

Revolutionizing life sciences: the nobel leap in artificial intelligence-driven biomodeling

Edited by

Valentina Tozzini and Cecilia Giulivi

Published in

Frontiers in Molecular Biosciences



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 978-2-8325-5919-2
DOI 10.3389/978-2-8325-5919-2

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

Revolutionizing life sciences: the nobel leap in artificial intelligence-driven biomodeling

Topic editors

Valentina Tozzini — Nanoscience Institute, Department of Physical Sciences and Technologies of Matter, National Research Council (CNR), Italy

Cecilia Giulivi — University of California, Davis, United States

Citation

Tozzini, V., Giulivi, C., eds. (2025). *Revolutionizing life sciences: the nobel leap in artificial intelligence-driven biomodeling*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-5919-2

Table of contents

- 05 **Editorial: Revolutionizing life sciences: the nobel leap in artificial intelligence-driven biomodeling**
Valentina Tozzini and Cecilia Giulivi
- 10 **Discovery of Small-Molecule Inhibitors of SARS-CoV-2 Proteins Using a Computational and Experimental Pipeline**
Edmond Y. Lau, Oscar A. Negrete, W. F. Drew Bennett, Brian J. Bennion, Monica Borucki, Feliza Bourguet, Aidan Epstein, Magdalena Franco, Brooke Harmon, Stewart He, Derek Jones, Hyojin Kim, Daniel Kirshner, Victoria Lao, Jacky Lo, Kevin McLoughlin, Richard Mosesso, Deepa K. Muruges, Edwin A. Saada, Brent Segelke, Maxwell A. Stefan, Garrett A. Stevenson, Marisa W. Torres, Dina R. Weilhammer, Sergio Wong, Yue Yang, Adam Zemla, Xiaohua Zhang, Fangqiang Zhu, Jonathan E. Allen and Felice C. Lightstone
- 25 **An Extended C-Terminus, the Possible Culprit for Differential Regulation of 5-Aminolevulinate Synthase Isoforms**
Gregory A. Hunter and Gloria C. Ferreira
- 33 **PINet 1.0: A pathway network-based evaluation of drug combinations for the management of specific diseases**
Yongkai Hong, Dantian Chen, Yaqing Jin, Mian Zu and Yin Zhang
- 43 **Identifying SM-miRNA associations based on layer attention graph convolutional network and matrix decomposition**
Jie Ni, Xiaolong Cheng, Tongguang Ni and Jiuzhen Liang
- 59 **Biasing AlphaFold2 to predict GPCRs and kinases with user-defined functional or structural properties**
Davide Sala, Peter W. Hildebrand and Jens Meiler
- 67 **Identification of novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation**
Abdus Samad, Amar Ajmal, Arif Mahmood, Beenish Khurshid, Ping Li, Syed Mansoor Jan, Ashfaq Ur Rehman, Pei He, Ashraf N. Abdalla, Muhammad Umair, Junjian Hu and Abdul Wadood
- 84 **Discovery of a cryptic pocket in the AI-predicted structure of PPM1D phosphatase explains the binding site and potency of its allosteric inhibitors**
Artur Meller, Saulo De Oliveira, Aram Davtyan, Tigran Abramyan, Gregory R. Bowman and Henry van den Bedem
- 95 **Hybrid neural network approaches to predict drug–target binding affinity for drug repurposing: screening for potential leads for Alzheimer’s disease**
Xialin Wu, Zhuojian Li, Guanxing Chen, Yiyang Yin and Calvin Yu-Chian Chen
- 111 **SMG-BERT: integrating stereoscopic information and chemical representation for molecular property prediction**
Jiahui Zhang, Wenjie Du, Xiaoting Yang, Di Wu, Jiahe Li, Kun Wang and Yang Wang

- 121 **Structural insights into the C-terminus of the histone-lysine N-methyltransferase NSD3 by small-angle X-ray scattering**
Benny Danilo Belviso, Yunpeng Shen, Benedetta Carrozzini, Masayo Morishita, Eric di Luccio and Rocco Caliandro
- 134 **An automatic diagnosis model of otitis media with high accuracy rate using transfer learning**
Fangyu Qi, Zhiyu You, Jiayang Guo, Yongjun Hong, Xiaolong Wu, Dongdong Zhang, Qiyuan Li and Chengfu Cai
- 143 **Deep learning-based classification of the capillary ultrastructure in human skeletal muscles**
Marius Reto Bigler and Oliver Baum
- 155 **Active and machine learning-enhanced discovery of new FGFR3 inhibitor, Rhapontin, through virtual screening of receptor structures and anti-cancer activity assessment**
Qingxin Zeng, Haichuan Hu, Zhengwei Huang, Aotian Guo, Sheng Lu, Wenbin Tong, Zhongheng Zhang and Tao Shen
- 168 **AlphaFold2 in biomedical research: facilitating the development of diagnostic strategies for disease**
Hong Zhang, Jiajing Lan, Huijie Wang, Ruijie Lu, Nanqi Zhang, Xiaobai He, Jun Yang and Linjie Chen
- 184 **Shedding light on the *DICER1* mutational spectrum of uncertain significance in malignant neoplasms**
D. S. Bug, I. S. Moiseev, Yu. B. Porozov and N. V. Petukhova



OPEN ACCESS

EDITED AND REVIEWED BY:

Graça Soveral,
University of Lisbon, Portugal

*CORRESPONDENCE

Valentina Tozzini,
✉ valentina.tozzini@nano.cnr.it

RECEIVED 06 December 2024

ACCEPTED 09 December 2024

PUBLISHED 03 January 2025

CITATION

Tozzini V and Giulivi C (2025) Editorial:
Revolutionizing life sciences: the nobel leap in
artificial intelligence-driven biomodeling.
Front. Mol. Biosci. 11:1540823.
doi: 10.3389/fmolb.2024.1540823

COPYRIGHT

© 2025 Tozzini and Giulivi. 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.

Editorial: Revolutionizing life sciences: the nobel leap in artificial intelligence-driven biomodeling

Valentina Tozzini^{1,2*} and Cecilia Giulivi^{3,4}

¹Istituto Nanoscienze del Consiglio Nazionale delle Ricerche (CNR), Lab NEST-Scuola Normale Superiore, Pisa, Italy, ²Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Pisa, Pisa, Italy, ³Department of Molecular Biosciences, School of Veterinary Medicine, University of California Davis, Davis, CA, United States, ⁴MIND Institute, University of California at Davis Medical Center, Sacramento, CA, United States

KEYWORDS

deep-learning, neural networks, structure prediction, drug design, disordered proteins, biomolecules interactions

Editorial on the Research Topic

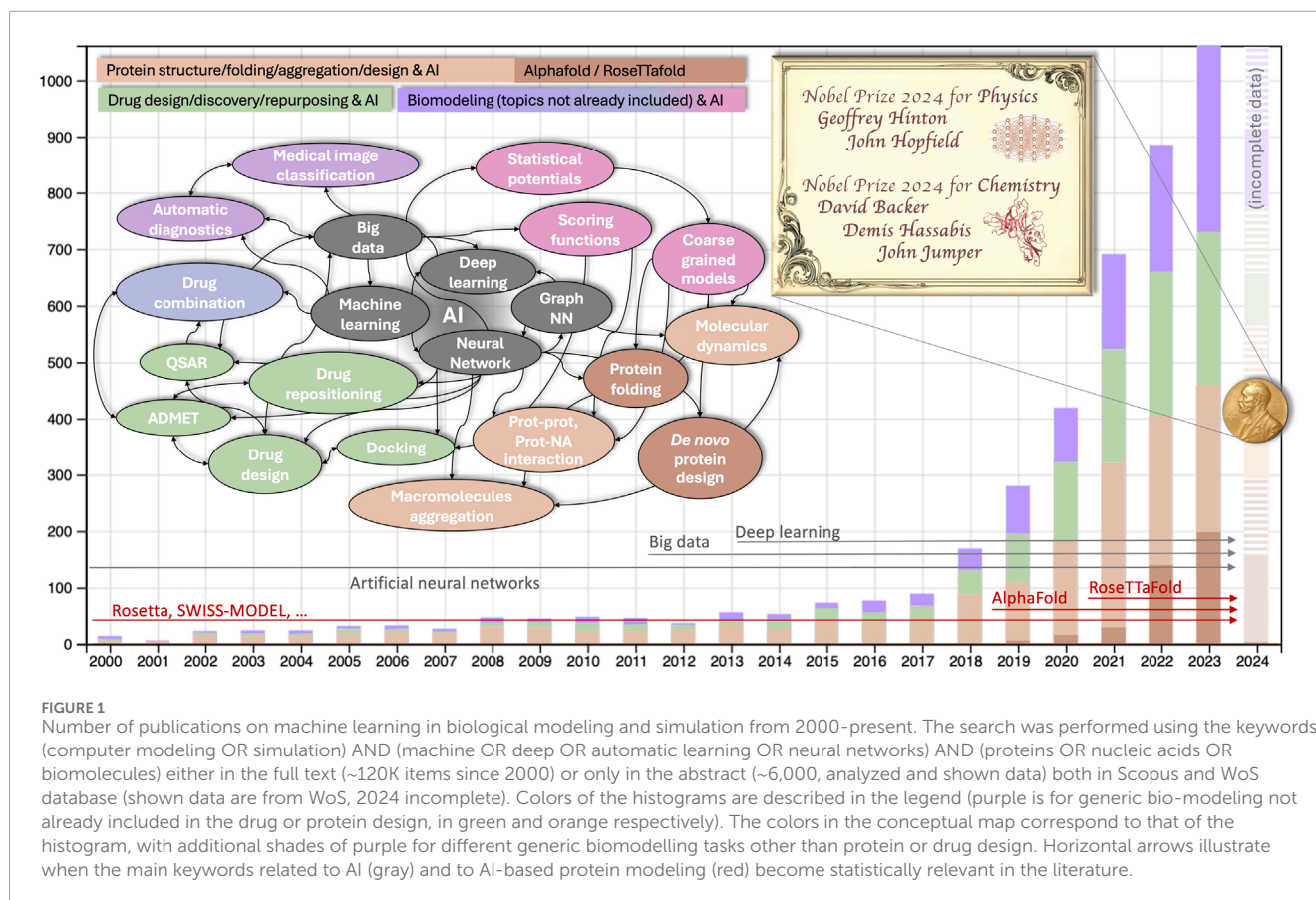
Revolutionizing life sciences: the nobel leap in artificial intelligence-driven biomodeling

1 Artificial intelligence's impact on biomolecular modeling

Within the research world, 2024 will be remembered as the year of Nobel Prizes for Artificial Intelligence (AI). The one for Physics, awarded to John Hopfield and Geoffrey Hinton *for foundational discoveries and inventions that enable machine learning with artificial neural networks*, has sealed the connection between physics and information science, now officially mating on a strongly interdisciplinary frontier field after over 50 years of fruitful interaction (Artificial, 2024). More specifically, connecting AI to biomolecular modeling relates to the Nobel Prize in Chemistry awarded to David Baker *for computational protein design* and to Demis Hassabis and John Jumper *for protein structure prediction*.

Numerous statistics illustrate the influence of artificial intelligence in the field of biomodeling. An inquiry conducted in scientific literature databases employing AI-related keywords pertinent to the computer modeling of biomolecules yields approximately 120,000 results (approximately 6,000 results if the search is confined to the abstract, as illustrated in Figure 1). The exponential rise observed starting from 2018–19 was the prelude to the Nobel, and approximately coincides with the appearance of the two software suites, AlphaFold (Senior et al., 2019) and RosettaFold (Humphreys et al., 2021), which implement the methods for proteins folding and proteins *de novo* design developed by Hassabis/Jumper and Baker, respectively.

Receiving a Nobel Prize just a few years after the awarded research is quite rare, but certainly not accidental. The methods for protein structure prediction based on homology modeling were developed starting in the 1990s and implemented in popular



software suites, including the early version of Rosetta (Bowers et al., 2000) and others [e.g., SWISS-MODEL (Guex and Peitsch, 1997)]. These methods heavily depend on statistical data. They involve aligning and ranking sequences and structures and parameterizing scoring functions through extensive analysis of sequence and structure databases. This process culminates in distilling the information into a few optimal structures or interaction models (Wang et al., 2019). Over the years, the growing volume of statistical data has necessitated the automation of tasks, particularly in searching and comparing information. Advancements in hardware architecture and storage capacity have supported this shift.

Meanwhile, automatically trained neural networks (NN) have emerged as a natural solution for the “distillation” of this data (Kanada et al., 2024). During the second decade of 2000s, the co-evolution of computer performance and algorithms led to the transition from *machine learning* (ML) to *deep learning* (DL). This shift involved adding layers to the neural networks, resulting in qualitative and quantitative predictive power improvements. The combination of an established supportive environment, the availability of big data, and the rise of DL has significantly contributed to the success of AI methods in bio-modeling.

Specifically regarding protein structure, AlphaFold now achieves an impressive 99% accuracy in predicting single-chain proteins, rendering the CASP challenge—historically focused on structure prediction—less relevant.

Besides the modeling of protein structures, a significant domain of artificial intelligence application elucidated by statistical analysis pertains to drug development. In particular, ML is used to address structure-activity relationships (Gupta et al., 2021) and uptake-toxicity of the drug (De Carlo et al., 2024), virtual screening, and structure-based design. While not claiming to cover all potential applications, we note that optimizing force fields for low-resolution models of biomolecules significantly benefits from machine learning (Kanada et al., 2024; Majewski et al., 2023; Mirarchi et al., 2024), whereas the application of graph neural networks for calculating molecular dynamical trajectories is a cutting-edge approach (Husic et al., 2020).

2 AI's impact on biological modeling and simulation in Frontiers in Molecular Biosciences

Frontiers in Molecular Biosciences (FMB) has witnessed an exponential rise of publications with the exact timing and similar topical distribution, currently counting several hundreds of publications on AI related topics. The section of Biological Modeling and Simulation (BMS) is one the most involved, having issued several Research Topic Collections (Research Topics, RT) on the diverse applications of neural networks in biomolecular simulations, on the prediction of protein structure and conformation, or focusing on data-driven applications, on drug design, even combined with

molecular studies of **metabolic pathways** also in relation to the **cancer treatment**.

A deeper look into the BMS section also reveals more specific topics out of the mainstream, such as the **prediction of protein-protein interactions** and the study of the **conformation of intrinsically disordered proteins**. Indeed, these are two aspects where ML algorithms show their weakness (Abramson et al., 2024), displaying decreased accuracy. This is attributed to the underrepresentation within the training dataset of crucial features, such as the conformational variability of disordered proteins and protein-protein interfaces (Saldano et al., 2022), especially when combined with sequence variability, e.g., in the study of antibodies (Yin et al., 2022). The decreased accuracy and predictive power in cases “too far” from those included in the learning dataset is considered one of the main drawbacks of automatic learning-based methods.

2.1 Beyond the stream and into the niches of AI applications

To explore unconventional AI methods for bio-modeling and showcase niche applications and challenging or problematic areas, we have compiled 15 “orphan” papers in this Research Topic. These papers, which are not part of any existing topical collection, have been published in the sections of Biological Modeling and Simulation or Structural Biology of FMB.

In the review by Zhang et al. it is noted that AlphaFold, along with other similar AI methods for structure prediction, such as RoseTTaFold and EMSFold, is widely used in various fields of biomedical research. In addition to drug design, the authors highlight its applications in immunology, particularly in predicting and designing immunoglobulin structures or developing structure-based vaccines. The work also emphasizes the development of biomarkers, the study of protein-protein and protein-nucleic acid interactions, and the investigation of missense mutations. However, the review points out some limitations of these methods, specifically the decreased accuracy in predicting the relative positioning of large protein domains and their intrinsically disordered regions and challenges in differentiating between various environmental conditions. In this regard, alternative approaches like AminoBERT, described in Zhang et al., demonstrate better performance in *de novo* design or when few homologous sequences are available. This improvement is achieved by omitting the multiple sequence alignment step and instead incorporating residue-based chemical and geometric information.

The absence of specific protein information in the training data and the resulting bias towards the included proteins are two sides of the same coin, which makes the neural network predictions contingent on the dataset's composition. Sala et al. transformed the challenge into an opportunity by introducing a controlled bias in AlphaFold2 toward specific user-defined subsets of structures. This can be achieved by incorporating genetic information to enhance accuracy for particular protein families. The algorithm has demonstrated improved performance on CPCRs and kinase protein families, which are notably difficult due to their multiple active conformations. Additionally, the capability of AlphaFold to address different or multiple structures was discussed in the mini-review by Hunter et al. This study focused on examining the structure of ALAS

synthase, specifically highlighting a predicted divergence in the C-terminal domain of the protein and its connection to the proposed allosteric regulation of protein activity.

2.2 Integrating AI and simulation techniques: advancing biomolecular structure prediction and drug discovery

Utilizing a diverse array of methods has demonstrated remarkable effectiveness in accurately predicting the structures of biomolecules. The structure predicted by AlphaFold, along with Molecular Dynamics (MD) simulations, served as the reference for evolutionary studies. Just to cite a few ones highlighting this link, the study by Bug et al. on the ribonuclease Dicer1 involved in miRNA biogenesis and hematological cancers progression, and that by Meller et al. to generate the structure of the unknown protein PPM1D phosphatase, an important marker in oncology involved in the regulation of DNA damage response. In these cases, the structure was combined with a graph convolutional network model trained over activity data, and with MD simulations to enhance the drug docking task, revealing an allosteric “cryptic” pocket, not immediately accessible and therefore escaping the structural-only analysis. Belviso et al. used AlphaFold and MD in combination with small-angle X-ray scattering to characterize the C-terminal region of NSD3 histone lysine methyltransferases, a marker in oncogenesis, showing that combined modeling techniques can be used to augment the low resolution experimental structural characterization techniques.

2.3 Advancing drug discovery: integrating AI, simulations, and experimental methods for targeted therapeutics

Drug design increasingly benefits from interdisciplinary approaches combining advanced computational techniques and ML with experimental validation to accelerate therapeutic discovery and innovation. Zeng et al. used a cascade of structure-based drug design methods combining MD and metadynamics of the drug-target complex with ML-based virtual screening and QSAR and ADMET evaluation. Combined with experimental procedures, this approach identified inhibitors of fibroblast growth factor receptors that were also tumor suppressors.

Drug design represents a promising frontier for advancing NN development, particularly at the algorithmic level. The complexity of molecular interactions, coupled with the need to predict binding affinities, toxicity, and pharmacokinetics, provides a fertile ground for refining and innovating NN architectures. Emerging techniques, such as graph-based neural networks and attention mechanisms, are poised to address these challenges by enabling more accurate modeling of molecular properties and interactions, paving the way for breakthroughs in computational drug discovery. Ni et al. developed a model of a Graph Convolutional Network with a layer attention mechanism and trained it to predict the association of small molecules to target miRNA. Despite the large number of hidden layers and advanced mechanisms to cope with data redundancies and reduce the noise, the authors claim

dissatisfaction with the specific task, possibly due to insufficient variability in the dataset. [Wu et al.](#) combined an NN with docking and virtual screening to repurpose drugs for Alzheimer's disease, which allows the optimization of a multi-target approach capable of identifying the network of proteins interacting with the receptor S1R, considered as the starting target, and subsequently identifying several leads, tested by docking and ADMET prediction. To a similar scope of finding effective combinations of drugs for multifactorial diseases, [Hong et al.](#) develop a different NN approach independent of structures and based on the Pathway Interaction Network (PINet), which was tested on acute myeloid leukemia, where it correctly predicted midostaurin and gemtuzumab as effective drug combinations and proved particularly effective when the training dataset is limited.

We should pay attention to the early research on antivirals targeting the main protease of SARS-CoV-2 in the context of structure-based drug design. [Lau et al.](#) combined molecular docking and MD with a convolutional neural network and spatial graph model trained on ligand-protein data, used to predict the ligand-protein score and identify from a library of 26 million molecules possible candidate compounds to target RBD domain of the Spike protein or Mpro. Using biolayer interferometry for the spike protein and a FRET-based reporter, their effective binding was tested. [Samad et al.](#) considered as the target the chymotrypsin-like protease (3CL^{PRO}) and used machine learning-based virtual screening of 4,000 phytochemicals. The Random Forest model, displaying 98% accuracy on the train and test set, identified several molecules that were subsequently docked into the target and analyzed by MD. The procedure identified 26 potential inhibitors.

Finally, we mention a couple of applications within the biological modeling area that are out of the mainstream, not on molecular modeling but on using images for diagnostics. [Bigler et al.](#) use a deep learning approach with transfer learning of a pre-trained convolutional neural network to identify pathological patterns in skeletal muscle biopsies, using transmission electron microscopy images showing that the learned network is proven superior in the classification concerning commonly used morphometric analyses. More specifically, [Qi et al.](#) trained an NN to automatically diagnose suppurative otitis media and middle ear cholesteatoma, proving a handy tool to help physicians discern these two chronic diseases displaying similar CT medical images.

3 Perspectives

In the last decade, AI has produced a massive acceleration in biomolecular modeling, making several tasks previously requiring a long time and specific expertise fast and easy. These are, in particular, those involving analyzing and synthesizing information from large amounts of data. The case of AlphaFold is an exemplar: the current version allows even nonexperts in the field to have a prediction of the fold of a protein from the sequence in minutes, a task which required weeks with the traditional homology modeling procedure, and reaching comparable or superior accuracy in most of the cases.

Despite its remarkable progress, AI-driven biomolecular modeling faces significant challenges highlighting the need for

caution and critical evaluation. One major issue lies in the bias and incompleteness of training databases. This risks to produce results that reflect the limitations or skewed composition of the input data, potentially leading to inaccurate predictions and amplifies the risk of “hallucinations” – outputs that are highly ranked, but scientifically invalid—possibly due to overfitting and extrapolation beyond known data. Beyond hallucinations, we already commented on the cases of disordered structures and inter-domain interface prediction, whose low confidence the ML models can autonomously evaluate. In addition, AI-driven platforms like DeepMind's AlphaFold have predicted novel drug candidates for various diseases, but still, several of these compounds need to be sufficiently followed up regarding their pharmacokinetics, such as IC50 values (the concentration needed to inhibit 50% of a target) or their ability to be administered effectively. In some cases, promising compounds identified by AI have yet to pass crucial stages in drug development, such as formulation stability, bioavailability, or FDA approval. A notable case is the identification of AI-generated inhibitors for the SARS-CoV-2 virus, which, while initially promising, failed to meet the necessary clinical standards and were ultimately not pursued for broader therapeutic use.

Furthermore, the need for explainability in many AI models compounds these challenges. Without transparent mechanisms to trace how predictions are made, it becomes difficult for researchers to assess their reliability or identify potential errors. This opacity raises concerns about the reproducibility and trustworthiness of AI-generated insights, particularly in high-stakes fields like drug discovery or biomolecular engineering. Adding explainability to the method, and not only in the biomodelling field, is currently one of the main challenges for developing automatic learning algorithms. On the technical level, one way to address this problem as far as that of (explicit or not) low reliability and bias, is to reduce the complete automatism by re-introducing into the procedure elements of symbolic artificial intelligence based on deductive rules into a hybrid approach known as neuro-symbolic AI ([Bhuyan et al., 2024](#)).

On a philosophical level, the growing reliance on AI may inadvertently foster excessive trust in its outputs, sometimes at the expense of scientific scrutiny. This overconfidence could lead to a diminished critical sense, where the technology's predictions are only accepted without adequate validation. For instance, some AI-predicted compounds have led to follow-up studies that overlook crucial aspects like side effects, toxicity, or long-term efficacy, which must be fully captured in the initial models. To mitigate these risks, fostering interdisciplinary collaboration, emphasizing data quality, and developing interpretable AI systems are essential to ensure that AI remains a robust and reliable tool for advancing biomolecular research.

In conclusion, while it is true that AI presents challenges and risks, it also offers transformative opportunities when wielded responsibly. We are at a juncture where AI is no longer just an optional tool but a cornerstone of modern modeling and problem-solving. Like any tool, its effectiveness depends on the skill and wisdom of its user. By combining the power of AI with the irreplaceable intuition and common sense of human

judgment, we can harness its potential for innovation and progress, ensuring a future where technology enhances, rather than replaces, our humanity.

Author contributions

VT: Conceptualization, Investigation, Methodology, Writing–original draft, Writing–review and editing. CG: Conceptualization, Investigation, Methodology, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was supported by Next Generation-EU (PNRR), through the project Tuscany Health Ecosystem (THE-Spoke 1, grant ECS 00000017), and INFN CSN5 through the MIRO project (VT) and partially by NIH NS128751 and discretionary funds (CG).

Acknowledgments

We thank Hannah Jacob (Content Specialist; Frontiers in Molecular Biosciences) and Emily Croft (Journal Manager; Frontiers in Molecular Biosciences) for their technical assistance, and contributing with materials.

References

- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. *NATURE* 630, 493–500. doi:10.1038/s41586-024-07487-w
- Artificial (2024). Artificial boundaries. *Nat. Phys.* 20, 1689. doi:10.1038/s41567-024-02717-4
- Bhuyan, P. B., Ramdane-Cherif, A., Tomar, R., and Singh, T. P. (2024). Neuro-symbolic artificial intelligence, a survey. *Neur. Comp. Applic.* 36, 12809–12844. doi:10.1007/s00521-024-09960-z
- Bowers, P., Strauss, C. E. M., and Baker, D. (2000). *De novo* protein structure determination using sparse nmr data. *J. Biomol. NMR* 18, 311–318. doi:10.1023/A:1026744431105
- De Carlo, A., Ronchi, D., Piastra, M., Tosca, E. M., and Magni, P. (2024). Predicting admet properties from molecule smile: a bottom-up approach using attention-based graph neural networks. *PHARMACEUTICS* 16, 776. doi:10.3390/pharmaceutics16060776
- Guex, N., and Peitsch, M. (1997). Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. *ELECTROPHORESIS* 18, 2714–2723. doi:10.1002/elps.1150181505
- Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., and Kumar, P. (2021). Artificial intelligenceto deep learning: machine intelligence approach for drug discovery. *Mol. Divers.* 25, 1315–1360. doi:10.1007/s11030-021-10217-3
- Humphreys, I. R., Pei, J., Baek, M., Anishchenko, I., Ovchinnikov, S., Zhang, J., et al. (2021). Computed structures of core eukaryotic protein complexes. *Science* 374, eabm4805. doi:10.1126/science.abm4805
- Husic, B. E., Charron, N. E., Lemm, D., Wang, J., Pérez, A., Majewski, M., et al. (2020). Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.* 153, 194101. doi:10.1063/5.0026133
- Kanada, R., Tokuhisa, A., Nagasaka, Y., Okuno, S., Amemiya, K., Chiba, S., et al. (2024). Enhanced coarse-grained molecular dynamics simulation with a smoothed hybrid potential using a neural network model. *J. Chem. Theory Comput.* 20, 7–17. doi:10.1021/acs.jctc.3c00889
- Majewski, M., Pérez, A., Thölke, P., Doerr, S., Charron, N. E., Giorgino, T., et al. (2023). Machine learning coarse-grained potentials of protein thermodynamics. *Nat. Comm.* 14, 5739. doi:10.1038/s41467-023-41343-1
- Mirarchi, A., Peláez, R. P., Simeon, G., and Fabritiis, G. D. (2024). Amaro: all heavy-atom transferable neural network potentials of protein thermodynamics. *J. Chem. Theor. Comput.* 20, 9871–9878. doi:10.1021/acs.jctc.4c01239
- Saldano, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Rueda, A. J. V., et al. (2022). Impact of protein conformational diversity on alphafold predictions. *BIOINFORMATICS* 38, 2742–2748. doi:10.1093/bioinformatics/btac202
- Senior, A. W., Evans, R., Jumper, J., Sifre, L., Green, T., et al. (2019). Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (caspl3). *Proteins* 18, 1141–1148. doi:10.1002/prot.25834
- Wang, J., Olsson, S., Wehmeyer, C., Perez, A., Charron, N. E., de Fabritiis, G., et al. (2019). Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.* 5, 755–767. doi:10.1021/acscentsci.8b00913
- Yin, R., Feng, B. Y., Varshney, A., and Pierce, B. G. (2022). Benchmarking alphafold for protein complex modeling reveals accuracy determinants. *PROTEIN Sci.* 31, e4379. doi:10.1002/pro.4379

Conflict of interest

All authors have disclosed any financial or other interests related to the submitted work that could impact the author's objectivity or influence the article's content. CG serves as an Editorial Board Member of Scientific Reports. She has received compensation as a Field Chief Editor for Frontiers in Molecular Biosciences and honoraria for participating in NIH peer review meetings. VT is the Specialty Chief Editor of the section Biological modeling and Simulation of Frontiers in Molecular Biosciences.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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.



Discovery of Small-Molecule Inhibitors of SARS-CoV-2 Proteins Using a Computational and Experimental Pipeline

Edmond Y. Lau¹, Oscar A. Negrete², W. F. Drew Bennett¹, Brian J. Bennion¹, Monica Borucki¹, Feliza Bourguet¹, Aidan Epstein³, Magdalena Franco¹, Brooke Harmon⁴, Stewart He³, Derek Jones³, Hyojin Kim⁵, Daniel Kirshner¹, Victoria Lao¹, Jacky Lo¹, Kevin McLoughlin³, Richard Mosesso⁴, Deepa K. Muruges¹, Edwin A. Saada⁴, Brent Segelke¹, Maxwell A. Stefan⁴, Garrett A. Stevenson⁶, Marisa W. Torres³, Dina R. Weilhammer¹, Sergio Wong¹, Yue Yang¹, Adam Zemla³, Xiaohua Zhang¹, Fangqiang Zhu¹, Jonathan E. Allen³ and Felice C. Lightstone^{1*}

OPEN ACCESS

Edited by:

Gennady Verkhivker,
Chapman University, United States

Reviewed by:

Luisa Di Paola,
Campus Bio-Medico University, Italy
Rakesh Kumar Tiwari,
Chapman University, United States

*Correspondence:

Felice C. Lightstone
lightstone1@llnl.gov

Specialty section:

This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: 12 March 2021

Accepted: 22 June 2021

Published: 09 July 2021

Citation:

Lau EY, Negrete OA, Bennett WFD, Bennion BJ, Borucki M, Bourguet F, Epstein A, Franco M, Harmon B, He S, Jones D, Kim H, Kirshner D, Lao V, Lo J, McLoughlin K, Mosesso R, Muruges DK, Saada EA, Segelke B, Stefan MA, Stevenson GA, Torres MW, Weilhammer DR, Wong S, Yang Y, Zemla A, Zhang X, Zhu F, Allen JE and Lightstone FC (2021) Discovery of Small-Molecule Inhibitors of SARS-CoV-2 Proteins Using a Computational and Experimental Pipeline. *Front. Mol. Biosci.* 8:678701. doi: 10.3389/fmolb.2021.678701

¹Lawrence Livermore National Laboratory, Physical and Life Sciences Directorate, Biotechnology and Biosciences Division, Livermore, CA, United States, ²Sandia National Laboratory, Department of Biotechnologies and Bioengineering, Livermore, CA, United States, ³Lawrence Livermore National Laboratory, Computing Directorate, Global Security Computing Division, Livermore, CA, United States, ⁴Sandia National Laboratory, Department Systems Biology, Livermore, CA, United States, ⁵Lawrence Livermore National Laboratory, Computing Directorate, Center for Applied Scientific Computing, Livermore, CA, United States, ⁶Lawrence Livermore National Laboratory, Engineering Directorate, Computational Engineering Division, Livermore, CA, United States

A rapid response is necessary to contain emergent biological outbreaks before they can become pandemics. The novel coronavirus (SARS-CoV-2) that causes COVID-19 was first reported in December of 2019 in Wuhan, China and reached most corners of the globe in less than two months. In just over a year since the initial infections, COVID-19 infected almost 100 million people worldwide. Although similar to SARS-CoV and MERS-CoV, SARS-CoV-2 has resisted treatments that are effective against other coronaviruses. Crystal structures of two SARS-CoV-2 proteins, spike protein and main protease, have been reported and can serve as targets for studies in neutralizing this threat. We have employed molecular docking, molecular dynamics simulations, and machine learning to identify from a library of 26 million molecules possible candidate compounds that may attenuate or neutralize the effects of this virus. The viability of selected candidate compounds against SARS-CoV-2 was determined experimentally by biolayer interferometry and FRET-based activity protein assays along with virus-based assays. In the pseudovirus assay, imatinib and lapatinib had IC₅₀ values below 10 μM, while candesartan cilexetil had an IC₅₀ value of approximately 67 μM against M^{pro} in a FRET-based activity assay. Comparatively, candesartan cilexetil had the highest selectivity index of all compounds tested as its half-maximal cytotoxicity concentration 50 (CC₅₀) value was the only one greater than the limit of the assay (>100 μM).

Keywords: COVID-19, molecular simulations, machine-learning, protein assays, FRET, live virus, main protease, spike protein

INTRODUCTION

In December 2019, the first cases of a novel coronavirus (SARS-CoV-2) were reported in Wuhan city, Hubei province of China (World Health Organization, 2020). Symptoms of the first patients were flu-like and included fever, dry cough, headache, and myalgia, but with a tendency to develop into potentially fatal dyspnea and acute respiratory distress syndrome (Huang et al., 2020). Within a matter of weeks this coronavirus had spread to many parts of China and preliminary evidence suggests its ability to pass between people without showing outward symptoms (Rothe et al., 2020). Additionally, its transmissibility is higher than that of SARS-CoV (Xia et al., 2020). These features and likely others in the coronavirus as well as the ease of international travel has allowed the outbreak to reach every populated continent. Many countries have taken the extraordinary measure of locking down cities with populations in the millions to slow the spread of the virus. As of this writing, over 98,000,000 people have contracted SARS-CoV-2 with more than 2,100,000 fatalities worldwide (WHO Coronavirus Disease, 2020). Phylogenetic analysis of the genomic sequence of SARS-CoV-2 has shown that it is a member of the betacoronavirus genus and related to SARS-CoV and MERS-CoV (Letko et al., 2020). SARS-CoV-2 has so far has been shown to be resistant to treatments developed for its related viruses although the compound remdesivir has shown some promise and has been approved for emergency use (Beigel et al., 2020).

A concerted effort worldwide has been placed on solving protein structures from SARS-CoV-2 to better understand the lifecycle of the virus and to provide targets for vaccines and drugs (Scudellari, 2020). The trimeric spike protein was the first protein from SARS-CoV-2 to be solved and was shown to be very similar in structure to the homologous protein in SARS-CoV (Wrapp et al., 2020). Coronaviruses utilize the spike protein to recognize binding sites on cells and anchor themselves to invade their host (Belouzard et al., 2012). The spike protein has been solved by X-ray crystallography and cryo-electron microscopy with its receptor binding domain (RBD) in complex with the human receptor protein angiotensin-converting enzyme 2 (ACE2) (Lan et al., 2020; Wrapp et al., 2020). The binding of RBD to human ACE2 that allows the virus to enter the cell is very strong at 4.7–14.7 nM but surprisingly the binding interaction does not occur over a large surface area (Lan et al., 2020; Wrapp et al., 2020). Many of the ACE2-RBD interactions are located within two large loop regions in the RBD and primarily through sidechain-sidechain interactions.

The other solved protein structure from SARS-CoV-2 used in this study is the main protease (Mpro). The Mpro is a cysteine protease with a catalytic dyad consisting of Cys145 and His41. The dimeric main protease is ubiquitous in coronaviruses and plays a pivotal role in viral gene expression and replication through proteolytic processing of replicase polyproteins (Ullrich and Nitsche, 2020). The SARS-CoV-2 Mpro structure has recently been solved with the covalent inhibitor N3 and released in the Protein Data Bank (PDB, 6LU7) (Jin et al., 2020). A second structure of the SARS-CoV-2 Mpro was made available without a bound inhibitor (6Y84) (Owen et al., 2020). The main

protease has a large gorge that binds and cleaves polypeptides that are critical for maturation of the virus and is an attractive site for new inhibitors.

The RBD domain of the spike protein and Mpro are promising targets for *in silico* small molecule studies to find molecules with inhibitory properties. We have performed a combined molecular docking, molecular dynamics simulation, and machine learning study in an effort to identify molecules that may bind to the RBD domain and/or Mpro. These bound molecules may attenuate or neutralize the effects of this virus. These predicted ligands were then tested experimentally for their ability to bind their partner protein using biolayer interferometry for the spike protein and a FRET-based reporter substrate for Mpro. Compounds that were found to bind were further tested in virus-based assays to determine their ability to neutralize SARS-CoV-2.

MATERIALS AND METHODS

Molecular Dynamics Simulations of the Apo-Proteins of the RBD of Spike and Main Protease

Classical molecular dynamics simulations were performed using the program OpenMM (Version 7.4) (Eastman et al., 2017). The AMBER force field was used for the proteins in the system (Maier et al., 2015). The individual proteins (RBD of the spike protein or the dimer of the main protease) were solvated in a TIP3P water box (Jorgensen et al., 1983) and the appropriate numbers of ions (Na^+ or Cl^-) were added to neutralize the system. M^{pro} was modeled as its biologically-appropriate dimer. AM1-BCC charges (Jakalian et al., 2002) were used to model the thiolate of Cys145 and His41 was modeled as protonated in M^{pro} . The density of the water was simulated at 1.0 g/ml. The energy of the system was minimized before dynamics. The molecular dynamics simulations were performed in an NPT ensemble using the Langevin integrator (Salomon-Ferrer et al., 2013b). The system was coupled to a Monte Carlo thermostat at 300 K. Non-bonded interactions were cutoff at 8 Å. The electrostatics was treated using Particle Mesh Ewald summation with an 8 Å real space cutoff and a 1 Å grid (Darden et al., 1993). SHAKE was used to constrain bonds containing hydrogens (Ryckaert et al., 1977). A 2.0 fs timestep was used and each simulation was run to 100 ns. The temperature of the system was increased in increments of 50 K for 100 ps. Positional constraints were placed on backbone atoms (C, N, and CA) with a force constant of 1 kcal/mole•Å² while the temperature was increased. Once the system has reached 300 K, an additional 1.5 ns of dynamics was performed with the positional constraints, after this time period 100 ns of dynamics was performed without the constraints.

Molecular Docking and Rescoring Calculations

The in-house ConveyorLC toolchain (Zhang et al., 2014; Zhang et al., 2017) was used to automate the docking and rescoring of

compounds against each of the four binding sites identified (two spike sites and two M^{pro} structures/conformations). This toolchain comprises four parallel programs for protein preparation (CDT1Receptor), ligand preparation (CDT2Ligand), molecular docking (CDT3Docking), and Molecular Mechanics/Generalized Born-Solvent Accessible Surface Area (MM/GBSA) rescoring (CDT4mmgbsa). The ConveyorLC toolchain depends on a number of external libraries, including the Message Passing Interface (MPI) library, the C++ Boost library, the Conduit library, the HDF5 library, and several molecular simulation packages, including Autodock Vina, (Trott and Olson, 2010) the AMBER molecular simulation package (Salomon-Ferrer et al., 2013a), and MGLTOOLS (Morris et al., 2009). Computational results are aggregated and saved in a series of HDF5 files. A few auxiliary tools are included in the toolchain to query and extract data in the HDF5 files.

Over 26 million compounds were selected from four publicly available compound libraries for docking. The ZINC database (Sterling and Irwin, 2015) FDA-approved and “world-not-FDA” drugs were assembled into a “world-approved 2018” set. From ChEMBL, approximately 1.5 million unique compounds were used (Gaulton et al., 2012). From EMOLECULES, approximately 18 M compounds were used (eMOLECULES, 2020). The remaining compounds were selected from the Enamine “REAL” database of over 1.2 billion enumerated structures of drug-like compounds predicted to be synthetically feasible (Enamine, 2020).

The CDT3Docking in the ConveyorLC toolchain is based on Autodock Vina (Version 1.1.2) and uses MPI and a multithreading hybrid parallel scheme (Trott and Olson, 2010; Zhang et al., 2013). The docking grids of the binding sites were determined by the protein preparation program in the toolchain. Compounds were prepared for docking in the following manner. SMILES strings and 2D SDF structures were imported into the Molecular Operating Environment (MOE) [Molecular Operating Environment (MOE), 2020] for removal of salts and metal-containing ligands, protonation states were set to the dominant form at pH 7, 3D structures were created and minimized, and relevant MOE descriptors were calculated. The final structures were exported from MOE as SDF files. These structures were then further processed by the ligand preparation in the toolchain by utilizing antechamber and the GAFF force field from the AMBER simulation package (Salomon-Ferrer et al., 2013a).

The over 26 million compounds described above were individually docked into each binding site for a total of more than 100 million docking simulations. An exhaustiveness of 16 was used for ligand pose sampling. The top 10 poses were kept for each docking calculation. Compounds that had a docking score equal to or better than -7.5 kcal/mole were saved in HDF5 files for further study. Using this score threshold, we selected $\sim 1\%$ of total compounds or approximately 1 million protein-compound complexes for each binding site.

The selected protein-compound complexes were rescored using CDT4mmgbsa in the ConveyorLC toolchain. A total of ~ 10 million poses were rescored for each binding site because each complex typically had 10 docking poses. CDT4mmgbsa

employs a master-worker parallel scheme, where the master is in charge of job dispatching and each worker receives jobs from the master and performs an MM/GBSA calculation using the AMBER sander program. The AMBER force field (amberff14SB) (Maier et al., 2015) was used for the proteins; the apo proteins' MM/GBSA energies were previously determined in the CDT1Receptor step. Partial atomic charges for the compounds were computed by antechamber using the AM1-BCC method (Jakalian et al., 2002); each compound's charges were previously calculated by the CDT2Ligand step. An energy minimization—1,000 steps of steepest descent and 1,000 additional steps of conjugate gradient—was performed on each docked compound-protein complex using the modified generalized Born model of Onufriev, Bashford, and Case with model 2 radii (igb = 5) (Onufriev et al., 2000) with a nonbonded cutoff of 25 Å. The MM/GBSA energy of the minimized protein-compound complex structure was calculated using an infinite cutoff (999 Å) and a protein dielectric constant of 4. The binding affinity was computed by MM/GBSA energy of the complex subtracted from the sum of the MM/GBSA energies of the apo protein and the isolated compound.

Molecular Dynamics Simulations of World-Approved 2018 Co-Complexes

Molecular dynamics (MD) simulations were performed for each of the world-approved 2018 complexes down-selected from the top 1% of docked compounds (see **Supplementary Table S1**). The best scoring single-point MM/GBSA co-complex structure was selected as a starting conformation for the MD simulations. The MD simulations were performed using the pmemd_cuda program in AMBER (Salomon-Ferrer et al., 2013b). The catalytic dyad (His41-Cys145) of the main protease was modeled as charged residues. Charges for the thiolate of Cys145 were obtained from AM1-BCC calculations (Jakalian et al., 2002). The General Amber Force Field (GAFF) was used to model the ligands (Wang et al., 2004). The ligand-protein complex was solvated into a truncated octahedron of TIP3P water (Jorgensen et al., 1983), 50 Na⁺ ions with a neutralizing number of Cl[−] ions were added to the solution. The system was energy minimized with 500 steps of steepest descents and 1,500 steps of conjugate gradients. Initial equilibration was performed with NVT dynamics at 300 K for 200 ps with positional constraints ($K = 1$ kcal/mole•Å²) on the CA atoms in residues. Electrostatic interactions were treated using Particle Mesh Ewald (PME) summation (Darden et al., 1993). The nonbonded interactions were cut off at 8 Å. Further equilibration was performed with NPT dynamics for 4.8 ns. The pressure was set at 1 atm using a Monte Carlo barostat (Salomon-Ferrer et al., 2013b). The positional constraints were reduced to 0.5 kcal/mole•Å². Production dynamics was performed for 200 ns without positional constraints. The MM/GBSA energies were calculated using MMPBSA.py (Miller et al., 2012) utilizing the Generalized Born model of Onufriev, Bashford, and Case (igb = 5) (Onufriev et al., 2000) on coordinates saved every 20 ps.

Machine Learning

To assist in determining promising compounds that may have missed the energy cutoff and complement MM/GBSA rescoring, we utilized our Structure-Based Deep Fusion Inference models. We will only briefly describe the Fusion methods, which is described in detail in a previous publication (Jones et al., 2021).

The Deep Fusion models are based on 3D convolutional neural network (3D-CNN) and spatial graph (SG-CNN) models trained independently on ligand-protein co-crystal structure data from PDBBind 2016 (Liu et al., 2017). Two types of fusion models are then built on top of the CNN layers. In the “Mid-Fusion” model, the intermediate CNN features extracted from each model are combined using a series of fully connected layers and then used to predict a ligand-protein binding score. Batch normalization and ReLU-based non-linearities are applied in each fully connected layer. In the “Late-Fusion” model, we combined the constituent CNN models’ predictions rather than their features to produce the final prediction. We used the two fusion models along with the two component CNN models to rank compounds for spike and M^{pro} inhibition.

We used the 3D configurations from the docking calculations in our pipeline as input for our structure-based deep learning methods. Since these models are trained using the protein binding pocket coupled with the ligand, it was necessary to develop a protocol to extract binding pockets from the SARS-CoV-2 proteins. We considered multiple volumes for the bounding box centered on the ligand centroid. We validated our choices by considering correlation (Pearson and Spearman) of the model predictions across bounding box size for all structure-based machine learning methods while additionally considering consensus with the MM/GBSA rescoring method via Pearson and Spearman correlation. Our results showed that given these metrics, the optimal bounding box configuration varied significantly and suggested that the optimal approach would be to combine results across all configurations.

Using these methods, we computed rankings of the SARS-CoV-2 protein inhibitors by scoring each compound for each target for each candidate bounding box. The predictions were then averaged across all bounding boxes to produce the final score for each protein-ligand combination. Then, for each of the models, the compounds were sorted according to predicted activity and ranked in descending order. The sum of the reciprocal rankings was then used to aggregate the rankings across all methods. The top five unique spike protein inhibitors along with the top 25 unique M^{pro} inhibitors were then chosen for experimental validation.

The pharmacokinetic and safety properties of the 26 million compounds used in this study were predicted with the ATOM Modeling PipeLine (AMPL) (Minnich et al., 2020), a data-driven pipeline for drug discovery, and the Maestro workflow manager (Di Natale, 2017). Chemical descriptors were computed with MOE and Mordred from 2D and 3D structures and graph (Ramsundar et al., 2019) and fingerprint representations. Fully connected neural networks, graph convolution, and random forest models were considered, and the best models selected using AMPL. A total of 30 models with 23 distinct targets

were used for property prediction and are summarized in **Supplementary Table S2**. Results for the 9 models trained on public data are available at <https://covid19drugscreen.llnl.gov>.

Spike RBD and ACE2-Fc Protein Production and Purification

The gene for the SARS-CoV-2 spike protein (NC_045512.2) was codon-optimized for expression in mammalian cells and subcloned into pcDNA3.4 with the native secretion signal and a C-terminal His₈ tag. The plasmid was transfected into Expi293 cells and cultured for 5 days according to the manufacturer (ThermoFisher Scientific). Cells were harvested by centrifugation and the spike-containing culture medium was sterile-filtered, pH adjusted to 7.4 using PBS, and captured on a HisTrap Excel (Cytiva) using the Akta Pure FPLC system. The column was washed with wash buffer (20 mM sodium phosphate, 300 mM sodium chloride, 40 mM imidazole, pH 7.4) and eluted with wash buffer containing 500 mM imidazole. Fractions containing spike RBD were pooled and concentrated using a 10 kDa MWCO centrifugal concentrator (ThermoFisher). The concentrated protein was loaded onto a Superdex 200 Increase 10/300 GL equilibrated with PBS, pH 7.4. Fractions containing spike RBD were pooled and concentrated as before.

The ACE2-Fc fusion construct was made by subcloning the ectodomain of the human ACE2 gene (Sino Biological) into the pCR3-Fc vector, which contains the CH2 and CH3 domains of human IgG1 as previously described (Negrete et al., 2006). The ACE2-Fc containing plasmid was transfected into ExpiCHO cells and cultured for 7 days according to the manufacturer (ThermoFisher Scientific). Cells were harvested by centrifugation and the ACE2-Fc-containing culture medium was sterile-filtered, pH adjusted to 7.4 using PBS, and captured on a MabSelect PrismaA column (Cytiva) using the Akta Pure FPLC system. The column was washed with wash buffer (50 mM sodium phosphate, 150 mM sodium chloride, pH 7.4) and eluted with 100 mM sodium citrate pH 3. Fractions containing ACE2-Fc were pooled and concentrated using a 10 MWCO centrifugal concentrator (ThermoFisher). The concentrated protein was loaded onto a Superdex 200 Increase 10/300 GL equilibrated with PBS, pH 7.4. Fractions containing ACE2-Fc were pooled and concentrated as before.

Biolayer Interferometry Competition Assay for Spike Protein binding Compound

The competitive binding assays were performed by biolayer interferometry using the Octet RED96 system (FortéBio). All experiments were performed using 96 well microplates (Greiner Bio-One) at 30°C with the shaking speed of 1,000 rpm and samples were diluted in kinetic buffer (PBS containing 0.02% Tween 20, 0.1% bovine serum albumin). Octet anti-human Fc (AHC) biosensors were pre-equilibrated in biosensor buffer [kinetic buffer (KB) containing 10 µg/ml biocytin] for 30 min before use in experiments. SARS-CoV-2 RBD was pretreated with candidate compounds for 30 min prior to assay start. Human ACE2-Fc protein was immobilized on the surface of the AHC

biosensor tip and followed by a baseline step of 120 s in KB. ACE2-captured biosensors were immersed in wells containing different concentrations (5–100 μ M) of small molecule and SARS-CoV-2 RBD for 180 s followed by dissociation step for 200 s. The raw data was analyzed using Octet Data Analysis High Throughput software (FortéBio). Binding sensorgrams were aligned at the beginning of the binding cycle, double reference subtracted and Savitzky Golay filtered data were globally fit to a 1:1 binding model. A total of 32 compounds (see **Supplementary Table S3**) were tested against the RBD. All compounds were purchased from TargetMol at 97% purity or higher and used without further purification.

M^{Pro} and FRET Substrate Protein Production and Purification

The gene for the SARS-Cov-2 M^{Pro} (from Genbank MN908947.3) was codon-optimized for expression in *E. coli* and subcloned into a pET-32 vector, with a N-terminal GST tag connected by an auto-cleavage sequence and a C-terminal His₆ tag. The plasmid was transformed into BL21 DE3 *E. coli* and streaked onto ampicillin plates. Individual colonies were picked and used to inoculate 50 ml starter cultures, which were grown in lysogeny broth (LB) containing ampicillin overnight at 37°C. The 50 ml starter cultures were then used to inoculate 1 L of LB, which was incubated at 37°C until OD = 0.6 to 0.9, at which point IPTG was added to a final concentration of 400 μ M and cells were incubated with gentle shaking at 16°C overnight. Cells were then pelleted, flash frozen in liquid nitrogen, and stored at –80°C. The pellet from 100 ml of culture was thawed, resuspended in 10 ml BugBuster master mix (Millipore Sigma), and gently inverted at 4°C for 1 h to lyse. The insoluble fraction of the lysate was then spun down and the supernatant was sterile-filtered prior to capture on a Ni NTA column. The lysate was diluted with Buffer A (20 mM Tris, 100 mM NaCl, 5 mM Bme, pH 8.0), and Ni NTA Buffer B (20 mM Tris, 100 mM NaCl, 5 mM Bme, 500 mM imidazole, pH 8.0) was added to a final concentration of 10 mM imidazole. The lysate was then loaded onto a 5 ml HisTrap Ni NTA column (GE Healthcare) using an FPLC system (Bio-Rad), and eluted with Ni NTA Buffer B. Fractions containing the eluted protein were pooled and spin-exchanged into Buffer A using 10 kDa MWCO Amicon Ultra centrifugal filters (Millipore Sigma). The C-terminal His₆ tag was then cleaved off by incubating the concentrated protein with 30 μ g of N-terminally His-tagged HRV-3C protease (Sigma-Aldrich) overnight at 4°C. The digested protein was applied again to the Ni NTA column, and the flowthrough was collected and used directly for ion exchange chromatography.

The flowthrough was loaded onto a 5 ml High Q anion exchange column (Bio-Rad) and proteins were eluted with a linear gradient of IEX Buffer B (20 mM Tris, 1 M NaCl, 5 mM Bme, pH 8.0). To our surprise, the M^{Pro} was found in the flowthrough rather than the eluted fractions. The flowthrough was collected, buffer exchanged into storage buffer (20 mM Tris, 150 mM NaCl, 1 mM TCEP, pH 7.8), flash-frozen, and stored at –80°C. Purity appeared to be >99% by SDS-PAGE and staining with SimplyBlue SafeStain (ThermoFisher).

The fluorescence resonance energy transfer (FRET)-based M^{Pro} substrate was cloned into pET bacterial expression vector starting from a pcDNA.31-Clover-mRuby2 plasmid with a cloned linker sequence FGAARAVLQSGFRAADP between the Clover and mRuby2 FRET protein pairs. The cloned linker sequence is a protease substrate and cleaves the peptide backbone between residues QS. pcDNA3.1-Clover-mRuby2 was a gift from Kurt Beam (Addgene plasmid # 49089; <http://addgene.org/49089>; RRID:Addgene_49089). The kanamycin-resistant pET plasmid was transformed into BL21(DE3) cells (NEB) and cultures were induced with IPTG (0.5 mM) at 15°C overnight with gentle shaking (150 RPM). The FRET substrate was subsequently purified by standard Ni-NTA affinity techniques, as described above.

M^{Pro} FRET-Based Activity Assay

M^{Pro} inhibitor screening and half maximal inhibitory concentration (IC₅₀) analysis were performed in 384 well assay plates, in 25 ml final volumes using 1875 ng of substrate and 375 ng of M^{Pro} diluted in assay buffer (0.0033% Triton-X100, 50 mM Tris-HCl, 150 mM NaCl, pH 7.4). All compounds were diluted with DMSO to volumes of 2.5 μ l to obtain a 10% final concentration of DMSO in the 25 μ l reaction. Percent cleavage of the FRET substrate was measured on a Tecan Spark[®]. Fluorescence emission at 620 nm was measured for each well using excitations at 560 nm (excite mRuby2, emit mRuby2), and 485 nm (FRET from Clover to mRuby2). The FRET signal was normalized to the fluorescence of mRuby2 for each well. All assays were run in technical replicates and averaged. This data was then normalized to the average of the -protease wells (16 replicates per plate). The data was then analyzed in GraphPad Prism 9, wherein the “Normalize” tool was used to set the %FRET values for the +protease control to 0 and the -protease controls to 1.0. Both protease controls utilized 16 replicates per plate. The Z-factor is calculated using the + and -protease control wells (Zhang et al., 1999). This sets the min/max signals for normalization. All wells had DMSO, as compounds were in DMSO. Complete reactions were run on SDS-PAGE gels to assess protein cleavage independently of FRET measurements. Gel densitometry analysis (analyzed using ImageJ) justified the 100 and 0% cleavage in the +protease and -protease controls, respectively, at the time points used for analysis. In each experiment, measurements were taken at several time points, however only end-point data (at which time the +protease control reactions have gone to completion) has been presented herein, at about 4 h post addition of protease. A total of 91 compounds (see **Supplementary Table S4**) were tested against M^{Pro}. All compounds were purchased from TargetMol at 97% purity or higher and used without further purification.

Viral Infection Assays

A pseudotyped, replication-competent vesicular stomatitis virus (VSV) expressing the SARS-CoV-2 spike gene (VSV-SARS2) in place of its own VSV-G gene was provided by Dr. Sean Whelan (Case et al., 2020) and used to screen compounds predicted to target the SARS-CoV-2 spike. VSV-SARS2 also expresses GFP allowing for rapid analysis of infection based on reporter

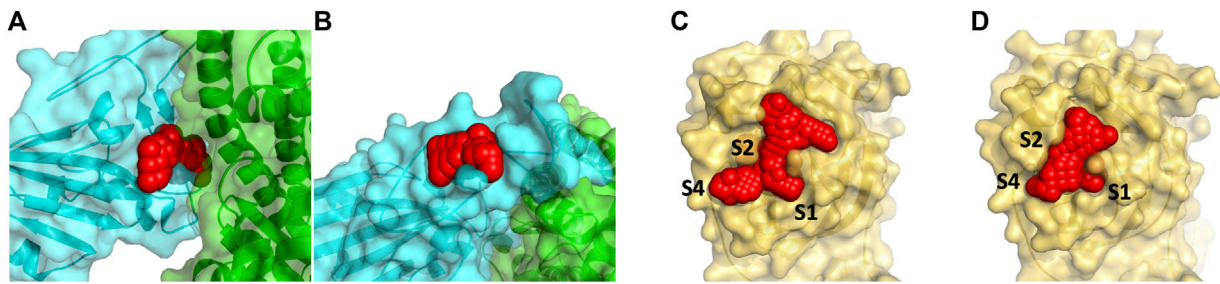


FIGURE 1 | Panel (A) shows the docking site on the RBD of the spike protein in red (by residues 501–505) that are at the interface with ACE2 (show in green) and denoted spike1 in the text. A smaller secondary binding site (denoted spike2) in the spike protein in receptor binding motif domain was detected and used for docking studies (B). Panel (C) and (D) show the binding site of the M^{pro} with the N3 inhibitor removed (6LU7) is protease1 and the apo protein (6Y84) is protease2. The S2 binding pocket is below the sidechains of Met49 and Gln189 and is not visible in the picture.

expression under BSL-2 containment. Initial drug screening was performed by incubating the compounds at 10 μ M with VSV-SARS2 or VSV-GFP (VSV) as a specificity control for 30 min prior to their addition to Vero cells seeded in a 96-well plate. Infection was performed for 1 h with a multiplicity of infection (MOI) of 0.5 for VSV-SARS2 and 0.1 for VSV. The infection media was subsequently removed, replaced with fresh media and fluorescent protein measurements were collected 18–24 h post-infection. Down-selected compounds were subjected to IC₅₀ analysis using dilutions of drug starting at 100 mM concentrations following a similar infection protocol against VSV-SARS2 under BSL-2 containment or recombinant SARS-CoV-2 expressing the mNeon reporter gene (provided by Dr. Pei Yong Shi) (Xie et al., 2020) under BSL-3 containment. The compounds were screened starting at 100 μ M using an 8-point, 1:2 dilution series with infections being performed at a MOI of 0.2. In addition, Presto-Blue cytotoxicity assays were performed using a similar dilution series in uninfected cells to determine relative cell viability to drug-only treatments. Fluorescent values were background subtracted using no-infection controls and normalized to no-treatment infection values. IC₅₀ curves and values generated using GraphPad Prism 9.

RESULTS AND DISCUSSIONS

Computational Predictions

Molecular dynamics simulations were performed on both the RBD of the spike protein and the M^{pro} to identify alternative conformations from the crystal structure (PDB, 6M0J) (Lan et al., 2020). For the RBD structure, a total of twelve 100 ns simulations were performed (aggregate 1.2 μ s of dynamics). The structures from the last 20 ns from each simulation was collected and clustered. There were only slight changes in the conformation of residues that would form interactions with ACE2. The most variable region within RBD was located at the opposite end of the protein relative to the ACE2 binding sites. The stability of the ACE2 binding regions likely is not surprising given the high binding constant of RBD to ACE2 and relatively small contact region (Lan et al., 2020). The dynamics of the M^{pro} dimer shows the residues near the active site are stable except for the loop

formed by residues Cys44-Pro52 (Bzowka et al., 2020). This loop shifts position in both monomers and moves the associated residues further from the active site.

We identified two binding sites within the RBD of the spike protein and within the M^{pro} proteins binding sites as shown in Figure 1. In the RBD, two sites were chosen that are proximal to critical residues that bind human ACE2. Both sites in the RBD are formed by stable loop areas. The first site is in the proximity of a beta-turn formed by residues 501–505 and denoted spike1 below. This region forms several interactions with ACE2 and the corresponding residues in the SARS-CoV-2 spike protein form the major recognition site for neutralizing antibodies. We used the crystal structure (PDB, 6M0J) for docking to this site since the protein conformations sampled from MD simulations did not significantly differ from the crystal structure. The second site is stabilized by a disulfide (Cys480-Cys488) that connects the loop at the end of the receptor-binding motif (RBM) and denoted spike2 below. These two regions include the two key mutations of the variants of concern—E484K and N501Y (Voloach et al., 2020; Fiorentini et al., 2021). During the MD simulations, it was observed that the sidechains of residues Lys458 and Glu471 become solvent-exposed. In the crystal structure, these two residues are in close proximity and divide a potential binding site into two small sites. In the MD structure, these residues are solvent-exposed and a single larger binding site is present (Figure 1B). We used the MD structure for docking to this site. We limited our drug discovery efforts on the spike protein to two sites in the proximity of the RBD-ACE2 interface where the small molecule would directly interfere with formation of the protein complex. There are likely other drug binding sites within the spike protein that can affect ACE2 binding (Olotu et al., 2020; Verkhiver, 2020) but determining their locations experimentally is non-trivial.

The main protease is a cysteine protease with a catalytic dyad consisting of Cys145-His41. To accommodate its natural polypeptide substrates, a large gorge is present on the surface of the enzyme. The covalent inhibitor N3 is based on the tripeptide Ala-Val-Leu and reacts with the thiolate of Cys145. Two crystal structures of M^{pro} have been solved recently. The 6LU7 crystal structure was solved with the covalent inhibitor N3 in the active site (denoted protease1 in the text) (Jin et al., 2020).

A second structure 6Y84 (denoted protease2 in the text) was solved as an apo protein in a different space group relative to 6LU7 (Owen et al., 2020). This crystal structure's active site differs from 6LU7 with N3 removed. The sidechains of Met49 and Met165 change positions depending upon having N3 present. The shifts in positions of these methionine residues enlarge the active site. In the MD simulations, Met49 shifts position away from the active site to also enlarge this region. We chose to use the crystal structure of 6Y84 as another site for docking since the changes relative to 6LU7 are small but the positional change in Met49 changes/enlarges the active site. In **Figures 1C,D** we show the two conformations of the active site, one from each of these crystal structures of M^{Pro}, were used for our docking study.

We docked over 26 million compounds to these four sites (two spike sites and two M^{Pro} structures/conformations) to find possible binders that could either interfere with protein binding (RBD of spike protein) or inhibit substrate binding (M^{Pro}). Although all the compounds docked to these four sites are supposed to be commercially available or can be synthesized, to expedite experimental testing we will focus our discussion on the world-approved 2018 set. The computational results on the other 26 million compounds docked to the four sites are freely available online at <https://covid19drugscreen.llnl.gov>. The docking score energy cutoff of -7.5 kcal/mole reduced the number of compounds to 136 in the spike1 site and 50 in the spike2 site in the RBD of the spike protein. The larger binding site of the main protease had a greater number of ligands for further testing, 916 for the protease1 site of the main protease2 site. All these compounds were interrogated for activity using our ML Fusion model and MM/GBSA single point calculations to identify the most promising compounds. Each compound bound to its respective site was ranked from highest to lowest by energies for Vina docking score, MM/GBSA energy, and Fusion model. The final ranking of the compounds in their respective sites were inverted (i.e., $1/\text{rank}$) and weighted by $1.2 \bullet (\text{MM/GBSA}) + 1.0 \bullet (\text{Fusion model}) + 0.8 \bullet (\text{Vina docking})$. We believe the physics-based MM/GBSA to be our most accurate method and molecular docking the least predictive method relative to experiment. Because of the modest number of compounds remaining after the energy cutoff, molecular dynamics simulations were performed on all the ligand-protein complexes to obtain an average MM/GBSA energy and to investigate whether the protein dynamics were altered by formation of the complex.

Disruption of RBD binding to ACE2 would prevent infection by SARS-CoV-2. Docking to the spike1 site on the RBD puts the ligand in direct conflict with ACE2 binding when the protein complex is formed. A relatively small number of compounds were able to make the MM/GBSA rescoring energy cutoff for further molecular dynamics simulations since this binding site is relatively shallow. 134 compounds were simulated in the spike1 site using their five lowest-energy docking poses and their average MM/GBSA was determined from the ligand-protein conformations from the MD trajectory. The root mean squared deviation (RMSD) of the protein backbone from the crystal structure was used as an additional criterion to determine the stability of the ligand-protein complex. To successfully

disrupt formation of the protein complex, the compounds must have a low MM/GBSA binding energy and be stable within the binding site. Twenty-eight compounds had an average MM/GBSA below -30 kcal/mole and an RMSD 4 \AA or less (recentering and was only performed on the protein) for at least one of their simulations. Some compounds on this list that are of additional interest additional interest are accolate, tasosartan, and olmesartan medoxmil. Accolate is used to control and prevent symptoms of asthma such as wheezing and shortness of breath. Tasosartan is an angiotensin II receptor agonist. Olesartan medoxomil is an angiotensin II receptor blocker. Several studies have pointed to improved outcomes when COVID19 patients have used angiotensin II receptor blockers/inhibitors (Meng et al., 2020; Sanchis-Gomar et al., 2020; Zhang et al., 2020).

The spike2 binding site is located in the receptor binding motif (RBM) of the RBD. This binding site does not directly interfere with formation of a protein-protein complex, however it is in close proximity with a group of aromatic residues (Phe456, Tyr473, and Tyr489) that form interactions with ACE2. We speculated that having a bound compound proximal to these residues might disrupt the positioning of these aromatic residues and affect ACE2 binding. From an initial 134 compounds, only 50 compounds had a MM/GBSA below -30 kcal/mole and an RMSD less than 4 \AA during at least one of the simulations. Interestingly, several of the best binding compounds are diuretics or metabolites (glucuronides). The considerable number of polar and charged residues in the vicinity makes this a favorable environment for the glucuronic acid.

