pop, 2009), BacMed (Pal et al., 2014), and Comprehensive Antibiotic Resistance Database (CARD, McArthur et al., 2013 and Alcock et al., 2020).

MODERN APPROACHES

As reviewed already by Durrant and Amaro in 2014 (Durrant and Amaro, 2015) up to now, the medicinal chemistry community

and pharmaceutical industry are adopting machine learning techniques in medicinal chemistry and drug design in general (Ekins et al., 2019) and antibacterial drug development (Patel et al., 2020; Serafim et al., 2020). Of special mention would be the acknowledgment of enormous data availability, its application toward drug design (Burki, 2020), and utilization of modern artificial intelligence approaches (David et al., 2021). Specifically, the applications of modern deep learning methods in antibacterial drug design are evident from a multitude of published reports in scientific literature, tailored offerings by commercial drug design software developers, and emergence of deep-learning in drug design-focused CROs and start-ups (Schroedl, 2019; Chang et al., 2019; Gupta et al., 2021; da Silva et al., 2021).

Deep-Learning and Artificial Neural Networks

An excellent example of the development and use of deep learning supervised, semi-supervised, or unsupervised models in the area of novel antibacterial drug development and discovery was recently reported (Stokes et al., 2020). The group initially generated the dataset by computing graph representations, Morgan fingerprints, and molecular features computed using RDKit (internal training set of 2560 compounds, 120 positive controls; with a test set: Broad's Drug Repurposing Hub of 6111 compounds) and used a Directed Message Passing Neural Network (D-MPNN; Chemprop implementation available on Github), a type of graph convolutional neural network for model development. After prioritization by toxicity prediction, the authors identified one promising new antibiotic, halicin (SU -3327, **Figure 2**), and eight (ZINC000098210492, ZINC000001735150, ZINC000225434673, ZINC000019771150, ZINC000004481415, ZINC000004623615, ZINC000238901709, and ZINC000100032716) other potential antibiotic candidates and experimentally validated the obtained hits to have an antibiotic activity on *E. coli*.

K-Nearest Neighbor

kNN is a supervised learning method that can be applied for classification and regression tasks and is effectively utilized in medicinal chemistry for novel antibacterial drug design. A classification application of kNN was reported by Karakoc et al. for classification of small molecules based on selecting the most relevant set of chemical descriptors used for ultimate discrimination between active and inactive compounds on various biological systems (Karakoc et al., 2007). A comprehensive list of kNN applications in classification and regression tasks all applied toward drug delivery for infectious disease treatment, treatment regimen optimization, drug delivery system and administration route design, and drug delivery outcome prediction was reported by He et al. (2021).

Support Vector Machines

SVM supervised learning models are also widely applied for classification, regression, and ranking/virtual screening tasks in medicinal chemistry in a range of fields such as novel anticancer research, design of antivirals, protein-protein interaction

research etc. (Romero-Molina et al., 2019). Focusing on antibacterial drug design, Li et al. reported SVM model development from the fingerprint-featurized ChEMBL database in order to identify novel antibacterial compounds (Li et al., 2021). SVM model applications in antibacterial design and antibacterial drug resistance research were reviewed by Serafim et al. (2020). In a broader scope, recent advances in SVMs and their numerous drug discovery applications are summarized by Maltarollo et al. (2019).

Random Forest and Decision Trees

RF is a supervised ensemble learning method that consists of a multitude of decision trees, constructed at a training phase. Upon reviewing literature on novel antibacterial design supported by machine learning, RF models were found to be one of the most commonly applied for classification, regression, and other tasks and represent a performance and computationally lean approach. In this review, a number of RF applications are presented, for small molecules, peptides (Bhadra et al., 2018), natural product-based antibacterial design, and studying antibacterial drug resistance (Dias et al., 2019; Maltarollo et al., 2019; Shi et al., 2020; Li et al., 2021; Wani et al., 2021). A good example of underlying supervised learning DT method was reported by Suay-Garcia et al. (2020). The authors created a QSAR model to predict antibacterial activity against *E. coli*. The compounds were classified using a tree-based method and linear discriminant analysis. A comprehensive review on other DT applications is also provided by Serafim et al., (2020).

Coupling to Big Data

Needless to say, we must emphasize the coupling of modern machine learning approaches to valuable data sources. Sripriya Akondi et al., (2022) emphasize the use of compound and protein conformational data which in its abundance classifies as big data in all respects. However, common problems with big data sources such as data quality, over-fitting, and difficult or lengthy protocols should be taken in consideration (Motamedi et al., 2022). Taken together, the big data era will walk hand-in-hand with future drug design and will have a significant impact on how to approach a drug discovery campaign (Zhu, 2020; Bhattarai, et al., 2022; Lee et al., 2022). Zhao et al. point out in a wonderful report "10 Vs." or characteristics that are intrinsic in drug discovery big data that we should be aware of and utilize, namely: volume (size of data), velocity (data growth), variety (lots of data sources), veracity (variable data quality), validity (authenticity of data), vocabulary (aware of the terminology), venue (numerous data platforms), visualization (presentation and patterns in data), volatility (time domain of the data and usefulness time window), and value (associated economic and added value, Zhao et al., 2020).

CONCLUSION

In conjunction with antibacterial compound databases (**Table 1**) and general (big) data sources such as ChEMBL and CO-ADD (SPARK), efficient research in the area of new antibacterial drug design and

target identification is possible (Gaulton et al., 2017; Wishart et al., 2018). Incorporating novel machine learning methods can successfully boost the traditional medicinal chemistry approaches, and this review highlights a host of applications and machine learning model deployments. The examples include synthetic and natural small molecules, as well as peptides, ranging from a narrow spectrum of Gram-positive or Gram-negative bacteria to a broad spectrum of compounds acting on mycobacteria and eventually even MDR bacteria. However, in reviewing the literature, it is immediately apparent that medicinal chemistry is currently still in the introductory phase of exploring modern (and also established) machine learning methods and adapting them to the field. Most of the reports are proof-of-concept works where the models are only deployed to test the data and no experimental biological evaluation is performed. However, the analysis of the best performing featurization approaches and the methods themselves may be even more important takeaways.

Input data is of critical importance, and the available tailored or focused antibacterial data libraries, especially public resources, leave much to be desired. The good availability of antimicrobial peptide data and general relational databases, such as the ones

mentioned above, improves the situation. In conclusion, the immense value of modern machine learning methods is obvious—coupled with classical and experimental approaches in medicinal chemistry—and new advances in antibacterial drug design and mode of action research are possible.

AUTHOR CONTRIBUTIONS

MJ and UB interpreted the data from the literature. MJ and UB wrote the original draft. MJ and UB reviewed, edited, and drafted the manuscript and approved the final version.

FUNDING

This work was supported by the Slovenian Ministry of Science and Education infrastructure project grants HPC-RIVR and RI-SI-ELIXIR and by the Slovenian Research Agency (ARRS) program and project grants P2-0046, J1-2471, J1-1715, N1-0209, P1-0403, L2-3175 and J1-9186.

REFERENCES

- Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., et al. (2020). CARD 2020: Antibiotic Resistance Surveillance with the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* 48 (D1), D517–D525. doi:10.1093/nar/gkz935
- Aleksandrov, A., and Myllykallio, H. (2019). Advances and Challenges in Drug Design against Tuberculosis: Application of In Silico Approaches. *Expert Opin. Drug Discov.* 14 (1), 35–46. doi:10.1080/17460441.2019.1550482
- Amirkia, V. D., and Qiubao, P. (2011). The Antimicrobial Index: a Comprehensive Literature-Based Antimicrobial Database and Reference Work. *Bioinformation* 5 (8), 365–366. doi:10.6026/97320630005365
- Anahar, M. N., Yang, J. H., and Kanjilal, S. (2021). Applications of Machine Learning to the Problem of Antimicrobial Resistance: An Emerging Model for Translational Research. *J. Clin. Microbiol.* 59 (7), e0126020. doi:10.1128/jcm.01260-20
- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.* 32 (Suppl. 1_1), D115–D119. doi:10.1093/nar/gkh131
- Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S. W. I. (2018). AmPEP: Sequence-Based Prediction of Antimicrobial Peptides Using Distribution Patterns of Amino Acid Properties and Random forest. *Sci. Rep.* 8 (1), 1697–1710. doi:10.1038/s41598-018-19752-w
- Bhattarai, S., Kumar, R., Nag, S., and Namasivayam, V. (2022). “Big Data in Drug Discovery,” in *Machine Learning and Systems Biology in Genomics and Health* (Singapore: Springer), 17–48. doi:10.1007/978-981-16-5993-5_2
- Boone, K., Wisdom, C., Camarda, K., Spencer, P., and Tamerler, C. (2021). Combining Genetic Algorithm with Machine Learning Strategies for Designing Potent Antimicrobial Peptides. *BMC bioinformatics* 22 (1), 239–317. doi:10.1186/s12859-021-04156-x
- Brown, E. D., and Wright, G. D. (2016). Antibacterial Drug Discovery in the Resistance Era. *Nature* 529 (7586), 336–343. doi:10.1038/nature17042
- Burki, T. (2020). A New Paradigm for Drug Development. *The Lancet Digital Health* 2 (5), e226–e227. doi:10.1016/S2589-7500(20)30088-1
- Bzdok, D., Krzywinski, M., and Altman, N. (2017). Points of Significance: Machine Learning: a Primer. *Nat. Methods* 14 (12), 1119–1120. doi:10.1038/nmeth.4526
- Cardoso, M. H., Orozco, R. Q., Rezende, S. B., Rodrigues, G., Oshiro, K. G. N., Cândido, E. S., et al. (2020). Computer-aided Design of Antimicrobial Peptides: Are We Generating Effective Drug Candidates? *Front Microbiol.* 10, 3097. doi:10.3389/fmicb.2019.03097
- Chang, C. H., Hung, C. L., and Tang, C. Y. (2019). “November). A Review of Deep Learning in Computer-Aided Drug Design,” in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), 1856–1861.
- Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., et al. (2005). VFDB: a Reference Database for Bacterial Virulence Factors. *Nucleic Acids Res.* 33 (Suppl. 1_1), D325–D328. doi:10.1093/nar/gki008
- Cooper, M. A. (2015). A Community-Based Approach to New Antibiotic Discovery. *Nat. Rev. Drug Discov.* 14 (9), 587–588. doi:10.1038/nrd4706
- Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., Asiedu, J., Narayan, R., Mader, C. C., Subramanian, A., and Golub, T. R. (2017). The Drug Repurposing Hub: a Next-Generation Drug Library and Information Resource. *Nat. Med.* 23 (4), 405–408. doi:10.1038/nm.4306
- da Silva, T. H., Hachigian, T. Z., Lee, J., and King, M. D. (2022). Using Computers to ESKAPE the Antibiotic Resistance Crisis. *Drug Discov. Today* 27 (2), 456–470. doi:10.1016/j.drudis.2021.10.005
- David, L., Brata, A. M., Mogosan, C., Pop, C., Czako, Z., Muresan, L., Ismaiel, A., Dumitrascu, D. I., Leucuta, D. C., Stanculete, M. F., Iaru, I., and Popa, S. L. (2021). Artificial Intelligence and Antibiotic Discovery. *Antibiotics (Basel)* 10 (11), 1376. doi:10.3390/antibiotics10111376
- de Avila, M. B., and de Azevedo, W. F., Jr (2018). Development of Machine Learning Models to Predict Inhibition of 3-dehydroquinate Dehydratase. *Chem. Biol. Drug Des.* 92 (2), 1468–1474.
- de Kraker, M. E., Stewardson, A. J., and Harbarth, S. (2016). Will 10 million people die a year due to antimicrobial resistance by 2050? *Plos Med.* 13 (11), e1002184. doi:10.1371/journal.pmed.1002184
- Di Luca, M., Maccari, G., Maisetta, G., and Batoni, G. (2015). BaAMPs: the Database of Biofilm-Active Antimicrobial Peptides. *Biofouling* 31 (2), 193–199. doi:10.1080/08927014.2015.1021340
- Dias, T., Gaudêncio, S. P., and Pereira, F. (2019). A Computer-Driven Approach to Discover Natural Product Leads for Methicillin-Resistant *Staphylococcus aureus* Infection Therapy. *Mar. Drugs* 17 (1), 16. doi:10.3390/md17010016
- Doster, E., Lakin, S. M., Dean, C. J., Wolfe, C., Young, J. G., Boucher, C., et al. (2020). MEGARes 2.0: a Database for Classification of Antimicrobial Drug, Biocide and Metal Resistance Determinants in Metagenomic Sequence Data. *Nucleic Acids Res.* 48 (D1), D561–D569. doi:10.1093/nar/gkz1010
- Durrant, J. D., and Amaro, R. E. (2015). Machine-learning Techniques Applied to Antibacterial Drug Discovery. *Chem. Biol. Drug Des.* 85 (1), 14–21. doi:10.1111/cbdd.12423

- Ejalonibu, M. A., Ogundare, S. A., Elrashedy, A. A., Ejalonibu, M. A., Lawal, M. M., Mhlongo, N. N., et al. (2021). Drug Discovery for Mycobacterium tuberculosis Using Structure-Based Computer-Aided Drug Design Approach. *Int. J. Mol. Sci.* 22 (24), 13259. doi:10.3390/ijms222413259
- Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., et al. (2019). Exploiting Machine Learning for End-To-End Drug Discovery and Development. *Nat. Mater.* 18 (5), 435–441. doi:10.1038/s41563-019-0338-z
- Farrell, L. J., Lo, R., Wanford, J. J., Jenkins, A., Maxwell, A., and Piddock, L. J. V. (2018). Revitalizing the Drug Pipeline: AntibioticDB, an Open Access Database to Aid Antibacterial Research and Development. *J. Antimicrob. Chemother.* 73 (9), 2284–2297. doi:10.1093/jac/dky208
- Fjell, C. D., Jenssen, H., Hilpert, K., Cheung, W. A., Panté, N., Hancock, R. E., et al. (2009). Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. *J. Med. Chem.* 52 (7), 2006–2015. doi:10.1021/jm8015365
- Frece, V. (2006). QSAR Analysis of Antimicrobial and Haemolytic Effects of Cyclic Cationic Antimicrobial Peptides Derived from Protegrin-1. *Bioorg. Med. Chem.* 14 (17), 6065–6074. doi:10.1016/j.bmc.2006.05.005
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2017). The ChEMBL Database in 2017. *Nucleic Acids Res.* 45 (D1), D945–D954. doi:10.1093/nar/gkw1074
- Gogoladze, G., Grigolava, M., Vishnepolsky, B., Chubinidze, M., Duroux, P., Lefranc, M. P., et al. (2014). DBAASP: Database of Antimicrobial Activity and Structure of Peptides. *FEMS Microbiol. Lett.* 357 (1), 63–68. doi:10.1111/1574-6968.12489
- Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., and Kumar, P. (2021). Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery. *Mol. Divers.* 25 (3), 1315–1360. doi:10.1007/s11030-021-10217-3
- He, S., Leanse, L. G., and Feng, Y. (2021). Artificial Intelligence and Machine Learning Assisted Drug Delivery for Effective Treatment of Infectious Diseases. *Adv. Drug Deliv. Rev.* 178, 113922. doi:10.1016/j.addr.2021.113922
- Ivanenkov, Y. A., Zhavoronkov, A., Yamidanov, R. S., Osterman, I. A., Sergiev, P. V., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Veselov, M. S., Ayginin, A. A., Kartsev, V. G., Skvortsov, D. A., Chemeris, A. V., Baimiev, A. K., Sofronova, A. A., Malyshev, A. S., Filkov, G. I., Bezrukov, D. S., Zagribelnyy, B. A., Putin, E. O., Puchinina, M. M., and Dontsova, O. A. (2019). Identification of Novel Antibacterials Using Machine Learning Techniques. *Front Pharmacol.* 10, 913. doi:10.3389/fphar.2019.00913
- Jaeger, S., Fulle, S., and Turk, S. (2018). Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* 58 (1), 27–35. doi:10.1021/acs.jcim.7b00616
- Jeon, W., and Kim, D. (2019). FP2VEC: a New Molecular Featurizer for Learning Molecular Properties. *Bioinformatics* 35 (23), 4979–4985. doi:10.1093/bioinformatics/btz307
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The Large-Scale Organization of Metabolic Networks. *Nature* 407 (6804), 651–654. doi:10.1038/35036627
- Jesus, T. F., Ribeiro-Gonçalves, B., Silva, D. N., Bortolaia, V., Ramirez, M., and Carriço, J. A. (2019). Plasmid ATLAS: Plasmid Visual Analytics and Identification in High-Throughput Sequencing Data. *Nucleic Acids Res.* 47 (D1), D188–D194. doi:10.1093/nar/gky1073
- Jhong, J. H., Yao, L., Pang, Y., Li, Z., Chung, C. R., Wang, R., et al. (2022). dbAMP 2.0: Updated Resource for Antimicrobial Peptides with an Enhanced Scanning Method for Genomic and Proteomic Data. *Nucleic Acids Res.* 50 (D1), D460–D470. doi:10.1093/nar/gkab1080
- Jorge, P., Alves, D., and Pereira, M. O. (2019). Catalysing the Way towards Antimicrobial Effectiveness: A Systematic Analysis and a New Online Resource for Antimicrobial-Enzyme Combinations against *Pseudomonas aeruginosa* and *Staphylococcus aureus*. *Int. J. Antimicrob. Agents* 53 (5), 598–605. doi:10.1016/j.ijantimicag.2019.01.001
- Jorge, P., Pérez-Pérez, M., Pérez Rodríguez, G., Fdez-Riverola, F., Olivia Pereira, M., and Lourenço, A. (2016). Reconstruction of the Network of Experimentally Validated AMP-Drug Combinations against *Pseudomonas aeruginosa* Infections. *Cbio* 11 (5), 523–530. doi:10.2174/1574893611666160617093955
- Kang, X., Dong, F., Shi, C., Liu, S., Sun, J., Chen, J., et al. (2019). DRAMP 2.0, an Updated Data Repository of Antimicrobial Peptides. *Sci. Data* 6 (148), 1–10. doi:10.1038/s41597-019-0154-y
- Karakoc, E., Cherkasov, A., and Sahinalp, S. C. (2007). Novel Approaches for Small Biomolecule Classification and Structural Similarity Search. *SIGKDD Explor. Newsl.* 9 (1), 14–21. doi:10.1145/1294301.1294307
- Khosravian, M., Faramarzi, F. K., Beigi, M. M., Behbahani, M., and Mohabatkari, H. (2013). Predicting Antibacterial Peptides by the Concept of Chou's Pseudo-amino Acid Composition and Machine Learning Methods. *Protein Pept. Lett.* 20 (2), 180–186. doi:10.2174/092986613804725307
- Korbee, C. J., Heemskerk, M. T., Kocev, D., van Strijen, E., Rabiee, O., Franken, K. L. M. C., et al. (2018). Combined Chemical Genetics and Data-Driven Bioinformatics Approach Identifies Receptor Tyrosine Kinase Inhibitors as Host-Directed Antimicrobials. *Nat. Commun.* 9 (358), 1–14. doi:10.1038/s41467-017-02777-6
- Laxminarayanan, R., Van Boeckel, T., Frost, I., Kariuki, S., Khan, E. A., Limmathurotsakul, D., et al. (2020). The Lancet Infectious Diseases Commission on Antimicrobial Resistance: 6 Years Later. *Lancet Infect. Dis.* 20 (4), e51–e60. doi:10.1016/S1473-3099(20)30003-7
- Lee, E. Y., Lee, M. W., Fulan, B. M., Ferguson, A. L., and Wong, G. C. L. (2017). What Can Machine Learning Do for Antimicrobial Peptides, and what Can Antimicrobial Peptides Do for Machine Learning? *Interf. Focus* 7 (6), 20160153. doi:10.1098/rsfs.2016.0153
- Lee, E. Y., Wong, G. C. L., and Ferguson, A. L. (2018). Machine Learning-Enabled Discovery and Design of Membrane-Active Peptides. *Bioorg. Med. Chem.* 26 (10), 2708–2718. doi:10.1016/j.bmc.2017.07.012
- Lee, J. W., Maria-Solano, M. A., Vu, T. N. L., Yoon, S., and Choi, S. (2022). Big Data and Artificial Intelligence (AI) Methodologies for Computer-Aided Drug Design (CADD). *Biochem. Soc. Trans.* 50 (1), 241–252. doi:10.1042/bst20211240
- Li, W. X., Tong, X., Yang, P. P., Zheng, Y., Liang, J. H., Li, G. H., et al. (2021). Screening of Antibacterial Compounds with Novel Structure from the FDA Approved Drugs Using Machine Learning Methods.
- Liu, B., and Pop, M. (2009). ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* 37 (Suppl. 1_1), D443–D447. doi:10.1093/nar/gkn656
- Macesic, N., Polubriaginof, F., and Tatonetti, N. P. (2017). Machine Learning: Novel Bioinformatics Approaches for Combating Antimicrobial Resistance. *Curr. Opin. Infect. Dis.* 30 (6), 511–517. doi:10.1097/QCO.0000000000000406
- Maltarollo, V. G., Kronenberger, T., Espinoza, G. Z., Oliveira, P. R., and Honorio, K. M. (2019). Advances with Support Vector Machines for Novel Drug Discovery. *Expert Opin. Drug Discov.* 14 (1), 23–33. doi:10.1080/17460441.2019.1549033
- Maltarollo, V. G. (2019). Classification of *Staphylococcus aureus* FabI Inhibitors by Machine Learning Techniques. *Int. J. Quantitative Structure-Property Relationships (Ijqspr)* 4 (4), 1–14. doi:10.4018/ijqspr.2019100101
- Mansbach, R. A., Leus, I. V., Mehla, J., Lopez, C. A., Walker, J. K., Rybenkov, V. V., Hengartner, N. W., Zgurskaya, H. I., and Gnanakaran, S. (2020). Machine Learning Algorithm Identifies an Antibiotic Vocabulary for Permeating Gram-Negative Bacteria. *J. Chem. Inf. Model.* 60 (6), 2838–2847. doi:10.1021/acs.jcim.0c00352
- Masalha, M., Rayan, M., Adawi, A., Abdallah, Z., and Rayan, A. (2018). Capturing Antibacterial Natural Products with In Silico Techniques. *Mol. Med. Rep.* 18 (1), 763–770. doi:10.3892/mmr.2018.9027
- Matamoros-Recio, A., Franco-Gonzalez, J. F., Forgione, R. E., Torres-Mozas, A., Silipo, A., and Martín-Santamaría, S. (2021). Understanding the Antibacterial Resistance: Computational Explorations in Bacterial Membranes. *ACS omega* 6 (9), 6041–6054. doi:10.1021/acsomega.0c05590
- McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The Comprehensive Antibiotic Resistance Database. *Antimicrob. Agents Chemother.* 57 (7), 3348–3357. doi:10.1128/AAC.00419-13
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2019). ChEMBL: towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* 47 (D1), D930–D940. doi:10.1093/nar/gky1075
- Motamedi, F., Pérez-Sánchez, H., Mehridehnavi, A., Fassihi, A., and Ghasemi, F. (2022). Accelerating Big Data Analysis through LASSO-Random Forest Algorithm in QSAR Studies. *Bioinformatics* 38 (2), 469–475. doi:10.1093/bioinformatics/btab659
- Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Recurrent Neural Network Model for Constructive Peptide Design. *J. Chem. Inf. Model.* 58 (2), 472–479.
- Naas, T., Oueslati, S., Bonnin, R. A., Dabos, M. L., Zavala, A., Dortet, L., et al. (2017). Beta-lactamase Database (BLDB) - Structure and Function. *J. Enzyme Inhib. Med. Chem.* 32 (1), 917–919. doi:10.1080/14756366.2017.1344235

- Nava Lara, R. A., Aguilera-Mendoza, L., Brizuela, C. A., Peña, A., and Del Rio, G. (2019). Heterologous Machine Learning for the Identification of Antimicrobial Activity in Human-Targeted Drugs. *Molecules* 24 (7), 1258. doi:10.3390/molecules24071258
- Ndagi, U., Falaki, A. A., Abdullahi, M., Lawal, M. M., and Soliman, M. E. (2020). Antibiotic Resistance: Bioinformatics-Based Understanding as a Functional Strategy for Drug Design. *RSC Adv.* 10 (31), 18451–18468. doi:10.1039/d0ra01484b
- Nocedo-Mena, D., Cornelio, C., Camacho-Corona, M. D. R., Garza-González, E., Waksman de Torres, N., Arrasate, S., et al. (2019). Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *J. Chem. Inf. Model.* 59 (3), 1109–1120. doi:10.1021/acs.jcim.9b00034
- Öztürk, H., Ozkirimli, E., and Özgür, A. (2018). A Novel Methodology on Distributed Representations of Proteins Using Their Interacting Ligands. *Bioinformatics* 34 (13), i295–i303.
- Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., and Larsson, D. G. (2014). BacMet: Antibacterial Biocide and Metal Resistance Genes Database. *Nucleic Acids Res.* 42 (D1), D737–D743. doi:10.1093/nar/gkt1252
- Patel, L., Shukla, T., Huang, X., Ussery, D. W., and Wang, S. (2020). Machine Learning Methods in Drug Discovery. *Molecules* 25 (22), 5277. doi:10.3390/molecules25225277
- Pires, D. E. V., and Ascher, D. B. (2020). mycoCSM: Using Graph-Based Signatures to Identify Safe Potent Hits against Mycobacteria. *J. Chem. Inf. Model.* 60 (7), 3450–3456. doi:10.1021/acs.jcim.0c00362
- Pirtskhalava, M., Gabrielian, A., Cruz, P., Griggs, H. L., Squires, R. B., Hurt, D. E., et al. (2016). DBAASP v.2: an Enhanced Database of Structure and Antimicrobial/Cytotoxic Activity of Natural and Synthetic Peptides. *Nucleic Acids Res.* 44 (D1), 6503–D1112. doi:10.1093/nar/gkw243
- Pushkaran, A. C., Biswas, R., and Mohan, C. G. (2019). “Impact of Target-Based Drug Design in Anti-bacterial Drug Discovery for the Treatment of Tuberculosis,” in *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process* (Cham: Springer), 307–346. doi:10.1007/978-3-030-05282-9_10
- Qureshi, A., Thakur, N., Tandon, H., and Kumar, M. (2014). AVDPdb: a Database of Experimentally Validated Antiviral Peptides Targeting Medically Important Viruses. *Nucleic Acids Res.* 42 (D1), D1147–D1153. doi:10.1093/nar/gkt1191
- Rajasekhar, S., Karupphasamy, R., and Chanda, K. (2021). Exploration of Potential Inhibitors for Tuberculosis via Structure-Based Drug Design, Molecular Docking, and Molecular Dynamics Simulation Studies. *J. Comput. Chem.* 42 (24), 1736–1749. doi:10.1002/jcc.26712
- Rayan, A., Marcus, D., and Goldblum, A. (2010). Predicting Oral Druglikeness by Iterative Stochastic Elimination. *J. Chem. Inf. Model.* 50 (3), 437–445. doi:10.1021/ci9004354
- Reimer, L. C., Vetcinova, A., Carbasse, J. S., Söhngen, C., Gleim, D., Ebeling, C., et al. (2019). BacDive in 2019: Bacterial Phenotypic Data for High-Throughput Biodiversity Analysis. *Nucleic Acids Res.* 47 (D1), D631–D636. doi:10.1093/nar/gky879
- Rodrigues, T., Reker, D., Schneider, P., and Schneider, G. (2016). Counting on Natural Products for Drug Design. *Nat. Chem.* 8 (6), 531–541. doi:10.1038/nchem.2479
- Romero-Molina, S., Ruiz-Blanco, Y. B., Harms, M., Münch, J., and Sanchez-Garcia, E. (2019). PPI-detect: A Support Vector Machine Model for Sequence-based Prediction of Protein–Protein Interactions. *J. Comput. Chem.* 40 (11), 1233–1242.
- Schroedl, S. (2019). Current Methods and Challenges for Deep Learning in Drug Discovery. *Drug Discov. Today Technol.* 32–33, 9–17. doi:10.1016/j.ddtec.2020.07.003
- Seebah, S., Suresh, A., Zhuo, S., Choong, Y. H., Chua, H., Chuon, D., et al. (2007). Defensins Knowledgebase: a Manually Curated Database and Information Source Focused on the Defensins Family of Antimicrobial Peptides. *Nucleic Acids Res.* 35 (Suppl. 1), D265–D268. doi:10.1093/nar/gkl866
- Serafim, M. S. M., Kronenberger, T., Oliveira, P. R., Poso, A., Honório, K. M., Mota, B. E. F., et al. (2020). The Application of Machine Learning Techniques to Innovative Antibacterial Discovery and Development. *Expert Opin. Drug Discov.* 15 (10), 1165–1180. doi:10.1080/17460441.2020.1776696
- Shi, C., Dong, F., Zhao, G., Zhu, N., Lao, X., and Zheng, H. (2020). Applications of Machine-Learning Methods for the Discovery of NDM-1 Inhibitors. *Chem. Biol. Drug Des.* 96 (5), 1232–1243. doi:10.1111/cbdd.13708
- Sripriya Akondi, V., Menon, V., Baudry, J., and Whittle, J. (2022). Novel Big Data-Driven Machine Learning Models for Drug Discovery Application. *Molecules* 27 (3), 594. doi:10.3390/molecules27030594
- Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell* 180 (4), 688–e13. doi:10.1016/j.cell.2020.01.021
- Suay-Garcia, B., Falcó, A., Bueso-Bordils, J. I., Anton-Fos, G. M., Pérez-Gracia, M. T., and Alemán-López, P. A. (2020). Tree-based QSAR Model for Drug Repurposing in the Discovery of New Antibacterial Compounds against *Escherichia coli*. *Pharmaceuticals* 13 (12), 431. doi:10.3390/ph13120431
- Szaboova, A., Kuželka, O., and Železný, F. (2012). Prediction of Antimicrobial Activity of Peptides Using Relational Machine Learning In IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE, 575–580.
- Thomas, J., Navre, M., Rubio, A., and Coukell, A. (2018). Shared Platform for Antibiotic Research and Knowledge: a Collaborative Tool to SPARK Antibiotic Discovery. *ACS Infect. Dis.* 4 (11), 1536–1539. doi:10.1021/acsinfecdis.8b00193
- Tian, S., Wang, J., Li, Y., Xu, X., and Hou, T. (2012). Drug-likeness Analysis of Traditional Chinese Medicines: Prediction of Drug-Likeness Using Machine Learning Approaches. *Mol. Pharm.* 9 (10), 2875–2886. doi:10.1021/mp300198d
- van Heel, A. J., de Jong, A., Song, C., Viel, J. H., Kok, J., and Kuipers, O. P. (2018). BAGEL4: a User-Friendly Web Server to Thoroughly Mine RiPPs and Bacteriocins. *Nucleic Acids Res.* 46 (W1), W278–W281. doi:10.1093/nar/gky383
- Van Oort, C. M., Ferrell, J. B., Remington, J. M., Wshah, S., and Li, J. (2021). AMPGAN V2: Machine Learning-Guided Design of Antimicrobial Peptides. *J. Chem. Inf. Model.* 61 (5), 2198–2207. doi:10.1021/acs.jcim.0c01441
- Veltri, D., Kamath, U., and Shehu, A. (2018). Deep Learning Improves Antimicrobial Peptide Recognition. *Bioinformatics* 34 (16), 2740–2747. doi:10.1093/bioinformatics/bty179
- Veltri, D., Kamath, U., and Shehu, A. (2015). Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming. *Ieee/acm Trans. Comput. Biol. Bioinform* 14 (2), 300–313. doi:10.1109/TCBB.2015.2462364
- Vila, J., Moreno-Morales, J., and Ballesté-Delpierre, C. (2020). Current Landscape in the Discovery of Novel Antibacterial Agents. *Clin. Microbiol. Infect.* 26 (5), 596–603. doi:10.1016/j.cmi.2019.09.015
- Waghu, F. H., Barai, R. S., Gurung, P., and Idicula-Thomas, S. (2016). CAMPR3: a Database on Sequences, Structures and Signatures of Antimicrobial Peptides. *Nucleic Acids Res.* 44 (D1), D1094–D1097. doi:10.1093/nar/gkv1051
- Waghu, F. H., Gopi, L., Barai, R. S., Ramteke, P., Nizami, B., and Idicula-Thomas, S. (2014). CAMP: Collection of Sequences and Structures of Antimicrobial Peptides. *Nucleic Acids Res.* 42 (D1), D1154–D1158. doi:10.1093/nar/gkt1157
- Wang, C., Garlick, S., and Zloh, M. (2021). Deep Learning for Novel Antimicrobial Peptide Design. *Biomolecules* 11 (3), 471. doi:10.3390/biom11030471
- Wang, G., Li, X., and Wang, Z. (2016). APD3: the Antimicrobial Peptide Database as a Tool for Research and Education. *Nucleic Acids Res.* 44 (D1), D1087–D1093. doi:10.1093/nar/gkv1278
- Wani, M. A., Garg, P., and Roy, K. K. (2021). Machine Learning-Enabled Predictive Modeling to Precisely Identify the Antimicrobial Peptides. *Med. Biol. Eng. Comput.* 59 (11), 2397–2408. doi:10.1007/s11517-021-02443-6
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–D1082. doi:10.1093/nar/gkx1037
- Wu, X., Wang, Z., Li, X., Fan, Y., He, G., Wan, Y., Yu, C., Tang, J., Li, M., Zhang, X., Zhang, H., Xiang, R., Pan, Y., Liu, Y., Lu, L., and Yang, L. (2014). *In Vitro* and *In Vivo* Activities of Antimicrobial Peptides Developed Using an Amino Acid-Based Activity Prediction Method. *Antimicrob. Agents Chemother.* 58 (9), 5342–5349. doi:10.1128/AAC.02823-14
- Yang, X. G., Chen, D., Wang, M., Xue, Y., and Chen, Y. Z. (2009). Prediction of Antibacterial Compounds by Machine Learning Approaches. *J. Comput. Chem.* 30 (8), 1202–1211. doi:10.1002/jcc.21148
- Yates, A. D., Allen, J., Amode, R. M., Azov, A. G., Barba, M., Becerra, A., et al. (2022). Ensembl Genomes 2022: an Expanding Genome Resource for Non-vertebrates. *Nucleic Acids Res.* 50 (D1), D996–D1003. doi:10.1093/nar/gkab1007
- Yoshida, M., Hinkley, T., Tsuda, S., Abul-Hajja, Y. M., McBurney, R. T., Kulikov, V., et al. (2018). Using Evolutionary Algorithms and Machine Learning to

- Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem* 4 (3), 533–543. doi:10.1016/j.chempr.2018.01.005
- Zhang, R., Li, X., Zhang, X., Qin, H., and Xiao, W. (2021). Machine Learning Approaches for Elucidating the Biological Effects of Natural Products. *Nat. Prod. Rep.* 38 (2), 346–361. doi:10.1039/d0np00043d
- Zhang, R., Wang, Z., Tian, Y., Yin, Q., Cheng, X., Lian, M., Zhou, B., Zhang, X., and Yang, L. (2019). Efficacy of Antimicrobial Peptide DP7, Designed by Machine-Learning Method, against Methicillin-Resistant *Staphylococcus aureus*. *Front Microbiol.* 10, 1175. doi:10.3389/fmicb.2019.01175
- Zhao, L., Ciallella, H. L., Aleksunes, L. M., and Zhu, H. (2020). Advancing Computer-Aided Drug Discovery (CADD) by Big Data and Data-Driven Machine Learning Modeling. *Drug Discov. Today* 25 (9), 1624–1638. doi:10.1016/j.drudis.2020.07.005
- Zhu, H. (2020). Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annu. Rev. Pharmacol. Toxicol.* 60, 573–589. doi:10.1146/annurev-pharmtox-010919-023324

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Jukič and Bren. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Structure-Based Design of 2-Aminopurine Derivatives as CDK2 Inhibitors for Triple-Negative Breast Cancer

Hanzhi Liang^{1†}, Yue Zhu^{1†}, Zhiyuan Zhao^{1†}, Jintong Du², Xinying Yang¹, Hao Fang^{1*} and Xuben Hou^{1*}

¹Key Laboratory of Chemical Biology (Ministry of Education), School of Pharmaceutical Science, Cheeloo College of Medicine, Shandong University, Ji'nan, China, ²Shandong Cancer Hospital and Institute, Shandong First Medical University, Jinan, China

OPEN ACCESS

Edited by:

Adriano D. Andricopulo,
University of Sao Paulo, Brazil

Reviewed by:

Wade Russu,
The University of the Pacific,
United States
Chung-Hang Leung,
University of Macau, China

*Correspondence:

Xuben Hou
hxb@sdu.edu.cn
Hao Fang
haofangcn@sdu.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 28 January 2022

Accepted: 24 March 2022

Published: 03 May 2022

Citation:

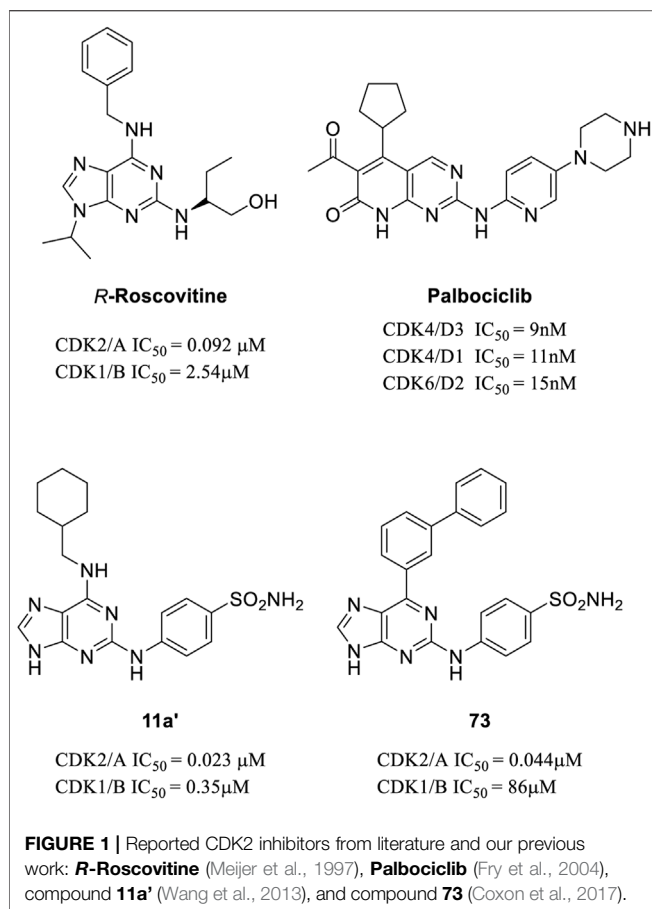
Liang H, Zhu Y, Zhao Z, Du J, Yang X,
Fang H and Hou X (2022) Structure-
Based Design of 2-Aminopurine
Derivatives as CDK2 Inhibitors for
Triple-Negative Breast Cancer.
Front. Pharmacol. 13:864342.
doi: 10.3389/fphar.2022.864342

Cyclin-dependent kinase 2 (CDK2) regulates the progression of the cell cycle and is critically associated with tumor growth. Selective CDK2 inhibition provides a potential therapeutic benefit against certain tumors. Purines and related heterocycle (e.g., *R*-Roscovitine) are important scaffolds in the development of CDK inhibitors. Herein, we designed a new series of 2-aminopurine derivatives based on the fragment-centric pocket mapping analysis of CDK2 crystal structure. Our results indicated that the introduction of polar substitution at the C-6 position of purine would be beneficial for CDK2 inhibition. Among them, compound **111** showed good CDK2 inhibitory activity (IC₅₀ = 19 nM) and possessed good selectivity against other CDKs. Further *in vitro* tests indicated that compound **111** possesses anti-proliferation activity in triple-negative breast cancer (TNBC) cells. Moreover, molecular dynamics simulation suggested the favorable binding mode of compound **111**, which may serve as a new lead compound for the future development of CDK2 selective inhibitors.

Keywords: structure-based drug design, CDK2 inhibitor, purine, anticancer, triple-negative breast cancer

1 INTRODUCTION

Cyclin-dependent kinases (CDKs) are essential kinases that drive cell cycle transformation and transcriptional regulation (Wood and Endicott, 2018; Rice, 2019). CDKs involve in a variety of biological processes, including cell metabolism, differentiation, and development. Human CDKs are mainly divided into two categories: 1) One group is involved in cell cycle regulation and related to mitosis, and the subtypes are CDK1, 2, 3, 4, and 6. 2) Another group is mainly involved in transcriptional regulation, regulating phosphorylation of RNA polymerase II, and the subtypes are CDK7, 8, 9, and 11 (Satyanarayana and Kaldis, 2009). Other subtypes, such as CDK5, have long been thought to be neuron-specific kinases that play an important role in cellular activity (survival, motility, etc.) (Roufayel and Murshid, 2019). Heretofore, several CDK4/6 inhibitors (e.g., Palbociclib (Fry et al., 2004), Ribociclib (Tripathy et al., 2017), and Abemaciclib (Lee et al., 2019)) have been approved by the FDA for the treatment of breast cancer and other solid tumors (Yuan et al., 2021). However, the long-term use of CDK4/6 inhibitors results in drug resistance and poor therapeutic effect on Rb-deficient tumors, especially some malignant tumors, which limits the clinical application of CDK4/6 inhibitors (Braul et al., 2021; Gomatou et al., 2021; Julve et al., 2021).



CDK2 plays a key role in cell cycle regulation (Aleem et al., 2004; Tadesse et al., 2019a; Volkart et al., 2019). CDK2 forms a complex with Cyclin E and then phosphorylates Rb, therefore activates E2F (Narasimha et al., 2014). CDK2-Cyclin A complex promotes cells to pass through the S/G2 checkpoint (Kimball and Webster, 2001). CDK2 also controls the phosphorylation of many transcription factors including Smad3 (Liu, 2006), FoxM1, FoxO1 (Adams, 2001), NFY, B-Myb (Joaquin and Watson, 2003), Myc (Hydbring and Larsson, 2010) and promotes the cell cycle. In addition, CDK2 also plays an important role in DNA replication (Fagundes and Teixeira, 2021), adaptive immune response, cell differentiation (Adams, 2001), and apoptosis (Golsteyn, 2005; Satyanarayana and Kaldis, 2009). CDK2 is an important regulatory factor of various carcinogenic signaling pathways (Jin et al., 2020). The overexpression of CDK2 and its related Cyclin A or Cyclin E is closely related to the development of tumors (Sviderskiy et al., 2020). Especially, the inhibition of CDK2 is a potential therapeutic strategy for those tumors that are considered to be ineffective by CDK4/6 inhibitors (Pandey et al., 2019; Tadesse et al., 2020). Inhibition of CDK2 resulted in increased Smad3 activity and decreased triple-negative breast cancer (TNBC) cell migration (Tarasewicz et al., 2014). Recently, CDK2 has been found to mediate phosphorylation of EZH2, which drives tumorigenesis of TNBC (Nie et al., 2019a). Nowadays, CDK2 has been recognized as a potential target for

anticancer drug development (Chohan et al., 2015; Zhang et al., 2015; Sánchez-Martínez et al., 2019).

Previously, we have reported a series of purine-2,6-diamine derivatives as CDK2 selective inhibitors (**Figure 1**) (Wang et al., 2013). We also developed purine-8-one derivatives that displayed good antitumor activities (Lu et al., 2019). In 2016, Coxon et al. designed a series of 6-substituted 2-arylamino purines, which also possessed good CDK2 selectivity (**Figure 1**) (Coxon et al., 2017). Based on the crystal structures of the CDK2-inhibitor **73** complex (PDB: 5NEV), we performed fragment-centric topographic mapping using AlphaSpace and analyzed the binding pocket of **73**. As shown in **Figure 2**, we identified an unoccupied polar pocket (pocket 5) besides the biphenyl group of **73**. Moreover, the partially polar binding pocket (pocket 2, nonpolar rate = 73%) for the biphenyl group is also not fully occupied (occupancy = 79%). Based on the structural analysis above, we designed a new series of 2-arylamino purines by introducing various substitutions in the C-6 position of the purine scaffold to further explore the structure-activity relationships.

2 RESULTS AND DISCUSSION

2.1 Chemistry

The synthesis routes of compounds **5a-5k** are depicted in **Scheme 1**. The 6-substituted purine derivatives were synthesized from the THP-protected 2,6-dichloropurine via a Suzuki coupling reaction with aryl boric acid or aryl pinacol boric acid ester. Coupling by Buchwald-Harwing Reaction with 3-Nitroaniline, employing $Pd(OAc)_2/Xantphos$ afforded the THP-protected 2-aminopurine derivatives in excellent yield. Then the N9-THP group was removed under the acidic condition to give the final compound.

And the synthesis routes of compounds **11a-11r** are depicted in **Scheme 2**. Ortho- or para-bromo benzylamine were protected by the Boc group, respectively. Then, through the Miyaura borylation reaction, the Boc-protected aryl borate esters were prepared. And then, similar to the synthetic route of **Scheme 1**, compounds **10a-10r** were obtained by Suzuki coupling, Buchwald-Harwing Coupling (Yin et al., 2002), and the removal of protection groups.

2.2 CDK2 Inhibitory Activities

All compounds were screened for CDK2 inhibitory activities at 0.5 μ M. Compounds with inhibition rates higher than 50% were further tested at different concentrations to determine IC_{50} values. And results are summarized in **Table 1**. The 6-position benzene substituted purine derivative (**5a**) showed good potency against CDK2 (IC_{50} = 0.31 μ M). When the benzene ring at the C6 position of compound **5a** was changed to naphthalene ring (**5b**), pyrrole ring (**5c**), benzo[d][1,3]dioxole (**5d**), and thiophene (**5e**), the CDK2 inhibitory activity decreased (**Table 1**). Among these compounds, **5e** was inactive, **5b** and **5d** showed weak activity against CDK2, whereas **5c** exhibited a 37% inhibition rate at 0.5 μ M. Then, we sought to investigate the impacts of different substituted benzenes at the C-6 position. As shown in **Table 1**, the

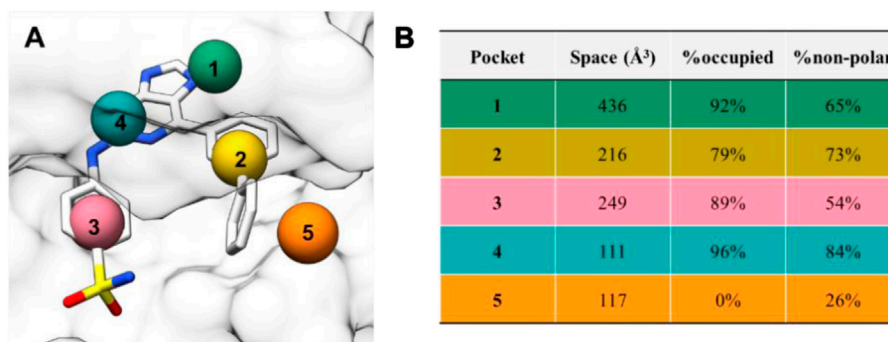
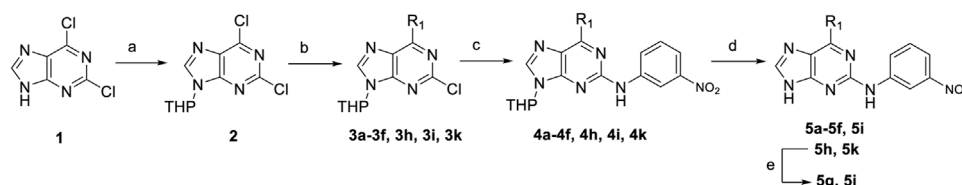
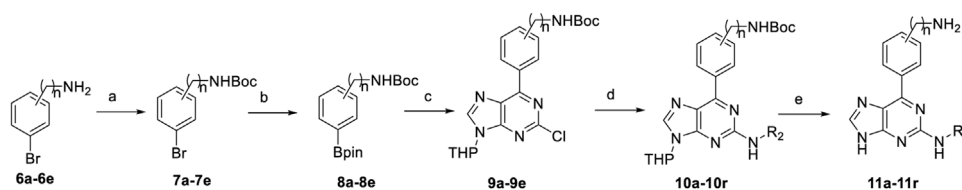


FIGURE 2 | (A) Calculated binding pockets of compound **73** in CDK2. Pockets are represented using spheres located at the centroid of each alpha-cluster. **(B)** The table presents pocket features including space, occupancy, and nonpolar rate.



SCHEME 1 | Synthetic route of target compounds **5a-5k**. Reagents and conditions: **(A)** 3,4-dihydro-2H-pyran, *DL*-Camphorsulfonic acid, EA, 65°C, 18 h; **(B)** aryl borate ester, Pd(PPh₃)₄, K₂CO₃, 1,4-dioxane/H₂O = 4:1, 80°C, 9 h; **(C)** 3-Nitroaniline, Pd(OAc)₂, Xantphos, Cs₂CO₃, 1,4-dioxane, 100°C, 9 h; **(D)** HCl/EA, rt, 4 h; **(E)** LiOH, THF/H₂O = 4:1, rt, 4 h.



SCHEME 2 | Synthetic route of target compounds **11a-11r**. Reagents and conditions: **(A)** (Boc)₂O, K₂CO₃, DCM, rt, 4 h; **(B)** Bis(pinacolato)diboron, Pd(dppf)₂Cl₂, KOAc, DMSO, 80°C, 9 h; **(C)** **2**, Pd(PPh₃)₄, K₂CO₃, 1,4-dioxane:H₂O = 4:1, 80°C, 9 h; **(D)** substituted anilines, Pd(OAc)₂, Xantphos, Cs₂CO₃, 1,4-dioxane, 100°C, 9 h; **(E)** HCl/EA, rt, 4 h.

introduction of methyl formate (**5h** and **5k**), fluorine (**5i**), or nitro (**5f**) substituted the benzene-abolished CDK2 inhibitory activity. Interestingly, the meta-substituted carboxylic group (**5g**) is beneficial for CDK2 inhibition, whereas the para-substituted carboxylic group leads to a compound with low activity (**5j**). The introduction of phenylamino or benzylamine group (**11a-11d**) in the C-6 position increases the CDK2 inhibitory activity. Compound **11a** (IC₅₀ = 0.31 μM) exhibited a similar CDK2 inhibitory activity with **5a**. Importantly, compound **11c** (IC₅₀ = 0.11 μM) possessed a better CDK2 inhibitory activity than **5a**. When the para-amino group was changed to meta-amino group (**11d**), its CDK2 inhibitory activity was decreased slightly (IC₅₀ = 0.23 μM). The above SAR result is consistent with our hypothesis that the introduction of a polar group at the C-6 site would be beneficial for binding against CDK2 protein. Next, we sought to optimize the R₂ substitutions and got compound

11f-11r. When we introduced different substituents to the benzene ring, such as the electron-donating methyl group (**11i**) and tert-butyl group (**11j**), the activity decreased obviously. The biphenyl group (**11h**) seems to be too bulky to occupy the active site and cause a decrease in activity. The introduction of fluorine (**11n** and **11o**), sulfonamide groups (**11l**, **11p**, and **11q**), and pyridine group (**11r**) is beneficial for CDK2 inhibition, and compound **11l** exhibited the best activity (IC₅₀ = 0.019 μM).

2.3 Isoform Selectivity

Three potent CDK2 inhibitors (**11c**, **11l**, and **11p**) were further selected to evaluate their inhibitory activities against other CDKs isoforms. As shown in **Table 2**, compounds **11c**, **11l**, and **11p** showed potent activity against CDK1 (IC₅₀ = 0.12–0.24 μM), weak activity against CDK6 (IC₅₀ = 2.2–4.8 μM) and are nearly

TABLE 1 | The inhibitory activities of compounds **5a-5k** and **11a-11r** against CDK2.

| Compound | R ₁ | R ₂ | IC ₅₀ ^a (μM) or inhibition rate (%) @0.5 μM |
|--------------------|--|--|---|
| 5a | Ph- | 3-NO ₂ -Ph- | 0.31 ± 0.01 |
| 5b | naphthyl | 3-NO ₂ -Ph- | 3% |
| 5c | pyrrole-2-yl | 3-NO ₂ -Ph- | 37% |
| 5d | 4-benzo[d][1,3]dioxole | 3-NO ₂ -Ph- | 7% |
| 5e | thiophene-1-yl | 3-NO ₂ -Ph- | NA |
| 5f | 3-NO ₂ -Ph- | 3-NO ₂ -Ph- | NA |
| 5g | 3-COOH-Ph- | 3-NO ₂ -Ph- | 0.15 ± 0.01 |
| 5h | 3-COOCH ₃ -Ph- | 3-NO ₂ -Ph- | 11% |
| 5i | 4-F-Ph- | 3-NO ₂ -Ph- | NA |
| 5j | 4-COOH-Ph- | 3-NO ₂ -Ph- | 7% |
| 5k | 4-COOCH ₃ -Ph- | 3-NO ₂ -Ph- | 3% |
| 11a | 3-NH ₂ -Ph- | 3-NO ₂ -Ph- | 0.31 ± 0.02 |
| 11b | 4-NH ₂ -Ph- | 3-NO ₂ -Ph- | 40% |
| 11c | 3-NH ₂ -Bn- | 3-NO ₂ -Ph- | 0.11 ± 0.01 |
| 11d | 4-NH ₂ -Bn- | 3-NO ₂ -Ph- | 0.23 ± 0.01 |
| 11e | 3-CH ₂ NH ₂ -Bn- | 3-NO ₂ -Ph- | 35% |
| 11f | 3-NH ₂ -Bn- | Ph- | 35% |
| 11g | 4-NH ₂ -Bn- | Ph- | 0.13 ± 0.02 |
| 11h | 3-NH ₂ -Bn- | biphenyl | 13% |
| 11i | 3-NH ₂ -Bn- | 4-Me-Ph- | 0.28 ± 0.02 |
| 11j | 3-NH ₂ -Bn- | 4-t-Bu-Ph- | 28% |
| 11k | 3-NH ₂ -Bn- | 4-piperazine-1-yl-Ph- | 26% |
| 11l | 3-NH ₂ -Bn- | 4-SO ₂ NH ₂ -Ph- | 0.019 ± 0.001 |
| 11m | 3-NH ₂ -Bn- | 3-NH ₂ -Ph- | 33% |
| 11n | 3-NH ₂ -Bn- | 4-F-Ph- | 0.32 ± 0.06 |
| 11o | 4-NH ₂ -Bn- | 4-F-Ph- | 0.24 ± 0.01 |
| 11p | 3-NH ₂ -Bn- | 4-SO ₂ N(Me)H-Ph- | 0.032 ± 0.001 |
| 11q | 3-NH ₂ -Bn- | 4-SO ₂ N(Me)H-Ph- | 0.18 ± 0.02 |
| 11r | 3-NH ₂ -Bn- | pyridin-3-yl | 0.19 ± 0.01 |
| Roscovitine | - | - | 0.073 ± 0.022 |

^aValues are geometric means of n ≥ 3 experiments, with a range of less than 20% of the mean value.

TABLE 2 | Inhibitory activity of selected compounds against different CDKs.

| Compound | CDK2/cyclin A | CDK1/cyclin B | CDK6/cyclin D3 | CDK8/cyclin C |
|------------|-----------------------|-----------------------|-----------------------|---------------------------|
| | IC ₅₀ (μM) | IC ₅₀ (μM) | IC ₅₀ (μM) | Inhibition rate @5 μM (%) |
| 11c | 0.117 ± 0.01 | 0.24 ± 0.04 | 2.2 ± 0.1 | 19 |
| 11l | 0.019 ± 0.01 | 0.12 ± 0.02 | 2.7 ± 0.5 | 11 |
| 11p | 0.032 ± 0.01 | 0.15 ± 0.02 | 4.8 ± 0.1 | 15 |

inactive against CDK8 (inhibition rate <20% @ 5 μM). Compounds **11l** and **11p** possess good selectivity for CDK2 over CDK6 and CDK8 (more than 140-fold), whereas their selectivity against CDK1 is lower (4.6 to 6.3-fold). Compared with compounds **11l** and **11p**, compound **11c** is a less selective CDK2 inhibitor (2-fold for CDK1, 18-fold for CDK6, more than 42-fold for CDK8). Taking the above results together, our newly designed compound **11l** is a potent and selective CDK2 inhibitor.

2.4 Anti-Triple-Negative Breast Cancer Activity of Selected Compounds

Previous studies have proved that CDK2 plays a critical role in breast cancer progression by phosphorylating and activating hormone receptors (Pierson-Mullany and Lange, 2004; Tadesse et al.,

2019b). In triple-negative breast cancer (TNBC), inhibition of CDK2 has shown synergistic effects with chemotherapy and radiotherapy (Deans et al., 2006; Rao et al., 2017; Nie et al., 2019b; Zhu et al., 2022). In the current study, we further investigated the antitumor activity of three compounds using MDA-MB-231 cells, which were derived from TNBC patients. As shown in **Figure 3A**, compounds **11c**, **11l**, and **11p** (IC₅₀ = 8.11–15.66 μM) exhibited better anti-proliferation activities than R-Roscovitine (IC₅₀ = 24.07 μM) in MDA-MB-231 cells. Furthermore, we also evaluated the cytotoxicity of compound **11l** in human embryonic kidney cell (293T) using the MTT assay. This compound showed low cytotoxicity with an IC₅₀ value higher than 100 μM.

To explore the mechanism of action of our newly designed compound, we further investigated their effects on the cell cycle regulation. As shown in **Figure 3B**, treatment of compounds **11c**,

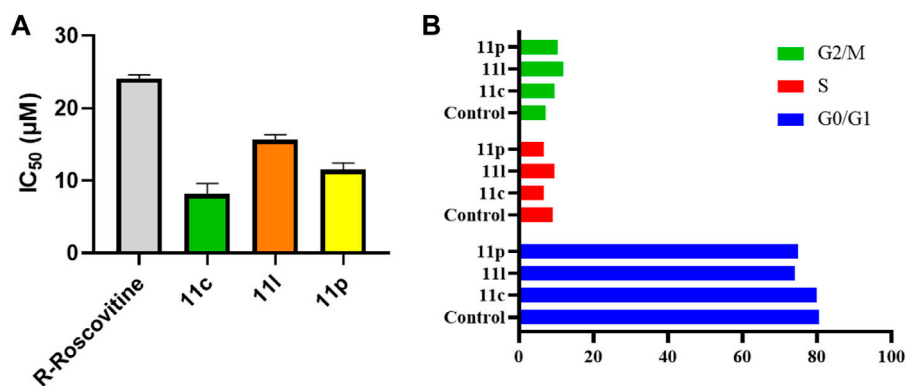


FIGURE 3 | (A) Anti-proliferation activities of compounds **11c**, **11l**, and **11p** against MDA-MB-231 cells. *R*-Roscovitine was employed as the positive control. **(B)** Impacts of compounds **11c**, **11l**, and **11p** on the cell cycle of MDA-MB-231 cells.