In docking calculations of the main protease, two different crystal structures were utilized for docking because the sidechain positions of Met49 and Met165 in the active site vary due to one structure had the ligand N3 (6LU7) present while the other was empty (6Y84). Although the shape of the active site differs, there were 535 compounds that were common to both structures out of the more than 900 compounds that made the initial -7.5 kcal/mole single point energy cutoff for each protein structure. Since there is no indication which structure is preferred, the compounds were ranked by the sum of their average MM/GBSA energies. Several of the top-scoring compounds that bind to both active site conformations are described here. Cefoperazone is a semi-synthetic beta-lactam antibiotic. Irinotecan is a plant alkaloid that acts as a topoisomerase inhibitor used to treat colon and small-cell lung cancers. Its relatively rigid structure allows it to span the length of the active site. Rutin is a citrus flavonoid consisting of quercetin and the disaccharide rutinose and used as an alternative medicine. Several compounds are protease inhibitors or metabolites of drugs. AFN911 is a metabolite of imatinib (benzylic hydroxylation). Losartan n2-glucuronide is the metabolite of losartan (an angiotensin II receptor antagonist). Saquinavir is an antiretroviral drug (protease inhibitor) used to treat HIV/AIDS. Teniposide is a topoisomerase II inhibitor used for treatment of several childhood cancers. Cabozantinib is a tyrosine kinase inhibitor that is used as mediation for medullary thyroid

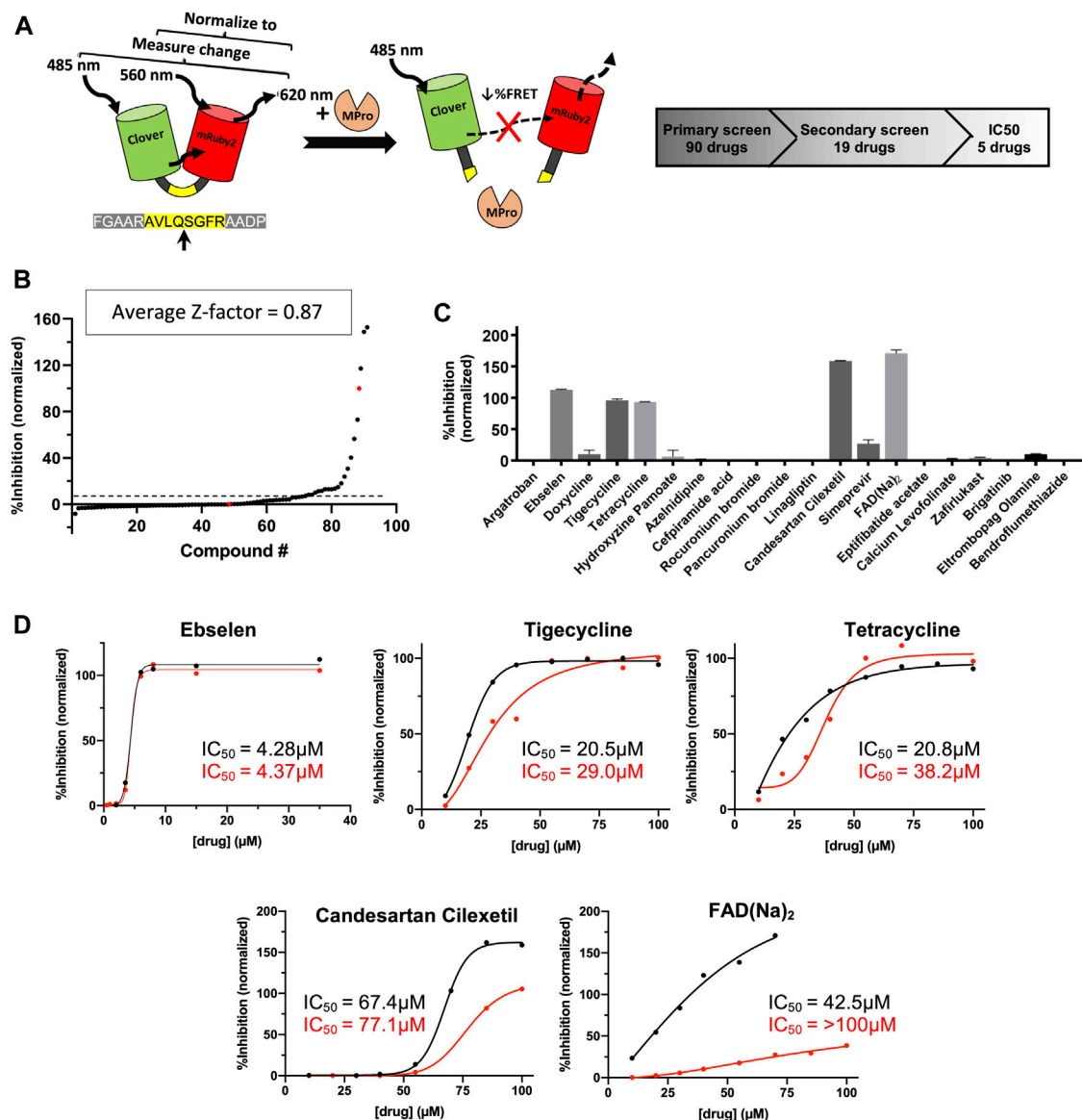


FIGURE 2 | Predicted M^{pro} drug inhibitors screened using a FRET-based protease assay with five down-selected hits. **(A)** A schematic of the FRET-based SARS-CoV-2 main protease assay is shown along with the hit identification overview. **(B)** Purified M^{pro} and FRET substrate proteins were incubated in the presence of 100 μ M of drugs from a library of computationally predicted M^{pro} inhibitors. No drug, no protease, and Ebselen were used as controls to calculate the Z-factor for each plate and an average score is displayed above. Red dots indicated no drug (0% inhibition) or no protease (100% inhibition) conditions, while the black dots are the ordered percent inhibition values. **(C)** Identified hits from the primary screen were re-screened at 100 μ M and the FRET values were normalized as percent inhibition values in the bar graph. Experiments were performed in duplicate and the presented results are the average values. **(D)** Verified compounds from rescreening were subjected to half-maximal inhibitory concentration (IC_{50}) analysis. Presented values are averaged from technical duplicate experiments. Black lines and values represent normalized data from FRET values while the red lines and values represent normalized data from gel electrophoresis (Supplementary Figure S1) and densitometry.

cancer and renal cell carcinoma. Intriguingly, the angiotensin II receptor blocker olmesartan medoxomil was also predicted to bind well to the spike protein. The compounds rutin, losartan, imatinib, saquinavir, and teniposide have been seen in other computational screens (Bello et al., 2020; Huynh et al., 2020; Pant et al., 2020; Nejat and Sadt, 2021). Losartan and imatinib have undergone clinical trials with COVID19 patients (Aman et al., 2021; Puskarich et al., 2021). Most of the metabolites

found in the computational screens unfortunately were not available for purchase.

Experimental Validation

Experimental testing of the predicted binders for Mpro was performed by utilizing a fluorescence resonance energy transfer (FRET) based activity assay (Figure 2A). This FRET assay consisted of a substrate composed of two fluorescent

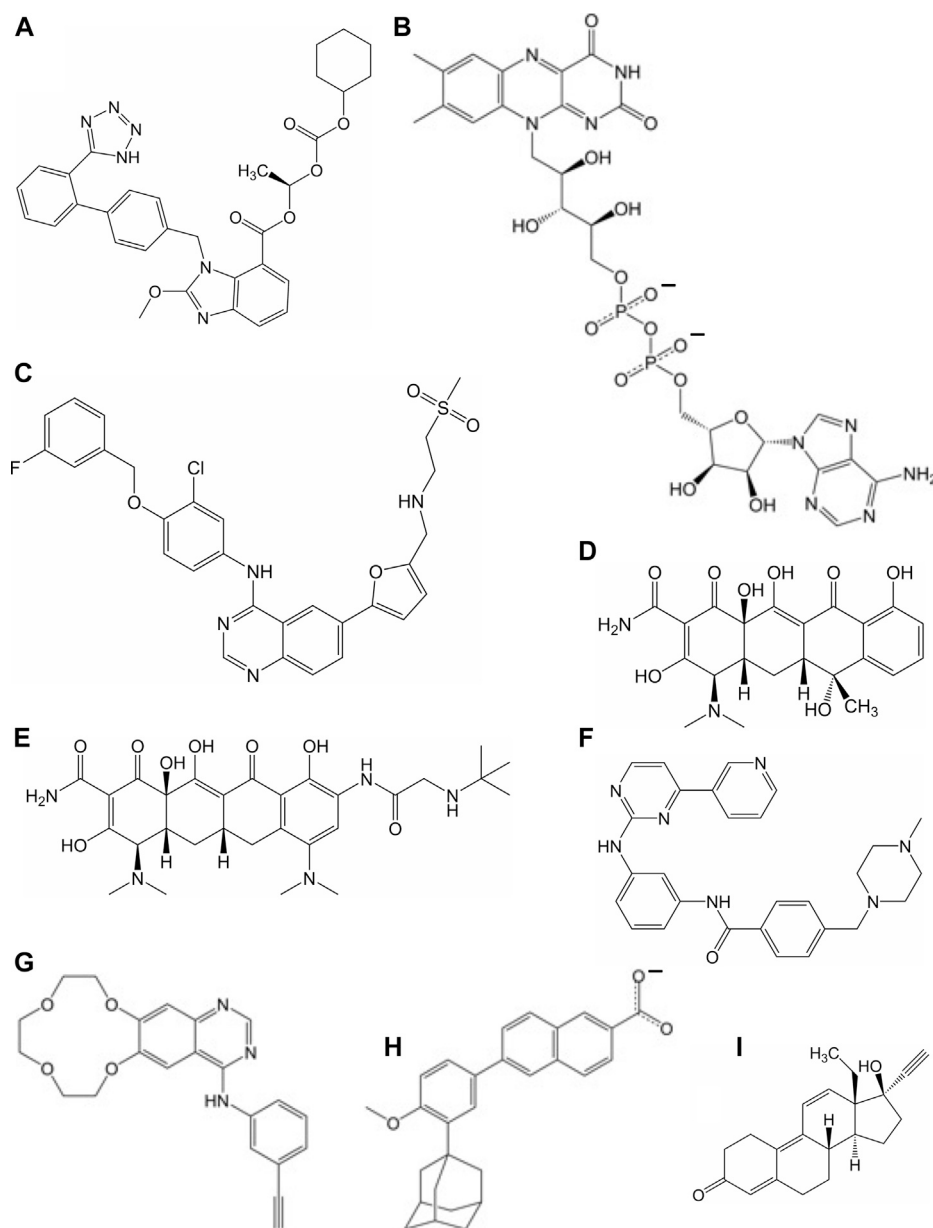


FIGURE 3 | Molecular of structures for compounds that have a repressive effect on some aspect of the virus activity: **(A)** candesartan cilexetil, **(B)** flavin adenosine dinucleotide, **(C)** lapatinib, **(D)** tetracycline, **(E)** tigecycline, **(F)** imatinib, **(G)** icotinib, **(H)** adapalene, and **(I)** gestrione.

proteins, Clover and mRuby2, linked through a Mpro recognition sequence. **Supplementary Figure S1** shows the advantage of a protein-based substrate over standard peptide-based methods was to allow for verification of FRET values by independent, FRET-independent gel electrophoresis. The assay was optimized using a positive control compound called Ebsele, a low micromolar Mpro inhibitor (Jin et al., 2020). **Supplementary Table S4** shows the results from our initial screen, from which, 19 compounds were down-selected and tested in a secondary screen where four compounds were found to completely inhibit the activity of Mpro at 100 μ M concentrations and are shown in

Figure 2C. These identified compounds included candesartan cilexetil, FAD, tigecycline and tetracycline (see **Figure 3**). Candesartan cilexetil is an angiotensin II receptor antagonist prodrug. Flavin adenosine dinucleotide is a redox-active coenzyme. Tigecycline is a glycylcycline antibiotic and closely related to tetracycline. These were the only two compounds that bind Mpro and had a similar molecular structure. In **Figure 2D** we show that these four compounds were relatively weak inhibitors of Mpro compared to Ebsele as the IC₅₀ values were calculated to be 67.4 μ M for candesartan cilexetil, 42.5 μ M for FAD disodium, 21.5 μ M for tigecycline, and 20.8 μ M for tetracycline. The IC₅₀

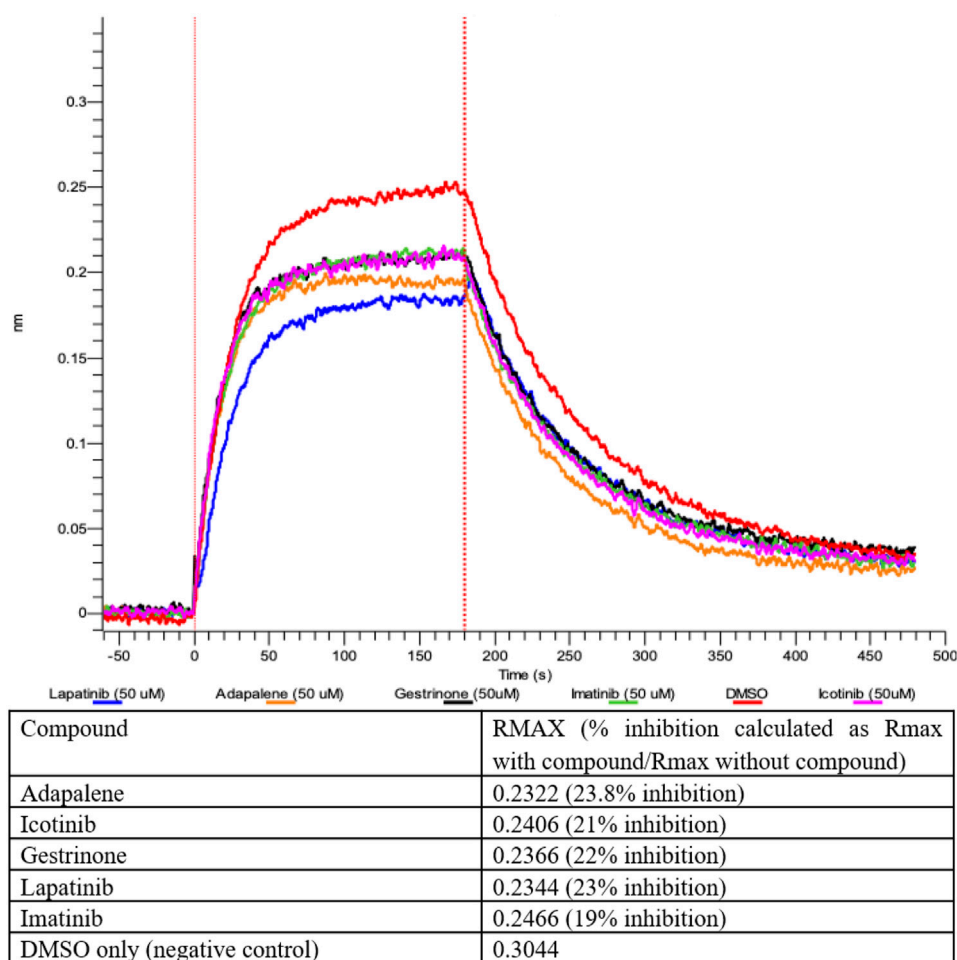


FIGURE 4 | Inhibition of ACE2-RBD binding after pre-treatment with 50 μ M compound measured by Biolayer interferometry.

values were comparable to gel electrophoresis-based analysis of the cleaved substrate products with the exception of FAD as shown in **Figure 2D** and **Supplementary Figure S1**. Importantly, candesartan cilexetil has been previously identified as a Mpro inhibitor with a IC_{50} of 2.8 μ M (Li et al., 2020) although the fluorogenic substrate used was slightly shorter than the substrate utilized in this study. Additionally, candesartan cilexetil has inherent fluorescent properties that make determining its cleavage inhibition difficult.

The compounds computationally predicted to target the SARS-CoV-2 spike protein RBD were screened by pseudotyped virus assay and biolayer interferometry competitive assay (BLI). Compounds were tested for their ability to inhibit ACE2-spike binding via BLI competitive assays on Octet RED96 platform (Forte Bio). In this assay, human ACE2-Fc was immobilized on AHC biosensors and binding to soluble SARS-CoV-2 RBD was detected. The RBD was pre-treated with candidate compounds at increasing concentrations prior to assay. In **Figure 4**, we show an inhibitor concentration-dependent decrease in ACE2-RBD binding in samples pretreated with adapalene, imatinib,

lapatinib, gestrinone, and icotinib. Adapalene is a topical retinoid used to treat acne. Icotinib and lapatinib are inhibitors of the tyrosine kinase EGFR. Imatinib is used to treat chronic myelogenous leukemia (CML). Gestrinone is a synthetic steroid used to treat endometriosis. An imatinib metabolite (AFN911) has previously been identified in this study as also a possible Mpro inhibitor.

In parallel, the computationally-predicted spike binding compounds were screened using a cell-based infection assay. The spike compound library set was screened against a BSL-2 surrogate virus encoding the SARS-CoV-2 spike protein that mimics ACE2-dependent SARS-CoV-2 fusion and cell entry (Case et al., 2020). The replication-competent pseudotyped virus, termed VSV-SARS2 (see Methods), expresses a GFP reporter upon cell infection and replication that was used as an indicator of infection in the drug screen. From the initial library set of 32 compounds, only imatinib and lapatinib were found to inhibit VSV-SARS2 at 10 μ M at ~50% or greater efficacy as shown in **Figure 5A** and **Supplementary Table S3**. To check for specificity, the compounds were screened against VSV and none were found to have a significant impact on infection thus

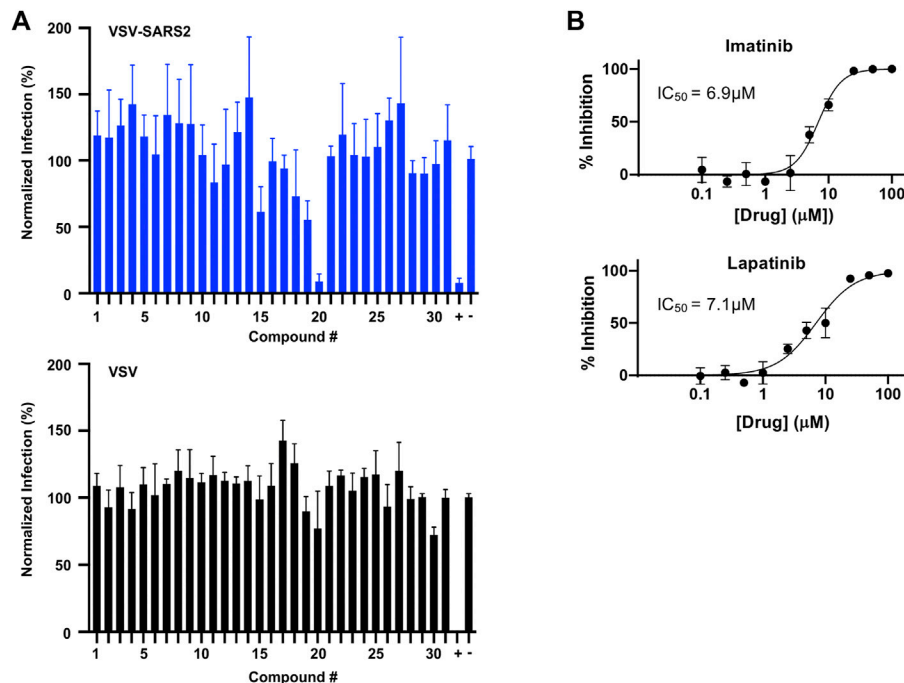


FIGURE 5 | Predicted spike drug inhibitors screened using a VSV-SARS2 infection assay reveals two promising hits. **(A)** Individual drugs from the library set were used at 10 μM to treat GFP reporter viruses, VSV-SARS2 and VSV, for 30 min prior to infection of Vero cells at 0.5 MOI or 0.1 MOI respectively. The infection media was replaced with fresh media at 1 h post-infection and fluorescent reporter values were measured the next day. **(B)** Half-maximal inhibitory concentration (IC_{50}) curves and values were obtained for Imatinib (compound 20) and Lapatinib (compound 19) using the same VSV-SARS infection assay performed for library screening. All data were normalized as percent infection or inhibition for drug-treated conditions vs. no-treatment control. The values are means, with error bars displaying standard deviation between the triplicate wells.

indicating the two hits were spike-dependent. The IC_{50} values of imatinib and lapatinib were 6.9 and 7.1 μM against VSV-SARS2, respectively (**Figure 5B**).

Finally, to further validate the anti-viral effects of identified Mpro and spike hits, the compounds were evaluated under BSL-3 containment using a SARS-CoV-2 reporter virus expressing mNeon (Xie et al., 2020). **Figure 6** shows the plotted IC_{50} and half-maximal cytotoxicity concentration 50 (CC_{50}) graphs for four compounds where virus inhibition was not simply due to the cytotoxicity induced by the drug alone. Imatinib, adapalene and candesartan cilexetil had IC_{50} values of approximately 10 μM against SARS-CoV-2 in a cell-based assay, while lapatinib had an IC_{50} value of 31.1 μM . The best scoring conformation of these four compounds with their target protein is shown in **Figure 7**. Comparatively, candesartan cilexetil had the highest selectivity index of all four compounds as its CC_{50} value was the only one greater than the limit of the assay ($>100 \mu\text{M}$, **Figure 6**). Similar results for candesartan cilexetil were obtained against Vero-E6 cells (Alnajjar et al., 2020). Interestingly, candesartan cilexetil is only effective as the prodrug. Candesartan cilexetil is rapidly ester hydrolyzed in the gastrointestinal tract into the angiotensin II receptor antagonist candesartan. Candesartan was tested in the FRET-based activity assay and found to have no effect. The active agent against the virus is either the intact prodrug or just the cyclohexyl-1-hydroxyethyl carbonate is required. Additionally,

to our knowledge, this is the first time the retinoid adapalene has been shown to be effective against SARS-CoV-2.

CONCLUSION

The COVID19 pandemic is the worst in the last century and has highlighted the critical need for a rapid response for identifying inhibitors to combat biological outbreaks before they become unmanageable. Leveraging high performance computing, we combined molecular simulations and machine learning to identify compounds that could possibly bind to the selected protein targets. Yet, computational identification of possible compounds is only the first step to finding an inhibitor. The viability of these selected compounds to inhibit protein function is critical and must be tested *in vitro* and *in vivo*. Through experimental binding assay studies between the identified compounds and the selected proteins and virus assays, four compounds (candesartan cilexetil, imatinib, lapatinib, and adapalene) have been shown to inhibit SARS-CoV-2 virus *in vitro*. Interestingly, compounds predicted to bind to the spike protein affected the virus more strongly than the predicted Mpro inhibitors even though the binding site of Mpro is deeper and better defined than the spike binding site. Imatinib, adapalene and candesartan cilexetil had IC_{50} values of approximately 10 μM against SARS-CoV-2 in an *in vitro* cellular

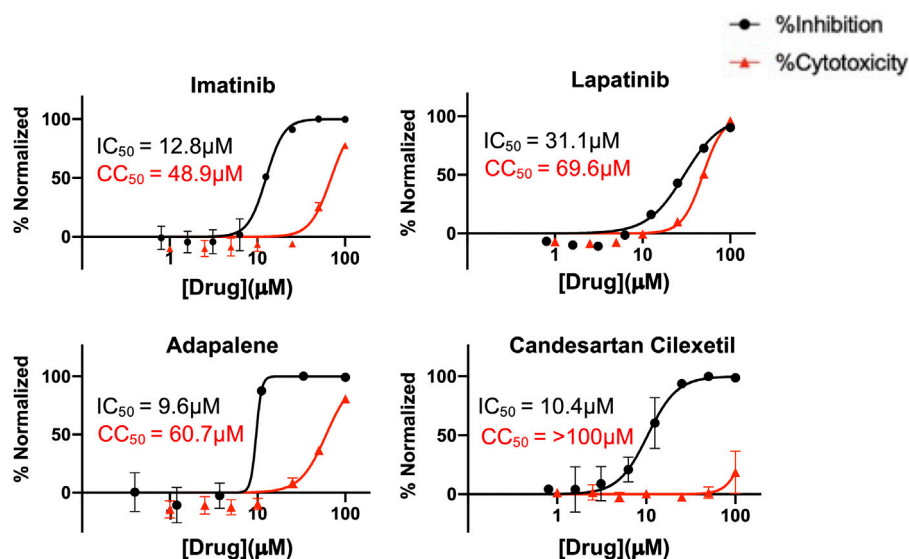


FIGURE 6 | Percentage inhibition and percentage cytotoxicity graphs from SARS-CoV-2 infection studies that show large therapeutic indexes in three hits. Varying concentrations of imatinib, lapatinib, and adapalene were used to treat virus for 30 min prior to infection in Vero cells, while candesartan cilexetil was added directly to cells without pre-treatment to virus. Infections were performed using SARS CoV-2mNeon at an MOI of 0.2. At 1 h post-infection, the media was removed and replaced with fresh media. Fluorescent reporter values were recorded 18 h post-infection. Similarly, Vero cells were treated with varying concentrations of indicated drugs, incubated for 18 h prior to analysis by Presto-Blue assays to assess cytopathic effect. Data were normalized to percent inhibition or percent cytotoxicity for drug-treated cells vs. no-treatment control. The values are means, with error bars displaying standard deviation between the triplicate wells. Half-maximal inhibitory concentration (IC_{50}) curves and values are represented in black while half-maximal cytotoxicity concentration 50 (CC_{50}) curves and values are represented in red.

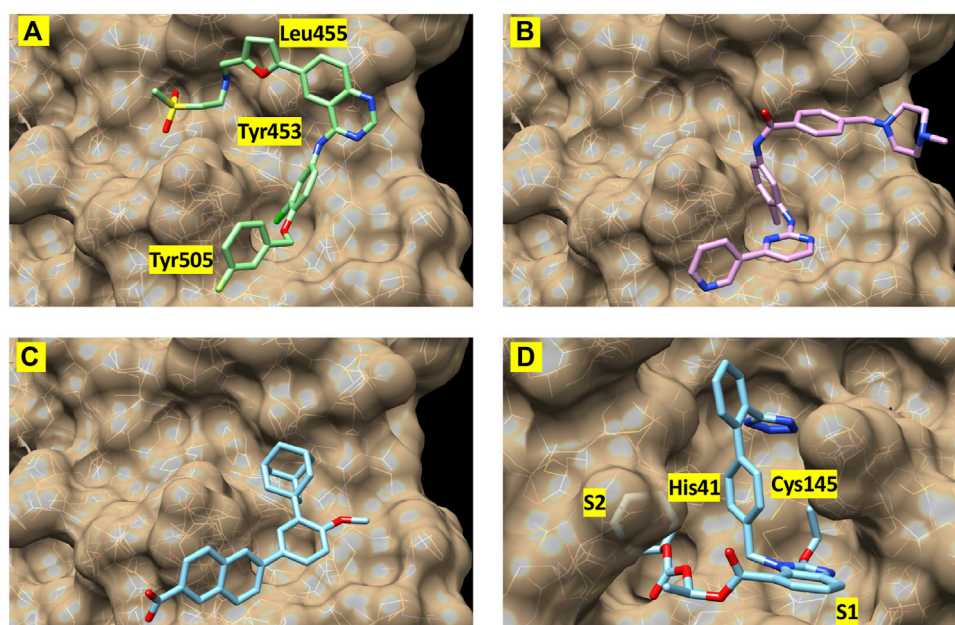


FIGURE 7 | Best-scoring pose from docking for (A) lapatinib, (B) imatinib, and (C) adapalene to the receptor binding domain of the spike protein (spike1 site). Panel (D) shows the best-scoring dock pose for candesartan cilexetil to M^{pro} . Labels identify protein residues neighboring the docked compounds.

infection assay, but the prodrug candesartan cilexetil shows the most promise as its selectivity index is greater than the limit of the assay.

DATA AVAILABILITY STATEMENT

Data and results from this study are publicly available at <https://covid19drugscreen.llnl.gov>.

AUTHOR CONTRIBUTIONS

FL and JA conceptualized the research work and reviewed and edited the manuscript. EYL, OAN, BJB, FB, MF, BH, DJ, HK, DK, KM, EAS, BS, GAS, MWT, DRW, SW, YY, AZ, XZ, WFDB, MB, AE, SH, VL, JL, RM, DKM, MAS, FZ collected and analyzed the data and wrote the manuscript.

FUNDING

Part of this research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. A portion of this work was funded by Lawrence Livermore National Laboratory through the Laboratory Directed Research and Development projects 20-ERD-065 and 20-ERD-062. Part of this research was also supported by the American Heart Association under CRADA TC02274 and the National Nuclear Security Administration through the Accelerating Therapeutics

for Opportunities in Medicine (ATOM) Consortium under CRADA TC02349.9.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Professor Xinquan Wang (Tsinghua University) for providing early access to his crystal structure of the ACE2-RBD complex. We thank Dr. Pei Yong Shi and the World Reference Center for Emerging Viruses and Arboviruses (WRCEVA) at UTMB for providing the SARS-CoV-2 mNG virus. The authors thank Livermore Computing for providing extensive computer time. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA00 03525. All work performed at Lawrence Livermore National Laboratory is performed under the auspices of the U.S. Department of Energy under Contract DE-AC52-07NA27344. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. LLNL-JRNL-819778.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.678701/full#supplementary-material>

REFERENCES

- Alnajjar, R., Mostafa, A., Kandeil, A., and Al-Karmalawy, A. A. (2020). Molecular Docking, Molecular Dynamics, and *In Vitro* Studies Reveal the Potential of Angiotensin II Receptor Blockers to Inhibit the COVID-19 Main Protease. *Heliyon* 6, e05641. doi:10.1016/j.heliyon.2020.e05641
- Aman, J., Duijvelaar, E., Botros, L., Kianzad, A., Schippers, J. R., Smele, P. J., et al. (2021). Imatinib in Patients With Severe COVID-19: A Randomised, Double-Blind, Placebo-Controlled, Clinical Trial. *Lancet Respir. Med.* doi:10.1016/S2213-2600(21)00237-X
- Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., et al. (2020). Remdesivir for the Treatment of Covid-19 - Final Report. *N. Engl. J. Med.* 383, 1813–1826. doi:10.1056/nejmoa2007764
- Bello, M., Martinez-Munoz, A., and Balbuena-Rebolledo, I. (2020). Identification of Saquinavir as a Potent Inhibitor of Dimeric SARS-CoV2 Main Protease through MM/GBSA. *J. Mol. Model.* 26, 340. doi:10.1007/s00894-020-04600-4
- Belouzard, S., Millet, J. K., Licitra, B. N., and Whittaker, G. R. (2012). Mechanisms of Coronavirus Cell Entry Mediated by the Viral Spike Protein. *Viruses* 4, 1011–1033. doi:10.3390/v4061011
- Bzówka, M., Mitusińska, K., Raczynska, A., Samol, A., Tuszyński, J. A., and Góra, A. (2020). Structural and Evolutionary Analysis Indicate that the SARS-CoV-2 Mpro is a Challenging Target for Small-Molecule Inhibitor Design. *Int. J. Mol. Sci.* 21, 3099. doi:10.3390/ijms21093099
- Case, J. B., Rothlauf, P. W., Chen, R. E., Liu, Z., Zhao, H., Kim, A. S., et al. (2020). Neutralizing Antibody and Soluble ACE2 Inhibition of a Replication-Competent VSV-SARS-CoV-2 and a Clinical Isolate of SARS-CoV-2. *Cell Host Microbe* 28, 475–485. doi:10.1016/j.chom.2020.06.021
- Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald: AnN-Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397
- Di Natale, F. (2017). Maestro Workflow Conductor. Available at: <https://github.com/LLNL/maestro> (Accessed March 1, 2020).
- Eastman, P., Swalis, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., et al. (2017). OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *Plos Comput. Biol.* 13, 1005659. doi:10.1371/journal.pcbi.1005659
- eMolecules (2020). *eMolecules*. San Diego, CA. 3430 Carmel Mountain Road, Suite 250. Available at: <https://www.emolecules.com/info/products-data-downloads.html> (Accessed March 1, 2020).
- Enamine (2020). *Enamine1 Distribution Way*. NJ: Monmouth Jct. Available at: <https://enamine.net/compound-collections/real-compounds> (Accessed March 1, 2020).
- Fiorentini, S., Messali, S., Zani, A., Caccuri, F., Giocanetti, M., Ciccozzi, M., et al. (2021). First Detection of SARS-CoV-2 Spike from N501 Mutation in Italy in August 2020. *Lancet Infect. Dis.* 21, e147. doi:10.1016/S1473-3099(21)00007-4
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi:10.1093/nar/gkr777
- Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *The Lancet* 395, 497–506. doi:10.1016/S0140-6736(20)30183-5
- Huynh, T., Wang, H., and Luan, B. (2020). Structure-based lead Optimization of Herbal Medicine Rutin for Inhibiting SARS-CoV-2's Main Protease. *Phys. Chem. Chem. Phys.* 22, 25335–25343. doi:10.1039/d0cp03867a

- Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* 23, 1623–1641. doi:10.1002/jcc.10128
- Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., et al. (2020). Structure of Mpro from SARS-CoV-2 and Discovery of its Inhibitors. *Nature* 582, 289–293. doi:10.1038/s41586-020-2223-y
- Jones, D., Kim, H., Zhang, X., Zemla, A., Stevenson, G., Bennett, W. F. D., et al. (2021). Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *J. Chem. Inf. Model.* 61, 1583–1592. doi:10.1021/acs.jcim.0c01306
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869
- Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., et al. (2020). Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor. *Nature* 581, 215–220. doi:10.1038/s41586-020-2180-5
- Letko, M., Marzi, A., and Munster, V. (2020). Functional Assessment of Cell Entry and Receptor Usage for SARS-CoV-2 and Other Lineage B Betacoronaviruses. *Nat. Microbiol.* 5, 562–569. doi:10.1038/s41564-020-0688-y
- Li, Z., Li, X., Huang, Y.-Y., Wu, Y., Liu, R., Zhou, L., et al. (2020). Identify Potent SARS-CoV-2 Main Protease Inhibitors via Accelerated Free Energy Perturbation-Based Virtual Screening of Existing Drugs. *Proc. Natl. Acad. Sci. USA*. 117, 27381–27387. doi:10.1073/pnas.2010470117
- Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., et al. (2017). Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Acc. Chem. Res.* 50, 302–309. doi:10.1021/acs.accounts.6b00491
- 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, 3696–3713. doi:10.1021/acs.jctc.5b00255
- Meng, J., Xiao, G., Zhang, J., He, X., Ou, M., Bi, J., et al. (2020). Renin-angiotensin System Inhibitors Improve the Clinical Outcomes of COVID-19 Patients with Hypertension. *Emerging Microbes & Infections* 9, 757–760. doi:10.1080/22221751.2020.1746200
- Miller, B. R., III, McGee, T. D., Jr., Swails, J. M., Homeyer, N., Gohlke, H., and Roitberg, A. E. (2012). MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theor. Comput.* 8, 3314–3321. doi:10.1021/ct300418h
- Minnich, A. J., McLoughlin, K., Tse, M., Deng, J., Weber, A., Murad, N., et al. (2020). AMPL: A Data-Driven Modeling Pipeline for Drug Discovery. *J. Chem. Inf. Model.* 60, 1955–1968. doi:10.1021/acs.jcim.9b01053
- Molecular Operating Environment (MOE) (2020). *Chemical Computing Group ULC*. Montreal, QC, Canada.
- 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
- Negrete, O. A., Wolf, M. C., Aguilar, H. C., Enterlein, S., Wang, W., Mühlberger, E., et al. (2006). Two Key Residues in EphrinB3 Are Critical for its Use as an Alternative Receptor for Nipah Virus. *Plos Pathog.* 2, e7. doi:10.1371/journal.ppat.0020007
- Nejat, R., and Sadt, A. S. (2021). Are Losartan and Imatinib Effective against SARS-CoV2 Pathogenesis? A Pathophysiologic-Based In Silico Study. *Silico Pharmacol.* 9, 1. doi:10.1007/s40203-020-00058-7
- Olotu, F. A., Omolabi, K. F., and Soliman, M. E. S. (2020). Leaving No Stone Unturned: Allosteric Targeting of SARS-CoV-2 Spike Protein at Putative Druggable Sites Disrupts Human Angiotensin-Converting Enzyme Interactions at the Receptor Binding Domain. *Inform. Med. Unlocked* 21, 100451. doi:10.1016/j.imu.2020.100451
- Onufriev, A., Bashford, D., and Case, D. A. (2000). Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* 104, 3712–3720. doi:10.1021/jp994072s
- Owen, C. D., Lukacik, P., Strain-Damerell, C. M., Douangamath, A., Powell, A. J., Fearon, D., et al. (2020). SARS-CoV-2 Main Protease with Unliganded Active Site (2019-nCoV, Coronavirus Disease 2019, COVID-19). doi:10.2210/pdb6Y84/pdb (Accessed March 7, 2020).
- Pant, S., Singh, M., Ravichandiran, V., Murty, U. S. N., and Srivastava, H. K. (2020). Peptide-like and Small-Molecule Inhibitors against Covid-19. *J. Biomol. Struct. Dyn.* 39, 2904–2913. doi:10.1080/07391102.2020.1757510
- Puskasich, M. A., Cummins, N. W., Ingraham, N. E., Wacker, D. A., Reiloff, R. A., Driver, B. E., et al. (2021). A Multi-Center Phase II Randomized Clinical Trial of Losartan on Symptomatic Outpatients With COVID-19. *EClinicalMedicine* 37, 100957. doi:10.1016/j.eclinm.2021.100957
- Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., and Wu, Z. (2019). *Deep Learning for the Life Sciences*. Sebastopol, CA: O'Reilly Media. doi:10.1183/13993003.congress-2019.pa1338
- Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., et al. (2020). Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *N. Engl. J. Med.* 382, 970–971. doi:10.1056/nejmc2001468
- Ryckaert, J.-P., Cicotti, 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, 327–341. doi:10.1016/0021-9991(77)90098-5
- Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013a). An Overview of the Amber Biomolecular Simulation Package. *Wires Comput. Mol. Sci.* 3, 198–210. doi:10.1002/wcms.1121
- Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S., and Walker, R. C. (2013b). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theor. Comput.* 9, 3878–3888. doi:10.1021/ct400314y
- Sanchis-Gomar, F., Lavie, C. J., Perez-Quilis, C., Henry, B. M., and Lippi, G. (2020). Angiotensin-Converting Enzyme 2 and Antihypertensives (Angiotensin Receptor Blockers and Angiotensin-Converting Enzyme Inhibitors) in Coronavirus Disease 2019. *Mayo Clinic Proc.* 95, 1222–1230. doi:10.1016/j.mayocp.2020.03.026
- Scudellari, M. (2020). The Sprint to Solve Coronavirus Protein Structures - and Disarm Them with Drugs. *Nature* 581, 252–255. doi:10.1038/d41586-020-01444-z
- 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
- 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
- Ullrich, S., and Nitsche, C. (2020). The SARS-CoV-2 Main Protease as Drug Target. *Bioorg. Med. Chem. Lett.* 30, 127377. doi:10.1016/j.bmcl.2020.127377
- Verkhiwer, G. M. (2020). Molecular Simulations and Network Modeling Reveal an Allosteric Signaling in the SARS-CoV-2 Spike Proteins. *J. Proteome Res.* 19, 4587–4608.
- Voloch, C. M., da Silva Francisco, R., Jr., de Almedia, L. G. P., Cardoso, C. C., Brustonlini, O. J., Gerber, A. L., et al. (2020). Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* 95, e00119. doi:10.1128/JVI.00119-21
- 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, 1157–1174. doi:10.1002/jcc.20035
- WHO Coronavirus Disease (2020). (COVID-19) Dashboard. Available at: <https://covid19.who.int> (Accessed January 28, 2021).
- World Health Organization (2020). Novel Coronavirus (2019-nCoV). situation report - 1 21 January 2020, 1–7. WHO Bull.
- 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.abb2507
- Xia, S., Lan, Q., Su, S., Wang, X., Xu, W., Liu, Z., et al. (2020). The Role of Furin Cleavage Site in SARS-CoV-2 Spike Protein-Mediated Membrane Fusion in the Presence or Absence of Trypsin. *Signal. Transduct. Target. Ther.* 5, 92. doi:10.1038/s41392-020-0184-0
- Xie, X., Muruato, A., Lokugamage, K. G., Narayanan, K., Zhang, X., Zou, J., et al. (2020). An Infectious cDNA Clone of SARS-CoV-2. *Cell Host Microbe* 27, 841–848. doi:10.1016/j.chom.2020.04.004
- Zhang, J.-H., Chung, T. D. Y., and Oldenburg, K. R. (1999). A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screen.* 4, 67–73. doi:10.1177/108705719900400206

- Zhang, P., Zhu, L., Cai, J., Lei, F., Qin, J.-J., Xie, J., et al. (2020). Association of Inpatient Use of Angiotensin-Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers with Mortality Among Patients with Hypertension Hospitalized with COVID-19. *Circ. Res.* 126, 1671–1681. doi:10.1161/circresaha.120.317134
- Zhang, X., Perez-Sanchez, H., and Lightstone, F. C. (2017). A Comprehensive Docking and MM/GBSA Rescoring Study of Ligand Recognition upon Binding Antithrombin. *Curr. Top. Med. Chem.* 17, 1–9. doi:10.2174/1568026616666161117112604
- Zhang, X., Wong, S. E., and Lightstone, F. C. (2013). Message Passing Interface and Multithreading Hybrid for Parallel Molecular Docking of Large Databases on Petascale High Performance Computing Machines. *J. Comput. Chem.* 34, 915–927. doi:10.1002/jcc.23214
- Zhang, X., Wong, S. E., and Lightstone, F. C. (2014). Toward Fully Automated High Performance Computing Drug Discovery: A Massively Parallel Virtual Screening Pipeline for Docking and Molecular Mechanics/Generalized Born Surface Area Rescoring to Improve Enrichment. *J. Chem. Inf. Model.* 54, 324–337. doi:10.1021/ci4005145

Conflict of Interest: Authors OAN, BH, RM, EAS, and MAS were employed by Sandia National Laboratories.

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.

Copyright © 2021 Lau, Negrete, Bennett, Bennion, Borucki, Bourguet, Epstein, Franco, Harmon, He, Jones, Kim, Kirshner, Lao, Lo, McLoughlin, Mosesso, Muruges, Saada, Segelke, Stefan, Stevenson, Torres, Weillhammer, Wong, Yang, Zemla, Zhang, Zhu, Allen and Lightstone. 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.



An Extended C-Terminus, the Possible Culprit for Differential Regulation of 5-Aminolevulinate Synthase Isoforms

Gregory A. Hunter^{1*} and Gloria C. Ferreira^{1,2,3*}

¹Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, United States,

²Department of Chemistry, College of Arts and Sciences, University of South Florida, Tampa, FL, United States, ³Global and Planetary Health, College of Public Health, University of South Florida, Tampa, FL, United States

OPEN ACCESS

Edited by:

Robert Stephen Phillips,
University of Georgia, United States

Reviewed by:

Andrea Mozzarelli,
University of Parma, Italy
Tim Mueser,
University of Toledo, United States
Giada Rossignoli,
University of Padua, Italy

*Correspondence:

Gregory A. Hunter
ghunter@usf.edu
Gloria C. Ferreira
gferreir@usf.edu

Specialty section:

This article was submitted to
Structural Biology,
a section of the journal
Frontiers in Molecular Biosciences

Received: 14 April 2022

Accepted: 30 May 2022

Published: 14 July 2022

Citation:

Hunter GA and Ferreira GC (2022) An
Extended C-Terminus, the Possible
Culprit for Differential Regulation of 5-
Aminolevulinate Synthase Isoforms.
Front. Mol. Biosci. 9:920668.
doi: 10.3389/fmolb.2022.920668

5-Aminolevulinate synthase (ALAS; E.C. 2.3.1.37) is a pyridoxal 5'-phosphate (PLP)-dependent enzyme that catalyzes the key regulatory step of porphyrin biosynthesis in metazoa, fungi, and α -proteobacteria. ALAS is evolutionarily related to transaminases and is therefore classified as a fold type I PLP-dependent enzyme. As an enzyme controlling the key committed and rate-determining step of a crucial biochemical pathway ALAS is ideally positioned to be subject to allosteric feedback inhibition. Extensive kinetic and mutational studies demonstrated that the overall enzyme reaction is limited by subtle conformational changes of a hairpin loop gating the active site. These findings, coupled with structural information, facilitated early prediction of allosteric regulation of activity via an extended C-terminal tail unique to eukaryotic forms of the enzyme. This prediction was subsequently supported by the discoveries that mutations in the extended C-terminus of the erythroid ALAS isoform (ALAS2) cause a metabolic disorder known as X-linked protoporphyria not by diminishing activity, but by enhancing it. Furthermore, kinetic, structural, and molecular modeling studies demonstrated that the extended C-terminal tail controls the catalytic rate by modulating conformational flexibility of the active site loop. However, the precise identity of any such molecule remains to be defined. Here we discuss the most plausible allosteric regulators of ALAS activity based on divergences in AlphaFold-predicted ALAS structures and suggest how the mystery of the mechanism whereby the extended C-terminus of mammalian ALASs allosterically controls the rate of porphyrin biosynthesis might be unraveled.

Keywords: 5-aminolevulinate synthase, pyridoxal 5'-phosphate, heme regulatory motif, allostery, redox sensor, porphyrin, regulation, AlphaFold

INTRODUCTION

5-Aminolevulinate synthase (ALAS; EC 2.3.1.37) catalyzes the initial and key regulatory step of heme biosynthesis in metazoa, fungi, and the α -subclass of proteobacteria (Stojanovski et al., 2019; Taylor and Brown, 2022). Pyridoxal 5'-phosphate (PLP) is an essential cofactor for the reaction, which involves the condensation of the α -carbon of glycine with the succinyl group of succinyl-Coenzyme A (SCoA) to produce 5-aminolevulinate (ALA), carbon dioxide, and Coenzyme A (Hunter and Ferreira, 2011) (**Supplementary Figure S1**). In metazoa and fungi, ALAS is translated as a precursor with an N-terminal signal sequence that codes for import into the mitochondrial matrix. Following import, the signal sequence is cleaved, and the mature enzyme has access to the substrate SCoA,

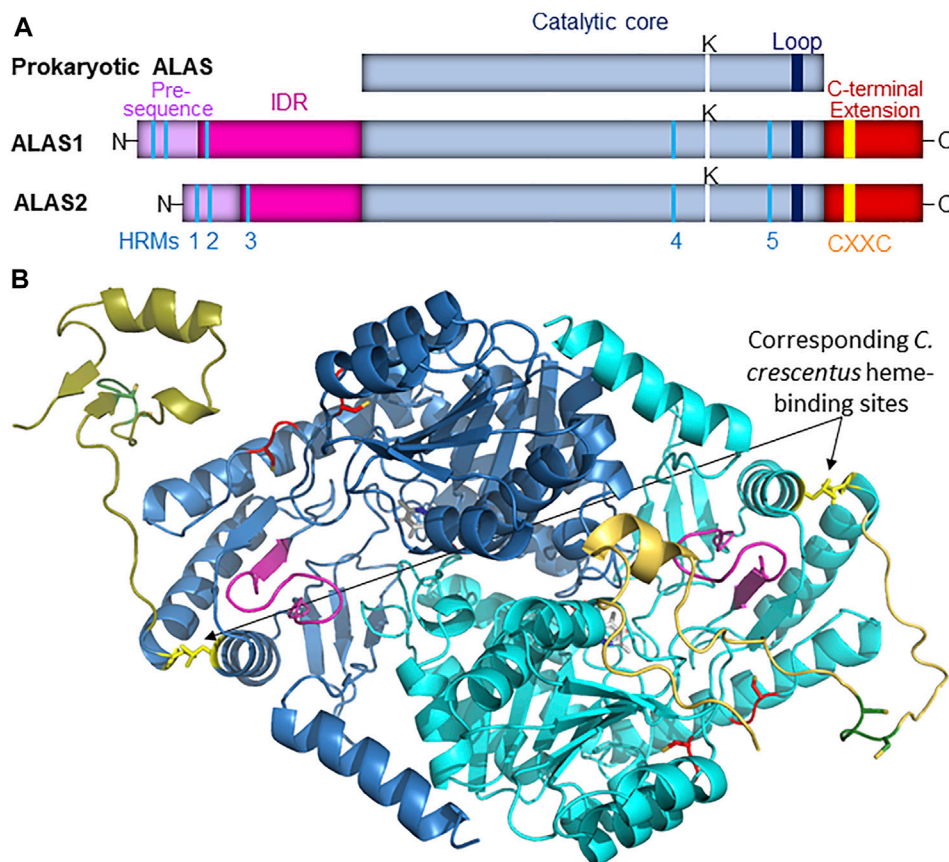


FIGURE 1 | (A). Schematic representation of ALAS monomeric structure. In vertebrate species the ALAS gene is duplicated, and the protein catalytic core (light blue) observed in prokaryotes is bracketed by extended N- and C-termini. The mitochondrial targeting sequence is illustrated in light purple, while the intrinsically disordered N-terminal extension (IDR) is in magenta, and the C-terminal extension is in dark red. Five conserved Heme Regulatory Motifs (HRMs, colored in cyan) are conserved in vertebrate ALAS isoforms, as is a CXXC motif (yellow) in the extended C-terminus. The position of the active site lysine residue that binds PLP in the active site is denoted by a white line, and the loop that gates the active site is represented by a dark blue line. **(B).** The position corresponding to the heme-binding site in *C. crescentus* ALAS modeled into mammalian ALAS2. AlphaFold-predicted structures for human (light blue and gold; AlphaFold entry P22557) and murine (dark blue and gold; AlphaFold entry P08680) ALAS2 were aligned with *R. capsulatus* ALAS crystal structure (PDB code 2bwn; not shown) using Pymol. The modeled site depicted in yellow here is not expected to bind heme in mammals due to evolutionary divergences, and this site is illustrated solely for perspective on its spatial relationship to the mammalian ALAS2 active site loop (purple), the C-terminal extension (shades of gold), and HRMs 4 and 5 (red). Additionally, the CXXC motifs are in green with the cysteines shown as sticks.

which is produced in mitochondria as part of the citric acid cycle. The requirement of SCoA as a substrate integrates heme biosynthesis with oxidative respiration, and as a result the two pathways are synchronized under normal healthy conditions. ALAS activity is additionally synchronized with cellular iron transport as porphyrin biosynthesis and iron transport unite in the final step of heme production wherein the enzyme ferrochelatase inserts ferrous iron into protoporphyrin IX to yield heme (Kafina and Paw, 2017; Poli et al., 2021). As a result of the central position of ALAS in these fundamental biochemical pathways ALAS activity is highly regulated and new modes of ALAS regulation continue to be discovered (Tanimura et al., 2016; Zhang et al., 2017; Liu et al., 2018; Peoc'h et al., 2019; Bailey et al., 2020; Nomura et al., 2021; Rondelli et al., 2021).

Vertebrate genomes encode two chromosomally distinct copies of the ALAS gene: *ALAS1*, which acts as a

“housekeeping” gene and initiates heme biosynthesis in all cells for production of cytochromes and other heme-binding proteins, and *ALAS2*, which is expressed only in developing erythrocytes and produces, almost exclusively, the much larger quantities of heme required for hemoglobin formation (Riddle et al., 1989; Peoc'h et al., 2019). The catalytic cores of human *ALAS1* and *ALAS2* are 75% identical and 94% similar in terms of amino acid sequences, suggesting gene duplication and similar enzymology despite the different metabolic functionalities of the gene products. The high degree of similarity in the catalytic cores of *ALAS1* and *ALAS2* is lessened in the extended N- and C-termini of the enzymes (**Supplementary Figure S2**) but the precise extent to which the mature mitochondrial enzymes might be differentially regulated is still an open question. The monomeric primary structures of prokaryotic and vertebrate ALASs are illustrated schematically in **Figure 1A**.

The ALAS-catalyzed reaction not only represents the first committed step of heme production, but also the rate-determining step of porphyrin biosynthesis, as most poignantly evinced by the consistent observation that exogenous ALA administration to mammalian cells leads to rapid protoporphyrin IX accumulation (Hunter and Ferreira, 2011; Nokes et al., 2013). This is clinically important because it means aberrations in ALAS activity can change the overall rate of porphyrin production and cause porphyrin biosynthesis to decouple from oxidative respiration and iron transport, resulting in metabolic imbalances (Taylor and Brown, 2022). For instance, certain liver toxins, such as allylisopropylacetamide, have long been known to elevate ALAS1 activity beyond the rate of iron transport, resulting in porphyrin accumulation and chemically induced porphyria (Goldberg and Rimington, 1955; Granick, 1966). Conversely, genetic defects in ALAS2 that lead to lower enzymatic activity have been identified as the basis for X-linked sideroblastic anemia, a condition characterized by accumulation of iron in erythroblast mitochondria (Abu-Zeinah and DeSancho, 2020). Remarkably, however, loss-of-function mutations are not the only cause of ALAS2-associated metabolic disorder. A limited number of mutations causing premature truncation or frameshifts in the extreme C-terminal extension of ALAS2 lead to variants with increased catalytic efficiencies and a disorder known as X-linked protoporphyria (Whatley et al., 2008; Ducamp et al., 2013; Wang et al., 2020). Interestingly, mutations in ALAS1 have not been associated with any disorder (Stenson et al., 2003).

5-AMINOLEVULINATE SYNTHASE IS A FOLD TYPE I PYRIDOXAL 5'-PHOSPHATE-DEPENDENT ENZYME WITH A DISTINCT ACTIVE SITE LOOP

PLP-dependent enzymes are structurally classified into seven different fold types, of which fold type I, sometimes referred to as the transaminase family, is by far the largest, with over 170 different Enzyme Classification numbers currently assigned (Percudani and Peracchi, 2009). Like other members of the PLP-dependent fold type I family ALAS is a homodimer with the active site buried near the center of the enzyme at the interface between the two monomers, with residues from each monomer being critical for substrate recognition (Brown et al., 2018; Stojanovski et al., 2019). Even though fold type I PLP-dependent enzymes have very little overall primary sequence similarity the active sites are highly conserved and facilitate phylogenetic analyses demonstrating function-based evolutionary relationships (Catazarò et al., 2014). It is thus informative to compare the structure of aspartate aminotransferase (AATase), which has been extensively characterized and is generally considered to be a model for the fold type I family (Toney, 2014), with the ALAS catalytic core, as seen in **Supplementary Figure S3**. The aligned structures of AATase in the open and closed conformations reveal the structure collapses inwards towards the PLP cofactor upon

substrate binding (McPhalen et al., 1992a; McPhalen et al., 1992b). A short active site loop (green and gold in **Supplementary Figure S3A**) closes inward over the active site cleft upon substrate binding, culminating in an arginine residue that is highly conserved in fold type I enzymes, and functions to form an ionic bond with the carboxylate group of the amino acid substrate (Tan et al., 1998; Liang et al., 2019). In AATase, this arginine is one of only two amino acids that has been designated as a “closure-inducing residue”, meaning it is essential for substrate-induced conformational change from the open to the closed state in which catalysis is optimized (Hayward, 2004). Comparison of these structures to analogous structures of *Rhodobacter capsulatus* ALAS (**Supplementary Figure S3B**) reveals that in ALAS substrate-induced conformational changes are largely limited to the active site loop, which has become longer and is turned more inward over the active site cleft relative to AATase.

Detailed mutational, kinetic, and molecular modeling studies have found that the rate of ALAS catalysis, and hence the rate of porphyrin production, are controlled by the slow opening of this active site loop, which allows the products to rapidly dissociate from the enzyme (Hunter and Ferreira, 1999; Hunter and Ferreira, 2011; Hunter et al., 2007; Stojanovski et al., 2019). This rate-dependence on conformational dynamics would seem to be an ideal situation for allosteric feedback inhibition of the heme biosynthesis pathway via a mechanism wherein effector binding to ALAS would modulate the active site loop conformational dynamics, as we previously suggested (Hunter et al., 2007).

5-AMINOLEVULINATE SYNTHASE STRUCTURAL FEATURES REVEAL IMPORTANT CLUES TO THE POSSIBILITY OF ALLOSTERIC REGULATION

Feedback inhibition of ALAS activity by heme has been known for over 50 years (Granick, 1966), and since then this regulation has been found to occur at a variety of levels, including gene transcription (Yamamoto et al., 1982), transport into mitochondria (Lathrop and Timko, 1993; Munakata et al., 2004), and targeting for degradation (Cable et al., 1996; Yoshino et al., 2007; Tian et al., 2011; Nomura et al., 2021). However, as of this writing direct binding of heme leading to allosteric feedback inhibition of ALAS has only been reported for the enzyme from the prokaryote *Caulobacter crescentus*, in which axial heme binding by H340 and C398 near the C-terminus of the enzyme causes PLP dissociation (Ikushiro et al., 2018) (**Figure 1B**). While the authors reported that these residues are conserved in some other α -proteobacteria and did confirm that recombinant *R. capsulatus* ALAS could also be isolated as a mixture of PLP- and heme-bound forms, these residues are not conserved in eukaryotes, so if allosteric feedback inhibition of ALAS in higher species occurs it must be *via* a different site. The recently resolved crystal structure for human ALAS2 revealed that the extended C-terminus might act as an autoinhibitory

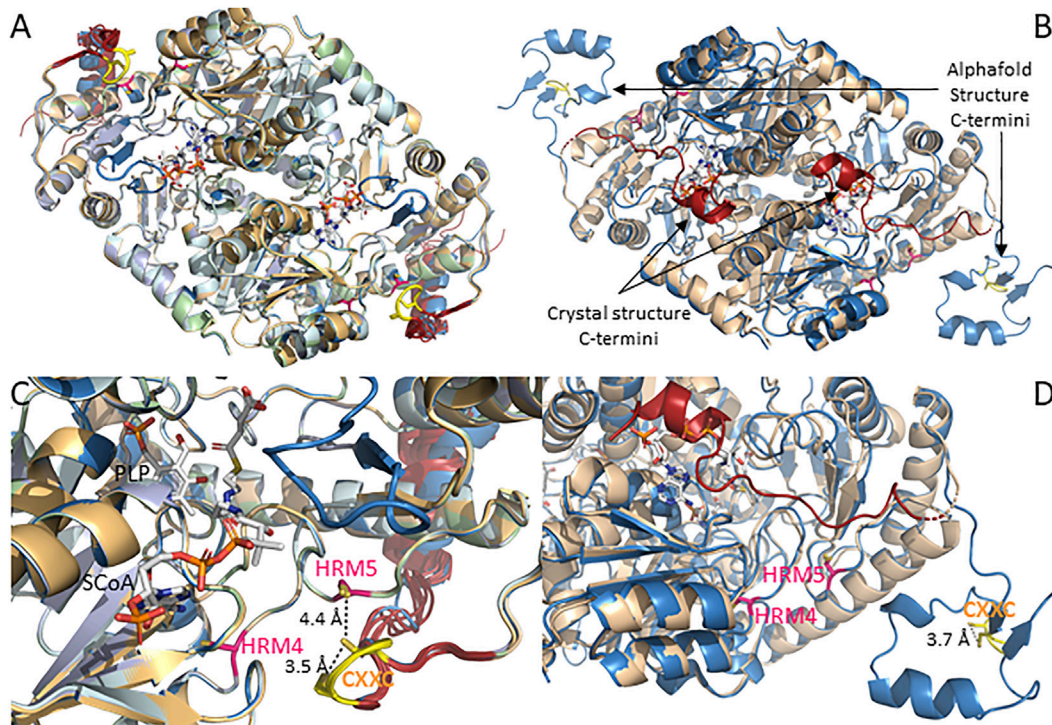


FIGURE 2 | AlphaFold structures for mammalian ALAS1 and ALAS2 reveal C-terminal divergences from the human ALAS2 crystal structure. **(A).** Alignment of AlphaFold-predicted structures of ALAS1 from human (UniProt accession # P13196), orangutan (UniProt accession Q5R9R9), bovine (UniProt accession A6QLI6), beluga whale (UniProt accession Q9XS79), mouse (UniProt accession Q8VC19), and rat (UniProt accession # P13195). **(B).** Alignment of human ALAS2 crystal and AlphaFold-predicted structures. AlphaFold-predicted structure (UniProt accession # P22557; blue) and crystal structure (PDB code 6HRH; beige with red C-termini). **(C).** Zoom of panel **(A)**. **(D).** Zoom of panel **(B)**.

element by folding directly over the active site cleft, clearly implying the existence of some allosteric modulator that alters the conformational dynamics about the C-terminus to allow substrates to access the active site (Bailey et al., 2020), and yet the identity of this effector remains a mystery.

Each of the vertebrate ALAS isozymes contains five heme-regulatory motifs (HRMs), consensus sequences containing a cysteine-proline dipeptide with the cysteine functioning as a ligand to Fe^{3+} -heme (Figure 1A) (Carter et al., 2017; Fleischhacker et al., 2020). HRMs are important in regulating the activity of a wide variety of enzymes controlling gene transcription (Hou et al., 2006; Fleischhacker et al., 2018; Arunachalam et al., 2021), protein synthesis (Igarashi et al., 2008), circadian rhythms (Yang et al., 2008), iron homeostasis (Nishitani et al., 2019), signal transduction (Shen et al., 2014; Schmalohr et al., 2021), and heme degradation (Fleischhacker et al., 2015; Fleischhacker et al., 2018). The first two ALAS HRMs reside in the mitochondrial import signal sequence, where they are positioned to bind excess labile heme and form a complex that is not imported into mitochondria, thus providing a form of feedback inhibition (Lathrop and Timko, 1993; Munakata et al., 2004). Following import the signal sequences are proteolytically removed to produce mature enzymes with intrinsically disordered N-termini (Stojanovski et al., 2016; Nomura et al., 2021). This N-terminal extension contains a third conserved

HRM that feedback inhibits ALAS1 by binding heme to form a complex targeting ALAS1 for proteolysis by the matrix peptidase chaperone subunit ClpX (Nomura et al., 2021). ClpX also controls ALAS2 turnover (Rondelli et al., 2021), and since HRM 3 is conserved in ALAS2, it seems likely that HRM 3 also mediates ClpX degradation of ALAS2 in a heme-dependent fashion, although this remains to be conclusively demonstrated.

The catalytic core of mammalian ALAS, which is approximately 44 kD in size, contains two additional conserved HRMs, which we designate HRMs 4 and 5. To the best of our knowledge, no studies have yet examined their potential biochemical significance. Along with the human ALAS2 crystal structure, the mammalian ALAS1 and ALAS2 AlphaFold-predicted structures reveal that even though HRMs 4 and 5 are ~132 amino acids apart in the primary sequence, in the three-dimensional structures the cysteine α -carbons are only 11 Å apart, and most importantly, they are near or at the enzyme surface in proximity to both the active site loop and the extended C-terminus, in conspicuous positions for heme-mediated feedback regulation of the mature enzyme (Figure 2). The positions of HRMs 4 and 5 in the AlphaFold-predicted ALAS structures are virtually indistinguishable from those in the human ALAS2 crystal structure (Figure 2B).

There are, however, important differences in the relative positions of the extended C-termini of the ALAS1 and ALAS2

isozymes as it relates to HRMs 4 and 5. As seen in **Figures 2A,C**, all six of the currently available AlphaFold-predicted structures for mammalian ALAS1 position the extended C-terminus such that the CXXC motif forms a hairpin loop that brings the cysteine sulfur atoms within ~ 3.5 Å of each other, suggesting disulfide bond formation and a possible redox sensing role. Furthermore, the CXXC loop is positioned almost directly over HRM5.

In contrast to the consensus positioning of the ALAS1 extended C-terminus over HRMs 4 and 5, the AlphaFold-predicted mammalian ALAS2 structures have more conformational heterogeneity about the C-terminal extension (**Figures 2B,D**). Moreover, none of the ALAS2 C-terminal extensions align with the ALAS1 C-terminus. Instead, the ALAS2 C-terminal extensions fall into one of three different conformations. In the AlphaFold-predicted structures for orangutan, bovine, beluga whale, and rat ALAS2s, the extended C-terminus folds over the active site to form an “autoinhibited” structure, in excellent alignment with the recently solved human ALAS2 crystal structure (Bailey et al., 2020), but the AlphaFold-predicted human ALAS2 structure places the extended C-terminus away from the catalytic core in what would presumably correspond to an active enzyme conformation. Meanwhile, in the mouse ALAS2, the extended C-terminus adopts a conformation between these two extremes. In all cases the cysteines of the ALAS2 CXXC motifs, like those of the ALAS1 CXXC motifs, are in sufficient proximity to reversibly form disulfides, and thus potentially act as redox sensors. But unlike ALAS1, HRMs 4 and 5 of ALAS2 are not occluded by the C-terminal extension and are thus more available to bind heme in what would presumably be a feedback-inhibited complex.

A CASE FOR DIFFERENTIAL REGULATION BY THE C-TERMINAL EXTENSIONS

Remarkably, in ten out of twelve different mammalian ALAS mitochondrial import presequences AlphaFold predicts the side chains of the cysteines in HRMs 1 and 2 to be almost ideally positioned to act as axial ligands for heme (**Supplementary Figure S4**). This agrees with experimental evidence demonstrating HRMs 1 and 2 bind heme to feedback inhibit mitochondrial import (Lathrop and Timko, 1993; Goodfellow et al., 2001; Munakata et al., 2004). Further, it leads us to suggest that the predicted conformational differences in the extended C-termini might in turn be experimentally revealed to be accurate predictors of important structural/functional divergences between the two ALAS isozymes.