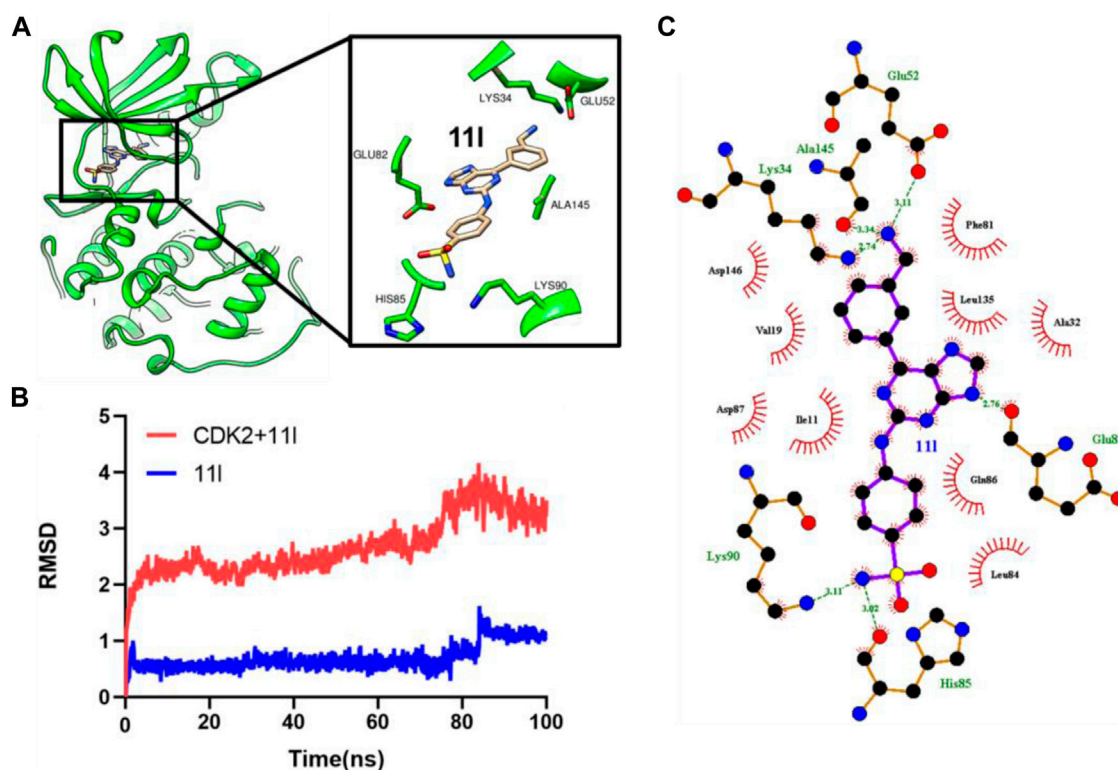


FIGURE 4 | (A) Predicted binding mode of **11l** in CDK2 from MD simulation. The key residues of CDK2 are highlighted and colored in green. **(B)** RMSD values of protein–ligand complex and **11l** during MD simulation. **(C)** Interacting residues between **11l** and CDK2.

11l, and **11p** increased the percentage of cells in the G2/M phase, compared with the negative control group. The results above suggested that our newly designed CDK2 inhibitors are potential antitumor agents for the treatment of TNBC.

2.5 Molecular Dynamics Simulation

To further decipher the binding mode of **11l**, we performed 100 ns molecular dynamics (MD) simulation based on the docking result. As shown in **Figure 4**, the RMSD values of the

protein–ligand complex are within 4 Å, while the RMSD values of compound **11l** are within 1.5 Å, indicating that the simulation system is stable during MD simulation. Then we extracted the representative binding mode from the MD trajectory and analyzed the key interacting residues. As shown in **Figure 4C**, compound **11l** forms multiple hydrogen bond interactions with surrounding residues in CDK2. The sulfonamide group forms hydrogen bonds with the side chain and backbone nitrogen of Lys90 and His85. The benzyl amine

TABLE 3 | Statistical analysis of the hydrogen bond interactions between **11l** and CDK2 during MD simulation.

| Donor | Acceptor | Occupancy (%) | Distance (Å) |
|------------|----------|---------------|--------------|
| 11l@N5-H | His85@O | 0.4580 | 2.8481 |
| 11l @N6-H | Ala145@O | 0.2140 | 2.8607 |
| Lys34@NZ-H | 11l @N6 | 0.2120 | 2.8698 |
| Lys34@NZ-H | 11l @N6 | 0.1740 | 2.8698 |
| 11l @N5-H | His85@O | 0.1120 | 2.8600 |
| Lys34@NZ-H | 11l @N6 | 0.1040 | 2.8717 |
| 11l @N3-H | Leu84@O | 0.1020 | 2.8160 |

group of **11l** locates at the entrance of the ATP-binding pocket, and form hydrogen bonds with Lys34, Glu52, and Ala145, respectively. The key hydrogen bonds were listed in **Table 3**. The hydrogen bond between **His85/Lys34** and **11l** is the most stable hydrogen bond interaction occupation values of 0.57 and 0.4, respectively (**Table 3**). The results above revealed the most favorable binding mode as well as key interactions of compound **11l** with CDK2, which would be helpful for further structural optimization.

3 CONCLUSION

In the current study, we designed a series of 2-aminopurine derivatives as new CDK2 inhibitors based on the fragment-centric pocket mapping of crystal structure. As expected, the introduction of polar groups in the C-6 position of the purine scaffold is beneficial for CDK2 inhibition. Among them, compound **11l** ($IC_{50} = 0.019 \mu M$) exhibited higher CDK2 inhibitory activity against CDK2 than known inhibitor *R*-Roscovitine ($IC_{50} = 0.073 \mu M$). Moreover, **11l** also possessed good selectivity against other CDK isoforms and showed better anti-proliferation activity in MDA-MB-231 cells than *R*-Roscovitine. Molecular dynamics simulation further suggested the binding mode of **11l** with CDK2, which would be helpful for the future development of more potent and selective CDK2 inhibitors.

4 EXPERIMENTAL SECTION

4.1 Chemistry

Chemical reagents were purchased commercially and were used without further purification. All reactions with air- or moisture-sensitive reagents were carried out under nitrogen and solvents were also dried before use. Reactions were monitored by thin-layer chromatography with preparative silica gel GF254 plates (UV lamp. or iodine), and column chromatography was performed on silica gel. The 1H -NMR spectra were obtained at 400 MHz. For 1H -NMR spectra, chemical shifts were given in parts per million (ppm) and were referenced to tetramethylsilane (TMS) peak as an internal standard or the residual solvent peak. ^{13}C NMR spectra were recorded at 101 MHz. Chemical shifts were reported in ppm and were referenced to the appropriate residual solvent peak. Splitting patterns were designed as s,

singlet; d, doublet; t, triplet; m, multiplet. High-resolution mass spectrometry (HRMS) data were recorded with a 1200RRLC-6520 Accurate-Mass Q-TOF LC/MS system at the Shandong Analysis and Test Center.

4.1.1 Preparation of

2,6-dichloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purine (**2**)

2,6-dichloropurine (5.29 mmol) and *DL*-Camphorsulfonic acid (0.05 mmol) were dissolved in ethyl acetate (20 ml), and heated to 65°C. 3,4-2H-dihydropyran (5.29 mmol) was added slowly and then the reaction mixture was stirred for 18 h at 65°C. After the completion, the reaction mixture is poured into H₂O (20 ml), extracted twice with ethyl acetate (50 ml), washed with brine, and dried with anhydrous Mg₂SO₄. The crude product was concentrated and purified by silica gel chromatography to obtain compound **2**. White solid; Yield: 70%; m.p.: 93–95°C; 1H NMR (600 MHz, CDCl₃) δ 8.33 (s, 1H), 5.76 (dd, $J = 10.8$, 2.4 Hz, 1H), 4.21–4.11 (m, 1H), 3.78 (td, $J = 11.8$, 2.6 Hz, 1H), 2.21–2.15 (m, 1H), 2.13–2.06 (m, 1H), 2.02–1.93 (m, 1H), 1.87–1.72 (m, 2H), 1.71–1.66 (m, 1H).

4.1.2 General Method for the Preparation of Compounds **3a-3f**, **3h**, **3i**, **3k**, **9a-9e**

Tert-butyl (4-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl) carbamate (**9d**). Compounds **8d** (4.5 mmol), compound **2** (4.5 mmol), Pd (PPh₃)₄ (0.05 mmol), and K₂CO₃ (13.5 mmol) were mixed in a two-neck flask. Under the protection of N₂, the solution of 1,4-dioxane and water (4:1) was added and the mixture reacted at 80°C for 12 h. After the completion, the reaction mixture was filtered through a pad of Celite. Spinned the filtrate dry and then dissolved it with ethyl acetate (15 ml) and water (20 ml), extracted twice with ethyl acetate (50 ml), washed with brine, and dried with anhydrous Mg₂SO₄. The crude product was concentrated and purified by silica gel chromatography (eluting with petroleum ether/ethyl acetate 3/1 to 1/1) to obtain compound **9d**. White solid; Yield: 75%; m.p.: 175–177°C; 1H NMR (400 MHz, CDCl₃) δ 8.76 (d, $J = 7.9$ Hz, 2H), 8.32 (s, 1H), 7.46 (d, $J = 7.9$ Hz, 2H), 5.83 (d, $J = 10.4$ Hz, 1H), 4.92 (s, 1H), 4.41 (d, $J = 5.0$ Hz, 2H), 4.20 (d, $J = 11.3$ Hz, 1H), 3.81 (t, $J = 11.0$ Hz, 1H), 2.18 (d, $J = 12.4$ Hz, 1H), 2.08 (s, 1H), 1.99 (dd, $J = 22.9$, 11.4 Hz, 1H), 1.80 (td, $J = 22.9$, 12.2 Hz, 2H), 1.68 (d, $J = 9.8$ Hz, 1H), 1.48 (s, 9H).

Compounds **3a-3f**, **3h**, **3i**, **3k**, **9a-9c**, **9e** were synthesized following the procedure described above.

2-chloro-6-phenyl-9-(tetrahydro-2H-pyran-2-yl)-9H-purine (**3a**)

Light yellow solid; Yield: 95%; m.p.: 132–134°C; 1H NMR (600 MHz, CDCl₃) δ 8.86–8.73 (m, 2H), 8.32 (s, 1H), 7.60–7.50 (m, 2H), 5.84 (dd, $J = 10.8$, 2.4 Hz, 1H), 4.27–4.15 (m, 1H), 3.81 (td, $J = 11.8$, 2.5 Hz, 1H), 2.18 (dd, $J = 12.5$, 2.0 Hz, 1H), 2.13–2.06 (m, 1H), 2.00 (ddd, $J = 23.5$, 12.5, 4.0 Hz, 1H), 1.89–1.73 (m, 2H), 1.68 (d, $J = 12.1$ Hz, 1H).

2-chloro-6-(naphthalen-1-yl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purine (**3b**)

White solid; Yield: 67%; m.p.: 149–151°C; 1H NMR (400 MHz, CDCl₃) δ 8.31 (s, 1H), 8.30–8.24 (m, 1H), 8.03 (t, $J = 7.6$ Hz, 2H), 7.97–7.89 (m, 1H), 7.66–7.60 (m, 1H), 7.57–7.49 (m, 2H), 5.88 (d, $J = 10.6$ Hz, 1H), 4.21 (d, $J = 11.2$

Hz, 1H), 3.83 (t, $J = 11.5$ Hz, 1H), 2.11 (dt, $J = 11.4, 9.6$ Hz, 3H), 1.89–1.65 (m, 3H).

***Tert*-butyl 2-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)-1H-pyrrole-1-carboxylate (3c).** White solid; Yield: 60%; m.p.: 142–144°C; ^1H NMR (400 MHz, CDCl_3) δ 8.24 (s, 1H), 7.50 (s, 1H), 7.27 (s, 1H), 7.11 (d, $J = 1.6$ Hz, 1H), 6.36 (s, 1H), 5.79 (d, $J = 10.7$ Hz, 1H), 4.19 (d, $J = 11.1$ Hz, 1H), 3.78 (d, $J = 11.3$ Hz, 1H), 2.30–1.93 (m, 3H), 1.76 (ddd, $J = 34.6, 22.7, 11.2$ Hz, 3H).

6-(benzo[d][1,3]dioxol-5-yl)-2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purine (3d). White solid; Yield: 74%; m.p.: 177–179°C; ^1H NMR (400 MHz, CDCl_3) δ 8.53 (d, $J = 8.3$ Hz, 1H), 8.30 (d, $J = 17.9$ Hz, 2H), 6.98 (d, $J = 8.3$ Hz, 1H), 6.06 (s, 2H), 5.81 (d, $J = 10.5$ Hz, 1H), 4.19 (d, $J = 11.4$ Hz, 1H), 3.80 (t, $J = 11.2$ Hz, 1H), 2.17 (d, $J = 12.1$ Hz, 1H), 2.08 (s, 1H), 1.98 (dd, $J = 24.3, 13.3$ Hz, 1H), 1.79 (td, $J = 23.0, 12.0$ Hz, 2H), 1.68 (d, $J = 9.4$ Hz, 1H).

2-chloro-9-(tetrahydro-2H-pyran-2-yl)-6-(thiophen-3-yl)-9H-purine(3e). Yellow solid; Yield: 98%; m.p.: 156–158°C; ^1H NMR (400 MHz, CDCl_3) δ 8.97–8.90 (m, 1H), 8.31–8.23 (m, 2H), 7.45 (q, $J = 4.55, 4.04$ Hz, 1H), 5.81 (d, $J = 10.49$ Hz, 1H), 4.19 (d, $J = 10.81$ Hz, 1H), 3.80 (t, $J = 11.05$ Hz, 1H), 2.23–1.96 (m, 3H), 1.78 (dt, $J = 24.39, 11.95$ Hz, 3H).

2-chloro-6-(3-nitrophenyl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purine (3f). White solid; Yield: 57%; m.p.: 105–107°C; ^1H NMR (400 MHz, CDCl_3) δ 9.72 (s, 1H), 9.18 (d, $J = 7.8$ Hz, 1H), 8.40 (d, $J = 3.5$ Hz, 2H), 7.74 (t, $J = 8.0$ Hz, 1H), 5.85 (d, $J = 10.4$ Hz, 1H), 4.21 (d, $J = 11.0$ Hz, 1H), 3.82 (t, $J = 10.9$ Hz, 1H), 2.21 (d, $J = 12.5$ Hz, 1H), 2.11 (d, $J = 6.4$ Hz, 1H), 2.02 (dd, $J = 12.4, 9.3$ Hz, 1H), 1.87–1.64 (m, 3H).

Methyl 3-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzoate (3h). White solid; Yield: 42%; m.p.: 110–112°C; ^1H NMR (400 MHz, CDCl_3) δ 9.42 (s, 1H), 9.00 (d, $J = 7.8$ Hz, 1H), 8.35 (s, 1H), 8.22 (d, $J = 7.7$ Hz, 1H), 7.64 (t, $J = 7.8$ Hz, 1H), 5.84 (d, $J = 10.4$ Hz, 1H), 4.20 (d, $J = 10.0$ Hz, 1H), 3.81 (t, $J = 11.3$ Hz, 1H), 2.19 (d, $J = 12.5$ Hz, 1H), 2.09 (d, $J = 6.4$ Hz, 1H), 1.98 (dd, $J = 16.8, 7.0$ Hz, 1H), 1.90–1.71 (m, 3H).

2-chloro-6-(4-fluorophenyl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purine (3i). Oil; Yield: 88%. The product was used for the next step without purification.

Methyl 4-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzoate (3k). Oil; Yield: 80%; ^1H NMR (400 MHz, CDCl_3) δ 8.87 (d, $J = 8.0$ Hz, 2H), 8.36 (s, 1H), 8.20 (d, $J = 8.0$ Hz, 2H), 5.84 (d, $J = 10.3$ Hz, 1H), 4.20 (d, $J = 10.7$ Hz, 1H), 3.97 (s, 3H), 3.81 (t, $J = 11.0$ Hz, 1H), 2.19 (d, $J = 12.3$ Hz, 1H), 2.10 (d, $J = 6.7$ Hz, 1H), 2.03–1.94 (m, 1H), 1.79 (dt, $J = 23.7, 11.7$ Hz, 2H), 1.69 (d, $J = 9.7$ Hz, 1H).

***Tert*-butyl 3-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)phenylcarbamate (9a).** White solid; Yield: 58%; m.p.: 158–160°C; ^1H NMR (400 MHz, CDCl_3) δ 8.53 (s, 1H), 8.47 (d, $J = 7.8$ Hz, 1H), 8.32 (s, 1H), 7.84 (s, 1H), 7.49 (t, $J = 8.0$ Hz, 1H), 6.75 (s, 1H), 5.82 (d, $J = 10.2$ Hz, 1H), 4.19 (d, $J = 12.4$ Hz, 1H), 3.80 (t, $J = 10.8$ Hz, 1H), 2.17 (d, $J = 12.9$ Hz, 1H), 2.09 (d, $J = 10.1$ Hz, 1H), 2.02–1.96 (m, 1H), 1.88–1.74 (m, 2H), 1.70 (d, $J = 11.3$ Hz, 1H), 1.54 (s, 9H).

***Tert*-butyl 4-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)phenylcarbamate (9b).** White solid; Yield: 62%; m.p.: 199–201°C; ^1H NMR (400 MHz, CDCl_3) δ 8.79 (d,

$J = 7.9$ Hz, 2H), 8.29 (s, 1H), 7.54 (d, $J = 8.2$ Hz, 2H), 6.70 (s, 1H), 5.82 (d, $J = 10.6$ Hz, 1H), 4.19 (d, $J = 10.9$ Hz, 1H), 3.80 (t, $J = 11.2$ Hz, 1H), 2.16 (d, $J = 12.5$ Hz, 1H), 2.06 (d, $J = 10.4$ Hz, 1H), 1.98 (dd, $J = 22.7, 12.0$ Hz, 1H), 1.87–1.72 (m, 2H), 1.67 (d, $J = 10.1$ Hz, 1H), 1.54 (s, 9H).

***Tert*-butyl 3-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzylcarbamate (9c).** White solid; Yield: 60%; mp: 102–104°C; ^1H NMR (400 MHz, CDCl_3) δ 8.74 (d, $J = 7.2$ Hz, 1H), 8.64 (s, 1H), 8.31 (s, 1H), 7.56–7.46 (m, 2H), 5.83 (d, $J = 10.5$ Hz, 1H), 4.97 (s, 1H), 4.45 (d, $J = 4.3$ Hz, 2H), 4.20 (d, $J = 10.8$ Hz, 1H), 3.81 (t, $J = 10.9$ Hz, 1H), 2.18 (d, $J = 12.5$ Hz, 1H), 2.08 (s, 1H), 1.99 (d, $J = 11.0$ Hz, 1H), 1.80 (td, $J = 23.3, 12.3$ Hz, 2H), 1.68 (d, $J = 10.7$ Hz, 1H), 1.24 (s, 9H).

***Tert*-butyl 4-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzylcarbamate (9d).** White solid; Yield: 75%; m.p.: 175–177°C; ^1H NMR (400 MHz, CDCl_3) δ 8.76 (d, $J = 7.9$ Hz, 2H), 8.32 (s, 1H), 7.46 (d, $J = 7.9$ Hz, 2H), 5.83 (d, $J = 10.4$ Hz, 1H), 4.92 (s, 1H), 4.41 (d, $J = 5.0$ Hz, 2H), 4.20 (d, $J = 11.3$ Hz, 1H), 3.81 (t, $J = 11.0$ Hz, 1H), 2.18 (d, $J = 12.4$ Hz, 1H), 2.08 (s, 1H), 1.99 (dd, $J = 22.9, 11.4$ Hz, 1H), 1.80 (td, $J = 22.9, 12.2$ Hz, 2H), 1.68 (d, $J = 9.8$ Hz, 1H), 1.48 (s, 9H).

***Tert*-butyl 3-(2-chloro-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)phenethyl carbamate (9e).** Yellow solid; Yield: 65%; m.p.: 149–151°C; ^1H NMR (400 MHz, CDCl_3) δ 8.69 (d, $J = 7.7$ Hz, 1H), 8.57 (s, 1H), 8.31 (s, 1H), 7.50 (t, $J = 7.7$ Hz, 1H), 7.39 (d, $J = 7.3$ Hz, 1H), 5.83 (d, $J = 10.4$ Hz, 1H), 4.60 (s, 1H), 4.20 (d, $J = 12.4$ Hz, 1H), 3.81 (t, $J = 11.0$ Hz, 1H), 3.46 (d, $J = 5.7$ Hz, 2H), 2.94 (t, $J = 6.7$ Hz, 2H), 2.18 (d, $J = 12.5$ Hz, 1H), 2.08 (s, 1H), 2.00 (dd, $J = 24.6, 13.5$ Hz, 1H), 1.80 (td, $J = 23.3, 12.3$ Hz, 2H), 1.68 (d, $J = 10.7$ Hz, 1H), 1.43 (s, 9H).

4.1.3 General Method for the Preparation of Compounds 4a–4f, 4h, 4i, 4k, 10a–10r

***Tert*-butyl 4-(2-((4-fluorophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzylcarbamate (10o).** Compound **9d** (1.0 mmol), 4-fluoroaniline (2 mmol), $\text{Pd}(\text{OAc})_2$ (0.05 mmol), Xantphos (0.10 mmol), and Cs_2CO_3 (13.5 mmol) were mixed in a two-neck flask. Under the protection of N_2 , the anhydrous 1,4-dioxane was added and the mixture reacted at 100°C for 18 h. After the completion, the reaction mixture was filtered through a pad of Celite. **Spinned the filtrate dry and then dissolved it with ethyl acetate (15 ml) and water (20 ml), extracted twice with ethyl acetate (50 ml), washed with brine, and dried with anhydrous MgSO_4 . The crude product was concentrated and purified by silica gel chromatography (eluting with dichloromethane/menthol 100/1 to 40/1) to obtain compounds 10o.**

Compounds **4a–4f**, **4h**, **4i**, **4k**, **10a–10n**, and **10p–10r** were synthesized following the procedure described above.

N-(3-nitrophenyl)-6-phenyl-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-2-amine (4a). Light yellow solid; Yield: 76%; m.p.: 164–166°C; ^1H NMR (400 MHz, $\text{DMSO}-d_6$) δ 10.29 (s, 1H), 9.38 (s, 1H), 8.85 (d, $J = 7.2$ Hz, 2H), 8.61 (s, 1H), 8.01 (d, $J = 8.1$ Hz, 1H), 7.81

(d, $J = 7.9$ Hz, 1H), 7.60 (t, $J = 8.5$ Hz, 4H), 5.73 (d, $J = 10.9$ Hz, 1H), 4.10 (d, $J = 11.4$ Hz, 1H), 3.76 (dd, $J = 15.6, 6.5$ Hz, 1H), 2.41 (dd, $J = 21.4, 10.6$ Hz, 1H), 2.08 (d, $J = 11.3$ Hz, 2H), 1.84–1.58 (m, 3H).

6-(naphthalen-1-yl)-*N*-(3-nitrophenyl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-2-amine (4b). White solid; Yield: 81%. m.p.: 190–192°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.42 (s, 1H), 9.30 (s, 1H), 8.53 (s, 1H), 8.21 (d, $J = 8.4$ Hz, 1H), 8.14 (d, $J = 8.2$ Hz, 1H), 8.05 (t, $J = 8.0$ Hz, 2H), 7.97 (d, $J = 7.0$ Hz, 1H), 7.80 (d, $J = 8.0$ Hz, 1H), 7.70 (t, $J = 7.6$ Hz, 1H), 7.57 (ddd, $J = 25.1, 12.7, 7.3$ Hz, 3H), 5.78 (d, $J = 10.9$ Hz, 1H), 4.12 (d, $J = 11.1$ Hz, 1H), 3.80 (dd, $J = 15.8, 6.6$ Hz, 1H), 2.45 (d, $J = 9.5$ Hz, 1H), 2.10 (t, $J = 14.2$ Hz, 2H), 1.87–1.60 (m, 3H).

***Tert*-butyl 2-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)-1H-pyrrole-1-carboxylate (4c).** White solid; Yield: 80%; m.p.: 185–187°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.27 (s, 1H), 9.26 (s, 1H), 8.51 (d, $J = 12.6$ Hz, 1H), 8.01 (d, $J = 8.3$ Hz, 1H), 7.79 (d, $J = 8.1$ Hz, 1H), 7.63–7.52 (m, 2H), 6.86 (s, 1H), 6.44 (s, 1H), 5.70 (d, $J = 10.8$ Hz, 1H), 4.09 (d, $J = 11.0$ Hz, 1H), 3.75 (dd, $J = 15.5, 6.6$ Hz, 1H), 2.43 (d, $J = 10.7$ Hz, 1H), 2.04 (d, $J = 9.4$ Hz, 2H), 1.85–1.70 (m, 1H), 1.65 (d, $J = 15.4$ Hz, 2H), 1.23 (s, 9H).

6-(benzo[d][1,3]dioxol-5-yl)-*N*-(3-nitrophenyl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-2-amine (4d). Yellow solid; Yield: 65%; m.p.: 207–209°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.20 (s, 1H), 9.34 (s, 1H), 8.58 (d, $J = 10.4$ Hz, 2H), 8.36 (s, 1H), 7.97 (d, $J = 8.3$ Hz, 1H), 7.80 (d, $J = 8.1$ Hz, 1H), 7.59 (t, $J = 8.1$ Hz, 1H), 7.15 (d, $J = 8.2$ Hz, 1H), 6.16 (s, 2H), 5.71 (d, $J = 10.8$ Hz, 1H), 4.09 (d, $J = 11.4$ Hz, 1H), 3.73 (d, $J = 8.7$ Hz, 1H), 2.39 (dd, $J = 22.0, 10.6$ Hz, 1H), 2.06 (d, $J = 10.5$ Hz, 2H), 1.84–1.57 (m, 3H).

***N*-(3-nitrophenyl)-9-(tetrahydro-2H-pyran-2-yl)-6-(thiophen-3-yl)-9H-purin-2-amine (4e).** Light yellow solid; Yield: 60%; m.p.: 208–210°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.23 (s, 1H), 9.39 (t, $J = 2.12$ Hz, 1H), 8.96 (dd, $J = 2.93, 0.97$ Hz, 1H), 8.59 (s, 1H), 8.27 (dd, $J = 5.07, 0.89$ Hz, 1H), 8.00 (dd, $J = 8.17, 1.38$ Hz, 1H), 7.87–7.77 (m, 2H), 7.59 (t, $J = 8.17$ Hz, 1H), 5.71 (dd, $J = 10.91, 1.69$ Hz, 1H), 4.09 (d, $J = 11.31$ Hz, 1H), 3.74 (td, $J = 11.36, 3.74$ Hz, 1H), 2.40 (ddd, $J = 16.28, 12.65, 4.22$ Hz, 1H), 2.07 (d, $J = 10.72$ Hz, 2H), 1.82–1.59 (m, 3H).

***N*,6-bis(3-nitrophenyl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-2-amine (4f).** Yellow solid; Yield: 77%; The product was put into the next step without purification.

Methyl 3-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl) benzoate (4h). Yellow solid; Yield: 40%; m.p.: 222–224°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.35 (d, $J = 9.6$ Hz, 1H), 9.41 (s, 1H), 9.31 (s, 1H), 9.11 (d, $J = 6.7$ Hz, 1H), 8.65 (d, $J = 6.2$ Hz, 1H), 8.17 (d, $J = 6.5$ Hz, 1H), 8.04 (d, $J = 7.4$ Hz, 1H), 7.85–7.68 (m, 2H), 7.67–7.54 (m, 1H), 5.75 (d, $J = 10.1$ Hz, 1H), 4.11 (d, $J = 11.3$ Hz, 1H), 3.75 (d, $J = 9.5$ Hz, 1H), 2.42 (d, $J = 12.0$ Hz, 1H), 2.08 (t, $J = 11.5$ Hz, 2H), 1.84–1.60 (m, 3H).

6-(4-fluorophenyl)-*N*-(3-nitrophenyl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-2-amine (4i). Yellow solid; Yield: 69%; m.p.: 206–208°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.30 (s, 1H), 9.34 (s, 1H), 8.97–8.87 (m, 2H), 8.62 (s, 1H), 8.01 (d, $J = 8.4$ Hz, 1H), 7.82 (d, $J = 8.1$ Hz, 1H), 7.60 (t, $J = 8.1$ Hz, 1H), 7.47 (t, $J = 8.4$ Hz, 2H), 5.73 (d, $J = 10.8$ Hz, 1H), 4.10 (d, $J = 11.3$ Hz, 1H), 3.74 (d, $J = 8.6$ Hz, 1H), 2.47–2.34 (m, 1H), 2.08 (d, $J = 11.4$ Hz, 2H), 1.75 (s, 1H), 1.65 (s, 2H).

Methyl 4-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl) benzoate (4k). Yellow oil; Yield: 41%; ^1H NMR (400 MHz, DMSO- d_6) δ 10.38 (s, 1H), 9.36 (s, 1H), 8.96 (d, $J = 7.9$ Hz, 2H), 8.66 (s, 1H), 8.19 (d, $J = 8.0$ Hz, 2H), 8.00 (d, $J = 8.0$ Hz, 1H), 7.82 (d, $J = 8.1$ Hz, 1H), 7.61 (t, $J = 8.1$ Hz, 1H), 5.74 (d, $J = 10.9$ Hz, 1H), 4.10 (d, $J = 10.7$ Hz, 1H), 3.92 (s, 3H), 3.74 (d, $J = 9.5$ Hz, 1H), 2.47–2.35 (m, 1H), 2.09 (d, $J = 11.0$ Hz, 2H), 1.83–1.59 (m, 3H).

***Tert*-butyl(3-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)phenyl)carbamate (10a).** Yellow solid; Yield: 22%; m.p.: 192–194°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.25 (s, 1H), 9.54 (s, 1H), 9.23 (s, 1H), 8.90 (s, 1H), 8.60 (s, 1H), 8.44 (d, $J = 7.7$ Hz, 1H), 8.17 (d, $J = 8.1$ Hz, 1H), 7.81 (d, $J = 8.1$ Hz, 1H), 7.54 (m, $J = 32.5, 16.0, 8.0$ Hz, 3H), 5.73 (d, $J = 10.9$ Hz, 1H), 4.10 (d, $J = 10.9$ Hz, 1H), 3.75 (t, $J = 10.8$ Hz, 1H), 2.41 (dd, $J = 20.7, 11.2$ Hz, 1H), 2.07 (d, $J = 9.6$ Hz, 2H), 1.76 (d, $J = 9.2$ Hz, 1H), 1.64 (s, 2H), 1.50 (s, 9H).

***Tert*-butyl(4-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)phenyl)carbamate (10b).** Yellow solid; Yield: 68%; m.p.: 220–222°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.22 (s, 1H), 9.74 (s, 1H), 9.41 (s, 1H), 8.80 (d, $J = 8.2$ Hz, 2H), 8.57 (s, 1H), 7.98 (d, $J = 8.2$ Hz, 1H), 7.80 (d, $J = 8.0$ Hz, 1H), 7.70 (d, $J = 8.3$ Hz, 2H), 7.59 (t, $J = 8.2$ Hz, 1H), 5.72 (d, $J = 10.9$ Hz, 1H), 4.09 (d, $J = 11.5$ Hz, 1H), 3.75 (t, $J = 10.8$ Hz, 1H), 2.39 (dd, $J = 21.4, 10.8$ Hz, 1H), 2.06 (d, $J = 10.7$ Hz, 2H), 1.84–1.60 (m, 3H), 1.51 (s, 9H).

***Tert*-butyl(3-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10c).** Yellow solid; Yield: 77%; m.p.: 163–165°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.29 (s, 1H), 9.36 (s, 1H), 8.75 (d, $J = 7.8$ Hz, 1H), 8.67 (s, 1H), 8.59 (s, 1H), 8.03 (d, $J = 8.3$ Hz, 1H), 7.81 (d, $J = 8.0$ Hz, 1H), 7.65–7.54 (m, 2H), 7.47 (d, $J = 6.9$ Hz, 2H), 5.74 (d, $J = 11.0$ Hz, 1H), 4.27 (d, $J = 5.5$ Hz, 2H), 4.10 (d, $J = 11.4$ Hz, 1H), 3.74 (d, $J = 9.2$ Hz, 1H), 2.40 (dd, $J = 31.6, 20.3$ Hz, 1H), 2.08 (d, $J = 9.8$ Hz, 2H), 1.73 (d, $J = 14.4$ Hz, 1H), 1.66 (d, $J = 15.1$ Hz, 2H), 1.35 (d, $J = 40.1$ Hz, 9H).

***Tert*-butyl(4-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10d).** Yellow solid; Yield: 53%; m.p.: 195–197°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.27 (s, 1H), 9.40 (s, 1H), 8.80 (d, $J = 7.8$ Hz, 2H), 8.61 (s, 1H), 7.99 (d, $J = 8.1$ Hz, 1H), 7.81 (d, $J = 8.0$ Hz, 1H), 7.60 (t, $J = 7.9$ Hz, 1H), 7.56–7.40 (m, 3H), 5.73 (d, $J = 10.9$ Hz, 1H), 4.25 (d, $J = 5.8$ Hz, 2H), 4.10 (d, $J = 11.6$ Hz, 1H), 3.75 (t, $J = 8.7$ Hz, 1H), 2.47–2.32 (m, 1H), 2.05 (t, $J = 19.1$ Hz, 2H), 1.69 (m, $J = 38.5, 16.9$ Hz, 3H), 1.38 (d, $J = 36.5$ Hz, 9H).

***Tert*-butyl(3-(2-((3-nitrophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)phenethyl)carbamate (10e).** Yellow solid; Yield: 73%; m.p.: 104–106°C; ^1H NMR (400 MHz, CDCl $_3$) δ 9.52 (s, 1H), 8.69 (d, $J = 7.3$ Hz, 1H), 8.55 (s, 1H), 8.13 (s, 1H), 7.87 (d, $J = 7.4$ Hz, 1H), 7.60 (s, 1H), 7.47 (dd, $J = 19.6, 7.2$ Hz, 3H), 7.37 (d, $J = 6.3$ Hz, 1H), 5.79 (d, $J = 9.8$ Hz, 1H), 4.68 (s, 1H), 4.23 (d, $J = 11.8$ Hz, 1H), 3.89 (t, $J = 11.4$ Hz, 1H), 3.48 (s, 2H), 2.96 (s, 2H), 2.17 (dd, $J = 19.4, 10.6$ Hz, 3H), 1.85 (ddd, $J = 36.7, 24.7, 12.0$ Hz, 2H), 1.71 (d, $J = 12.5$ Hz, 1H), 1.43 (s, 9H).

Tert-butyl(3-(2-(phenylamino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10f). Yellow solid; Yield: 62%; m.p.: 110–112°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.68 (s, 1H), 8.70 (d, *J* = 7.7 Hz, 1H), 8.64 (s, 1H), 8.51 (s, 1H), 7.93 (d, *J* = 7.9 Hz, 2H), 7.54 (dd, *J* = 15.2, 7.3 Hz, 2H), 7.44 (d, *J* = 7.3 Hz, 1H), 7.35 (t, *J* = 7.5 Hz, 2H), 6.96 (t, *J* = 7.2 Hz, 1H), 5.69 (d, *J* = 10.8 Hz, 1H), 4.25 (d, *J* = 5.7 Hz, 2H), 4.08 (d, *J* = 11.0 Hz, 1H), 3.70 (s, 1H), 2.39 (dd, *J* = 22.1, 10.9 Hz, 1H), 2.03 (d, *J* = 11.2 Hz, 2H), 1.77 (s, 1H), 1.63 (s, 2H), 1.49–1.17 (s, 9H).

Tert-butyl(3-(2-([1,1'-biphenyl]-4-ylamino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate(10h). Yellow solid; Yield: 77%; m.p.: 120; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.82 (s, 1H), 8.72 (d, *J* = 7.8 Hz, 1H), 8.65 (s, 1H), 8.53 (s, 1H), 8.04 (d, *J* = 8.1 Hz, 2H), 7.69 (d, *J* = 7.7 Hz, 4H), 7.60–7.41 (m, 5H), 7.31 (t, *J* = 7.0 Hz, 1H), 5.71 (d, *J* = 10.7 Hz, 1H), 4.26 (d, *J* = 5.6 Hz, 2H), 4.09 (d, *J* = 9.9 Hz, 1H), 3.73 (s, 1H), 2.42–2.32 (m, 1H), 2.05 (d, *J* = 11.9 Hz, 2H), 1.79 (s, 1H), 1.64 (s, 2H), 1.38 (s, 9H).

Tert-butyl(3-(9-(tetrahydro-2H-pyran-2-yl)-2-(p-tolylamino)-9H-purin-6-yl)benzyl)carbamate(10i). Yellow solid; Yield: 60; m.p.: 106–108°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.57 (s, 1H), 8.68 (d, *J* = 7.6 Hz, 1H), 8.63 (s, 1H), 8.48 (s, 1H), 7.81 (d, *J* = 7.9 Hz, 2H), 7.53 (dd, *J* = 16.2, 8.2 Hz, 2H), 7.44 (d, *J* = 7.6 Hz, 1H), 7.15 (d, *J* = 7.9 Hz, 2H), 5.67 (d, *J* = 10.8 Hz, 1H), 4.25 (d, *J* = 5.7 Hz, 2H), 4.07 (d, *J* = 11.3 Hz, 1H), 3.72 (d, *J* = 13.1 Hz, 1H), 2.39 (dd, *J* = 22.0, 11.7 Hz, 1H), 2.28 (s, 3H), 2.02 (d, *J* = 11.0 Hz, 2H), 1.76 (s, 1H), 1.63 (s, 2H), 1.48–1.19 (s, 9H).

Tert-butyl(3-(2-((4-(tert-butyl)phenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10j). Oil; Yield: 64%; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.54 (s, 1H), 8.67 (d, *J* = 7.6 Hz, 1H), 8.64 (s, 1H), 8.47 (s, 1H), 7.84 (d, *J* = 8.0 Hz, 2H), 7.54 (t, *J* = 7.7 Hz, 1H), 7.44 (d, *J* = 7.4 Hz, 2H), 7.36 (d, *J* = 8.1 Hz, 2H), 5.69 (d, *J* = 10.9 Hz, 1H), 4.25 (d, *J* = 5.3 Hz, 2H), 4.07 (d, *J* = 11.2 Hz, 1H), 3.72 (s, 1H), 2.45–2.33 (m, 1H), 2.03 (d, *J* = 11.2 Hz, 2H), 1.78 (s, 1H), 1.63 (s, 2H), 1.40 (s, 9H), 1.30 (s, 9H).

Tert-butyl 4-(4-((6-(3-(((tert-butoxycarbonyl)amino)methyl)phenyl)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-2-yl)amino)phenyl)piperazine-1-carboxylate (10k). Yellow solid; Yield: 72%; m.p.: 120–124°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.45 (s, 1H), 8.67 (d, *J* = 7.3 Hz, 1H), 8.61 (s, 1H), 8.45 (s, 1H), 7.78 (d, *J* = 8.2 Hz, 2H), 7.49 (dt, *J* = 25.1, 7.7 Hz, 3H), 6.98 (d, *J* = 8.4 Hz, 2H), 5.66 (d, *J* = 11.0 Hz, 1H), 4.24 (d, *J* = 5.4 Hz, 2H), 4.07 (d, *J* = 11.0 Hz, 1H), 3.70 (s, 1H), 3.48 (s, 4H), 3.03 (s, 4H), 2.43–2.30 (m, 1H), 2.02 (d, *J* = 10.2 Hz, 2H), 1.76 (s, 1H), 1.63 (s, 2H), 1.35 (dd, *J* = 55.6, 18.5 Hz, 18H).

Tert-butyl(3-(2-((4-sulfamoylphenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10l). Yellow solid; Yield: 35%; m.p.: 145–147°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 10.12 (s, 1H), 8.76–8.62 (m, 2H), 8.57 (s, 1H), 8.09 (d, *J* = 8.3 Hz, 2H), 7.80 (d, *J* = 8.3 Hz, 2H), 7.57 (t, *J* = 7.6 Hz, 1H), 7.47 (t, *J* = 8.2 Hz, 2H), 7.16 (s, 2H), 5.73 (d, *J* = 10.7 Hz, 1H), 4.26 (d, *J* = 5.5 Hz, 2H), 4.09 (d, *J* = 11.2 Hz, 1H), 3.74 (s, 1H), 2.47–2.32 (m, 1H), 2.05 (d, *J* = 10.6 Hz, 2H), 1.80 (s, 1H), 1.65 (s, 2H), 1.36 (d, *J* = 36.9 Hz, 9H).

Tert-butyl(3-(2-((4-fluorophenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10n).

Yellow solid; Yield: 52%; m.p.: 159–161°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.57 (s, 1H), 8.68 (d, *J* = 7.6 Hz, 1H), 8.63 (s, 1H), 8.48 (s, 1H), 7.81 (d, *J* = 7.9 Hz, 2H), 7.53 (dd, *J* = 16.2, 8.2 Hz, 2H), 7.44 (d, *J* = 7.6 Hz, 1H), 7.15 (d, *J* = 7.9 Hz, 2H), 5.67 (d, *J* = 10.8 Hz, 1H), 4.25 (d, *J* = 5.7 Hz, 2H), 4.07 (d, *J* = 11.3 Hz, 1H), 3.72 (d, *J* = 13.1 Hz, 1H), 2.39 (dd, *J* = 22.0, 11.7 Hz, 1H), 2.28 (s, 3H), 2.02 (d, *J* = 11.0 Hz, 2H), 1.76 (s, 1H), 1.63 (s, 2H), 1.48–1.19 (s, 9H).

Tert-butyl(3-(2-((4-(N-methylsulfamoyl)phenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10p). Yellow solid; Yield: 30%; ¹H NMR (400 MHz, DMSO-*d*₆) δ 10.21 (s, 1H), 8.70 (d, *J* = 7.5 Hz, 1H), 8.65 (s, 1H), 8.59 (s, 1H), 8.14 (d, *J* = 8.3 Hz, 2H), 7.76 (d, *J* = 8.3 Hz, 2H), 7.62–7.43 (m, 3H), 7.23 (d, *J* = 5.0 Hz, 1H), 5.74 (d, *J* = 10.9 Hz, 1H), 4.26 (d, *J* = 5.8 Hz, 2H), 4.08 (d, *J* = 11.3 Hz, 1H), 3.75 (s, 1H), 2.46–2.30 (m, 4H), 2.04 (dd, *J* = 24.5, 13.0 Hz, 2H), 1.80 (s, 1H), 1.64 (s, 2H), 1.40 (s, 9H).

Tert-butyl(3-(2-((4-(N,N-dimethylsulfamoyl)phenyl)amino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10q). Yellow solid; Yield: 57%; m.p.: 123–125°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 10.27 (s, 1H), 8.71 (d, *J* = 7.7 Hz, 1H), 8.62 (d, *J* = 15.9 Hz, 2H), 8.20 (d, *J* = 8.1 Hz, 2H), 7.74 (d, *J* = 8.1 Hz, 2H), 7.57 (t, *J* = 7.6 Hz, 1H), 7.54–7.42 (m, 2H), 5.74 (d, *J* = 10.8 Hz, 1H), 4.25 (d, *J* = 5.5 Hz, 2H), 4.08 (d, *J* = 11.1 Hz, 1H), 3.74 (d, *J* = 10.8 Hz, 1H), 2.61 (s, 6H), 2.37 (dd, *J* = 22.3, 11.4 Hz, 1H), 2.05 (d, *J* = 10.0 Hz, 2H), 1.80 (s, 1H), 1.64 (s, 2H), 1.40 (s, 9H).

Tert-butyl(3-(2-(pyridin-3-ylamino)-9-(tetrahydro-2H-pyran-2-yl)-9H-purin-6-yl)benzyl)carbamate (10r). Yellow solid; Yield: 46%; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.89 (s, 1H), 9.01 (s, 1H), 8.67 (d, *J* = 7.6 Hz, 1H), 8.63 (s, 1H), 8.54 (s, 1H), 8.42 (d, *J* = 8.2 Hz, 1H), 8.17 (s, 1H), 7.56 (t, *J* = 7.5 Hz, 1H), 7.50 (s, 1H), 7.45 (d, *J* = 7.5 Hz, 1H), 7.40 (d, *J* = 6.4 Hz, 1H), 5.70 (d, *J* = 10.7 Hz, 1H), 4.25 (d, *J* = 5.7 Hz, 2H), 4.08 (d, *J* = 11.2 Hz, 1H), 3.71 (t, *J* = 8.4 Hz, 1H), 2.38 (dd, *J* = 22.4, 11.3 Hz, 1H), 2.04 (d, *J* = 10.4 Hz, 2H), 1.78 (m, 1H), 1.63 (m, 2H), 1.35 (s, 9H).

4.1.4 General Method for the Preparation of Compounds 5a-5f, 5h, 5i, 5k, 11a-11r

6-(4-(aminomethyl)phenyl)-N-(4-fluorophenyl)-9H-purin-2-amine hydrochloride (11o). Compounds **10o** (1.0 mmol) were dissolved in HCl saturated ethyl acetate solution (15 ml) and stirred at room temperature for 4 h and then filtered to get compounds **11o**. Light yellow solid; Yield: 90%; m.p.: >300°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 9.56 (s, 1H), 8.80 (d, *J* = 8.06 Hz, 2H), 8.36 (d, *J* = 14.57 Hz, 6H), 7.91–7.82 (m, 3H), 7.69 (d, *J* = 8.22 Hz, 3H), 7.16 (t, *J* = 8.90 Hz, 2H), 4.14 (q, *J* = 5.94 Hz, 3H), 4.10 (s, 16H). ¹³C NMR (101 MHz, DMSO-*d*₆) δ 155.46, 155.24, 152.79, 148.27, 144.12, 140.34, 137.29, 136.04, 129.84, 129.59, 125.56, 117.53, 42.49. HRMS (AP-ESI) *m/z* Calcd for C₁₈H₁₅FN₆ [M + H]⁺ 335.1415, found: 335.1418.

Compounds **5a-5f**, **5h**, **5i**, **5k**, **11a-11n**, and **11p-11r** were synthesized following the procedure described above.

N-(3-nitrophenyl)-6-phenyl-9H-purin-2-amine (5a). Light yellow solid; Yield: 85%; m.p.: 201–203°C; ¹H NMR (400 MHz, DMSO-*d*₆) δ 10.18 (s, 1H), 9.18 (s, 1H), 8.80 (d, *J* = 7.1 Hz, 2H), 8.53 (s, 1H), 8.12 (d, *J* = 8.2 Hz, 1H), 7.80 (d, *J* = 8.0 Hz, 1H), 7.61 (dd, *J* = 12.3, 7.1 Hz, 4H); ¹³C NMR (101 MHz, DMSO-*d*₆) δ

155.77, 155.21, 153.43, 148.72, 143.21, 143.07, 136.26, 131.34, 130.09, 129.82, 128.99, 126.03, 124.61, 115.39, 112.23. HRMS (AP-ESI) m/z Calcd for $C_{17}H_{12}N_6O_2$ $[M + H]^+$ 333.1095, found: 333.1090.

6-(naphthalen-1-yl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine (5b). Light yellow solid. Yield: 92%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.31 (s, 1H), 8.98 (s, 1H), 8.54 (s, 1H), 8.18 (dd, $J = 20.0, 9.5$ Hz, 3H), 8.07 (d, $J = 8.0$ Hz, 1H), 7.95 (d, $J = 7.0$ Hz, 1H), 7.77 (d, $J = 8.3$ Hz, 1H), 7.71 (t, $J = 7.5$ Hz, 1H), 7.56 (m, $J = 15.1, 14.2, 7.2$ Hz, 3H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.27, 148.72, 143.93, 142.79, 133.86, 130.85, 130.67, 130.20, 129.36, 128.80, 127.23, 126.72, 126.19, 125.70, 124.73, 115.71, 112.46. HRMS (AP-ESI) m/z Calcd for $C_{21}H_{14}N_6O_2$ $[M + H]^+$ 383.1251, found: 383.1250.

***N*-(3-nitrophenyl)-6-(1*H*-pyrrol-2-yl)-9*H*-purin-2-amine hydrochloride (5c).** Yellow solid; Yield: 91%; m.p.: 236°C (Dec.); 1H NMR (400 MHz, DMSO- d_6) δ 11.41 (s, 1H), 9.87 (s, 1H), 9.14 (s, 1H), 8.31 (s, 1H), 8.15 (d, $J = 8.1$ Hz, 1H), 7.77 (d, $J = 8.0$ Hz, 1H), 7.58 (t, $J = 8.1$ Hz, 1H), 7.42 (s, 1H), 7.20 (s, 1H), 6.35 (s, 1H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 155.98, 154.38, 148.73, 147.23, 143.24, 142.59, 130.08, 128.69, 124.54, 124.45, 122.78, 115.11, 113.89, 112.32, 110.54. HRMS (AP-ESI) m/z Calcd for $C_{15}H_{11}N_7O_2$ $[M + H]^+$ 322.1047, found: 322.1044.

6-(benzo[d][1,3]dioxol-5-yl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine (5d). Yellow solid; Yield: 97%; m.p.: 260°C (Dec.); 1H NMR (400 MHz, DMSO- d_6) δ 10.12 (s, 1H), 9.14 (s, 1H), 8.75–8.44 (m, 2H), 8.33 (s, 1H), 8.08 (d, $J = 8.2$ Hz, 1H), 7.79 (d, $J = 8.0$ Hz, 1H), 7.59 (t, $J = 8.1$ Hz, 1H), 7.16 (d, $J = 8.2$ Hz, 1H), 6.17 (s, 2H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 155.92, 154.27, 152.61, 150.62, 148.62, 148.32, 143.28, 142.42, 130.16, 128.88, 125.40, 124.89, 121.08, 115.92, 112.53, 109.00, 108.95, 102.33. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{12}N_6O_4$ $[M + H]^+$ 377.0993, found: 377.0993.

***N*-(3-nitrophenyl)-6-(thiophen-3-yl)-9*H*-purin-2-amine (5e).** Yellow solid; Yield: 93%; m.p.: > 280°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.09 (s, 1H), 9.22 (s, 1H), 8.97 (s, 1H), 8.40 (s, 1H), 8.28 (d, $J = 5.02$ Hz, 1H), 8.09 (d, $J = 8.04$ Hz, 1H), 7.79 (d, $J = 5.09$ Hz, 2H), 7.59 (t, $J = 8.13$ Hz, 1H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 155.89, 154.93, 149.24, 148.66, 143.55, 142.89, 138.35, 131.01, 130.14, 127.79, 127.55, 124.73, 115.54, 112.32. HRMS (AP-ESI) m/z Calcd for $C_{15}H_{10}N_6O_2S$ $[M + H]^+$ 339.0659, found: 339.0661.

***N*,6-bis(3-nitrophenyl)-9*H*-purin-2-amine (5f).** Brown solid; Yield: 84%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 13.44 (s, 1H), 10.29 (s, 1H), 9.79 (s, 1H), 9.34 (d, $J = 7.7$ Hz, 1H), 9.12 (s, 1H), 8.57–8.45 (m, 2H), 8.20 (d, $J = 8.0$ Hz, 1H), 7.98 (t, $J = 8.0$ Hz, 1H), 7.86 (d, $J = 7.9$ Hz, 1H), 7.66 (t, $J = 8.1$ Hz, 1H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 155.81, 154.95, 152.96, 147.88, 144.25, 140.57, 135.49, 135.07, 132.55, 130.41, 129.76, 129.62, 125.58, 117.81, 42.73. HRMS (AP-ESI) m/z Calcd for $C_{17}H_{11}N_7O_4$ $[M + H]^+$ 378.0945, found: 378.0964.

Methyl 3-(2-((3-nitrophenyl)amino)-9*H*-purin-6-yl)benzoate (5h). Light yellow solid; Yield: 53%; m.p.: 257°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.20 (s, 1H), 9.45 (s, 1H), 9.19–9.04 (m, 2H), 8.46 (s, 1H), 8.17 (d, $J = 8.0$ Hz, 2H), 7.84–7.77 (m, 2H), 7.60 (t, $J = 8.1$ Hz, 1H), 3.93 (s, 3H); ^{13}C NMR (101 MHz,

DMSO- d_6) δ 166.51, 155.96, 155.30, 152.22, 148.69, 143.74, 142.84, 136.50, 134.22, 131.87, 130.66, 130.34, 130.15, 129.64, 124.81, 115.67, 112.48, 52.84. HRMS (AP-ESI) m/z Calcd for $C_{19}H_{14}N_6O_4$ $[M + H]^+$ 391.1149, found: 391.1147.

6-(4-fluorophenyl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine (5i). Yellow solid; Yield: 97%; m.p.: 245–247°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.13 (s, 1H), 9.13 (s, 1H), 8.97–8.89 (m, 2H), 8.41 (s, 1H), 8.11 (d, $J = 8.1$ Hz, 1H), 7.79 (d, $J = 7.9$ Hz, 1H), 7.59 (t, $J = 8.1$ Hz, 1H), 7.46 (t, $J = 8.3$ Hz, 2H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 165.58, 163.10, 155.97, 154.99, 152.26, 148.67, 143.51, 142.77, 132.21, 132.12, 130.17, 124.77, 116.26, 116.05, 115.67, 112.41. HRMS (AP-ESI) m/z Calcd for $C_{17}H_{11}FN_6O_2$ $[M + H]^+$ 351.1000, found: 351.0998.

Methyl 4-(2-((3-nitrophenyl)amino)-9*H*-purin-6-yl)benzoate (5k). Light yellow solid; Yield: 95%; m.p.: 212–214°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.26 (s, 1H), 9.17 (s, 1H), 8.93 (d, $J = 8.1$ Hz, 2H), 8.57 (s, 1H), 8.18 (d, $J = 8.0$ Hz, 2H), 8.11 (d, $J = 8.1$ Hz, 1H), 7.80 (d, $J = 7.9$ Hz, 1H), 7.60 (t, $J = 8.1$ Hz, 1H), 3.92 (s, 3H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 166.37, 155.88, 155.66, 151.78, 148.72, 144.04, 142.88, 140.37, 131.73, 130.15, 129.89, 129.76, 125.53, 124.73, 115.59, 112.35, 52.81. HRMS (AP-ESI) m/z Calcd for $C_{19}H_{14}N_6O_4$ $[M + H]^+$ 391.1149, found: 391.1148.

6-(3-aminophenyl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine hydrochloride (11a). Yellow brown solid; Yield: 90%; mp: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 13.16 (s, 1H), 10.05 (s, 1H), 9.16 (s, 1H), 8.34 (s, 1H), 8.17–8.09 (m, 2H), 8.05 (d, $J = 7.7$ Hz, 1H), 7.78 (d, $J = 8.0$ Hz, 1H), 7.58 (dd, $J = 18.0, 9.8$ Hz, 1H), 7.24 (t, $J = 7.8$ Hz, 1H), 6.78 (d, $J = 7.8$ Hz, 1H), 5.26 (s, 2H).

^{13}C NMR (101 MHz, DMSO- d_6) δ 155.67, 149.17, 148.74, 143.18, 136.84, 130.13, 129.36, 124.53, 117.78, 116.99, 115.26, 112.19. HRMS (AP-ESI) m/z Calcd for $C_{17}H_{13}N_7O_2$ $[M + H]^+$ 348.1203, found: 348.1200.

6-(4-aminophenyl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine hydrochloride (11b). Yellow solid; Yield: 85%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.46 (s, 1H), 9.13 (s, 1H), 8.74 (d, $J = 7.7$ Hz, 3H), 8.08 (d, $J = 8.1$ Hz, 1H), 7.83 (d, $J = 8.1$ Hz, 1H), 7.62 (t, $J = 8.1$ Hz, 1H), 7.32 (d, $J = 7.7$ Hz, 2H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 155.36, 154.78, 151.97, 148.69, 143.66, 142.35, 131.41, 130.30, 124.96, 120.21, 116.05, 112.60, 40.60, 40.40, 40.19, 39.98, 39.77, 39.56, 39.35. HRMS (AP-ESI) m/z Calcd for $C_{17}H_{13}N_7O_2$ $[M + H]^+$ 348.1203, found: 348.1205.

6-(3-(aminomethyl)phenyl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine hydrochloride (11c). White solid; Yield: 81%; m.p.: 236–238°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.20 (s, 1H), 9.29 (s, 1H), 8.84 (d, $J = 31.7$ Hz, 2H), 8.52 (s, 4H), 8.07 (d, $J = 8.2$ Hz, 1H), 7.81 (d, $J = 8.1$ Hz, 1H), 7.76 (s, 1H), 7.70 (t, $J = 7.6$ Hz, 1H), 7.61 (t, $J = 8.1$ Hz, 1H), 4.19 (d, $J = 5.4$ Hz, 2H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.32, 152.77, 148.70, 143.88, 142.63, 135.05, 132.47, 130.37, 130.21, 129.83, 129.59, 124.92, 115.85, 112.51, 42.76. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{15}N_7O_2$ $[M + H]^+$ 362.1360, found: 362.1358.

6-(4-(aminomethyl)phenyl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine hydrochloride (11d). Light yellow solid; Yield: 92%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.24 (s, 1H), 9.23 (s, 1H), 8.82 (d, $J = 7.7$ Hz, 2H), 8.63 (s, 4H), 8.08 (d, $J = 8.1$ Hz, 1H), 7.79 (dd, $J = 16.3, 7.9$ Hz, 3H), 7.60 (t, $J = 8.1$ Hz, 1H). ^{13}C NMR

(101 MHz, DMSO- d_6) δ 156.07, 154.90, 152.60, 148.64, 143.76, 142.66, 137.67, 135.36, 130.16, 129.72, 129.68, 124.84, 122.81, 115.77, 112.43, 42.34. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{15}N_7O_2$ $[M + H]^+$ 362.1360, found: 362.1356.

6-(3-(2-aminoethyl)phenyl)-*N*-(3-nitrophenyl)-9*H*-purin-2-amine hydrochloride (11e). Yellow solid; Yield: 54%; mp: >230°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.24 (s, 1H), 9.21 (s, 1H), 8.77–8.52 (m, 3H), 8.19 (s, 3H), 8.11 (d, J = 8.2 Hz, 1H), 7.81 (d, J = 7.9 Hz, 1H), 7.61 (dd, J = 12.0, 7.7 Hz, 2H), 7.53 (d, J = 7.4 Hz, 1H), 3.23–3.00 (m, 4H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.41, 154.21, 153.39, 153.02, 148.60, 146.95, 143.64, 142.37, 138.68, 138.57, 135.14, 132.34, 130.19, 129.84, 129.59, 128.16, 124.99, 116.04, 112.60, 40.25 33.35. HRMS (AP-ESI) m/z Calcd for $C_{19}H_{17}N_7O_2$ $[M + H]^+$ 376.1516, found: 376.1520.

6-(3-(aminomethyl)phenyl)-*N*-phenyl-9*H*-purin-2-amine hydrochloride (11f). Light yellow solid; Yield: 92%; m.p.: 203–205°C; 1H NMR (400 MHz, DMSO- d_6) δ 9.97 (s, 1H), 9.09 (s, 1H), 8.77 (s, 3H), 8.58 (s, 1H), 8.48 (d, J = 7.5 Hz, 1H), 7.87 (d, J = 7.8 Hz, 2H), 7.81 (d, J = 7.4 Hz, 1H), 7.69 (t, J = 7.5 Hz, 1H), 7.34 (t, J = 7.4 Hz, 2H), 7.00 (t, J = 7.1 Hz, 1H), 4.18 (d, J = 4.5 Hz, 2H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 157.16, 154.55, 152.94, 143.13, 140.84, 135.06, 132.67, 130.59, 129.66, 129.44, 129.07, 122.15, 119.43, 42.69. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{16}N_6$ $[M + H]^+$ 317.1509, found: 317.1507.