The AlphaFold structural database currently has nearly a million protein structures available, including complete proteomes for *Homo sapiens* and 47 other species (Jumper et al., 2021; Tunyasuvunakool et al., 2021). These structures are rapidly facilitating an unprecedented understanding of structural biology (Hegedus et al., 2022; Porta-Pardo et al., 2022; Varadi et al., 2022; Wehrspan et al., 2022). Yet, the accuracy of AlphaFold in terms of predicting otherwise unsolved structures is relatively untested since it only became publicly available less than a year ago. AlphaFold is

reported to accurately predict not just the highly organized structures observed in crystallized proteins, but also the extent of conformational dynamics or even intrinsic disorder in individual residues or peptides by calculating a per residue confidence score referred to as a predicted local distance difference test (pLDDT) (Tunyasuvunakool et al., 2021). The current interpretation of this score is that it predicts the extent to which a residue is unstructured, meaning a low score should be seen not so much as an indication the structure is inaccurate, but more as an accurate indication of greater conformational dynamics. Because of this AlphaFold should provide important insight into dynamic regulatory structures that have been difficult to crystallize.

The ALAS1/2 conserved CXXC motif is of particular interest since similar motifs act as allosteric redox switches *via* reversible formation of a disulfide bond in many enzymes, including the PLP-dependent enzymes cystathionine β -synthase and human mitochondrial branched chain aminotransferase (Conway et al., 2004; Wouters et al., 2010; Niu et al., 2018; Herbert et al., 2020). The CXXC motif-containing region was only partially resolved in the human ALAS2 crystal structure, implying a high degree of conformational mobility. The AlphaFold pLDDT scores for the six mammalian ALAS2 (and six ALAS1) structures in the public database agree, as they drop from very high confidence to low or even very low for the corresponding amino acids in all species except human ALAS2 (**Supplementary Figure S5**), in which the extended C-terminus adopts what is presumably an activated enzyme conformation. In this “activated” ALAS2 structure the scores for the CXXC motif are mostly confident, indicating greater structural organization, and with the cysteine side chain sulfur atoms within 3.7 Å of each other, disulfide bond formation is possible. Given all these considerations, if the CXXC motif in the extended C-terminus of ALAS2 acts as a redox switch we would predict that the “activated” structure would be oxidized to the disulfide, while the more disordered autoinhibited structure would be reduced.

The positioning of the human ALAS2 extended C-terminus over the active site leads us to raise the questions as to what the active conformation might look like and how the interconversion between the inhibited and activated conformations might be triggered. The corresponding AlphaFold structure appears to provide a plausible answer to the first of these two questions, but only hints at the answer to the second. Binding of the β -subunit of succinyl-CoA synthetase (Furuyama and Sassa, 2000; Bishop et al., 2012; Bishop et al., 2013) and/or other heme biosynthetic enzymes might promote activation (Medlock et al., 2015). A novel, but certainly not mutually exclusive, possibility supported by the structures analyzed here is that the CXXC motif acts as a redox sensor to modulate conformational dynamics about the extended C-terminus.

In contrast to ALAS2, a crystal structure for ALAS1 has not yet been reported, and the AlphaFold-predicted structures indicate only one conformation for the ALAS1 extended C-terminus. Yet, the CXXC motif is conserved in ALAS1, and if it has a redox switching function then some degree of conformational perturbation presumably occurs to form an autoinhibited conformation or to alter the dynamics about the active site

loop, which controls the catalytic rate. This latter possibility is attractive as it would be consistent with the anti-correlation between the active site loop and C-terminal extension of ALAS2 during molecular dynamics simulations (Na et al., 2018). Additionally, the shielding of the otherwise solvent exposed HRMs 4 and 5 by the ALAS1 C-terminal extension suggests an alternative conformation that would allow heme access to feedback inhibit the enzyme. Given these considerations we posit that the ALAS1 structures represent an activated form wherein the CXXC motif is oxidized to the disulfide and positioned to prevent allosteric feedback inhibition by heme. Reduction of the CXXC motif would then facilitate a conformation change allowing heme to allosterically feedback inhibit ALAS1 *via* HRMs 4 and/or 5. A more prominent role of redox sensing in ALAS1 is in part attractive due to the role of ALAS1 in producing heme specifically for hemoproteins catalyzing redox chemistry, such as cytochrome P450 enzymes, catalase, and superoxide dismutase.

CONCLUSION AND OUTLOOK

In summary, based upon the alignment of the ALAS1 structures we put forth the following postulates: 1) HRMs 4 and/or 5 facilitate feedback inhibition of ALAS1; 2) under oxidizing conditions, the CXXC motif forms a disulfide bond that causes the C-terminal extension to fold over HRMs 4 and 5 such that it sterically prevents hemin binding and feedback inhibition; 3) under non-oxidizing conditions, the CXXC motif is reduced and adopts an alternative conformation wherein HRMs 4 and 5 are exposed to provide feedback inhibition by excess heme. Stated more concisely, feedback inhibition of ALAS1 by heme is dependent upon cellular redox status.

Based on the alignment of the ALAS2 structures we put forth the following postulates: 1) the C-terminal extension of ALAS2 adopts two different conformations, neither of which prevents feedback regulation via heme binding to HRMs 4 and 5. 2) In

ALAS2 oxidizing conditions cause disulfide bond formation in the CXXC motif and movement of the extended C-terminus not over HRMs 4&5 but instead to a more equatorial and activated position relative to the enzyme, thereby relieving the autoinhibition observed when the extended C-terminus folds over the active site. Stated more succinctly, heme and redox status independently regulate ALAS2 activity.

These postulates are not incompatible with the possibility of protein-protein interactions regulating activity. Of course, experimental data will be required to further support or refine the views presented here, but whatever the outcome the remarkably divergent structures discussed here will likely represent a key test of the capacity of AlphaFold to discern fine structural differences and facilitate prediction of allostery in all enzymes, including those dependent upon PLP for functionality.

AUTHOR CONTRIBUTIONS

GH performed the structural analyses and analyzed the data. GH and GF conceptualized, wrote, and edited the manuscript. GH and GF approved the submitted version of the manuscript.

FUNDING

During the writing of this review, the authors were partially supported by a grant from the Florida Department of Health, Bankhead-Coley Cancer Research Program (#9BC14).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2022.920668/full#supplementary-material>

REFERENCES

- Abu-Zeinah, G., and DeSancho, M. T. (2020). Understanding Sideroblastic Anemia: An Overview of Genetics, Epidemiology, Pathophysiology and Current Therapeutic Options. *Jbm* 11, 305–318. doi:10.2147/jbm.s232644
- Arunachalam, A., Lakshmanan, D. K., Ravichandran, G., Paul, S., Manickam, S., Kumar, P. V., et al. (2021). Regulatory Mechanisms of Heme Regulatory Protein BACH1: A Potential Therapeutic Target for Cancer. *Med. Oncol.* 38, 122. doi:10.1007/s12032-021-01573-z
- Bailey, H. J., Bezerra, G. A., Marcero, J. R., Padhi, S., Foster, W. R., Rembeza, E., et al. (2020). Human Aminolevulinic Synthase Structure Reveals a Eukaryotic-Specific Autoinhibitory Loop Regulating Substrate Binding and Product Release. *Nat. Commun.* 11, 2813. doi:10.1038/s41467-020-16586-x
- Bishop, D. F., Tchaikovskii, V., Hoffbrand, A. V., Fraser, M. E., and Margolis, S. (2012). X-Linked Sideroblastic Anemia Due to Carboxyl-Terminal ALAS2 Mutations that Cause Loss of Binding to the β -Subunit of Succinyl-CoA Synthetase (SUCLA2). *J. Biol. Chem.* 287, 28943–28955. doi:10.1074/jbc.m111.306423
- Bishop, D. F., Tchaikovskii, V., Nazarenko, I., and Desnick, R. J. (2013). Molecular Expression and Characterization of Erythroid-Specific 5-Aminolevulinic Synthase Gain-Of-Function Mutations Causing X-Linked Protoporphyrria. *Mol. Med.* 19, 18–25. doi:10.2119/molmed.2013.00003
- Brown, B. L., Kardon, J. R., Sauer, R. T., and Baker, T. A. (2018). Structure of the Mitochondrial Aminolevulinic Acid Synthase, a Key Heme Biosynthetic Enzyme. *Structure* 26, 580–589. doi:10.1016/j.str.2018.02.012
- Cable, E. E., Gildemeister, O. S., Pepe, J. A., Donohue, S. E., Lambrecht, R. W., and Bonkovsky, H. L. (1996). Hepatic 5-Aminolevulinic Acid Synthase mRNA Stability Is Modulated by Inhibitors of Heme Biosynthesis and by Metalloporphyrins. *Eur. J. Biochem.* 240, 112–117. doi:10.1111/j.1432-1033.1996.0112h.x
- Carter, E. L., Ramirez, Y., and Ragsdale, S. W. (2017). The Heme-Regulatory Motif of Nuclear Receptor Rev-Erb β Is a Key Mediator of Heme and Redox Signaling in Circadian Rhythm Maintenance and Metabolism. *J. Biol. Chem.* 292, 11280–11299. doi:10.1074/jbc.m117.783118
- Catazaro, J., Caprez, A., Guru, A., Swanson, D., and Powers, R. (2014). Functional Evolution of PLP-Dependent Enzymes Based on Active-Site Structural Similarities. *Proteins* 82, 2597–2608. doi:10.1002/prot.24624
- Conway, M. E., Poole, L. B., and Hutson, S. M. (2004). Roles for Cysteine Residues in the Regulatory CXXC Motif of Human Mitochondrial Branched Chain Aminotransferase Enzyme. *Biochemistry* 43, 7356–7364. doi:10.1021/bi0498050

- Ducamp, S., Schneider-Yin, X., de Rooij, F., Clayton, J., Fratz, E. J., Rudd, A., et al. (2013). Molecular and Functional Analysis of the C-Terminal Region of Human Erythroid-Specific 5-Aminolevulinic Synthase Associated with X-Linked Dominant Protoporphyrin (XLDPP). *Hum. Mol. Genet.* 22, 1280–1288. doi:10.1093/hmg/dd531
- Fleischhacker, A. S., Carter, E. L., and Ragsdale, S. W. (2018). Redox Regulation of Heme Oxygenase-2 and the Transcription Factor, Rev-Erb, Through Heme Regulatory Motifs. *Antioxidants Redox Signal.* 29, 1841–1857. doi:10.1089/ars.2017.7368
- Fleischhacker, A. S., Gunawan, A. L., Kochert, B. A., Liu, L., Wales, T. E., Borowy, M. C., et al. (2020). The Heme-Regulatory Motifs of Heme Oxygenase-2 Contribute to the Transfer of Heme to the Catalytic Site for Degradation. *J. Biol. Chem.* 295, 5177–5191. doi:10.1074/jbc.ra120.012803
- Fleischhacker, A. S., Sharma, A., Choi, M., Spencer, A. M., Bagai, I., Hoffman, B. M., et al. (2015). The C-Terminal Heme Regulatory Motifs of Heme Oxygenase-2 Are Redox-Regulated Heme Binding Sites. *Biochemistry* 54, 2709–2718. doi:10.1021/acs.biochem.5b00266
- Furuyama, K., and Sassa, S. (2000). Interaction Between Succinyl CoA Synthetase and the Heme-Biosynthetic Enzyme ALAS-E Is Disrupted in Sideroblastic Anemia. *J. Clin. Invest.* 105, 757–764. doi:10.1172/jci6816
- Goldberg, A., and Rimington, C. (1955). Experimentally Produced Porphyrin in Animals. *Proc. R. Soc. Lond B Biol. Sci.* 143, 257–279. doi:10.1098/rspb.1955.0009
- Goodfellow, B. J., Dias, J. S., Ferreira, G. C., Henklein, P., Wray, V., and Macedo, A. L. (2001). The Solution Structure and Heme Binding of the Presequence of Murine 5-Aminolevulinic Synthase. *FEBS Lett.* 505, 325–331. doi:10.1016/s0014-5793(01)02818-6
- Granick, S. (1966). The Induction *In Vitro* of the Synthesis of δ -Aminolevulinic Acid Synthetase in Chemical Porphyrin: A Response to Certain Drugs, Sex Hormones, and Foreign Chemicals. *J. Biol. Chem.* 241, 1359–1375. doi:10.1016/s0021-9258(18)96783-9
- Hayward, S. (2004). Identification of Specific Interactions That Drive Ligand-Induced Closure in Five Enzymes with Classic Domain Movements. *J. Mol. Biol.* 339, 1001–1021. doi:10.1016/j.jmb.2004.04.004
- Hegedűs, T., Geisler, M., Lukács, G. L., and Farkas, B. (2022). Ins and Outs of AlphaFold2 Transmembrane Protein Structure Predictions. *Cell Mol. Life Sci.* 79, 73. doi:10.1007/s00018-021-04112-1
- Herbert, D., Gibbs, S., Riddick, A., Conway, M., and Dong, M. (2020). Crystal Structure of an Oxidized Mutant of Human Mitochondrial Branched-Chain Aminotransferase. *Acta Cryst. Sect. F* 76, 14–19. doi:10.1107/s2053230x19016480
- Hou, S., Reynolds, M. F., Horrigan, F. T., Heinemann, S. H., and Hoshi, T. (2006). Reversible Binding of Heme to Proteins in Cellular Signal Transduction. *Acc. Chem. Res.* 39, 918–924. doi:10.1021/ar040020w
- Hunter, G. A., and Ferreira, G. C. (2011). Molecular Enzymology of 5-Aminolevulinic Synthase, the Gatekeeper of Heme Biosynthesis. *Biochimica Biophysica Acta (BBA) - Proteins Proteomics* 1814, 1467–1473. doi:10.1016/j.bbapap.2010.12.015
- Hunter, G. A., and Ferreira, G. C. (1999). Pre-Steady-State Reaction of 5-Aminolevulinic Synthase. *J. Biol. Chem.* 274, 12222–12228. doi:10.1074/jbc.274.18.12222
- Hunter, G. A., Zhang, J., and Ferreira, G. C. (2007). Transient Kinetic Studies Support Refinements to the Chemical and Kinetic Mechanisms of Aminolevulinic Synthase. *J. Biol. Chem.* 282, 23025–23035. doi:10.1074/jbc.m609330200
- Igarashi, J., Murase, M., Iizuka, A., Pichierr, F., Martinkova, M., and Shimizu, T. (2008). Elucidation of the Heme Binding Site of Heme-Regulated Eukaryotic Initiation Factor 2 α Kinase and the Role of the Regulatory Motif in Heme Sensing by Spectroscopic and Catalytic Studies of Mutant Proteins. *J. Biol. Chem.* 283, 18782–18791. doi:10.1074/jbc.m801400200
- Ikushiro, H., Nagami, A., Takai, T., Sawai, T., Shimeno, Y., Hori, H., et al. (2018). Heme-Dependent Inactivation of 5-Aminolevulinic Synthase from *Caulobacter crescentus*. *Sci. Rep.* 8, 14228. doi:10.1038/s41598-018-32591-z
- 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
- Kafina, M. D., and Paw, B. H. (2017). Intracellular Iron and Heme Trafficking and Metabolism in Developing Erythroblasts. *Metallomics* 9, 1193–1203. doi:10.1039/c7mt00103g
- Lathrop, J. T., and Timko, M. P. (1993). Regulation by Heme of Mitochondrial Protein Transport Through a Conserved Amino Acid Motif. *Science* 259, 522–525. doi:10.1126/science.8424176
- Liang, J., Han, Q., Tan, Y., Ding, H., and Li, J. (2019). Current Advances on Structure-Function Relationships of Pyridoxal 5'-Phosphate-Dependent Enzymes. *Front. Mol. Biosci.* 6, 4. doi:10.3389/fmolb.2019.00004
- Liu, J., Li, Y., Tong, J., Gao, J., Guo, Q., Zhang, L., et al. (2018). Long Non-Coding RNA-Dependent Mechanism to Regulate Heme Biosynthesis and Erythrocyte Development. *Nat. Commun.* 9, 4386. doi:10.1038/s41467-018-06883-x
- McPhalen, C. A., Vincent, M. G., and Jansonius, J. N. (1992a). X-Ray Structure Refinement and Comparison of Three Forms of Mitochondrial Aspartate Aminotransferase. *J. Mol. Biol.* 225, 495–517. doi:10.1016/0022-2836(92)90935-d
- McPhalen, C. A., Vincent, M. G., Picot, D., Jansonius, J. N., Lesk, A. M., and Chothia, C. (1992b). Domain Closure in Mitochondrial Aspartate Aminotransferase. *J. Mol. Biol.* 227, 197–213. doi:10.1016/0022-2836(92)90691-c
- Medlock, A. E., Shiferaw, M. T., Marcero, J. R., Vashisht, A. A., Wohlschlegel, J. A., Phillips, J. D., et al. (2015). Identification of the Mitochondrial Heme Metabolism Complex. *PLoS One* 10, e0135896. doi:10.1371/journal.pone.0135896
- Munakata, H., Sun, J. Y., Yoshida, K., Nakatani, T., Honda, E., Hayakawa, S., et al. (2004). Role of the Heme Regulatory Motif in the Heme-Mediated Inhibition of Mitochondrial Import of 5-Aminolevulinic Synthase. *J. Biochem.* 136, 233–238. doi:10.1093/jb/mvh112
- Na, I., Catena, D., Kong, M. J., Ferreira, G. C., and Uversky, V. N. (2018). Anti-Correlation Between the Dynamics of the Active Site Loop and C-Terminal Tail in Relation to the Homodimer Asymmetry of the Mouse Erythroid 5-Aminolevulinic Synthase. *Int. J. Mol. Sci.* 19. doi:10.3390/ijms19071899
- Nishitani, Y., Okutani, H., Takeda, Y., Uchida, T., Iwai, K., and Ishimori, K. (2019). Specific Heme Binding to Heme Regulatory Motifs in Iron Regulatory Proteins and its Functional Significance. *J. Inorg. Biochem.* 198, 110726. doi:10.1016/j.jinorgbio.2019.110726
- Niu, W., Wang, J., Qian, J., Wang, M., Wu, P., Chen, F., et al. (2018). Allosteric Control of Human Cystathionine β -Synthase Activity by a Redox Active Disulfide Bond. *J. Biol. Chem.* 293, 2523–2533. doi:10.1074/jbc.ra117.000103
- Nokes, B., Apel, M., Jones, C., Brown, G., and Lang, J. E. (2013). Aminolevulinic Acid (ALA): Photodynamic Detection and Potential Therapeutic Applications. *J. Surg. Res.* 181, 262–271. doi:10.1016/j.jss.2013.02.002
- Nomura, K., Kitagawa, Y., Aihara, M., Ohki, Y., Furuyama, K., and Hirokawa, T. (2021). Heme-Dependent Recognition of 5-Aminolevulinic Synthase by the Human Mitochondrial Molecular Chaperone ClpX. *FEBS Lett.* 595, 3019–3029. doi:10.1002/1873-3468.14214
- Peoc'h, K., Nicolas, G., Schmitt, C., Mirmiran, A., Daher, R., Lefebvre, T., et al. (2019). Regulation and Tissue-Specific Expression of δ -Aminolevulinic Acid Synthases in Non-Syndromic Sideroblastic Anemias and Porphyrrias. *Mol. Genet. Metabolism* 128, 190–197. doi:10.1016/j.ymgme.2019.01.015
- Percudani, R., and Peracchi, A. (2009). The B6 Database: A Tool for the Description and Classification of Vitamin B6-Dependent Enzymatic Activities and of the Corresponding Protein Families. *BMC Bioinforma.* 10, 273. doi:10.1186/1471-2105-10-273
- Poli, A., Schmitt, C., Moulouel, B., Mirmiran, A., Puy, H., Lefebvre, T., et al. (2021). Iron, Heme Synthesis and Erythropoietic Porphyrins: A Complex Interplay. *Metabolites* 11 (12), 798. doi:10.3390/metabo11120798
- Porta-Pardo, E., Ruiz-Serra, V., Valentini, S., and Valencia, A. (2022). The Structural Coverage of the Human Proteome Before and After AlphaFold. *PLoS Comput. Biol.* 18, e1009818. doi:10.1371/journal.pcbi.1009818
- Riddle, R. D., Yamamoto, M., and Engel, J. D. (1989). Expression of Delta-Aminolevulinic Synthase in Avian Cells: Separate Genes Encode Erythroid-Specific and Nonspecific Isozymes. *Proc. Natl. Acad. Sci. U.S.A.* 86, 792–796. doi:10.1073/pnas.86.3.792
- Rondelli, C. M., Perfetto, M., Danoff, A., Bergonia, H., Gillis, S., O'Neill, L., et al. (2021). The Ubiquitous Mitochondrial Protein Unfoldase CLPX Regulates Erythroid Heme Synthesis by Control of Iron Utilization and Heme Synthesis Enzyme Activation and Turnover. *J. Biol. Chem.* 297, 100972. doi:10.1016/j.jbc.2021.100972
- Schmalohr, B. F., Mustafa, A. H. M., Krämer, O. H., and Imhof, D. (2021). Structural Insights into the Interaction of Heme with Protein Tyrosine Kinase JAK2*. *Chembiochem* 22, 861–864. doi:10.1002/cbic.202000730

- Shen, J., Sheng, X., Chang, Z., Wu, Q., Wang, S., Xuan, Z., et al. (2014). Iron Metabolism Regulates P53 Signaling Through Direct Heme-P53 Interaction and Modulation of P53 Localization, Stability, and Function. *Cell Rep.* 7, 180–193. doi:10.1016/j.celrep.2014.02.042
- Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., et al. (2003). Human Gene Mutation Database (HGMD): 2003 Update. *Hum. Mutat.* 21, 577–581. doi:10.1002/humu.10212
- Stojanovski, B. M., Breydo, L., Uversky, V. N., and Ferreira, G. C. (2016). Murine Erythroid 5-Aminolevulinate Synthase: Truncation of a Disordered N-Terminal Extension Is Not Detrimental for Catalysis. *Biochimica Biophysica Acta (BBA) - Proteins Proteomics* 1864, 441–452. doi:10.1016/j.bbapap.2016.02.002
- Stojanovski, B. M., Hunter, G. A., Na, I., Uversky, V. N., Jiang, R. H. Y., and Ferreira, G. C. (2019). 5-Aminolevulinate Synthase Catalysis: The Catcher in Heme Biosynthesis. *Mol. Genet. Metabolism* 128, 178–189. doi:10.1016/j.ymgme.2019.06.003
- Tan, D., Harrison, T., Hunter, G. A., and Ferreira, G. C. (1998). Role of Arginine 439 in Substrate Binding of 5-aminolevulinate Synthase. *Biochemistry* 37, 1478–1484. doi:10.1021/bi971928f
- Tanimura, N., Miller, E., Igarashi, K., Yang, D., Burstyn, J. N., Dewey, C. N., et al. (2016). Mechanism Governing Heme Synthesis Reveals a GATA Factor/heme Circuit that Controls Differentiation. *EMBO Rep.* 17, 249–265. doi:10.15252/embr.2015141465
- Taylor, J. L., and Brown, B. L. (2022). Structural Basis for Dysregulation of Aminolevulinic Acid Synthase in Human Disease. *J. Biol. Chem.* 298 (3), 101643. doi:10.1016/j.jbc.2022.101643
- Tian, Q., Li, T., Hou, W., Zheng, J., Schrum, L. W., and Bonkovsky, H. L. (2011). Lon Peptidase 1 (LONP1)-Dependent Breakdown of Mitochondrial 5-Aminolevulinic Acid Synthase Protein by Heme in Human Liver Cells. *J. Biol. Chem.* 286, 26424–26430. doi:10.1074/jbc.M110.215772
- Toney, M. D. (2014). Aspartate Aminotransferase: an Old Dog Teaches New Tricks. *Archives Biochem. Biophysics* 544, 119–127. doi:10.1016/j.abb.2013.10.002
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., et al. (2021). Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* 596, 590–596. doi:10.1038/s41586-021-03828-1
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2022). AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* 50, D439–D444. doi:10.1093/nar/gkab1061
- Wang, T., Wang, Y., Dong, Q., Xu, C., Zhou, X., Ouyang, Y., et al. (2020). X-Linked Dominant Protoporphyrinemia in a Chinese Pedigree Reveals a Four-Based Deletion of ALAS2. *Ann. Transl. Med.* 8, 344. doi:10.21037/atm.2020.02.80
- Wehrspan, Z. J., McDonnell, R. T., and Elcock, A. H. (2022). Identification of Iron-Sulfur (Fe-S) Cluster and Zinc (Zn) Binding Sites within Proteomes Predicted by DeepMind's AlphaFold2 Program Dramatically Expands the Metalloproteome. *J. Mol. Biol.* 434, 167377. doi:10.1016/j.jmb.2021.167377
- Whatley, S. D., Ducamp, S., Gouya, L., Grandchamp, B., Beaumont, C., Badminton, M. N., et al. (2008). C-Terminal Deletions in the ALAS2 Gene Lead to Gain of Function and Cause X-Linked Dominant Protoporphyrinemia Without Anemia or Iron Overload. *Am. J. Hum. Genet.* 83, 408–414. doi:10.1016/j.ajhg.2008.08.003
- Wouters, M. A., Fan, S. W., and Haworth, N. L. (2010). Disulfides as Redox Switches: From Molecular Mechanisms to Functional Significance. *Antioxidants Redox Signal.* 12, 53–91. doi:10.1089/ars.2009.2510
- Yamamoto, M., Hayashi, N., and Kikuchi, G. (1982). Evidence for the Transcriptional Inhibition by Heme of the Synthesis of δ -Aminolevulinate Synthase in Rat Liver. *Biochem. Biophysical Res. Commun.* 105, 985–990. doi:10.1016/0006-291x(82)91067-1
- Yang, J., Kim, K. D., Lucas, A., Drahos, K. E., Santos, C. S., Mury, S. P., et al. (2008). A Novel Heme-Regulatory Motif Mediates Heme-Dependent Degradation of the Circadian Factor Period 2. *Mol. Cell Biol.* 28, 4697–4711. doi:10.1128/mcb.00236-08
- Yoshino, K., Munakata, H., Kuge, O., Ito, A., and Ogishima, T. (2007). Haeme-Regulated Degradation of δ -Aminolevulinate Synthase 1 in Rat Liver Mitochondria. *J. Biochem.* 142, 453–458. doi:10.1093/jb/mvm159
- Zhang, Y., Zhang, J., An, W., Wan, Y., Ma, S., Yin, J., et al. (2017). Intron 1 GATA Site Enhances ALAS2 Expression Indispensably During Erythroid Differentiation. *Nucleic Acids Res.* 45, 657–671. doi:10.1093/nar/gkw901

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 Hunter and Ferreira. 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

Fuhai Li,
Washington University in St. Louis,
United States

REVIEWED BY

Nenad Filipovic,
University of Kragujevac, Serbia
Jing Tang,
University of Helsinki, Finland

*CORRESPONDENCE

Mian Zu,
rabbitzumian@outlook.com
Yin Zhang,
aring2010@163.com

[†]These authors have contributed equally
to this work

SPECIALTY SECTION

This article was submitted to Biological
Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 17 June 2022

ACCEPTED 03 October 2022

PUBLISHED 18 October 2022

CITATION

Hong Y, Chen D, Jin Y, Zu M and Zhang Y
(2022), PINet 1.0: A pathway network-
based evaluation of drug combinations
for the management of
specific diseases.
Front. Mol. Biosci. 9:971768.
doi: 10.3389/fmolb.2022.971768

COPYRIGHT

© 2022 Hong, Chen, Jin, Zu and Zhang.
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.

PINet 1.0: A pathway network-based evaluation of drug combinations for the management of specific diseases

Yongkai Hong, Dantian Chen, Yaqing Jin, Mian Zu^{*†} and
Yin Zhang^{*†}

Institute of Health Service and Transfusion Medicine, Academy of Military Medical Sciences, Beijing, China

Drug combinations can increase the therapeutic effect by reducing the level of toxicity and the occurrence of drug resistance. Therefore, several drug combinations are often used in the management of complex diseases. However, due to the exponential growth in drug development, it would be impractical to evaluate all combinations through experiments. In view of this, we developed Pathway Interaction Network (PINet) biological model to estimate the optimal drug combinations for various diseases. The random walk with restart (RWR) algorithm was used to capture the “disease state” and “drug state,” while PINet was used to evaluate the optimal drug combinations and the high-order drug combination¹. The model achieved a mean area under the curve of a receiver operating characteristic curve of 0.885. In addition, for some diseases, PINet predicted the optimal drug combination. For example, in the case of acute myeloid leukemia, PINet correctly predicted midostaurin and gemtuzumab as effective drug combinations, as demonstrated by the results of a Phase-I clinical trial. Moreover, PINet also correctly predicted the potential drug combinations for diseases that lacked a training dataset that could not be predicted using standard machine learning models.

KEYWORDS

pathway, gene, drug combination, network pharmacology, random walk with restart

1 Introduction

Compared with the “one disease, one gene” drug paradigm, drug combinations can more effectively cope with multifactorial diseases such as infections, cardiovascular diseases, and tumors (Bayat Mokhtari et al., 2017) (Huffman et al., 2017). Drug combinations can also delay the development of drug resistance and are often used in the treatment of acquired immunodeficiency syndrome (AIDS) and multi-drug resistant

¹ Combinations of three or more drugs.

bacteria (Liu et al., 2021) (Cihlar and Fordyce, 2016). Network or multi-pharmacology involves the combinations of several drugs used for different targets to create a synergistic effect that can perturb the biological networks and thus increase the clinical benefits (Jia et al., 2009).

The development of optimal drug combinations typically involves three stages: the intuition phase, the clinical trials phase, and the biological data mining phase. However, since the development of the current drug combination is based on the researchers' intuition and expertise, the process is often inefficient. As a result, it is now gradually being replaced by the high-throughput screening method (Shinn et al., 2019). Nevertheless, as the number of approved drugs increases, the number of drug combinations requiring high-throughput screening verification has increased exponentially, eventually leading to a significant prolongation of the verification process and research costs. Machine learning and deep learning, which can mine the correlation between massive amounts of biological data, are increasingly being used in the discovery of effective drug combinations (Shi et al., 2018) (Li et al., 2020) (Kim et al., 2021) (Zagidullin et al., 2021). Since machine learning depends on training datasets, it is mostly used for tumors. However, for diseases that lack training datasets, the model is difficult to optimize because it is not possible to fit the parameters into the model. In addition, the results provided by the machine learning algorithms are often difficult to explain, and therefore clinicians find it difficult to apply the machine learning solution in clinical practice. An alternative approach to the data-driven machine learning method is to use theory-driven methods based on the knowledge of biological systems and networks (Wang et al., 2021) (Jafari et al., 2022). Compared with data-driven methods, theory-driven methods are more explanatory, and their performance is not affected by the quality of the training dataset. The limitation of theory-driven methods is that they rely on the accurate generation of a theoretical hypothesis.

(Yang et al., 2008) define two network biological states: the disease and normal states. According to (Yang et al., 2008), the transition from the disease state to the normal state is achieved through the perturbation of specific target combinations within the arachidonic acid network (a kind of inflammation-related network). This approach has several limitations. First of all, there is a lack of uniform standards to define the disease and normal states. Therefore, the definition of these states often requires the subjective input of expert professionals. In addition, not all disease targets have corresponding drugs available, and more than one pathway may be involved in the development of a specific disease (Geva-Zatorsky et al., 2010). found that the protein responses to drug combinations can be accurately described by a linear superposition (weighted sum) of each protein's response to each specific individual drug. Based on this finding (Lee et al., 2012), made use of gene set enrichment analysis to convert the gene expression profile of

specific cancers (non-small cell lung cancer and triple-negative breast cancer) into related signaling pathways. The data about the linear drug superposition combinations was combined with the disease pathways data to obtain the optimal drug combination. Through this method (Lee et al., 2012), found two combination drug pairs with a synergistic effect on lung cancer cells. However, this method still has a number of shortcomings since it ignores the relationship between pathways. Moreover, the theory of linear superposition does not fit all kinds of protein. Because drugs acting on the same pathway through different targets or drugs regulating a relatively small number of highly-connected pathways are more likely to produce synergistic effects (Chen et al., 2016), proposed a "pathway to pathway interaction" network model to predict the therapeutic effect of synergistic drug combinations. This model resulted in an area under the curve (AUC) of a receiver operating characteristic curve of 0.75. The method proposed by (Chen et al., 2016) still has some shortcomings. This method ignores the disease condition, and only the pathway associations of gene overlap are retained, while the pathway associations of protein interactions and function associations are discarded. In addition, the drug combinations are evaluated based on the shortest path without considering the global topology features². Therefore (Cheng et al., 2019) quantified the network-based relationship between drug targets and the diseased human protein to protein interaction. Although this method revealed the existence of six distinct potential drug combinations, only one of these six drug combinations correlated with therapeutic effects. Eventually, a beneficial therapeutic effect was noted when the drug targets hit the same disease module located in separate neighborhoods. Still, the application of this model is limited as it ignores the pathway information and uses the shortest path to evaluate the optimal drug combinations without considering the global topology features.

In view of this, we constructed a Pathway Interaction Network (PINet) model to overcome the limitations of the models described in previous studies (Table 1). This new model abstracts the human body as a two-layer network containing gene and pathway information and describes the influence of a disease or drug on the human as a probability distribution in the network, which is called "disease state" and "drug state." In addition, it predicts the optimal drug combinations by combining "disease state" and "drug state".

The main advantage of the PINet model over the other models is that it can evaluate 5-drug combinations, while most models can only evaluate 2-drug combinations. In addition, PINet is also sensitive to various diseases.

² The regulatory distance of upstream targets to downstream targets may exceed the shortest path (usually 3).

TABLE 1 Optimization of previous research.

Inadequacies of predecessors	Improvement measures
Ignore global topology features Chen et al. (2016) , Cheng et al. (2019)	Analyzing networks using RWR
Ignore pathway information Yang et al. (2008) , Cheng et al. (2019)	Building a two-layer heterogeneous network
It is difficult for users to select indicators Yang et al. (2008)	Redefine disease states without user selection
Only applicable to 1 or 2 diseases Yang et al. (2008) , Lee et al. (2012) Chen et al. (2016) , Cheng et al. (2019)	The new model incorporated multiple diseases and the sensitivity of the specific disease was validated

TABLE 2 Data source.

Data	Number	Source
Pathway	345	KEGG
Gene	18,532	STRING, KEGG, HVIDB, DrugBank, BindingDB, CTD
Drug	6,259	DrugBank, BindingDB
Disease	8	CTD, KEGG
Pathway-pathway	1,659	KEGG
Pathway-gene	34,426	KEGG
Gene-gene	5,680,317	STRING, HVIDB
Drug-gene	39,805	DrugBank, BindingDB
Drug-pathway	57,067	KEGG enrichment analysis
Disease-gene	683	CTD
Disease-pathway	10	KEGG
Drug-disease	257	Clinical guidelines (Table 3)

TABLE 3 Disease-specific drug combinations.

Disease	Drug combinations	Clinical guidelines	References
acquired immunodeficiency syndrome (AIDS)	13	Office of AIDS Research Advisory Council (OARAC)	https://clinicalinfo.hiv.gov/en/guidelines/adult-and-adolescent-arv
inflammatory bowel disease (IBD)	34	The American Gastroenterological Association (AGA)	Terdiman et al. (2013) , Ko et al. (2019) , Feuerstein et al. (2020) , Feuerstein et al. (2021)
Diabetes*	32	the American Diabetes Association (ADA)	American Diabetes (2021)
Atherosclerosis	63	the American College of Cardiology (ACC)	Grundy et al. (2019) , Kumbhani et al. (2021) , Virani et al. (2021)
acute myeloid leukemia (AML)	25	The National Comprehensive Cancer Network (NCCN)	https://www.nccn.org/guidelines/category_1
Breast cancer	60		
Non-small cell lung cancer (NSCLC)	30		

Diabetes including type 1 and type 2 diabetes.

2 Dataset

PINet is composed of four types of entities and eight types of relationships³. The four types of entities include pathways, genes,

drugs and diseases, while the eight types of interactions include pathway to pathway, pathway to gene, gene to gene, drug to gene, disease to gene, disease to pathway, drug to disease and drug to pathway. Except for drug to disease, other data come from databases ([Table 2](#)). The specific data cleaning and processing methods are described in the [Supplementary Material S1.1](#); [Supplementary Material S1.2](#).

Databases include KEGG ([Kanehisa et al., 2021](#)), STRING ([Szkłarczyk et al., 2021](#)), DrugBank ([Wishart et al., 2018](#)),

³ Relationship between drugs is predicted by the model. So it does not appear in the model. We assume that the patient has only one disease, so the relationship between diseases does not exist in the model.

BindingDB (Gilson et al., 2016), CTD (Davis et al., 2021) and HVIDB (Yang et al., 2021).

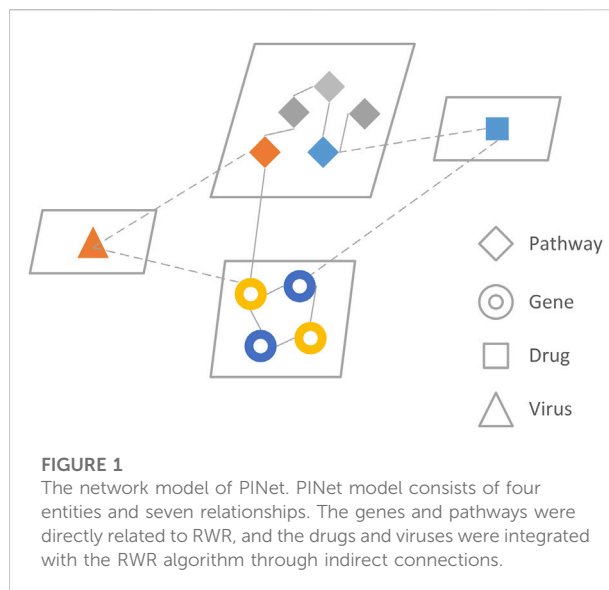
3 Methods

3.1 The theoretical basis of the model

The theoretical basis of the model was built based on the findings of four studies (Yang et al., 2008). Showed that perturbing the targets can shift the disease state to the normal state. Based on this study, we introduce the probability distribution of different drugs or diseases in the network as drug states or disease states and the higher degree of overlap between the drug state and the disease state, the better the efficacy of the drug. Chen et al. (2016). showed that the effect of a disease or drug on the body is achieved through the manipulation of genetic pathways. Therefore, our model included information on the genes and pathways. We also made use of the work of Geva-Zatorsky et al. (2010), which simplified the drug combinations as a linear summation of drug targets. The targets of drug A within our model were denoted as (a_1 , a_2 , and a_3), and the targets of drug B were denoted as (b_1 , b_2). Based on the study of (Geva-Zatorsky et al., 2010), the drug state of the combination of drugs A and B was deemed to be equivalent to the drug state of the virtual drug V, of which targets are (a_1 , a_2 , a_3 , b_1 , b_2). Finally, to narrow down the scope of potential drug combinations and reduce the computational power costs, we used the research of Cheng et al. (2019), which demonstrated that drug synergy is more likely to occur when the drugs act on different disease targets at the same time.

3.2 Construct network model

PINet consists of seven networks⁴ (pathway to pathway, gene to gene, pathway to gene, drug to gene, drug to pathway, disease to pathway, disease to gene), each stored in an adjacency matrix (Figure 1). The main part of the PINet model was based on the restart random walks (RWR) algorithm built on the pathway to pathway, gene to gene, and pathway to gene networks. Further details about the model constructions are provided in the Supplementary Material S1.3.



3.3 Capture state

The effect of a drug or disease on the body can be represented by a vector that contains both pathway and genetic information, which is called a drug state or disease state. These two states were obtained by selecting specific initial nodes on the model to perform the RWR, and the stable probability distribution was defined as the drug or disease state. The specific state capture is described in more detail in the Supplementary Material S2.

3.3.1 Random walk with restart

Biological systems can be simplified into heterogeneous networks, and the RWR algorithm is widely used in the analysis of heterogeneous networks (Cho et al., 2016) (Luo et al., 2017). The RWR algorithm was developed by determining the initial probability, the transition matrix, and the stable probability distribution threshold as follows. More detail about the RWR algorithm is available in the Supplementary Material S2.

3.3.1.1 Determination of the initial probability

The initial nodes were composed of disease or drug-related genes and pathways. The initial probability in a specific network was composed of the initial gene to gene and pathway to pathway networks and can be calculated according to a specific node. For example, in the case of influenza, the initial gene was associated with influenza, and the initial pathway path: hsa05164 was identified from the KEGG database and was fixed to 1. On the other hand, for a drug, the original gene was considered as the drug target, the initial pathway was identified through pathway enrichment analysis, and the number of potential initial pathways was not fixed.

⁴ The drug-disease relationship is the data used to evaluate the model. So it doesn't appear in the model.

The initial probability of the pathway to pathway network a_0 was formed so that equal probabilities were assigned to the initial nodes in the pathway to pathway network, and the sum of the nodes' probabilities was equal to 1. Therefore if the probabilities of non-initial nodes are 0, then the initial probability of the gene to gene network b_0 is the same. This relationship is summarized by the equation.

$$p_0 = 0.5 \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} \quad (1)$$

Whereby a_0 is the pathway initial probability, and b_0 is the gene initial probability. Both a_0 and b_0 are vectors.

3.3.1.2 Determination of the transition matrix

The transition matrix describes the transition characteristics of all nodes within the network model. There are four transfer modes in PINet: pathway to pathway, pathway to gene, gene to gene, and gene to the pathway. Each transfer mode requires a transition matrix. The description of the PINet transition node requires a large transition matrix M composed of four small transition matrices M_i .

The (t) th probability distribution was obtained by mapping the $(t-1)$ th probability distribution through the transition matrix as follows:

$$(1-r) \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix} \begin{bmatrix} a_t \\ b_t \end{bmatrix} + r p_0 = \begin{bmatrix} a_{t+1} \\ b_{t+1} \end{bmatrix} = p_{t+1} \quad (2)$$

Whereby $M1$ is the pathway to pathway, $M2$ is the gene to pathway, $M3$ is the pathway to gene, and $M4$ is the gene to gene. r is the restart probability which is generally equal to 0.5.

3.3.1.3 Determination of the stable probability distribution threshold

The initial node was selected to perform the RWR. As the number of iterations increased, the probability distribution gradually became stable. When the difference in the probability distribution between the (n) th and the $(n+1)$ th was less than the given threshold, the (n) th probability distribution was considered to be a stable probability distribution, and the threshold was generally set to 10^{-10} .

3.3.2 Capturing the disease state

The disease state was then captured through the identification of the initial nodes of the disease in the pathway to pathway network and, subsequently, the gene-gene network. The initial probability p_0 of the disease was constructed, and then RWR was performed until the probability distribution became stable. The stable probability of the disease site p_n was then captured for the disease state.

3.3.3 Capturing the drug state

The drug state was captured through the identification of the virtual drug corresponding to the drug combination. The

initial probability p_0 of the drug was determined according to the target and enrichment pathway of the virtual drug. Finally, RWR was performed until the probability distribution became stable, and the stable probability p_n was captured for the drug state.

3.4 The drug combination score

Since the drug combinations have certain indications, we evaluated the drug combinations under specific disease conditions by "drug state" and "disease state." The same drug combinations have different scores on different disease conditions in PINet. The absolute drug score value was obtained by calculating the difference between the "drug state" and the "disease state".

$$score = |S_{di} - S_{dr}| \quad (3)$$

S_{di} is the disease state, S_{dr} is the drug state.

A lower score indicates a higher likelihood of a synergistic drug combination. Further details on the calculation of the drug combination score can be found in [Supplementary Material S3](#).

3.5 Evaluation of pathway interaction network

During the development of PINet, it was assumed that the drug combination contained two types of information: the drug composition and the indication. Therefore two tests were performed to evaluate the sensitivity of PINet to detect disease and drug quantity. The disease sensitivity analysis assessed whether PINet can correctly identify the indications for the different drug combinations. For example, whether PINet will wrongly judge a drug designed to treat AIDS as a drug used to treat cancer. The drug quantity sensitivity analysis evaluated the ability of PINet to identify the n -drugs combination ($n = 2, 3, 4$, and 5).

3.5.1 Disease sensitivity

The drug combination highlighted in the clinical guidelines of each disease was regarded as the positive gold standard treatment. The clinical indications of the drug combinations used to manage a specific disease were then modified to represent a negative example, i.e., another disease. All positive and negative examples were entered into the PINet for scoring, and the AUC under the ROC was calculated for each example. An AUC below 0.5 indicates that the PINet model was not sensitive enough to detect the disease and corresponding drug combinations, and these were therefore excluded from the model. The remaining diseases and drug combinations in the clinical guidelines were evaluated again in the next step.

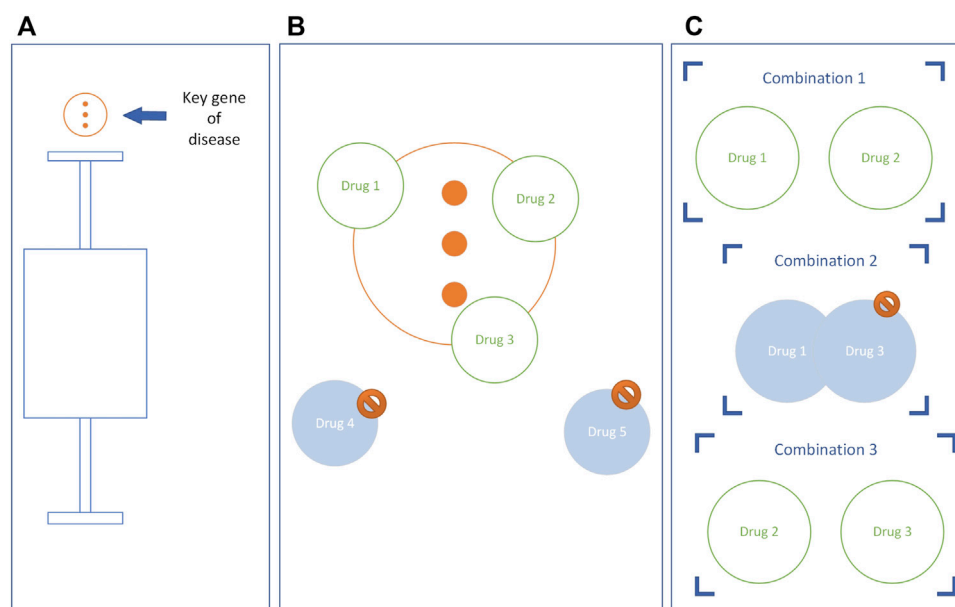


FIGURE 2

The construction of the potential drug combinations. Taking the key genes of diseases as an example, the key pathways are the same. **(A)** Genes above the upper limit are key genes. **(B)** Eliminate drugs that do not have intersections with key disease genes. **(C)** A drug combination is constructed, and if the drugs in the combination have the same target, the combination is eliminated.

3.5.2 Drug quantity sensitivity

The drug combinations may include four possible options with 2, 3, 4, or 5 drugs. The sensitivity of PINet to different drug combinations was calculated as follows. First, the drug combination in the clinical guidelines was used as a positive example, and the randomly generated drug combination was used as a negative example. Subsequently, the drug status and disease status were calculated according to the drug composition and indications, respectively, as explained in Section 3.3. Then, the score for each drug combination was calculated, as explained in Section 3.4. Finally, based on the calculated score, the AUC was calculated for each drug combination.

3.6 Prediction of the drug combinations

3.6.1 Primary potential drug combination

Outliers of disease state are identified by Quartile, and these outliers are key genes and key pathways of the disease. The potential drugs were selected if the target of the drug had an intersection with the key gene of the disease and the enriched pathway of the drug had an intersection with the key pathway of the disease. We assumed that for N potential drugs, there are C_N^i primary potential drug combinations (i is the number of drugs in the drug combination. Refer to Figures 2A–C). More detail about Quartile is available in the Supplementary Material S4.

3.6.2 Secondary potential drug combinations

The drug combinations with overlapping drug targets were removed from the primary potential drug combination to obtain the secondary potential drug combination (Figure 2C).

3.6.3 Evaluation of the potential drug combinations

To improve the prediction accuracy of the model, we used the score corresponding to the false positive rate of 10% on the ROC of the “Drug quantity sensitivity” as the threshold. The scores of the secondary potential drug combinations were calculated, and those below the threshold were classified as synergistic drug combinations.

4 Results

4.1 Disease sensitivity

The PINet had a high sensitivity for NSCLC, AML, breast cancer, and IBD and low sensitivity for diabetes type 1, diabetes type 2, AIDS, and atherosclerosis (Figure 3).

4.2 Drug quantity sensitivity

Figure 4 illustrates the drug quantity sensitivity after excluding the diseases with a low PINet sensitivity. The

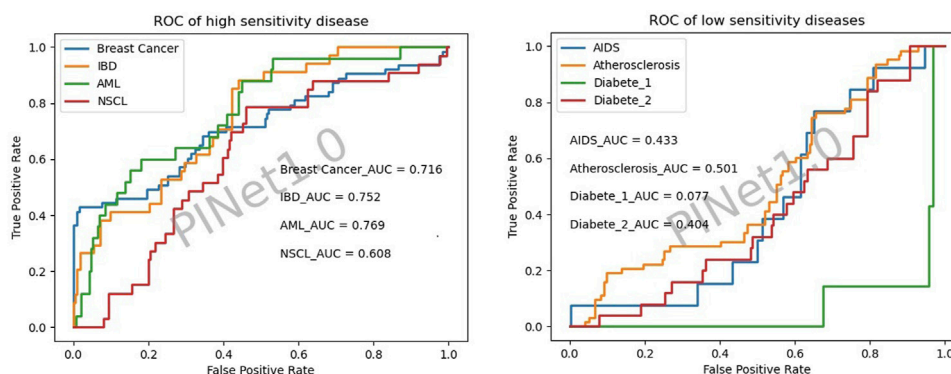


FIGURE 3
Disease sensitivity of PINet.

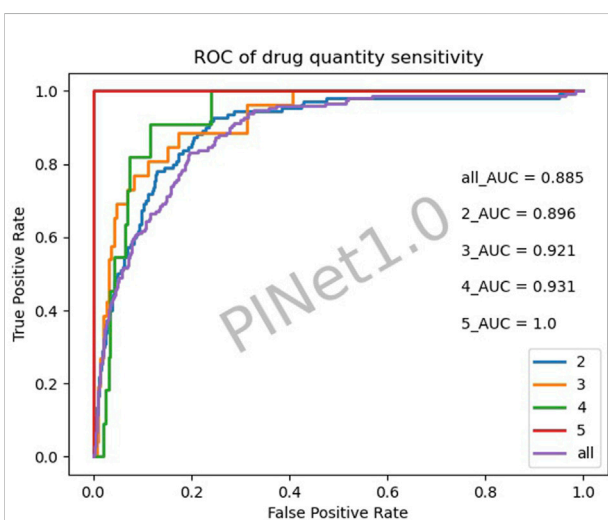


FIGURE 4
Drug quantity sensitivity of PINet.

sensitivity of PINet increased as the order of drug combinations increased. PINet also achieved good results in the identification of high-order drug combinations. However, since the sample was too small (2 positive cases and 58 negative cases in the fifth-order drug combination), the ROC may not be accurate.

4.3 Prediction accuracy

Since PINet had the highest sensitivity for predicting AML, we decided to use PINet to predict the optimal drug combinations for this disease. PINet was first used to identify the key genes and pathways of AML. Subsequently, the drugs based on these genes and pathways were identified and used to construct the primary drug combinations. This revealed a total of

26,106 possible primary drug combinations. The drug combinations with the same target were eliminated, and the remaining drug combinations ($n = 17,713$) were scored to identify the optimal drug combinations ($n = 2,590$). After excluding the unapproved drugs, 1,221 possible drug combinations were identified. The efficacy of two of the drug combinations identified by PINet has been validated in clinical trials or *in vivo* studies. Röllig et al. (2021) demonstrated the synergy between gemtuzumab ozogamicin and midostaurin in newly diagnosed AML in a phase-I clinical trial. Tian et al. (2018) found that Emricasan and Ponatinib can synergistically reduce ischemia-reperfusion injury in rat brains.

5 Discussion

As the development of new drugs continues to increase, there is a need to develop novel methods to identify optimal drug combinations for managing specific diseases. In this study, we proposed a novel model PINet to make it easier for clinicians to identify optimal drug combinations. When compared with other machine learning models, PINet has several advantages and limitations.

5.1 Advantages of pathway interaction network

5.1.1 Interpretability

PINet is a theory-driven method for evaluating drug combinations based on the assumption that “drugs can correct disease states.” A low PINet score means that the drug combination is more applicable to a specific disease. This simple scoring system used in PINet is easily understood by researchers in the non-data science fields, making PINet easy to generalize.

TABLE 4 Comparison of different models.

	Yang et al. (2008)	Lee et al. (2012)	Chen et al. (2016)	Cheng et al. (2019)	PINet1.0
Indications ^{aa}	inflammation	NSCLC; TNBC	\	hypertension	Breast cancer; NSCLC; AML; IBD
order of drug combination ^{bb}	2	2	2	2	≥2
drug range ^{cc}	++	++	+++	+++	+++

aa: Applicable diseases of the model. bb: The number of drugs in a specific drug combination. cc: Drugs within the model. (Yang et al. (2008) only considered targets and ignored the multi-target phenomenon of drugs. Lee et al. (2012)'s drug relied on transcriptome data). NSCLC, non-small cell lung cancer; TNBC, triple-negative breast cancer; AML, acute myeloid leukemia; IBD, inflammatory bowel disease.

5.1.2 Non-training set dependency

Unlike machine learning, there is no need to fit all parameters in PINet, and therefore, PINet does not require a training dataset. This is crucial for drug combination prediction for some diseases that lack a training dataset.

5.1.3 High-order drug combinations

Most drug combination prediction models focus on 2-drug combinations since high-order drug combinations are computationally expensive to calculate. PINet takes the same time to evaluate 2-drug combinations as higher-order drug combinations by narrowing the range of candidate drugs based on theory to maintain the computational power consumption within an acceptable range.

5.1.4 Applicable to multiple diseases

A variety of diseases are already included in PINet, and the model's effectiveness in predicting optimal drug combinations in breast cancer, IBD, AML, and NSCL has already been verified. With the advancement of disease pathway research in KEGG, the applicability of PINet will be extended to more diseases.

5.2 Disadvantages of pathway interaction network

5.2.1 Poor sensitivity to some diseases

The sensitivity of PINet in some diseases, such as AIDS and diabetes, was found to be low in our study. A possible explanation for this could be that the effect of these diseases on genes is expressed as either an up-regulation or down-regulation gene expression. However, PINet simplifies the relationship between diseases and genes to 0 or 1, resulting in the loss of information. Furthermore, most anti-infective drugs target pathogens, and the targets of these drugs do not have corresponding genes in KEGG.

5.2.2 Drug antagonism is not considered

The drug-to-target relationship was simplified to 0 or 1, and the antagonist effects of drug combinations were not considered when assessing the drug sensitivity on PINet. This means that PINet cannot distinguish between synergy and antagonism. Although we avoided competitive antagonism by narrowing

down the drug candidates, this does not solve the problem on a theoretical level.

5.2.3 Poor validation

The validation of PINet is not sufficient for the following reasons: Various theoretical models are suitable for different diseases, and there are certain differences in the range of drugs that can be selected, so it is difficult to make an objective comparison (Table 4). In fact, the drug combinations in PINet 1.0 are all derived from clinical guidelines, and many of these drugs lack transcriptome data and cannot be evaluated by the method of (Lee et al., 2012). There are differences between other methods (Cheng et al., 2019) (Chen et al., 2016) (Yang et al., 2008) and PINet1.0 in the indication, which makes it impossible to compare. On the other hand, due to a lack of experimental conditions, it was not possible to validate the accuracy of the PINet predictions.

5.3 Recommendations for future practice

Several aspects can be improved on PINet to increase its prediction accuracy and applicability.

5.3.1 Differentiate between synergies and indications for drug combinations

In PINet, we evaluate drug combinations by comparing disease states and drug states, considering both synergy and indications of the drug combination together. First, we found that PINet has moderate disease sensitivity but can accurately distinguish synergistic drug combinations from random drug combinations, during the evaluation of the model. In addition, the combination of drugs predicted to treat AML is suitable for ischemia-reperfusion injury, which may be related to the multi-targets phenomenon of drugs and multi-phenotypes phenomenon of diseases (Tian et al., 2018). Furthermore, synergy was identified by relying only on the shortest path in the pathway network without disease information (Chen et al., 2016). Based on the above facts, we suggest that synergy and indication should be two relatively independent attributes of a drug combination and these attributes are relatively independent and may provide a new theoretical basis for the development of a

repository for the rapid identification of drug combinations. If the conjecture is correct, PINet could be used in the future to evaluate drug combinations independently of the disease state, eventually increasing the scope of application of the model. As a result, the indications can be isolated and analyzed separately in finer divisions according to the drug function (e.g., anti-inflammatory, or anti-viral) rather than the entire disease.

We plan to elucidate the synergistic effect of drug combinations through information theory. This will enable us to locate key pathways and key genes to define the indications of drug combinations and verify whether the conjecture is correct.

5.3.2 Increase disease sensitivity

The relationship between diseases and genes can be optimized as -1 , 0 , and 1 to achieve differentiation of different diseases, thereby improving the disease sensitivity of PINet.

5.3.3 Identify antagonism

The drug-to-target relationship can also be optimized to -1 , 0 , and 1 to simulate the antagonistic relationship between drugs. In follow-up studies, we will additionally evaluate the ability of PINet to identify antagonistic drug combinations. Chen et al., 2012, Hopkins, 2008, Hsieh et al., 2021, Zhang et al., 2021.

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

- A D A (2021). 9. Pharmacologic approaches to glycemic treatment: Standards of medical care in diabetes-2021. *Diabetes Care* 44 (1), S111–S124. doi:10.2337/dc21-S009
- Bayat Mokhtari, R., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., et al. (2017). Combination therapy in combating cancer. *Oncotarget* 8 (23), 38022–38043. doi:10.18632/oncotarget.16723
- Chen, D., Zhang, H., Lu, P., Liu, X., and Cao, H. (2016). Synergy evaluation by a pathway-pathway interaction network: A new way to predict drug combination. *Mol. Biosyst.* 12 (2), 614–623. doi:10.1039/c5mb00599j
- Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* 8 (7), 1970–1978. doi:10.1039/c2mb00002d
- Cheng, F., Kovacs, I. A., and Barabasi, A. L. (2019). Network-based prediction of drug combinations. *Nat. Commun.* 10 (1), 1197. doi:10.1038/s41467-019-09186-x
- Cho, H., Berger, B., and Peng, J. (2016). Compact integration of multi-network topology for functional analysis of genes. *Cell Syst.* 3 (6), 540–548. e545. doi:10.1016/j.cels.2016.10.017
- Cihlar, T., and Fordyce, M. (2016). Current status and prospects of HIV treatment. *Curr. Opin. Virol.* 18, 50–56. doi:10.1016/j.coviro.2016.03.004
- Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegiers, J., Wiegiers, T. C., et al. (2021). Comparative toxicogenomics database (CTD): Update 2021. *Nucleic Acids Res.* 49 (D1), D1138–D1143. doi:10.1093/nar/gkaa891
- Feuerstein, J. D., Ho, E. Y., Shmidt, E., Singh, H., Falck-Ytter, Y., Sultan, S., et al. (2021). AGA clinical practice guidelines on the medical management of moderate to severe luminal and perianal fistulizing crohn's disease. *Gastroenterology* 160 (7), 2496–2508. doi:10.1053/j.gastro.2021.04.022
- Feuerstein, J. D., Isaacs, K. L., Schneider, Y., Siddique, S. M., Falck-Ytter, Y., Singh, S., et al. (2020). AGA clinical practice guidelines on the management of moderate to severe ulcerative colitis. *Gastroenterology* 158 (5), 1450–1461. doi:10.1053/j.gastro.2020.01.006
- Geva-Zatorsky, N., Dekel, E., Cohen, A. A., Danon, T., Cohen, L., and Alon, U. (2010). Protein dynamics in drug combinations: A linear superposition of individual-drug responses. *Cell* 140 (5), 643–651. doi:10.1016/j.cell.2010.02.011
- 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 (D1), D1045–D1053. doi:10.1093/nar/gkv1072
- Grundey, S. M., Stone, N. J., Bailey, A. L., Beam, C., Birtcher, K. K., Blumenthal, R. S., et al. (2019). 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APHA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: A report of the American college of cardiology/American heart association task force on clinical practice guidelines. *J. Am. Coll. Cardiol.* 73 (24), e285–e350. doi:10.1016/j.jacc.2018.11.003
- Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* 4 (11), 682–690. doi:10.1038/nchembio.118

Author contributions

YH was responsible for the theoretical design and code reproduction. DC took part in data cleaning and model building. MZ was involved in the theoretical design, YJ performed the data collection, and YZ performed the feasibility analysis. All authors have read and agreed to the published version of the manuscript.

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2022.971768/full#supplementary-material>

- Hsieh, K., Wang, Y., Chen, L., Zhao, Z., Savitz, S., Jiang, X., et al. (2021). Drug repurposing for COVID-19 using graph neural network and harmonizing multiple evidence. *Sci. Rep.* 11 (1). doi:10.1038/s41598-021-02353-5
- Huffman, M. D., Xavier, D., and Perel, P. (2017). Uses of polypills for cardiovascular disease and evidence to date. *Lancet* 389 (10073), 1055–1065. doi:10.1016/s0140-6736(17)30553-6
- Jafari, M., Mirzaie, M., Bao, J., Barneh, F., Zheng, S., Eriksson, J., et al. (2022). Bipartite network models to design combination therapies in acute myeloid leukaemia. *Nat. Commun.* 13 (1), 2128. doi:10.1038/s41467-022-29793-5
- Jia, J., Zhu, F., Ma, X., Cao, Z., Cao, Z. W., Li, Y., et al. (2009). Mechanisms of drug combinations: Interaction and network perspectives. *Nat. Rev. Drug Discov.* 8 (2), 111–128. doi:10.1038/nrd2683
- Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). Kegg: Integrating viruses and cellular organisms. *Nucleic Acids Res.* 49 (D1), D545–D551. doi:10.1093/nar/gkaa970
- Kim, Y., Zheng, S., Tang, J., Jim Zheng, W., Li, Z., and Jiang, X. (2021). Anticancer drug synergy prediction in understudied tissues using transfer learning. *J. Am. Med. Inf. Assoc.* 28 (1), 42–51. doi:10.1093/jamia/ocaa212
- Ko, C. W., Singh, S., Feuerstein, J. D., Falck-Ytter, C., Falck-Ytter, Y., Cross, R. K., et al. (2019). AGA clinical practice guidelines on the management of mild-to-moderate ulcerative colitis. *Gastroenterology* 156 (3), 748–764. doi:10.1053/j.gastro.2018.12.009
- Kumbhani, D. J., Cannon, C. P., Beavers, C. J., Bhatt, D. L., Cuker, A., Gluckman, T. J., et al. (2021). 2020 acc expert consensus decision pathway for anticoagulant and antiplatelet therapy in patients with atrial fibrillation or venous thromboembolism undergoing percutaneous coronary intervention or with atherosclerotic cardiovascular disease: A report of the American college of cardiology solution set oversight committee. *J. Am. Coll. Cardiol.* 77 (5), 629–658. doi:10.1016/j.jacc.2020.09.011
- Lee, J. H., Kim, D. G., Bae, T. J., Rho, K., Kim, J. T., Lee, J. J., et al. (2012). Cda: Combinatorial drug discovery using transcriptional response modules. *PLoS One* 7 (8). doi:10.1371/journal.pone.0042573
- Li, J., Tong, X. Y., Zhu, L. D., and Zhang, H. Y. (2020). A machine learning method for drug combination prediction. *Front. Genet.* 11, 1000. doi:10.3389/fgene.2020.01000
- Liu, Y., Tong, Z., Shi, J., Li, R., Upton, M., and Wang, Z. (2021). Drug repurposing for next-generation combination therapies against multidrug-resistant bacteria. *Theranostics* 11 (10), 4910–4928. doi:10.7150/thno.56205
- Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8 (1), 573. doi:10.1038/s41467-017-00680-8
- Röllig, C., Schliemann, C., Mikesch, J.-H., Fransecky, L., Baldus, C. D., Heydrich, B.-N., et al. (2021). Gemtuzumab ozogamicin plus midostaurin in combination with standard intensive induction therapy in newly diagnosed AML: Results from a phase-I study. *Blood* 138 (1), 2324. doi:10.1182/blood-2021-150069
- Shi, J. Y., Huang, H., Li, J. X., Lei, P., Zhang, Y. N., Dong, K., et al. (2018). Tmfuf: A triple matrix factorization-based unified framework for predicting comprehensive drug-drug interactions of new drugs. *BMC Bioinforma.* 19 (14), 411. doi:10.1186/s12859-018-2379-8
- Shinn, P., Chen, L., Ferrer, M., Itkin, Z., Klumpp-Thomas, C., McKnight, C., et al. (2019). High-throughput screening for drug combinations. *Methods Mol. Biol.* 1939, 11–35. doi:10.1007/978-1-4939-9089-4_2
- Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49 (D1), D605–D612. doi:10.1093/nar/gkaa1074
- Terdiman, J. P., Gruss, C. B., Heidelbaugh, J. J., Sultan, S., Falck-Ytter, Y. T., Practice, A. G. A. I. C., et al. (2013). American Gastroenterological Association Institute guideline on the use of thiopurines, methotrexate, and anti-TNF-alpha biologic drugs for the induction and maintenance of remission in inflammatory Crohn's disease. *Gastroenterology* 145 (6), 1459–1463. doi:10.1053/j.gastro.2013.10.047
- Tian, J., Guo, S., Chen, H., Peng, J. J., Jia, M. M., Li, N. S., et al. (2018). Combination of emricasan with Ponatinib synergistically reduces ischemia/reperfusion injury in rat brain through simultaneous prevention of apoptosis and necroptosis. *Transl. Stroke Res.* 9 (4), 382–392. doi:10.1007/s12975-017-0581-z
- Virani, S. S., Morris, P. B., Agarwala, A., Ballantyne, C. M., Birtcher, K. K., Kris-Etherton, P. M., et al. (2021). 2021 acc expert consensus decision pathway on the management of ascvd risk reduction in patients with persistent hypertriglyceridemia: A report of the American college of cardiology solution set oversight committee. *J. Am. Coll. Cardiol.* 78 (9), 960–993. doi:10.1016/j.jacc.2021.06.011
- Wang, Y., Yang, H., Chen, L., Jafari, M., and Tang, J. (2021). Network-based modeling of herb combinations in traditional Chinese medicine. *Brief. Bioinform.* 22 (5), bbab106. doi:10.1093/bib/bbab106
- 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
- Yang, K., Bai, H., Ouyang, Q., Lai, L., and Tang, C. (2008). Finding multiple target optimal intervention in disease-related molecular network. *Mol. Syst. Biol.* 4, 228. doi:10.1038/msb.2008.60
- Yang, X., Lian, X., Fu, C., Wuchty, S., Yang, S., and Zhang, Z. (2021). Hvidb: A comprehensive database for human-virus protein-protein interactions. *Brief. Bioinform.* 22 (2), 832–844. doi:10.1093/bib/bbaa425
- Zagidullin, B., Wang, Z., Guan, Y., Pitkanen, E., and Tang, J. (2021). Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Brief. Bioinform.* 22 (6), bbab291. doi:10.1093/bib/bbab291
- Zhang, T., Zhang, L., Payne, P. R. O., and Li, F. (2021). "Synergistic drug combination prediction by integrating multiomics data in deep learning models," in *Translational bioinformatics for therapeutic development*. Editor J. Markowitz New York, NY, USA: Springer US, 223–238.