6-(4-(aminomethyl)phenyl)-*N*-phenyl-9*H*-purin-2-amine hydrochloride (11g). Light yellow solid; Yield: 92%; m.p.: 200–202°C; 1H NMR (400 MHz, DMSO- d_6) δ 9.53 (s, 1H), 8.81 (d, J = 8.09 Hz, 2H), 8.36 (s, 1H), 7.87 (d, J = 8.05 Hz, 2H), 7.70 (d, J = 8.21 Hz, 2H), 7.32 (t, J = 7.85 Hz, 2H), 6.95 (t, J = 7.27 Hz, 1H), 4.15 (q, J = 5.90 Hz, 2H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.05, 155.04, 152.76, 144.44, 143.60, 137.54, 136.35, 135.56, 129.73, 129.66, 127.06, 123.46, 117.98, 42.38. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{16}N_6$ $[M + H]^+$ 317.1509, found: 317.1507.

***N*-([1,1'-biphenyl]-4-yl)-6-(3-(aminomethyl)phenyl)-9*H*-purin-2-amine hydrochloride (11h).** Yellow solid; Yield: 89%; m.p.: 210°C (Dec.); 1H NMR (400 MHz, DMSO- d_6) δ 9.80 (s, 1H), 8.75 (d, J = 7.7 Hz, 2H), 8.57 (s, 1H), 8.49 (s, 3H), 8.00 (d, J = 7.8 Hz, 2H), 7.70 (dt, J = 13.3, 7.0 Hz, 6H), 7.45 (t, J = 7.2 Hz, 2H), 7.32 (t, J = 7.4 Hz, 1H), 4.18 (d, J = 5.2 Hz, 2H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 157.13, 154.45, 153.07, 143.16, 140.41, 135.09, 133.75, 132.71, 130.59, 129.69, 129.49, 129.34, 128.34, 127.29, 127.22, 126.60, 124.40, 119.66, 42.69. HRMS (AP-ESI) m/z Calcd for $C_{24}H_{20}N_6$ $[M + H]^+$ 393.1822, found: 393.1826.

6-(3-(aminomethyl)phenyl)-*N*-(*p*-tolyl)-9*H*-purin-2-amine hydrochloride (11i). Light yellow solid, Yield: 90%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 9.85 (s, 1H), 9.03 (s, 1H), 8.72 (s, 3H), 8.59 (s, 1H), 8.50 (d, J = 7.5 Hz, 1H), 7.73 (dq, J = 23.2, 7.6 Hz, 4H), 7.16 (d, J = 7.9 Hz, 2H), 4.18 (d, J = 5.0 Hz, 2H), 2.29 (s, 3H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 157.20, 154.70, 152.84, 143.01, 138.28, 135.16, 135.03, 132.59, 131.02, 130.57, 129.65, 129.48, 119.57, 60.23, 42.72. HRMS (AP-ESI) m/z Calcd for $C_{19}H_{18}N_6$ $[M + H]^+$ 331.1666, found: 331.1667.

6-(3-(aminomethyl)phenyl)-*N*-(4-(*tert*-butyl)phenyl)-9*H*-purin-2-amine hydrochloride (11j). Yellow solid; Yield: 64%; 260°C (Dec.); 1H NMR (400 MHz, DMSO- d_6) δ 9.50 (s, 1H), 8.78–8.69 (m, 2H),

8.48 (d, J = 9.8 Hz, 4H), 7.77 (d, J = 7.9 Hz, 2H), 7.74–7.65 (m, 2H), 7.34 (d, J = 8.0 Hz, 2H), 4.16 (d, J = 5.2 Hz, 3H), 1.31 (d, J = 12.0 Hz, 9H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 157.29, 154.55, 152.88, 144.65, 143.05, 138.07, 135.06, 134.95, 132.70, 130.59, 129.68, 129.42, 125.68, 119.45, 42.68, 34.42, 31.78. HRMS (AP-ESI) m/z Calcd for $C_{22}H_{24}N_6$ $[M + H]^+$ 373.2135, found: 373.213.

6-(3-(aminomethyl)phenyl)-*N*-(4-(piperazin-1-yl)phenyl)-9*H*-purin-2-amine hydrochloride (11k). Yellow solid; Yield: 91%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 9.68 (s, 1H), 9.42 (s, 2H), 8.78 (s, 1H), 8.69–8.53 (m, 4H), 7.83–7.72 (m, 3H), 7.68 (t, J = 7.7 Hz, 1H), 7.08 (d, J = 8.3 Hz, 2H), 4.17 (d, J = 5.2 Hz, 2H), 3.27 (s, 4H), 1.99 (s, 4H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 157.79, 156.51, 154.95, 152.82, 142.99, 134.96, 132.41, 130.49, 129.62, 129.52, 120.64, 117.93, 47.38, 42.79. HRMS (AP-ESI) m/z Calcd for $C_{22}H_{24}N_8$ $[M + H]^+$ 401.2197, found: 401.2193.

4-((6-(3-(aminomethyl)phenyl)-9*H*-purin-2-yl)amino)benzenesulfonamide hydrochloride (11l). Yellow solid; Yield: 96%; 1H NMR (400 MHz, DMSO- d_6) δ 10.04 (s, 2H), 8.79 (s, 3H), 8.73–8.59 (m, 1H), 8.50 (d, J = 5.42 Hz, 2H), 8.44 (s, 7H), 8.09–8.02 (m, 4H), 7.82–7.63 (m, 7H), 7.20 (s, 3H), 4.18 (d, J = 6.00 Hz, 4H), 4.03 (t, J = 6.87 Hz, 1H), 2.00 (d, J = 1.84 Hz, 1H), 1.26–1.14 (m, 1H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.14, 152.84, 144.57, 143.72, 136.21, 134.93, 132.10, 130.39, 129.81, 129.53, 127.11, 117.89, 42.89. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{17}N_7O_2S$ $[M + H]^+$ 396.1237, found: 396.1234.

***N*'-(6-(3-(aminomethyl)phenyl)-9*H*-purin-2-yl)benzene-1,3-diamine dihydrochloride (11m).** Yellow solid; Yield: 96%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.52 (s, 2H), 9.95 (s, 1H), 8.84 (d, J = 6.78 Hz, 1H), 8.73 (s, 1H), 8.54 (s, 3H), 8.22 (s, 1H), 7.73 (ddd, J = 25.27, 17.03, 7.59 Hz, 3H), 7.43 (t, J = 8.13 Hz, 1H), 6.99 (d, J = 7.67 Hz, 1H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.47, 152.82, 143.75, 142.57, 135.93, 135.01, 132.49, 132.38, 130.37, 130.26, 130.15, 129.55, 118.43, 115.95, 113.36, 42.78. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{17}N_7$ $[M + H]^+$ 332.1618, found: 332.1621.

6-(3-(aminomethyl)phenyl)-*N*-(4-fluorophenyl)-9*H*-purin-2-amine hydrochloride (11n). Light yellow solid; Yield: 99%; m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 9.93 (s, 1H), 8.99 (s, 1H), 8.72 (s, 3H), 8.59 (s, 1H), 8.52 (d, J = 7.3 Hz, 1H), 7.87 (dd, J = 7.6, 5.3 Hz, 3H), 7.79 (d, J = 7.4 Hz, 1H), 7.69 (t, J = 7.7 Hz, 2H), 7.20 (t, J = 8.5 Hz, 3H), 4.18 (d, J = 5.2 Hz, 2H); ^{13}C NMR (101 MHz, DMSO- d_6) δ 158.88, 157.13, 156.51, 154.70, 153.03, 143.05, 137.39, 135.37, 135.02, 132.53, 130.49, 129.62, 129.52, 121.07, 120.99, 115.68, 115.46, 60.23, 42.72. HRMS (AP-ESI) m/z Calcd for $C_{18}H_{15}FN_6$ $[M + H]^+$ 335.1415, found: 335.1418.

4-((6-(3-(aminomethyl)phenyl)-9*H*-purin-2-yl)amino)-*N*-methylbenzenesulfonamide hydrochloride (11p). Yellow solid; Yield: 91%, m.p.: >300°C; 1H NMR (400 MHz, DMSO- d_6) δ 10.16 (s, 1H), 8.77–8.71 (m, 1H), 8.63 (s, 0H), 8.56 (s, 1H), 8.14–8.07 (m, 1H), 7.73 (dt, J = 14.44, 7.08 Hz, 2H), 4.74 (s, 9H), 4.18 (q, J = 5.95 Hz, 1H), 2.41 (s, 3H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.04, 152.84, 145.17, 136.21, 134.93, 132.09, 130.79, 130.35, 129.88, 129.55, 128.28, 118.04, 42.89, 29.18. HRMS (AP-ESI) m/z Calcd for $C_{19}H_{19}N_7O_2S$ $[M + H]^+$ 409.4680, found: 409.4675.

4-((6-(3-(aminomethyl)phenyl)-9H-purin-2-yl)amino)-N,N-dimethylbenzene sulfonamide hydrochloride (11q). Yellow solid; Yield: 91%, m.p.: >300°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.34 (s, 1H), 8.86 (s, 1H), 8.67 (t, J = 11.0 Hz, 5H), 8.16 (d, J = 8.1 Hz, 2H), 7.86–7.67 (m, 4H), 4.19 (d, J = 5.2 Hz, 2H), 2.61 (s, 6H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 156.20, 155.13, 152.97, 145.64, 143.91, 135.76, 135.03, 132.38, 130.39, 129.77, 129.62, 129.24, 126.10, 118.14, 42.78, 38.18. HRMS (AP-ESI) m/z Calcd for $\text{C}_{20}\text{H}_{21}\text{N}_7\text{O}_2\text{S}$ [$\text{M} + \text{H}$] $^+$ 424.1550, found: 424.1554.

6-(3-(aminomethyl)phenyl)-N-(4-(piperazin-1-yl)phenyl)-9H-purin-2-amine dihydrochloride (11r). Yellow solid; Yield: 49%; m.p.: >300°C; ^1H NMR (400 MHz, DMSO- d_6) δ 10.70 (s, 1H), 9.73 (s, 1H), 8.86 (d, J = 7.1 Hz, 1H), 8.72 (s, 1H), 8.67 (d, J = 8.6 Hz, 1H), 8.54 (d, J = 7.2 Hz, 4H), 8.08–8.01 (m, 1H), 7.75 (d, J = 7.3 Hz, 1H), 7.69 (t, J = 7.6 Hz, 1H), 4.24 (d, J = 5.4 Hz, 2H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 152.82, 142.99, 134.96, 132.41, 130.49, 129.62, 129.52, 120.64, 117.93, 47.38, 42.79. HRMS (AP-ESI) m/z Calcd for $\text{C}_{17}\text{H}_{15}\text{N}_7$ [$\text{M} + \text{H}$] $^+$ 318.1462, found: 318.1457.

4.1.5 General Method for the Preparation of Compounds 5g, 5j

3-(2-((3-nitrophenyl)amino)-9H-purin-6-yl)benzoic acid (5g). To a solution of compound **5h** (0.48 mmol) in THF/ H_2O solution (4:1, 10 ml), LiOH (1.5 mmol) was added, and the mixture was stirred at rt for 4 h. The mixture was adjusted to around pH 2 with HCl solution (2M), and the mixture was extracted twice with ethyl acetate (50 ml), washed with brine, and dried with anhydrous Mg_2SO_4 . The solution was concentrated to get compound **5g**. White solid; Yield: 69%; m.p.: >300°C; ^1H NMR (400 MHz, DMSO- d_6) δ 13.30 (s, 1H), 13.15 (s, 1H), 10.19 (s, 1H), 9.51 (s, 1H), 9.11 (d, J = 7.8 Hz, 1H), 9.08 (s, 1H), 8.42 (s, 1H), 8.16 (dd, J = 16.5, 7.9 Hz, 2H), 7.80 (d, J = 8.0 Hz, 1H), 7.75 (t, J = 7.6 Hz, 1H), 7.60 (t, J = 8.1 Hz, 1H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 167.63, 155.84, 155.28, 152.52, 148.72, 143.46, 142.97, 136.56, 133.80, 131.96, 131.76, 131.12, 130.84, 130.14, 129.38, 126.06, 124.70, 115.53, 112.38. HRMS (AP-ESI) m/z Calcd for $\text{C}_{18}\text{H}_{12}\text{N}_6\text{O}_4$ [$\text{M} + \text{H}$] $^+$ 377.0993, found: 377.0998.

Compound **5j** was synthesized following the procedure described above.

4-(2-((3-nitrophenyl)amino)-9H-purin-6-yl)benzoic acid (5j). Yellow solid; Yield: 97%; m.p.: >300°C; ^1H NMR (400 MHz, DMSO- d_6) δ 13.31 (s, 2H), 10.21 (s, 1H), 10.17 (s, 2H), 9.17 (s, 2H), 9.00 (t, J = 8.00 Hz, 4H), 8.42 (s, 2H), 8.17 (t, J = 7.15 Hz, 4H), 8.11 (d, J = 8.38 Hz, 2H), 7.80 (d, J = 8.20 Hz, 2H), 7.60 (t, J = 8.27 Hz, 2H), 7.56–7.47 (m, 1H). ^{13}C NMR (101 MHz, DMSO- d_6) δ 166.40, 155.80, 155.55, 148.74, 143.86, 142.94, 140.53, 131.68, 130.68, 130.16, 129.90, 129.76, 124.70, 115.54, 112.31, 52.81. HRMS (AP-ESI) m/z Calcd for $\text{C}_{18}\text{H}_{12}\text{N}_6\text{O}_4$ [$\text{M} + \text{H}$] $^+$ 377.0993, found: 377.0998.

4.1.6 General Method for the Preparation of Compounds 7a–7e

Tert-butyl (4-bromophenethyl)carbamate (7d). Compounds **6d** (2 mmol) and Di-tert-butyl dicarbonate (2.4 mmol) were dissolved in dichloromethane (25 ml). K_2CO_3 (6 mmol) was added and stirred

for 4 h at room temperature. After the completion, the reaction mixture is extracted with ethyl acetate (15 ml \times 3), washed with water, 1M citric acid solution, and brine, and dried with anhydrous Mg_2SO_4 . The crude product were concentrated and purified by silica gel chromatography to obtain compound **7d**. Oil; Yield: 90%; m.p.: 44–46°C; ^1H NMR (400 MHz, CDCl_3) δ 7.36 (d, J = 7.7 Hz, 2H), 7.20–7.10 (m, 2H), 4.56 (s, 1H), 3.36 (d, J = 6.2 Hz, 2H), 2.77 (t, J = 6.5 Hz, 2H), 1.44 (s, 9H).

Compounds **7a–7c** and **7e** were synthesized following the procedure described above.

Tert-butyl (3-bromophenyl)carbamate (7a). White solid; Yield: 70%; m.p.: 85–87°C; ^1H NMR (400 MHz, CDCl_3) δ 7.67 (s, 1H), 7.21 (d, J = 7.1 Hz, 1H), 7.18–7.10 (m, 2H), 6.49 (s, 1H), 1.52 (s, 9H).

Tert-butyl (4-bromophenyl)carbamate (7b). White solid; Yield: 77%; m.p.: 62–64°C; ^1H NMR (400 MHz, CDCl_3) δ 7.39 (d, J = 8.3 Hz, 2H), 7.25 (d, J = 7.2 Hz, 2H), 6.46 (s, 1H), 1.51 (s, 9H).

Tert-butyl (3-bromobenzyl)carbamate (7c). White solid; Yield: 60%; m.p.: 55–57°C; ^1H NMR (400 MHz, CDCl_3) δ 7.43 (s, 1H), 7.39 (d, J = 6.4 Hz, 1H), 7.20 (d, J = 6.0 Hz, 2H), 4.86 (s, 1H), 4.29 (d, J = 5.1 Hz, 2H), 1.46 (s, 9H).

Tert-butyl (3-bromophenethyl)carbamate (7e). White solid; Yield: 100%; m.p.: 44–46°C; ^1H NMR (400 MHz, CDCl_3) δ 7.36 (d, J = 7.7 Hz, 2H), 7.20–7.10 (m, 2H), 4.56 (s, 1H), 3.36 (d, J = 6.2 Hz, 2H), 2.77 (t, J = 6.5 Hz, 2H), 1.44 (s, 9H).

4.1.7 General Method for the Preparation of Compounds 8a–8e

Tert-butyl(4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)benzyl)carbamate (8d). Compound **7d** (4.5 mmol), bis(pinacolato)diboron (4.5 mmol), Pd (dppf) $_2\text{Cl}_2$ (0.05 mmol), and KOAc (13.5 mmol) were mixed in a two-neck flask. Under the protection of N_2 , anhydrous DMSO (10 ml) was added and the mixture reacted at 80°C for 12 h. After the completion, the reaction mixture was filtered through a pad of Celite. Spinned the filtrate dry and then dissolved it with ethyl acetate (15 ml) and water (20 ml), extracted twice with ethyl acetate (50 ml), washed with brine, and dried with anhydrous Mg_2SO_4 . The crude product was concentrated and purified by silica gel chromatography to obtain compounds **8d**. Oil; Yield: 90%; ^1H NMR (400 MHz, CDCl_3) δ 7.83 (d, J = 7.4 Hz, 2H), 7.09 (d, J = 7.4 Hz, 2H), 2.56 (s, 2H), 1.37 (s, 9H), 1.34 (s, 12H).

Compounds **8a–8c** and **8e** were synthesized following the procedure described above.

Tert-butyl (3-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)phenyl)carbamate (8a). White solid; Yield: 93%; m.p.: 108–110°C; ^1H NMR (400 MHz, CDCl_3) δ 7.61 (s, 2H), 7.47 (d, J = 7.2 Hz, 1H), 7.31 (t, J = 7.8 Hz, 1H), 6.46 (s, 1H), 1.51 (s, 9H), 1.33 (s, 12H).

Tert-butyl (4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)phenyl)carbamate (8b). White solid; Yield: 63%. The product is put into the next step without purification.

Tert-butyl (3-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)benzyl)carbamate (8c). White solid; Yield: 70%; ^1H NMR (400 MHz, DMSO- d_6) δ 7.57 (s, 1H), 7.52 (d, J = 6.4 Hz, 1H), 7.40 (s, 1H), 7.34 (t, J = 8.1 Hz, 2H), 4.12 (d, J = 5.8 Hz, 2H), 1.39 (s, 9H), 1.29 (s, 12H).

Tert-butyl (3-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)phenethyl)carbamate (8e). Oil; Yield: 58%; ^1H NMR (400 MHz,

CDCl₃) δ 7.67 (d, J = 6.6 Hz, 1H), 7.64 (s, 1H), 7.38–7.28 (m, 2H), 4.52 (s, 1H), 3.38 (d, J = 6.1 Hz, 2H), 2.80 (t, J = 6.6 Hz, 2H), 1.43 (s, 9H), 1.35 (s, 12H).

4.2 Cyclin-Dependent Kinases Inhibition Test

Experiments were carried out using the Kinase-Glo[®] Luminescent Kinase Assays as described previously (Kashem et al., 2007). Briefly, all enzymatic reactions were conducted at 30°C for 40 min. The 50 μ l reaction mixture contains 40 mM Tris, pH 7.4, 10 mM MgCl₂, 0.1 mg/ml BSA, 1 mM DTT, 10 μ M ATP, 0.2 μ g/ml CDKs, and 100 μ M lipid substrate. The compounds were diluted with 10% DMSO and then 5 μ l of the dilution was removed and put into the subsequent reaction. The kinase activities were measured by detecting the content of remaining ATP. The luminescent signal was correlated with the amount of residual ATP and negatively correlated with the amount of kinase activity. The IC₅₀ values were calculated using Prism GraphPad software.

4.3 Anti-proliferation Test

Standard MTT (thiazolyl blue; 3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyltetrazolium bromide) assays were performed as 5 mg/ml. Briefly, MDA-MB-231 or 293T cells were seeded into 96-well plates and incubated for 24 h at 37°C. All compounds were dissolved in DMSO, and a gradient dilution series were prepared in 100 μ l of cell medium, added to cells (in triplicates), and incubated for 48 h at 37°C with 5% CO₂. MTT was added (5 mg/ml, 20 μ l) to each plate and these mixtures were incubated for another 4 h. Then, the medium was removed, and the mixture was completely dissolved in DMSO (200 μ L) after shaking for 10 min. The absorbance was recorded at 490 nm (detection wavelength) and 630 nm (reference wavelength) and inhibition rates were calculated to determine IC₅₀ values.

4.4 Cell Cycles

MDA-MB-231 cells were seeded in six-well plates and incubated with 20 μ M compounds **11c**, **11l**, **11p**, and vehicle (0.2% DMSO) for 24 h. Subsequently, cells were centrifugated and washed with cold PBS buffer. After the centrifugation, the supernatants were removed, and the cells were resuspended in PBS buffer. Then, 10 μ l of PI were added and the cells were incubated in the dark for 15 min at room temperature. The stained cells were analyzed by a flow cytometer (BD Accuri C6).

4.5 Molecular Dynamics Simulation

Based on the crystal structure of CDK2–inhibitor complex (PDB: 5NEV), we performed molecular docking used by

AutoDock Vina to obtain the initial structure complex for molecular dynamics simulation. Molecular dynamics simulations of CDK2–**11l** complex were carried out employing Amber16 package. The Amber14SB force field was used for proteins, and the TIP3P model was used for water molecules. The partial charge of **11l** was assigned using AM1-BCC methods via antechamber. The system was neutralized with Cl-counterions and solvated in a rectangular periodic box with explicit TIP3P water using AmberTools17. The solvation system consists of ~30,000 atoms. The Particle-mesh Ewald method for nonbonded interactions is used for MD simulation. After a series of minimization and equilibration, standard molecular dynamics simulations were performed on the GPU using the CUDA version of PMEMD (Particle Mesh Ewald Molecular Dynamics) for 50 ns with periodic boundary conditions. The SHAKE algorithm is used to constrain all the bonds involving hydrogen atoms. A time step of 2 fs was used and the system temperature was controlled at 300K using the Berendsen thermostat method. The snapshots were saved every 10 ps for analysis. All other parameters are default.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

HL and YZ performed synthetic work and wrote the manuscript. ZZ designed target compound and performed molecular docking and molecular dynamics simulation. JD performed *in vitro* biological experiments. XY performed data analysis. XH and HF designed and supervised the study, revised manuscript and provided materials. HL, YZ, and ZZ contributed equally to this work.

FUNDING

This work was supported by the National Natural Science Foundation of China (81874288, 82003590, and 92053105), the Natural Science Foundation of Shandong Province (ZR2019LZL004 and ZR2020QH342), and the Young Scholars Program of Shandong University.

REFERENCES

- Adams, P. D. (2001). Regulation of the Retinoblastoma Tumor Suppressor Protein by Cyclin/cdk. *Biochim. Biophys. Acta* 1471, M123–M133. doi:10.1016/s0304-419x(01)00019-1
- Aleem, E., Berthet, C., and Kaldis, P. (2004). Cdk2 as a Master of S Phase Entry: Fact or Fake? *Cell Cycle* 3, 35–37. doi:10.4161/cc.3.1.632
- Braal, C. L., Jongbloed, E. M., Wilting, S. M., Mathijssen, R. H. J., Koolen, S. L. W., and Jager, A. (2021). Inhibiting CDK4/6 in Breast Cancer with Palbociclib, Ribociclib, and Abemaciclib: Similarities and Differences. *Drugs* 81, 317–331. doi:10.1007/s40265-020-01461-2
- Chohan, T. A., Qian, H., Pan, Y., and Chen, J. Z. (2015). Cyclin-Dependent Kinase-2 as a Target for Cancer Therapy: Progress in the Development of CDK2 Inhibitors as Anti-cancer Agents. *Curr. Med. Chem.* 22, 237–263. doi:10.2174/0929867321666141106113633

- Coxon, C. R., Anscombe, E., Harnor, S. J., Martin, M. P., Carbain, B., Golding, B. T., et al. (2017). Cyclin-Dependent Kinase (CDK) Inhibitors: Structure-Activity Relationships and Insights into the CDK-2 Selectivity of 6-Substituted 2-Arylamino-purines. *J. Med. Chem.* 60, 1746–1767. doi:10.1021/acs.jmedchem.6b01254
- Deans, A. J., Khanna, K. K., Mcnees, C. J., Mercurio, C., Heierhorst, J., and Mcarthur, G. A. (2006). Cyclin-dependent Kinase 2 Functions in normal DNA Repair and Is a Therapeutic Target in BRCA1-Deficient Cancers. *Cancer Res.* 66, 8219–8226. doi:10.1158/0008-5472.CAN-05-3945
- Fagundes, R., and Teixeira, L. K. (2021). Cyclin E/CDK2: DNA Replication, Replication Stress and Genomic Instability. *Front. Cel. Dev. Biol.* 9, 774845. doi:10.3389/fcell.2021.774845
- Fry, D. W., Harvey, P. J., Keller, P. R., Elliott, W. L., Meade, M., Trachet, E., et al. (2004). Specific Inhibition of Cyclin-dependent Kinase 4/6 by PD 0332991 and Associated Antitumor Activity in Human Tumor Xenografts. *Mol. Cancer Ther.* 3, 1427–1438. doi:10.1158/1535-7163.1427.3.11
- Golsteyn, R. M. (2005). Cdk1 and Cdk2 Complexes (Cyclin Dependent Kinases) in Apoptosis: a Role beyond the Cell Cycle. *Cancer Lett.* 217, 129–138. doi:10.1016/j.canlet.2004.08.005
- Gomatou, G., Trontzas, I., Ioannou, S., Drizou, M., Syrigos, N., and Kotteas, E. (2021). Mechanisms of Resistance to Cyclin-dependent Kinase 4/6 Inhibitors. *Mol. Biol. Rep.* 48, 915–925. doi:10.1007/s11033-020-06100-3
- Hydbring, P., and Larsson, L. G. (2010). Cdk2: a Key Regulator of the Senescence Control Function of Myc. *Aging (Albany NY)* 2, 244–250. doi:10.18632/aging.100140
- Jin, X., Ge, L. P., Li, D. Q., Shao, Z. M., Di, G. H., Xu, X. E., et al. (2020). LncRNA TROJAN Promotes Proliferation and Resistance to CDK4/6 Inhibitor via CDK2 Transcriptional Activation in ER+ Breast Cancer. *Mol. Cancer* 19, 87. doi:10.1186/s12943-020-01210-9
- Joaquin, M., and Watson, R. J. (2003). Cell Cycle Regulation by the B-Myb Transcription Factor. *Cell Mol. Life Sci.* 60, 2389–2401. doi:10.1007/s00018-003-3037-4
- Julve, M., Clark, J. J., and Lythgoe, M. P. (2021). Advances in Cyclin-dependent Kinase Inhibitors for the Treatment of Melanoma. *Expert Opin. Pharmacother.* 22, 351–361. doi:10.1080/14656566.2020.1828348
- Kashem, M. A., Nelson, R. M., Yingling, J. D., Pullen, S. S., Prokopowicz, A. S., Jones, J. W., et al. (2007). Three Mechanistically Distinct Kinase Assays Compared: Measurement of Intrinsic ATPase Activity Identified the Most Comprehensive Set of ITK Inhibitors. *J. Biomol. Screen.* 12, 70–83. doi:10.1177/1087057106296047
- Kimball, S. D., and Webster, K. R. (2001). Chapter 14. Cell Cycle Kinases and Checkpoint Regulation in Cancer. *Annu. Rep. Med. Chem.* Vol. 36, 139–148. ed. A.M. Doherty. doi:10.1016/S0065-7743(01)36054-2
- Lee, K. A., Shepherd, S. T., and Johnston, S. R. (2019). Abemaciclib, a Potent Cyclin-dependent Kinase 4 and 6 Inhibitor, for Treatment of ER-Positive Metastatic Breast Cancer. *Future Oncol.* 15, 3309–3326. doi:10.2217/fon-2019-0169
- Liu, F. (2006). Smad3 Phosphorylation by Cyclin-dependent Kinases. *Cytokine Growth Factor. Rev.* 17, 9–17. doi:10.1016/j.cytogfr.2005.09.010
- Lu, M. J., Li, W., Yang, X. Y., and Fang, H. (2019). Synthesis and Antitumor Activity of N9 Position Aromatic Substituted Purine-8-One Derivatives. *Chem. J. Chin. Universities* 40, 254–261. doi:10.7503/cjcu20180573
- Meijer, L., Borgne, A., Mulner, O., Chong, J. P., Blow, J. J., Inagaki, N., et al. (1997). Biochemical and Cellular Effects of Roscovitine, a Potent and Selective Inhibitor of the Cyclin-dependent Kinases Cdc2, Cdk2 and Cdk5. *Eur. J. Biochem.* 243, 527–536. doi:10.1111/j.1432-1033.1997.t01-2-00527.x
- Narasimha, A. M., Kaulich, M., Shapiro, G. S., Choi, Y. J., Sicinski, P., and Dowdy, S. F. (2014). Cyclin D Activates the Rb Tumor Suppressor by Mono-Phosphorylation. *Elife* 3, e02872. doi:10.7554/eLife.02872
- Nie, L., Wei, Y., Zhang, F., Hsu, Y. H., Chan, L. C., Xia, W., et al. (2019a). CDK2-mediated Site-specific Phosphorylation of EZH2 Drives and Maintains Triple-Negative Breast Cancer. *Nat. Commun.* 10, 5114. doi:10.1038/s41467-019-13105-5
- Nie, L., Wei, Y., Zhang, F., Hsu, Y. H., Chan, L. C., Xia, W., et al. (2019b). CDK2-mediated Site-specific Phosphorylation of EZH2 Drives and Maintains Triple-Negative Breast Cancer. *Nat. Commun.* 10, 5114. doi:10.1038/s41467-019-13105-5
- Pandey, K., An, H. J., Kim, S. K., Lee, S. A., Kim, S., Lim, S. M., et al. (2019). Molecular Mechanisms of Resistance to CDK4/6 Inhibitors in Breast Cancer: A Review. *Int. J. Cancer* 145, 1179–1188. doi:10.1002/ijc.32020
- Pierson-Mullany, L. K., and Lange, C. A. (2004). Phosphorylation of Progesterone Receptor Serine 400 Mediates Ligand-independent Transcriptional Activity in Response to Activation of Cyclin-dependent Protein Kinase 2. *Mol. Cell Biol.* 24, 10542–10557. doi:10.1128/MCB.24.24.10542-10557.2004
- Rao, S. S., Stoehr, J., Dokic, D., Wan, L., Decker, J. T., Konopka, K., et al. (2017). Synergistic Effect of Eribulin and CDK Inhibition for the Treatment of Triple Negative Breast Cancer. *Oncotarget* 8, 83925–83939. doi:10.18632/oncotarget.20202
- Rice, A. P. (2019). Roles of CDKs in RNA Polymerase II Transcription of the HIV-1 Genome. *Transcription* 10, 111–117. doi:10.1080/21541264.2018.1542254
- Roufayel, R., and Murshid, N. (2019). CDK5: Key Regulator of Apoptosis and Cell Survival. *Biomedicine* 7, 88. doi:10.3390/biomedicine7040088
- Sánchez-Martínez, C., Lallena, M. J., Sanfeliciano, S. G., and De Dios, A. (2019). Cyclin Dependent Kinase (CDK) Inhibitors as Anticancer Drugs: Recent Advances (2015–2019). *Bioorg. Med. Chem. Lett.* 29, 126637. doi:10.1016/j.bmcl.2019.126637
- Satyanarayana, A., and Kaldis, P. (2009). Mammalian Cell-Cycle Regulation: Several Cdks, Numerous Cyclins and Diverse Compensatory Mechanisms. *Oncogene* 28, 2925–2939. doi:10.1038/onc.2009.170
- Sviderskiy, V. O., Blumenberg, L., Gorodetsky, E., Karakousi, T. R., Hirsh, N., Alvarez, S. W., et al. (2020). Hyperactive CDK2 Activity in Basal-like Breast Cancer Imposes a Genome Integrity Liability that Can Be Exploited by Targeting DNA Polymerase ϵ . *Mol. Cell* 80, 682–e7. doi:10.1016/j.molcel.2020.10.016
- Tadesse, S., Anshabo, A. T., Portman, N., Lim, E., Tilley, W., Caldon, C. E., et al. (2020). Targeting CDK2 in Cancer: Challenges and Opportunities for Therapy. *Drug Discov. Today* 25, 406–413. doi:10.1016/j.drudis.2019.12.001
- Tadesse, S., Caldon, E. C., Tilley, W., and Wang, S. (2019a). Cyclin-Dependent Kinase 2 Inhibitors in Cancer Therapy: An Update. *J. Med. Chem.* 62, 4233–4251. doi:10.1021/acs.jmedchem.8b01469
- Tadesse, S., Caldon, E. C., Tilley, W., and Wang, S. (2019b). Cyclin-Dependent Kinase 2 Inhibitors in Cancer Therapy: An Update. *J. Med. Chem.* 62, 4233–4251. doi:10.1021/acs.jmedchem.8b01469
- Tarasiewicz, E., Rivas, L., Hamdan, R., Dokic, D., Parimi, V., Bernabe, B. P., et al. (2014). Inhibition of CDK-Mediated Phosphorylation of Smad3 Results in Decreased Oncogenesis in Triple Negative Breast Cancer Cells. *Cell Cycle* 13, 3191–3201. doi:10.4161/15384101.2014.950126
- Tripathy, D., Bardia, A., and Sellers, W. R. (2017). Ribociclib (LEE011): Mechanism of Action and Clinical Impact of This Selective Cyclin-dependent Kinase 4/6 Inhibitor in Various Solid Tumors. *Clin. Cancer Res.* 23, 3251–3262. doi:10.1158/1078-0432.CCR-16-3157
- Volkart, P. A., Bitencourt-Ferreira, G., Souto, A. A., and De Azevedo, W. F., Jr. (2019). Cyclin-Dependent Kinase 2 in Cellular Senescence and Cancer. A Structural and Functional Review. *Curr. Drug Targets* 20, 716–726. doi:10.2174/1389450120666181204165344
- Wang, J., Wang, Q., Zhang, L., and Fang, H. (2013). Design, Synthesis and Preliminary Biological Evaluation of Purine-2,6-Diamine Derivatives as Cyclin-dependent Kinase (CDK) Inhibitors. *Chin. J. Chem.* 31, 1181–1191. doi:10.1002/cjoc.201300420
- Wood, D. J., and Endicott, J. A. (2018). Structural Insights into the Functional Diversity of the CDK-Cyclin Family. *Open Biol.* 8, 180112. doi:10.1098/rsob.180112
- Yin, J., Zhao, M. M., Huffman, M. A., and Mcnamara, J. M. (2002). Pd-catalyzed N-Arylation of Heteroarylamines. *Org. Lett.* 4, 3481–3484. doi:10.1021/ol0265923
- Yuan, K., Wang, X., Dong, H., Min, W., Hao, H., and Yang, P. (2021). Selective Inhibition of CDK4/6: A Safe and Effective Strategy for Developing Anticancer Drugs. *Acta Pharm. Sin. B* 11, 30–54. doi:10.1016/j.apsb.2020.05.001

Zhang, J., Wang, Q., Hou, X., and Liu, H. (2015). Recent Advances in Cyclin-dependent Kinase Inhibitors with Purine Scaffold. *Chin. J. Org. Chem.* 35, 1022. doi:10.6023/cjoc201410039

Zhu, Y., Hu, Y., Tang, C., Guan, X., and Zhang, W. (2022). Platinum-based Systematic Therapy in Triple-Negative Breast Cancer. *Biochim. Biophys. Acta Rev. Cancer* 1877, 188678. doi:10.1016/j.bbcan.2022.188678

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Liang, Zhu, Zhao, Du, Yang, Fang and Hou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Probabilistic Pocket Druggability Prediction via One-Class Learning

Riccardo Aguti^{1,2†}, Erika Gardini^{1,2†}, Martina Bertazzo¹, Sergio Decherchi^{1*} and Andrea Cavalli^{1,2}

¹Computational and Chemical Biology, Fondazione Istituto Italiano di Tecnologia, Genoa, Italy, ²Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

OPEN ACCESS

Edited by:

Leonardo L. G. Ferreira,
University of São Paulo, Brazil

Reviewed by:

Neelima Arora,
University Grants Commission, India
Alan Talevi,
National University of La Plata,
Argentina

*Correspondence:

Sergio Decherchi
sergio.decherchi@iit.it

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Experimental Pharmacology and Drug
Discovery,
a section of the journal
Frontiers in Pharmacology

Received: 06 February 2022

Accepted: 24 March 2022

Published: 29 June 2022

Citation:

Aguti R, Gardini E, Bertazzo M,
Decherchi S and Cavalli A (2022)
Probabilistic Pocket Druggability
Prediction via One-Class Learning.
Front. Pharmacol. 13:870479.
doi: 10.3389/fphar.2022.870479

The choice of target pocket is a key step in a drug discovery campaign. This step can be supported by *in silico* druggability prediction. In the literature, druggability prediction is often approached as a two-class classification task that distinguishes between druggable and non-druggable (or less druggable) pockets (or voxels). Apart from obvious cases, however, the non-druggable class is conceptually ambiguous. This is because any pocket (or target) is only non-druggable until a drug is found for it. It is therefore more appropriate to adopt a one-class approach, which uses only unambiguous information, namely, druggable pockets. Here, we propose using the import vector domain description (IVDD) algorithm to support this task. IVDD is a one-class probabilistic kernel machine that we previously introduced. To feed the algorithm, we use customized DrugPred descriptors computed via NanoShaper. Our results demonstrate the feasibility and effectiveness of the approach. In particular, we can remove or mitigate biases chiefly due to the labeling.

Keywords: druggability prediction, drug design, machine learning, unsupervised methods, one-class classification, import vector domain description, conception

1 INTRODUCTION

Drug discovery is a time-consuming and complex task (Nicolaou, 2014). It requires a multistep pipeline from biological understanding to fine-tuning of the lead candidate (for small molecules), often via computational means (Csermely et al., 2013; Jamali et al., 2016). In the past 20 years, computation has significantly contributed to many drug discovery steps via physics-based simulation, machine learning modeling, and the combination of the two (Decherchi and Cavalli, 2020b; Decherchi et al., 2021).

In particular, computational modeling can help find a putatively druggable target and hence a pocket that may accept a small molecule. A protein of interest is considered druggable when a drug has been found to inhibit it. However, some authors consider ligandability to be a more appropriate term for the propensity of the target/protein to accept drug-like molecules, irrespective of the more complex pharmacokinetic and pharmacodynamic mechanisms implied by the term druggability (Edfeldt et al., 2011). Here, we use the term druggable pocket to indicate a region of a protein with a high probability of accepting a small molecule. The reliable *in silico* identification of potentially druggable pockets is important for drug discovery. Finding new druggable hot spots can be particularly relevant when searching for an allosteric binder and to boost selectivity. Selectivity, in turn, is particularly important when designing chemical entities like PROTACs (Shimokawa et al., 2017; Qi et al., 2021), even more relevant than optimizing the affinity of the warhead itself. While researchers often know about the orthosteric pocket of a specific protein, it requires geometric and

chemical insights to detect alternate druggable pockets, making it a much more complex task. Effective tools are therefore required to support the computational medicinal chemist in detecting and ranking new pockets in order to design highly selective drugs.

The literature contains many reports on the computational estimation of druggability (Agoni et al., 2020). The available tools for this task include standalone software [e.g., P2Rank (Krivák and Hoksza, 2018)] and web servers [e.g., PockDrug (Hussein et al., 2015)]. Prediction often involves defining geometric and chemical features to support machine learning techniques (Xie et al., 2009) [e.g., DrugPred (Krasowski et al., 2011)]. Alternatively, more recent deep learning methodologies often use 3D grids (voxels) of physicochemical fields. Indeed, there are several methods for predicting the probability of a pocket's druggability. DoGSiteScorer (Volkamer et al., 2012b) is an algorithm that detects pockets and estimates druggability by considering global and local pocket properties. It uses support vector machines to build a predictive model. PRANK (Krivák and Hoksza, 2015) uses decision trees and random forests to re-rank/re-score the pockets predicted by other software, such as ConCavity (Capra et al., 2009) and Fpocket (Le Guilloux et al., 2009). PRANK could help improve the performance of existing prediction methods; it aims to predict the ligandability of a specific point near the surface of the pocket. TRAPP is a powerful method for analyzing molecular dynamics trajectories. It was recently endowed with druggability assessment capabilities, extending its analysis to an entire ensemble of structures (Yuan et al., 2020).

Druggability can also be assessed with pharmacophores (Desaphy et al., 2012) by using either very simple geometric considerations (e.g., Cavity (Yuan et al., 2013)) or fully fledged deep learning approaches. There are many such deep learning approaches, which often leverage convolutional neural networks coupled to 3D grids. In Zhang et al. (2020), the authors used both the pocket and the ligand with DenseNet architecture. In contrast, Pu et al. (2019) used convolutional neural networks specialized for nucleotide and heme-binding sites, again starting from 3D grids. InDeep (Mallet et al., 2021) is another contribution based on a convolutional architecture. Here, the focus is on characterizing protein-protein interfaces (PPI) to allow designing of PPI disruptors. The capabilities of convolutional neural networks were boosted by pocket segmentation in Aggarwal et al. (2021). This work and others [e.g., Stepniewska-Dziubinska et al. (2020)] demonstrated that both prediction and other activities, such as segmentation, are beneficial, so one can devise a more complex framework than a pure predictor. Along these lines, PURESNet (Kandel et al., 2021) uses an interesting deep residual (skip connections) decoder/encoder architecture derived from the U-net concept. This work presented both an architecture and a cleanup procedure for the training set. This class of deep methods is very accurate but lacks native interpretability.

From the protein dataset perspective, some datasets used in published works are suitable benchmarks. They are often used to train and test machine learning protocols, thus creating a shared base. For instance, in Hajduk et al. (2005), the authors created an online dataset containing 72 unique protein-binding sites. The

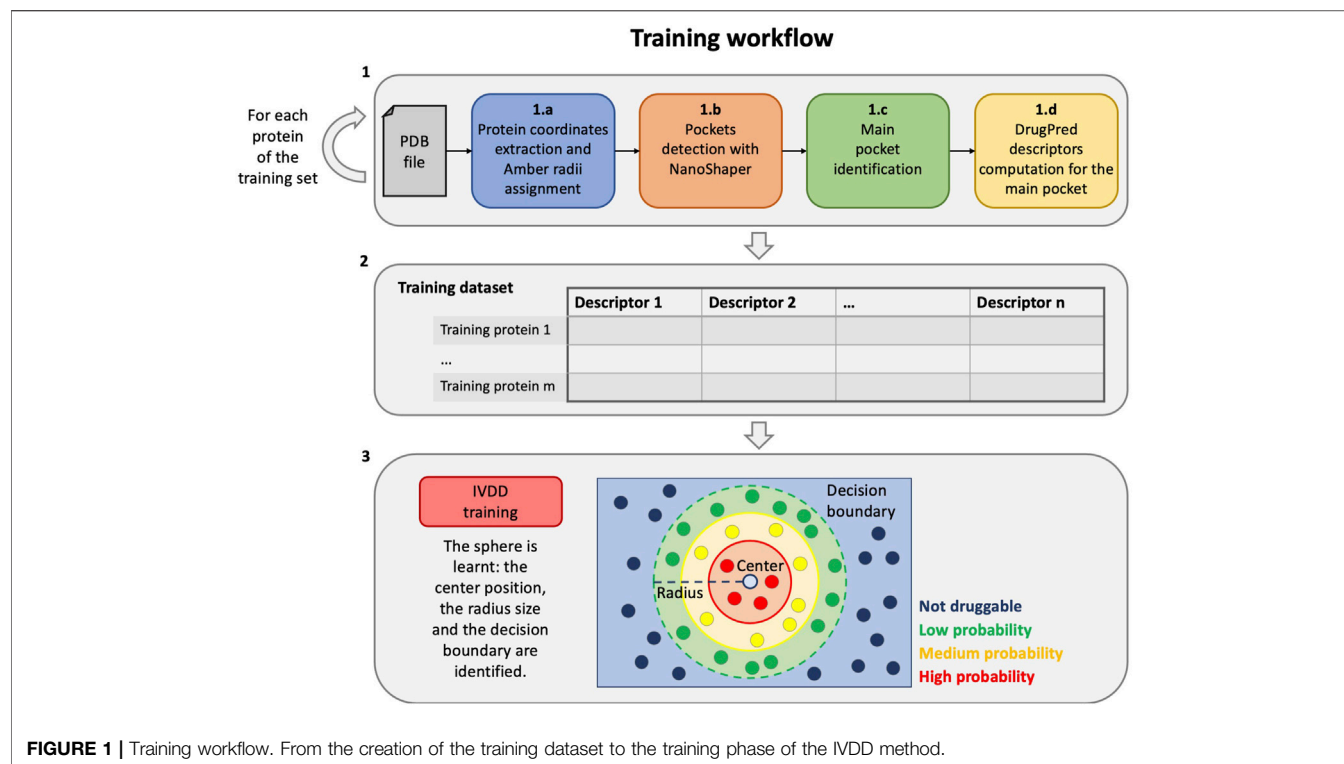
authors in Schmidtke and Barril (2010) published two datasets: a large but redundant dataset (DD, with 1,070 structures) and a non-redundant subset (70 binding sites).

Here, we address the problem of druggability estimation from the perspective of bias mitigation. The *a priori* dichotomy between druggable and less druggable (or non-druggable) pockets technically supports machine learning classifiers. Conceptually, however, it is questionable to use or define a non-druggable class. Indeed, apart from trivial cases (e.g., very small pockets), it is at best ambiguous to define such class. Defining a pocket as non-druggable (or less druggable) automatically creates a bias in the learned model, which may hamper the detection of a potentially useful pocket. Hence, we argue that druggability estimation should be approached as a one-class unsupervised learning task, not a classification one. This is because a classification task would inevitably create arbitrary user-dependent biases in the definition of the non-druggable (or less druggable) class. Starting from this observation, we devised a protocol that uses the import vector domain description method (a probabilistic one-class non-linear learner) to learn a hypersphere (a generalized minimum enclosing ball), which contains druggable pockets (Decherchi and Rocchia, 2016; Decherchi and Cavalli, 2020a). That is, only the definition of a druggable pocket is required during training, avoiding the creation of bias in the definition of the non-druggable class. To support the learner, we used a NanoShaper-based implementation of DrugPred (Krasowski et al., 2011) descriptors with minor modifications (the entrance area computed by NanoShaper is used as an additional descriptor). We employed the dataset in Krasowski et al. (2011) because it is widely used and explicitly defines a less druggable set of pockets. Furthermore, we defined a diversified new dataset of 100 protein targets to further validate the method. This dataset is a subset of the Potential Drug Target Database (PDTD (Gao et al., 2008)). Our results demonstrate the effectiveness of the approach. In the following, **Section 2** describes the method workflow, **Section 3** shows the results of the experiments, and **Section 4** introduces possible future developments and reports the final conclusions.

2 METHODS

In this section, we have described the proposed workflow for druggability prediction. For clarity, we have separated the training workflow from the testing (the operative phase) one. The training phase is a step that is required to estimate (learn) the model and comprises three main steps (see **Figure 1**):

- 1 First, we compute descriptors for the proteins of the training set, in particular, for each protein, as follows:
 - a) the protein part is filtered from the input PDB, and the radii of the Amber99SB-ildn force field are assigned to it;
 - b) the PDB file is thus converted to a .xyzr file and then passed to NanoShaper to detect all the pockets;
 - c) a main druggable pocket is identified (one for each training protein);



- d) the geometric and chemical descriptors of the pocket are computed.
- 2 All the information from the previous step is aggregated in order to form the training dataset, which is therefore composed by the descriptors of each main druggable pocket of the training targets.
- 3 Finally, the training dataset is used to train the import vector domain description (IVDD) machine learning method. In this phase, a sphere is learned and allows to assign a probability value to each pocket and consequently to distinguish druggable (probability ≥ 0.5) and non-druggable pockets (probability < 0.5).

On the other hand, the testing/operative protocol, that is, when the model is used for predictions only, comprises three main steps (see **Figure 2**):

- 1 First, we compute the descriptors for the current target protein, as follows:
 - a) the protein part is filtered from the input PDB and the radii of the Amber99SB-ildn force field are assigned to it;
 - b) the PDB file is thus converted to a .xyzr file and then passed to NanoShaper to detect all the pockets;
 - c) the geometric and chemical descriptors of the pockets are computed.
- 2 All the information from the previous step is aggregated obtaining a single file comprised of the descriptors of each pocket of the current target.
- 3 Finally, the previously estimated hypersphere is used to predict the probability of each of the newly detected pockets. The

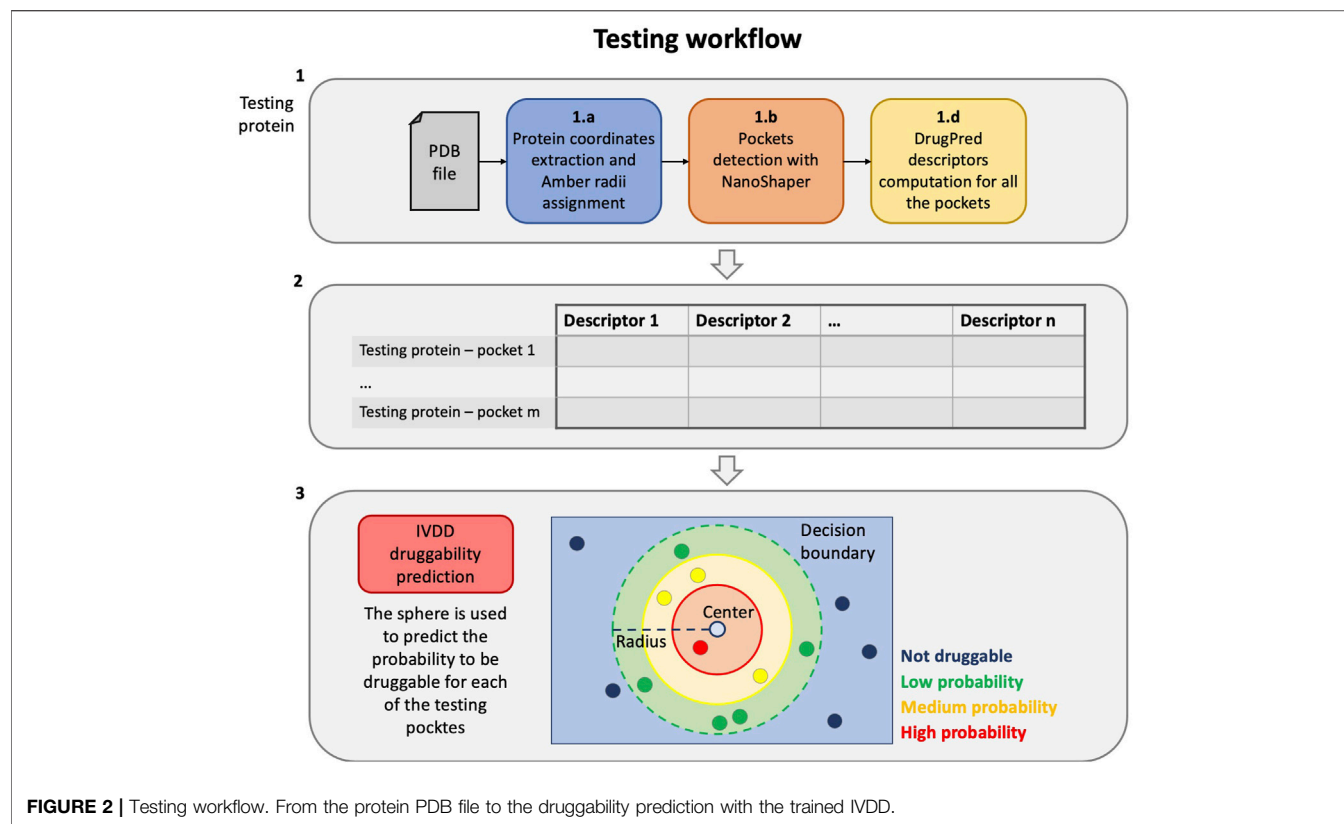
pockets with the highest probability are most likely to be druggable.

In the following sections we provide more details regarding the abovementioned steps. In particular, Section 2.1 describes steps 1b and 1c of the pipeline, Section 2.2 provides information regarding the descriptors building step (1d), and finally, Section 2.3 explains the IVDD method mentioned in step 3.

2.1 NanoShaper Pockets Detection and Main Pocket Identification

The detection of all the available pockets is instrumental for estimating the druggability of each pocket in the protein of interest. For this step, we used the NanoShaper tool (Decherchi and Rocchia, 2013; Decherchi et al., 2018) to efficiently deliver the set of pockets on a protein. NanoShaper was chosen as it accurately estimates the molecular surface (Wilson and Krasny (2021)); the detected pockets are triangulated with the same technique used for molecular surface triangulation, hence providing smooth triangulated meshes.

The detected pockets are saved as mesh files in MSMS or in the .off format, and they can be easily parsed to support the subsequent descriptors building step. NanoShaper also provides volume, surface area, and a list of the constituting atoms for all the internal cavities and pockets identified for the given molecular system. These are identified and computed *via* an intuitive approach, which involves a



volumetric difference of the regions of the space between system's solvent-excluded surfaces (SEs), with two probe radii, dubbed a large probe (with radius R) and a small one (with radius r) (Decherchi et al., 2018). The probe sizes encode the expectation onto the shape of the pockets. High R values allow the identification of shallow pockets, whereas high r values will smooth inner surface gaps. Default values are 3.0 Å and 1.4 Å for the large and small probes, respectively. The large radius is based on empirical evidence and the small radius mimics the water molecule. Here, we used the default value of the small radius but fine-tuned the large radius to a value of 3.5 Å. With respect to the default value of 3.0 Å, we found that this value allows a better detection of slightly more shallow pockets (a larger surface size of pocket entrance).

To create the training dataset, we needed an automated method to detect the orthosteric/main pocket, where the ligand is located, and discriminate it from the others (NanoShaper delivers several pockets). Because the orthosteric pocket is well-identified in the analyzed PDB, we used the surrounding atoms of the ligand. In detail, we used the Jaccard index on the atom indices to easily detect the orthosteric pocket; the Jaccard index of atoms is an accurate proxy of the discretized volume overlap, often found in druggability predictors. We defined the orthosteric pocket as the pocket detected by NanoShaper with the maximal Jaccard index with respect to the reference indices. This is easily achieved by localizing the atom indices around target's

natural substrate (or drug). The Jaccard index is defined as follows:

$$J(O, P_i) = \frac{|O \cap P_i|}{|O \cup P_i|}, \quad (1)$$

where O is the indices set for the orthosteric site and P_i is the set of detected atom indices in the i th pocket. The Jaccard index is hence a natural measure of the quality of the detected pocket with respect to ligand's envelope. One can note that the Jaccard index can be decomposed into two components, which account for the degree of overimposition of the pocket and reference ligand volume in two different ways. The first component is the normalized intersection component J_{int} :

$$J_{int}(O, P_i) = \frac{|O \cap P_i|}{|O|}, \quad (2)$$

and the second one is the normalized union component J_{or} :

$$J_{or}(O, P_i) = \frac{|O|}{|O \cup P_i|}. \quad (3)$$

They both belong to the interval (0,1). They account, respectively, for the ability to detect all the reference atoms (J_{int}) and the tightness of detection (J_{or}). Both properties are desirable and consistently lead to the Jaccard index upon multiplication. To fairly evaluate the results, we considered these metrics together with classification accuracy.

TABLE 1 | Descriptors of the datasets. The incidence is calculated for every amino acid X.

| Descriptor | Abbr |
|---|--------|
| Binding site volume | vol |
| Total surface area | area_b |
| Entrance area | area_e |
| Binding site compactness | cness |
| Relative hydrogen-bond donor surface area | dsa_r |
| Hydrogen-bond donor surface area | dsa_t |
| Relative hydrogen-bond acceptor surface area | asa_r |
| Hydrogen-bond acceptor surface area | asa_t |
| Relative hydrophobic surface area | hsa_r |
| Hydrophobic surface area | hsa_t |
| Relative occurrence of polar amino acids | paa |
| Relative occurrence of non-polar amino acids | haa |
| Relative occurrence of multifunctional amino acids | maa |
| Relative occurrence of charged amino acids | caa |
| Relative polar surface area (dsa_r + asa_r) | psa_r |
| Incidence of amino acid X in the binding site relative to the surface | in_X |

2.2 Descriptors Building

To characterize each pocket identified by NanoShaper, we used the descriptors defined by Krasowski et al. (2011) together with the entrance area provided by NanoShaper (Table 1).

Binding site properties describing size, shape, polarity, and amino acid composition were calculated using NanoShaper output files as input to the descriptors builder. In particular, to compute volume (vol), total surface area (area_b), and entrance area (area_e) (which describes the area of the pocket mouth), we directly used the estimations provided by NanoShaper. To calculate the other descriptors, we started from the NanoShaper output files describing the atoms and meshes of each pocket. The hydrogen-bond donor and acceptor properties of each pocket were calculated by considering the surface area surrounding all the polar atoms (dsa_t and asa_t). Based on these descriptors, the hydrophobic surface area (hsa_t) is defined as the difference between the total surface area and the sum of the hydrogen-bond donor and acceptor surface areas. Moreover, relative amplitude of the hydrogen-bond donor and acceptor surface areas (dsa_r and asa_r) and the hydrophobic surface area (hsa_r) were computed by dividing each descriptor by the total surface area of the binding site. Finally, the relative polar surface area (psa_r) is defined as the sum between the relative hydrogen-bond donor and acceptor surface areas. To characterize the shape of different cavities, we used the compactness descriptor, defined by Krasowski et al. (2011):

$$cness = \frac{4\pi \left(\sqrt[3]{\frac{vol}{\pi}} \right)^2}{area_b} \quad (4)$$

According to this equation, the closer the compactness is to 1, the more spherical is the pocket. The remaining descriptors, relating to amino acid composition, were calculated by considering the occurrence of different classes of amino acids grouped by their overall physicochemical properties. In particular, all the amino acids were grouped into the following classes:

- Apolar: Ala, Gly, Val, Ile, Leu, Met, Phe, and Pro.
- Polar: Thr, Lys, Arg, Glu, Asp, Gln, Asn, and Ser.
- Charged: Lys, Arg, His, Asp, and Glu.
- Multifunctional: Trp, Tyr, His, and Cys.

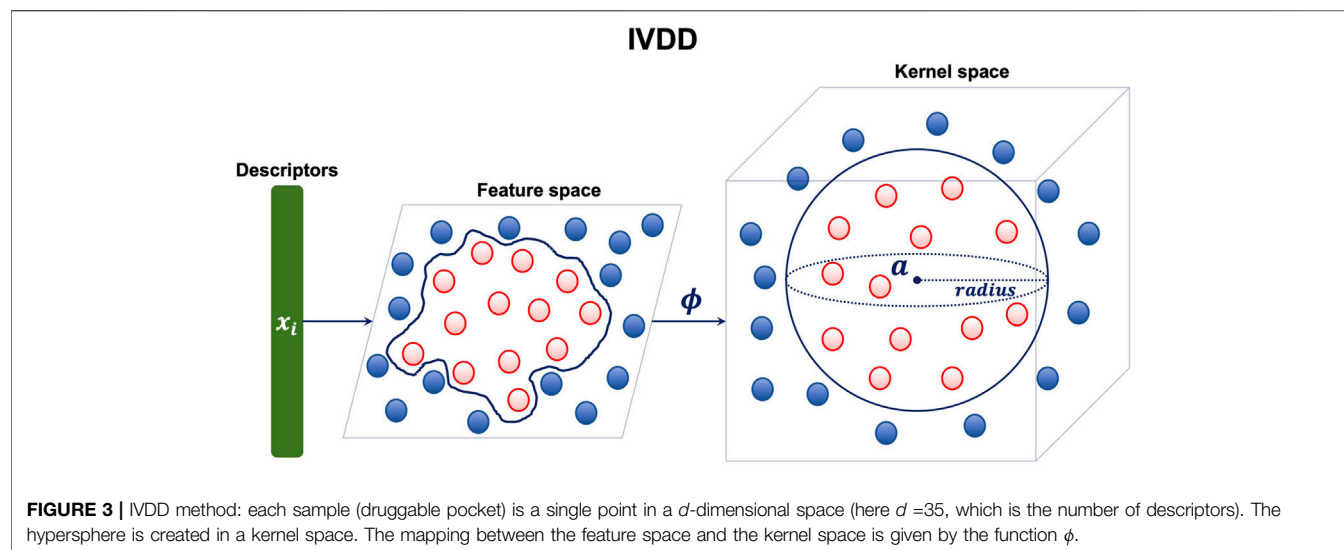
To define the relative occurrence of hydrophobic amino acids (haa), polar amino acids (paa), charged amino acids (caa), and multifunctional amino acids (maa), we computed the fraction of each group of amino acids with respect to the total number of amino acids comprising each cavity. Finally, we reported the incidence of each amino acid of type (in_X) as descriptors, defined as the sum of all the surface areas surrounding the amino acid X.