OPEN ACCESS

EDITED BY

Irina Sousa Moreira,
University of Coimbra, Portugal

REVIEWED BY

Panagiotis Alexiou,
Central European Institute of
Technology (CEITEC), Czechia
Congshan Jiang,
Xi'an Children's Hospital, China

*CORRESPONDENCE

Jiuzhen Liang,
jzliang@cczu.edu.cn

SPECIALTY SECTION

This article was submitted
to Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 05 August 2022

ACCEPTED 03 November 2022

PUBLISHED 23 November 2022

CITATION

Ni J, Cheng XL, Ni TG and Liang JZ
(2022), Identifying SM-miRNA
associations based on layer attention
graph convolutional network and
matrix decomposition.
Front. Mol. Biosci. 9:1009099.
doi: 10.3389/fmolb.2022.1009099

COPYRIGHT

© 2022 Ni, Cheng, Ni and Liang. 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.

Identifying SM-miRNA associations based on layer attention graph convolutional network and matrix decomposition

Jie Ni, Xiaolong Cheng, Tongguang Ni and Jiuzhen Liang*

School of Computer Science and Artificial Intelligence and Aliyun School of Big Data and School of Software, Changzhou University, Changzhou, China

The accurate prediction of potential associations between microRNAs (miRNAs) and small molecule (SM) drugs can enhance our knowledge of how SM cures endogenous miRNA-related diseases. Given that traditional methods for predicting SM-miRNA associations are time-consuming and arduous, a number of computational models have been proposed to anticipate the potential SM-miRNA associations. However, several of these strategies failed to eliminate noise from the known SM-miRNA association information or failed to prioritize the most significant known SM-miRNA associations. Therefore, we proposed a model of Graph Convolutional Network with Layer Attention mechanism for SM-MiRNA Association prediction (GCNLASMMMA). Firstly, we obtained the new SM-miRNA associations by matrix decomposition. The new SM-miRNA associations, as well as the integrated SM similarity and miRNA similarity were subsequently incorporated into a heterogeneous network. Finally, a graph convolutional network with an attention mechanism was used to compute the reconstructed SM-miRNA association matrix. Furthermore, four types of cross validations and two types of case studies were performed to assess the performance of GCNLASMMMA. In cross validation, global Leave-One-Out Cross Validation (LOOCV), miRNA-fixed LOOCV, SM-fixed LOOCV and 5-fold cross-validation achieved excellent performance. Numerous hypothesized associations in case studies were confirmed by experimental literatures. All of these results confirmed that GCNLASMMMA is a trustworthy association inference method.

KEYWORDS

microRNA, small molecule, deep learning, association prediction, matrix decomposition

1 Introduction

As a form of non-coding RNA (ncRNA), MicroRNA (miRNA), is roughly 22 nucleotides in length (Bartel, 2004; Hammond, 2015; Lu and Rothenberg, 2018). Lin-4 was the first human miRNA identified in 1993 by Lee *et al.* in *Caenorhabditis elegans* (Lee *et al.*, 1993; Wightman *et al.*, 1993). With the advent of high-throughput sequencing technologies, an increasing number of miRNAs with important functions in human gene expression have been identified (Denzler *et al.*, 2016; Tagliaferro *et al.*, 2017; Thomou *et al.*, 2017; Gam *et al.*, 2018; Ghini *et al.*, 2018; Liu *et al.*, 2018). Specifically, miRNAs can attach to the 3' UnTranslated Region (3' UTR) of target messenger RNAs (mRNAs) via base-pairing to control the degradation of target mRNAs and limit the translation of target mRNAs, hence regulating gene expression (Gorbea *et al.*, 2017). In the control of target mRNA gene expression by miRNA, one miRNA may regulate many target mRNAs, or numerous miRNAs regulate one target mRNA (Saikia *et al.*, 2020; Iwata *et al.*, 2021; Zhong *et al.*, 2021). Several studies demonstrated the role of miRNAs in the maturation of immune cells (Kumar Kingsley and Vishnu Bhat, 2017). Since the profound impact of miRNAs on biological development became apparent, numerous miRNA types have been identified to be involved in biological evolutionary processes (Rupaimoole and Slack, 2017; Cristino *et al.*, 2019).

Small Molecule (SM) drugs are mostly composed of molecules with molecular weights typically fewer than 1,000 g/mol. More than 98 percent of today's drugs are SMs (Geng and Craig, 2021). The development of SMs that target miRNAs is a current trend in drug research (Dai and Tan, 2015; Yu *et al.*, 2020). In previous drug development, protein enzymes and receptors were typically employed as therapeutic targets. Over 80 percent of drug development was intimately tied to protein enzymes and receptors (Deyle *et al.*, 2017; Yekkirala *et al.*, 2017; Nair *et al.*, 2018; Lai-Kwon *et al.*, 2021). In recent years, more scientific experiments have proven inextricable linkages between SMs and miRNAs (Healy *et al.*, 2012; Monroig *et al.*, 2015; Haniff *et al.*, 2021). When miRNAs fail to regulate the gene expression of an organism, specific disorders such as cardiovascular diseases, neurological diseases and cancers may develop (Kumari *et al.*, 2018; Xia *et al.*, 2019; Dragomir *et al.*, 2021). In addition, SMs are effective in regulating miRNA dysregulation to treat linked endogenous disorders, and numerous SMs have been created for clinical therapy of these diseases (Dragomir *et al.*, 2021).

The development of novel SMs is facilitated by the accurate identification of miRNA-related SMs. Recent studies have focused on discovering possible associations between SMs and miRNAs (Chen *et al.*, 2021; Li *et al.*, 2021; Wang *et al.*, 2021). Early identification approaches used high-throughput screening methods, such as mass spectrometry, fluorescence and reporter genes (Seth *et al.*, 2005; Parsons *et al.*, 2009; Carnevali *et al.*, 2010;

Chen *et al.*, 2012). The most frequent method for discovering potential SM-miRNA associations is the reporter genes. On the basis of the reporter genes, a functional novel drug screening method capable of screening lead compounds was proposed. By substituting biomacromolecules with tiny organic compounds, the screening process for drugs could be expedited dramatically. The use of tiny organic compounds throughout the screening procedure could provide information on the functional responses of cells. (Wen *et al.*, 2015). In drug screening research, luciferase reporter genes satisfy the requirements for high sensitivity, target specificity and high throughput (Thorne *et al.*, 2010).

However, it was discovered that biological screening approaches are stochastic and time-consuming. With the proliferation of bioinformatics databases, the number of known SM-miRNA associations increased, as did the calculational methodologies for SM and miRNA similarity. Consequently, machine learning techniques obtained more precise prediction outcomes (Qu *et al.*, 2019). Bioinformaticians have begun to employ machine learning techniques to predict probable SM-miRNA associations to circumvent time-consuming and labor-intensive biological investigations (Wang and Chen, 2019; Wang *et al.*, 2019).

Among the previous methods for predicting probable SM-miRNA associations, (Qu *et al.*, 2018), developed a model titled Triple Layer Heterogeneous Network based Small Molecule-MiRNA Association prediction (TLHNSMMA). TLHNSMMA first merged the known SM-miRNA associations, SM similarity and miRNA similarity into a three-layer heterogeneous network. The three-layer heterogeneous graph was then implemented with an iterative updating algorithm. Finally, the reconstructed SM-miRNA association matrix was obtained using an iterative propagation approach that made extensive use of global data. Based on the establishment of a three-layer SM-miRNA heterogeneous network, (Liu *et al.*, 2020), suggested a novel model for potential SM-miRNA association prediction called Random Walk with Negative Samples (RWNS). Firstly, RWNS obtained integrated similarities of SM and miRNA. Then, Liu *et al.* devised a Credible Negative Sample extraction method (CNSMiRS) to extract plausible negative SM-miRNA samples under the premise that dissimilar SMs/miRNAs are unlikely to be associated with each other's related miRNAs/SMs. Finally, the reconstructed SM-miRNA association matrix was obtained by implementing a random walk algorithm on the constructed small molecule-disease-miRNA association network. However, the performance of TLHNSMMA and RWNS is dependent on the known SM-miRNA association adjacency matrix. Consequently, (Yin *et al.*, 2019), suggested a model of Sparse Learning and Heterogeneous Graph Inference for Small Molecule-MiRNA Association prediction (SLHGISMMA). Yin *et al.* first used matrix decomposition on known SM-miRNA associations to obtain the new SM-miRNA associations. Then, the new SM-miRNA associations, integrated miRNA similarity and integrated SM similarity were incorporated into a heterogeneous network.

Finally, the reconstructed SM-miRNA association matrix was obtained using heterogeneous graph inference. Chen et al. (2021) recently proposed the Bounded Nuclear Norm Regularization for SM-miRNA Associations prediction (BNNRSMMA), which treated the problem of potential SM-miRNA association prediction as a matrix complementation problem. In addition, BNNRSMMA included a regularization term to remove the negative effects of data noise.

In recent years, improvements have been made to machine learning techniques, and deep learning has emerged as one of the brightest new stars (Wang et al., 2020). Deep learning has achieved exceptional results in traditional classification tasks, such as handwritten font recognition (Singh et al., 2021), computer vision (Borges Oliveira et al., 2021) and computational biology (Angermueller et al., 2016). In addition, deep learning has substantially affected the field of potential association prediction. For example, zeng et al. proposed a computational framework termed AOPEDF based on drug-target network and deep forest algorithm to predict potential drug-target associations (Zeng et al., 2020). AOPEDF attained excellent performance in identifying molecular targets among known drugs on two external validation datasets by comparison to other machine learning methods. Therefore, we proposed a model of Graph Convolutional Network with Layer Attention mechanism for SM-MiRNA Association prediction (GCNLASMMMA). To evaluate the performance of GCNLASMMMA, we used two types of cross validation, namely, 5-fold cross-validation and Leave-One-Out Cross Validation (LOOCV). Additionally, we also utilized two types of case studies to confirm the effectiveness of GCNLASMMMA in identifying potential miRNAs for investigated SMs. The results showed that GCNLASMMMA could accurately and effectively predict the SM-miRNA pairs most likely to be potentially associated.

2 Materials and methods

2.1 SM-miRNA associations

We named two datasets used in our work after dataset1 and dataset2. Eight hundred and thirty-one SMs in dataset1 were downloaded from three databases, namely SM2miR, DrugBank (Knox et al., 2011) and PubChem (Wang et al., 2009). Five hundred and forty-one miRNAs were downloaded from four databases, namely SM2miR, HMDD (Li et al., 2014), miR2Disease (Jiang et al., 2009) and PhenomiR (Ruepp et al., 2010). Six hundred and sixty-four known SM-miRNA associations were downloaded from a database, namely SM2miR V1.0 (Liu et al., 2013). On the basis of dataset1, we removed the SMs and miRNAs that did not constitute any known association. Then, we obtained dataset2 which included 286 different miRNAs, 39 different SMs and 664 known

SM-miRNA association pairs. Specifically, the known SM-miRNA association A_{ij} between the i_{th} SM and the j_{th} miRNA was stored as follows.

2.2 Integration of SM similarities

The integrated SM similarity was calculated by (Lv et al., 2015). In his method, a total of four SM similarities were used, namely SM side effect similarity (Gottlieb et al., 2011), gene functional consistency-based similarity for SMs (Lv et al., 2012), SM chemical structure similarity (Hattori et al., 2003) and disease phenotype-based similarity for SMs (Gottlieb et al., 2011). In Lv's article, the side effect properties of SM were first downloaded from SDe Effect Resource (SIDER) and calculated by Jaccard score to obtain SMs side effect similarities (Gottlieb et al., 2011). The calculation of gene functional consistency-based similarities for SMs was implemented on the target genes of SMs obtained from the DrugBank and Therapeutic Targets Database (TTD) (Liu et al., 2011). The Gene Set Functional Similarity (GSFS) method was given in the previous article (Lv et al., 2012). Specifically, we downloaded the SM chemical structure information. Then, a graph-based method, SIMilar COMpound (SIMCOMP) (Lv et al., 2012), was applied to obtain SMs' chemical structure similarities. Finally, the disease phenotype-based similarities for SMs were obtained by calculating the data downloaded from the DrugBank and TTD with the Jaccard score method.

After obtaining all four SM similarities, we named them after SS1, SS2, SS3 and SS4, respectively. Then, the scores of the four SM similarities were integrated by the following formula,

$$SSM = \frac{\sum_i \alpha_i SS_i}{\sum_i \alpha_i}, (i = 1, 2, 3, 4) \quad (1)$$

where α represents the weights of SM similarities. All of the measures are important in terms of biology. Thus, we set the values of all α to 1, which means that each SM similarity made an equal contribution to constituting the integrated SM similarity (Li et al., 2004). Finally, the integrated SM similarity $SSM(s_i, s_j)$ between the i_{th} and j_{th} SMs was obtained after normalization as follows.

$$SSM(s_i, s_j) = \frac{SSM(s_i, s_j)}{\sqrt{\sum_{l=1}^{ns} SSM(s_i, s_l)} \sqrt{\sum_{l=1}^{ns} SSM(s_l, s_j)}} \quad (2)$$

2.3 Integration of miRNA similarities

Two miRNA similarities, gene function consistency-based similarity (Lv et al., 2012) and indication phenotype-based similarity (Gottlieb et al., 2011), were used to obtain

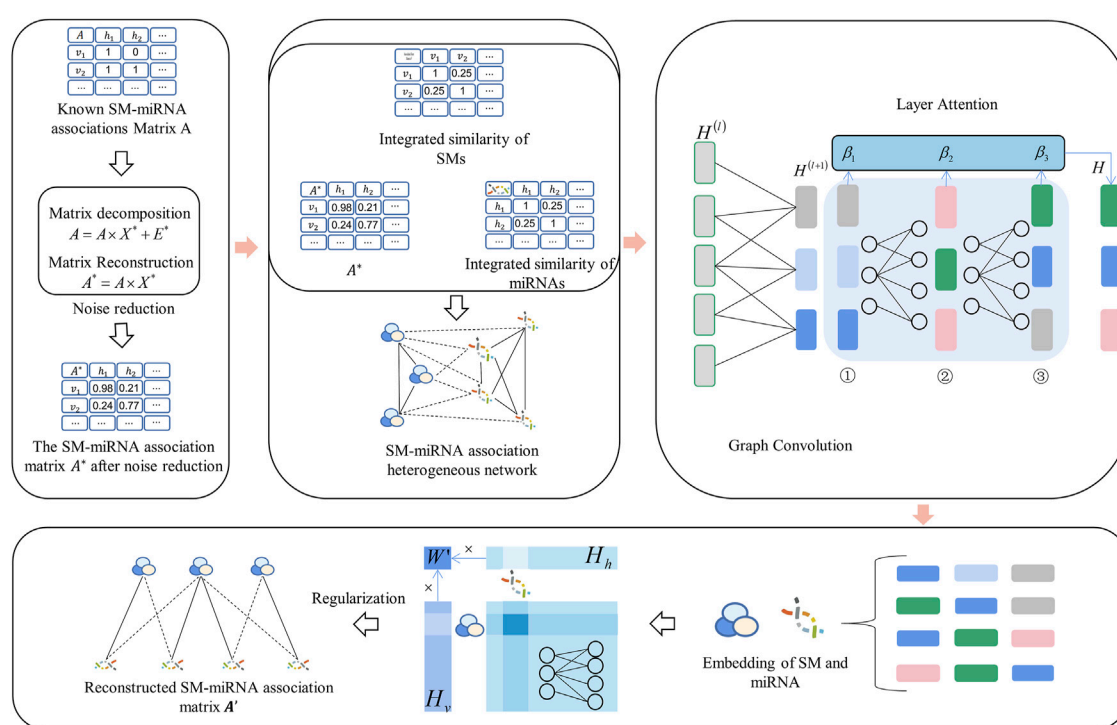


FIGURE 1

The flow chart of potential SM-miRNA association prediction based on GCNLASMMMA. Firstly, the matrix decomposition is applied to obtain the new SM-miRNA associations. Then the new SM-miRNA associations, integrated SM similarity and integrated miRNA similarity are constructed into an SM-miRNA association heterogeneous network. Finally, a graph convolutional network with layer attention mechanism is applied to obtain the reconstructed SM-miRNA association matrix.

integrated miRNA similarity. Specifically, we downloaded the target scores of each miRNA from the database TargetScan (Agarwal et al., 2015) and obtained gene function consistency-based similarity using the GSFS method (Lv et al., 2012). The indication phenotype-based similarity was obtained from the Human MicroRNA Disease Database (HMDD) version 2.0 (v 2.0), miR2Disease and PhenomiR databases using the GSFS method. Then, we combined the gene function consistency-based similarity and the indication phenotype-based similarity using the Jaccard score. Then, we named the two kinds of miRNA similarities after SM1 and SM2, respectively. Moreover, the integrated miRNA similarity SMR was obtained by the following equation,

$$SMR = \frac{\sum_j \beta_j SM_j}{\sum_j \beta_j}, (j = 1, 2) \quad (3)$$

where β_1 and β_2 represent the weights of miRNA similarities. Also, we set the values of β_1 and β_2 to 1, which means each miRNA similarity made an equal contribution to constituting the integrated miRNA similarity. Finally, the integrated miRNA similarity $SMR(m_i, m_j)$ between the i_{th} and j_{th} miRNAs was obtained after normalization as follows.

$$SMR(m_i, m_j) = \frac{SMR(m_i, m_j)}{\sqrt{\sum_{l=1}^m SMR(m_i, m_l)} \sqrt{\sum_{l=1}^m SMR(m_l, m_j)}} \quad (4)$$

2.4 GCNLASMMMA

GCNLASMMMA was separated into two steps. The known SM-miRNA association A was initially decomposed and reconstructed to obtain the new SM-miRNA association A^* . The reconstructed SM-miRNA association matrix A' was then obtained by calculating the new SM-miRNA association A^* using a graph convolutional network with an attention mechanism. More specifically, we obtained the new SM-miRNA associations by matrix decomposition. Then, the new SM-miRNA association matrix, integrated SM similarity and integrated miRNA similarity were constructed into a heterogeneous network. Finally, the graph convolutional network with layer attention mechanism was applied to obtain the reconstructed SM-miRNA association matrix. GCNLASMMMA is a model of a neural network with more hidden layers than other networks. The multi-layer calculation thoroughly considered the known

TABLE 1 The illustration of the IALM algorithm.

Algorithm: Inexact augmented lagrange multipliers

Input: Known SM-miRNA association matrix A and $\alpha = 0.1$

Initialize:
 $X = 0, E = 0, Y_1 = 0, Y_2 = 0, \mu = 10^{-4}, \max_{\mu} = 10^{10}, \rho = 1.1, \varepsilon = 10^{10}$

While true

1. Fix others and $J = \operatorname{argmin}_{\mu} \frac{1}{2} \|J\|_* + \frac{1}{2} \|J - (X + Y_2/\mu)\|_F^2$
2. Fix others and $X = (I + A^T A)(A^T A - A^T E + J + (A^T Y_1 - Y_2)/\mu)$
3. Fix others and $E = \operatorname{argmin}_{\mu} \frac{\alpha}{2} \|E\|_{2,1} + \frac{1}{2} \|E - (A - AX + Y_1/\mu)\|_F^2$
4. Update $Y_1 = Y_1 + \mu(A - AX - E); Y_2 = Y_2 + \mu(X - J)$
5. Update $\mu = \min(\rho\mu, \max_{\mu})$

If $\|A - AX - E\|_{\infty} < \varepsilon$ and $\|X - J\|_{\infty} < \varepsilon$

End while

Output: X^* and E^*

features and avoided overfitting. Moreover, the attention mechanism extracted significant information from each layer, thereby improving the accuracy of association prediction (Niu et al., 2021). The specific flow chart of GCNLASMMMA is shown in Figure 1.

2.4.1 Matrix decomposition

The existence of noise in known SM-miRNA associations tends to reduce prediction accuracy. Prior research has demonstrated that hidden features with considerable value can be extracted by applying dimension-reduction and noise-reduction to the data (Vidal, 2011). A low-rank matrix is a tool for efficiently obtaining hidden features with significant values (Peng et al., 2012). Therefore, we used matrix decomposition to learn a low-rank matrix from the known SM-miRNA association A . The decomposition of A was performed as follows:

$$A = A \times X + E \quad (5)$$

Since the above equation contains an infinite number of solutions, we applied the constraint to turn it into:

$$\min_{X,E} \|X\|_* + \alpha \|E\|_{2,1} \text{ s.t. } A = A \times X + E \quad (6)$$

where $\|X\|_* = \sum_i \sigma_i$, (σ_i is the singular value of matrix X), $\|E\|_{2,1} = \sum_{j=1}^n \sqrt{\sum_{i=1}^m (E_{ij})^2}$. In Eq. 6, the nuclear norm and sparse norm were applied to constrain X and E , which allowed X and E to be low-rank and sparse matrices, respectively. The balance parameter of low-rank and sparse matrices α was set to 0.1. According to earlier research, if A in Eq. 6 is transformed into an identity matrix, then the model is degenerated to the Robust Principal Component Analysis (RPCA), a convex optimization problem with constraints (Chandrasekaran et al., 2009).

$$\min_{X,E,J} \|J\|_* + \alpha \|E\|_{2,1} \text{ s.t. } A = A \times X + E, X = J \quad (7)$$

Based on the previous work (Meng et al., 2014), Eq. 7 can be converted into an unconstrained optimization problem. Therefore, the problem can be resolved using the Exact Augmented Lagrange Multipliers (EALM) algorithm.

$$L = \|J\|_* + \alpha \|E\|_{2,1} + \operatorname{tr}(Y_1^T (A - A \times X - E)) + \operatorname{tr}(Y_2^T (X - J)) + \frac{\delta}{2} (\|A - A \times X - E\|_F^2 + \|X - J\|_F^2) \quad (8)$$

In Eq. 8, the penalty parameter $\delta \geq 0$. According to the Inexact Augmented Lagrange Multipliers (IALM) algorithm (See Table 1), we fixed other variables and solved the minimum value of J , X and E by updating the Lagrange multipliers Y_1 and Y_2 . Moreover, we defined X^* and E^* as the solution of Eq. 8. X^* represents the similarity matrix of miRNA or SM. E^* represents the noise matrix. Then, the new SM-miRNA association A^* was expressed as:

$$A^* = A \times X^* \quad (9)$$

2.4.2 SM-miRNA heterogeneous network

In this study, the new SM-miRNA association A^* , integrated SM similarity SSM and integrated miRNA similarity SMR were combined into a heterogeneous network. There would be a known association between the i_{th} SM and the j_{th} miRNA if element A_{ij}^* in A^* equaled 1. $SSM(i, j)$ represented the integrated similarity between the i_{th} SM and the j_{th} SM. $SMR(i, j)$ represented the integrated similarity between the i_{th} miRNA and the j_{th} miRNA. The specific equation of the heterogeneous network A_H construction is as follows:

$$A_H = \begin{bmatrix} \sim SMR & A^* \\ A^{*T} & \sim SSM \end{bmatrix} \quad (10)$$

where A^{*T} represents the transpose matrix of A^* . In Eq. 10, we normalized the similarity matrix of SM and miRNA by $\sim SSM = D_s^{-\frac{1}{2}} SSM D_s^{-\frac{1}{2}}$ and $\sim SMR = D_m^{-\frac{1}{2}} SMR D_m^{-\frac{1}{2}}$, respectively. Specifically, $D_s = \operatorname{diag}(\sum_j SSM_{ij})$ and $D_m = \operatorname{diag}(\sum_j SMR_{ij})$.

2.4.3 Graph convolutional network

As classic network models, Long-Short Term Memory (LSTM) and Convolution Neural Network (CNN) are only applicable to grid-structured data. Nevertheless, the Graph Convolutional Network (GCN) can manage data with generalized topological graph structures and deeply explore the features of the data (Habib and Qureshi, 2020). In this paper, we constructed GCNLASMMMA, which is a model for graph convolution of biological information. Specifically, GCN was implemented on the SM-miRNA heterogeneous network A_H that was constructed by the known SM-miRNA associations, SM similarities and miRNA similarities. GCN is a neural network

TABLE 2 Validation of the random 50 SM-miRNAs associations. The first column records the random 1–25 associations. The second column records the random 26–50 associations.

SM	miRNA	Evidence	SM	miRNA	Evidence
CID 4116	hsa-mir-329-2	unconfirmed	CID 2662	hsa-mir-330	unconfirmed
CID 60726	hsa-mir-216b	unconfirmed	CID 7028	hsa-mir-592	unconfirmed
CID 4760	hsa-mir-520c	unconfirmed	CID 5656	hsa-mir-646	32083545
CID 3052	hsa-mir-193a	unconfirmed	CID 3520	hsa-mir-1266	unconfirmed
CID 444036	hsa-mir-199a-2	unconfirmed	CID 43008	hsa-mir-519a-1	unconfirmed
CID 3198	hsa-mir-216a	unconfirmed	CID 3343	hsa-mir-1469	unconfirmed
CID 157922	hsa-mir-1260a	unconfirmed	CID 5566	hsa-mir-548a-3	unconfirmed
CID 3698	hsa-mir-2110	unconfirmed	CID 5493444	hsa-mir-1285-2	unconfirmed
CID 4212	hsa-mir-219-2	unconfirmed	CID 60843	hsa-let-7d	unconfirmed
CID 8223	hsa-mir-98	unconfirmed	CID 110635	hsa-mir-216b	unconfirmed
CID 19861	hsa-mir-659	unconfirmed	CID 2801	hsa-mir-744	unconfirmed
CID 71329	hsa-mir-100	unconfirmed	CID 216239	hsa-mir-1273e	unconfirmed
CID 47641	hsa-mir-150	unconfirmed	CID 71398	hsa-mir-526a-1	unconfirmed
CID 443980	hsa-mir-760	unconfirmed	CID 4201	hsa-mir-153-2	unconfirmed
CID 5574	hsa-mir-512-2	unconfirmed	CID 5281040	hsa-mir-548a-2	unconfirmed
CID 8969	hsa-mir-543	unconfirmed	CID 444020	hsa-mir-320a	unconfirmed
CID 5282415	hsa-mir-619	unconfirmed	CID 3025	hsa-mir-24-1	unconfirmed
CID 65833	hsa-mir-760	unconfirmed	CID 3019	hsa-mir-1226	unconfirmed
CID 1775	hsa-mir-520f	unconfirmed	CID 1125	hsa-mir-27a	unconfirmed
CID 3749	hsa-mir-1285-2	unconfirmed	CID 1349907	hsa-mir-642a	unconfirmed
CID 2905	hsa-mir-96	unconfirmed	CID 656719	hsa-mir-611	unconfirmed
CID 3180	hsa-mir-148a	unconfirmed	CID 2795	hsa-mir-711	unconfirmed
CID 5566	hsa-mir-646	unconfirmed	CID 23994	hsa-mir-614	unconfirmed
CID 4212	hsa-mir-18a	31063487	CID 4099	hsa-mir-708	unconfirmed
CID 82146	hsa-mir-490	unconfirmed	CID 5281106	hsa-mir-1302-6	unconfirmed

structure consisting of an input layer, an output layer and many hidden layers that can represent nodes in a low-dimensional manner. Each hidden layer of GCN takes the output of the previous layer as input. The graph convolutional network propagation rule is as follows:

$$H^{(l+1)} = f(H^{(l)}, G) = \sigma(D^{-\frac{1}{2}}GD^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (11)$$

In Eq. 11, $H^{(l)}$ and $H^{(l+1)}$ denote the embeddings of nodes in the l_{th} and $(l+1)_{th}$ layers, respectively. $D = \text{diag}(\sum_j G_{ij})$ is a diagonal matrix of input graph G , $W^{(l)}$ represents the trainable weight matrix with a layer-specific value, $\sigma(\cdot)$ denotes the nonlinear activation function.

In the encoder part, to learn low-dimensional representations of miRNAs and SMs, we combined the new SM-miRNA association, integrated SM similarity and integrated miRNA similarity into SM-miRNA association heterogeneous network A_H . Firstly, we set a penalty factor μ in the input graph G during the propagation process as follows:

$$G = \begin{bmatrix} \mu \sim SMR & A^* \\ A^{*T} & \mu \sim SSM \end{bmatrix} \quad (12)$$

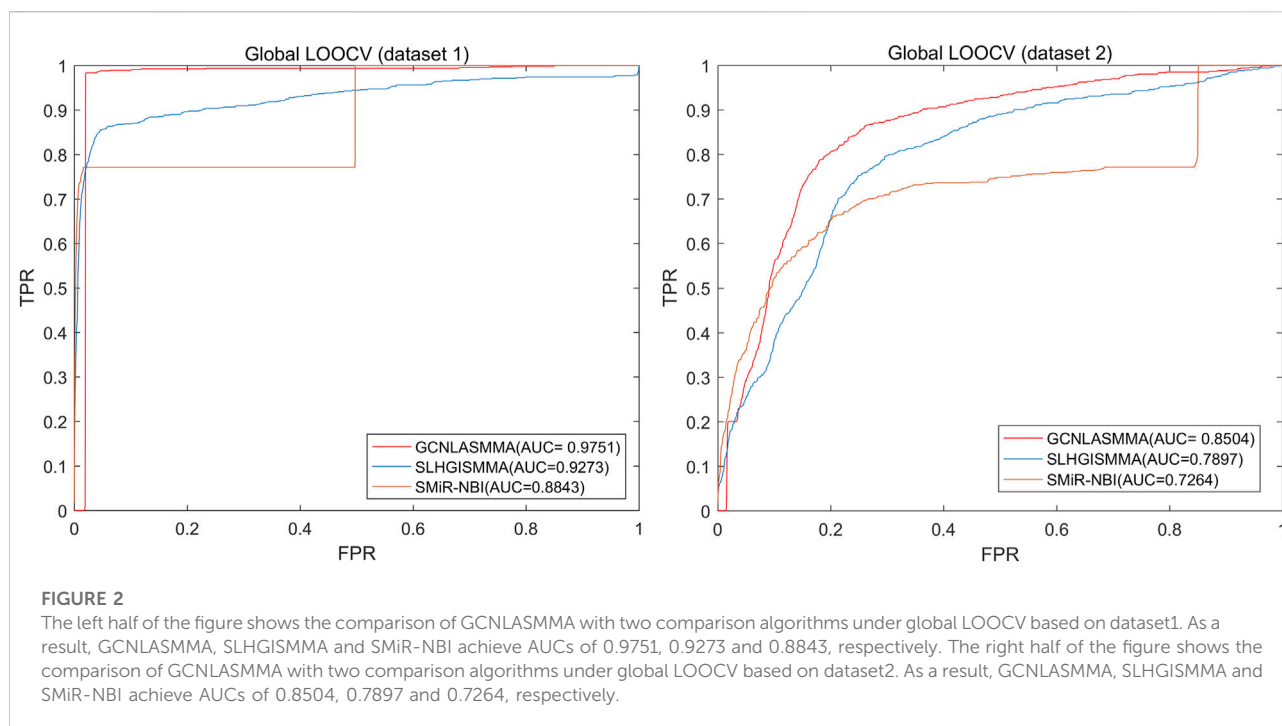
Then, we initialized the input layer embeddings as:

$$H^{(0)} = \begin{bmatrix} 0 & A^* \\ A^{*T} & 0 \end{bmatrix} \quad (13)$$

In this way, we obtained the propagation formula for the first layer from Eqs 11, 13:

$$H^{(1)} = \sigma(D^{-\frac{1}{2}}GD^{-\frac{1}{2}}H^{(0)}W^{(0)}) \quad (14)$$

In Eq. 12, $W^{(0)}$ is a weight matrix that acts only between the input layer and the first hidden layer. $H^{(1)}$ is the first-layer embeddings of the heterogeneous network A_H , k is the dimension of the embeddings. Similarly, the propagation rules for the subsequent layers of the GCN encoder followed Eq. 11, where $l = 1, 2, \dots, L$. After L iterations, we obtained L k -dimensional embeddings from different graph convolution



layers. Exponential linear elements were used as nonlinear activation functions in the graph convolution layer, which sped up the learning process and significantly improved the generalization performance.

In addition, we tried several different combinations of parameters from the range $\alpha \in \{400, 600, 800, 1000\}$, $lr \in \{0.00700, 0.00725, 0.00750, 0.00775, 0.00800\}$. By adjusting the parameters empirically, we set the dimensions of embeddings $k = 64$, the number of layers $L = 3$, the initial learning rate of optimizer $lr = 0.00725$, the total training epochs $\alpha = 600$, the two dropout rates $\beta = 0.6$ and $\gamma = 0.4$, the penalty factor $\mu = 6$ on both dataset1 and dataset2.

2.4.4 Layer attention mechanism

In addition, the layer attention mechanism was added to this model by introducing an attention mechanism between each layer and storing the position information in A_H . As a help for the attention mechanism, we extracted the pertinent information straight from the source data when constructing the embeddings of each layer output during the decoding process. Through this mechanism, we obtained the final SM embeddings and final miRNA embeddings from the fully connected layer:

$\begin{bmatrix} H_m \\ H_s \end{bmatrix} = \sum a_l H^l$, where H_m represents the final embeddings of miRNA, H_s is the final embeddings of SM. The neural network automatically adjusted the value of a_l by the initial input value $\frac{1}{(l+1)}$, $l = 1, 2, \dots, L$. Finally, we obtained the reconstructed SM-miRNA association matrix A' by an activation function as follows,

$$A' = \text{sigmoid}(H_m W' H_s^T) \quad (15)$$

where W' is a trainable matrix. The corresponding element A'_{ij} is the potential correlation score between miRNA m_i and SM s_j .

3 Results

To evaluate the performance of GCNLASMMMA, we used two types of cross validation, namely, 5-fold cross-validation and Leave-One-Out Cross Validation (LOOCV). The two different datasets include the same known 664 SM-miRNA associations. Specifically, dataset 1 has 831 SMs and 541 miRNAs. On the basis of dataset1, we removed the SMs and miRNAs that did not constitute any known association. Then, we obtained dataset2 which has only 286 different miRNAs, 39 different SMs. In this study, the Area Under the receiver operating characteristic Curves (AUCs) obtained under 5-fold cross-validation based on dataset1 and dataset2 were 0.9721 ± 0.0018 and 0.8393 ± 0.0047 , respectively. The global AUC and local AUC obtained under LOOCV based on dataset1 were 0.9751 (global LOOCV), 0.9746 (miRNA-fixed LOOCV) and 0.5014 (SM-fixed LOOCV), respectively. Based on dataset2, the AUCs of GCNLASMMMA were 0.8504 (global LOOCV), 0.8490 (miRNA-fixed LOOCV) and 0.6398 (SM-fixed LOOCV), respectively. Additionally, we utilized two types of case studies to confirm the effectiveness of GCNLASMMMA in identifying

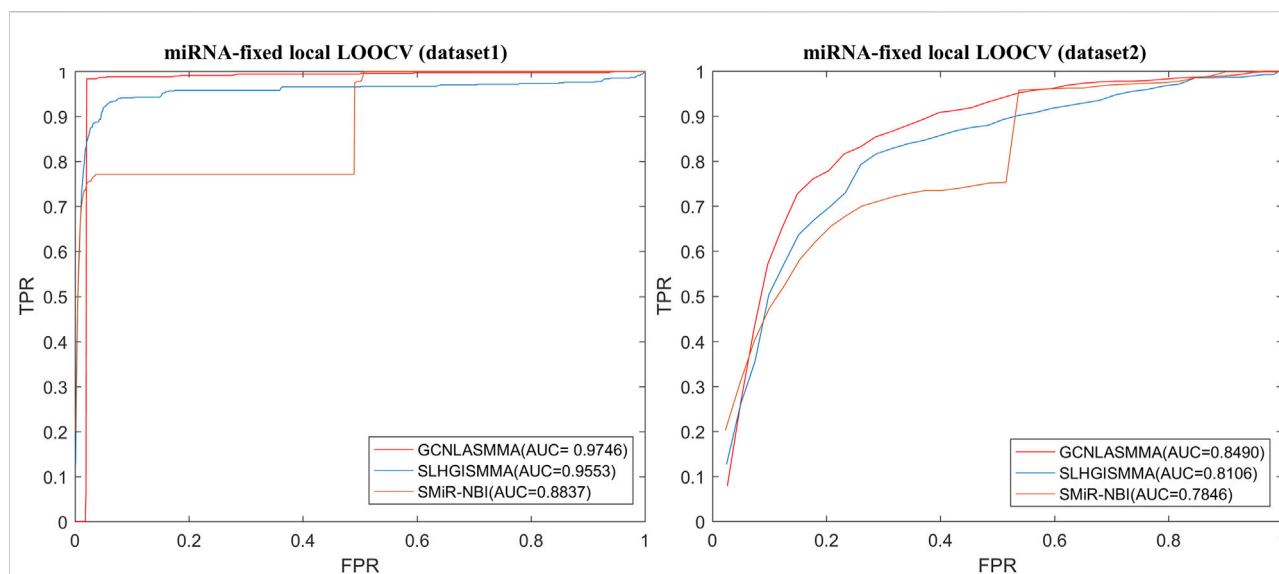


FIGURE 3

The left half of the figure shows the comparison of GCNLASMMMA with two comparison algorithms under miRNA-fixed LOOCV based on dataset1. As a result, GCNLASMMMA, SLHGISMMA and SMiR-NBI achieve AUCs of 0.9746, 0.9553 and 0.8837, respectively. The right half of the figure shows the comparison of GCNLASMMMA with two comparison algorithms under miRNA-fixed LOOCV based on dataset2. As a result, GCNLASMMMA, SLHGISMMA and SMiR-NBI achieve AUCs of 0.8490, 0.8106 and 0.7846, respectively.

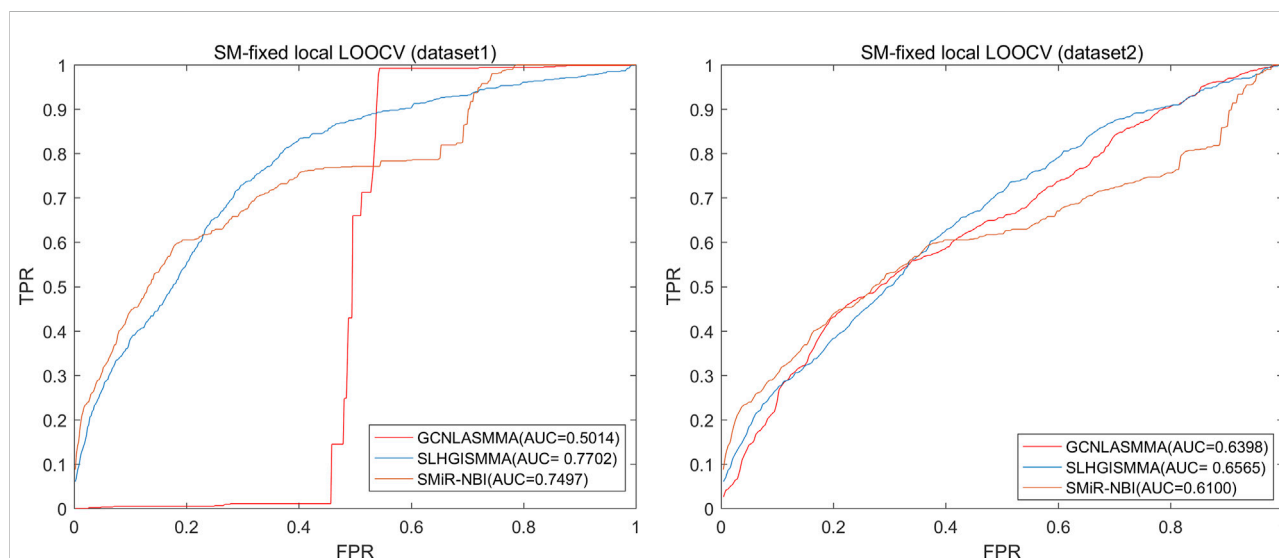


FIGURE 4

The left half of the figure shows the comparison of GCNLASMMMA with two comparison algorithms under SM-fixed LOOCV based on dataset1. As a result, GCNLASMMMA, SLHGISMMA and SMiR-NBI achieve AUCs of 0.5014, 0.7702 and 0.7497, respectively. The right half of the figure shows the comparison of GCNLASMMMA with two comparison algorithms under SM-fixed LOOCV based on dataset2. As a result, GCNLASMMMA, SLHGISMMA and SMiR-NBI achieve AUCs of 0.6398, 0.6565 and 0.6100, respectively.

potential miRNAs for investigated SMs. Specifically, GCNLASMMMA has predicted the potential miRNAs associated with 5-Fluorouracil (5-Fu, CID: 3385), 5-Aza-2'-deoxycytidine (5-Aza-CdR, CID: 451668) and 17 β -Estradiol (E2, CID: 5757).

For 5-Fu, the results showed that 9, 16 and 39 out of the top 10, 20 and 50 potential related miRNAs in the first type of case studies, 8, 15 and 39 out of the top 10, 20 and 50 potential related miRNAs in the second type of case studies were validated in other

TABLE 3 Validation of the top 50 miRNAs associated with 5-Fu in the first type of case studies. The first column records the top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

SM	miRNA	Evidence	SM	miRNA	Evidence
CID 3385	hsa-miR-151a	23220571	CID 3385	hsa-miR-126	26062749
CID 3385	hsa-miR-195	21947305	CID 3385	hsa-miR-128-1	23220571
CID 3385	hsa-let-7d	23220571	CID 3385	hsa-miR-337	unconfirmed
CID 3385	hsa-miR-195	21947305	CID 3385	hsa-miR-181c	unconfirmed
CID 3385	hsa-miR-125a	23220571	CID 3385	hsa-miR-30c-1	unconfirmed
CID 3385	hsa-miR-345	unconfirmed	CID 3385	hsa-miR-27a	23220571
CID 3385	hsa-miR-16-1	26198104	CID 3385	hsa-let-7a-1	23220571
CID 3385	hsa-miR-24-1	26198104	CID 3385	hsa-miR-139	27173050
CID 3385	hsa-miR-23b	23220571	CID 3385	hsa-miR-302b	26457704
CID 3385	hsa-miR-1226	26198104	CID 3385	hsa-let-7b	25789066
CID 3385	hsa-miR-151a	23220571	CID 3385	hsa-miR-26b	23220571
CID 3385	hsa-miR-132	23220571	CID 3385	hsa-miR-221	27501171
CID 3385	hsa-125b-1	unconfirmed	CID 3385	hsa-miR-338	28928082
CID 3385	hsa-let-7e	23220571	CID 3385	hsa-miR-130a	unconfirmed
CID 3385	hsa-miR-19a	23220571	CID 3385	hsa-miR-10b	22322955
CID 3385	hsa-miR-181a-1	unconfirmed	CID 3385	hsa-miR-204	27095441
CID 3385	hsa-miR-181b-1	unconfirmed	CID 3385	hsa-miR-26a-1	unconfirmed
CID 3385	hsa-miR-25	23220571	CID 3385	hsa-miR-92a-1	23220571
CID 3385	hsa-miR-106a	23220571	CID 3385	hsa-miR-299	31786874
CID 3385	hsa-miR-200c	23220571	CID 3385	hsa-miR-107	26636340
CID 3385	hsa-miR-22	25449431	CID 3385	hsa-miR-181a-2	24462870
CID 3385	hsa-miR-20a	23220571	CID 3385	hsa-miR-205	24396484
CID 3385	hsa-let-7d	23220571	CID 3385	hsa-miR-23a	23220571
CID 3385	hsa-miR-34b	unconfirmed	CID 3385	hsa-miR-199b	unconfirmed
CID 3385	hsa-miR-205	24396484	CID 3385	hsa-miR-93	23220571

literature or databases, respectively. For 5-Aza-CdR, the results showed that 8, 13 and 26 out of the top 10, 20 and 50 potential related miRNAs in the first type of case studies, 8, 14 and 28 out of the top 10, 20 and 50 potential related miRNAs in the second type of case studies were validated in other literature or databases, respectively. For E2, the results showed that 6, 14 and 29 out of the top 10, 20 and 50 potential related miRNAs in the first type of case studies, 4, 11 and 29 out of the top 10, 20 and 50 potential related miRNAs in the second type of case studies were validated in other literature or databases, respectively.

3.1 Performance evaluation

In 5-fold cross-validation, all known SM-miRNA associations were randomly separated into five subsets of nearly comparable size. Then, each subset was in turn considered as the test sample, and the rest four subsets were treated as training samples. Moreover, all unknown SM-miRNA pairs were regarded as candidate samples. Subsequently, we obtained a predicted association score matrix by

GCNLASMMMA, and ranked the scores of each test sample against those of the candidate samples. This partition-prediction-ranking procedure was repeated 100 times to obtain a sound estimate of the mean and variance of GCNLASMMMA’s prediction accuracy. Finally, the prediction of a test sample was deemed successful if the sample’s rank was higher than the given threshold. Therefore, we utilized the threshold to calculated the false positive rate (FPR, 1-specificity) and the true positive rate (TPR, sensitivity). The FPR and TPR represented the percentage of candidate samples that lower than the threshold and the percentage of test samples that higher than the threshold, respectively. Then, we regarded FPR and TPR as horizontal and vertical axis. The Receiver Operating Characteristic (ROC) curve were plotted. Finally, we attained the Area Under the ROC Curve (AUC) by computing the area under the ROC curves. In this investigation, GCNLASMMMA achieved the AUCs of 0.9721 ± 0.0018 and 0.8393 ± 0.0047 under 5-fold cross-validation based on dataset1 and dataset2, respectively.

LOOCV was further classified as either global and local. Then, the local-LOOCV was subdivided into miRNA-fixed

TABLE 4 Validation of the top 50 miRNAs associated with 5-Fu in the second type of case studies. The first column records the top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

SM	miRNA	Evidence	SM	miRNA	Evidence
CID 3385	hsa-miR-151a	23220571	CID 3385	hsa-miR-195	21947305
CID 3385	hsa-let-7d	23220571	CID 3385	hsa-miR-27a	23220571
CID 3385	hsa-miR-205	24396484	CID 3385	hsa-miR-204	27095441
CID 3385	hsa-miR-181a-2	24462870	CID 3385	hsa-miR-181a-1	unconfirmed
CID 3385	hsa-miR-23a	23220571	CID 3385	hsa-miR-25	23220571
CID 3385	hsa-miR-1226	26198104	CID 3385	hsa-miR-199b	unconfirmed
CID 3385	hsa-miR-181c	unconfirmed	CID 3385	hsa-miR-139	27173050
CID 3385	hsa-miR-151a	23220571	CID 3385	hsa-miR-195	21947305
CID 3385	hsa-miR-26a-1	unconfirmed	CID 3385	hsa-miR-132	23220571
CID 3385	hsa-miR-26b	23220571	CID 3385	hsa-miR-20a	23220571
CID 3385	hsa-miR-130a	unconfirmed	CID 3385	hsa-miR-126	26062749
CID 3385	hsa-miR-345	unconfirmed	CID 3385	hsa-125b-1	unconfirmed
CID 3385	hsa-miR-128-1	23220571	CID 3385	hsa-miR-200c	23220571
CID 3385	hsa-let-7d	23220571	CID 3385	hsa-miR-299	31786874
CID 3385	hsa-miR-181b-1	unconfirmed	CID 3385	hsa-miR-30c-1	unconfirmed
CID 3385	hsa-miR-205	24396484	CID 3385	hsa-miR-24-1	26198104
CID 3385	hsa-miR-125a	23220571	CID 3385	hsa-miR-93	23220571
CID 3385	hsa-miR-22	25449431	CID 3385	hsa-let-7e	23220571
CID 3385	hsa-miR-16-1	26198104	CID 3385	hsa-let-7b	25789066
CID 3385	hsa-miR-106a	23220571	CID 3385	hsa-miR-221	27501171
CID 3385	hsa-miR-23b	23220571	CID 3385	hsa-miR-19a	23220571
CID 3385	hsa-miR-338	28928082	CID 3385	hsa-miR-92a-1	23220571
CID 3385	hsa-miR-10b	22322955	CID 3385	hsa-miR-302b	26457704
CID 3385	hsa-let-7a-1	23220571	CID 3385	hsa-miR-107	26636340
CID 3385	hsa-miR-337	unconfirmed	CID 3385	hsa-miR-34b	unconfirmed

LOOCV and SM-fixed LOOCV. In LOOCV, each known SM-miRNA association was in turn considered to be the test sample and the others were treated as the training samples. Moreover, all unknown SM-miRNA pairs were treated as candidate samples. In miRNA-fixed LOOCV and SM-fixed LOOCV, test samples and training samples were chosen similarly. However, in SM-fixed LOOCV, only unknown SM-miRNA pairs containing the selected SM were regarded as candidate samples. Similarly, in miRNA-fixed LOOCV, candidate samples only included those involving the chosen miRNA. Then, we ranked the score of the test sample against those of the candidate samples. Finally, the prediction of a test sample was deemed successful if the rank of this test sample was higher than the given threshold. Based on dataset1, GCNLASMMMA attained the AUCs of 0.9751, 0.9746 and 0.5014 under global LOOCV, miRNA-fixed LOOCV and SM-fixed LOOCV, respectively. Based on dataset2, GCNLASMMMA attained the AUCs of 0.8504, 0.8490 and 0.6398 under global LOOCV, miRNA-fixed LOOCV and SM-fixed LOOCV, respectively.

The AUC comparison figures based on dataset1 (dataset2) were plotted to determine the differences between

GCNLASMMMA and other models' outcomes. AUC = 0.5 would suggest that the model was only capable of random prediction, whereas AUC = 1 would indicate that all test samples were accurately predicted. Figure 2 demonstrates that the results of GCNLASMMMA under global LOOCV are significantly better than that of SMiR-NBI. Figures 3, 4 show that the results of GCNLASMMMA under miRNA-fixed local LOOCV and SM-fixed local LOOCV were significantly better than those of SLHGISMMA and SMiR-NBI. Furthermore, the AUC of miRNA-fixed local LOOCV based on dataset1 is 0.9746, which means almost all potential SM-miRNA associations in dataset1 were predicted successfully.

3.2 Case studies

To further illustrate the GCNLASMMMA's applicability to identify potential miRNAs, we conducted two types of case studies on three essential SMs, namely 5-Fluorouracil (5-Fu, CID: 3385), 5-Aza-2'-deoxycytidine (5-Aza-CdR, CID:

TABLE 5 Validation of the top 50 miRNAs associated with 5-Aza-CdR in the first type of case studies. The first column records the top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

SM	miRNA	Evidence	SM	miRNA	Evidence
CID 451668	hsa-miR-20a	23220571	CID 451668	hsa-miR-30a	unconfirmed
CID 451668	hsa-miR-320a	26198104	CID 451668	hsa-miR-107	23220571
CID 451668	hsa-miR-125a	23220571	CID 451668	hsa-miR-199b	24659709
CID 451668	hsa-miR-182	23220571	CID 451668	hsa-let-7a-1	unconfirmed
CID 451668	hsa-miR-204	unconfirmed	CID 451668	hsa-miR-92a-1	unconfirmed
CID 451668	hsa-miR-200b	23626803	CID 451668	hsa-miR-181a-1	23220571
CID 451668	hsa-miR-23a	unconfirmed	CID 451668	hsa-let-7e	22053057
CID 451668	hsa-let-7f-1	23220571	CID 451668	hsa-miR-26a-1	unconfirmed
CID 451668	hsa-let-7b	26708866	CID 451668	hsa-miR-1233-1	unconfirmed
CID 451668	hsa-miR-200c	23626803	CID 451668	hsa-miR-130a	23220571
CID 451668	hsa-miR-25	23220571	CID 451668	hsa-miR-30c-1	unconfirmed
CID 451668	hsa-miR-128-1	27705931	CID 451668	hsa-miR-22	23220571
CID 451668	hsa-miR-145	26198104	CID 451668	hsa-miR-301a	unconfirmed
CID 451668	hsa-miR-221	unconfirmed	CID 451668	hsa-let-7g	23220571
CID 451668	hsa-miR-19b-1	unconfirmed	CID 451668	hsa-miR-195	23333942
CID 451668	hsa-miR-197	unconfirmed	CID 451668	hsa-miR-302b	unconfirmed
CID 451668	hsa-let-7i	23220571	CID 451668	hsa-miR-26b	unconfirmed
CID 451668	hsa-miR-181b-1	unconfirmed	CID 451668	hsa-miR-205	unconfirmed
CID 451668	hsa-miR-338	unconfirmed	CID 451668	hsa-miR-218-1	unconfirmed
CID 451668	hsa-let-7d	26802971	CID 451668	hsa-miR-93	23220571
CID 451668	hsa-miR-139	unconfirmed	CID 451668	hsa-miR-124-1	unconfirmed
CID 451668	hsa-miR-328	unconfirmed	CID 451668	hsa-miR-15b	unconfirmed
CID 451668	hsa-miR-126	23220571	CID 451668	hsa-miR-10b	unconfirmed
CID 451668	hsa-miR-17	23220571	CID 451668	hsa-miR-128-2	unconfirmed
CID 451668	hsa-miR-19a	23220571	CID 451668	hsa-miR-27a	23220571

451668) and 17 β -Estradiol (E2, CID: 5757). On the basis of all known SM-miRNA associations, the first type was applied to forecast potential miRNAs for investigated SMs. As the training set, we utilized the known SM-miRNA associations from dataset1. Then, for each investigated SM, we ranked all candidate miRNAs according on their predicted scores. The second type was used to forecast potential miRNAs for investigated SMs without any known SM-miRNA association. Therefore, we removed all verified associations related to the investigated SMs before the prediction and ranked them as the first type of case studies. After ranking all candidate miRNAs for each investigated SM based on their predicted scores, the top 50 predicted miRNAs were picked out and verified in other literature or databases. Moreover, we selected 10, 20 and 50 associations randomly from all potential associations to further demonstrate the validity of GCNLASMMMA. The results show that only 0, 0 and 2 out of random 10, 20 and 50 associations are confirmed in other literature or databases (See Table 2), which significantly worse than the top 10, 20 and 50 miRNAs related to investigated SMs.

3.2.1 5-Fu

5-Fu, one of the earliest anticancer drugs, can be fully absorbed by tumor cells. Moreover, 5-Fu can decrease tumor cell proliferation by interfering with the formation of DeoxyriboNucleic Acid (DNA) and RiboNucleic Acid (RNA) in tumor cells. It has been demonstrated that 5-Fu has considerable inhibitory effects on various cancer cells. Therefore, 5-Fu is frequently used as a positive control in anticancer drug effect experiments and clinical adjuvant treatment of gastric cancer (Longley et al., 2003). The first type of case studies' results show that 9, 16 and 39 out of the top 10, 20 and 50 potential 5-Fu-associated miRNAs are confirmed in other literature or databases (See Table 3). The second type of case studies' results show that 8, 15 and 39 out of the top 10, 20 and 50 potential 5-Fu-associated miRNAs are confirmed in other literature or databases (See Table 4). For example, 5-Fu is the most common chemotherapeutic agent for colorectal cancer. On the one hand, over-expression of hsa-miR-23a causes the resistance to 5-Fu in microsatellite instability colorectal cancer, which results in a diminished effect of 5-Fu chemotherapy (Shang et al., 2014). On the other hand, Ectopic expression of hsa-miR-23a increased the viability and survival of microsatellite stability

TABLE 6 Validation of the top 50 miRNAs associated with 5-Aza-CdR in the second type of case studies. The first column records the top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

SM	miRNA	Evidence	SM	miRNA	Evidence
CID 451668	hsa-miR-20a	23220571	CID 451668	hsa-miR-92a-1	unconfirmed
CID 451668	hsa-miR-181b-1	unconfirmed	CID 451668	hsa-miR-125a	23220571
CID 451668	hsa-miR-205	unconfirmed	CID 451668	hsa-let-7b	26708866
CID 451668	hsa-miR-19a	23220571	CID 451668	hsa-miR-302b	unconfirmed
CID 451668	hsa-miR-181a-1	23220571	CID 451668	hsa-miR-30a	unconfirmed
CID 451668	hsa-miR-130a	23220571	CID 451668	hsa-miR-23b	23220571
CID 451668	hsa-let-7g	23220571	CID 451668	hsa-miR-199b	24659709
CID 451668	hsa-miR-200b	23626803	CID 451668	hsa-miR-128-2	unconfirmed
CID 451668	hsa-miR-126	23220571	CID 451668	hsa-miR-15b	unconfirmed
CID 451668	hsa-miR-320a	26198104	CID 451668	hsa-miR-124-1	unconfirmed
CID 451668	hsa-miR-30c-1	unconfirmed	CID 451668	hsa-miR-26b	unconfirmed
CID 451668	hsa-miR-328	unconfirmed	CID 451668	hsa-miR-128-1	27705931
CID 451668	hsa-let-7e	22053057	CID 451668	hsa-let-7a-1	unconfirmed
CID 451668	hsa-miR-10b	unconfirmed	CID 451668	hsa-miR-218-1	unconfirmed
CID 451668	hsa-let-7f-1	23220571	CID 451668	hsa-miR-200c	23626803
CID 451668	hsa-miR-221	unconfirmed	CID 451668	hsa-miR-26a-1	unconfirmed
CID 451668	hsa-miR-182	23220571	CID 451668	hsa-miR-338	unconfirmed
CID 451668	hsa-let-7i	23220571	CID 451668	hsa-miR-93	23220571
CID 451668	hsa-miR-195	23333942	CID 451668	hsa-miR-139	unconfirmed
CID 451668	hsa-miR-27a	23220571	CID 451668	hsa-miR-145	26198104
CID 451668	hsa-miR-204	unconfirmed	CID 451668	hsa-miR-107	23220571
CID 451668	hsa-miR-25	23220571	CID 451668	hsa-let-7d	26802971
CID 451668	hsa-miR-23a	unconfirmed	CID 451668	hsa-miR-19b-1	unconfirmed
CID 451668	hsa-let-7f-1	23220571	CID 451668	hsa-miR-22	23220571
CID 451668	hsa-miR-17	23220571	CID 451668	hsa-miR-197	unconfirmed

colorectal cancer cells, thereby leading to the apoptosis of colorectal cancer cells (Li et al., 2015).

3.2.2 5-Aza-CdR

5-Aza-CdR can bind to DNA methyltransferases to reduce methylation levels, reducing the biological activity of methyltransferase inhibitors and regulating gene expression. In clinical usage, 5-Aza-CdR is frequently used in clinical settings to treat diseases caused by gene variants (Do Amaral et al., 2019). Additionally, 5-Aza-CdR can suppress tumor cell proliferation via demethylation, making it one of the most potent inhibitors currently available *in vitro* (Lemaire et al., 2008). Meanwhile, 5-Aza-CdR can enhance the sensitivity of targeted drugs in non-small cell lung cancer chemotherapy, inhibit cell proliferation, accelerate the apoptosis of cancer cells, induce cell differentiation and activate quiescent anticancer cells in the human body. The first type of case studies' results show that 8, 13 and 26 out of the top 10, 20 and 50 potential 5-Aza-CdR-associated miRNAs are confirmed in other literature or databases (See Table 5). The second type of case studies' results show that 8,

14 and 28 out of the top 10, 20 and 50 potential 5-Aza-CdR-associated miRNAs are confirmed in other literature or databases (See Table 6). For example, quantitative methylation-specific Polymerase Chain Reaction analysis showed hypermethylation of the choline phosphoglyceride island adjacent to hsa-let-7e, and demethylation treatment with 5-Aza-CdR or transfection of pYr-let-7e-shRNA plasmid containing unmethylated hsa-let-7e DNA sequence could restore hsa-let-7e expression and partly reduce the chemoresistance (Cai et al., 2013).

3.2.3 E2

In addition to stimulating the growth and maintenance of the reproductive system, E2 exerts protective effects on cardiovascular and other organs. Specifically, E2 can reduce blood cholesterol levels by decreasing Low-Density Lipoprotein (LDL), increasing High-Density Lipoprotein (HDL) and boosting apolipoprotein content (Oh et al., 2019). Moreover, researchers are paying more attention to the anti-inflammatory, antioxidant and anti-apoptotic properties of E2 on cardiovascular diseases such as coronary heart disease and atherosclerosis, are getting more attention from researchers (Tse

TABLE 7 Validation of the top 50 miRNAs associated with E2 in the first type of case studies. The first column records the top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

SM	miRNA	Evidence	SM	miRNA	Evidence
CID 5757	hsa-miR-183	unconfirmed	CID 5757	hsa-miR-181b-1	unconfirmed
CID 5757	hsa-let-7g	23220571	CID 5757	hsa-miR-19b-1	unconfirmed
CID 5757	hsa-miR-181a-2	unconfirmed	CID 5757	hsa-miR-141	unconfirmed
CID 5757	hsa-miR-125a	21914226	CID 5757	hsa-miR-15a	unconfirmed
CID 5757	hsa-miR-107	23220571	CID 5757	hsa-miR-17	23220571
CID 5757	hsa-miR-26b	24735615	CID 5757	hsa-miR-10b	23220571
CID 5757	hsa-miR-19a	29416771	CID 5757	hsa-miR-30a	29331043
CID 5757	hsa-miR-195	unconfirmed	CID 5757	hsa-let-7f-1	23220571
CID 5757	hsa-miR-128-2	23220571	CID 5757	hsa-miR-302b	23220571
CID 5757	hsa-miR-181a-1	unconfirmed	CID 5757	hsa-miR-199b	unconfirmed
CID 5757	hsa-miR-128-1	23220571	CID 5757	hsa-miR-181c	unconfirmed
CID 5757	hsa-miR-130a	unconfirmed	CID 5757	hsa-miR-106b	28422740
CID 5757	hsa-miR-338	22996663	CID 5757	hsa-miR-23a	23220571
CID 5757	hsa-let-7e	23220571	CID 5757	hsa-miR-9-2	23220571
CID 5757	hsa-miR-20a	21914226	CID 5757	hsa-miR-182	28678802
CID 5757	hsa-miR-200c	23220571	CID 5757	hsa-miR-139	unconfirmed
CID 5757	hsa-miR-27a	23220571	CID 5757	hsa-let-7b	23220571
CID 5757	hsa-miR-200b	23220571	CID 5757	hsa-miR-25	unconfirmed
CID 5757	hsa-miR-221	21057537	CID 5757	hsa-miR-218-1	unconfirmed
CID 5757	hsa-miR-151a	unconfirmed	CID 5757	hsa-miR-22	24715036
CID 5757	hsa-miR-204	29789714	CID 5757	hsa-miR-15b	23220571
CID 5757	hsa-miR-106a	unconfirmed	CID 5757	hsa-miR-130a	unconfirmed
CID 5757	hsa-miR-205	unconfirmed	CID 5757	hsa-miR-23b	23220571
CID 5757	hsa-miR-92a-1	unconfirmed	CID 5757	hsa-miR-26a-1	unconfirmed
CID 5757	hsa-miR-130b	unconfirmed	CID 5757	hsa-miR-30c-1	23220571

et al., 1999; Rachoní et al., 2002). The first type of case studies’ results show that 6, 14 and 29 out of the top 10, 20 and 50 potential E2-associated miRNAs are confirmed in other literature or databases (See Table 7). The second type of case studies’ results show that 4, 11 and 29 out of the top 10, 20 and 50 potential E2-associated miRNAs are confirmed in other literature or databases (See Table 8). For example, hsa-miR-23a could be negatively regulated by E2 in both myocardium and cultured cardiomyocytes. Moreover, hsa-miR-23a could directly down-regulate peroxisome proliferator-activated receptor γ coactivator- α (PGC-1 α) expression in cardiomyocytes via binding to its 3’-untranslated regions, which implied that hsa-miR-23a could be critical for the down-regulation of PGC-1 α under E2 deficiency (Sun et al., 2014).

4 Discussion

Deep learning offers a wide range of applications in major areas of computer science, such as computer vision, natural language processing and machine translation. More effective models can be obtained by adding hidden layers to standard neural networks. Deep

learning also contributes to medication development and precision medicine by predicting potential SM-miRNA associations. Furthermore, deep learning models have more hidden layer nodes than conventional neural networks. The number of hidden layers can even reach ten for extremely complex problems. After multiple layers of calculation, the results of deep learning-based algorithms are often closer to the actual situation than those of traditional machine learning-based algorithms. Initially, we utilized matrix decomposition to reduce noise from known SM-miRNA associations. Then, the layer attention mechanism was introduced to the deep learning model, which significantly improved the performance of our model by integrating the SM-miRNA association feature vectors used for calculation.

GCNLASMMA is a model of a neural network with numerous hidden layers. Multiple layers computations allowed the results to completely consider known features and avoid overfitting. The attention mechanism extracted vital information from each layer of the neural network. Besides, the matrix decomposition module reduced the noise of known SM-miRNA associations, significantly enhancing GCN’s performance. GCNLASMMA was an attempt to identify

TABLE 8 Validation of the top 50 miRNAs associated with E2 in the second type of case studies. The first column records the top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

SM	miRNA	Evidence	SM	miRNA	Evidence
CID 5757	hsa-miR-183	unconfirmed	CID 5757	hsa-miR-19a	29416771
CID 5757	hsa-miR-30c-1	23220571	CID 5757	hsa-miR-19b-1	unconfirmed
CID 5757	hsa-miR-15a	unconfirmed	CID 5757	hsa-miR-125a	21914226
CID 5757	hsa-miR-181a-1	unconfirmed	CID 5757	hsa-miR-15b	23220571
CID 5757	hsa-let-7f-1	23220571	CID 5757	hsa-miR-128-2	23220571
CID 5757	hsa-miR-181b-1	unconfirmed	CID 5757	hsa-miR-20a	21914226
CID 5757	hsa-miR-205	unconfirmed	CID 5757	hsa-miR-26b	24735615
CID 5757	hsa-miR-181a-2	unconfirmed	CID 5757	hsa-miR-10b	23220571
CID 5757	hsa-miR-9-2	23220571	CID 5757	hsa-miR-181c	unconfirmed
CID 5757	hsa-miR-23a	23220571	CID 5757	hsa-miR-22	24715036
CID 5757	hsa-miR-128-1	23220571	CID 5757	hsa-miR-139	unconfirmed
CID 5757	hsa-let-7e	23220571	CID 5757	hsa-miR-106a	unconfirmed
CID 5757	hsa-let-7b	23220571	CID 5757	hsa-miR-141	unconfirmed
CID 5757	hsa-miR-130a	unconfirmed	CID 5757	hsa-let-7g	23220571
CID 5757	hsa-miR-338	22996663	CID 5757	hsa-miR-107	23220571
CID 5757	hsa-miR-30a	29331043	CID 5757	hsa-miR-23b	23220571
CID 5757	hsa-miR-302b	23220571	CID 5757	hsa-miR-195	unconfirmed
CID 5757	hsa-miR-130b	unconfirmed	CID 5757	hsa-miR-27a	23220571
CID 5757	hsa-miR-106b	28422740	CID 5757	hsa-miR-25	unconfirmed
CID 5757	hsa-miR-199b	unconfirmed	CID 5757	hsa-miR-204	29789714
CID 5757	hsa-miR-200b	23220571	CID 5757	hsa-miR-221	21057537
CID 5757	hsa-miR-182	28678802	CID 5757	hsa-miR-151a	unconfirmed
CID 5757	hsa-miR-26a-1	unconfirmed	CID 5757	hsa-miR-218-1	unconfirmed
CID 5757	hsa-miR-17	23220571	CID 5757	hsa-miR-130a	unconfirmed
CID 5757	hsa-miR-200c	23220571	CID 5757	hsa-miR-92a-1	unconfirmed

potential SM-miRNA associations using deep learning. The advantages above enabled GCNLASMMA to accurately anticipate potential SM-miRNA associations.

Deep learning’s spectacular performance is contingent on a vast number of known SM-miRNA associations. The number of known SM-miRNA associations utilized in this investigation was apparently insufficient to fulfill GCNLASMMA. Therefore, the performance of GCNLASMMA was still unsatisfactory. In addition, the parameters used in GCNLASMMA may not be ideal. Moreover, the construction of heterogeneous networks will yield better results if other biological information, such as long non-coding RNA or disease, is utilized. These factors will motivate researchers to develop more effective deep learning models to predict potential SM-miRNA associations using more trustworthy biological datasets.

Data availability statement

The Python code and datasets of GCNLASMMA are publicly available at <https://github.com/1054366388/GCNLASMMMA>.

Author contributions

JZL led the project and supervised the writing. JN did the experiments, searched the literature and wrote the article. XLC and TGN did the subsequent revisions. All authors participated in the revisions and approved the final version of the manuscript.

Funding

This study was supported by the Postgraduate Research and Practice Innovation Program of Jiangsu Province (Grant Number KYCX21_2836).

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

References

- Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4. doi:10.7554/eLife.05005
- Angermueller, C., Pärnamäa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12 (7), 878. doi:10.15252/msb.20156651
- Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* 116 (2), 281–297. doi:10.1016/s0092-8674(04)00045-5
- Borges Oliveira, D. A., Ribeiro Pereira, L. G., Bresolin, T., Pontes Ferreira, R. E., and Rebouças Dorea, J. R. (2021). A review of deep learning algorithms for computer vision systems in livestock. *Livest. Sci.* 253, 104700. doi:10.1016/j.livsci.2021.104700
- Cai, J., Yang, C., Yang, Q., Ding, H., Jia, J., Guo, J., et al. (2013). Deregulation of let-7e in epithelial ovarian cancer promotes the development of resistance to cisplatin. *Oncogenesis* 2 (10), e75. doi:10.1038/oncsis.2013.39
- Carnevali, M., Parsons, J., Wyles, D. L., and Hermann, T. (2010). A modular approach to synthetic RNA binders of the hepatitis C virus internal ribosome entry site. *ChemBioChem* 11 (10), 1364–1367. doi:10.1002/cbic.201000177
- Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2009). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* 21 (2), 572–596. doi:10.1137/090761793
- Chen, C. Z., Sobczak, K., Hoskins, J., Southall, N., Marugan, J. J., Zheng, W., et al. (2012). Two high-throughput screening assays for aberrant RNA–protein interactions in myotonic dystrophy type 1. *Anal. Bioanal. Chem.* 402 (5), 1889–1898. doi:10.1007/s00216-011-5604-0
- Chen, X., Zhou, C., Wang, C. C., and Zhao, Y. (2021). Predicting potential small molecule-miRNA associations based on bounded nuclear norm regularization. *Brief. Bioinform.* 22 (6), bbab328. doi:10.1093/bib/bbab328
- Cristino, A. S., Nourse, J., West, R. A., Sabdia, M. B., Law, S. C., Gunawardana, J., et al. (2019). EBV microRNA-BHRF1-2-5p targets the 3'UTR of immune checkpoint ligands PD-L1 and PD-L2. *Blood* 134 (25), 2261–2270. doi:10.1182/blood.2019000889
- Dai, X., and Tan, C. (2015). Combination of microRNA therapeutics with small-molecule anticancer drugs: Mechanism of action and co-delivery nanocarriers. *Adv. Drug Deliv. Rev.* 81, 184–197. doi:10.1016/j.addr.2014.09.010
- Denzler, R., McGeary, S. E., Title, A. C., Agarwal, V., Bartel, D. P., and Stoffel, M. (2016). Impact of MicroRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. *Mol. Cell* 64 (3), 565–579. doi:10.1016/j.molcel.2016.09.027
- Deyle, K., Kong, X. D., and Heinis, C. (2017). Phage selection of cyclic peptides for application in research and drug development. *Acc. Chem. Res.* 50 (8), 1866–1874. doi:10.1021/acs.accounts.7b00184
- Do Amaral, G., Planello, A. C., Borgato, G., de Lima, D. G., Guimarães, G. N., Marqueso, M. R., et al. (2019). 5-Aza-CdR promotes partial MGMT demethylation and modifies expression of different genes in oral squamous cell carcinoma. *Oral Surg Oral Med Oral Pathol Oral Radiol* 127 (5), 424–432. doi:10.1016/j.oooo.2019.01.006
- Dragomir, M. P., Knutsen, E., and Calin, G. A. (2021). Classical and noncanonical functions of miRNAs in cancers. *Trends Genet.* 38, 379–394. doi:10.1016/j.tig.2021.10.002
- Gam, J. J., Babb, J., and Weiss, R. (2018). A mixed antagonistic/synergistic miRNA repression model enables accurate predictions of multi-input miRNA sensor activity. *Nat. Commun.* 9 (1), 2430. doi:10.1038/s41467-018-04575-0
- Geng, B., and Craig, T. J. (2021). Small molecule drugs for atopic dermatitis, rheumatoid arthritis, and hereditary angioedema. *Ann. Allergy Asthma Immunol.* 128, 263–268. doi:10.1016/j.anaai.2021.10.015
- Ghini, F., Rubolino, C., Climent, M., Simeone, I., Marzi, M. J., and Nicassio, F. (2018). Endogenous transcripts control miRNA levels and activity in mammalian cells by target-directed miRNA degradation. *Nat. Commun.* 9 (1), 3119. doi:10.1038/s41467-018-05182-9
- Gorbea, C., Mosbruger, T., and Cazalla, D. (2017). A viral Sm-class RNA base-pairs with mRNAs and recruits microRNAs to inhibit apoptosis. *Nature* 550 (7675), 275–279. doi:10.1038/nature24034
- Gottlieb, A., Stein, G. Y., Rupp, E., and Sharan, R. (2011). Predict: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7, 496. doi:10.1038/msb.2011.26
- Habib, G., and Qureshi, S. (2020). Optimization and acceleration of convolutional neural networks: A survey. *J. King Saud Univ. - Comput. Inf. Sci.* 34, 4244. doi:10.1016/j.jksuci.2020.10.004
- Hammond, S. M. (2015). An overview of microRNAs. *Adv. Drug Deliv. Rev.* 87, 3–14. doi:10.1016/j.addr.2015.05.001
- Haniff, H. S., Liu, X., Tong, Y., Meyer, S. M., Knerr, L., Lemurell, M., et al. (2021). A structure-specific small molecule inhibits a miRNA-200 family member precursor and reverses a type 2 diabetes phenotype. *Cell Chem. Biol.* 29, 300–311.e10. doi:10.1016/j.chembiol.2021.07.006
- Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* 125 (39), 11853–11865. doi:10.1021/ja036030u
- Healy, N., Schiff, R., Osborne, C. K., and Kerin, M. (2012). Mirnas: Small molecules, big players in tamoxifen resistance in breast cancer. *Int. J. Surg.* 10 (8), S4. doi:10.1016/j.ijsu.2012.06.025
- Iwata, T., Mizuno, N., Nagahara, T., Kaneda-Ikeda, E., Kajiya, M., Sasaki, S., et al. (2021). Cytokines regulate stemness of mesenchymal stem cells via miR-628-5p during periodontal regeneration. *J. Periodontol.* 93, 269–286. doi:10.1002/jper.21-0064
- Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104. doi:10.1093/nar/gkn714
- Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2011). DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 39, D1035–D1041. doi:10.1093/nar/gkq1126
- Kumar Kingsley, S. M., and Vishnu Bhat, B. (2017). Role of MicroRNAs in the development and function of innate immune cells. *Int. Rev. Immunol.* 36 (3), 154–175. doi:10.1080/08830185.2017.1284212
- Kumari, R., Kumar, S., and Kant, R. (2018). Role of circulating miRNAs in the pathophysiology of CVD: As a potential biomarker. *Gene Rep.* 13, 146–150. doi:10.1016/j.genrep.2018.10.003
- Lai-Kwon, J., Tiu, C., Pal, A., Khurana, S., and Minchom, A. (2021). Moving beyond epidermal growth factor receptor resistance in metastatic non-small cell lung cancer - a drug development perspective. *Crit. Rev. Oncol. Hematol.* 159, 103225. doi:10.1016/j.critrevonc.2021.103225
- Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75 (5), 843–854. doi:10.1016/0092-8674(93)90529-y
- Lemaire, M., Chabot, G. G., Raynal, N. J., Momparler, L. F., Hurtubise, A., Bernstein, M. L., et al. (2008). Importance of dose-schedule of 5-aza-2'-deoxycytidine for epigenetic therapy of cancer. *BMC Cancer* 8, 128. doi:10.1186/1471-2407-8-128
- Li, X., Rao, S., Wang, Y., and Gong, B. (2004). Gene mining: A novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. *Nucleic Acids Res.* 32 (9), 2685–2694. doi:10.1093/nar/gkh563
- Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi:10.1093/nar/gkt1023
- Li, X., Li, X., Liao, D., Wang, X., Wu, Z., Nie, J., et al. (2015). Elevated microRNA-23a expression enhances the chemoresistance of colorectal cancer cells with microsatellite instability to 5-fluorouracil by directly targeting ABCF1. *Curr. Protein Pept. Sci.* 16 (4), 301–309. doi:10.2174/138920371604150429153309
- Li, J., Peng, D., Xie, Y., Dai, Z., Zou, X., and Li, Z. (2021). Novel potential small molecule-miRNA-cancer associations prediction model based on fingerprint, sequence, and clinical symptoms. *J. Chem. Inf. Model.* 61 (5), 2208–2219. doi:10.1021/acs.jcim.0c01458

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.

- Liu, X., Zhu, F., Ma, X., Tao, L., Zhang, J., Yang, S., et al. (2011). The therapeutic target database: An internet resource for the primary targets of approved, clinical trial and experimental drugs. *Expert Opin. Ther. Targets* 15 (8), 903–912. doi:10.1517/14728222.2011.586635
- Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., et al. (2013). SM2miR: A database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* 29 (3), 409–411. doi:10.1093/bioinformatics/bts698
- Liu, R., Lu, Z., Gu, J., Liu, J., Huang, E., Liu, X., et al. (2018). MicroRNAs 15A and 16-1 activate signaling pathways that mediate chemotaxis of immune regulatory B cells to colorectal tumors. *Gastroenterology* 154 (3), 637–651. e637. doi:10.1053/j.gastro.2017.09.045
- Liu, F., Peng, L., Tian, G., Yang, J., Chen, H., Hu, Q., et al. (2020). Identifying small molecule-miRNA associations based on credible negative sample selection and random walk. *Front. Bioeng. Biotechnol.* 8, 131. doi:10.3389/fbioe.2020.00131
- Longley, D. B., Harkin, D. P., and Johnston, P. G. (2003). 5-fluorouracil: Mechanisms of action and clinical strategies. *Nat. Rev. Cancer* 3 (5), 330–338. doi:10.1038/nrc1074
- Lu, T. X., and Rothenberg, M. E. (2018). MicroRNA. *J. Allergy Clin. Immunol.* 141 (4), 1202–1207. doi:10.1016/j.jaci.2017.08.034
- Lv, S., Li, Y., Wang, Q., Ning, S., Huang, T., Wang, P., et al. (2012). A novel method to quantify gene set functional association based on gene ontology. *J. R. Soc. Interface* 9 (70), 1063–1072. doi:10.1098/rsif.2011.0551
- Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* 31 (22), 3638–3644. doi:10.1093/bioinformatics/btv417
- Meng, F., Yang, X., and Zhou, C. (2014). The augmented Lagrange multipliers method for matrix completion from corrupted samplings with application to mixed Gaussian-impulse noise removal. *Plos One* 9 (9), e108125. doi:10.1371/journal.pone.0108125
- Monroig, P. d. C., Chen, L., Zhang, S., and Calin, G. A. (2015). Small molecule compounds targeting miRNAs for cancer therapy. *Adv. Drug Deliv. Rev.* 81, 104–116. doi:10.1016/j.addr.2014.09.002
- Nair, A., Chung, H. C., Sun, T., Tyagi, S., Dobrolecki, L. E., Dominguez-Vidana, R., et al. (2018). Combinatorial inhibition of PTPN12-regulated receptors leads to a broadly effective therapeutic strategy in triple-negative breast cancer. *Nat. Med.* 24 (4), 505–511. doi:10.1038/nm.4507
- Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing* 452, 48–62. doi:10.1016/j.neucom.2021.03.091
- Oh, J. Y., Choi, G. E., Lee, H. J., Jung, Y. H., Chae, C. W., Kim, J. S., et al. (2019). 17 β -Estradiol protects mesenchymal stem cells against high glucose-induced mitochondrial oxidants production via Nrf2/Sirt3/MnSOD signaling. *Free Radic. Biol. Med.* 130, 328–342. doi:10.1016/j.freeradbiomed.2018.11.003
- Parsons, J., Castaldi, M. P., Dutta, S., Dibrov, S. M., Wyles, D. L., and Hermann, T. (2009). Conformational inhibition of the hepatitis C virus internal ribosome entry site RNA. *Nat. Chem. Biol.* 5 (11), 823–825. doi:10.1038/nchembio.217
- Peng, Y., Ganesh, A., Wright, J., Xu, W., and Ma, Y. (2012). Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (11), 2233–2246. doi:10.1109/TPAMI.2011.282
- Qu, J., Chen, X., Sun, Y. Z., Li, J. Q., and Ming, Z. (2018). Inferring potential small molecule-miRNA association based on triple layer heterogeneous network. *J. Cheminform.* 10 (1), 30. doi:10.1186/s13321-018-0284-9
- Qu, J., Chen, X., Sun, Y. Z., Zhao, Y., Cai, S. B., Ming, Z., et al. (2019). *In silico* prediction of small molecule-miRNA associations based on the HeteSim algorithm. *Mol. Ther. Nucleic Acids* 14, 274–286. doi:10.1016/j.omtn.2018.12.002
- Rachoń, D., Myśliwska, J., Suchecka-Rachoń, K., Wieckiewicz, J., Myśliwski, A., and Myśliwski, A. (2002). Effects of oestrogen deprivation on interleukin-6 production by peripheral blood mononuclear cells of postmenopausal women. *J. Endocrinol.* 172 (2), 387–395. doi:10.1677/joe.0.1720387
- Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., et al. (2010). PhenoMiR: A knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.* 11 (1), R6. doi:10.1186/gb-2010-11-1-r6
- Rupaimoole, R., and Slack, F. J. (2017). MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* 16 (3), 203–222. doi:10.1038/nrd.2016.246
- Saikia, M., Paul, S., and Chakraborty, S. (2020). Role of microRNA in forming breast carcinoma. *Life Sci.* 259, 118256. doi:10.1016/j.lfs.2020.118256
- Seth, P. P., Miyaji, A., Jefferson, E. A., Sannes-Lowery, K. A., Osgood, S. A., Propp, S. S., et al. (2005). SAR by MS: Discovery of a new class of RNA-binding small molecules for the hepatitis C virus: Internal ribosome entry site IIA subdomain. *J. Med. Chem.* 48 (23), 7099–7102. doi:10.1021/jm050815o
- Shang, J., Yang, F., Wang, Y., Wang, Y., Xue, G., Mei, Q., et al. (2014). MicroRNA-23a antisense enhances 5-fluorouracil chemosensitivity through APAF-1/caspase-9 apoptotic pathway in colorectal cancer cells. *J. Cell. Biochem.* 115 (4), 772–784. doi:10.1002/jcb.24721
- Singh, S., Sharma, A., and Chauhan, V. K. (2021). Online handwritten Gurmukhi word recognition using fine-tuned Deep Convolutional Neural Network on offline features. *Mach. Learn. Appl.* 5, 100037. doi:10.1016/j.mlwa.2021.100037
- Sun, L. Y., Wang, N., Ban, T., Sun, Y. H., Han, Y., Sun, L. L., et al. (2014). MicroRNA-23a mediates mitochondrial compromise in estrogen deficiency-induced concentric remodeling via targeting PGC-1 α . *J. Mol. Cell. Cardiol.* 75, 1–11. doi:10.1016/j.yjmcc.2014.06.012
- Tagliafierro, L., Glenn, O. C., Zamora, M. E., Beach, T. G., Woltjer, R. L., Lutz, M. W., et al. (2017). Genetic analysis of α -synuclein 3' untranslated region and its corresponding microRNAs in relation to Parkinson's disease compared to dementia with Lewy bodies. *Alzheimers Dement.* 13 (11), 1237–1250. doi:10.1016/j.jalz.2017.03.001
- Thomou, T., Mori, M. A., Dreyfuss, J. M., Konishi, M., Sakaguchi, M., Wolfrum, C., et al. (2017). Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature* 542 (7642), 450–455. doi:10.1038/nature21365
- Thorne, N., Inglese, J., and Auld, D. S. (2010). Illuminating insights into firefly luciferase and other bioluminescent reporters used in chemical biology. *Chem. Biol.* 17 (6), 646–657. doi:10.1016/j.chembiol.2010.05.012
- Tse, J., Martin-McNulty, B., Halks-Miller, M., Kauser, K., DelVecchio, V., Vergona, R., et al. (1999). Accelerated atherosclerosis and premature calcified cartilaginous metaplasia in the aorta of diabetic male Apo E knockout mice can be prevented by chronic treatment with 17 β -estradiol. *Atherosclerosis* 144 (2), 303–313. doi:10.1016/S0021-9150(98)00325-6
- Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine* 28 (2), 52–68. doi:10.1109/msp.2010.939739
- Wang, C. C., and Chen, X. (2019). A unified framework for the prediction of small molecule-MicroRNA association based on cross-layer dependency inference on multilayered networks. *J. Chem. Inf. Model.* 59 (12), 5281–5293. doi:10.1021/acs.jcim.9b00667
- Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633. doi:10.1093/nar/gkp456
- Wang, C. C., Chen, X., Qu, J., Sun, Y. Z., and Li, J. Q. (2019). Rfsmma: A new computational model to identify and prioritize potential small molecule-MiRNA associations. *J. Chem. Inf. Model.* 59 (4), 1668–1679. doi:10.1021/acs.jcim.9b00129
- Wang, X., Zhao, Y., and Pourpanah, F. (2020). Recent advances in deep learning. *Int. J. Mach. Learn. Cybern.* 11 (4), 747–750. doi:10.1007/s13042-020-01096-5
- Wang, S. H., Wang, C. C., Huang, L., Miao, L. Y., and Chen, X. (2021). Dual-Network Collaborative Matrix Factorization for predicting small molecule-miRNA associations. *Brief. Bioinform.* 23, bbab500. doi:10.1093/bib/bbab500
- Wen, D., Danquah, M., Chaudhary, A. K., and Mahato, R. I. (2015). Small molecules targeting microRNA for cancer therapy: Promises and obstacles. *J. Control. Release* 219, 237–247. doi:10.1016/j.jconrel.2015.08.011
- Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell* 75 (5), 855–862. doi:10.1016/0092-8674(93)90530-4
- Xia, X., Wang, Y., Huang, Y., Zhang, H., Lu, H., and Zheng, J. C. (2019). Exosomal miRNAs in central nervous system diseases: Biomarkers, pathological mediators, protective factors and therapeutic agents. *Prog. Neurobiol.* 183, 101694. doi:10.1016/j.pneurobio.2019.101694
- Yekkirala, A. S., Roberson, D. P., Bean, B. P., and Woolf, C. J. (2017). Breaking barriers to novel analgesic drug development. *Nat. Rev. Drug Discov.* 16 (8), 545–564. doi:10.1038/nrd.2017.87
- Yin, J., Chen, X., Wang, C. C., Zhao, Y., and Sun, Y. Z. (2019). Prediction of small molecule-MicroRNA associations by sparse learning and heterogeneous graph inference. *Mol. Pharm.* 16 (7), 3157–3166. doi:10.1021/acs.molpharmaceut.9b00384
- Yu, A. M., Choi, Y. H., and Tu, M. J. (2020). RNA drugs and RNA targets for small molecules: Principles, progress, and challenges. *Pharmacol. Rev.* 72 (4), 862–898. doi:10.1124/pr.120.019554
- Zeng, X., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., et al. (2020). Network-based prediction of drug-target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* 36 (9), 2805–2812. doi:10.1093/bioinformatics/btaa010
- Zhong, H., Zhou, Y., Xu, Q., Yan, J., Zhang, X., Zhang, H., et al. (2021). Low expression of miR-19a-5p is associated with high mRNA expression of diacylglycerol O-acyltransferase 2 (DGAT2) in hybrid tilapia. *Genomics* 113 (4), 2392–2399. doi:10.1016/j.ygeno.2021.05.016



OPEN ACCESS

EDITED BY

Yuanpeng Janet Huang,
Rensselaer Polytechnic Institute,
United States

REVIEWED BY

Marcus Fischer,
St. Jude Children's Research Hospital,
United States
Lim Heo,
Michigan State University, United States
Swapna Gurla,
Rensselaer Polytechnic Institute,
United States

*CORRESPONDENCE

Davide Sala,
✉ davide.sala@uni-leipzig.de
Jens Meiler,
✉ jens@meilerlab.org

SPECIALTY SECTION

This article was submitted to Biological
Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 12 December 2022

ACCEPTED 31 January 2023

PUBLISHED 16 February 2023

CITATION

Sala D, Hildebrand PW and Meiler J (2023),
Biasing AlphaFold2 to predict GPCRs and
kinases with user-defined functional or
structural properties.
Front. Mol. Biosci. 10:1121962.
doi: 10.3389/fmolb.2023.1121962

COPYRIGHT

© 2023 Sala, Hildebrand 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.

Biasing AlphaFold2 to predict GPCRs and kinases with user-defined functional or structural properties

Davide Sala^{1*}, Peter W. Hildebrand² and Jens Meiler^{1,3,4*}

¹Institute of Drug Discovery, Faculty of Medicine, University of Leipzig, Leipzig, Germany, ²Institute of Medical Physics and Biophysics, Faculty of Medicine, University of Leipzig, Leipzig, Germany, ³Center for Structural Biology, Vanderbilt University, Nashville, TN, United States, ⁴Department of Chemistry, Vanderbilt University, Nashville, TN, United States

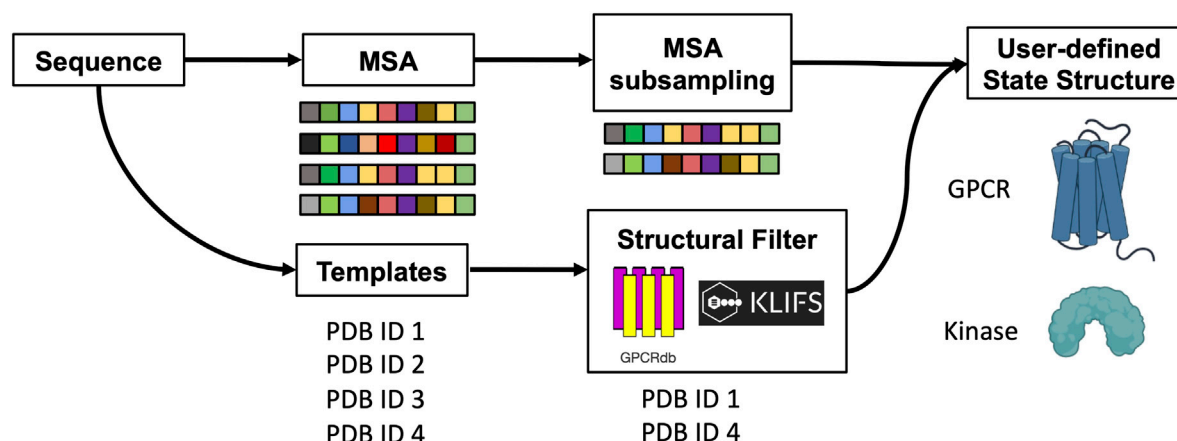
Determining the three-dimensional structure of proteins in their native functional states has been a longstanding challenge in structural biology. While integrative structural biology has been the most effective way to get a high-accuracy structure of different conformations and mechanistic insights for larger proteins, advances in deep machine-learning algorithms have paved the way to fully computational predictions. In this field, AlphaFold2 (AF2) pioneered *ab initio* high-accuracy single-chain modeling. Since then, different customizations have expanded the number of conformational states accessible through AF2. Here, we further expanded AF2 with the aim of enriching an ensemble of models with user-defined functional or structural features. We tackled two common protein families for drug discovery, G-protein-coupled receptors (GPCRs) and kinases. Our approach automatically identifies the best templates satisfying the specified features and combines those with genetic information. We also introduced the possibility of shuffling the selected templates to expand the space of solutions. In our benchmark, models showed the intended bias and great accuracy. Our protocol can thus be exploited for modeling user-defined conformational states in an automatic fashion.

KEYWORDS

AlphaFold, GPCRs (G-protein-coupled receptors), kinases, structure prediction, protein function

Introduction

X-ray crystallography and cryogenic electron microscopy (cryo-EM) are two widely used techniques for determining the detailed structures of biomolecules at the atomic level (Vénien-Bryan et al., 2017; Wang and Wang, 2017). For structure-based drug discovery and design, having at least one high-accuracy structure is essential (Congreve et al., 2020). Despite recent advances in technology have made more protein structures available (Callaway, 2020), their experimental determination is still a difficult and costly process with a high risk of failure (Lyumkis, 2019). In fact, experimental protein structures represent only a small fraction of the complete set of known protein sequences (The Uniprot Consortium, 2019; Burley et al., 2021). Furthermore, one structure only represents a snapshot of a certain protein state, and may not necessarily be sufficient to understand the overall mechanism of operation. This limitation has important implications for drug discovery, especially for common drug targets such as G-protein-coupled receptors (GPCRs) and kinases, which are known to modulate cellular behavior by switching among multiple structurally different functional states (Attwood et al., 2021; Yang et al., 2021).



SCHEME 1

Schematic representation of the method. The protein sequence is used to collect MSA and templates. A subset of sequences and templates are collected by randomly subsampling the MSA and by interrogating webserver to filter templates with user-defined structural properties. The predicted ensemble of structures is biased toward the intended conformation.

The 14th edition of Critical Assessment of protein Structure Prediction (CASP14) has recognized AlphaFold2 (AF2) for its impressive accuracy in predicting monomeric protein structures *de novo* (Jumper et al., 2021). AF2 makes it straightforward to predict a protein structure from a protein sequence and has provided millions of protein models with estimated accuracy (Tunyasuvunakool et al., 2021). Since the emergence of AF2, a number of deep learning-based methods have been developed with the same goal of predicting protein structures at experimental accuracy (AlQuraishi, 2021; Baek et al., 2021; Chowdhury et al., 2022; Lin et al., 2022). Among them, RoseTTAFold was the first approach that was able to predict both active and inactive GPCR conformations by using templates in a uniform functional state, outperforming comparative homology modeling methods (Baek et al., 2021). This achievement has sparked interest in developing workflows to predict multiple native conformations of a protein target with the state-of-the-art AF2 implementation.

To date, a number of AF2 customizations that adopted different concepts are available (Del Alamo et al., 2022; Heo and Feig, 2022; Stein and Mchaourab, 2022; Wayment-Steele et al., 2022). Del Alamo and co-authors took advantage of a shallow multiple sequence alignment (sMSA) to collect an ensemble of structures, among which multiple native conformations of GPCRs and transporters were identified (Del Alamo et al., 2022). Alternatively, SPEACH-AF (hereafter SPEACH) masked multiple positions in the multiple sequence alignment (MSA) to switch the prediction toward alternative conformational states that were less represented in the MSA (Stein and Mchaourab, 2022). Another protocol removed the MSA (noMSA) and prepared a local database of state-annotated GPCRs to perform AF2 template-based modeling (Heo and Feig, 2022). These methods for sampling conformational changes in proteins have shown great potential, but also have some limitations, such as a reduced breadth of sampled conformations or a high dependence on the structural features of selected templates.

Here, we update our previous protocol (sMSA) to facilitate the collection of templates with user-defined functional or structural properties of GPCRs and kinases. Templates are automatically filtered and retrieved from an annotated database in accord with

the specified functional or structural criteria. Through a calibrated balancing of genetic and template-based features, our protocol samples equal or better active GPCR states than all the peer-reviewed methods for sampling alternative states. On a difficult target, randomizing templates to explore the available structural space significantly improved accuracy. In modeling kinase conformations, our protocol enriched the predicted ensemble with models carrying user-defined structural features.

Methods

We updated our previous modified ColabFold version (Del Alamo et al., 2022; Mirdita et al., 2022) and our python interface to allow users to specify functional or structural properties of templates for modeling GPCRs and kinases. The new implementation and accompanying documentation can be found at https://github.com/meilerlab/AF2_GPCR_Kinase.

GPCRs benchmark

Target PDBs for Lutropin-choriogonadotropic hormone receptor (LSHR), Melatonin receptor type 1A (MTR1A), Prostaglandin E2 receptor EP4 subtype (PE2R4), Beta-1 adrenergic receptor (ADRB1), Parathyroid hormone/parathyroid hormone-related peptide receptor (PTH1R) and Frizzled-7 (FZD7) were 7FII, 7VGY, 7D7M, 7JJO, 6NBF and 6WW2 respectively (Su et al., 2020; Duan et al., 2021; Nojima et al., 2021; Wang et al., 2022). The protein regions corresponding to transmembrane helices (TM-RMSD) were retrieved from GPCRdb (Kooistra et al., 2021). Four workflows were evaluated to predict the active state of GPCRs: ActTemp+sMSA was run with eight sequence clusters and 16 extra cluster sequences combined with the automatic detection of “Active” templates not belonging to the same subfamily. Those number of sequences were chosen to provide evolution-based structural information without changing the activation state inferred from templates. In particular, the script takes the

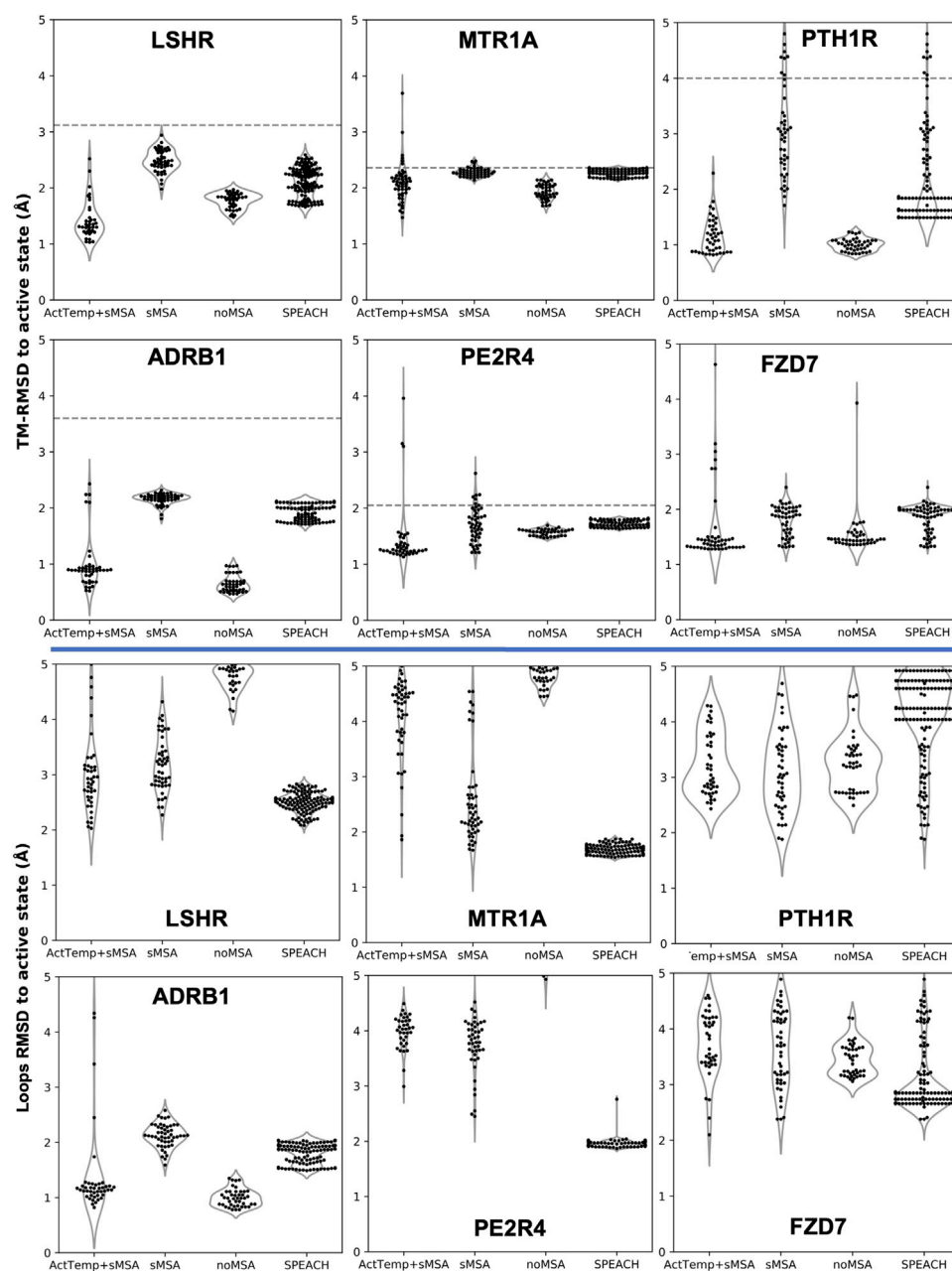


FIGURE 1

AF2 accuracy in predicting active state GPCRs with different protocols. ActTemp+sMSA was predicted with templates in the active state and a shallow MSA, sMSA with a shallow MSA only, noMSA without a MSA for templates aligned regions, SPEACH with a sliding window masked MSA. TM-RMSD between experimental active and inactive structures is shown as a dashed line.

AF2 generated list of templates ranked by sequence identity and filters out all the PDBs not matching the user-defined activation state in accord to GPCRdb annotation. Here, the top 4 templates were used. For LSHR, MTR1A, PE2R4, ADRB1, PTH1R and FZD7 those were (sequence identity in parenthesis): 6H7L_A (20.6%)-6IBL_A(15.9%)-6K41_R(23.1%)-6K42_R(23.7%), 6H7L_A(26.6%)-7P00_R(23.7%)-6IBL_A(19.9%)-7RMG_R(22.7%), 7E32_R(21.9%)-7CKY_R(20.4%)-7CKW_R(19.2%)-7JVP_R(20.4%), 6MXT_A(37.1%)-7CKY_R(36.8%)-7CKW_R(36.8%)-7JVP_R(37.4%), 7F16_R(35.8%)-6M1I_A(26.0%)-6P9Y_R(30.5%)-6VN7_R(32.0%) and 6XBM_R(25.7%)-6XBK_R(19.0%)-6OT0_R(27.2%)-7D76_R(18.3%) respectively. Other AF2 parameters

were kept as in our previous pipeline - named sMSA - that used 16 sequence clusters and 32 extra cluster sequences without any template and no recycling (Del Alamo et al., 2022). To remove the MSA (noMSA run), the same implementation published previously was adopted (Heo and Feig, 2022). These runs were then carried out using the GPCRdb API (Application Programming Interface) rather than a local GPCR database to avoid mismatches between the pool of available templates. The SPEACH protocol was applied with a sliding window of 10 masked residues (Stein and Mchaourab, 2022). Thus, the number of models collected with SPEACH was higher than the 50 models collected with other protocols. Unfolded models were discharged.

To assess the impact of randomizing templates, the inactive state structure of Leukotriene B4 receptor 1 (LT4R1, PDB 7K15) was used as a target (Michaelian et al., 2021). The MSA for the aligned regions was removed, and 50 models were generated with and without randomizing templates. The templates used for the models without randomization were 6VI4_A(27.5%)-4ZUD_A(20.0%)-4YAY_A(20.1%)-4N6H_A(20.2%).

EIF2AK4 kinase benchmark

All the experimental structures available were absent from the AF2 training set. Models were predicted by using exactly the same ActTemp+sMSA protocol adopted for GPCRs predictions but with 20 templates instead of 4. The DFG, aC_helix, and Salt bridge $K^{III.17}$ and $E^{aC.24}$ structural features as well as the activation loop orientation used to collect templates were defined according to the KLIF database (Kanev et al., 2021). Unfolded models were discarded.

Results

The original pipeline that was developed to sample alternative conformations was expanded to improve the prediction of GPCRs and kinases in a specific conformational state. Here, templates are selected through structural filters and the resulting structures are combined with genetic information coming from a subset of the MSA to predict models carrying the desired structural properties at high accuracy (Scheme 1). In particular, users can now specify the activation state of GPCRs and the script will look for templates that match that state or are bound to a signaling protein. To do so, one of the following labels must be declared: “Active”, “Inactive”, “Intermediate”, “G protein”, “Arrestin”. For kinases, users can select specific structural feature values and the script will search for templates that match those criteria. Allowed values for the corresponding structural feature are 1) DFG: “out”, “in”, “out-like”, “all”; 2) aC_helix: “out”, “in”, “all”; 3) Salt bridge $K^{III.17}$ $E^{aC.24}$: “yes”, “no”, “all” (McClendon et al., 2014). Optionally, the list of templates that pass the sequence and structural filters can be randomized to explore the available structural space.

In the sections below, we demonstrate how selecting templates in accord with functional or structural properties and combining those with genetic information can influence the predicted structural features of the models. We also show the results of randomizing templates on a difficult target.

Combining a shallow MSA with state-annotated templates achieves state-of-the-art accuracy in predicting GPCRs active state

Our new pipeline was used to predict GPCR models by combining a very shallow MSA with the automatic detection of the best 4 active templates from GPCRD (ActTemp+sMSA). The benchmark set of these GPCRs consisted of six proteins: LSHR, MTR1A, PE2R4, PTH1R, FZD7 and ADRB1. The first three class A receptors were predicted with the lowest accuracy in a broad benchmark in which the active state was modeled without MSA (Heo and Feig, 2022). PTH1R and FZD7 are members of class B and class F family, respectively. Instead, the active state of ADRB1 was included because the inactive

state was part of the neural networks training set. Thus, we targeted the active state with the specific aim of assessing the ability of our implementation to overcome the neural networks preference for the inactive state. For each method, we measured the accuracy as Ca-RMSD (root-mean-square deviation) of the transmembrane helices (TM-RMSD) as well as of the loops with respect to the experimentally determined structure. Our implementation was compared to AF2 workflows designed to sample alternative protein conformations. ActTemp+sMSA consistently generated models with near or subangstrom accuracy for all the GPCRs TM helices, showing state-of-the-art accuracy (Figure 1). Interestingly, our approach and noMSA were the only methods able to overcome the ADRB1 inactive state bias and accurately model the active state with an average accuracy of 0.5 Å on TM helices and 1 Å on loops. On the remaining targets, loops were in general better modeled by protocol leveraging on genetic information than those on templates. In particular, SPEACH—that does not reduce the MSA depth—has shown a consistent good accuracy. By comparing the two methods that leverage on templates (ActTemp+sMSA and noMSA), loops were on average better modeled by the former probably due to the contribution of genetic information compensating for missing or poorly conserved loops in the selected templates.

Given the separated evaluation of TM helices and loops accuracy, we measured the pTM score per model and assessed Spearman correlation between pTM and global RMSD for each ensemble (Figure 2). Overall, ActTemp+sMSA generated equally or better active state models than noMSA mainly due to higher accuracy in loops modeling. Within each ensemble, correlation is often reasonable and more importantly the best models are often assigned with the highest pTM scores with very few exceptions. However, pTM scores between the two protocols do not seem correlating well with accuracy. In other words, pTM scores often cannot correctly discriminate which protocol generated best active structures.

Shuffling templates in a homogenous functional state can improve accuracy

Given that subsampling the sequence space (i.e., the MSA) returns different models, we hypothesized that randomly selecting a subset of templates can potentially yield more accurate models. To test this, we removed the genetic information within the AF2 pipeline and generated 50 models with and without randomizing inactive templates. For each model, our script selected 4 random inactive state structures from GPCRD that passed the sequence similarity filter. Accuracy was measured as TM-RMSD from the inactive state structure of LT4R1 (PDB 7K15). The exploration of the structural space defined by the ensemble of all the inactive templates resulted in more accurate models compared to using the top 4 templates (Figure 3A).

The superposition of the best model in the two ensembles shows improved fitting of the long TM7 helix and better modeling of TM1 and TM6 when using random templates (Figure 3B).

User-defined structural features to bias kinase modeling

The concept of allowing users to define structural features of GPCR templates was also applied to kinases using the KLIF webserver

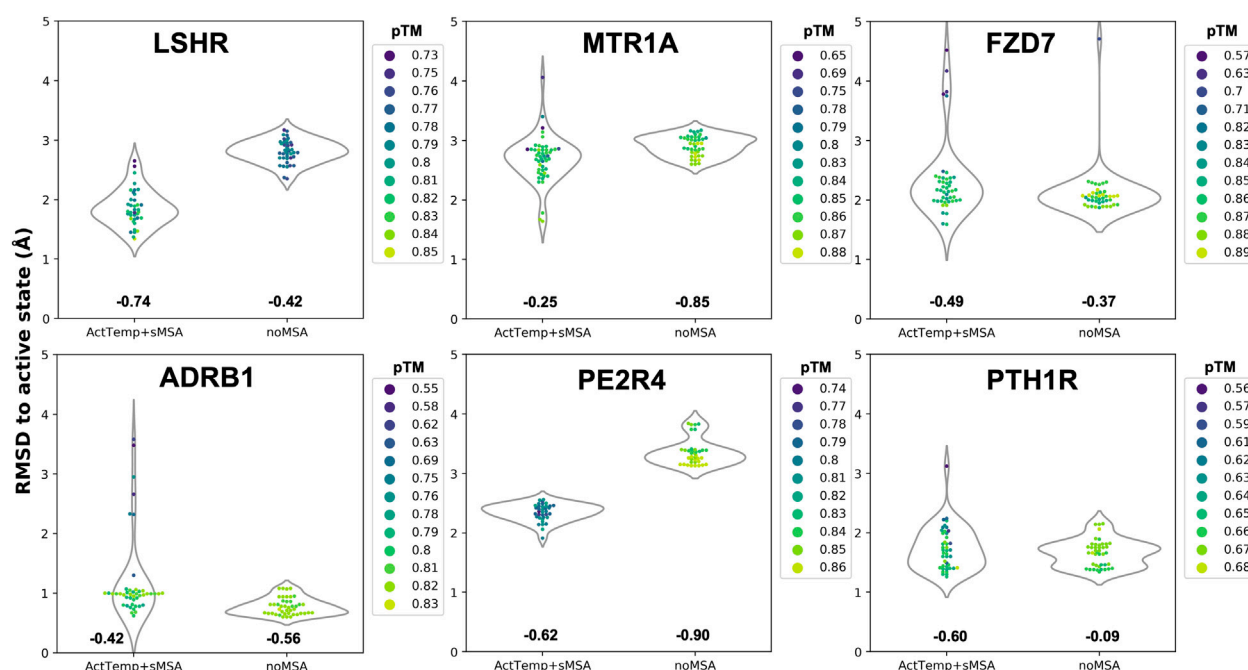


FIGURE 2

Correlation between pTM and global RMSD per target. Spearman correlation for each ensemble is indicated below each violin plot.

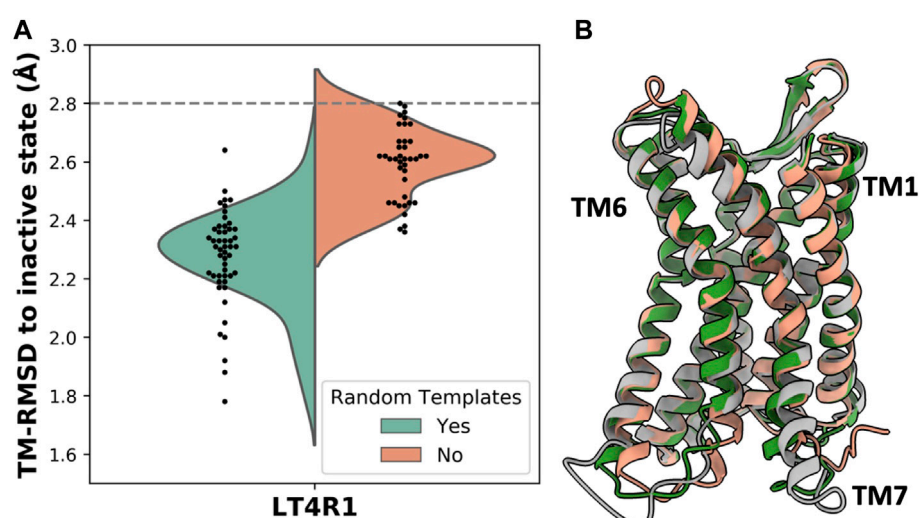


FIGURE 3

Accuracy in predicting the LT4R1 inactive state with and without randomizing templates. (A) TM-RMSD distribution of models. TM-RMSD between experimental active and inactive structures is shown as a dashed line. (B) Superposition of the best model from the random templates ensemble (green) and without randomizing templates (orange) to the experimental structure (gray).

(Kanev et al., 2021). We implemented the possibility to choose templates differing on three conformational properties: DFG, α C-helix (α C_H), and salt bridge $K^{III.17}E^{\alpha C.24}$. The script automatically selects and retrieves templates satisfying user-defined values for these three structural criteria. We assessed the effect on the predicted conformations by modeling the EIF2AK4 (GCN2) kinase. We generated four ensembles of 50 models each with the following

templates biased features: 1) “DFG=all/ α C_H=all”, i.e. all templates are allowed; 2) “DFG=in/ α C_H=in” and 3) “DFG=in/ α C_H=out” which differ in the α C-helix position regardless of its rotation, i.e. templates have DFG=in but differ in the α C_H conformation; 4) “DFG=out/ α C_H=all”, all the selected templates have DFG=out but α C_H is allowed in any conformation. Because DFG is a multi-criteria parameter, instead of measuring whether the predicted DFG corresponds to the selected

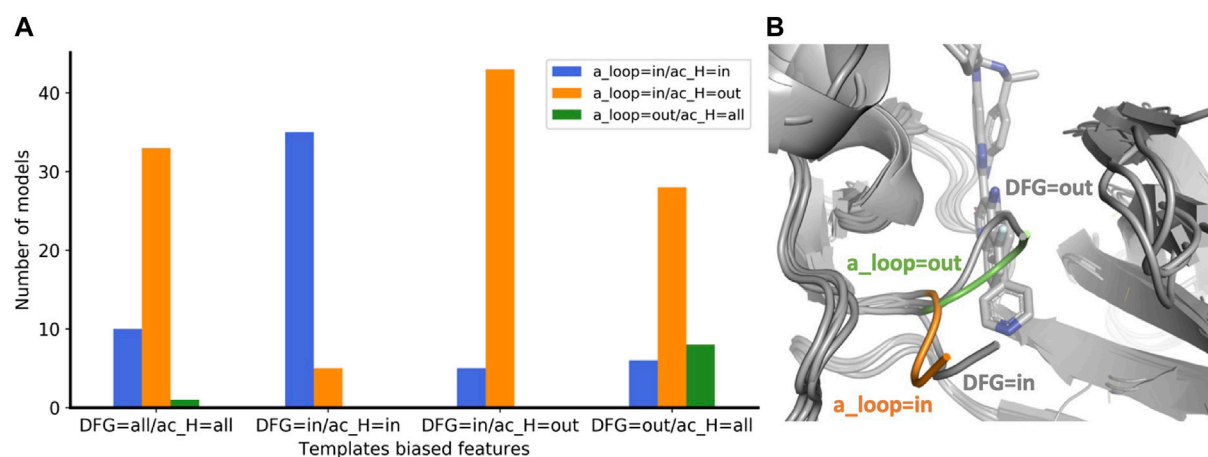


FIGURE 4

(A) Enrichment of eif2k4 kinase models with structural properties corresponding to the biased template features used. The four ensembles were calculated with a different “DFG/ac_H” templates bias. For each ensemble, the number of models with the three “a_loop/ac_H” conformational feature combinations are shown with a different color bar. (B) Superposition of two models with a_loop=in and a_loop=out to the two corresponding “DFG=in” and “DFG=out” experimental structures. DFG residues of models with “out” and “in” orientations are shown in green and orange, respectively. Experimental structures of eif2k4 are shown in gray.

DFG templates bias, we evaluated the activation loop (a_loop) position which is well-defined and mostly corresponds to DFG. Without biasing the prediction (DFG=all/ac_H=all), most of the models were found in the “a_loop=in/ac_H=out” conformation, while 20% of the pool was in the “a_loop=in/ac_H=in” conformation, and only one model was found with “a_loop=out” (Figure 4A). By biasing the prediction through the selection of ac_H=in and ac_H=out templates in two different ensembles (DFG=in/ac_H=in and DFG=in/ac_H=out), AF2 generated most of the models in agreement with the templates ac_H position. Accordingly, “DFG=in” templates generated only “a_loop=in” conformations (blue and orange bars) while in the only “DFG=out” ensemble we found a significant number of models carrying the “a_loop=out” conformation (green bar). The superposition of “a_loop=out” and “a_loop=in” models onto the corresponding experimental “DFG=out” (PDB 7QWK) and “DFG=in” structures (PDB 7QQ6) shows an excellent fitting of DFG loops, with a small discrepancy for ‘DFG/a_loop=out’ likely due to the presence of the inhibitor in the experimental structure (Figure 4B) (Maia de Oliveira et al., 2020).

Discussion

The prediction of user-defined conformational states of proteins has been a challenge even after the advent of AF2. Previous workflows attempting to solve this problem either do not explicitly predict user-defined structural properties or require the creation of state-annotated local structure databases (Del Alamo et al., 2022; Heo and Feig, 2022; Stein and Mchaourab, 2022; Wayment-Steele et al., 2022). In this work, we propose a pipeline that biases AF2 predictions toward the intended functional state of GPCRs or specific structural properties of kinases. One key aspect of our method is its simplicity in use. By leveraging on the API (Application Programming Interface) of two popular web servers, GPCRdb and KLIFS (Kanev et al., 2021; Kooistra et al., 2021), our

script filters templates according to pre-defined structural or functional parameters, allowing for a fully automatic selection of templates without the need for manual inspection or for downloading and updating of databases.

Our results in predicting the active structures of several challenging GPCRs show that combining a shallow multiple sequence alignment (MSA) with templates in a user-defined activation state (i.e. structure annotated as Active, Inactive or Intermediate) outperforms existing AF2 workflows. A direct comparison with models predicted without an MSA (noMSA) suggests that the balanced combination of genetic (MSA) and structural (templates) features may be crucial for achieving high accuracy, especially on loops that are usually less conserved and feature higher structural variance. This balanced mixture enables structural refinement of the desired conformational state while avoiding the overwhelming effect coming from a deep MSA, as previously reported (Del Alamo et al., 2022). Another advantage of a balanced mixture of genetic and structural information is its reduced sensitivity to neural network biases, i.e. the conformational preference of the neural network. In our benchmark, target conformations were four class A and one class B1 GPCRs for which inactive structures were more prevalent than active ones in the AF2 training set. Furthermore, the inactive structure of ADRB1 was directly part of the AF2 training set, thus representing a very strong bias. Indeed, protocols relying solely on genetic information (sMSA and SPEACH) were on average less accurate and completely missed the target conformation for ADRB1. On the other side, ActTemp + sMSA and noMSA depend on the presence of high-accuracy templates. Indeed, ADRB1 was predicted with an astonishing low RMSD value due to the high accuracy of the active state templates on both TM helices and loops.

Shuffling templates to predict the inactive state structure of LT4R1 generated better models than by taking the top four sequence identity templates in the inactive state. Regions that were better modeled were indeed different in the top four templates. Suggesting that despite a lower sequence identity, templates

randomly chosen from the remaining pool of inactive state structures may have been more suitable to model this conformational state. This kind of approach can be used to expand sampling without changing the desired structural features, like the activation state of a GPCR.

Our efforts to bias the prediction of a kinase toward user-defined structural properties exploited two important structural components that define its activation state: DFG and α C-helix. While the latter was easier to direct toward the intended position, the former was more difficult likely due to the neural network bias in the training set composition. Despite this, we successfully generated multiple models with “DFG=out” conformation. Given that “DFG=out” structures are needed for structure-based drug design and discovery of type-II inhibitors (Ung and Schlessinger, 2015), our script is well positioned to generate models carrying this crucial structural feature. Frequency of sampling the desired structural features may change protein by protein due to multiple factors such as neural network biases, templates features and MSA composition.

Our work expands the portfolio of AlphaFold2 customizations developed with the aim of predicting multiple conformational states of proteins. Our python interface facilitates the prediction of intended functional or structural properties of GPCRs and kinases and can be further extended to include more properties as needed. We also emphasize the importance that structure- and function-annotated databases had for this work. The expansion of existing databases to include additional annotations and the development of new protein family-based databases would improve or enable automatic calibrated modeling, respectively. This is particularly relevant for receptors and transporters that are known to span multiple conformations in their functional cycle. Together, curated databases and machine learning offer a powerful combination for high throughput modeling at high accuracy and, ultimately, for structure-based drug discovery (Sala et al., 2022).

Data availability statement

Models generated with the described protocol are made available at <https://doi.org/10.5281/zenodo.7602488>. The python script and corresponding documentation can be found at https://github.com/meilerlab/AF2_GPCR_Kinase.

References

- AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* 65, 1–8. Elsevier Current Trends. doi:10.1016/j.cbpa.2021.04.005
- Attwood, M. M., Fabbro, D., Sokolov, A. V., Knapp, S., and Schioth, H. B. (2021). Trends in kinase drug discovery: Targets, indications and inhibitor design. *Nat. Rev. Drug Discov.* 20, 839–861. doi:10.1038/s41573-021-00252-y
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754
- Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichtlow, G. V., et al. (2021). RCSB protein data bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* 49, D437–D451. doi:10.1093/nar/gkaa1038
- Callaway, E. (2020). Revolutionary cryo-EM is taking over structural biology. *Nature* 578, 201. doi:10.1038/d41586-020-00341-9
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* 40, 1617–1623. doi:10.1038/s41587-022-01432-w
- Congreve, M., de Graaf, C., Swain, N. A., and Tate, C. G. (2020). Impact of GPCR structures on drug discovery. *Cell* 181, 81–91. doi:10.1016/j.cell.2020.03.003
- Del Alamo, D., Sala, D., Mchaourab, H. S., and Meiler, J. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* 11, e75812. doi:10.7554/eLife.75751
- Duan, J., Xu, P., Cheng, X., Mao, C., Croll, T., He, X., et al. (2021). Structures of full-length glycoprotein hormone receptor signalling complexes. *Nature* 598, 688–692. doi:10.1038/s41586-021-03924-2
- Heo, L., and Feig, M. (2022). Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins Struct. Funct. Bioinforma.* 90, 1873–1885. doi:10.1002/prot.26382
- 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
- Kanev, G. K., de Graaf, C., Westerman, B. A., de Esch, I. J. P., and Kooistra, A. J. (2021). Klifs: An overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res.* 49, D562–D569. doi:10.1093/nar/gkaa895

Author contributions

DS and JM conceived the idea, with a contribution from PWH. DS designed the framework, wrote the code, performed the calculations, analyzed the data, prepared figures, and wrote the manuscript. JM and PWH further revised the manuscript.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1423, project number 421152132, subproject A07 and Z04. JM is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. The work was further supported by NIH NIGMS R01 GM080403, NIH NIHL R01 HL122010, and NIH NIDA R01 DA046138.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (SPP 2363, “Molecular Machine Learning”) for generous financial support. The authors would like to thank Dr. Ben Brown for useful discussions.

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.

- Kooistra, A. J., Mordalski, S., Pandey-Szekeres, G., Esguerra, M., Mamyrbekov, A., Munk, C., et al. (2021). GPCRdb in 2021: Integrating GPCR sequence, structure and function. *Nucleic Acids Res.* 49, D335–D343. doi:10.1093/nar/gkaa1080
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2022) Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*. doi:10.1101/2022.07.20.500902
- Lyumkis, D. (2019). Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* 294, 5181–5197. doi:10.1074/jbc.REV118.005602
- Maia de Oliveira, T., Korboukh, V., Caswell, S., Winter Holt, J. J., Lamb, M., Hird, A. W., et al. (2020). The structure of human GCN2 reveals a parallel, back-to-back kinase dimer with a plastic DFG activation loop motif. *Biochem. J.* 477, 275–284. doi:10.1042/BCJ20190196
- McClendon, C. L., Kornev, A. P., Gilson, M. K., and Taylor, S. S. (2014). Dynamic architecture of a protein kinase. *Proc. Natl. Acad. Sci.* 111, E4623–E4631. doi:10.1073/pnas.1418402111
- Michaelian, N., Sadybekov, A., Besserer-Offroy, E., Han, G. W., Krishnamurthy, H., Zamylny, B. A., et al. (2021). Structural insights on ligand recognition at the human leukotriene B4 receptor 1. *Nat. Commun.* 12, 2971. doi:10.1038/s41467-021-23149-1
- Mirdita, M., Schütze, K., Moriawaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi:10.1038/s41592-022-01488-1
- Nojima, S., Fujita, Y., Kimura, K. T., Nomura, N., Suno, R., Morimoto, K., et al. (2021). Cryo-EM structure of the prostaglandin E receptor EP4 coupled to G protein. *Structure* 29, 252–260. e6. doi:10.1016/j.str.2020.11.007
- Sala, D., Batebi, H., Ledwith, K., Hildebrand, P. W., and Meiler, J. (2022). Targeting *in silico* GPCR conformations with ultra-large library screening for hit discovery. *Trends Pharmacol. Sci.* S0165-6147, 00280–00282. doi:10.1016/j.tips.2022.12.006
- Stein, R. A., and Mchaourab, H. S. (2022). SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLOS Comput. Biol.* 18, e1010483. doi:10.1371/journal.pcbi.1010483
- 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 β_1 -adrenergic receptor. *Mol. Cell* 80, 59–71. e4. doi:10.1016/j.molcel.2020.08.001
- The Uniprot Consortium (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596. doi:10.1038/s41586-021-03828-1
- Ung, P. M. U., and Schlessinger, A. (2015). DFGmodel: Predicting protein kinase structures in inactive states for structure-based discovery of type-II inhibitors. *ACS Chem. Biol.* 10, 269–278. doi:10.1021/cb500696t
- Vénien-Bryan, C., Li, Z., Vuillard, L., and Boutin, J. A. (2017). Cryo-electron microscopy and X-ray crystallography: Complementary approaches to structural biology and drug discovery. *Acta Crystallogr. Sect. Struct. Biol. Commun.* 73, 174–183. doi:10.1107/S2053230X17003740
- Wang, H. W., and Wang, J. W. (2017). How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci.* 26, 32–39. doi:10.1002/pro.3022
- Wang, Q., Lu, Q., Guo, Q., Teng, M., Gong, Q., Li, X., et al. (2022). Structural basis of the ligand binding and signaling mechanism of melatonin receptors. *Nat. Commun.* 13, 454. doi:10.1038/s41467-022-28111-3
- Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L., and Kern, D. (2022). Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. doi:10.1101/2022.10.17.512570
- 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. *Signal Transduct. Target. Ther.* 6, 7. Nature Publishing Group, 1–27. doi:10.1038/s41392-020-00435-w



OPEN ACCESS

EDITED BY

Parimal Kar,
Indian Institute of Technology Indore,
India

REVIEWED BY

Kemal Yelekci,
Kadir Has University, Türkiye
Md. Fulbabu Sk,
University of Illinois at Urbana-Champaign,
United States

*CORRESPONDENCE

Junjian Hu,
✉ hujunjian79@163.com
Abdul Wadood,
✉ awadood@awkum.edu.pk

SPECIALTY SECTION

This article was submitted to Biological
Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 04 October 2022

ACCEPTED 11 January 2023

PUBLISHED 07 March 2023

CITATION

Samad A, Ajmal A, Mahmood A, Khurshid B,
Li P, Jan SM, Rehman AU, He P, Abdalla AN,
Umair M, Hu J and Wadood A (2023),
Identification of novel inhibitors for SARS-
CoV-2 as therapeutic options using
machine learning-based virtual screening,
molecular docking and MD simulation.
Front. Mol. Biosci. 10:1060076.
doi: 10.3389/fmolb.2023.1060076

COPYRIGHT

© 2023 Samad, Ajmal, Mahmood,
Khurshid, Li, Jan, Rehman, He, Abdalla,
Umair, Hu and Wadood. 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 novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation

Abdus Samad¹, Amar Ajmal¹, Arif Mahmood^{2,3}, Beenish Khurshid¹,
Ping Li⁴, Syed Mansoor Jan¹, Ashfaq Ur Rehman⁵, Pei He⁶,
Ashraf N. Abdalla⁷, Muhammad Umair⁸, Junjian Hu^{9*} and
Abdul Wadood^{1*}

¹Department of Biochemistry, Abdul Wali Khan University, Mardan, KPK, Pakistan, ²Center for Medical Genetics and Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China, ³Institute of Molecular Precision Medicine, Xiangya Hospital, Central South University, Changsha, Hunan, China, ⁴Institutes of Biomedical Sciences, Shanxi university, Taiyuan, China, ⁵Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, United States, ⁶Department of Obstetrics and Gynecology, Nanfang Hospital, Southern Medical University, Guangzhou, China, ⁷Department of Pharmacology and Toxicology, College of Pharmacy, Umm Al-Qura University, Makkah, Saudi Arabia, ⁸Department of Life Sciences, School of Science, University of Management and Technology (UMT), Lahore, Pakistan, ⁹Department of Central Laboratory, SSL Central Hospital of Dongguan City, Affiliated Dongguan Shilong People's Hospital of Southern Medical University, Dongguan, China

The new coronavirus SARS-CoV-2, which emerged in late 2019 from Wuhan city of China was regarded as causing agent of the COVID-19 pandemic. The primary protease which is also known by various synonymous i.e., main protease, 3-Chymotrypsin-like protease (3CL^{PRO}) has a vital role in the replication of the virus, which can be used as a potential drug target. The current study aimed to identify novel phytochemical therapeutics for 3CL^{PRO} by machine learning-based virtual screening. A total of 4,000 phytochemicals were collected from deep literature surveys and various other sources. The 2D structures of these phytochemicals were retrieved from the PubChem database, and with the use of a molecular operating environment, 2D descriptors were calculated. Machine learning-based virtual screening was performed to predict the active phytochemicals against the SARS-CoV-2 3CL^{PRO}. Random forest achieved 98% accuracy on the train and test set among the different machine learning algorithms. Random forest model was used to screen 4,000 phytochemicals which leads to the identification of 26 inhibitors against the 3CL^{PRO}. These hits were then docked into the active site of 3CL^{PRO}. Based on docking scores and protein-ligand interactions, MD simulations have been performed using 100 ns for the top 5 novel inhibitors, ivermectin, and the APO state of 3CL^{PRO}. The post-dynamic analysis i.e., Root means square deviation (RMSD), Root mean square fluctuation analysis (RMSF), and MM-GBSA analysis reveal that our newly identified phytochemicals form significant interactions in the binding pocket of 3CL^{PRO} and form stable complexes, indicating that these phytochemicals could be used as potential antagonists for SARS-CoV-2.

KEYWORDS

SARS-CoV-2, COVID, 19, machine learning, molecular docking, MD simulation, Corona virus

1 Introduction

SARS-CoV-2 is a single-strand RNA, positive sense, and enveloped beta coronavirus that causes respiratory, nervous, hepatic, and human gastrointestinal diseases (Tahir ul Qamar et al., 2020). Wuhan, a city in China, was the first city to be infected by the virus in December 2019 (Zhu et al., 2019; Zhou et al., 2020). COVID-19 outbreak was declared a pandemic by the World Health Organization (WHO). The infection spreads rapidly across the World. By the end of October 2020, more than 60 million people were infected by COVID-19, resulting in more than 1.4 million fatalities. The number of patients and fatalities was rising, posing a major threat to global health. High temperature, coughing, shortness of breath, and severe cases that can result in renal failure and even death are some of the symptoms of COVID-19 infections (Rothan and Byrareddy, 2020; Asif et al., 2022), until now, there is no effective treatment available yet.

SARS-CoV-2 is a member of the beta coronavirus family (Marty and Jones, 2020), usually, during the process of transcription, beta coronaviruses produce an 800 kDa polypeptide (Xu et al., 2020). The genome of the novel SARS-CoV-2 was recently sequenced and compared with those of existing coronaviruses (CoVs) by Wu et al. who identified that the novel SARS-CoV-2 belonged to the β -CoVs, which were initially discovered in bats and have now evolved to infect humans (Wu et al., 2020a). The SARS-CoV-2 genome is approximately 30 kb in size and is comprised of at least six open reading frames (ORFs) which are responsible for encoding the whole proteome of the virus. The coding RNA contains the structural, non-structural protein (Nsp) coding regions and the accessory protein-coding region (Durojaiye et al., 2020). The genes on the 3'-terminus encode the four structural proteins including the spike protein, membrane, envelope, nucleocapsid, and many accessory proteins. The membrane, envelope, and nucleocapsid protein protect the virus before entering the host cell. The Spike protein of SARS-CoV-2 comprises S1 and S2 subunits. The receptor-binding domain is a part of the S1 subunit that plays role in the attachment of the virus with the receptor while viral cell membrane fusion is mediated by the S2 subunit, thus facilitating the virus entry (Alanagreh et al., 2020; Jackson et al., 2021). The SARS-CoV-2 virus's replication and ability to spread are facilitated by numerous crucial proteins and enzymes. Two essential proteases, main protease (3CL^{PRO}) and papain-like protease (PLpro) are necessary for viral replication (Huang et al., 2020; Mouffouk et al., 2021). The non-structural proteins nsp1, nsp2, and nsp3 are known to be cleaved by PLpro, while the remaining 13 are cleaved by 3CL^{PRO} (Klemm et al., 2020). The 3CL^{PRO} cleaves polypeptide sequences after a glutamine residue, making it a perfect drug target as no human host-cell proteases with this cleavage specificity are identified (Hilgenfeld and Hilgenfeld, 2014; Ullrich and Nitsche, 2020).

The structure of the 3CL^{PRO} comprises three important domains, domain-I ranges from 8–101, while domains-II corresponds to position 102–184, followed by the connecting loop from 185–200, which links domain-II and domain-III, domain-III has a total number of 103 residues which lies after the connecting loop from 201–303 (Wu et al., 2020b). Furthermore, the His-41 and Cys-145 form an essential catalytic dyad (Kneller et al., 2020). Small compounds that target conserved viral proteases, such as the major protease, may thus be able to inhibit crucial phases of the SARS-CoV-2 life cycle while causing few adverse effects (Mengist et al., 2021). Approved drugs have been developed for viral infections such as those caused by Hepatitis C virus and human immunodeficiency virus for the target's serine proteases and

aspartyl protease respectively which employ that viral proteases are well-established therapeutic targets (Agbowuro et al., 2018). Antiviral drugs are required in this situation to prevent infection in high-risk populations as well as to treat infected patients. Developing inhibitors that stop coronavirus replication can recover millions of people globally. In the clinical investigations, efforts to repurpose the majority of approved drugs have discovered several promising candidates (such as remdesivir and hydroxychloroquine) but these drugs had little to no effect on mortality and the duration of hospital stay (Luttens et al., 2022). Hence, it is crucial to find new drug candidates that would target various SARS-CoV-2 proteins for increased COVID-19 therapeutic effectiveness (Elmaaty et al., 2022). Despite the significant cost and time required for the development of the new drug, clinical trials only yield a 13 percent success rate, while in 40%–60% of cases, drugs failed to reach the market because of the lack of optimum pharmacokinetic properties (Gurung et al., 2021).

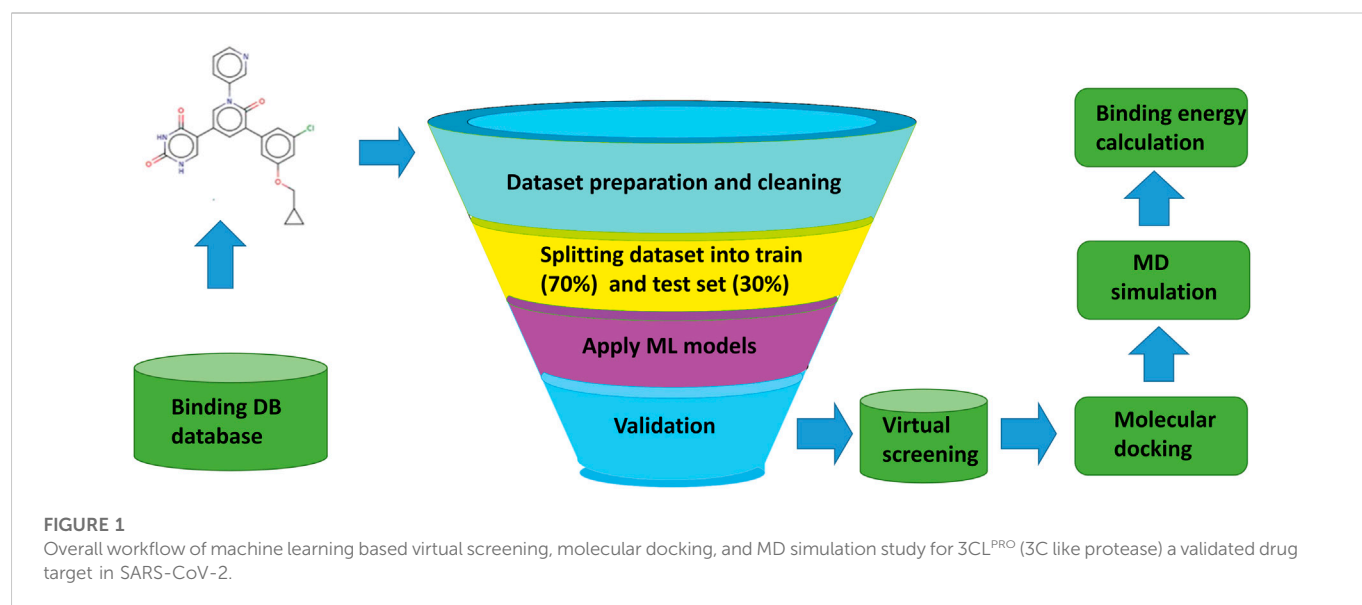
The use of computer-aided drug discovery (CADD) tools helps to accelerate the process of drug discovery and to reduce costs (Macalino et al., 2015). In addition, the advent of supercomputing facilities, algorithms, and tools has enhanced lead identification in pharmaceutical research (Macalino et al., 2018). Artificial intelligence (AI) and machine learning approaches have substantially assisted the analysis of pharmaceutical-related large data in the drug discovery process (Floresta et al., 2022). Furthermore, the structure-based drug development method is specific and successful in identifying lead compounds and optimizing them, and it has aided in the understanding of disease at the molecular level (Yang et al., 2022). In the current study, we employed different machine learning (ML) models for the virtual screening of phytochemicals against the 3CL^{PRO} drug target in SARS-CoV-2. The active hits obtained from ML-based were passed through an electronic filter called PAINS filter and their ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties were examined. The active phytochemicals that passed through the PAINS filter and have enhanced properties were further considered for the molecular docking analysis. Furthermore, the stability and binding energy of these compounds in the active site of 3CL^{PRO} were investigated by 100 ns of MD simulations. Based on our findings we suggest these phytochemicals as potent inhibitors of SARS-CoV-2 3CL^{PRO}. *In vitro* evaluation of these compounds, is essential for the understanding of their action and mechanism to cope with such a pandemic.

2 Methodology

The overall workflow of the current study, from the collection and preparation of the dataset of active and inactive compounds, screening of compounds, molecular docking, and binding energy calculations are represented in Figure 1.

2.1 Preparing and cleaning the dataset

From the binding DB database (Sandhu et al., 2022) a total of 101 molecules were retrieved for 3CL^{PRO} (3C like protease) a drug target in SARS-CoV-2. A total of 500 decoys molecules, which are considered to be inactive, were generated using the DUDE database (Mysinger et al., 2012). Out of the total 601 compounds (Supplementary Table S1), 101 compounds from the binding DB



database were labeled as “1” active, and the 500 decoys were labeled as “0” inactive. The Pandas library of python was used for data preprocessing and data cleaning (Santos et al., 2020). The dataset was split into train set (70%) and a test set (30%).

2.2 Features calculation

The 2D features of all the compounds were calculated using MOE (2016) software (Wadood et al., 2022a). Total 206 features were calculated. Feature with 0 or null values were removed from the dataset to reduce the computation time.

2.3 Principal component analysis (PCA)

The dataset was uploaded to iRaPCA v1.0 implemented in the LideB tool in CSV format. The optimum subsets of descriptors were selected from the dataset. The dimensionality was reduced by performing the PCA. The process is based on the principle of feature bagging (Prada Gori et al., 2022). The conventional feature extraction and data representation method used extensively in the fields of pattern recognition is principal component analysis (PCA), generally called as Karhunen-Loeve expansion. PCA is a method for reducing high-dimension data to low-dimension while preserving the majority of the relevant data. The main benefits of PCA are its low noise sensitivity, lower capacity and memory requirements, and increased performance (Karamizadeh et al., 2013).

2.4 Machine learning models

2.4.1 K nearest neighbor model

The distance-based classification algorithm is called k-Nearest Neighbors (kNN), which is an effective and simple machine learning algorithm widely used for the classification of active and inactive

compounds in the dataset (Wadood et al., 2022b). The accuracy of the kNN model depends entirely on the quality of the data. One of the most difficult parts of KNN is figuring out how many neighbors to consider. The KNN can be used for both classification and regression (Sarker, 2021a).

2.4.2 Support vector machine (SVM)

SVM is generally used for the classification of data. SVM is based on the principle of calculating margins between two classes. This classifier reduced the error by drawing the margins in a manner where the distance between the margin and the classes is as large, as possible (Noreen et al., 2016). The SVM classifier depends on the kernel function and is more effective for high-dimensional data classification. When the dataset contains additional noise, such as overlapping target classes, SVM does not perform effectively (Sarker, 2021b).

2.4.3 Random forest

Random forest (RF) is an ensemble algorithm made up of several decision trees, similar to how a forest is made up of many trees (Breiman, 2001). To train, the decision trees of a random forest various subsets of the training dataset are used. To classify a new sample, the sample's input vector must be passed down from each decision tree of the forest. This algorithm classifies the data using majority voting. In terms of performance, RF performs better than a decision tree. For huge datasets, it works effectively. The classifier also calculates which variables or attributes are most significant in the classification (Ul Hassan et al., 2018). The sklearn library of python was used for developing the three machine learning models.

2.4.4 Naïve bayes

The naive Bayesian algorithm is based on the Bayes theorem and is a reliable classification method. A data set can be classified by NB classifier assuming that every feature contributes equally and independently (Patel et al., 2020). In this study, the NB classifier was built using python v.3.9.

2.4.5 Cross-validation and performance evaluation

We used 10-fold cross-validation in this study. The performance of the models was accessed by using several statistical parameters including accuracy, sensitivity, specificity, F1 score, MCC (Ahmad et al., 2021).

2.5 Virtual screening of the asian phytochemicals

A list of Asian plants with notable therapeutic properties was compiled, and then a thorough literature search was performed to determine the phytochemical contents. The compound collection was carried out using Google Scholar, PubMed, MEDLINE, and other web-based resources. A total of 4,000 phytochemical libraries was generated, and the 2D structure of these phytochemicals was retrieved from the PubChem database. Before adding to the library all these phytochemicals were cleaned and energy minimized using the mmff94 force field.

2.6 PAIN filter

Pre-filtering large databases using appropriate molecular properties is a typical approach to reduce computing and get rid of unwanted compounds (Baell and Holloway, 2010). All the active hits were filtered by an online tool PAINS (Wadood et al., 2022c) and only those compounds were further selected for docking that was passed from the PAINS filters.

2.7. Molecular docking study

2.7.1 Preparation and validation of target protein

The 3D structure of SARS-CoV-2 3CL^{PRO} (PDB ID: 6LU7; Resolution: 2.16 Å; Organism: SARS-CoV-2; Method: X-ray diffraction) was downloaded from the RCSB Protein Data Bank (Hatada et al., 2020). There are two chains in the crystal structure: A and C. The macromolecule chain A was chosen as the target receptor. Pymol was used to remove water molecules and heteroatoms from the protein structure (Janson et al., 2017). The structure was then energy minimized using ff14sb implemented in the molecular operating environment (MOE) (Ashraf et al., 2021). The PROCHECK (Laskowski et al., 1996) and ERRAT (Colovos and Yeates, 1993) tools from the Structural Analysis and Verification Server (SAVES) (<http://nihserver.mbi.ucla.edu/SAVES>) were used to validate the crystal structure. The stereo chemical quality of the protein structure was evaluated using PROCHECK.

2.7.2 Molecular docking protocol

All the phytochemicals predicted as active by the machine learning method were docked into the active site of a SARS-CoV-2 3CL^{PRO} for molecular interaction studies. The crystal structure of the SARS-CoV-2 3CL^{PRO} (PDB ID: 6LU7) is complex with an N3 inhibitor was retrieved from the PDB database. The Inhibitor N3 is linked to the protease at site one of this crystal structure, which contains five cavities for ligand binding (Das et al., 2021). We used the N3 binding site (site 1) for virtual screening of these phytochemicals' library. For the molecular docking study, MOE

v2016 was used to run a docking protocol using rigid and ligand-based docking parameters. The Triangular Matching docking method (default) was used and a total of ten poses were generated for each Phytochemical (Thuy et al., 2020). The best S score hits against 3CL^{PRO} were considered for the molecular interactions study and their 3D images were generated by PyMol software. A total of 05 top-ranked compounds were shortlisted for further molecular dynamic simulations analysis based on the docking score. These phytochemicals are structurally diverse, effective, and new inhibitors for the main protease, according to the docking score, binding mode, and visual ligand interaction.

2.8 MD simulations

Molecular dynamics simulation is a powerful tool to understand the dynamics and interaction behavior of the reference complex and the selected top hits were used. The ff14SB protein force field in Amber 20 package was employed (Salomon-Ferrer et al., 2013a). For solvation of each system, the tip3p water model with box dimension 8.0 was used. All of the systems were adequately solvated and neutralized by adding four Na⁺ ions to counterbalance the charges on the systems. Afterward, energy minimization for 6,000 steps of neutralized complexes was carried out using the steepest descent minimization algorithm, then progressively heated to 300 K before equilibrating density for 2 ns with weak constraints. The whole system was equilibrated at constant pressure for another 2 ns. A Langevin thermostat was used to control the temperature 300 K. Further, a 100-ns MD was performed on the equilibrated systems. For long-range electrostatic interactions, Particle Mesh Ewald (PME) algorithm was used (Darden et al., 1998). For covalent bonds including hydrogen, the SHAKE algorithm was utilized. Finally, a 100 ns MD simulation of all equilibrated complexes at constant pressure and temperature was carried out by using PMEMD.cuda (Salomon-Ferrer et al., 2013b).

2.9 DCCM

The dynamic cross-correlation analysis is useful for explaining the correlation among the residues represented by a three-dimensional matrix. The cross-correlation was calculated by the formula (Junaid et al., 2018)

$$C_{ij} = \langle \Delta r_i \Delta r_j \rangle / (\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle)^{(1/2)} \quad (1)$$

Where the mean position of *i*th and *j*th atom is represented by Δr_i , Δr_j respectively. Where the angular brackets are used to measure the average time of the entire trajectories produced as a result of MD simulations. Positive Correlated movement such as movement in the same direction is represented by the positive value of C_{ij} , while the negative value of C_{ij} reflects strong anti-correlation movements between the residues. Cpptraj was used to perform DCCM analysis while origin 2021 was used for graphical representations (Perez-Lemus et al., 2022).

2.10 Binding affinity calculations

To study the interaction between protein and ligand, binding free energy calculations play an important role. Using MMPBSA. PY

TABLE 1 Train and test set used in the study.

Dataset	Inhibitors	Non-inhibitors	Total
Train	32	388	420
Test	33	148	181

script, the binding free energy between main protease and phytochemicals inhibitors was calculated (Gul et al., 2021). The following equation was used to calculate the free energy of each energy term:

$$\Delta G_{bind} = \Delta G_{complex} - [\Delta G_{receptor} + \Delta G_{ligand}] \quad (2)$$

In the equation, ΔG_{bind} represents the total binding free energy, $\Delta G_{complex}$ denotes the free energy of complex, $\Delta G_{receptor}$ and ΔG_{ligand} represents the free energy of receptor protein and ligand respectively. The following equation was used to calculate the individual free energy of complex, protein and ligand.

$$G_X = E_{MM} - (TS) + (G_{solvation}) \quad (3)$$

Where x denotes complex, protein or ligand, the average molecular mechanic potential in a vacuum is given by E_{MM} , the entropic and temperature contribution is represented by TS, while the free energy of the solvation is given by $G_{solvation}$.

3 Results

3.1 Data preparation

A total of 101 molecules were retrieved from the binding databank database for 3CL^{PRO} a drug target in SARS-CoV-2. The 101, molecules were categorized as active molecules. The remaining 500 decoys molecules were labeled as inactive. The dataset was split into a train set (70%) and test set (30%). Out of the total 601 molecules, the train set contains 420 compounds while the test set contains 181 compounds. The active and inactive compounds of the train and test set are present in Table 1.

3.2 Principle component analysis

Total 208 2D features were calculated with the help of MOE software. The feature with 0 values were removed. As, not every extracted feature will necessarily depict the optimal properties of molecules. Therefore, optimization was carried out to get rid of duplication. Additionally, after applying the PCA the features that have higher significance were used to train the models (Araki et al., 2016). After applying PCA the data size (N) of the dataset was decreased. To evaluate how the PCA manages to maintain the dominant properties throughout the classification tasks. The models were generated by using the entire dataset without optimum features selection and the performance of the models was evaluated. It was found that the accuracy of SVM was very low as 61% and the MCC was 0.27. The accuracy of KNN model was 70% with an MCC value of 0.58 while the accuracy of RF model was 90% with an MCC value of 0.78. However, after the optimum features selection and the reduction of

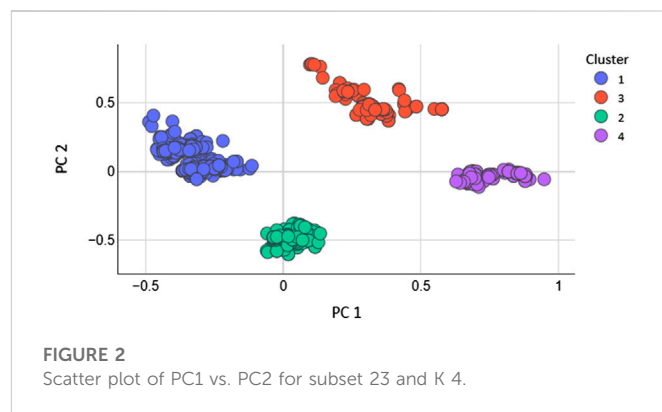


FIGURE 2
Scatter plot of PC1 vs. PC2 for subset 23 and K 4.

the dimension of the dataset the performance of all the models was greatly improved. If we want to reveal variance in a dataset having x-y coordinates, PCA finds a new coordinate system in which x, y coordinates have a different value. A new coordinate is created by the axes PC1 and PC2. These are combinations of the x-y coordinate system. Figure 2 shows the scatter plot of PC1 vs. PC2 for K = 4.

3.2.1 Chemical space and diversity analysis

The machine learning model's accuracy is predicted by the chemical diversity of the samples from the training and test sets. The applicability of machine learning models is restricted by a small number of samples. As a result, in the present study's physiochemical distribution analysis of the training set and test set for the molecular weight (MW) and LogP was conducted (Figures 3, 4) with MW ranging from 50 to 800 Da and LogP ranging from -2 to 15.

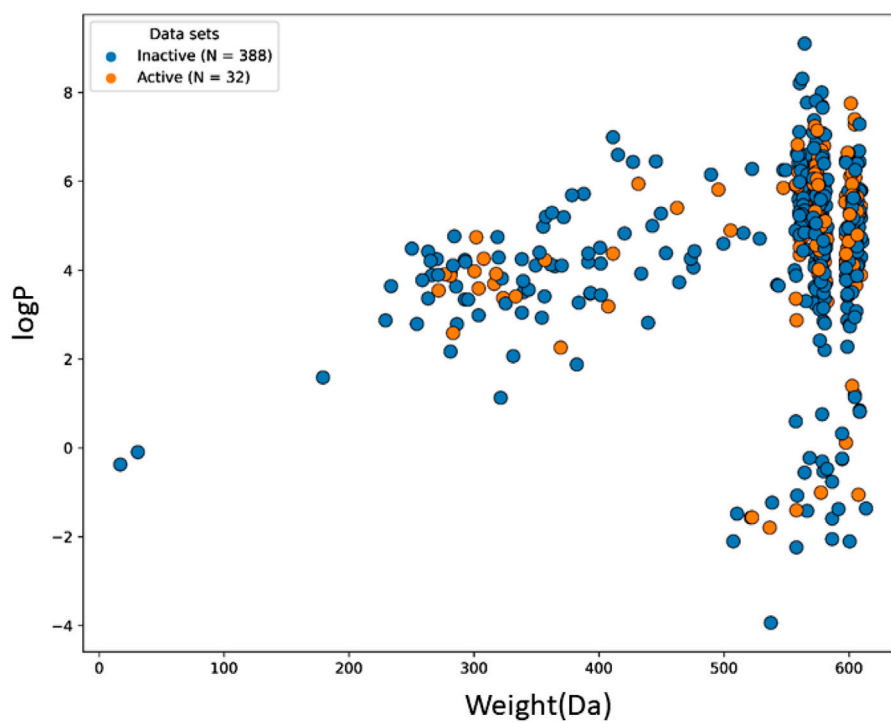
3.3 Models generation and validation

Machine learning algorithms such as kNN, SVM, RF and GNB were used for the classification of the active inhibitors against 3CL^{PRO}. The sklearn library of python was used for developing the models. All the models were trained on the dataset downloaded from the binding DB database. The performance of the models was accessed by using a number of statistical parameters including accuracy, sensitivity, specificity, and MCC. Table 2 displays the over-all performance of the models on the train set while Table 3 displays the performance of all the models on the test set.

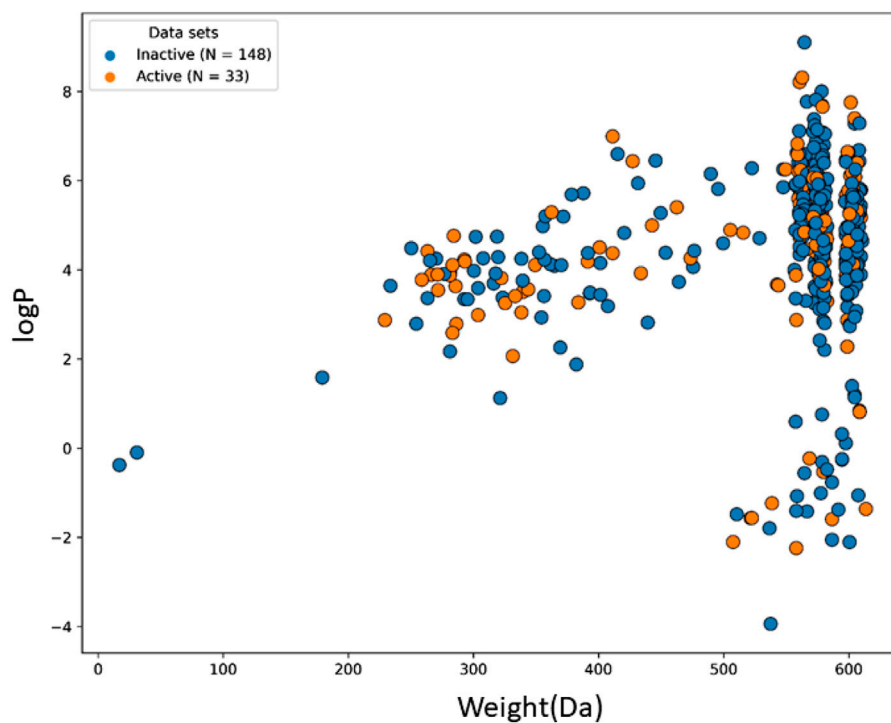
Compared to other machine learning models random forest model achieved better accuracy and MCC value. Model performance is proportional to the area under the curve (AUC). RF has the highest AUC, followed by SVM on the training and test set Figures 5, 6. Further, we used RF model to classify the active phytochemicals against the 3CL^{PRO} enzyme. Out of 4,000 phytochemicals, a total of 26 phytochemicals were predicted as active against the 3CL^{PRO}.

3.4 PAIN filter

Using the online PAINS tool all the hits were examined for their ADMET (absorption, distribution, metabolism, excretion, and toxicity) (Supplementary Table S2) properties. A total of seven compounds were passed from the PAINS filter and only two

**FIGURE 3**

The chemical space and diversity distribution of the train set. The molecular weight and LogP define the chemical space.

**FIGURE 4**

The chemical space and diversity distribution of the test set. The molecular weight and LogP define the chemical space.

TABLE 2 Overall performance of machine learning models on the train set.

Model	Accuracy (%)	Sensitivity	Specificity	MCC
KNN	97	0.88	0.99	0.91
SVM	98	0.90	0.99	0.93
RF	98	0.97	0.99	0.96
GNB	94	0.83	0.96	0.79

TABLE 3 Performance of models on the test set.

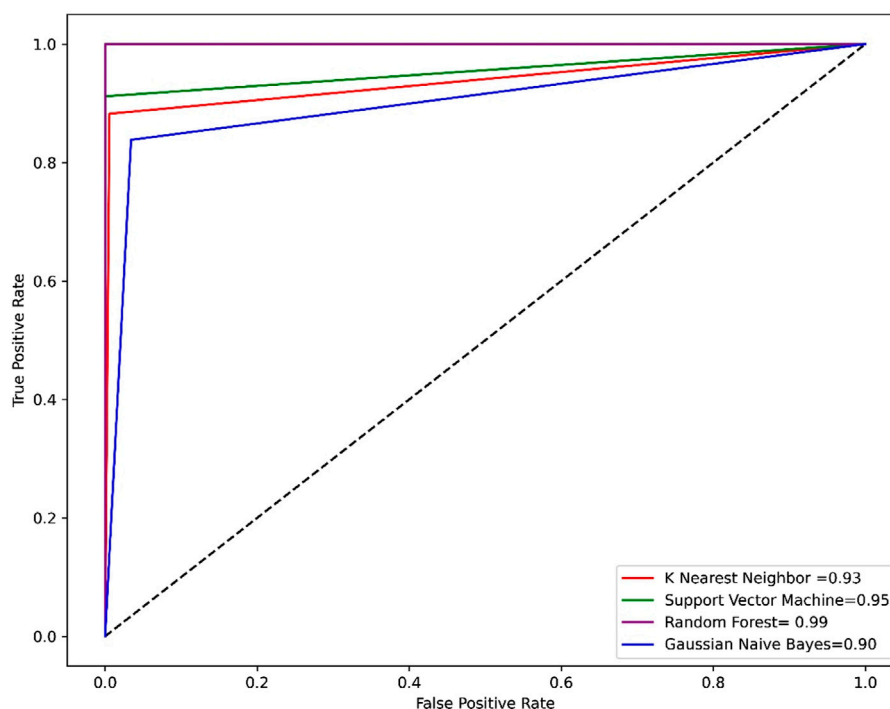
Model	Accuracy (%)	Sensitivity	Specificity	MCC
KNN	94	0.75	0.98	0.78
SVM	96	0.82	0.99	0.87
RF	98	0.95	0.99	0.95
GNB	96	0.86	0.98	0.85

compounds were out of the limit. The structure of compound along with IUPAC name of the compounds passed from the PAIN filter are given in Table 4.

3.5 Molecular docking analysis

The hits obtained from ML based virtual screening were further used for molecular docking study. The crystal structure of the SARs-CoV-2

3CL^{PRO} (PDB ID: 6LU7) is complex with an N3 inhibitor was retrieved from the PDB database. PROCHECK tool was used to assess the 3D model's quality of the 3CL^{PRO} structure using the Ramachandran plot (Figure S2a). The Ramachandran plot for the 3CL^{PRO} structure showed that 84.5% of residues were in the most favored region, while 14.3% were in the additional allowed region, 1.1% residues were in the generously allowed region and 0% residues were in the disallowed region demonstrating the high quality of the 3CL^{PRO} structure. For non-bonded atomic interactions, ERRAT is also known as the "overall quality factor," with higher scores reflecting the high quality. For a high-quality model, the accepted range is > 50 (Messaoudi et al., 2013). The ERRAT server predicted an overall quality factor of 85.90 for the 3CL^{PRO} structure used in our study (Figure S2b). The interaction of top hits and the reference compound were analyzed, and it was found that all of the compounds have potent inhibitory effects on 3CL^{PRO}. In order to study the interactions of these compounds in detail, the 3D visualization and compound interaction analysis was carried out. According to the interaction details Table 5, Compound 1 has stronger interaction among all of the docked compounds, it has 04 hydrogen bond donor interactions with the active site residues i.e., CYS145, SER46, and MET49, with four hydrogen bond acceptor interactions with HIS41, LEU141, and HIS163, along with one π -stacking interaction with residue THR25 with the docking score of -12.0321. Similarly, the interactions details of Compound 2 reveal that it shares five hydrogen bond donor interactions with key active site residues of the main protease i.e., THR26, MET49, ASN142, CYS145, and MET165, and two π -H interactions with residues with SER46 and THR90 respectively. The interaction table indicates that Compound 3 forms 6 hydrogen bond interactions with His41, Met49, Cys145, His163, and Gln189, and one π -H interaction with Leu 141. Compound 4 shows 04 hydrogen bond donor interactions with the catalytic residues i.e., Thr 25, Thr26, Met49, and His164, and one

**FIGURE 5**

The ROC-AUC curve of all the models on the train set. The graph shows the TP against FP rate.

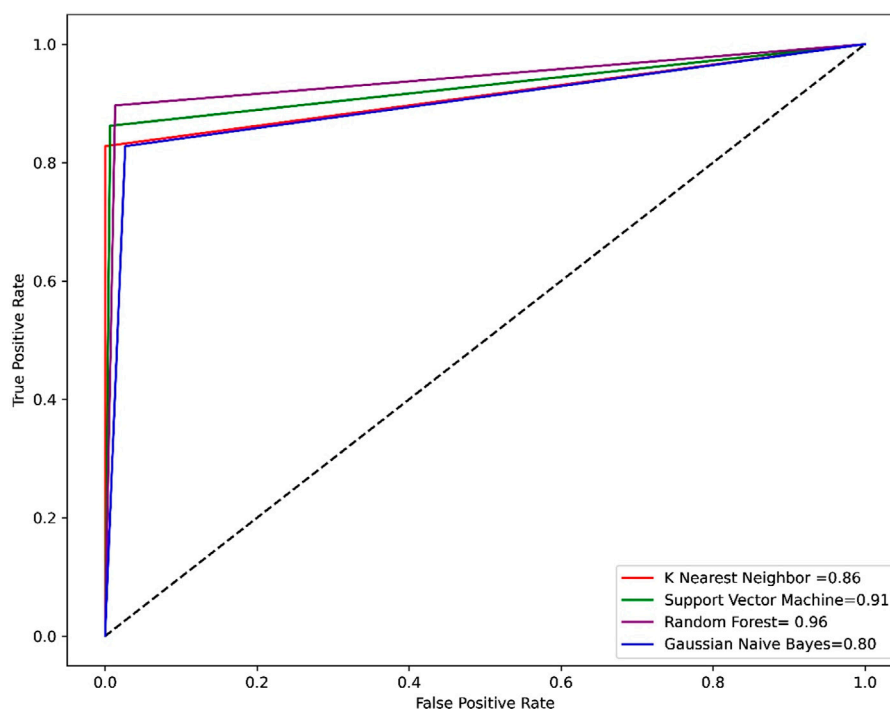


FIGURE 6

ROC-AUC curve of all the models on the test set. The graph shows the TP against FP rate.

hydrogen bond acceptor interaction with Gly143, with one π - π interaction with residue His41. Afterward, we analyzed the interaction of Compound 5, the finding of interaction analysis indicates that Compound 5 interacts *via* four hydrogen bond donor interactions with the key residues including Thr26, Met49, Asn142, and Gln189, while Thr26, and His41 were found in hydrogen bond donor interactions with Compound 5 with a docking score of -10.7164 . It has recently been demonstrated that ivermectin inhibits SARS-CoV-2 by up to 5000-fold *in vitro* with an IC₅₀ value of ~ 2 μ M (Jan et al., 2021; Kaur et al., 2021). In the docking study, ivermectin was selected as a standard reference inhibitor. The interaction details for the control compound are listed in Figure 7H. The control compound forms 05 hydrogen bonds with the key catalytic residues of main protease Asn119, Cys145, and Met165. The co-crystallized ligand (PDB ID: 6LU7) was removed from the active site and re-docked into the binding site of 3CL^{PRO} in order to evaluate the precision of MOE-Dock. The RMSD value between the top-ranked docked conformation and the co-crystallized ligand was 0.6532 (Figure S3), indicating the strong accuracy of the MOE-Dock procedure (Wadood et al., 2022c).

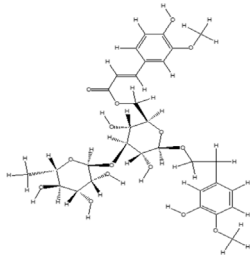
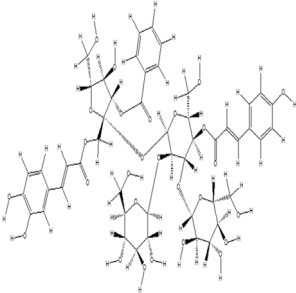
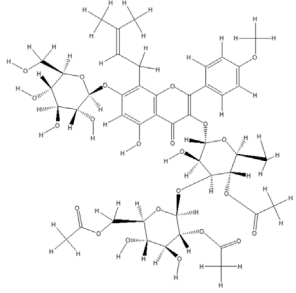
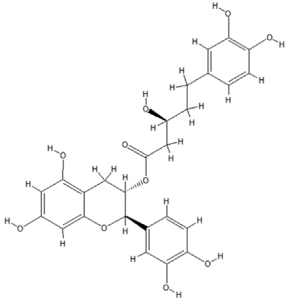
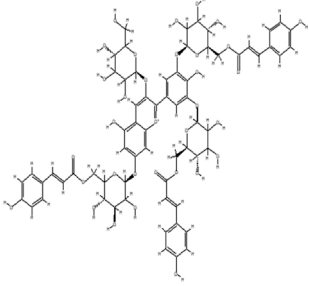
3.6 MD simulation analysis

3.6.1 Root means square deviation

Root means square deviation (RMSD) analysis was performed to calculate the stability of the top five phytochemicals and reference compound (ivermectin) in the active site of the main protease. We examined and compared the stability of these compounds with the reference and APO protein. The RMSD finding indicates that all these

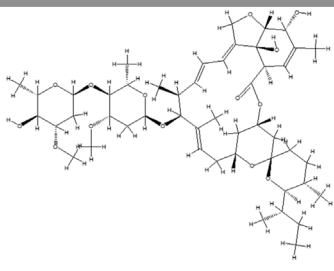
five phytochemicals show stable behavior but some minor deviation. For all the systems the averaged RMSD ranges between 1 and 3 Å. The average RMSD for ivermectin was initially 2.0 Å. Then a small increase was observed in RMSD up to 40 ns, soon after reaching 40ns the system acquired stability and remained stable for the rest of the simulation period. The complex Compound 1 shows significant stability as can be observed, however after 60 ns, the system briefly displayed a small variation. Then the system achieved stability and moved into the production phase. For Compound 2, RMSD reveals that the system shows highly stable behavior in the entire period of simulation, at 20ns minor fluctuations from its mean position were observed, afterward, the system gained stability and no more significant deviations were observed with the average RMSD value of 1.7 Å. For complex Compound 3, the system initially shows stable behavior, at around 15 ns a gradual increase in the RMSD curve was observed followed by a slight decrease in the RMSD curve at 20 ns. After that the system equilibrates with an average RMSD value of 2.1 Å as shown in Figure 8. The Compound 4 complex RMSD analysis reveals that the system initially shows an increase in the RMSD curve but soon after reaching 25 ns the system equilibrates and no significant fluctuations were observed for the rest of the simulation period which indicates the stable binding of Compound 4 compound in the catalytic pocket of 3CL^{PRO} with the average RMSD value of 1.4 Å. Afterward, we analyzed the RMSD of Compound 5 in the active site of 3CL^{PRO}, the RMSD curve of the corresponding complex has minor fluctuations at different time intervals, with an average RMSD value of 1.7 Å. The backbone RMSD for the phytochemical bound 3CL^{PRO} was slightly lower than the control indicating the stable binding of these phytochemicals which was further validated by RMSF and MM-GBSA analysis.

TABLE 4 PubChem ID of the compound, IUPAC name of compound and the PAIN filter result of the compounds.

Compound ID	Structure	IUPAC name	PAINS filter
91895373 (Compound 1)		[(2 <i>R</i> ,3 <i>R</i> ,4 <i>S</i> ,5 <i>R</i> ,6 <i>R</i>)-3,5-dihydroxy-6-[2-(3-hydroxy-4-methoxyphenyl) ethoxy]-4-[(2 <i>S</i> ,3 <i>R</i> ,4 <i>R</i> ,5 <i>R</i> ,6 <i>S</i>)-3,4,5-trihydroxy-6-methyloxan-2-yl] oxyoxan-2-yl] methyl (<i>E</i>)-3-(4-hydroxy-3-methoxyphenyl) prop-2-enoate	Passed
10606127 (Compound 2)		[(2 <i>S</i> ,3 <i>S</i> ,4 <i>R</i> ,5 <i>R</i>)-2-[[(<i>E</i>)-3-(3,4-dihydroxyphenyl) prop-2-enoyl] oxymethyl]-4-hydroxy-5-(hydroxymethyl)-2-[(2 <i>R</i> ,3 <i>R</i> ,4 <i>S</i> ,5 <i>R</i> ,6 <i>R</i>)-6-(hydroxymethyl)-5-[(<i>E</i>)-3-(4-hydroxyphenyl) prop-2-enoyl] oxy-3,4-bis[[[(2 <i>S</i> ,3 <i>R</i> ,4 <i>S</i> ,5 <i>S</i> ,6 <i>R</i>)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxy] oxan-2-yl] oxyoxolan-3-yl] benzoate	Passed
5318857 (Compound 3)		(5 <i>R</i> ,10 <i>S</i> ,13 <i>R</i> ,16 <i>R</i> ,19 <i>S</i>)-10-[(4 <i>S</i> ,5 <i>S</i>)-4-[(4 <i>S</i> ,6 <i>R</i>)-4,5-dihydroxy-6-(hydroxymethyl)-3-[(2 <i>S</i> ,3 <i>R</i> ,5 <i>S</i>)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxyoxan-2-yl] oxy-3,5-dihydroxyoxan-2-yl] oxy-16,19-dihydroxy-4,5,9,9,13,19,20-heptamethyl-21-oxahexacyclo[18.2.2.01,18.04,17.05,14.08,13] tetracos-17-en-22-one	Passed
457885 (Compound 4)		[(2 <i>R</i> ,3 <i>S</i>)-2-(3,4-dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2 <i>H</i> -chromen-3-yl] (3 <i>S</i>)-5-(3,4-dihydroxyphenyl)-3-hydroxypentanoate	Passed
44256914 (Compound 5)		[(3 <i>S</i> ,4 <i>S</i> ,6 <i>S</i>)-3,4,5-trihydroxy-6-[5-hydroxy-2-[4-hydroxy-3,5-bis[[[(2 <i>S</i> ,5 <i>S</i> ,6 <i>R</i>)-3,4,5-trihydroxy-6-[[[(<i>E</i>)-3-(4-hydroxyphenyl) prop-2-enoyl] ox methyl] oxan-2-yl] oxy] phenyl]-3-[(2 <i>S</i> ,5 <i>S</i>)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxychromenylium-7-yl] oxyoxan-2-yl] methyl (<i>E</i>)-3-(4-hydroxyphenyl) prop-2-enoate	Passed

(Continued on following page)

TABLE 4 (Continued) PubChem ID of the compound, IUPAC name of compound and the PAIN filter result of the compounds.

Compound ID	Structure	IUPAC name	PAINS filter
6321424 (Reference compound)		(1R,4S,5'S,6R,6'R,8R,10E,12S,13S,14E,16E,20R,21R,24S)-6'-[(2S)-butan-2-yl]-21,24-dihydroxy-12-[(2R,4S,5S,6S)-5-[(2S,4S,5S,6S)-5-hydroxy-4-methoxy-6-methyloxan-2-yl]oxy-4-methoxy-6-methyloxan-2-yl]oxy-5',11,13,22-tetramethylspiro[3,7,19-trioxatetracyclo[15.6.1.14,8,020,24]pentacos-10,14,16,22-tetraene-6,2'-oxane]-2-one	Passed

3.6.2 Root mean square fluctuation

The individual amino acid fluctuations of the main protease in complex with ligands were computed by RMSF analysis to assess the stability of the active site residues toward the compounds in the entire 100 ns MD trajectories (Figure 9). The RMSF of the main protease in the APO state, reference compound, and all five phytochemicals bounds to the main protease were analyzed and compared to each other, the black line in each plot represents the apo state while the red line indicated the residual flexibility of reference compound bounds to the target protein. Figure 9 shows that residues 51 and 250–260 show higher fluctuations. All these fluctuating residues were not found in the active site and these residues were far away from the active site indicating the stable binding of phytochemicals in the active site of the target protein.

3.6.3 Radius of gyration

The radius of gyration is useful for exploring the compactness and folding of the protein. Higher Rg values are indicative of less compactness (more unfolded), while lower Rg values indicate more structural rigidity and strong compactness. The MD simulation study serves to illustrate the effects of inhibitors binding upon the conformation of protein molecules. As illustrated in Figure 10 the results of Rg analysis indicate that these phytochemicals bound to 3CL^{PRO} have less radius of gyration values compared to the apo state, which demonstrates the 3CL^{PRO} stability, and compactness after ligand binding. The reference compound, Compound 1, and Compound 4 have almost similar Rg values, with an average Rg value of 22–22.3 and 22–22.4 Å while Compound 2, Compound 3, and Compound 5 showed an average gyration of 22–22.5, 22–23.3 and 22–22.4 Å respectively. The compactness of the protein was significantly affected by the binding and unbinding of these phytochemical inhibitors.

3.6.4 Dynamic cross-correlation matrix (DCCM) analysis

The extent of correlation motion between the residues imposed by the binding of compounds in the active site of 3CL^{PRO} was elucidated by the inter-residue correlation analysis. The results indicate that compound 1 in complex with the receptor active site residues showed significantly stronger parallel correlations motions in comparison with the control compound, which further validates that these positive correlation motions may be induced by the acquired interaction of these compounds with the key residues (25–50, 141–145, 163), like hydrophilic, hydrogen and hydrophobic. Overall, the DCCM findings demonstrate that the control compound and our identified compound displayed comparable patterns of highly positive correlation. Furthermore, for compound 3 and compound

5 the nearby loops regions were also found in strong positive correlations as shown in Figure 11. The dark green color demonstrates a positive correlation in residues of protein while the dense brown color indicates a negative correlation between the protein residues. The negatively correlated residues move in an anti-parallel direction while the positively correlated residues move in a parallel direction.

3.7 GBSA results

3.7.1 MM-GBSA analysis

Protein-ligand complexes from the MD simulation trajectories were used to calculate the energy parameters to assess the energetics of 3CL^{PRO} to the ligands. The binding free energies of each system were calculated using the MM-GBSA method. Table 6 display the computed average binding free energies and specific energetic contribution components of the final 500 frames. As can be observed, compound 1 has smaller free energy (–56.94 kcal/mol) followed by compound 2 (–55.65 kcal/mol), compound 3 (–53.58 kcal/mol), and compound 4 (–46.95 kcal/mol). It was observed that, as compared to the control system, all the ligands in complex with 3CL^{PRO} revealed high binding affinity demonstrating that all the systems are stable. Out of all, the binding affinity of system one was very high for the receptor. This outcome is consistent with the conclusion drawn from the earlier RMSD and docking analysis i.e., compound 1 showed stable dynamic behavior and established a greater number of non-covalent interactions (Figure 8A; Table 5).

4 Discussion

The increased mortality rate of SARS-CoV-2 has created a pandemic situation globally, no effective drugs and treatments are available to treat COVID-19, however, many clinical trials are undergoing. New infectious agents, like SARS and MERS, have emerged in the last 20 years and have created epidemics. The functional significance of 3CL^{PRO} in the viral life cycle and the lack of closely comparable human homologs make 3CL^{PRO} an attractive target for the development of antiviral medications (Jin et al., 2020). By targeting the 3CL^{PRO} most of the natural compounds play a significant role in the treatment of COVID-19 infections (Jin et al., 2020; Mengist et al., 2020). *In vitro*, animal models, and clinical trials are all used to study natural compounds that are extracted from medicinal plants, animals, and marine species for the treatment of COVID-19 (Wu et al., 2019; Wei et al., 2020; Sahoo et al., 2021). One of the most promising and

TABLE 5 Docking score and interaction of top five hits against the 3CL^{pro}.

C. No	Docking score	Ligand	Receptor	Residues	Interaction	Distance	Energy (kcal/mol)
Compound 1	−12.0321	O 4	SG	CYS 145	H-donor	4.06	−0.5
		O 8	SG	CYS 145	H-donor	4.04	−0.8
		O 14	OG	SER 46	H-donor	2.96	−0.6
		C 28	SD	MET 49	H-donor	3.89	−0.8
		O 2	NE2	HIS 41	H-acceptor	3.29	−0.7
		O 8	NE2	HIS 163	H-acceptor	3.05	−0.7
		O 9	NE2	HIS 163	H-acceptor	3.28	−1.8
		O 11	CA	LEU 141	H-acceptor	3.49	−0.6
		6-ring	CA	THR 25	π -H	4.07	−0.6
Compound 2	−11.4527	O 13	SG	CYS 145	H-donor	4.40	−0.7
		O 15	SD	MET 49	H-donor	3.84	−0.5
		O 18	O	THR 26	H-donor	2.86	−1.4
		O 21	OD1	ASN 142	H-donor	2.84	−0.6
		O 25	SD	MET 165	H-donor	3.60	−1.2
		O 12	NE2	HIS 41	H-acceptor	2.96	−0.8
		O 19	NE2	HIS 163	H-acceptor	3.07	−1.9
		6-ring	N	SER 46	π -H	4.24	−1.4
		6-ring	N	THR 90	π -H	4.33	−0.6
Compound 3	−11.2783	O 8	SD	MET 49	H-donor	3.79	−0.5
		O 22	SG	CYS 145	H-donor	3.19	−1.1
		C 26	OE1	GLN 189	H-donor	3.13	−0.9
		O 22	NE2	HIS 41	H-acceptor	3.15	−1.0
		O 23	NE2	HIS 163	H-acceptor	3.19	−1.0
		6-ring	CA	LEU 141	π -H	3.80	−0.5
Compound 4	−10.9628	O 4	O	THR 26	H-donor	2.80	−2.2
		O 6	ND1	HIS 164	H-donor	2.95	−1.8
		O 9	OG1	THR 25	H-donor	3.05	−1.6
		C 13	SD	MET 49	H-donor	3.81	−0.6
		O 5	N	GLY 143	H-acceptor	3.16	−2.7
		6-ring	5-ring	HIS 41	π - π	3.27	−0.0
Compound 5	−10.7164	O 10	OD1	ASN 142	H-donor	3.11	−1.9
		O 15	O	GLN 189	H-donor	3.07	−1.0
		O 18	O	THR 26	H-donor	3.01	−1.8
		C 57	SD	MET 49	H-donor	3.94	−0.6
		O 18	N	THR 26	H-acceptor	2.95	−0.9
		O 30	NE2	HIS 41	H-acceptor	3.10	−0.6
IVERMECTIN	−9.5398	O 5	SG	CYS 145	H-donor	3.77	−0.6
		O 6	O	ASP 187	H-donor	2.91	−0.4
		C 35	SD	MET 165	H-donor	3.81	−0.5
		C 45	SD	MET 49	H-donor	3.49	−0.2
		O 13	ND2	ASN 119	H-acceptor	3.43	−0.6

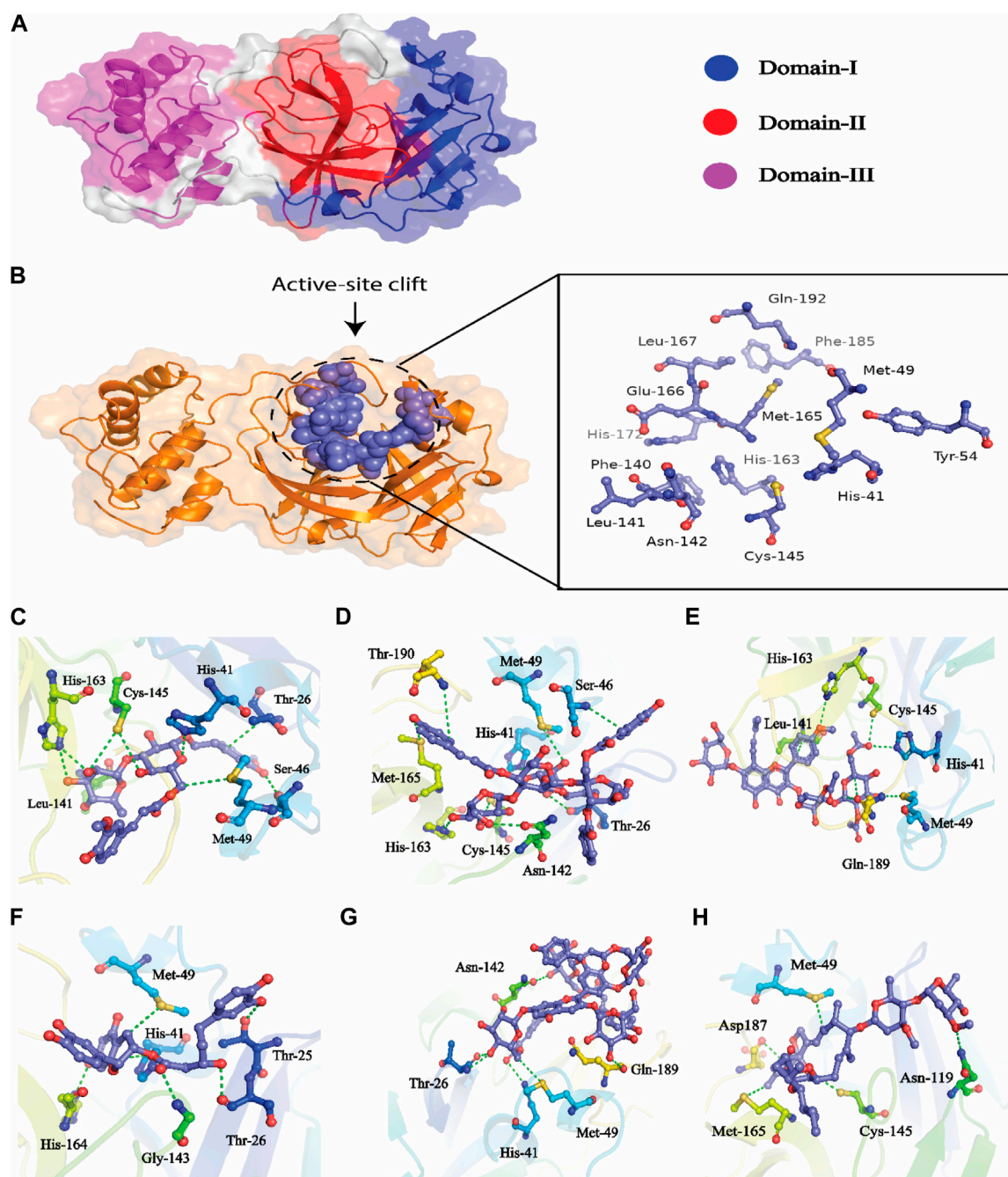


FIGURE 7

(A) All the three domains of 3CL^{PRO}, (B) active site of the main protease and (C) indicates the interaction of Compound 1 in the active site of 3CL^{PRO}, (D) represents the 3D interactions of Compound 2, (E) indicates the 3D interaction of Compound 3, (F) indicates interactions of Compound 4, (G) indicates the interaction of Compound 5, (H) indicating the three-dimensional interactions of the Control compounds (Ivermectin) with the 3CL^{PRO}.

effective strategies for combating the current pandemic is still seen to be the use of natural products (ying et al., 2001). Extractions from medicinal plants and their secondary metabolites frequently show strong antiviral properties. Some *in vitro* studies showed that PSM and viral incubation had direct interference. The viral protein, its lipid layers, and the cell's lysis can be destroyed by the plants' metabolites (Akram et al., 2018). There are about six to seven thousand different plant species in Pakistan, of which 700 are regularly used as medicines (Khan et al., 2022). The SARS CoV 2 RdRp was chosen as a receptor for computational drug development in the previous study in which 200 phytochemicals were used for virtual

screening. The top 10 ligands among 200 total ligands were chosen based on drug discovery criteria such as S-score, ligand interactions, hydrophobic interactions, and drug-likeness (Mahrosh and Mustafa, 2021).

Developing a new drug against the virus is time-consuming and costly. The ability of computer-aided drug design, on the other hand, to screen a large library of small molecules quickly and accurately may help the researcher to develop a new therapeutic agent against SARS-CoV-2 (Wang, 2020). The virtual screening workflow has made it possible to screen the enormous, diverse chemical library for

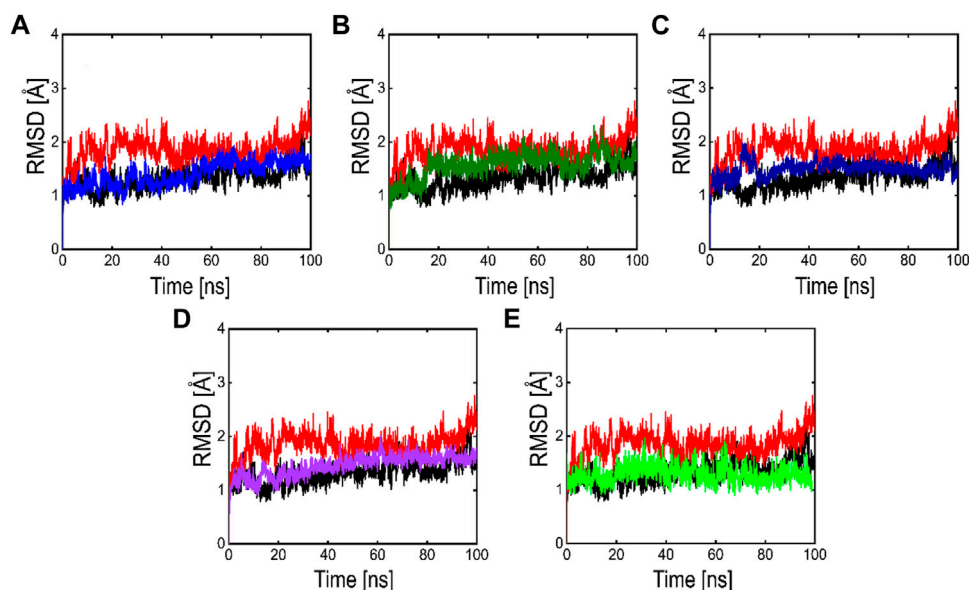


FIGURE 8

RMSD plots of the APO form (Black color), reference complex (Red color) and the top active phytochemicals (A) Compound 1 (B) Compound 2 (C) Compound 3 (D) Compound 4 and (E) Compound 5 bound to 3CL^{PRO}.

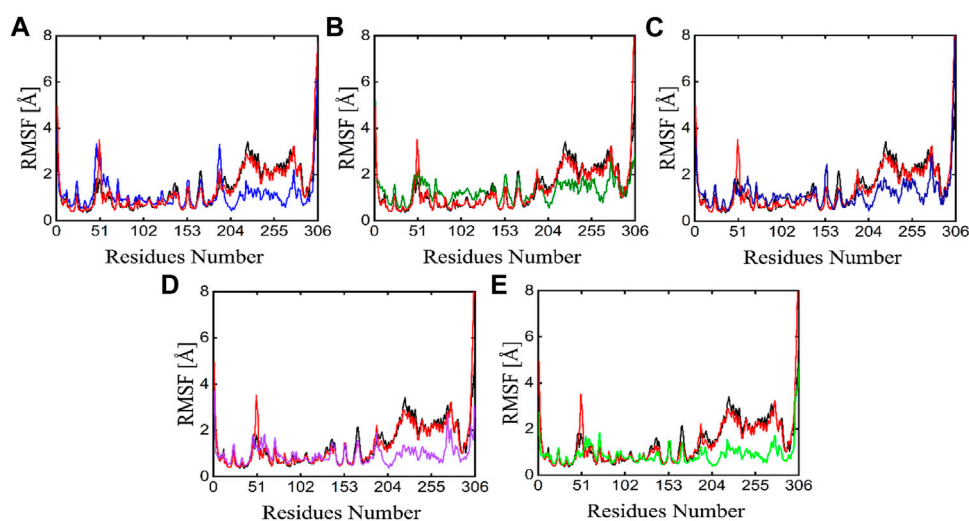


FIGURE 9

RMSF plots of the APO state (Black), control (Red) and the potent phytochemicals (A) Compound 1 (B) Compound 2 (C) Compound 3 (D) Compound 4 and (E) Compound 5.

the identification of powerful inhibitors (Neves et al., 2018). In the drug development processes, machine learning (ML) techniques are frequently used to categorize compounds as potentially active or inactive against a given protein target (Patel et al., 2020). Structure and ligand-based virtual screening frequently yield a high proportion of false positive hits (Deng et al., 2015). To reduce the false positive hits in this work, we used to machine-learning-base virtual screening for the prediction of new inhibitors against the 3CL^{PRO}. K-nearest neighbor (KNN), support vector machine (SVM), and Random Forest (RF) algorithm three of the most popular ML algorithms were chosen for virtual screening workflow. In general, classifier

performance is evaluated in terms of accuracy. KNN achieved 0.93% accuracy SVM achieved 0.96% accuracy, whereas RF produced 0.99% accuracy on the train set. Our results revealed the best performance of the RF model, so we used the RF model to classify the Asian phytochemicals. Out of 4,000 phytochemicals, a total of 26 phytochemicals were predicted as active against the 3CL^{PRO}. These active hits were further docked into the active site of the main protease. Among the 26 docked compounds, Compound 1 was found as the most potent with a docking score of −12.03 and it formed four H-donor interaction with CYS145, SER46, MET49, and four H-acceptor interactions with HIS41, HIS163, LEU141 one pi-H

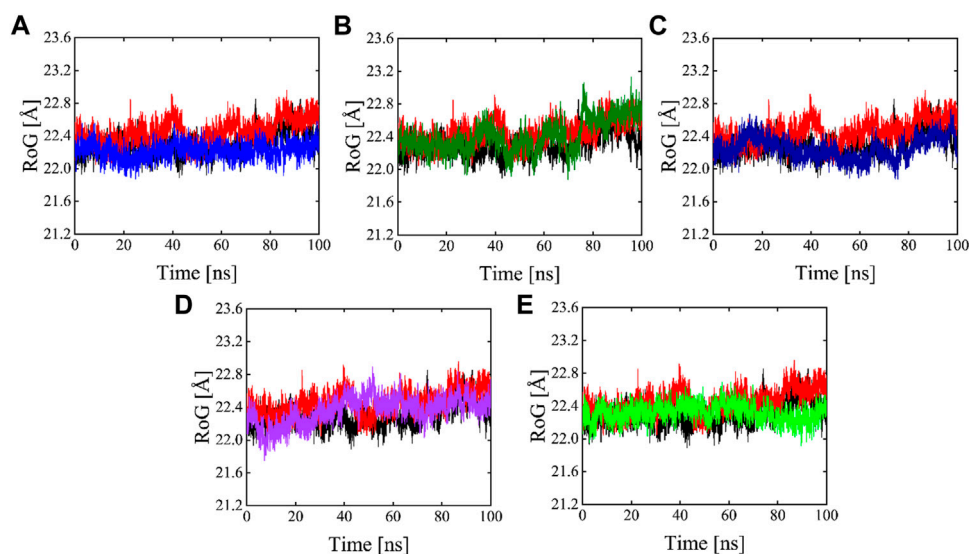


FIGURE 10

Rg plots of Apo (Black), red (reference), and Compound 1-5 are labeled different colors as (A–E) Respectively.

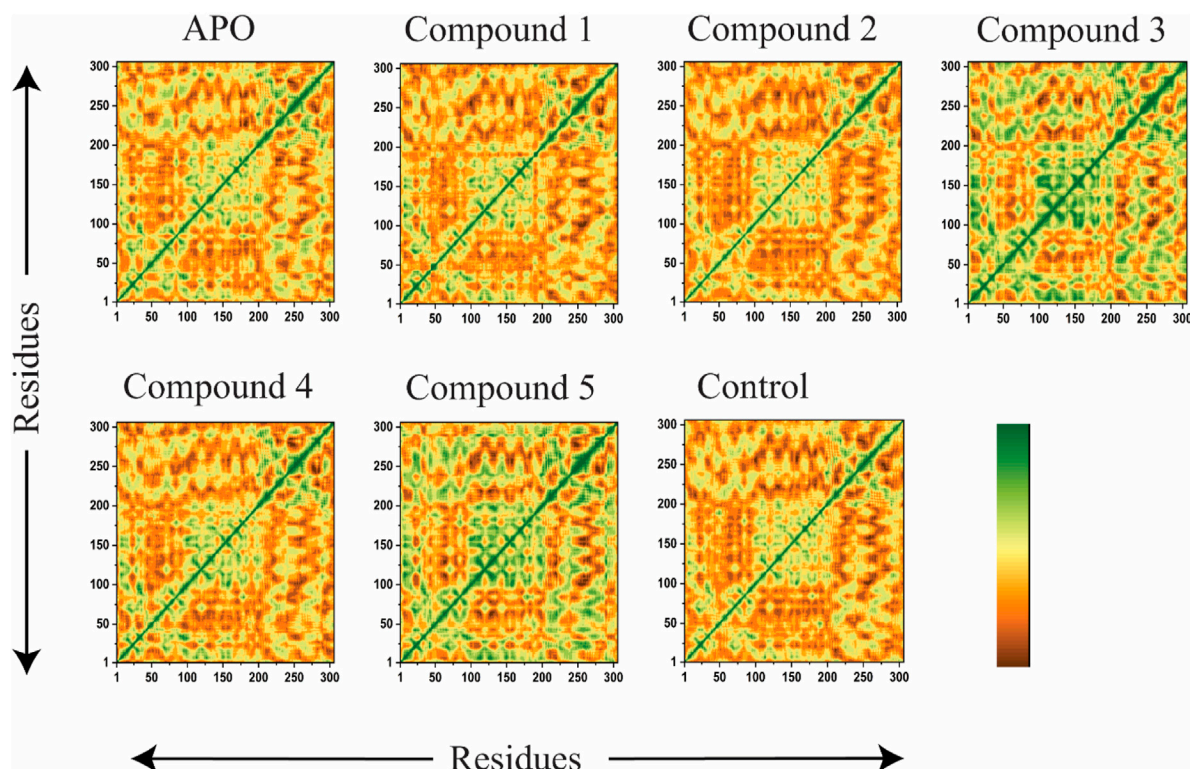


FIGURE 11

DCCM of the Apo state, Compound 1, Compound 2, Compound 3, Compound 4, Compound 5, and ivermectin (control) bound to 3CL^{PRO}. The positively correlated movement is represented by green color, while negatively correlated motion is indicated by deep brown color.

interaction with THR25 active site residues. Compound 2 was found as the second most potent hit with a docking score of -11.45 followed by Compound 3. Compound 2 formed a total of five hydrogen bonds donor interactions with the active site residues including CYS145, MET49, THR26, ASN142, MET165, and two H-acceptor

interactions with HIS41, and HIS163. The docking scores as well as interactions of Compound 3, 4 and 5 were also good as compared to the standard compound. The docking score of reference compound ivermectin was -9.53 and it formed a total of four H-donor interactions with CYS 145, MET 165 and one

TABLE 6 Represents MMGBSA Binding Free Energy (kcal/mol) calculation for the selected phytochemicals and control compound.

S. No	Compound name	VDWAALS	EEL	EGB	ESURF	-TΔS	DELTA TOTAL
1	Compound 1	-83.4745	-20.3304	56.6693	-9.8094	-18.4312	-56.9450
2	Compound 2	-79.3325	-20.6400	52.7843	-8.4635	-17.8254	-55.6517
3	Compound 3	-73.1537	-19.5693	51.8532	-8.5177	-19.2984	-53.5835
4	Compound 4	-64.4348	-16.3432	41.7462	-6.8571	-13.9835	-46.9500
5	Compound 5	-42.2227	-4.3191	13.2240	-4.7141	-10.8921	-38.0319
6	Ivermectin	-38.9027	-6.3834	20.7589	-4.3827	-14.5924	-28.9100

vdW = the van der Waals energy, EEL, electrostatic energy; ESURF, surface areas energy; EGB, the electrostatic contribution to the solvation free energy calculated by GB.

H-acceptor interaction with ASN 119 active site residue. Additionally, dynamics simulation was carried out to comprehend and support the molecular docking study. For all the systems the averaged RMSD was found between 1 and 3 Å. The averaged RMSD for ivermectin was 2.0 Å, initially, up to 40 ns the system undergoes raised up in the RMSD value up to 40 ns, and soon after reaching 40 ns the system acquired stability and remained stable for the rest of the simulation period. The complex Compound 1 shows significant stability as can be observed, however after 60 ns, the system briefly displayed a tolerable variation. The system thereafter became stable and moved into the production phase. For Compound 2, the finding of the stability index in terms of RMSD reveals that the system shows highly stable behavior in the entire period of simulation, at 20 ns minor fluctuations from its mean position were observed, afterward, the system gained stability and no more significant deviations were observed with the average RMSD value of 1.7 Å. For complex Compound 3, the system initially shows invariant behavior, up to 15 ns a gradual increase in the RMSD curve was observed followed by a slight decrease in the RMSD curve at 20 ns afterward the system attains the equilibrated with the averaged RMSD value of 2.1 Å. The protein structure's compactness as a function of time can be evaluated by the radius of gyration (Ajmal et al., 2022). The RoG analysis revealed compound 1, and compound 4 have almost similar Rg values, with an average Rg value of 22–22.3 and 22–22.4 Å while compound 2, compound 3, and compound 5 showed an average gyration of 22–22.5, 22–23.3 and 22–22.4 Å respectively. The Rg analysis of all the simulated complexes revealed that these phytochemicals formed stable and compact complexes with 3CL^{PRO}. All the short-listed phytochemicals revealed good binding affinity for 3CL^{PRO}. Compound 1 has smaller free energy (–56.94 kcal/mol) followed by compound 2 (–55.65 kcal/mol), compound 3 (–53.58 kcal/mol), and compound 4 (–46.95 kcal/mol). It was observed that, as compared to the control system, all the ligands in complex with 3CL^{PRO} revealed high binding affinity demonstrating that all the systems are stable. The RMSF analysis revealed that Domain II had a stable behavior, whereas Domain I and Domain III's amino acid residues had more flexibility in the helix and turn regions. The overall finding of RMSD and binding energy indicates that our novel phytochemicals have higher binding capacity toward the active site of the main protease. ML-based workflow combined with molecular docking and molecular dynamics approach reveals that the predicted new active phytochemicals may disrupt the SARS-CoV-2 3CL^{PRO} function.

5 Conclusion

We used *in silico* machine learning tools for drug designing against the SARS-CoV-2 3CL^{PRO}. The phytochemical dataset with more than 4,000 chemicals derived from the PubChem database was used for virtual screening against 3CL^{PRO}. Furthermore, the compound's inhibitory potential was explored using the molecular docking and MD simulation study. Using these advanced approaches, we found high-potential therapeutic compounds that can possibly inhibit SARS-CoV-2 pathogenesis. The virtual screening process, which includes MM-GBSA methods assists in reducing the list from over 4,000 possible lead compounds to 26 compounds. This research relies only on various computational tools and further it is recommended to evaluate the *in-vitro* inhibitory potential of these short-listed compounds. Successful assessment and *in vitro* evaluation of these compounds will help us to use them as a therapeutic option to treat and cope with COVID-19.

Data availability statement

Data will be provided upon reasonable request from the corresponding author of this manuscript. Requests to access these datasets should be directed to awadood@awkum.edu.pk.

Author contributions

AS, AJ, and SMJ performed experiments, analyzed data, and drafted the manuscript. AM and BK analyzed data, interpreted the results, drafted the manuscript, and revised the manuscript. AM, PL, AR, AA, MU, and PH revised the manuscript, drafted the methods, performed proofreading and improved discussion. MU and AS draw figures and tables. HJ, AM, and AW, designed, conceptualized, drafted the manuscript, analyzed and interpreted the results and revised the manuscript.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4331128DSR60).

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2023.1060076/full#supplementary-material>

References

- Agbowuro, A. A., Huston, W. M., Gamble, A. B., and Tyndall, J. D. A. (2018). Proteases and protease inhibitors in infectious diseases. *Med. Res. Rev.* 38, 1295–1331. doi:10.1002/med.21475
- Ahmad, A., Akbar, S., Khan, S., Hayat, M., Ali, F., Ahmed, A., et al. (2021). Deep-AntiFP: Prediction of antifungal peptides using distant multi-informative features incorporating with deep neural networks. *Chemom. Intelligent Laboratory Syst.* 208, 104214. doi:10.1016/j.chemolab.2020.104214
- Ajmal, A., Ali, Y., Khan, A., Wadood, A., and Rehman, A. U. (2022). Identification of novel peptide inhibitors for the KRas-G12C variant to prevent oncogenic signaling. *J. Biomol. Struct. Dyn.* 1–10. doi:10.1080/07391102.2022.2138550
- Akram, M., Tahir, I. M., Shah, S. M. A., Mahmood, Z., Altaf, A., Ahmad, K., et al. (2018). Antiviral potential of medicinal plants against HIV, HSV, influenza, hepatitis, and coxsackievirus: A systematic review. *Phytotherapy Res.* 32(5), 811–822. doi:10.1002/ptr.6024
- Alanagreh, L., Alzoughool, F., and Atoum, M. (2020). The human coronavirus disease COVID-19: Its origin, characteristics, and insights into potential drugs and its mechanisms. *Pathogens* 9, 331. doi:10.3390/pathogens9050331
- Araki, T., Ikeda, N., Shukla, D., Jain, P. K., Londhe, N. D., Shrivastava, V. K., et al. (2016). PCA-based polling strategy in machine learning framework for coronary artery disease risk assessment in intravascular ultrasound: A link between carotid and coronary grayscale plaque morphology. *Comput. Methods Programs Biomed.* 128, 137–158. doi:10.1016/j.cmpb.2016.02.004
- Ashraf, S., Ranaghan, K. E., Woods, C. J., Mulholland, A. J., and Ul-Haq, Z. (2021). Exploration of the structural requirements of Aurora Kinase B inhibitors by a combined QSAR, modelling and molecular simulation approach. *Sci. Rep.* 11, 18707. doi:10.1038/s41598-021-97368-3
- Asif, A., Ilyas, I., Abdullah, M., Sarfraz, S., Mustafa, M., and Mahmood, A. (2022). The comparison of mutational progression in SARS-CoV-2: A short updated overview. *J. Mol. Pathology* 3, 201–218. doi:10.3390/jmp3040018
- Baell, J. B., and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740. doi:10.1021/jm901137j
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 261–277. doi:10.1023/a:1017934522171
- Colovos, C., and Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Sci.* 2(9), 1511–1519. doi:10.1002/pro.5560020916
- Darden, T., York, D., and Pedersen, L. (1998). Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98 (12), 10089–10092. doi:10.1063/1.464397
- Das, P., Majumder, R., Mandal, M., and Basak, P. (2021). In-silico approach for identification of effective and stable inhibitors for COVID-19 main protease (Mpro) from flavonoid based phytochemical constituents of *Calendula officinalis*. *J. Biomol. Struct. Dyn.* 39(16), 6265–6280. doi:10.1080/07391102.2020.1796799
- Deng, N., Forli, S., He, P., Perryman, A., Wickstrom, L., Vijayan, R. S. K., et al. (2015). Distinguishing binders from false positives by free energy calculations: Fragment screening against the flap site of HIV protease. *ACS Publ.* 119(3), 5. doi:10.1021/jp506376z
- Durojaiye, A. B., Clarke, J. R. D., Stamatiades, G. A., and Wang, C. (2020). Repurposing cefuroxime for treatment of COVID-19: A scoping review of in silico studies. *J. Biomol. Struct. Dyn.* 39 (6489), 1–8. doi:10.1080/07391102.2020.1777904
- Elmaaty, A. A., Darwish, K. M., Khattab, M., Elhady, S. S., Salah, M., Hamed, M. I. A., et al. (2022). In a search for potential drug candidates for combating COVID-19: Computational study revealed salvianolic acid B as a potential therapeutic targeting 3CLpro and spike proteins. *J. Biomol. Struct. Dyn.* 40 (19), 8866–8893. doi:10.1080/07391102.2021.1918256
- Floresta, G., Zagni, C., Gentile, D., Patamia, V., and Rescifina, A. (2022). Artificial intelligence technologies for COVID-19 De Novo drug design. *Int. J. Mol. Sci.* 23, 3261. doi:10.3390/ijms23063261
- Gul, S., Ozcan, O., Asar, S., Okyar, A., Baris, I., and Kavakli, I. H. (2021). In silico identification of widely used and well-tolerated drugs as potential SARS-CoV-2 3C-like protease and viral RNA-dependent RNA polymerase inhibitors for direct use in clinical trials. *J. Biomol. Struct. Dyn.* 39(17), 6772–6791. doi:10.1080/07391102.2020.1802346
- Gurung, A. B., Ali, M. A., Lee, J., Farah, M. A., and Al-Anazi, K. M. (2021). An updated review of computer-aided drug design and its application to COVID-19. *Biomed. Res. Int.* 2021, 8853056. doi:10.1155/2021/8853056
- Hatada, R., Okuwaki, K., Mochizuki, Y., Handa, Y., Fukuzawa, K., Komeiji, Y., et al. (2020). Fragment molecular orbital based interaction analyses on COVID-19 main protease - inhibitor N3 complex (PDB ID: 6LU7). *J. Chem. Inf. Model.* 60, 3593–3602. doi:10.1021/acs.jcim.0c00283
- Hilgenfeld, R., and Hilgenfeld, C. R. (2014). From SARS to MERS: Crystallographic studies on coronavirus proteases enable antiviral drug design. *FEBS J. [Internet]* 281 (18), 4085–4096. doi:10.1111/febs.12936
- Huang, J., Song, W., Huang, H., and Sun, Q. (2020). Pharmacological therapeutics targeting RNA-dependent RNA polymerase, proteinase and spike protein: From mechanistic studies to clinical trials for COVID-19. *J. Clin. Med.* 9, 1131. doi:10.3390/jcm9041131
- Jackson, C. B., Farzan, M., Chen, B., and Choe, H. (2021). Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell. Biol.* 23, 3–20. doi:10.1038/s41580-021-00418-x
- Jan, J. T., Cheng, T. J. R., Juang, Y. P., Ma, H. H., Wu, Y. T., Yang, W. B., et al. (2021). Identification of existing pharmaceuticals and herbal medicines as inhibitors of SARS-CoV-2 infection. *Proc. Natl. Acad. Sci. U. S. A.* 118(5), e2021579118. doi:10.1073/pnas.2021579118
- Janson, G., Zhang, C., Prado, M. G., and Paiardini, A. (2017). PyMod 2.0: Improvements in protein sequence-structure analysis and homology modeling within PyMOL. *Bioinformatics* 33(3), 444–446. doi:10.1093/bioinformatics/btw638
- Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., et al. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582, 289–293. doi:10.1038/s41586-020-2223-y
- Junaid, M., Shah, M., and Khan, A. (2018). CLJ of, 2018 undefined. Structural-dynamic insights into the *H. pylori* cytotoxin-associated gene A (CagA) and its abrogation to interact with the tumor suppressor protein ASP2 using decoy. *Taylor Francis* 37(15), 4035–4050. doi:10.1080/07391102.2018.1537895
- Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., and Hooman, A. (2013). An overview of principal component analysis. *J. Signal Inf. Process.* 04 (3), 173–175. doi:10.4236/jsip.2013.43b031
- Kaur, H., Shekhar, N., Sharma, S., Sarma, P., Prakash, A., and Medhi, B. (2021). Ivermectin as a potential drug for treatment of COVID-19: An in-silico review with clinical and computational attributes. *Pharmacol. Rep.* 73(3), 736–749. doi:10.1007/s43440-020-00195-y
- Khan, M., Zaeem, A., Munir, A., Ulfat, A., and Ampak, J. B. (2022). Undefined. Plants secondary metabolites (psms), as an investigational source against Covid-19 from flora of Pakistan. *Pakbs. Org.*, 1485–1493. Available from: <https://pakbs.org/pjbot/papers/1650356086.pdf>.
- Klemm, T., Ebert, G., Calleja, D. J., Allison, C. C., Richardson, L. W., Bernardini, J. P., et al. (2020). Mechanism and inhibition of the papain-like protease, PLpro, of SARS-CoV-2. *EMBO J.* 39 (18), e106275. doi:10.15252/embj.2020106275
- Kneller, D. W., Phillips, G., Weiss, K. L., Pant, S., Zhang, Q., O'Neill, H. M., et al. (2020). Unusual zwitterionic catalytic site of SARS-CoV-2 main protease revealed by neutron crystallography. *J. Biol. Chem.* 295, 17365–17373. doi:10.1074/jbc.ac120.016154
- Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* 8(4), 477–486. doi:10.1007/BF00228148
- Lutten, A., Gullberg, H., Abdurakhmanov, E., Vo, D. D., Akaberi, D., Talibov, V. O., et al. (2022). Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors

- with broad-spectrum activity against coronaviruses. *J. Am. Chem. Soc.* 144 (7), 2905–2920. doi:10.1021/jacs.1c08402
- Macalino, S. J. Y., Gosu, V., Hong, S., and Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. *Archives Pharmacol. Res.* 38, 1686–1701. doi:10.1007/s12272-015-0640-5
- Macalino, S. J. Y., Basith, S., Clavio, N. A. B., Chang, H., Kang, S., and Choi, S. (2018). Evolution of in silico strategies for protein-protein interaction drug discovery. *Molecules* 23, 1963. doi:10.3390/molecules23081963
- Mahrosh, H. S., and Mustafa, G. (2021). An *in silico* approach to target RNA-dependent RNA polymerase of COVID-19 with naturally occurring phytochemicals. *Environ. Dev. Sustain* 23 (11), 16674–16687. doi:10.1007/s10668-021-01373-5
- Marty, A. M., and Jones, M. K. (2020). The novel Coronavirus (SARS-CoV-2) is a one health issue. *One Health* 9. doi:10.1016/j.onehlt.2020.100123
- Mengist, H. M., Fan, X., and Jin, T. (2020). Designing of improved drugs for COVID-19: Crystal structure of SARS-CoV-2 main protease Mpro. *Signal Transduct. Target. Ther.* 5, 67. doi:10.1038/s41392-020-0178-y
- Mengist, H. M., Dilnessa, T., and Jin, T. (2021). Structural basis of potential inhibitors targeting SARS-CoV-2 main protease. *Front. Chem.* 9, 7. doi:10.3389/fchem.2021.622898
- Messaoudi, A., Belguith, H., and ben Hamida, J. (2013). Homology modeling and virtual screening approaches to identify potent inhibitors of VEB-1 β -lactamase. *Theor. Biol. Med. Model.* 10(1), 1–10. doi:10.1186/1742-4682-10-22
- Mouffouk, C., Mouffouk, S., Mouffouk, S., Hambaba, L., and Haba, H. (2021). Flavonols as potential antiviral drugs targeting SARS-CoV-2 proteases (3CLpro and PLpro), spike protein, RNA-dependent RNA polymerase (RdRp) and angiotensin-converting enzyme II receptor (ACE2). *Eur. J. Pharmacol.* 891, 173759. doi:10.1016/j.ejphar.2020.173759
- 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
- Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., and Andrade, C. H. (2018). QSAR-based virtual screening: Advances and applications in drug discovery. *Front. Pharmacol.* 9, 1275. doi:10.3389/fphar.2018.01275
- Noreen, K., Azween, A., Belhaouari, S. B., Sellapan, P., Saeed, A. B., and Nilanjan, D. (2016). Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. *J. Med. Imaging Health Inf.* 6 (1), 78–87. doi:10.1166/jmhi.2016.1593
- Patel, L., Shukla, T., Huang, X., Ussery, D. W., and Wang, S. (2020). Machine learning methods in drug discovery. *Molecules* 25, 5277. doi:10.3390/molecules25225277
- Perez-Lemus, G. R., Menéndez, C. A., Alvarado, W., Bylén, F., and de Pablo, J. J. (2022). Toward wide-spectrum antivirals against coronaviruses: Molecular characterization of SARS-CoV-2 NSP13 helicase inhibitors. *Sci. Adv.* 8. doi:10.1126/sciadv.abj4526
- Prada Gori, D. N., Llanos, M. A., Bellera, C. L., Talevi, A., and Alberca, L. N. (2022). iRaPCA and SOMoC: Development and validation of web applications for new approaches for the clustering of small molecules. *J. Chem. Inf. Model.* 62 (12), 2987–2998. doi:10.1021/acs.jcim.2c00265
- Rothan, H. A., and Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J. Autoimmun.* 109, 102433. doi:10.1016/j.jaut.2020.102433
- Sahoo, A., Fuloria, S., Swain, S. S., Panda, S. K., Sekar, M., Subramaniyan, V., et al. (2021). Potential of marine terpenoids against sars-cov-2: An *in silico* drug development approach. *Biomedicine* 9(11), 1505. doi:10.3390/biomedicine9111505
- Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3(2), 198–210. doi:10.1002/wcms.1121
- Salomon-Ferrer, R., Götz, A. W., Poole, D., le Grand, S., and Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. Theory Comput.* 9, 3878–3888. doi:10.1021/ct400314y
- Sandhu, H., RajaramKumar, N., and Garg, P. (2022). Machine learning-based modeling to predict inhibitors of acetylcholinesterase. *Mol. Divers.* 26, 1–10. doi:10.1007/s11030-021-10223-5
- Santos, B. S., Silva, I., Ribeiro-Dantas, M. da C., Alves, G., Endo, P. T., and Lima, L. (2020). COVID-19: A scholarly production dataset report for research analysis. *Data Brief* 32, 106178. doi:10.1016/j.dib.2020.106178
- Sarker, I. H. (2021). CyberLearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. *Internet Things* 14, 100393. doi:10.1016/j.iot.2021.100393
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* 2(3), 1–21. doi:10.1007/s42979-021-00592-x
- Tahir ul Qamar, M., Alqahtani, S. M., Alamri, M. A., and Chen, L. L. (2020). Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants. *J. Pharm. Anal.* 10 (4), 313–319. doi:10.1016/j.jpha.2020.03.009
- Thuy, B. T. P., My, T. T. A., Hai, N. T. T., Hieu, L. T., Hoa, T. T., Thi Phuong Loan, H., et al. (2020). Investigation into SARS-CoV-2 resistance of compounds in garlic essential oil. *ACS Omega* 5(14), 8312–8320. doi:10.1021/acsomega.0c00772
- Ul Hassan, C. A., Khan, M. S., and Shah, M. A. (2018). “Comparison of machine learning algorithms in data classification” in Proceedings of the ICAC 2018 - 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing, September 2018, Newcastle Upon Tyne, UK. doi:10.23919/ICAC.2018.8748995
- Ullrich, S., and Nitsche, C. (2020). The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* 30 (17), 127377. doi:10.1016/j.bmcl.2020.127377
- Wadood, A., Ajmal, A., Junaid, M., Rehman, A. U., Uddin, R., Azam, S. S., et al. (2022). Machine learning-based virtual screening for STAT3 anticancer drug target. *Curr. Pharm. Des.* 28, 3023–3032. doi:10.2174/1381612828666220728120523
- Wadood, A., Ajmal, A., Junaid, M., Rehman, A. U., Uddin, R., Azam, S. S., et al. (2022). Machine learning-based virtual screening for STAT3 anticancer drug target-. *Curr. Pharm. Des.* 28, 3023–3032. doi:10.2174/1381612828666220728120523
- Wadood, A., Shareef, A., Ur Rehman, A., Muhammad, S., Khurshid, B., Khan, R. S., et al. (2022). Silico drug designing for ala438 deleted ribosomal protein S1 (RpsA) on the basis of the active compound Zrl15. *ACS Omega* 7(1), 397–408. doi:10.1021/acsomega.1c04764
- Wang, J. (2020). Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. *J. Chem. Inf. Model.* 60, 3277–3286. doi:10.1021/acs.jcim.0c00179
- Wei, X., Zhao, M., Zhao, C., Zhang, X., Qiu, R., Lin, Y., et al. (2020). TMR modern herbal medicine. *Glob. registry COVID-19 Clin. trials Indic. Des. traditional Chin. Med. Clin. trials* 3 (3), 140.
- Wu, H., Gao, S., and Terakawa, S. 2019 Inhibitory effects of fucoidan on NMDA receptors and I-type Ca2+ channels regulating the Ca2+ responses in rat neurons. *Pharm. Biol.* 57(1), 1–7. doi:10.1080/13880209.2018.1548626
- Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi:10.1038/s41586-020-2008-3
- 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 (5), 766–788. doi:10.1016/j.apsb.2020.02.008
- Xu, X., Chen, P., Wang, J., Feng, J., Zhou, H., Li, X., et al. (2020). Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* 63, 457–460. doi:10.1007/s11427-020-1637-5
- Yang, J., Cai, Y., Zhao, K., Xie, H., and Chen, X. (2022). Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discov. Today* 27 (11), 103356. doi:10.1016/j.drudis.2022.103356
- Ying, A. T., Qiu, H. R., Yang, Z., Zhang, D. K., Ren, L. G., Cheng, Y. Y., et al. (2001). Alkaloids from *Cynanchum komarovii* with inhibitory activity against the tobacco mosaic virus. *Phytochemistry* 58 (8), 1267–1269. doi:10.1016/s0031-9422(01)00382-x
- Zhou, P., Lou, Y. X., 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/s41586-020-1012-7
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2019). A novel coronavirus from patients with pneumonia in China. *N. Engl. J. Med.* 382 (8). doi:10.1056/nejmoa2001017



OPEN ACCESS

EDITED BY

Gyu Rie Lee,
University of Washington, United States

REVIEWED BY

Laura Orellana,
Karolinska Institutet (KI), Sweden
Pilar Cossio,
Flatiron Institute, United States

*CORRESPONDENCE

Gregory R. Bowman,
✉ grbowman@seas.upenn.edu
Henry van den Bedem,
✉ vdbedem@atomwise.com

[†]These authors have contributed equally to this work

RECEIVED 21 February 2023

ACCEPTED 07 April 2023

PUBLISHED 18 April 2023

CITATION

Meller A, De Oliveira S, Davtyan A, Abramyan T, Bowman GR and van den Bedem H (2023), Discovery of a cryptic pocket in the AI-predicted structure of PPM1D phosphatase explains the binding site and potency of its allosteric inhibitors.
Front. Mol. Biosci. 10:1171143.
doi: 10.3389/fmolb.2023.1171143

COPYRIGHT

© 2023 Meller, De Oliveira, Davtyan, Abramyan, Bowman and van den Bedem. 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.

Discovery of a cryptic pocket in the AI-predicted structure of PPM1D phosphatase explains the binding site and potency of its allosteric inhibitors

Artur Meller^{1,2†}, Saulo De Oliveira^{3†}, Aram Davtyan³, Tigran Abramyan³, Gregory R. Bowman^{4*} and Henry van den Bedem^{3,5*}

¹Department of Biochemistry and Molecular Biophysics, Washington University in St. Louis, St. Louis, MO, United States, ²Medical Scientist Training Program, Washington University in St. Louis, St. Louis, MO, United States, ³Atomwise, Inc., San Francisco, CA, United States, ⁴Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA, United States, ⁵Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, United States

Virtual screening is a widely used tool for drug discovery, but its predictive power can vary dramatically depending on how much structural data is available. In the best case, crystal structures of a ligand-bound protein can help find more potent ligands. However, virtual screens tend to be less predictive when only ligand-free crystal structures are available, and even less predictive if a homology model or other predicted structure must be used. Here, we explore the possibility that this situation can be improved by better accounting for protein dynamics, as simulations started from a single structure have a reasonable chance of sampling nearby structures that are more compatible with ligand binding. As a specific example, we consider the cancer drug target PPM1D/Wip1 phosphatase, a protein that lacks crystal structures. High-throughput screens have led to the discovery of several allosteric inhibitors of PPM1D, but their binding mode remains unknown. To enable further drug discovery efforts, we assessed the predictive power of an AlphaFold-predicted structure of PPM1D and a Markov state model (MSM) built from molecular dynamics simulations initiated from that structure. Our simulations reveal a cryptic pocket at the interface between two important structural elements, the flap and hinge regions. Using deep learning to predict the pose quality of each docked compound for the active site and cryptic pocket suggests that the inhibitors strongly prefer binding to the cryptic pocket, consistent with their allosteric effect. The predicted affinities for the dynamically uncovered cryptic pocket also recapitulate the relative potencies of the compounds ($\tau_b = 0.70$) better than the predicted affinities for the static AlphaFold-predicted structure ($\tau_b = 0.42$). Taken together, these results suggest that targeting the cryptic pocket is a good strategy for drugging PPM1D and, more generally, that conformations selected from simulation can improve virtual screening when limited structural data is available.

KEYWORDS

allosteric inhibition, cryptic site, molecular dynamics simulation, markov state models, deep learning, virtual high throughput screening (vHTS)

Introduction

Virtual screening is a common tool for identifying novel inhibitors of proteins with known structures (Wallach et al., 2015; Lyu et al., 2019; Bender et al., 2021). Conventional, structure-based virtual high throughput screening approaches use an empirical- or force-field-based scoring function to dock ligands to mostly rigid receptors and rank compounds (Trott and Olson, 2010). Docking to structures that deviate from the ligand-bound state can result in inaccurate predictions of the bound complex and poor compound ranking. For example, it is often difficult to recover active compounds when docking against ligand-free experimental structures (e.g., an *apo* state), or when the cognate ligand is small (Abagyan et al., 2010). Even worse, experimentally derived structures are unavailable for many targets with disordered or flexible domains. AlphaFold (AF) has the potential to accelerate drug discovery thanks to accurate structure prediction for such proteins (Jumper et al., 2021). However, these are still just rigid structures, and their utility will be limited if they do not represent bound-like structures (Vijayan et al., 2015; Wankowicz et al., 2022).

Phosphatases are a protein family with many potential therapeutic targets, but few are currently drugged (Mullard, 2018; Köhn, 2020) owing to a highly conserved and charged active site. Phosphatases are distinguished by different functional domains that can be exploited for the design of selective therapeutics (e.g., SH2 domain in SHP2 (Chen et al., 2016)). Often, these domains are highly flexible (Miller et al., 2022). Human protein phosphatase, Mg²⁺/Mn²⁺ dependent 1D PPM1D, also known as Wip1, is an important therapeutic target in oncology (Pecháková et al., 2017). PPM1D negatively regulates p53 and other components of the DNA damage response pathway (Lu et al., 2008). Overactivation of

PPM1D, either through duplication or loss of its degradation domain, is present in several human cancers, including breast cancer (Li et al., 2002), ovarian clear cell carcinoma (Tan et al., 2009), and brain cancers (Castellino et al., 2008).

Several allosteric inhibitors of PPM1D have been discovered through experimental screens (Gilmartin et al., 2014), but they remain difficult to improve upon because PPM1D has defied structure determination. A dual biophysical and biochemical screen targeting PPM1D revealed a novel class of inhibitors called the capped amino acids (CAA) (Gilmartin et al., 2014). These compounds selectively and non-competitively inhibit the phosphatase activity of PPM1D towards FDP and natural substrates. Efforts to crystallize PPM1D alone or PPM1D in complex with these inhibitors were repeatedly unsuccessful, likely due to a highly disordered loop or a flexible flap domain.

In the absence of this structural information, two distinct binding modes have been proposed based on indirect evidence. Photoaffinity labeling experiments suggested that the allosteric compounds bind at the PPM1D flap domain, in the vicinity of P219 and M236 (Figure 1). (Gilmartin et al., 2014) In support of this model, the authors demonstrated that swapping the flap domain of PPM1D into another phosphatase rendered that protein sensitive to the PPM1D inhibitors. However, this finding was later disputed by several experiments that implicated the hinge domain in the binding of the allosteric compounds (Miller et al., 2022). Deletion of the flap domain did not have an impact on the thermal shift, binding affinity, or the deuterium exchange profile caused by one of the allosteric compounds. Conversely, deletion of the hinge contributed to a substantial decrease in binding affinity and inhibition (i.e., an increase in IC₅₀). Thus, the lack of experimental structures as well as competing binding modes makes PPM1D a uniquely challenging target for computational drug design.

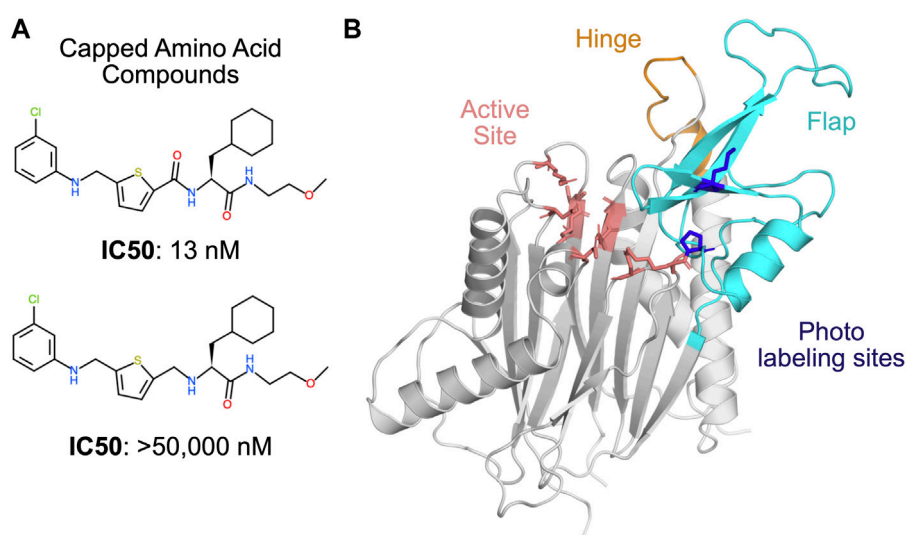


FIGURE 1

PPM1D phosphatase is allosterically inhibited by the capped amino acid (CAA) compounds, but the precise binding site is unknown. **(A)** The capped amino acid compounds have a common amino acid-like substructure, and small differences in their chemical structure (i.e., the absence of a carbonyl) can contribute to very large differences in their potency. **(B)** The AlphaFold-predicted structure of PPM1D highlights key regions that have been implicated in the binding of the capped amino acid compounds. The active site is shown in salmon sticks while two residues identified as proximal to the binding site based on photolabeling experiments are shown in blue sticks. The flap domain, a region hypothesized to be the primary CAA compound binding site, is shown in cyan. Another region hypothesized to be the primary CAA compound binding site, the hinge, is shown in orange.

Here, we use AlphaFold, molecular dynamics simulations (Karplus and McCammon, 2002; Hollingsworth and Dror, 2018), and machine learning to generate distinct conformations of PPM1D to investigate the molecular mechanisms of allosteric inhibition.

Results

PPM1D's AlphaFold structure lacks high scoring pockets at the flap and the hinge

Given the lack of available PPM1D experimental structures, we first tested if a structure predicted by AlphaFold (AF) could help determine the preferred binding site for its allosteric inhibitors. The high accuracy of AF predictions (Jumper et al., 2021) suggests that structures predicted by AF can be used for determining binding sites and conducting virtual high throughput screening campaigns. Therefore, we analyzed the PPM1D AF structure to determine if there were binding sites with a high probability of ligand binding.

The PPM1D AlphaFold structure lacks clear pockets at the flap and the hinge, which are the two binding sites proposed in the literature. In contrast to previous homology models constructed for PPM1D, the AF structure of PPM1D includes a structured flap domain. The predicted local distance difference test (pLDDT) score, a useful proxy for how ordered a region is (Wilson et al., 2022), is high in the flap domain (Supplementary Figure S1). Despite the structured nature of the flap domain, there are few obvious pockets for an allosteric inhibitor to bind. Using the P2rank algorithm (Krivák and Hoksza, 2018), we evaluated pockets on the protein surface and found two pockets with high scores (Supplementary Figure S2). One is at the active site, which cannot be the preferred binding mode for the capped amino acid compounds given the non-competitive nature of PPM1D inhibition. The second high scoring pocket is found opposite the flap domain where helix 323–326 and helix 347–360 interface with one of the β -strands in the PPM1D β -sandwich (Supplementary Figure S1). This pocket has no overlap with either of the proposed binding sites found in the literature for the PPM1D allosteric compounds. Both the flap and the hinge lack high scoring pockets in their vicinity. Similarly, when we searched for pockets using the LIGSITE algorithm (Hendlich et al., 1997), we do not find pockets at either of the proposed binding sites (Supplementary Figure S3). These findings suggest that the binding site of the allosteric inhibitors is possibly cryptic or transient, or simply not captured by the AlphaFold structure—thus posing a challenge for a successful docking campaign. Hence, we decided to investigate whether molecular dynamics simulations might reveal cryptic pockets at the flap or the hinge.

PPM1D *apo* simulations reveal a cryptic pocket at the flap-hinge interface

Next, inspired by recent success in capturing cryptic pocket formation in molecular dynamics simulations, (Hollingsworth et al., 2019; Sztain et al., 2021; Zimmerman et al., 2021; Cruz et al., 2022; Meller et al., 2023b; Meller et al., 2023c), we tested whether simulations launched from the AF structure could reveal cryptic

pockets that encompass the flap or the hinge. We used an adaptive sampling algorithm FAST (Zimmerman and Bowman, 2015) to search for cryptic pockets. FAST balances exploration with exploitation to efficiently search conformational space for conformations with desired traits. FAST does this by launching swarms of simulations and then selecting the most promising states as evaluated by an objective function for further simulations. In our case, we defined an objective function that included LIGSITE pocket volume to favor states with large pockets and another term to reward conformations which had been rarely observed (see Methods). Following each round of simulations, we created Markov State Models (MSMs) (Pande et al., 2010; Bowman et al., 2015) of the protein's conformational ensemble after clustering conformations using C- α RMSD as a distance metric.

In our simulations, the flap domain is extremely dynamic, sampling closed and highly open conformations (Figure 2A). An MSM-weighted distribution of flap domain to active site distances reveals two modes, one centered roughly on the distance found in the AF starting structure (~ 23 Å) and another around 27 Å (Figure 2A). In the closed conformations with a small active site-flap distance, the flap domain approaches a helix (residues 346–361) whose minimum distance to the flap domain in the AF structure is 11 Å (structure I in Figure 2B; Supplementary Figure S4). This behavior is consistent with experiments which showed that flap deletion leads to an increase in deuterium incorporation, implying an increase in backbone solvent exposure, at peptides spanning residues 328–362. (Miller et al., 2022). Not only can the flap close in on the active site, it can also dissociate dramatically as seen in the long tail on the right of the active site-flap distance distribution (structure iii in Figure 2B). In this extended conformation, K218 and other residues involved in substrate recognition are far from the active site (i.e., the distance between K218s sidechain to D105s sidechain grows from 9 Å in the AF structure to as much as 29 Å in simulations). The two peaks seen in the flap domain to active site distance distribution are consistent with both hydrogen deuterium exchange mass spectrometry and sedimentation velocity ultracentrifugation experiments (Miller et al., 2022), which showed that PPM1D exists in an equilibrium between two different flap domain conformations.

The highly dynamic nature of the flap domain is not captured in the AlphaFold predictions. As predicted by the high pLDDT estimates for the flap domain, the β -strands in the flap remain structured as β -strands throughout the simulations (Supplementary Figure S5). However, neither AF's pLDDT nor the predicted aligned error for the flap domain suggest that flap domain dissociation is possible or likely. We speculate that AF underestimates flap domain flexibility because it is trained with static structures from the Protein Databank (PDB), and thus simulations are a useful means to identify functionally important excited states.

Our simulations revealed a cryptic pocket at the flap-hinge interface between the two proposed binding sites. We calculated pockets for each structure in the MSM using P2Rank (see Methods). We then found the difference in each residue's maximum ligand-binding probability in the ensemble and its ligand-binding probability in the AlphaFold structure. This analysis revealed that the flap domain, especially a flap domain loop (residues 276–290), is enriched for residues with large increases in ligand-binding probability (Supplementary Figure S6, S7). To visualize this flap

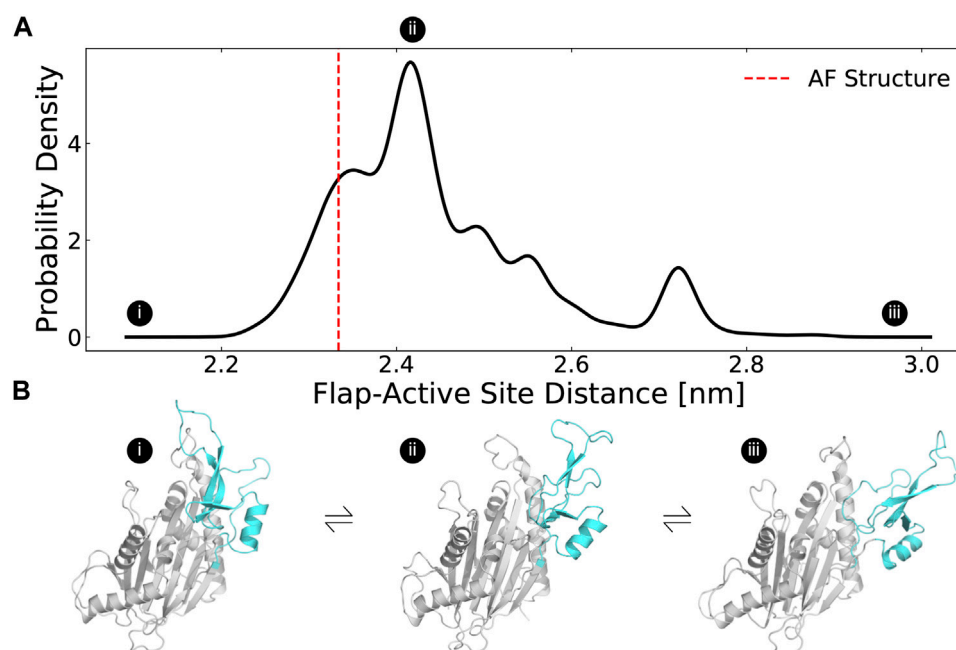


FIGURE 2

The distribution of flap domain to active site distances from MD simulations highlights that the flap is a highly flexible domain that can adopt more open conformations than seen in the AlphaFold-predicted structure. **(A)** The MSM-weighted distribution of average distances between the flap domain (defined as residue 219–295) and the active site (residues 105, 192, 314, and 366) backbones shows two peaks as well as long tails that highlight low probability highly closed and highly open conformations. The dashed red line indicates the same distance measured for the AlphaFold-predicted structure. **(B)** These structures depict a highly closed, an intermediate, and a highly open MSM cluster center. The flap domain is colored in cyan. Circles with Roman numerals indicate where these structures fall in the distribution.

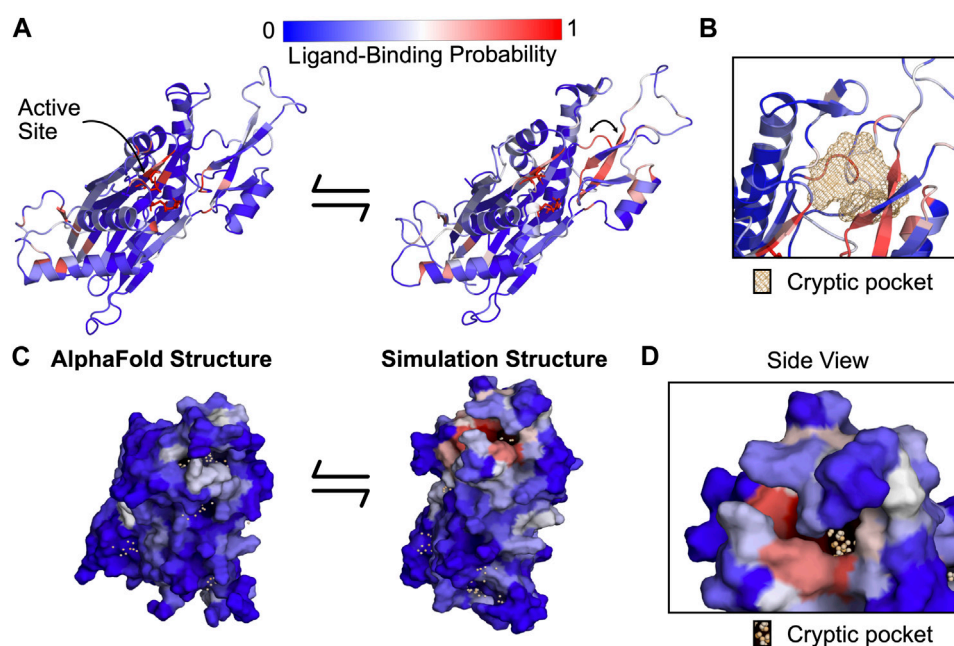
domain cryptic pocket, we found the simulation structure with the largest increase in predicted ligand-binding probability relative to the AlphaFold structure. This structure shows conformational changes in the orientation of the central β -strand in the flap as well as the loop spanning residues 269–295 (Figure 3A). Collectively, these lead to the formation of a deep pocket (Figures 3B,D) with a P2Rank-predicted ligand-binding probability of 0.87. There are other regions of the protein with increases in predicted ligand-binding probability, including the hinge (Supplementary Figure S8) and the photoaffinity labeling sites (Supplementary Figure S8), but these increases are not as substantial as those in the flap domain loop. Taken together, these results suggested that relevant binding modes for the PPM1D allosteric compounds may be hidden in the ground state AlphaFold structure.

The AtomNet PoseRanker neural network predicts a single preferred cryptic binding site between the flap and hinge

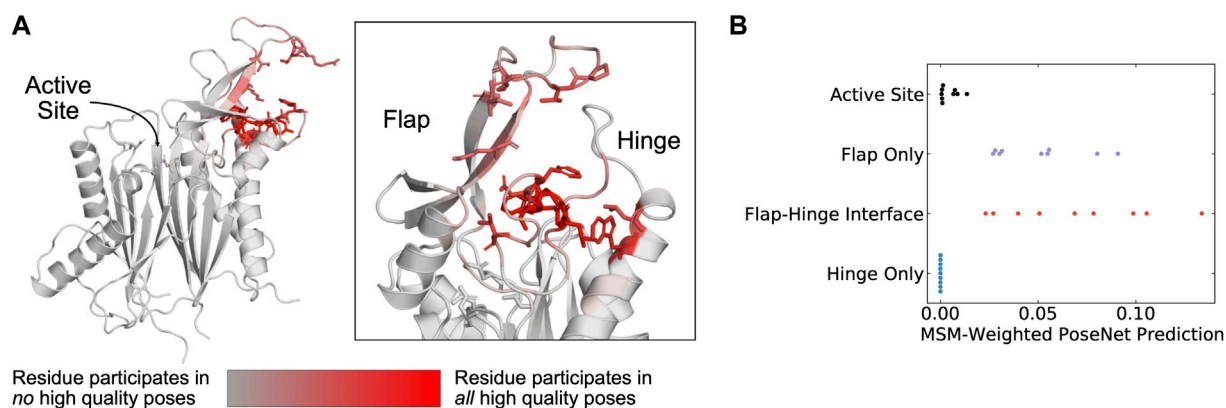
To help determine which cryptic site was the most likely binding site, we docked the PPM1D allosteric compounds across the ensemble of structures in our MSMs. Traditional rigid body docking can often produce high quality poses (root mean square deviation from a crystal pose less than 2 Å), but these methods struggle to rank the poses correctly (Su et al., 2019); the highest quality poses rarely correspond to the highest scoring poses. To circumvent this limitation, deep learning methods often re-rank

conventional docking poses and achieve improved performance. We used one of these methods, AtomNet PoseRanker (ANPR), to re-rank the poses from molecular docking (Stafford et al., 2022). ANPR was trained on existing data on the PDB and demonstrated to have an implicit understanding of physical interactions and protein dynamics. ANPR is trained as a binary classifier, and outputs a probability score between 0 and 1 (scores greater than 0.5 are usually indicative that ANPR has confidence that the pose in question is of high quality). We hypothesized that correctly assigned binding sites for ligands would admit better poses than incorrect sites. We therefore used ANPR scores to evaluate and identify the most likely binding site of the PPM1D allosteric inhibitors. We expected the most likely binding site to have higher ANPR scores across the simulated conformations with a relevant cryptic pocket.

We docked compounds to all states from the PPM1D MSMs using CUina (Gniewek et al., 2021; Stafford et al., 2022), a GPU-efficient implementation of smina (Koes et al., 2013), and evaluated the quality of the resulting docked poses with ANPR. For every state from the MSM, we used P2Rank to identify possible binding sites in that state's representative structure. A significant number of conformations presented a cryptic pocket between the hinge and the flap. A smaller number of conformations presented a pocket almost exclusively at the hinge. We used the pockets identified by P2Rank to design a box centered around these pockets. We padded the box by 5 Å on each dimension, and we used that box to define the search space of our molecular docking runs. As a control, two additional bounding boxes were created for the active site and photolabeling site described in the Gilmartin

**FIGURE 3**

PPM1D apo simulations reveal a cryptic pocket at the flap-hinge interface. **(A)** The AlphaFold-predicted PPM1D structure and a simulation structure where each residue is colored by its P2Rank-predicted ligand-binding probabilities show an increase in ligand-binding probability at the flap domain near the hinge. This simulation structure was selected because it had the largest increases in ligand-binding probability relative to the starting structure across the ensemble of states. Active site residues are shown in sticks. Arrow indicates the backbone motion that is required to form the cryptic pocket. **(B)** Mesh representation of the cryptic pocket shows that it forms between a flap domain loop (residues 276–279), two of the β -strands in the flap (residues 243–247 and 268–271), and a flap domain helix (residues 227–234). **(C)** Surface representation looking onto the AlphaFold structure and the open simulation structure highlights that a deep trench forms between the flap domain and hinge. The surface is colored by P2Rank-predicted ligand-binding probability. **(D)** A zoom-in of the surface representation of the open state reveals that the cryptic pocket lies in a deep groove. The orange spheres are the pocket grid points identified by P2Rank.

**FIGURE 4**

The AtomNet PoseRanker neural network predicts that poses found at the flap-hinge interface are more crystal-like. **(A)** A PPM1D AlphaFold structure colored by the frequency with which residues participate in high-quality poses indicates that residues needed for high-quality poses are found at the flap-hinge interface. Residues in dark red most frequently contact the GSK2830371 compound in its high-quality poses. High-quality poses were those poses that received a PoseRanker score of 0.5 or higher. A contact was defined when a ligand heavy atom was within 4 Å of a protein heavy atom. **(B)** The MSM-weighted AtomNet PoseRanker (ANPR) predictions across different binding sites show that the flap-hinge interface receives higher ANPR scores. Each point represents a different CAA compound. When there were multiple poses in one of the binding site categories, we selected the pose with the highest ANPR score. We defined the hinge as residues 150–166 and the flap as residues 219–295.

publication by defining the boundaries based on the catalytic residues or the photo labeling residues respectively. These boxes were also padded by 5 Å in each dimension (see Methods). In total,

we docked nine capped amino acid compounds against four possible sites (two proposed sites around the hinge, the photolabeling site, and the active site as a negative control). These compounds were

docked against all MSM states where the relevant cryptic pocket was detected by P2Rank. For each compound + binding site pair, we re-ranked the top 64 poses (as ranked by the vina scoring function) using ANPR. The pose for each compound and binding site with the highest ANPR ranking was selected for subsequent analyses. Interestingly, none of the poses where PPM1D allosteric compounds were docked to the AF structure scored above 0.5, indicating that these were unfavorable poses (Supplementary Table S1). This corroborates our pocket assessment results, suggesting that the static AF structure is not amenable to docking of the PPM1D allosteric inhibitors.

Across the PPM1D MSM ensemble, we found that ANPR assigns the highest scores to poses where the compounds bind between the flap and hinge. For each compound, we assessed which poses were given a ANPR probability score greater than 0.5. We defined those as predicted high-quality poses. We found that residues found at the interface of the hinge and flap domain are most likely to make contacts with high-quality poses (Figure 4A). Specifically, residues in the flap domain loop from D277 to V289 are most likely to form contacts with these poses. When we overlaid all high-quality poses of the compounds onto the AF starting structure, we found that they cluster in a single region between the flap and hinge (Supplementary Figure S10). Next, we classified poses by the protein contacts that they form into the following categories: flap domain only, hinge only, flap-domain interface, and active site (see “Pose classification” in Methods). There are no high-quality poses that form contacts only with the hinge and rarely did any high-quality poses form contacts with the active site. This is true across all compounds. Considering that the PPM1D allosteric inhibitors are non-competitive, our negative control results (docking against the active site) bolster our confidence that the ANPR probability scores can distinguish between correct and incorrect sites. We used the equilibrium probabilities from the MSM to calculate a weighted average of the ANPR score across the PPM1D ensemble (Supplementary Figure S12). We find that the ensemble-weighted ANPR probability is highest at the flap domain and flap-hinge interface (Figure 4B; Supplementary Figure S12). Thus, these ANPR predictions strongly suggest that PPM1D allosteric compounds bind between the flap and hinge.

Combining MSM-docking with pKi predictions from a neural network accurately ranks compounds

While an estimate of pose quality might be helpful in virtual screening, the decision to select compounds for synthesis and testing with *in vitro* assays relies on an estimate of a compound's bioactivity or affinity. The deep learning-based pKi predictor AtomNet has been shown to be physics-aware and to be sensitive to pose perturbations. (Wallach et al., 2015; Gniewek et al., 2021). Considering that the CAA compounds have known affinities, we can assess whether MSM-docking (Meller et al., 2023b) can have an impact on the retrospective performance of the AtomNet pKi predictor.

We applied the AtomNet pKi predictor to each of the docked poses in our MSM ensemble. The AtomNet pKi predictor was trained using a combination of public and proprietary structural

data. It outputs a value for the predicted pKi of a compound for a particular target given a particular pose provided as input. We docked each compound to several sites for each structure in the ensemble. We used the ANPR score to select the highest scoring pose per compound-structure pair in the ensemble (Figure 5A). We then passed that compound-state pair as input to the AtomNet pKi predictor, resulting in one prediction of the compound's potency per MSM state.

We find that taking an ensemble perspective that accounts for cryptic pockets outperforms results for the static AF structure. We first established a baseline by evaluating how well docking scores rank PPM1D allosteric compounds by potency. Docking scores for the AF structure alone and MSM-weighted docking scores for the ensemble (see Methods) generated very poor predictions of compound potency, demonstrating that ranking these compounds is a non-trivial task. In fact, compounds with better docking scores were less potent in general (Kendall $\tau_b = -0.59$, Figure 5B); we noticed negative correlation between docking scores and their measured potency. On the other hand, the AtomNet pKi predictor ranks more potent compounds higher using docked poses against the AF structure alone ($\tau_b = 0.42$, Figure 5B). The ability to rank compounds based on their predicted affinity further improves when we dock to all MSM states and weight the pKi predictions based on the equilibrium probability of each state (see Methods). Indeed, we achieve an impressive τ_b of 0.70 when using MSM-weighted pKi predictions (Figure 5B). Thus, combining MSMs with the AtomNet pKi predictor may improve the performance of virtual screening.

Discussion

Protein phosphatases are a challenging class of drug targets that broadly illustrate the advantages of using allosteric compounds (Köhn, 2020). There are nearly 200 phosphatases in the human genome, and many are implicated in human diseases, including diabetes (Krishnan et al., 2018), neurodegeneration (Vieira et al., 2017), and multiple cancers (Pecháčková et al., 2017). Phosphatases are downstream targets of several signaling pathways that integrate various cellular signals (Lu et al., 2008). This suggests that targeting of phosphatases may be useful across numerous cancer subtypes caused by mutations of upstream proteins or in cases where tumors develop resistance to upstream therapies. However, to the best of our knowledge, there are no approved therapies that target phosphatases. Previous drug discovery efforts have focused on active site inhibitors. Targeting the active site has proved challenging because high sequence conservation limits the selectivity of compounds. Furthermore, compounds targeting the active site need to be highly charged, limiting their bioavailability. Hence, allosteric compounds, like the CAA compounds that target PPM1D and novel allosteric inhibitors of SHP2 (Chen et al., 2016), may be needed to successfully inhibit phosphatases in clinical settings.

Definitively establishing the binding site of the PPM1D allosteric compounds remains challenging, but our results predict a plausible binding site that agrees with most previous experiments. Photoaffinity labeling experiments and flap swap experiments, which showed that introducing the PPM1D flap domain can sensitize other phosphatases to the PPM1D allosteric inhibitors, strongly implicate the flap domain as the primary compound binding site. Our proposed binding site at the flap-hinge

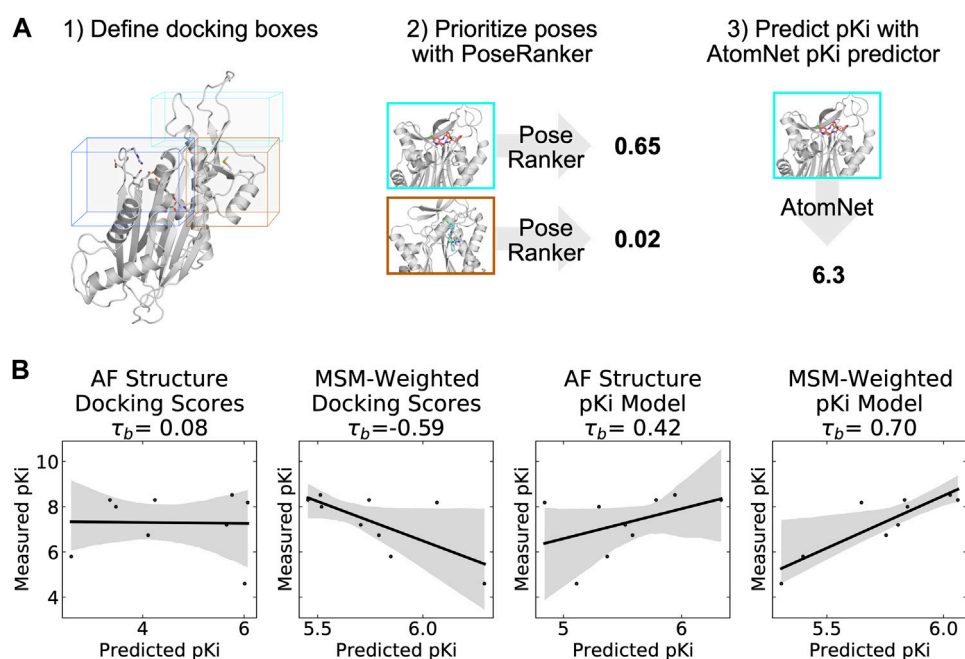


FIGURE 5

A neural network trained to predict pKi accurately ranks allosteric compounds by potency when applied to structures from a PPM1D ensemble. **(A)** Schematic highlighting the procedure that was used for selecting a single pose for each PPM1D cluster center in the MSM. For each MSM cluster center, we defined multiple docking boxes based on the active site, residues involved in photolabeling experiments, and P2Rank pockets at the flap and hinge. After performing docking, we selected a best pose per MSM state using the PoseRanker neural network. Finally, we fed this best docked pose to the AtomNet pKi predictor. **(B)** MSM-weighting of the pKi predictions from the AtomNet pKi predictor outperforms docking-based methods as well as a single pKi prediction based on the AlphaFold-predicted structure. For each scatter plot, we show the line of best fit in black as well as the 95% confidence interval based on bootstrapping in translucent grey bands. We report the Kendall rank correlation coefficient, a statistic that measures the ordinal association between the predicted pKi and the measured pKi and whose maximum value is 1.

interface is consistent with these results. Though our proposed binding mode does not directly involve the points of covalent attachment (i.e., P219 and M236), we speculate that the large photoactivatable benzophenone groups that were added to the compound scaffold enable compounds with these groups to bind at our proposed site but still reach these residues. Furthermore, Gilmartin et al. showed that residues 247–268 in the flap are not essential for PPM1D allosteric compound binding (Gilmartin et al., 2014). Consistent with these results, our proposed binding site does not involve these residues with the minor exception of K247. On the other hand, Miller et al. demonstrated that deletion of the hinge causes a ~1000-fold decrease in binding affinity and a 100-fold increase in IC₅₀ for one of the allosteric compounds. Our proposed binding site has substantial involvement from hinge residue L157 and an adjacent residue W154. As a result, our proposed binding site is consistent with the hinge deletion experiments. However, given that residues in the flap, especially residues D277 to V289, are commonly involved in high-quality poses, we cannot explain why Miller et al. report that flap deletion (specifically residues 219–287) has no effect on binding affinity or binding kinetics. We speculate that it may be possible for the allosteric inhibitors to bind even when most of the flap is deleted, but our analysis suggests further experiments are needed to disentangle the relative contributions of the flap and hinge to compound binding.

Furthermore, our results highlight the advantages of explicitly accounting for protein conformational heterogeneity when using

deep learning methods for predicting compound affinity. The AtomNet pKi predictor is designed and trained to be pose-sensitive (Wallach et al., 2015; Gniewek et al., 2021; Stafford et al., 2022). Its performance at ranking compounds varies widely between target structures in the MSM (Supplementary Figure S14). We noticed that even when the poses are likely of poor quality (e.g., the AF structure where the cryptic pocket is not present), we still often see relatively good predictive performance for the pKis. While some of the predictive power of the AtomNet pKi predictor is driven by the pose, we hypothesize that the ligand features might also play a part in and influence the predicted pKis that AtomNet pKi predictor outputs. For the cases where the pose is poor (e.g., docking against AF structure), we get a baseline for how well a ligand-based model would perform. The boost in performance seen with MSM-docking is likely due to better poses resulting from docking to structures with open cryptic pockets.

Our results also provide insight into how ligand features contribute to differences in predicted affinity. We inspected top-scoring poses for the compounds shown in Figure 1A to assess why these compounds have significantly different binding affinities despite differing by just a carbonyl (Supplementary Figure S14). While we cannot single out a specific interaction formed by this differing carbonyl across the inspected top poses, this moiety was consistently in contact with residues Y281 to F284. For certain states, we observed additional polar interactions between the carbonyl-containing compound and the protein. We also observed

that in states where the pKi prediction strongly favored the carbonyl-containing compounds, the carbonyl-containing compound buried substantially more protein solvent accessible surface area in its docked pose (Supplementary Figure S14). However, we acknowledge that comparing docked poses does not provide definitive insight into why the AtomNet pKi predictor makes higher predictions for active compound poses.

While a traditional docking scoring function does not accurately rank compounds based on their binding affinity (Figure 5B), it is well established in the literature that docking can be used to sample good binding poses. That is the case even though the poses that best capture the correct binding mode are usually not the best-ranked or highest-scoring ones. We and others have shown that this issue can be mitigated by sampling multiple poses and re-ranking them with a deep learning model (Stafford et al., 2022), often yielding sufficiently accurate poses for applications such as the one included in this manuscript. Recent advances in molecular docking (Corso et al., 2022) have led to docking tools that do away with explicitly defined docking scoring functions. If these tools could be run at the scale needed to generate training data for the AtomNet pKi predictor, it is possible that the model could achieve better predictive performance.

Nonetheless, our results show that MSMs can address some of the limitations of rigid docking against AlphaFold predicted protein structures. Rigid docking has lower performance when the protein structure(s) being used for docking corresponds to an *apo* or unbound state (Abagyan et al., 2010). Deep learning-based (DL-based) protein structure prediction methods like AlphaFold, are trained using all available data on the PDB, and there is data to support that output structures are somewhere in between *apo* and *holo*. (Saldan  et al., 2022). Docking efforts against AlphaFold structures show lower performance than against *holo* structures available on the PDB. (D az-Rovira et al., 2022; Wong et al., 2022). Here, we show that this can be mitigated by considering conformational heterogeneity using MSMs. Using a highly flexible system, we can sample conformations and identify cryptic pockets that can be successfully used in downstream virtual screening applications. While our work was based off a single AF structure as a starting point, we are aware of efforts to use these DL protein structure prediction tools to sample multiple conformations, thus better capturing protein flexibility (Saldan  et al., 2022; Meller et al., 2023a). To our knowledge, these methods have not been compared against MSM approaches and more research would be needed before conducting a similar analysis as described herein with a DL-generated structural ensemble.

Despite these encouraging results, there are notable limitations to our approach. Firstly, most of our pKi analyses included nine capped amino acid compounds. This is not a particularly large dataset, and we acknowledge that this is somewhat restrictive in terms of establishing robust statistical significance for our results. Ranking based on docking scores output by CUina does suggest that this is not a trivial ranking problem, and that achieving good predictive performance at random, despite the small data set size, is statistically unlikely. While in an ideal scenario we would hope to have a larger number of data points to validate our findings, affinity data is often relatively sparse at early stages of the pharmaceutical pipeline, so estimating the performance of virtual screening can be difficult. Secondly, our data suggests that the AtomNet pKi predictor tends to regress to the mean. Even though the ranking metrics are good, the dynamic range of predicted vs. observed

pKis differ significantly. We hypothesize that this is likely due to a data imbalance in the training data of the AtomNet pKi predictor, as data points in the extremes of the pKi distribution (either very high or very low) are rare, and our sampling strategy during training does not stratify on that property. Still, given that model accurately ranks compounds by potency, our approach represents a promising strategy for novel virtual screening campaigns.

Conclusion

In summary, we have uncovered a cryptic pocket at the PPM1D flap-hinge interface that improves the ability to predict the potency of PPM1D inhibitors. AlphaFold predicts a PPM1D structure that lacks high scoring allosteric pockets at proposed binding sites based on an analysis conducted using the P2Rank and LIGSITE pocket detection algorithms. Though the AF-predicted structure lacks allosteric pockets, molecular dynamics simulations of ligand-free PPM1D capture a cryptic pocket at the flap-hinge interface. A neural network trained to evaluate the quality of docked poses predicts that this site is the most likely binding mode for the PPM1D allosteric inhibitors. Finally, by docking compounds to this pocket and using a structure-based pKi predictor, we demonstrate that aggregating pKi predictions across a MSM is superior at ranking compounds than using docking scores or using the single predicted AlphaFold structure. Thus, our methodology provides a promising template for structure-based drug discovery and *in silico* binding site prediction.

Methods

Molecular dynamics simulations

The AlphaFold predicted structure (AF-O15297) was used as an initial structure for PPM1D simulations since no structures were available in the PDB. However, because several PPM1D domains (C-terminus domain and an internal loop stretching from residue 39–92) are predicted to be disordered (pLDDT <70) and because we were primarily interested in flap domain dynamics, we removed residues 39–92 and truncated the C-terminus (residue 396–end).

GROMACS (Abraham et al., 2015) was used to prepare and to simulate PPM1D using the CHARMM36m force fields (Huang et al., 2016). The protein structure was solvated in a dodecahedral box of TIP3P water (Jorgensen et al., 1983) that extended 1 nm beyond the protein in every dimension. Thereafter, sodium and chloride ions were added to the system to maintain charge neutrality and 0.1 M NaCl concentration. The system was minimized using steepest descents until the maximum force on any atom decreased below 1,000 kJ/(mol x nm). The system was then equilibrated with all atoms restrained in place at 310 K maintained by the Bussi-Parinello thermostat (Bussi et al., 2007) and the Parrinello-Rahman barostat (Parrinello and Rahman, 1998).

Production simulations were performed in the CHARMM36m forcefield. Simulations were run in the NPT ensemble at 310 K using the leapfrog integrator, Bussi-Parinello thermostat, and the Parrinello-Rahman barostat. A 12   cutoff distance was utilized with a force-based switching function starting at 10  . Periodic boundary conditions and the PME method were utilized to calculate the long-range

electrostatic interactions with a grid density greater than 1.2 \AA^{-3} . Hydrogen bonds were constrained with the LINCS algorithm (Hess et al., 1997) to enable the use of a constant integration timestep of 2 fs

Adaptive sampling

We used the Fluctuation Amplification of Specific Traits (FAST) algorithm (Zimmerman and Bowman, 2015) to explore a diverse ensemble of states with cryptic pockets. We performed 5 generations of simulations; each generation consisted of 10 parallel simulations 40 ns in length (total aggregate simulation time: 2 microseconds of adaptive sampling). After each completed generation, we selected seeds for the next round based on an objective function. We used an objective function that rewarded states based on their total pocket volume as measured by LIGSITE (Hendlich et al., 1997). The following LIGSITE parameters were used: a minimum rank of 7, a minimum cluster size of 3, and a probe radius of 0.14 nm. Our ranking function also included a term that penalizes states conformationally similar to others already selected (the width parameter for this term was 1.5 times the cluster radius) (Zimmerman et al., 2017). We performed *k*-centers clustering after each round of FAST with the RMSD of C-alpha positions of the entire protein as the distance metric. Clustering continued until the maximal distance from each point to its nearest cluster center was a maximum of 2 Å C-alpha RMSD. The top 10-scoring cluster centers based on the LIGSITE objective function were then selected for the next round of FAST.

To generate Markov state models from the MD simulations, we applied a $1/n$ pseudocount to each element of the transition counts matrix and then performed row normalization to generate a transition matrix as recommended in (Zimmerman et al., 2018). Markov state models were generated using the *enspara* software package (Porter et al., 2019).

P2Rank pocket detection

We used P2Rank v2.4 (Krivák and Hoksza, 2018) with default parameters to identify pockets across all of the representative states (cluster centroids) from our simulations. For subsequent analyses, we consider only pockets with a permissive pocket probability (as output by P2Rank) greater than 0.2.

Docking

We docked compounds using a proprietary GPU-enabled docking engine, CUina. CUina (Stafford et al., 2022) is a proprietary implementation of *smina* (Koes et al., 2013), which has been parallelized and refactored to operate more efficiently on a GPU. The scoring function (Vina scoring function) and sampling routines of CUina are analogous to those in *smina*. CUina requires a bounding box to restrict its search space. We defined four bounding boxes representing each of the three proposed binding sites for CAA compounds, and one negative control (active site). For the first two boxes, we used the coordinates of the pockets identified by P2Rank in the vicinity of the flap or the hinge of

PPM1D (where available). The minimum and maximum coordinates of the voxels output by P2Rank were used to define the box, and we padded these coordinates by 5 Å along each dimension. A third box was defined using the coordinates of the two residues (P219 and M236) that were part of the photolabeling experiment described by Gilmartin et al. The fourth and final boxed was defined based on the active site: we used the coordinates of all the catalytic residues to define the box. The box boundaries were calculated by taking the minimum and maximum coordinates of all photolabeling or catalytic residues and padding by 5 Å along each dimension.

We docked nine CAA compounds to all states (i.e., a representative structure for each MSM state) resulting from the MSM effort described above. For each compound, we dock the best (minimized) ligand conformation against all four proposed binding sites. In the MSM states where P2Rank failed to identify one of the pockets, docking against that pocket was omitted.

For each docking operation corresponding to a binding site + MSM representative structure + compound, we output 64 poses and imposed a 1 Å RMSD similarity cutoff, thus ensuring that the poses output are sufficiently different from one another.

Pose classification

Following docking, poses were classified based on the contacts that they formed. Specifically, we found residues whose heavy atoms were within 4 Å of a ligand heavy atom. Next, we classified poses into the following categories based on their list of contact residues: flap domain only, hinge only, flap-domain interface, and active site. The active site was defined as residues 18, 22, 23, 105, 106, 192, 218, 314, and 366 based on the annotation in (Gilmartin et al., 2014); the flap domain was defined as residues 219–288; and the hinge domain was defined as residues 150–167, which includes both a loop and half the helix spanning residues 136–158. If the compound made contacts with both a hinge domain and a flap domain residue, it was classified as binding in the flap-hinge interface.

pKi model predictions

We used AtomNet's pKi predictor to perform pKi predictions using the poses generated and selected by our pose generation pipeline (CUina + ANPR). AtomNet's global pKi model uses a graph-based convolutional neural network to regress over pKi.

Data: This model was trained using a combination of public and proprietary data, spanning more than 4,000 targets for which activity measurements were available. In total, several million activity data points were used to train the model. PPM1D was not part of the training data for the model, but the training set did include a number of other phosphatases.

Architecture

AtomNet's global pKi model uses the GRAPHite architecture (previously described in (Stafford et al., 2022)). The GRAPHite architecture is a directed Graph Convolutional Network (GCN)

comprised of four graph convolutional layers. The first two layers include both ligand and receptor features, whereas the last two layers are ligand-only. Nodes in the graph represent ligand and receptor atoms. Only receptor atoms within 7 Å of any ligand atom were used as part of the graph. Edges were defined by atoms within 4 Å of each other and edge weights were distance-dependent. The final layer is sum-pooled into an embedding. This embedding is then passed through two (independent) multilayer perceptrons to predict two outputs: the ANPR pose quality score, and the Vina docking score. Those outputs are then concatenated to the embedding and passed through a third multilayer perceptron which outputs the predicted pKi.

More details about the method and parameters can be found in (Gniewek et al., 2021; Stafford et al., 2022).

MSM-weighting of docking and pKi predictions

To determine an overall MSM-weighted pKi prediction from pKi predictions for each MSM state, we first selected a single highest scoring pose for each state based on the AtomNet PoseRanker predictions. Next, we converted the predicted pKi value to an association constant. Then, we found a macro-association constant from the individual micro-association constants:

$$K_a = \sum_i \pi_i K_{a_i}$$

We use association constants because this ensures that large contributions to the sum come from states with either a high equilibrium probability, a large association constant (i.e., favor ligand binding), or both. States that have small association constants or low equilibrium probabilities will have a minimal contribution to the overall association constant. Finally, we convert the overall association constant to a pKi by taking the -log10 of its inverse.

For docking scores which are in units of kcal/mol, we follow a similar procedure. Given there were multiple poses for each MSM state, we selected the pose with the highest ANPR prediction for that state. Docking scores are then converted to association constants:

$$K_a = e^{\frac{-\Delta G_{\text{docking}}}{RT}}$$

Then we follow the same aggregation procedure:

$$K_a = \sum_i \pi_i K_{a_i}$$

Finally, we convert this overall association constant into a pKi by taking the -log10 of its inverse.

References

Abagyan, R., Rueda, M., and Bottegoni, G. (2010). Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* 50, 186–193. doi:10.1021/ci9003943

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://alphafold.ebi.ac.uk/entry/B1WCA0>.

Author contributions

AM and SO carried out the simulations, analyzed the results, and wrote the manuscript. AD and TA analyzed the results and edited the manuscript. HB and GB conceived, designed, oversaw the research, and edited the manuscript.

Funding

AM was supported by the National Institutes of Health F30 Fellowship (1F30HL162431-01A1). GB holds a NSF grant MCB 2218156 (GB) and NIH grants R01 GM124007 (GB) and RF1AG067194 (GB). GB holds a Packard Fellowship for Science and Engineering from The David and Lucile Packard Foundation.

Acknowledgments

We would like to thank Atomwise for financing this project.

Conflict of interest

SO and HB are employees and hold equity in Atomwise, Inc. AD and TA are former employees of Atomwise.

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.

GRB is a co-founder of Decrypt Bio.

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2023.1171143/full#supplementary-material>

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

- 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
- Bowman, G. R., Bolin, E. R., Hart, K. M., Maguire, B. C., and Marqusee, S. (2015). Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc. Natl. Acad. Sci. U. S. A.* 112, 2734–2739. doi:10.1073/pnas.1417811112
- Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126, 014101. doi:10.1063/1.2408420
- Castellino, R. C., de Bortoli, M., Lu, X., Moon, S. H., Nguyen, T. A., Shepard, M. A., et al. (2008). Medulloblastomas overexpress the p53-inactivating oncogene WIP1/PPM1D. *J. Neurooncol.* 86, 245–256. doi:10.1007/s11060-007-9470-8
- Chen, Y. N. P., Lamarche, M. J., Chan, H. M., Fekkes, P., Garcia-Fortanet, J., Acker, M. G., et al. (2016). Allosteric inhibition of SHP2 phosphatase inhibits cancers driven by receptor tyrosine kinases. *Nature* 535, 148–152. doi:10.1038/nature18621
- Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. (2022). DiffDock: Diffusion steps, twists, and turns for molecular docking. Available at: <https://arxiv.org/abs/2210.01776v2> (Accessed April 1, 2023).
- Cruz, M. A., Frederick, T. E., Mallimadugula, U. L., Singh, S., Vithani, N., Zimmerman, M. I., et al. (2022). A cryptic pocket in Ebola VP35 allosterically controls RNA binding. *Nat. Commun.* 13, 2269–2310. doi:10.1038/s41467-022-29927-9
- Díaz-Rovira, A. M., Martín, H., Beuming, T., Díaz, L., Guallar, V., and Ray, S. S. (2022). Are deep learning structural models sufficiently accurate for virtual screening? Application of docking algorithms to AlphaFold2 predicted structures. *bioRxiv*, 2022.08.18.504412. doi:10.1101/2022.08.18.504412
- Gilmartin, A. G., Fagit, T. H., Richter, M., Groy, A., Seefeld, M. A., Darcy, M. G., et al. (2014). Allosteric Wip1 phosphatase inhibition through flap-subdomain interaction. *Nat. Chem. Biol.* 10, 181–187. doi:10.1038/nchembio.1427
- Gniewek, P., Worley, B., Stafford, K., van den Bedem, H., and Anderson, B. (2021). Learning physics confers pose-sensitivity in structure-based virtual screening.
- Hendlich, M., Rippmann, F., and Barnickel, G. (1997). LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph Model.* 15, 359–389. doi:10.1016/S1093-3263(98)00002-3
- Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h(199709)18:12
- Hollingsworth, S. A., and Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron* 99, 1129–1143. doi:10.1016/j.neuron.2018.08.011
- Hollingsworth, S. A., Kelly, B., Valant, C., Michaelis, J. A., Mastromihalis, O., Thompson, G., et al. (2019). Cryptic pocket formation underlies allosteric modulator selectivity at muscarinic GPCRs. *Nat. Commun.* 10, 3289–9. doi:10.1038/s41467-019-11062-7
- Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., et al. (2016). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14, 71–73. doi:10.1038/nmeth.4067
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869
- 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
- Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9, 646–652. doi:10.1038/nsb0902-646
- Koes, D. R., Baumgartner, M. P., and Camacho, C. J. (2013). Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* 53, 1893–1904. doi:10.1021/CJ300604Z
- Köhn, M. (2020). Turn and face the strange: A new view on phosphatases. *ACS Cent. Sci.* 6, 467–477. doi:10.1021/acscentsci.9b00909
- Krishnan, N., Konidaris, K. F., Gasser, G., and Tonks, N. K. (2018). A potent, selective, and orally bioavailable inhibitor of the protein-tyrosine phosphatase PTP1B improves insulin and leptin signaling in animal models. *J. Biol. Chem.* 293, 1517–1525. doi:10.1074/JBC.117.819110
- 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, 39–12. doi:10.1186/s13321-018-0285-8
- Li, J., Yang, Y., Peng, Y., Austin, R. J., van Eyndhoven, W. G., Nguyen, K. C. Q., et al. (2002). Oncogenic properties of PPM1D located within a breast cancer amplification epicenter at 17q23. *Nat. Genet.* 31, 133–134. doi:10.1038/ng888
- Lu, X., Nguyen, T. A., Moon, S. H., Darlington, Y., Sommer, M., and Donehower, L. A. (2008). The type 2C phosphatase Wip1: An oncogenic regulator of tumor suppressor and DNA damage response pathways. *Cancer Metastasis Rev.* 27, 123–135. doi:10.1007/s10555-008-9127-x
- Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., et al. (2019). Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224–229. doi:10.1038/s41586-019-0917-9
- Meller, A., Bhakat, S., Solieva, S., and Bowman, G. R. (2023a). Accelerating cryptic pocket discovery using AlphaFold. *J. Chem. Theory Comput.* doi:10.1021/ACS.JCTC.2C01189
- Meller, A., Lotthammer, J. M., Smith, L. G., Novak, B., Lee, L. A., Kuhn, C. C., et al. (2023b). Drug specificity and affinity are encoded in the probability of cryptic pocket opening in myosin motor domains. *Elife* 12, e83602. doi:10.7554/ELIFE.83602
- Meller, A., Ward, M., Borowsky, J., Kshirsagar, M., Lotthammer, J. M., Oviedo, F., et al. (2023c). Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat. Commun.* 14, 1177–1215. doi:10.1038/s41467-023-36699-3
- Miller, P. G., Sathappa, M., Moroco, J. A., Jiang, W., Qian, Y., Iqbal, S., et al. (2022). Allosteric inhibition of PPM1D serine/threonine phosphatase via an altered conformational state. *Nat. Commun.* 13, 3778–3816. doi:10.1038/s41467-022-30463-9
- Mullard, A. (2018). Phosphatases start shedding their stigma of undruggability. *Nat. Rev. Drug Discov.* 17, 847–849. doi:10.1038/NRD.2018.201
- 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, 99–105. doi:10.1016/j.jymeth.2010.06.002
- Parrinello, M., and Rahman, A. (1998). Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* 52, 7182–7190. doi:10.1063/1.328693
- Pecháková, S., Burdová, K., and Macurek, L. (2017). WIP1 phosphatase as pharmacological target in cancer therapy. *J. Mol. Med. Berl.* 95, 589–599. doi:10.1007/S00109-017-1536-2
- Porter, J. R., Zimmerman, M. I., and Bowman, G. R. (2019). Enspira: Modeling molecular ensembles with scalable data structures and parallel computing. *J. Chem. Phys.* 150, 044108. doi:10.1063/1.5063794
- Saldanõ, T., Escobedo, N., Marchetti, J., Zea, D. J., mac Donagh, J., Velez Rueda, A. J., et al. (2022). Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* 38, 2742–2748. doi:10.1093/BIOINFORMATICS/BTAC202
- Stafford, K. A., Anderson, B. M., Sorenson, J., and van den Bedem, H. (2022). AtomNet PoseRanker: Enriching ligand pose quality for dynamic proteins in virtual high-throughput screens. *J. Chem. Inf. Model.* 62, 1178–1189. doi:10.1021/ACS.JCIM.1C01250/ASSET/IMAGES/LARGE/CI1C01250_0005.JPEG
- Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., et al. (2019). Comparative assessment of scoring functions: The CASF-2016 update. *J. Chem. Inf. Model.* 59, 895–913. doi:10.1021/ACS.JCIM.8B00545/ASSET/IMAGES/LARGE/CI-2018-00545U_0010.JPEG
- Sztain, T., Amaro, R., and McCammon, J. A. (2021). Elucidation of cryptic and allosteric pockets within the SARS-CoV-2 main protease. *J. Chem. Inf. Model.* 61, 3495–3501. doi:10.1021/acs.jcim.1c00140
- Tan, D. S. P., Lambros, M. B. K., Rayter, S., Natrajan, R., Vatcheva, R., Gao, Q., et al. (2009). PPM1D is a potential therapeutic target in ovarian clear cell carcinomas. *Clin. Cancer Res.* 15, 2269–2280. doi:10.1158/1078-0432.CCR-08-2403
- 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
- Vieira, M. N. N., Lyra e Silva, N. M., Ferreira, S. T., and de Felice, F. G. (2017). Protein tyrosine phosphatase 1B (PTP1B): A potential target for alzheimer's therapy? *Front. Aging Neurosci.* 9, 7. doi:10.3389/fnagi.2017.00007
- Vijayan, R. S. K., He, P., Modi, V., Duong-Ly, K. C., Ma, H., Peterson, J. R., et al. (2015). Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. *J. Med. Chem.* 58, 466–479. doi:10.1021/jm501603h
- Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. doi:10.48550/arxiv.1510.02855
- Wankowicz, S. A., de Oliveira, S. H. P., Hogan, D. W., van den Bedem, H., and Fraser, J. S. (2022). Ligand binding remodels protein side chain conformational heterogeneity. *Elife* 11, e74114. doi:10.7554/ELIFE.74114
- Wilson, C. J., Choy, W. Y., and Karttunen, M. (2022). AlphaFold2: A role for disordered protein/region prediction? *Int. J. Mol. Sci.* 23, 4591. doi:10.3390/ijms23094591
- Wong, F., Krishnan, A., Zheng, E. J., Sté Ark, H., Manson, A. L., Earl, A. M., et al. (2022). Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* 18, e11081. doi:10.15252/MSB.202211081
- Zimmerman, M. I., and Bowman, G. R. (2015). FAST conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.* 11, 5747–5757. doi:10.1021/acs.jctc.5b00737
- Zimmerman, M. I., Hart, K. M., Sibbald, C. A., Frederick, T. E., Jimah, J. R., Knoverek, C. R., et al. (2017). Prediction of new stabilizing mutations based on mechanistic insights from Markov state models. *ACS Cent. Sci.* 3, 1311–1321. doi:10.1021/ACSCENTSCI.7B00465/ASSET/IMAGES/OC-2017-004659_M006.GIF
- Zimmerman, M. I., Porter, J. R., Sun, X., Silva, R. R., and Bowman, G. R. (2018). Choice of adaptive sampling strategy impacts state discovery, transition probabilities, and the apparent mechanism of conformational changes. *J. Chem. Theory Comput.* 14, 5459–5475. doi:10.1021/acs.jctc.8b00500
- Zimmerman, M. I., Porter, J. R., Ward, M. D., Singh, S., Vithani, N., Meller, A., et al. (2021). SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat. Chem.* 1, 651–659. doi:10.1038/s41557-021-00707-0



OPEN ACCESS

EDITED BY

Fabio Polticelli,
Roma Tre University, Italy

REVIEWED BY

Haiping Zhang,
Chinese Academy of Sciences (CAS),
China
Jian Wang,
The Pennsylvania State University,
United States

*CORRESPONDENCE

Calvin Yu-Chian Chen,
✉ chenyuchian@mail.sysu.edu.cn

RECEIVED 29 May 2023

ACCEPTED 13 June 2023

PUBLISHED 27 June 2023

CITATION

Wu X, Li Z, Chen G, Yin Y and Chen CY-C
(2023), Hybrid neural network
approaches to predict drug–target
binding affinity for drug repurposing:
screening for potential leads for
Alzheimer's disease.
Front. Mol. Biosci. 10:1227371.
doi: 10.3389/fmolb.2023.1227371

COPYRIGHT

© 2023 Wu, Li, Chen, Yin and Chen. 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.

Hybrid neural network approaches to predict drug–target binding affinity for drug repurposing: screening for potential leads for Alzheimer's disease

Xialin Wu^{1,2}, Zhuojian Li³, Guanxing Chen³, Yiyang Yin³ and Calvin Yu-Chian Chen^{3,4,5*}

¹School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China, ²Guangzhou University of Chinese Medicine, Guangzhou, China, ³Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen, China, ⁴Department of Medical Research, China Medical University Hospital, Taichung, Taiwan, ⁵Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan

Alzheimer's disease (AD) is a neurodegenerative disease that primarily affects elderly individuals. Recent studies have found that sigma-1 receptor (S1R) agonists can maintain endoplasmic reticulum stress homeostasis, reduce neuronal apoptosis, and enhance mitochondrial function and autophagy, making S1R a target for AD therapy. Traditional experimental methods are costly and inefficient, and rapid and accurate prediction methods need to be developed, while drug repurposing provides new ways and options for AD treatment. In this paper, we propose HNNDTA, a hybrid neural network for drug–target affinity (DTA) prediction, to facilitate drug repurposing for AD treatment. The study combines protein–protein interaction (PPI) network analysis, the HNNDTA model, and molecular docking to identify potential leads for AD. The HNNDTA model was constructed using 13 drug encoding networks and 9 target encoding networks with 2506 FDA-approved drugs as the candidate drug library for S1R and related proteins. Seven potential drugs were identified using network pharmacology and DTA prediction results of the HNNDTA model. Molecular docking simulations were further performed using the AutoDock Vina tool to screen haloperidol and bromperidol as lead compounds for AD treatment. Absorption, distribution, metabolism, excretion, and toxicity (ADMET) evaluation results indicated that both compounds had good pharmacokinetic properties and were virtually non-toxic. The study proposes a new approach to computer-aided drug design that is faster and more economical, and can improve hit rates for new drug compounds. The results of this study provide new lead compounds for AD treatment, which may be effective due to their multi-target action. HNNDTA is freely available at <https://github.com/lizhj39/HNNDTA>.

KEYWORDS

Alzheimer's disease, drug repurposing, hybrid neural network, molecular docking, sigma-1 receptor

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that mainly affects elderly people and whose etiology remains unclear. The symptoms of patients include a decline in cognitive abilities and a weakening of memory and thinking abilities (Hung and Fu, 2017; Srivastava et al., 2021; Briggs et al., 2016). Although there are some drugs currently used to treat AD, their effectiveness is limited. However, drug repurposing (DR) has provided a new approach and selection for the treatment of AD (Padhi and Govindaraju, 2022; Ihara and Saito, 2020). This method involves reanalyzing the biological effects of known drugs and applying them to new areas of disease treatment. DR can accelerate the development of new drugs, provide more treatment options, and reduce the risk of drug development.

Previous studies have suggested that sigma-1 receptor (S1R) has neuroprotective effects and that its physiological function has a direct impact on endogenous neuroprotective mechanisms (Voronin et al., 2023). As a protein chaperone, S1R locates on specialized lipid rafts of mitochondria-associated endoplasmic reticulum membranes (MAMs), which are known to form mitochondrial endoplasmic reticulum contacts (MERCs) with the outer mitochondrial membrane and play a role in various biochemical processes, such as autophagosome formation, cellular energy production, and maintenance of *IR3R3*-dependent calcium homeostasis. Thus, disruption of this structure is now considered an early stage in the pathogenesis of neurodegenerative diseases, including AD. Activation of S1R using agonists has been shown to maintain the structural and functional stability of MAMs and MERCs, thereby enhancing autophagic activity, restoring mitochondrial function, and regulating intracellular calcium balance (Barazzuol et al., 2021; Leal and Martins, 2021; Wilson and Metzakopian, 2021; Weng et al., 2017). In AD models, such as PS1-KI and APP-KI, dendritic spines of hippocampal neurons are lost both *in vitro* and *in vivo*, indicating that the loss of mushroom-shaped "memory spines" reflects cognitive decline, learning, and memory deficits in AD (Ryskamp et al., 2019; Fisher et al., 2015), suggesting the involvement of reduced S1R in AD pathology. The mixed muscarinic/S1R agonist AF710B stabilizes mature mushroom spines in hippocampal cultures derived from AD mice *in vitro*, while pridopidine, an S1R agonist, stabilizes mushroom spines in an Alzheimer's mouse model through its action on S1R. S1R agonists have demonstrated preclinical efficacy in AD animal models (Ryskamp et al., 2019; Fisher et al., 2015). Donepezil, a potent acetylcholinesterase inhibitor used for AD treatment, is also a high-affinity S1R ligand. Precise pharmacological studies on the interaction between donepezil and S1R suggest that the drug exerts anti-amnesic effects primarily through S1R activation against scopolamine, β -amyloid, or carbon monoxide-induced memory impairments (Hassan et al., 2017). Overall, S1R agonists exhibit neuroprotective effects and modulate synaptic plasticity, making S1R a potential target for AD treatment.

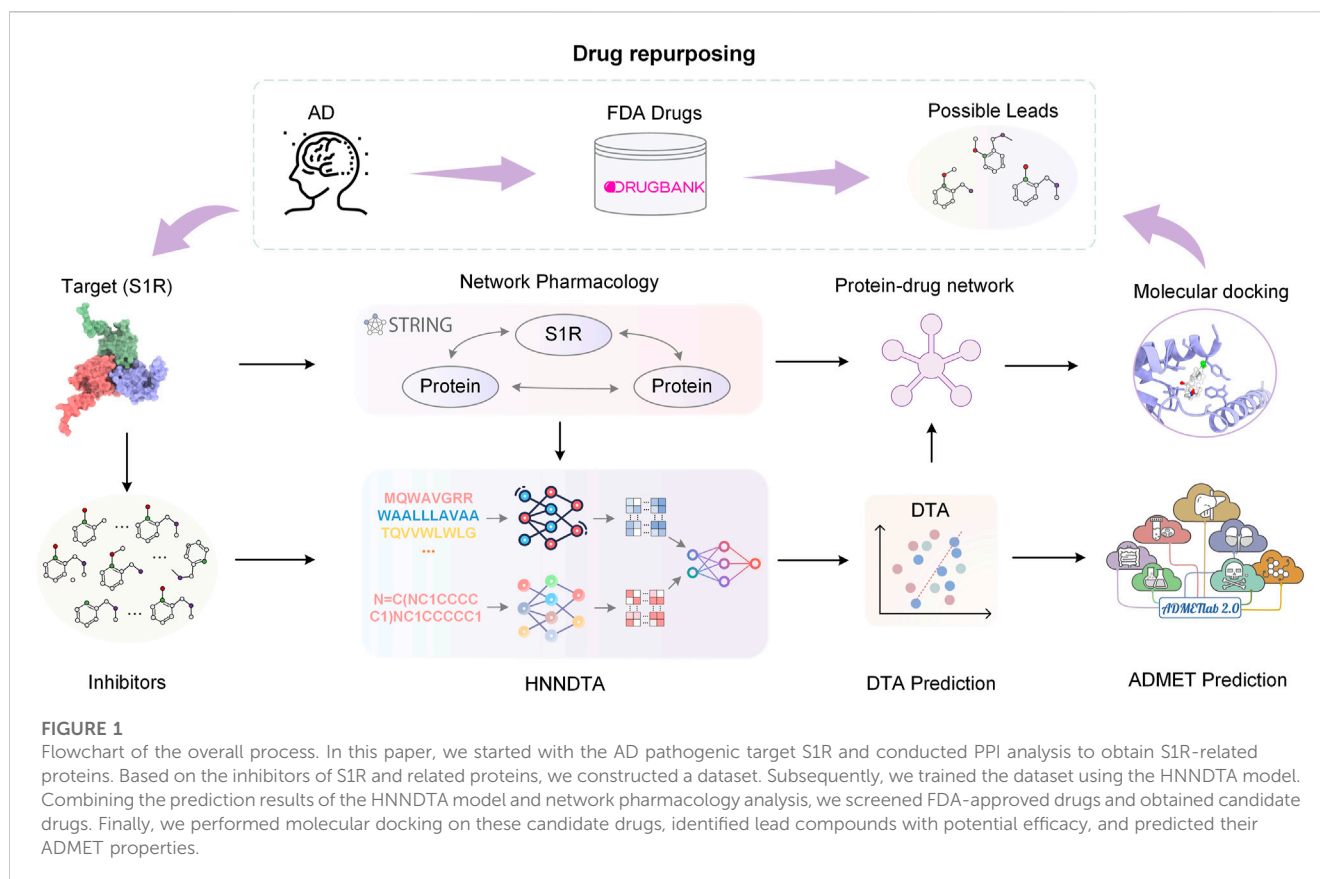
In the past decade, the "one disease–one target–one drug" paradigm has dominated the approach to drug discovery. However, this paradigm has certain limitations, and recent advances in systems biology have shifted the focus from "single-target drugs to "multi-target drugs" (Noor et al., 2023). When treating a particular disease, it is not feasible to rely solely on a

single target to identify drugs. Instead, a range of targets within an imbalanced pathway in the complex biological network must be considered as inhibiting a single enzyme alone may lead to cancer cells compensating by activating other enzymes (Ryskamp et al., 2019; Fisher et al., 2015; Hassan et al., 2017). Zhi et al. utilized network pharmacology and molecular docking to reveal dihydroorotate dehydrogenase (DHODH) as a therapeutic target for small-cell lung cancer. Subsequently, they constructed a prediction model using graph neural networks (GNNs) and traditional machine learning methods to screen for potential DHODH inhibitors (Noor et al., 2023; Zhi et al., 2021). Cantini et al. introduced a multi-network strategy by integrating multiple genomic information layers, particularly gene co-expression and protein–protein interactions, to identify cancer-related targets. They employed consensus clustering algorithms in a predictive network, revealing CD46, BTG2, ATF3, HDGF, and F11R as driver genes in cancer (Noor et al., 2023; Cantini et al., 2015).

In drug repurposing, artificial intelligence (AI) plays an important role. By analyzing data on existing drugs and diseases using machine learning and deep learning methods, potential drugs can be quickly and efficiently screened (Cheng and Cummings, 2022; Yin and Wong, 2021; Vatansever et al., 2021). In addition, simulating the interactions between drugs and proteins can predict drug activity and affinity, guiding drug repositioning research. In recent years, researchers have successfully screened many promising drugs using AI methods (Selvaraj et al., 2021; Malandraki-Miller and Riley, 2021; Patel et al., 2020). These studies indicate that drug repositioning has important clinical application prospects, and AI methods can provide more powerful support for drug repositioning.

The affinity between drugs and targets is the basis for drug action, and predicting the affinity between drugs and targets is an important part of drug repurposing (Pushpakom et al., 2019; Parisi et al., 2020). Traditional experimental methods have disadvantages such as high cost and low efficiency, making it necessary to develop a fast and accurate prediction method. In recent years, with the development of deep learning technology, using neural networks to predict the affinity between drugs and targets has gradually become a research hotspot (Thomas et al., 2022; Choudhury et al., 2022; Jiang et al., 2022; Wang and Dokholyan, 2022). Neural networks are powerful computational tools with the ability to deal with non-linear problems and have achieved some success in predicting the affinity between drugs and targets.

In recent years, more and more researchers have begun to explore the use of neural networks to construct computational models for drug repositioning prediction to screen drugs for treating AD (Chyr et al., 2022; Wu et al., 2022; Siavelis et al., 2016). Some related studies have made some progress. For example, Zhou et al. Fang et al. (2022) proposed an integrated network-based AI method that can quickly translate genome-wide association study findings and multi-omics data into genotype-based therapeutic discoveries in AD, and identified pioglitazone as a potential new method for treating AD using AI methods. Tsuji et al. (2021) developed a deep learning-based computational framework that can extract low-dimensional representations of high-dimensional protein–protein interaction network data and infer potential drug target genes using latent features and state-of-the-art machine learning techniques. The study inferred that tamoxifen, bosutinib, and dasatinib could serve as repositionable



candidate compounds against the disease. Rodriguez et al. (2021) proposed a machine learning framework, DRIAD (drug repositioning in AD), which quantifies potential associations between the pathological severity of AD and molecular mechanisms encoded in a list of gene names, and identified a ranked list of repositioning candidates for treating AD from 80 FDA-approved and clinically tested drugs.

Using AI methods for drug repurposing has become an important approach in AD drug research, providing important ideas and directions for new drug discovery. Although many studies have used neural networks to predict drug–target affinity, their application in the field of AD treatment is still relatively limited. This study aims to use neural networks to predict drug–target affinity and screen potential drugs for the treatment of AD, providing new ideas and choices for AD treatment. At the same time, we will compare and analyze different neural network models to find the best prediction model.

In this paper, we propose HNNDTA, a hybrid neural network for drug–target affinity prediction, thereby enabling drug repurposing for the treatment of AD. As shown in Figure 1, starting from the pathogenic target of AD, S1R, we conducted protein–protein interaction (PPI) analysis, screened out proteins related to S1R, and constructed a dataset based on inhibitors of S1R and related proteins. Subsequently, we used the HNNDTA model to train the dataset, combined with network pharmacology analysis to screen FDA-approved drugs, and obtained a batch of candidate drugs. Then, we use the molecular docking of candidate drugs with S1R and its related proteins to find potential effective lead

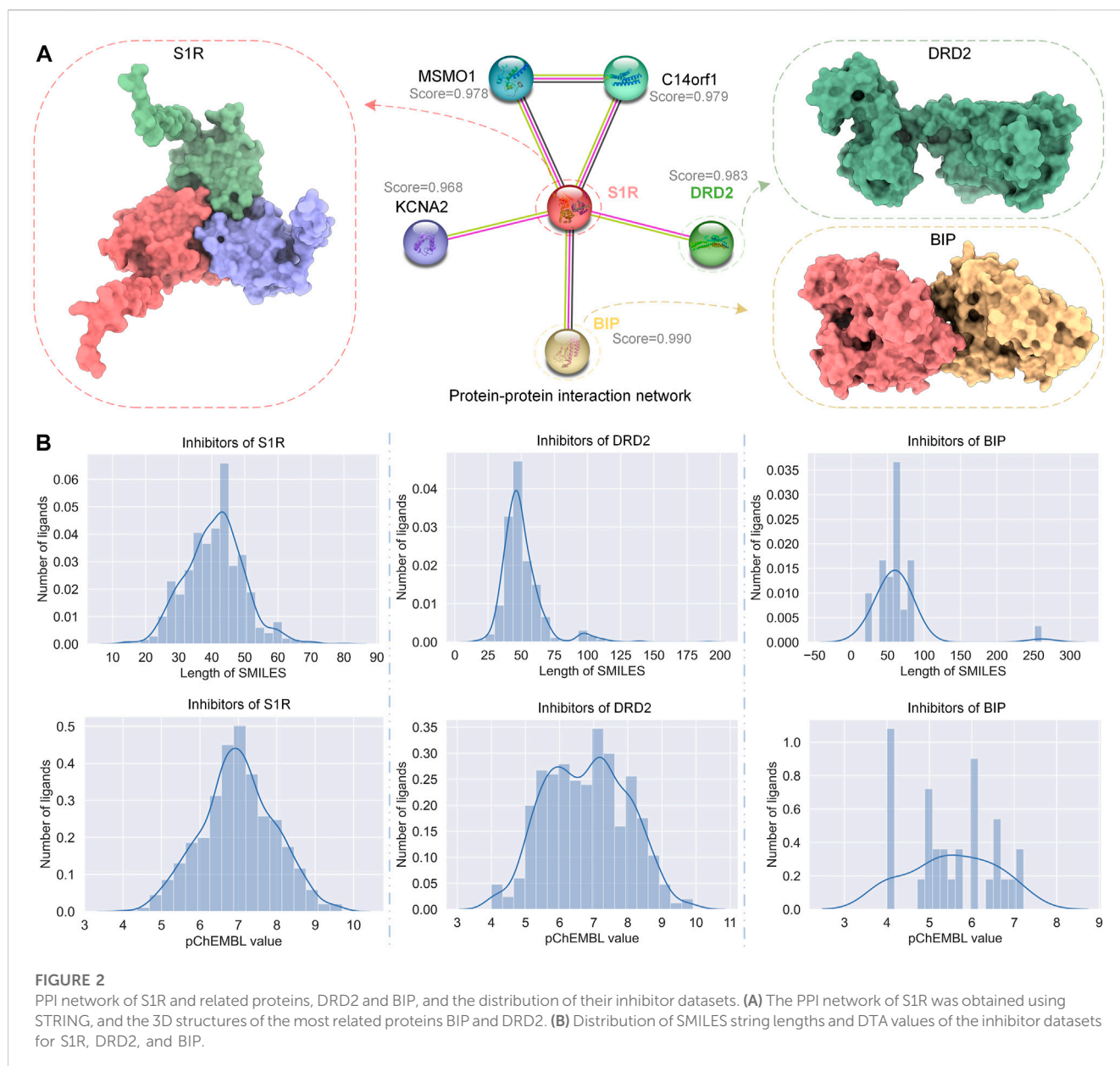
compounds, and predict their pharmacokinetics and toxicity to ensure the pharmacokinetics of these candidate drugs. The academic characteristics meet the requirements. Through this series of studies, we have obtained some lead compounds with potential therapeutic effects, which provide new ideas and options for the treatment of AD.

2 Materials and methods

2.1 Dataset

2.1.1 Target

STRING (Szklarczyk et al., 2023) is a database of known and predicted PPIs. We used STRING to get the PPI network of S1R, as shown in Figure 2A; we marked the correlation scores of proteins related to S1R in the network, among which the scores of dopamine D2 receptor (DRD2) and binding-immunoglobulin protein (BIP) are highest, 0.983 and 0.990, respectively, so we picked them as primary targets for network pharmacology analysis. We obtained the sequences of S1R (Q99720), DRD2 (P14416), and BIP (P11021) from the UniProt repository (Consortium, 2019). In addition, we obtained S1R (PDB ID: 5HK1) (Schmidt et al., 2016), DRD2 (6 PDB ID: LUQ) (Fan et al., 2020), and BIP (PDB ID: 3LDN) (Macias et al., 2011) from the RCSB Protein Data Bank (PDB) (Berman et al., 2000), which are 2.51 Å, 3.10 Å, and 2.20 Å, respectively, and their structures are shown in Figure 2A.



2.1.2 Inhibitors

The half-inhibitory concentration (IC_{50}) refers to the concentration of the drug or inhibitor required to inhibit half of the specified biological process, and the inhibition constant K_i reflects the inhibitory strength of the inhibitor on the target. The smaller the value, the stronger the inhibitory ability. pIC_{50} is the negative logarithm of the IC_{50} value, which is usually used to characterize the activity of the molecules in drug screening. The formula for converting IC_{50} values to pIC_{50} values is

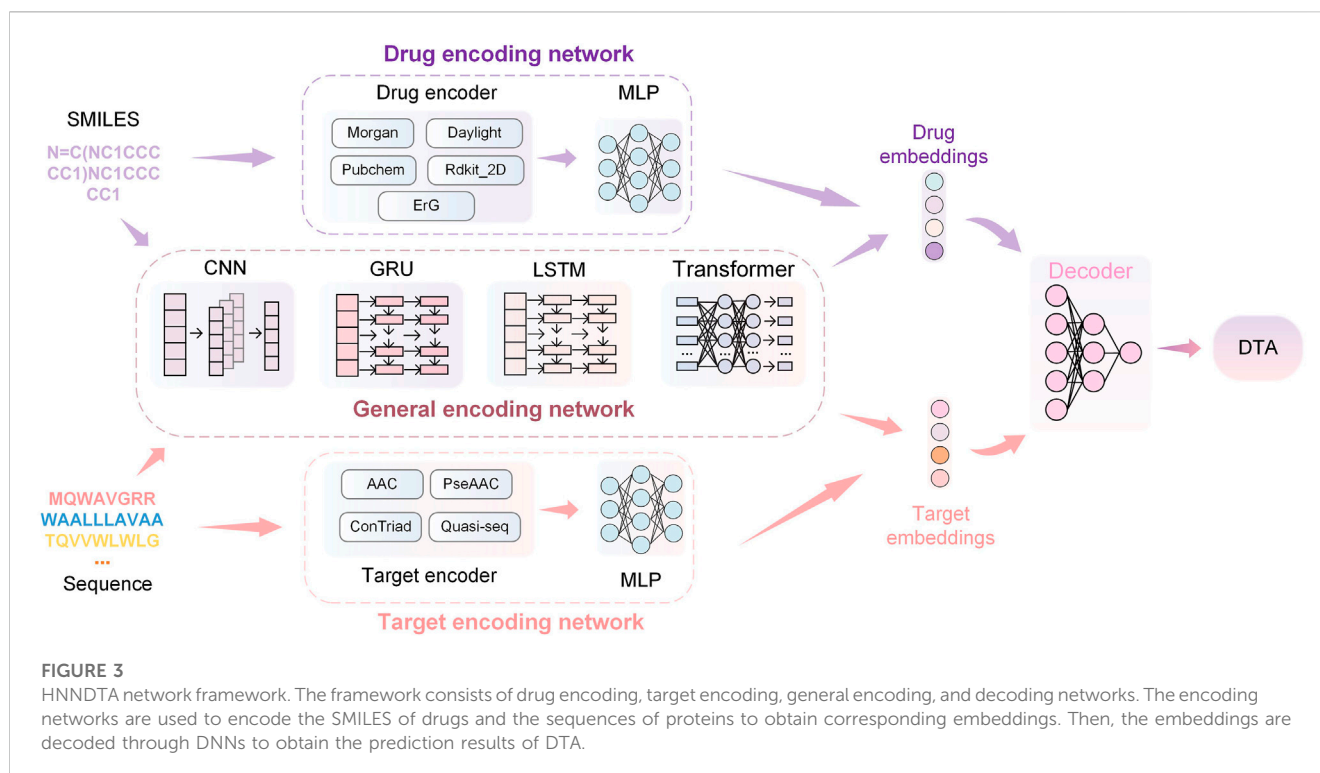
$$pIC_{50} = -\log_{10}(IC_{50}). \quad (1)$$

We obtained data on inhibitors of S1R, DRD2, and BIP and their binding abilities to their targets from the ChEMBL database (Gaulton et al., 2012). Although both IC_{50} and K_i can reflect the activity of the inhibitor, for data consistency, we screened the inhibitor data with IC_{50} as the subsequent drug–target affinity

(DTA) training data on the HNN. Similarly, under the premise of ensuring the number of datasets, we screened the data whose source description was scientific literature and excluded other data. Figure 2B shows the simplified molecular input line entry system (SMILES) length distribution and binding force distribution of the three protein inhibitors. The inhibitor distribution of S1R and DRD2 showed a Gaussian distribution trend, while the inhibitor distribution of BIP was relatively sparse.

2.1.3 Molecules for drug repurposing

The drug screening library used in this study comes from FDA-approved drugs in the DrugBank database (Wishart et al., 2008). DrugBank is a comprehensive pharmaceutical knowledge bank that provides pharmacists, pharmacologists, health professionals, and drug researchers with free academic resources to help advance drug development and clinical practice. We chose DrugBank as the



screening bank because it contains extensive drug information and a list of FDA-approved drugs, which can be used to screen potential drugs for the treatment of AD. These drugs have been proven to be safe and effective treatments in human clinical trials, so they are expected to be used in the treatment of AD. We selected FDA-approved drugs in the DrugBank database as screening libraries, and a total of 2509 drug molecules were available for drug repurposing studies.

2.2 HNNDTA

2.2.1 Overview of the framework

The overview of the HNNDTA framework proposed in this study is shown in Figure 3. First, we used a network pharmacology approach to find other targets in the same pathway as the AD target S1R, namely, DRD2 and BIP. We searched the ChEMBL website for inhibitor data for these three targets. The target protein is encoded as a one-dimensional target embedding, and the drug molecule is encoded as a one-dimensional drug embedding. The two encoding vectors are spliced in zero dimension, and after the calculation of the deep neural network (DNN), the final DTA is obtained, which can be expressed as follows:

$$DTA = DNN[cat(v_p, v_d)], \quad (2)$$

where the function $cat(a, b)$ represents the splicing operation of the 1D a and b vectors, and v_p and v_d represent the encoding vectors of the target protein and the drug molecule, respectively. In this paper, there are 13 kinds of target encoders and 9 kinds of drug encoders, all of which are built by DeepPurpose (Huang et al., 2020). A suitable

combined model will produce better prediction accuracy. During the training phase, the dataset was randomly divided into independent training, validation, and test sets in a ratio of 7:1:2. The training set was used to train the model, while the validation and test sets were used to evaluate its performance. Due to the nature of our HNNDTA framework, which was trained on datasets specific to individual targets, it exhibits higher predictive accuracy for single targets. We have observed that models trained on single targets exhibit higher accuracy than those trained on mixed-target datasets.

2.2.2 Drug encoding network

The drug encoder receives SMILES sequences as input. The Morgan encoder first uses the ECFP (Rogers and Hahn, 2010) algorithm to generate the feature representation sequence of the circular substructure of the drug, with a length of 1,024 bits. A multi-layer perceptron (MLP) then processes the sequence of feature representations to obtain a vector representation that can be fed into a neural network. The Morgan encoder is expressed as follows:

$$f_{\text{morg}}(\text{SMILES}) = \text{MLP}(\text{ECFP}(\text{SMILES})). \quad (3)$$

Similar to the Morgan encoder, the daylight encoder also uses the ECFP algorithm to generate a feature sequence based on the channel substructure of the drug, which is used as the input of the multi-layer perceptron to generate a feature sequence with a length of 2048 bits. The daylight encoder is represented as follows:

$$f_{\text{day}}(\text{SMILES}) = \text{MLP}(\text{ECFP}(\text{SMILES})). \quad (4)$$

The PubChem encoder (Kim et al., 2019) generates feature sequences using handcrafted important substructures and then generates a feature sequence with a length of 881 bits through a

multi-layer perceptron. The PubChem encoder is represented as follows:

$$f_{pub}(\text{SMILES}) = \text{MLP}(\text{Substructure}(\text{SMILES})). \quad (5)$$

The rdkit_2d_normalize encoder (Reczko and Bohr, 1994) generates a feature sequence with a length of 200 bits according to the global pharmacophore of the drug and then normalizes the feature sequence by fitting the cumulative density function of a given molecule sample. The rdkit_2d_normalize encoder is represented as follows:

$$\begin{aligned} f_{rdkit}(\text{SMILES}) &= \text{MLP}(\text{Normalize}(\text{Feature})) \\ \text{Feature} &= \text{GlobalPharmacophore}(\text{SMILES}). \end{aligned} \quad (6)$$

The extended reeb graph (ErG) method (Stiefl et al., 2006) mixes the simplified graph and the binding attribute pair to generate a feature sequence and uses the node description of the drug carrier type to encode the relevant molecular properties; the encoded features are obtained after the MLP calculation vector. The ErG coder is expressed as follows:

$$\begin{aligned} f_{erg}(\text{SMILES}) &= \text{MLP}(\text{Graph}(\text{Feature})) \\ \text{Feature} &= \text{Scaffold} - \text{BasedNodeDescriptor}(\text{SMILES}). \end{aligned} \quad (7)$$

MLP obtains the output value through feedforward propagation and updates the model parameters through reverse transmission so that the model output value gradually approaches the real value. The output of the MLP forward propagation is expressed as follows:

$$y = AC \left(\sum_{i=1}^{M_l} \omega_{il} \bullet AC \left(\sum_{i=1}^{M_{l-1}} \omega_{i,l-1} \bullet (\dots) \right) \right), \quad (8)$$

where AC is the activation function and the typical activation function is the modified linear unit ReLU; M_l is the number of neurons in the l th layer network, ω_{il} is the weight of the i th neuron in the l th layer network; and the termination condition of (\dots) in the aforementioned formula is the first layer of the neural network, that is, the input layer. The reverse transfer uses the Adam optimizer to update the model weights. The underlying algorithm is the gradient descent method. The update on the weight of $\omega_{i,l}$ can be expressed as follows:

$$\omega_{i,l} \leftarrow \omega_{i,l} - \eta \frac{\partial E}{\partial \omega_{i,l}}, \quad (9)$$

where E is the difference between the predicted value and the real value, $\frac{\partial E}{\partial \omega_{i,l}}$ is the partial derivative of E to $\omega_{i,l}$, and η is the learning rate.

2.2.3 Target encoding network

The input to a target encoder is the amino acid sequence of the target. The signature sequence generated by the amino acid composition (AAC) coder is 8420 positions in length, where each position is consistent with the maximum length of overlapping subsequences (k-mers) of one amino acid. The amino acid composition coder is expressed as follows:

$$\text{AAC}_i = \frac{f_i}{L}, \quad i = 1, 2, \dots, 20, \quad (10)$$

where f_i represents the number of occurrences of amino acid i in the protein and L represents the length of the amino acid sequence. The

AAC encoder concatenates 20 AAC values for each position in the amino acid sequence to obtain a signature sequence of 8420 elements in length.

The pseudo amino acid composition (PseAAC) encoder adds the hydrophobic and hydrophilic pattern information on the protein based on AAC to generate a 30-bit feature vector representation. The pseudo-amino acid composition encoder is expressed as follows:

$$\text{PseAAC}_{i,j} = \frac{\sum_{k=1}^L f_{k,i} w_{k,j}}{\sum_{k=1}^L f_{k,i}}, \quad (11)$$

$$i = 1, 2, \dots, 20; \quad j = 1, 2, \dots, 30,$$

where $f_{k,i}$ represents the frequency of amino acid i in the k th position in the protein sequence and $w_{k,j}$ represents the weight of the pattern of the k th amino acid and the relative position j . The PseAAC encoder concatenates 30 PseAAC values for each position in the amino acid sequence, resulting in a feature vector of 30 elements in length.

The conjoint triad (ConTriad) encoder (Shen et al., 2007) forms a 7-letter alphabet based on amino acid triplet features, generating a feature vector with a length of 343 elements. The ConTriad encoder is expressed as follows:

$$\text{ConTriad}_i = \frac{\sum_{j=1}^7 f_{j,i} w_j}{L-2}, \quad i = 1, 2, \dots, 3430, \quad (12)$$

where $f_{j,i}$ indicates that the three adjacent amino acids in the protein sequence are converted into a number according to the 7-letter alphabet, the i th element indicates the frequency of the j th triplet appearing in the protein sequence, and w_j is the weight of the j th triplet. The ConTriad encoder concatenates 343 ConTriad values for each position in the amino acid sequence, resulting in a feature vector of 343 elements in length.

The quasi-sequential encoder consists of a 100-element feature vector of quasi-sequential descriptors (Chou, 2000). The feature vectors generated by the aforementioned manual feature encoder will be further processed as input to MLP to obtain the feature vector of the target. The quasi-sequential encoder is expressed as follows:

$$\text{QuasiSeq}_i = \sum \frac{(j-1)^N \rho_j}{d_{ij}}, \quad i = 1, 2, \dots, 100, \quad (13)$$

where ρ_j represents the weight of the j th quasi-sequential descriptor and d_{ij} is the distance between the i th amino acid and the j th sequence descriptor. The quasi-sequential encoder concatenates 100 QuasiSeq values for each position in the amino acid sequence, resulting in a feature vector of 100 elements in length.

2.2.4 General encoding network

The aforementioned drug and target feature extraction methods are based on prior chemical knowledge and manual transformation, so these encoders cannot be mixed. The encoders introduced in this section are general-purpose encoders based on DNNs, including convolutional neural networks (CNNs) (Krizhevsky et al., 2017), gated recurrent units (GRUs) (Chung et al., 2014), long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), and transformers (Vaswani et al., 2017). These neural networks treat amino acid sequences as one-dimensional data.

The CNN encoder is a multilayer 1D CNN (Krizhevsky et al., 2017). After encoding the amino acid sequence character by character, the obtained deep feature vector will pass through multiple 1D convolutional layers and finally pass through the one-dimensional maximum pooling layer to obtain the output of the target feature vector. The output of the 1D convolutional layer is the result of convolving the input with the convolution kernel, which can be expressed as follows:

$$out = input \otimes kernel, \quad (14)$$

where \otimes represents a convolution operation. Assuming that the convolution kernel size is $2k + 1$, $k \in \mathbb{N}^+$, the i th convolution output can be expressed as follows:

$$out_i = \sum_{j=i-k}^{i+k} \sum_{a=1}^{2k+1} input_j \bullet kernel_a. \quad (15)$$

GRU and LSTM encoders are types of recurrent neural networks. In both networks, each node will get an output based on the state at the last moment and the current input and update the state of the node. This can solve the problem of traditional convolutional networks without long-term memory to a certain extent. Specifically, the SMILES sequence or amino acid sequence will first pass through the CNN for feature extraction and then use the output of the CNN as the input of the recurrent network.

The transformer encoder applies a self-attention mechanism (Vaswani et al., 2017). Due to the computational time and memory cost of the transformer, amino acid sequences are decomposed into moderately sized protein substructures, such as motifs, and each segmentation is then treated as a token and fed into a self-attention-based encoder. If a SMILES sequence or amino acid sequence is treated as a sentence, cut into several meaningful phrases, and encoded into several vectors with the same number of phrases, denoted as x , then the output of the transformer can be expressed as

$$\begin{aligned} x_1 &= \text{norm}(x + \text{attn}(x, \text{mask})), \\ out &= \text{norm}(x_1 + \text{feedforward}(x_1)), \end{aligned} \quad (16)$$

where attn is a self-attention function, mask is a Boolean value about whether the input x is eliminated, feedforward is a feedforward neural network, and norm is a layer normalization operation.

2.2.5 Evaluation metrics

In mathematical statistics, mean-squared error (MSE) is a method used to measure the difference between the predicted and real values. It calculates the mean of the squared difference between predicted and true values, which is the expected value of the squared difference between predicted and true values. The smaller the value of the MSE, the higher the prediction accuracy of the prediction model. Assuming there are n samples, MSE can be expressed by the following formula:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (17)$$

where y_i and \hat{y}_i are the true and estimated values of the i th sample, respectively. In this paper, the MSE is used to evaluate the accuracy of the model to predict the binding affinity of the drug to the target.

Harrell's C-index (also known as the concordance index, CI) is a widely used metric for evaluating the performance of risk models. It is commonly employed in survival analysis, especially when dealing with censored data (Harrell et al., 1982). The C-index measures the degree of concordance between predicted and observed rankings of survival times. It serves as an indicator of the model's accuracy with values closer to 1, indicating a higher level of consistency between the predicted outcomes and the actual observed outcomes.

Suppose the data are represented by vectors $(\tilde{T}_i, \Delta_i, X_{i1}, \dots, X_{ip})$, $i = 1, \dots, n$, where \tilde{T}_i is a possibly right-censored continuous survival time and $(X_{i1}, \dots, X_{ip})^T$ is a vector of predictor variables. It is assumed that \tilde{T}_i is the minimum of the true survival time T_i and an independent continuous censoring time C_i . The variable $\Delta_i = I(T_i \leq C_i)$ indicates whether T_i has been fully observed ($\Delta_i = 1$) or not ($\Delta_i = 0$). A one-dimensional score $\eta_i \in \mathbb{R}$ is estimated for each observation $i = 1, \dots, n$, by averaging the cumulative hazard estimates over all trees and all time points. The concordance index is given by

$$C = \frac{\sum_{i,j} I(\tilde{T}_i > \tilde{T}_j) \cdot I(\eta_i > \eta_j) \cdot \Delta_j}{\sum_{i,j} I(\tilde{T}_i > \tilde{T}_j) \cdot \Delta_j}, \quad (18)$$

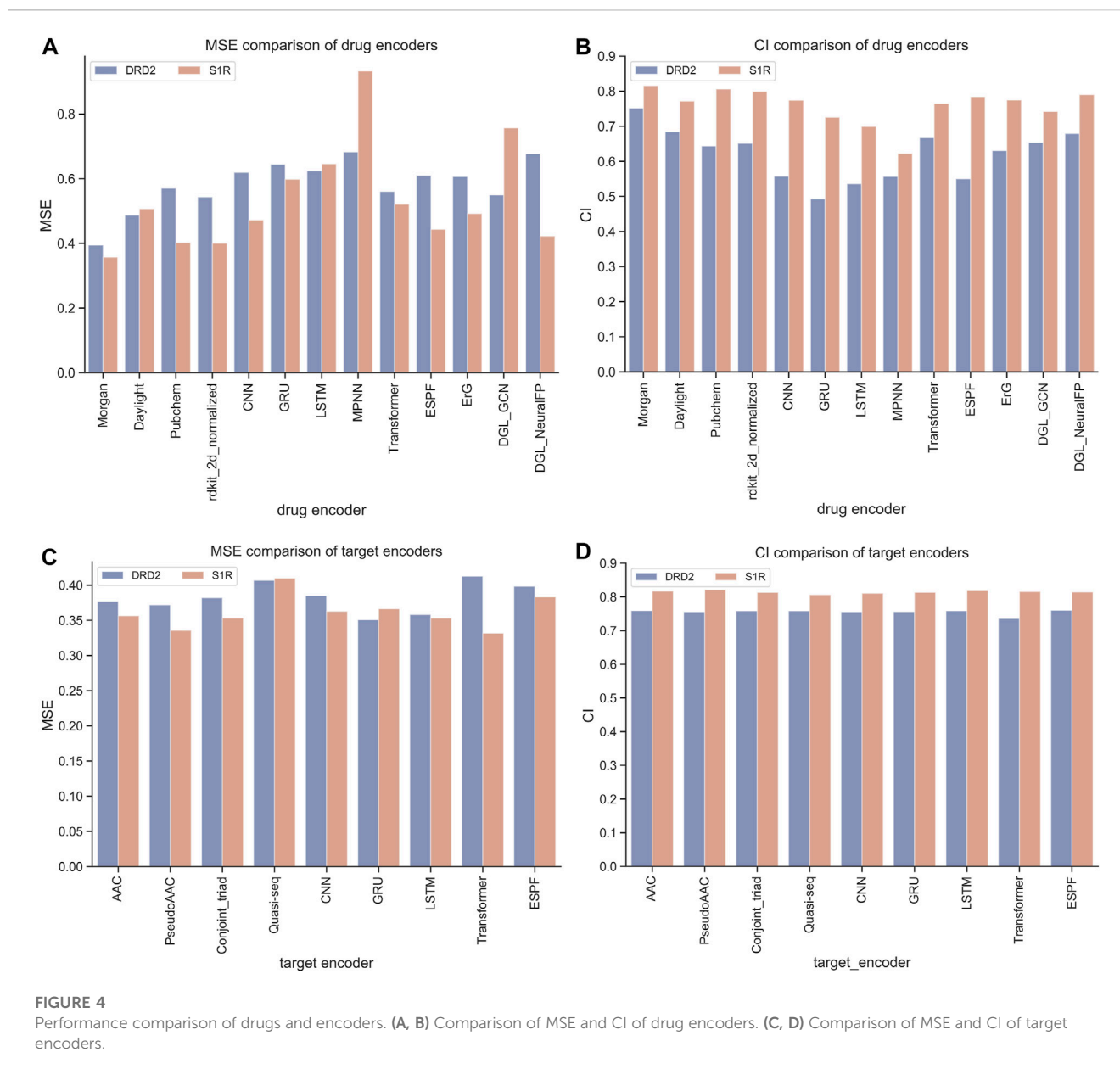
where the indices i and j refer to pairs of observations in the sample (Schmid et al., 2016).

2.3 Network pharmacology

Network pharmacology (NP) (Hopkins, 2008) is a new drug development method based on systems biology. It reveals the multi-target action mechanism of drugs by integrating protein interaction and drug compound networks. To construct a network pharmacology-based analysis, we mapped protein-protein and protein-drug interaction networks (Hasan et al., 2020). We fetch the PPI network from the STRING database and select the protein most related to the target we need to study. We then used the HNNDTA model to predict the binding forces between these proteins and compounds. To identify the best compounds, we picked the top 20 most binding proteins for each protein and mapped them into a protein-compound network. We use Sankey diagrams (Lee et al., 2019) to visualize drug-protein interaction networks to better understand and analyze the mechanism of action of drugs in biological systems. In this network, we can identify which compounds may be the most promising drug candidates by analyzing the interactions between proteins and compounds. In particular, for those compounds that bind strongly to multiple proteins, we can select them as our drug candidates.

2.4 Molecular docking

In molecular docking tasks, AutoDock Vina (Eberhardt et al., 2021) is one of the widely used docking engines in AutoDock Suite, and its open-source code and fast docking speed are favored. We use AutoDock Vina 1.1.2 for molecular docking experiments. First, we obtained the 3D molecular structure files of all receptor molecules



and processed them to remove crystal water and hydrogenate them to generate preprocessed receptors. Next, we first removed the crystal water and the original ligand, then added hydrogen and charge distribution, and manually set the active site area of the receptor as the grid box according to the feature information on the protein in UniProt. For the ligand file, we obtained its structure from PubChem (Kim et al., 2019) and then performed hydrogenation and charge addition to obtain the preprocessed ligand file. Then, we used AutoDock Vina for docking; exhaustiveness is set to 32; a total of 10 docking poses are generated; the top 5 best poses are kept, and finally, the binding energy value (in k/mol) of the best pose is used as the docking score. The results of molecular docking were output in pdbqt format and visualized and analyzed using PyMOL molecular visualization software. The docking results are evaluated by factors such as hydrogen bonds, van der Waals forces, and electron static energy.

3 Results

3.1 Performance evaluation

The HNNDTA framework was constructed using 13 drug encoders and 9 target encoders. We fixed the target encoder as AAC and constructed 13 different drug encoder models. The MSE and CI of the test models are shown in Figures 4A, B. The orange column is the test result of the ligand dataset of the S1R protein, and the sky blue column is the test result of the ligand dataset of the DRD2 protein. The smaller the MSE and the larger the CI, the smaller the difference between the predicted results of the model and the real results, and the higher the accuracy of the model. In the figure, the MSE value of the Morgan encoder is the smallest and the CI value is the largest, indicating that the Morgan encoder will make the model perform better, and the Morgan encoder should be

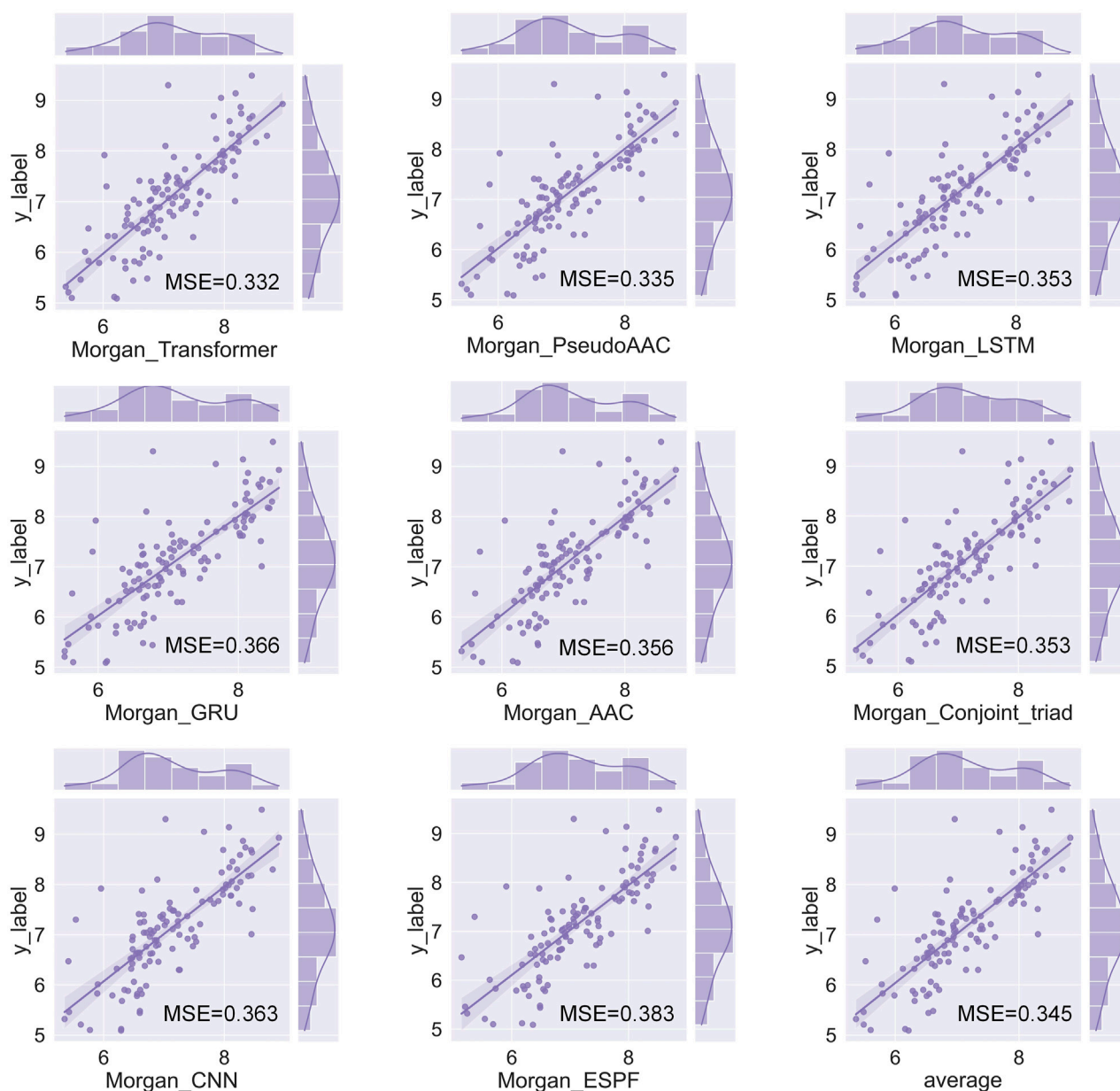


FIGURE 5
Fit plot of the best-performing model.

considered in the subsequent grid search for the best drug encoder–target encoder combination. We fixed the drug encoder as Morgan, constructed nine models with different target encoders, and compared the MSE and CI of the test models, as shown in Figures 4C, D. The MSE values of each encoder are basically at the same level because there is only a very small amount of target data in the dataset, and the difference in information provided by the target is less.

We have a total of 117 models of 13 drug encoders and 9 target encoders, and conduct a grid search on the ligand datasets of the three targets of S1R, DRD2, and BIP to find the best models. After testing, there are eight models with both MSE and CI in the top 10, as shown in Figure 5; Table 1. Among them, the Morgan encoder has

the best encoding effect on drugs, and the transformer and PseudoAAC encoders have better encoding effects on protein targets. Overall, the performance of these eight models is comparable and complements each other. In the next step of screening candidate drugs, the average of the votes predicted by the eight models is taken as the drug–target interaction score.

3.2 Virtual screening of HNNDTA and network pharmacology

In this study, 2506 FDA-approved drugs were used as drug candidates for the AD target protein S1R and other targets of the

TABLE 1 Best-performing model.

Drug encoder	Target encoder	MSE	CI
Morgan	Transformer	0.332	0.816
Morgan	PseudoAAC	0.335	0.822
Morgan	LSTM	0.353	0.818
Morgan	Conjoint_triad	0.353	0.813
Morgan	AAC	0.356	0.817
Morgan	CNN	0.363	0.811
Morgan	GRU	0.366	0.814
Morgan	ESPF	0.383	0.815

same pathway, DRD2 and BIP. For S1R and DRD2, the respective models were trained using ligand datasets obtained from the ChEMBL website. For BIP, due to the lack of ligand data on BIP on the ChEMBL website, it is not enough to train a good model. We can pre-train the model with a large amount of ligand data for the same pathway target of S1R and then fine-tune the model with the ligand data on BIP itself.

The HNNDTA model was used to predict the activities of FDA-approved drugs and targets S1R, DRD2, and BIP, and the 20 drugs

with the highest binding activities to these three targets are shown in **Figure 6A**. On the left side of the Sankey diagram are the three target proteins, and on the right side are the 20 drugs with the highest binding activity to these three targets. At the intersection, there are a total of 40 drugs. The prediction results show that most drugs can only have high activity with one or two targets, while the seven drugs DB13928, DB06287, DB00626, DB09265, DB00502, DB12401, and DB01369 have high binding activity with three targets, indicating that they can simultaneously inhibit these three AD-related targets. Therefore, these seven drugs can be used as alternative drugs for the treatment of AD. The DTA values of the aforementioned seven candidate drugs and S1R, DRD2, and BIP are shown in **Figure 6B**. The DTA values of these seven drugs and three targets stand out among more than 2,000 FDA drugs. The two drugs, DB00502 and DB12401, have the highest combined affinity for the three targets and are expected to become candidate drugs for the treatment of AD.

3.3 Benchmark testing

To assess the accuracy of the model predictions and validate the efficacy of the drugs identified through network pharmacology (i.e., haloperidol and bromperidol), benchmark testing was conducted. Known high-affinity ligands for S1R, DRD2, and BIP