2.3 Druggability Estimation via IVDD One-Class Learning

As anticipated, we used a one-class approach, that is, we require and consider for the training phase only the samples in the class from which we want to learn the concept. The aim is to learn the concept of a *druggable* pocket. This requires only samples (pockets) that are known to be druggable. To perform this step, we used the one-class learner dubbed import vector domain description (Decherchi and Rocchia (2016)). The import vector domain description method tries to embed the available training samples into an enclosing hypersphere. This sphere does not belong to the original input space but rather resides in a, possibly infinite dimensional, kernel space. This approach allows us to wrap the data in arbitrarily complex enclosing surfaces because the hypersphere in kernel space corresponds to a not necessarily spherical enclosing surface in the original space (see Figure 3).

This makes the method very flexible. Moreover, the enclosing surface is endowed with a probabilistic model, which assigns the probability of belonging (or not) to the enclosing sphere.

The aim of the training procedure of IVDD is to find a sphere configuration (center position and radius size) that best minimizes the cost function (see later). The cost function tries to maintain as much as possible the samples inside the sphere while at the same time keeping under control the radius size, possibly letting some training samples outside the sphere. One is eventually searching for a compact representation of the space spanned by the samples. We will call $[\pi_{low}, \pi_{high}]$ the range of acceptance of the fraction of training examples inside the sphere. It can be shown that the optimal sphere (the solution of the minimization problem) is unique, as the problem is convex. Once the final sphere configuration is found it determines predictions during the operative phase. The non-druggable nature of a pocket is just an interpretation over the probability values; strictly speaking, one-class learning just describes the adherence of a sample (a pocket) to a concept (druggability). If a crisp classification is needed, the probability threshold of 0.5 can be used. Samples outside the sphere (decision boundary) are predicted as non-druggable (with a corresponding probability lower than 0.5), while samples inside the sphere are predicted as druggable (with a corresponding probability higher than 0.5). Clearly, the inner and most central pockets are estimated to have



the highest probabilities of being druggable. Indeed, this probability is high at the core of the sphere and decreases toward the edges.

At a mathematical level, the training phase of the IVDD method is characterized by the following minimization problem:

$$\min_{\Gamma, \mathbf{a}} \Gamma^2 - \hat{C} \sum_{i=1}^n \log(p_i), \quad (5)$$

where Γ is the square of the radius of the hypersphere, constant $\hat{C} = C/n$ represents the trade-off between the radius size and the error minimization, and p_i is the probability defined by a logistic model:

$$p_i = \frac{1}{1 + \exp(\beta f_i)}, \quad (6)$$

where β is a fixed coefficient and f_i is the decision function defined as follows:

$$f_i = d^2(\Phi(\mathbf{x}_i), \mathbf{a}) - \Gamma, \quad (7)$$

where $d^2(\Phi(\mathbf{x}_i), \mathbf{a})$ is the distance function and \mathbf{a} is the center of the hypersphere. The cost function in Eq. 5 is optimized via an efficient learning algorithm that can be ascribed to a class of sequential minimal optimization (SMO) algorithms (Zeng et al., 2008). The introduced probability model is used to probe the druggability of each pocket. We refer the reader to Decherchi and Rocchia (2016) for further details.

3 RESULTS

3.1 Datasets

In this work, we used two different datasets to run the experiments. In both cases, we generated two versions of the dataset: with and without hydrogen atoms. The first dataset is the NRDL dataset, presented in Krasowski et al. (2011). It is the largest publicly accessible non-redundant dataset for model

building and validation of structure-based druggability methods. The dataset comprises 115 structures (protein-binding sites), including 71 druggable and 44 less druggable (which becomes 42 after the analysis in Krasowski et al. (2011)). For each binding site, 35 different descriptors are calculated, as described in section 2.2 and summarized in Table 1.

In addition to the NRDL dataset, we created another dataset comprising the binding sites of 100 different proteins. Those targets are taken from the PDTD (Potential Drug Target Database) (Gao et al., 2008), a free online collection of 1,100 3D structures of proteins. The targets in our 100-protein dataset include enzymes, receptors, antibodies, signaling proteins, and lipid-binding proteins. We thus obtained 5,692 and 4,807 binding sites without and with hydrogen atoms, respectively. Of these, 100 are orthosteric (one for each target). For each structure, we selected the pocket that hosts the drug or substrate. We avoided selecting pockets that host cofactors. We defined these pockets as orthosteric (or main) throughout the text (because the drug is co-crystallized in the orthosteric site in most cases). As for the NRDL dataset, we calculated previously defined descriptors for each binding site (see Table 1).

For more information on the targets of the NRDL and the PDTD datasets, see Supplementary Material Sections S1, S2.

3.2 Model Training

We trained IVDD considering the descriptors of $n = 70$ druggable structures in the NRDL dataset. The *Invj* structure was excluded since it represents a small oligonucleotide and we only considered proteins to calculate the descriptors. The following parameters were adopted: kernel used is RBF with $\sigma = \max_{ij} (d_{ij})/\log(n)$ (where d_{ij} is the distance between the i -th and the j -th sample); value of C is initialized as 0.5, the value of β is set as 25, while the range of accepted inner samples is set to $[\pi_{low}, \pi_{high}] = [0.8, 0.9]$. The values of $[\pi_{low}, \pi_{high}]$ may vary according to the reliability of the training dataset. In this case, we preferred a conservative approach, with 80–90% of samples included inside the sphere and the remaining peripheral

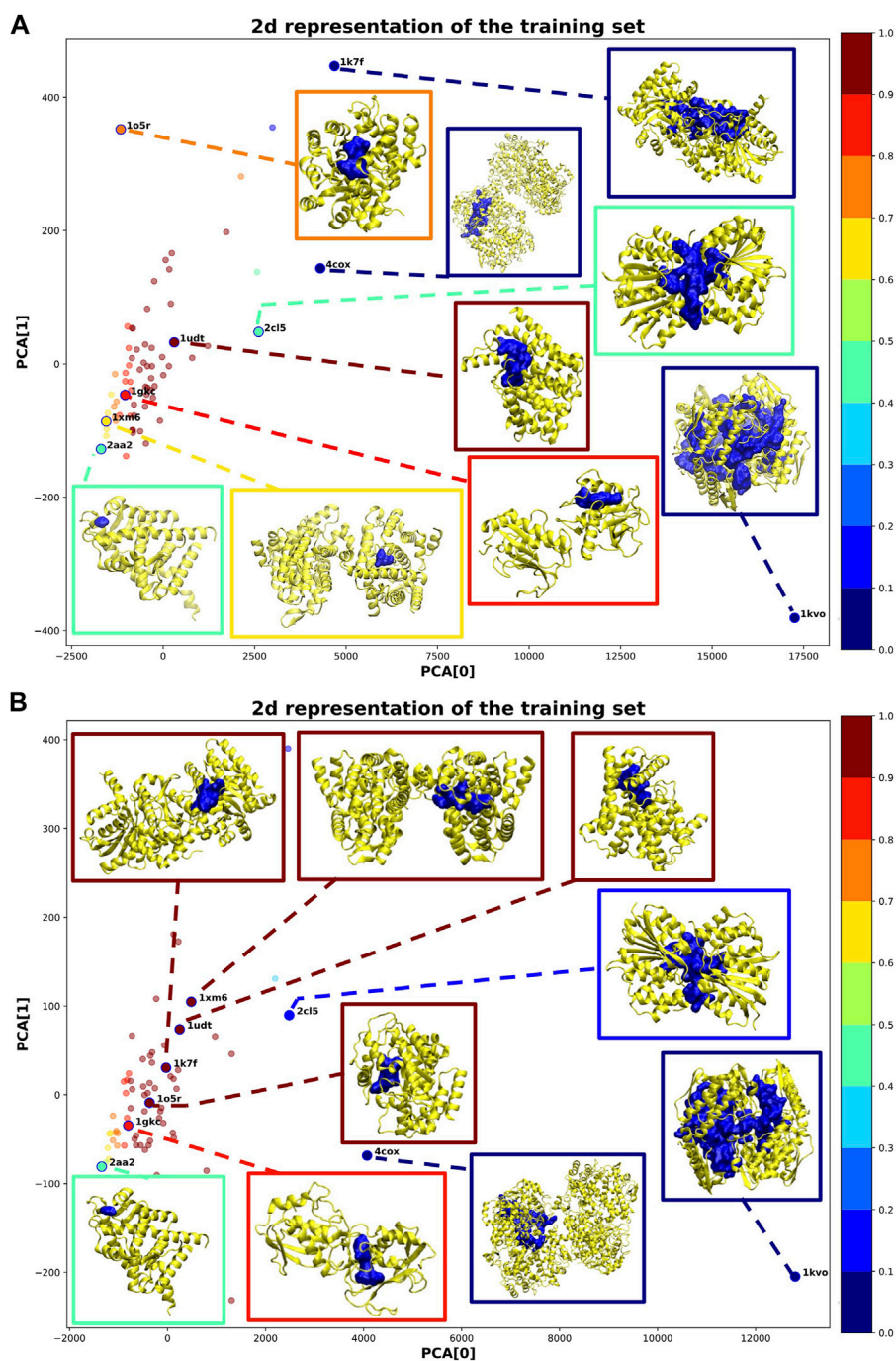
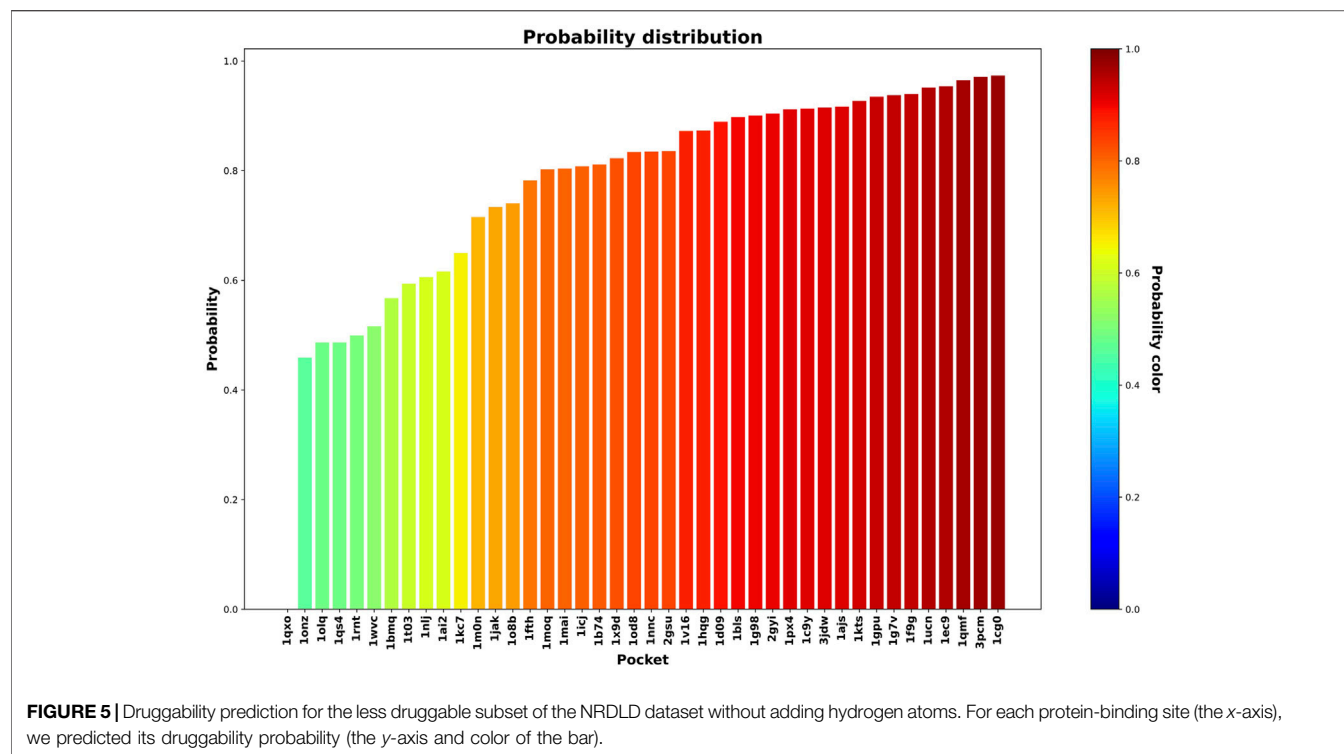


FIGURE 4 | 2D representation of the training samples *via* PCA dimensionality reduction. Each point corresponds to a training sample (protein-binding site). The color of each point corresponds to the probability assigned by IVDD (graded according to the color map on the right). For some training samples, the corresponding 3D structure is shown. **(A)** is without hydrogen atoms, whereas in **(B)** hydrogen atoms were added.

20–10% as outliers, in order to avoid overfitting. The learning phase is stopped when the range of inner samples is hit. Each time the training is repeated, the C is increased/reduced by 0.01 (increased if the percentage of samples inside the sphere is lower than the desired range, reduced otherwise). In our case, the training procedure ended with 90% samples inside the sphere

and a final C value of 0.1 for the solution without hydrogen atoms and with 90% of samples inside the sphere and a final C value of 0.12 for the solution with hydrogen atoms.

Figure 4 shows a 2D representation of the training set obtained by reducing the dimensionality *via* a principal component analysis (PCA) (Jolliffe, 1986). For some samples,



we additionally plotted the corresponding 3D structure. In both cases, most of the training samples coherently obtained a high probability of druggability (dark red points in **Figure 4**). This outcome is obtained because we imposed the solution to include at least 80% of the training samples inside the sphere.

Considering the solution without hydrogen atoms (see **Figure 4A**), the sample *1udt* has the highest probability and is the sample nearest to the center of the sphere. In this structure, the pocket identified by NanoShaper is very compact and well-defined. IVDD performs the best in cases where the pocket closely surrounds the ligand bound in it. The samples outside the sphere (corresponding to 10% of the samples) obtained low probability scores. These scores are explainable by looking at the pocket shape. Structures such as *1kvo*, *4cox*, and *1k7f* do not look like well-defined pockets but rather like a fusion of more than one pocket. This leads to descriptors that are quite distant from those that the algorithm is learning as the druggable reference. As a consequence, those structures are scored as outliers. This highlights that *ex post* segmentation can be a powerful preprocessing tool before the machine learning step. Nevertheless, IVDD can cope with this situation by excluding or marginalizing percolating pockets. It is possible to identify another case where NanoShaper did not correctly identify the orthosteric pocket (i.e., *2aa2*). Here, the pocket is very shallow and the bound ligand is not deeply buried. The identified pocket is much smaller than it should be, leading to a low probability. This effect is expected because NanoShaper can only detect shallow pockets *via* a proper tuning of the big probe, whereas the selected value is expected to work mainly for deep buried prototypical pockets.

The solution with hydrogen atoms (see **Figure 4B**) identifies the sample *1xm6* as having the highest probability. In contrast to the solution without hydrogen atoms, its structure is now more compact around the ligand with a greater J_{int} . Since the presence of hydrogen atoms better defined the orthosteric pocket, NanoShaper improved its accuracy, leading to a high IVDD probability. This happened similarly for *1k7f*, where the channel that led to a big pocket was closed by the presence of hydrogen atoms. In this specific case, NanoShaper identified the orthosteric pocket with a Jaccard index three times better than the solution without hydrogen atoms. Although the solution with hydrogen atoms solved some NanoShaper errors (wide percolation), pockets such as *1kvo*, *4cox*, and *2aa2* remained more or less unchanged, with very big or shallow structures. The option to use hydrogen atoms (or not) is partially data-dependent and is further studied in NRDL and new datasets.

3.3 Experiment on the NRDL Dataset

In this step, we used the 42 less druggable structures described in Krasowski et al. (2011) in order to test the previously trained model and perform druggability prediction. **Figures 5** and **6** show the probability assigned to each structure by the IVDD method for the solutions without and with hydrogen atoms, respectively.

Generally speaking, the following results are relatively similar. The resulting trend shows that IVDD predicts a probability greater than 0.8 for around half of the less druggable set. This points to a possible bias in the “less druggable” set. Indeed, a purely unsupervised approach such as this one, in which no *a priori* dichotomy is created, shows that several pockets are not judged to be less druggable. On the contrary, more than half are

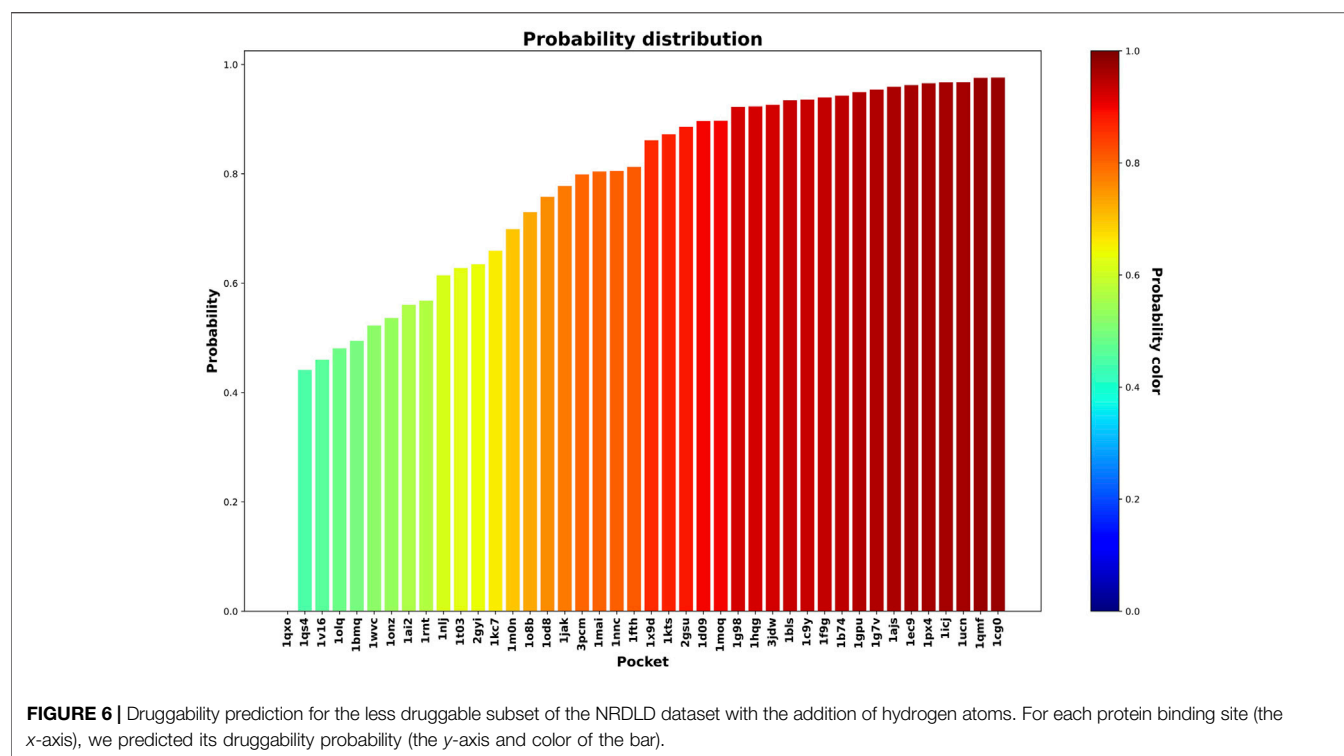


FIGURE 6 | Druggability prediction for the less druggable subset of the NRDL dataset with the addition of hydrogen atoms. For each protein binding site (the x-axis), we predicted its druggability probability (the y-axis and color of the bar).

scored with high probability values. The less druggable nature can be ascribed partially to the shallow nature of this set; however, thanks to the large probe set to 3.5 Å, NanoShaper can still detect them.

This result hence partially contrasts with the *less druggable* labeling of this dataset. One should consider the principles behind this previous classification. Krasowski et al. (2011) postulated that a protein (not just the pocket) can be ascribed to the less druggable realm if none of the following conditions are met: 1) at least one ligand is orally available as judged by the Lipinski's rule of five and 2) the ligands must have a $\text{clogP} \geq -2$. In addition, the ligand efficiency of at least one of the ligands fulfilling criteria 1) and 2) must be $\geq 0.3 \text{ kcal mol}^{-1}$ per heavy atom. To correctly fulfill the requirements one should be able to test all the chemical space before making any conclusion. Indeed, ideally, and more correctly, one could define the *true druggability* of a pocket as the value of the activity of the best possible ligand for that pocket in the chemical space. As the sampling of the chemical space is limited and further biases are due to the drug discovery community interest and efforts for a specific protein, this classification is questionable and not necessarily reliable. The problem of druggability classification of a pocket, or a protein, that is ligand-dependent is that it would require the true sampling of the chemical space. In our proposal, instead, we do not define *a priori* the labels but concentrate on the only reliable information that is, druggable pockets. The final result of this is that some pockets previously labeled as less druggable instead obtain high druggability probability values.

It is interesting to analyze the probability shift from lower to upper values, systematically. **Figure 7** shows the orthosteric

pockets found by NanoShaper for the less druggable proteins, where we subsampled the structures set with a ratio of one every five complexes. The pockets here tend to become deeper and more compact moving from lesser probability to higher. The shift is particularly evident comparing *1onz* and *1cg0*, where the first case is a very shallow pocket, in which a ligand can be found, but it is neither a prototypical nor ideal pocket; its probability value is 0.46. In contrast, *1cg0* shows a much better defined and large enough pocket that would host a potential ligand well; IVDD classifies it as druggable with a probability value of 0.97. Except for *1qxo* (a pocket detected by NanoShaper that is too large), one can observe that the lower the score, the smaller and more shallow the pocket is. This is also evident looking at the portion of solvent-exposed surface of the ligands, where the low probability pockets tend to have more solvent-floating ligands.

There are some particularly interesting cases in this less druggable set, also considering the ligands found in the crystal structures. In *1kts*, *1gpu*, *1ucn*, and *1cg0* the ligands are small molecules or small molecule-like ligands. Missing these pockets would be quite negative in a drug discovery campaign. All these pockets score quite high with our method. One should not restrict to the pure small molecule paradigm; in the case where one is concerned with the design of a molecular glue or a PROTAC, even a warhead relatively not too active can be sufficient to degrade the protein. Our method is agnostic to ligand-induced labeling and avoids to miss or undervalue this kind of pockets.

At a technical level, it is interesting to compare the pocket probabilities with and without hydrogen addition and to consider the NanoShaper's behaviour. As anticipated, adding or not adding hydrogen atoms does not change the detection of the

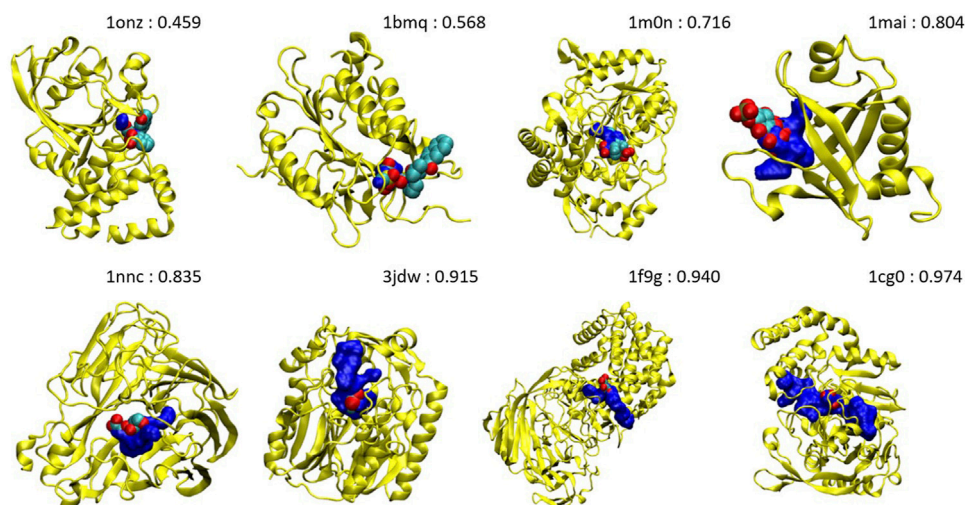


FIGURE 7 | Main pockets (computed without hydrogen atoms) of 1onz, 1bmq, 1m0n, 1mai, 1nnc, 3jdw, 1f9g, and 1cg0. The pocket surface is in blue and the complexed ligand in the pdb file is in the VdW style. The number is the estimated druggability probability value.

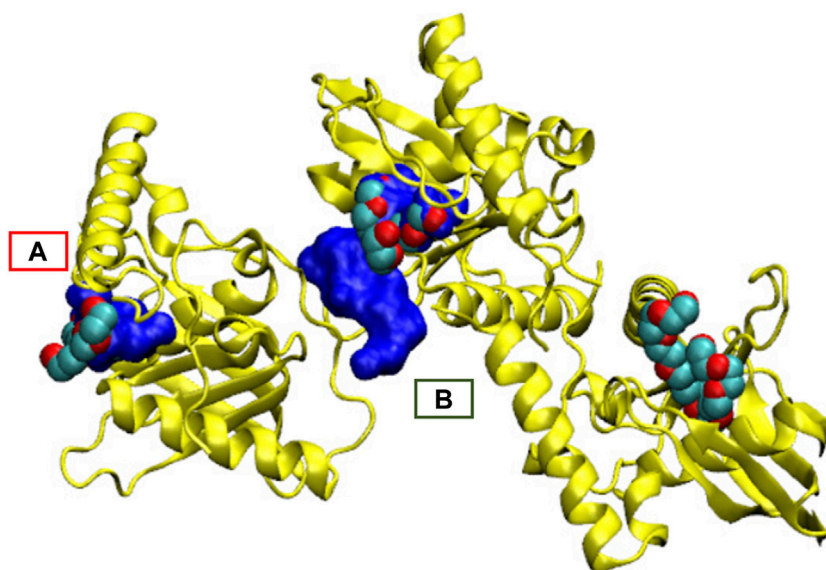
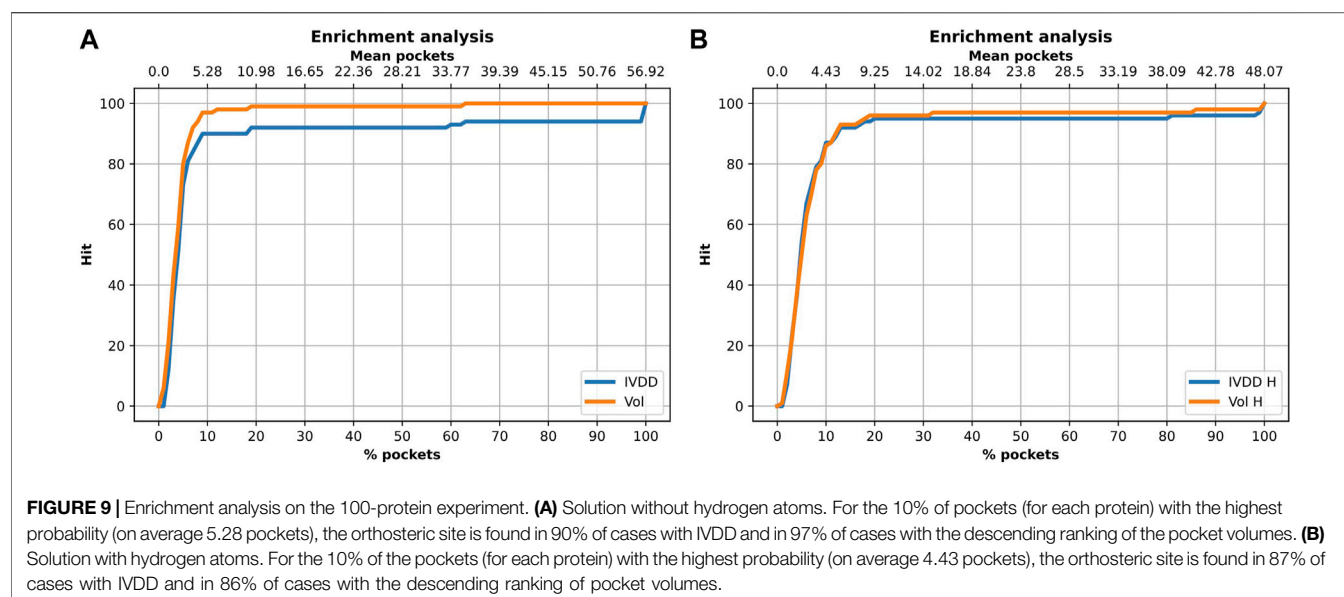


FIGURE 8 | Main pocket shift for 1icj together with the co-crystallized ligand. **(A)** Main pocket detected when adding hydrogen atoms. **(B)** Main pocket without adding hydrogen atoms. The pocket is semantically the same orthosteric pocket but changes from one monomer to another. The three structures of ligands bound in the PDB structure are also reported.

main pocket by NanoShaper (highest Jaccard index). However, the shape and relative probability ranking both change. A first observation is that, in some peculiar cases, the percolating behavior of NanoShaper pockets cannot be solved by adding hydrogen atoms. Indeed, 1qxo is still ranked last and, coherently, this pocket is percolating widely inside protein crevices. This global invariance is confirmed by analyzing 1icj (see **Figure 8**). In this case, the detection of the main pocket is geometrically, but not semantically, changed when the structure with and without hydrogens atoms are considered. That is, the main detected

pocket is the same but is in another monomer of the homotrimeric unit. Despite this finding, its druggability probability changes when adding hydrogen atoms. This demonstrates that the same pocket in two different conformations (monomers) is well-detected and always ranked as druggable. Indeed, without hydrogen atoms we can identify the orthosteric pocket in monomer A. Upon addition of hydrogen atoms, we instead identified the orthosteric pocket in monomer B. In this last case, the Jaccard index is higher with improved pocket quality (the pocket is more compact and located at the interface).



However, the probability value changes as the corresponding geometry (and presence or absence of hydrogen atoms) changes, leading to a way higher value for pocket B. Therefore, from one side, what is judged druggable remains druggable. However, inside the druggable set, conformational changes of the same pocket have a non-trivial role in shifting the probability value. This confirms that it is crucial to consider dynamical aspects, particularly the probability of a given site conformation (and hence its free energy), in order to obtain a complete picture of the overall druggability of a site, which may be dealt with as a physical observable value.

Overall, this analysis shows that the dataset definition can create non-trivial biases, including biases due to labeling and the presence or absence of hydrogen atoms, which can induce local changes. One-class learning can mitigate the first bias because it only uses the druggable class during training. In the next section, we discuss other possible sources of bias and further evaluate the accuracy on a wider and curated dataset, also considering the initial processing of the structures (hydrogen addition).

3.4 PDTD Subset Validation

In this analysis, we used the 100-protein dataset, which is our curated subset of the PDTD. Here, we again evaluated the accuracy of classification and also searched for other possible sources of biases. It is well-known that the volume value has a crucial role in determining the druggability of a site. Among others, in Naya and Honig (2006), the authors used SCREEN (Surface Cavity Recognition and Evaluation) to locate and analyze pockets in the NRDL dataset. They observed that just picking the pocket with the highest volume value had a success rate of 64%. However, just looking at the volume value may create further biases, some intrinsic, some operational, and some technical. An overly large volume could be erroneously ascribed to the main site just because a small fraction contains the true binding site. This can happen in dependence of the pocket detection engine (e.g., for the percolation effect). Fortunately, this can be evaluated well *via* overlapping volume metrics or by the

Jaccard index. Here, we performed this analysis by considering this issue. We compared our performance with that obtained by considering a simple descending ranking of the pocket volumes. **Figure 9** and **Table 2** show the results for the situations with and without hydrogen atoms. Using a simple ranking of the volume, we obtained a better performance at top 5, with an accuracy of 97%. This decreased to 89% when hydrogen atoms were added. In contrast, IVDD identified 90% of the orthosteric pockets in the top 5 highest probability pockets, which increased to 92% when hydrogen atoms were added. This shows that IVDD is more stable, although lower in accuracy in absolute terms.

It is important to consider the quality of the pockets identified in both cases. The presence of hydrogen atoms sometimes allows the fragmentation of some of the overly large pockets. This not only increases the accuracy in terms of the main pocket druggability estimation but also affects the overall shape, which often becomes too tight. This is a NanoShaper-dependent effect, which is documented in **Figures 10** and **11**. In **Figure 11**, we reported the cumulative scores, namely J , J_{int} , J_{or} for the volume and the IVDD ranking for the top 1 pockets, ordered respectively by volume and by probability. The trend shows a systematically higher value for all three scores for IVDD without hydrogen atoms and almost indistinguishable scores with hydrogen atoms. Interestingly, without hydrogen atoms, IVDD has a lower accuracy than that in the simple volume. This is

TABLE 2 | Results obtained on the PDTD subset (with and without hydrogen atoms) with the IVDD method and by a simple descending ranking of the pocket volumes. All results are referred to the orthosteric/main sites.

| Description | IVDD | Volume | IVDD + H | Volume + H |
|-------------|------|--------|----------|------------|
| Top 1 | 50 | 60 | 50 | 50 |
| Top 2 | 67 | 76 | 69 | 65 |
| Top 3 | 81 | 87 | 81 | 79 |
| Top 5 | 89 | 97 | 92 | 89 |
| Top 10% | 90 | 97 | 87 | 86 |

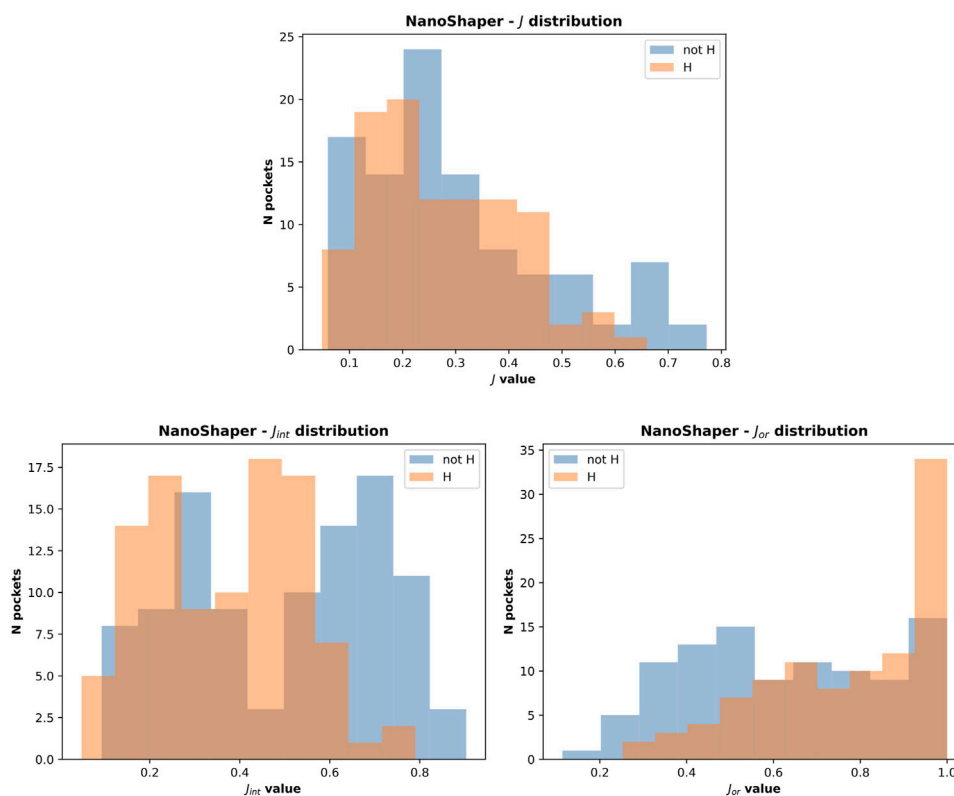


FIGURE 10 | NanoShaper score distribution with and without hydrogen atoms.

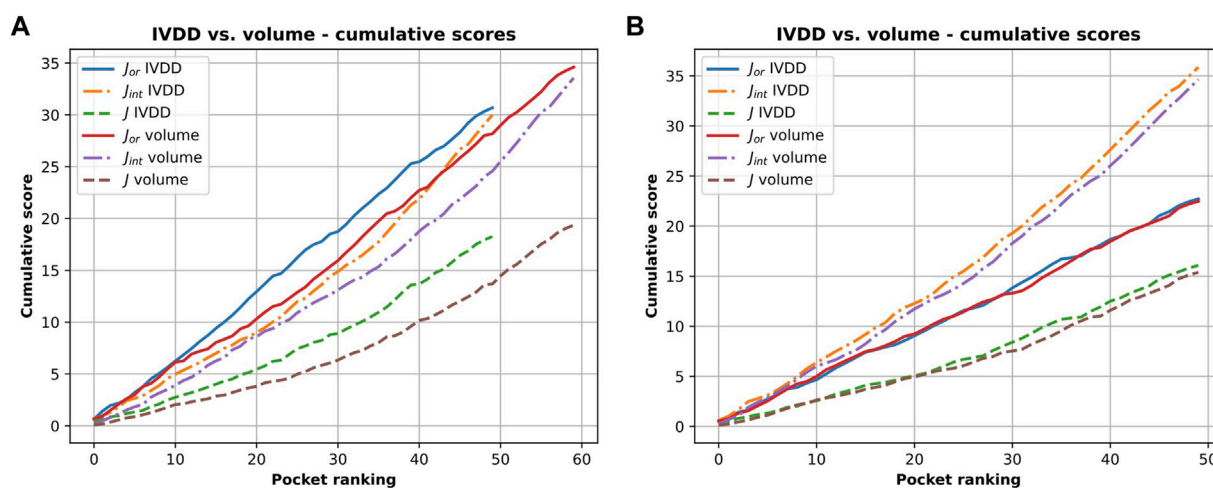


FIGURE 11 | Cumulative scores (J , J_{or} and J_{int}) for IVDD and volume ranking. Here, the orthosteric sites identified by IVDD and the volume ranking in top 1 are considered and ranked according to the probability score and the volume, respectively. Both the rankings are in descending order. Inset (A) is results without hydrogen atoms and inset (B) is with hydrogen atoms.

unsurprising since an overly percolating volume allows easier main pocket detection. However, when quality is considered, even if some pockets are lost with IVDD, the remaining pockets have significantly higher scores. Again, we can mitigate a bias by

not overfitting the volume-induced ranking. In the paradoxical case where one has a volume percolating throughout the protein, one would get a completely useless top 1 with 100% accuracy by using a pure volume ranking.

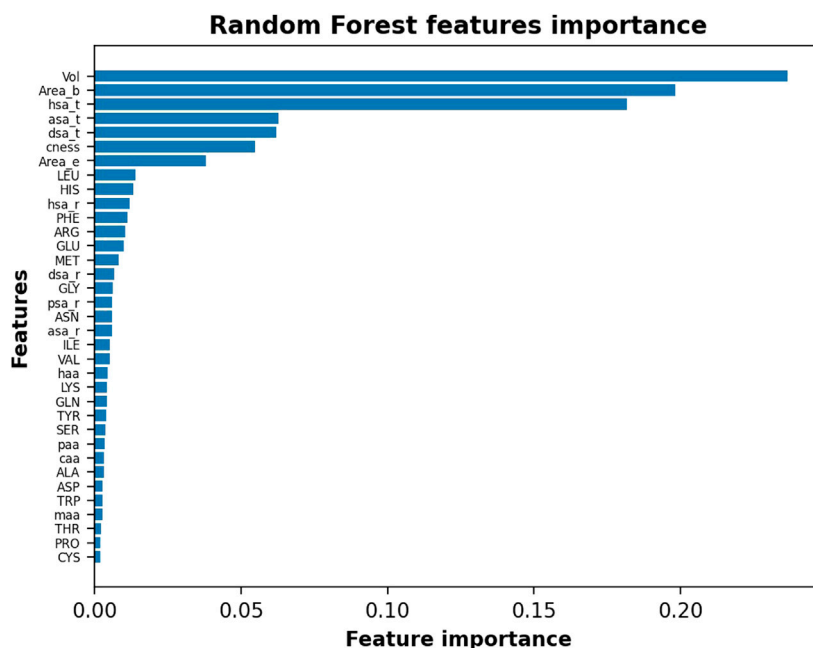


FIGURE 12 | Random Forest features importance in the descending order by assigning *ex post* labels to the IVDD predictions. Results shown are without hydrogen atoms; similar results are obtained with hydrogen atoms.

Within the IVDD results, it is also relevant to compare what happens with and without hydrogen atoms. Examining the structures that did not land in the top 5 positions with and without hydrogen atoms, one can conclude that most (e.g., *lvkg*, *lqpb*, and *lht8*) are large pockets with low or intermediate Jaccard index or with very low J_{or} value. In some cases, there are shallow pockets (e.g., *lgp6* and *li7g*) characterized by very high values of J_{or} . Some of those structures improve in the presence of hydrogen atoms, reducing the number of targets that fall outside the top five from 11 to 8. Some shared structures (e.g., *lht8*, *lh9u*, and *lv8b*) do not change the shape of the orthosteric pocket, leading to not significant changes in the probability.

We can compare the proposed solution to the many others in the literature. We have shown that by avoiding some of the possible biases (chiefly the labels) and considering the model without hydrogen atoms, we can obtain 81% detection accuracy in top 3 and 89% in top 5. We have also shown that a non-negligible fraction of the missed detections in top 5 can be ascribed to NanoShaper's behavior. In comparison, Volkamer et al. (2012a) obtained 88% accuracy in correctly assigning to the druggable or non-druggable class in the NRDLD dataset with DoGSiteScorer, where the support vector machine is used as machine learning backend. In contrast, DrugPred (Krasowski et al., 2011) obtained 91% accuracy for NRDLD. A widely used method is fpocket from Le Guilloux et al. (2009), which correctly identified 83% of ligandable pockets in top 3 of all analyzed proteins. Overall, we achieved an accuracy that is similar to that of several existing methods but with some *ab initio* safeguards such as avoiding biases due to labels and volume.

To further investigate the IVDD results, we identified how much each single feature affects the IVDD prediction. IVDD does not embed a feature selection method, so we used an *ex post* labeling strategy. We first estimated the probability obtained, on average, for each orthosteric site in the dataset, obtaining 0.852 and 0.877, respectively, without and with hydrogen atoms. These values represent two thresholds and allow a labeling for each binding site, which is 0 when its probability is lower than the threshold value, otherwise 1. This *ex post* labeling allows us to fit a classifier (here, we chose a random forest classifier (Breiman, 2001) with 100 estimators and the Gini index as criteria for the split) and to estimate the features importance. **Figure 12** shows the results of this additional experiment. Volume (Vol) is a major impacting feature, followed by area of the pocket surface (Area_b), hydrophobic surface area, hydrogen-bond acceptor surface area (asa_t), hydrogen-bond donor surface area (dsa_t), binding site compactness (cness), and entrance (mouth) surface area (Area_e). Similar results can be obtained with different classifiers and can be found in **Supplementary Material Section S3**.

To further check these results, we ran this experiment by normalizing data. We found that hydrophobic surface, polar surface, and volume still dominate the model. This means that IVDD is influenced by the volume, but it also considers other chemical aspects in predicting probability. Of less relevance is the fact that hydrophobic residues (LEU, PHE, MET, and GLY) and some charged residues (HIS, GLU) rank slightly higher. The presence of hydrophobic residues and volume as key factors is largely consistent with chemical intuition.

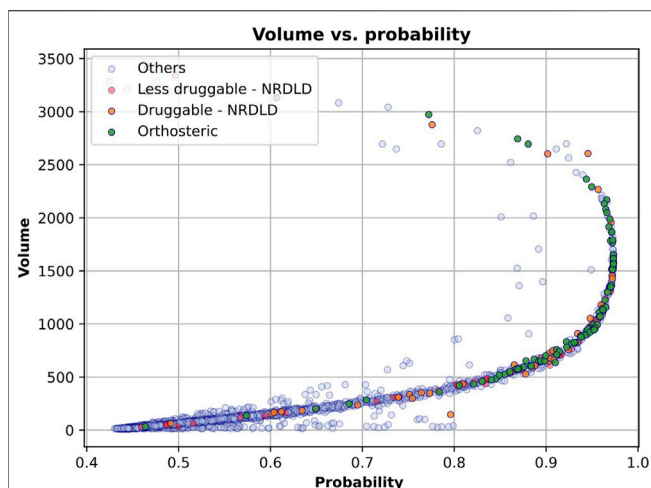


FIGURE 13 | IVDD probability scores vs. volume. Each sample represents a pocket (colored according to the corresponding dataset). The x-axis represents the probability that a pocket is druggable, while the y-axis represents the volume of each pocket. The plot is referred to the solution with hydrogen atoms. Similar results are obtained without hydrogen atoms.

The correlation between IVDD prediction and volume can be seen in **Figure 13**, in which we have plotted each binding site as a point in the 2D space, where the coordinates are the probability predicted by IVDD and the volume itself. In the presence (see **Figure 13**) and absence (data not shown) of hydrogen atoms, the samples with the highest probability have a volume between 500 and 2,000 Å³. The orthosteric sites and the training samples are condensed on the right side of the figure, meaning that they obtained high probability scores in most cases. Non-orthosteric binding sites are condensed in the bottom left of the figure since they are mostly small pockets and obtain low probability scores. However, both figures contain some non-orthosteric pockets with a volume between 1,000 and 2,000 Å³ and lower probability scores. In such cases, the IVDD decision has been influenced by factors other than volume.

4 DISCUSSION AND CONCLUSIONS

In this study, we presented an unsupervised one-class approach to build a druggability estimation model. We defined a pipeline to obtain all the pockets of a protein (NanoShaper), their corresponding descriptors, and druggability prediction. The method achieved 89% accuracy in top 5, in line with other methods. Although the method was less accurate than a trivial volume-based ranking by NanoShaper, it favors well-shaped pockets with higher J , J_{or} , and J_{int} scores. This has practical relevance since a relatively tight and well-shaped pocket reduces the ambiguity and difficulty of the subsequent virtual screening and docking campaigns. Crucially, the proposed method does not aim to distinguish between druggable and

less druggable pockets (binary classification). Rather, a probability for pocket is given, which is easily interpretable and comparable across different proteins. In contrast to a score, the probability estimation does not need *a posteriori* calibration. Rather, the logistic model of the hypersphere naturally delivers this information. Again, a probability allows the computational medicinal chemist to easily identify the most eligible pocket for subsequent drug discovery steps, without wondering if the score value is high or low in absolute terms. This is because any probability very close to one is inevitably a strong indicator. Most importantly, this approach does not need to define a less druggable or non-druggable class. This potentially ambiguous concept is bypassed by the one-class approach. The results show that druggability prediction is best considered as a concept learning problem, rather than a classification problem. This approach allows de-biasing from the start of the learning process, which is clear in the results from the less druggable dataset. We also found that the presence or absence of hydrogen atoms can change the overall modeling attempt in ways that are not always obvious. This is because the effects of NanoShaper are overimposed on the IVDD learning model. Our proposal to mitigate and reduce various biases, even at the cost of lower accuracy, is indebted to the fair machine learning field (Jiang and Nachum, 2019). While fairness concepts are usually applied to social aspects (e.g., demographic parity), we draw on this way of thinking to focus on certain label information only.

Together with explicit structural biases, technical aspects also have an important role. We tested several different values for the small and large NanoShaper probes (data not shown) to identify the pockets. The small probe was easy because there is no reason not to choose the water molecule-like size of 1.4 Å. For the large probe, there is no immediate physically driven quantity, with the convex hull being the extreme solution. We found that a value of 3.5 Å performed better than 3 Å in detecting relatively shallow pockets together with the more prototypical buried ones. Larger values generally led to poorer results in terms of shape, with a systematic decrease in Jaccard index values.

In terms of future developments, we envision several improvements of our methodology. A volume segmentation *ad hoc* algorithm could improve the accuracy, particularly when selecting the value of the large probe. Such a tool could provide more freedom of choice for this parameter. The work of Aggarwal et al. (2021), among others, has shown that many pieces of software for pocket identification tend to identify large pockets without segmentation techniques. Segmentation could be used to find subpockets that are better suited to virtual screening and docking. Another development would be a web server to easily access the tool. Finally, we plan to combine this method with the Pocketron method (La Sala et al., 2017) to not only track the pocket volume and residues over time but also to provide a dynamic druggability score that explicitly considers the probability of the conformation ultimately delivering a Boltzmann weighted estimator.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RA ran the experiments and wrote the manuscript. EG ran the experiments, developed the machine learning code, and wrote the manuscript. RA and EG equally contributed to the study. MB developed the molecular descriptors code and wrote the manuscript. SD designed the research and wrote the

manuscript. AC designed the research and wrote the manuscript.

ACKNOWLEDGMENTS

We thank the HPC team at IIT for computing time and support on the Franklin platform. We thank Grace Fox for proofreading.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2022.870479/full#supplementary-material>

REFERENCES

- Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C. V., and Priyakumar, U. D. (2021). DeepPocket: Ligand Binding Site Detection and Segmentation Using 3d Convolutional Neural Networks. *J. Chem. Inf. Model.* accepted. doi:10.1021/acs.jcim.1c00799
- Agoni, C., Olotu, F. A., Ramharack, P., and Soliman, M. E. (2020). Druggability and Drug-Likeness Concepts in Drug Design: Are Biomodelling and Predictive Tools Having Their Say? *J. Mol. Model.* 26, 120. doi:10.1007/s00894-020-04385-6
- Breiman, L. (2001). Random Forests. *Machine Learn.* 45, 5–32. doi:10.1023/a:1010933404324
- Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., and Funkhouser, T. A. (2009). Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3d Structure. *Plos Comput. Biol.* 5, e1000585. doi:10.1371/journal.pcbi.1000585
- Csermely, P., Kórcsmáros, T., Kiss, H. J., London, G., and Nussinov, R. (2013). Structure and Dynamics of Molecular Networks: A Novel Paradigm of Drug Discovery: A Comprehensive Review. *Pharmacol. Ther.* 138, 333–408. doi:10.1016/j.pharmthera.2013.01.016
- Decherchi, S., and Cavalli, A. (2020a). Fast and Memory-Efficient Import Vector Domain Description. *Neural Process. Lett.* 52, 511–524. doi:10.1007/s11063-020-10243-6
- Decherchi, S., and Cavalli, A. (2020b). Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chem. Rev.* 120, 12788–12833. doi:10.1021/acs.chemrev.0c00534
- Decherchi, S., and Rocchia, W. (2013). A General and Robust ray-casting-based Algorithm for Triangulating Surfaces at the Nanoscale. *PLOS ONE* 8, e59744–15. doi:10.1371/journal.pone.0059744
- Decherchi, S., and Rocchia, W. (2016). Import Vector Domain Description: A Kernel Logistic One-Class Learning Algorithm. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 1722–1729. doi:10.1109/TNNLS.2016.2547220
- Decherchi, S., Spitaleri, A., Stone, J., and Rocchia, W. (2018). NanoShaper-VMD Interface: Computing and Visualizing Surfaces, Pockets and Channels in Molecular Systems. *Bioinformatics* 35, 1241–1243. doi:10.1093/bioinformatics/bty761
- Decherchi, S., Grisoni, F., Tiwary, P., and Cavalli, A. (2021). Editorial: Molecular Dynamics and Machine Learning in Drug Discovery. *Front. Mol. Biosci.* 8, 231. doi:10.3389/fmolb.2021.673773
- Desaphy, J., Azdimousa, K., Kellenberger, E., and Rognan, D. (2012). Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* 52, 2287–2299. doi:10.1021/ci300184x
- Edfeldt, F. N., Folmer, R. H., and Breeze, A. L. (2011). Fragment Screening to Predict Druggability (Ligandability) and lead Discovery success. *Drug Discov. Today* 16, 284–710. doi:10.1016/j.drudis.2011.02.002
- Gao, Z., Li, H., Zhang, H., Liu, X., Kang, L., Luo, X., et al. (2008). Pdbt: a Web-Accessible Protein Database for Drug Target Identification. *BMC bioinformatics* 9, 1–7. doi:10.1186/1471-2105-9-104
- Hajduk, P. J., Huth, J. R., and Fesik, S. W. (2005). Druggability Indices for Protein Targets Derived from Nmr-Based Screening Data. *J. Med. Chem.* 48, 2518–2525. doi:10.1021/jm049131r
- Hussein, H. A., Borrel, A., Geneix, C., Petitjean, M., Regad, L., and Camproux, A.-C. (2015). Pockdrug-server: a New Web Server for Predicting Pocket Druggability on Holo and Apo Proteins. *Nucleic Acids Res.* 43, W436–W442. [pmid]. doi:10.1093/nar/gkv462.25956651
- Jamali, A. A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., and Ebrahimie, E. (2016). Drugminer: Comparative Analysis of Machine Learning Algorithms for Prediction of Potential Druggable Proteins. *Drug Discov. Today* 21, 718–724. doi:10.1016/j.drudis.2016.01.007
- Jiang, H., and Nachum, O. (2019). *Identifying and Correcting Label Bias in Machine Learning*. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, August 26–28, 2019. Editors S. Chiappa and R. Calandra. (Proceedings of Machine Learning Research) 108, 702–712.
- Jolliffe, I. T. (1986). “Principal Components in Regression Analysis,” in *Principal Component Analysis* (Springer), 129–155. doi:10.1007/978-1-4757-1904-8_8
- Kandel, J., Tayara, H., and Chong, K. T. (2021). Puresnet: Prediction of Protein-Ligand Binding Sites Using Deep Residual Neural Network. *J. Cheminformatics* 13, 65. doi:10.1186/s13321-021-00547-7
- Krasowski, A., Muthas, D., Sarkar, A., Schmitt, S., and Brenk, R. (2011). Druggpred: a Structure-Based Approach to Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* 51, 2829–2842. doi:10.1021/ci200266d
- Krivák, R., and Hoksza, D. (2015). Improving Protein-Ligand Binding Site Prediction Accuracy by Classification of Inner Pocket Points Using Local Features. *J. cheminformatics* 7, 1–13. doi:10.1186/s13321-015-0059-5
- Krivák, R., and Hoksza, D. (2018). P2rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure. *J. Cheminform.* 10, 39. doi:10.1186/s13321-018-0285-8
- La Sala, G., Decherchi, S., De Vivo, M., and Rocchia, W. (2017). Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. *ACS Cent. Sci.* 3, 949–960. doi:10.1021/acscentsci.7b00211
- Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an Open Source Platform for Ligand Pocket Detection. *BMC bioinformatics* 10, 1–11. doi:10.1186/1471-2105-10-168
- Mallet, V., Checa Ruano, L., Moine Franel, A., Nilges, M., Druart, K., Bouvier, G., et al. (2021). InDeep: 3D Fully Convolutional Neural Networks to Assist In Silico Drug Design on Protein-Protein Interactions. *Bioinformatics* 38, 1261–1268. doi:10.1093/bioinformatics/btab849
- Nayal, M., and Honig, B. (2006). On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites. *Proteins: Struct. Funct. Bioinformatics* 63, 892–906. doi:10.1002/prot.20897

- Nicolaou, K. C. (2014). Advancing the Drug Discovery and Development Process. *Angew. Chem. Int. Edition* 53, 9128–9140. doi:10.1002/anie.201404761
- Pu, L., Govindaraj, R. G., Lemoine, J. M., Wu, H.-C., and Brylinski, M. (2019). Deepdrug3d: Classification of Ligand-Binding Pockets in Proteins with a Convolutional Neural Network. *PLOS Comput. Biol.* 15, 1–23. doi:10.1371/journal.pcbi.1006718
- Qi, S.-M., Dong, J., Xu, Z.-Y., Cheng, X.-D., Zhang, W.-D., and Qin, J.-J. (2021). Protac: An Effective Targeted Protein Degradation Strategy for Cancer Therapy. *Front. Pharmacol.* 12, 1124. doi:10.3389/fphar.2021.692574
- Schmidtke, P., and Barril, X. (2010). Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *J. Med. Chem.* 53, 5858–5867. doi:10.1021/jm100574m
- Shimokawa, K., Shibata, N., Sameshima, T., Miyamoto, N., Ujikawa, O., Nara, H., et al. (2017). Targeting the Allosteric Site of Oncoprotein Bcr-Abl as an Alternative Strategy for Effective Target Protein Degradation. *ACS Med. Chem. Lett.* 8, 1042–1047. doi:10.1021/acsmedchemlett.7b00247
- Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. (2020). Improving Detection of Protein-Ligand Binding Sites with 3d Segmentation. *Scientific Rep.* 10, 5035. doi:10.1038/s41598-020-61860-z
- Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., and Rarey, M. (2012a). Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* 52, 360–372. doi:10.1021/ci200454v
- Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2012b). DoGSiteScorer: a Web Server for Automatic Binding Site Prediction, Analysis and Druggability Assessment. *Bioinformatics* 28, 2074–2075. doi:10.1093/bioinformatics/bts310
- Wilson, L., and Krasny, R. (2021). Comparison of the Msms and Nanoshaper Molecular Surface Triangulation Codes in the Tabi Poisson–Boltzmann Solver. *J. Comput. Chem.* 42, 1552–1560. doi:10.1002/jcc.26692
- Xie, L., Li, J., Xie, L., and Bourne, P. E. (2009). Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network to Explain the Side Effects of Ceti Inhibitors. *PLoS Comput. Biol.* 5, e1000387. doi:10.1371/journal.pcbi.1000387
- Yuan, Y., Pei, J., and Lai, L. (2013). Binding Site Detection and Druggability Prediction of Protein Targets for Structure-Based Drug Design. *Curr. Pharm. Des.* 19, 2326–2333. doi:10.2174/1381612811319120019
- Yuan, J.-H., Han, S. B., Richter, S., Wade, R. C., and Kokh, D. B. (2020). Druggability Assessment in Trapp Using Machine Learning Approaches. *J. Chem. Inf. Model.* 60, 1685–1699. doi:10.1021/acs.jcim.9b01185
- Zeng, Z.-Q., Yu, H.-B., Xu, H.-R., Xie, Y.-Q., and Gao, J. (2008). “Fast Training Support Vector Machines Using Parallel Sequential Minimal Optimization,” in 2008 3rd international conference on intelligent system and knowledge engineering (Xiamen: IEEE), 997–1001. doi:10.1109/iske.2008.4731075
- Zhang, H., Saravanan, K. M., Lin, J., Liao, L., Ng, J. T.-Y., Zhou, J., et al. (2020). Deepbindpoc: a Deep Learning Method to Rank Ligand Binding Pockets Using Molecular Vector Representation. *PeerJ* 8, e8864. doi:10.7717/peerj.8864

Conflict of Interest: AC and SD are co-founders of BiKi Technologies s.r.l. a company that commercializes the drug discovery software BiKi Life Sciences.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Aguti, Gardini, Bertazzo, Decherchi and Cavalli. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership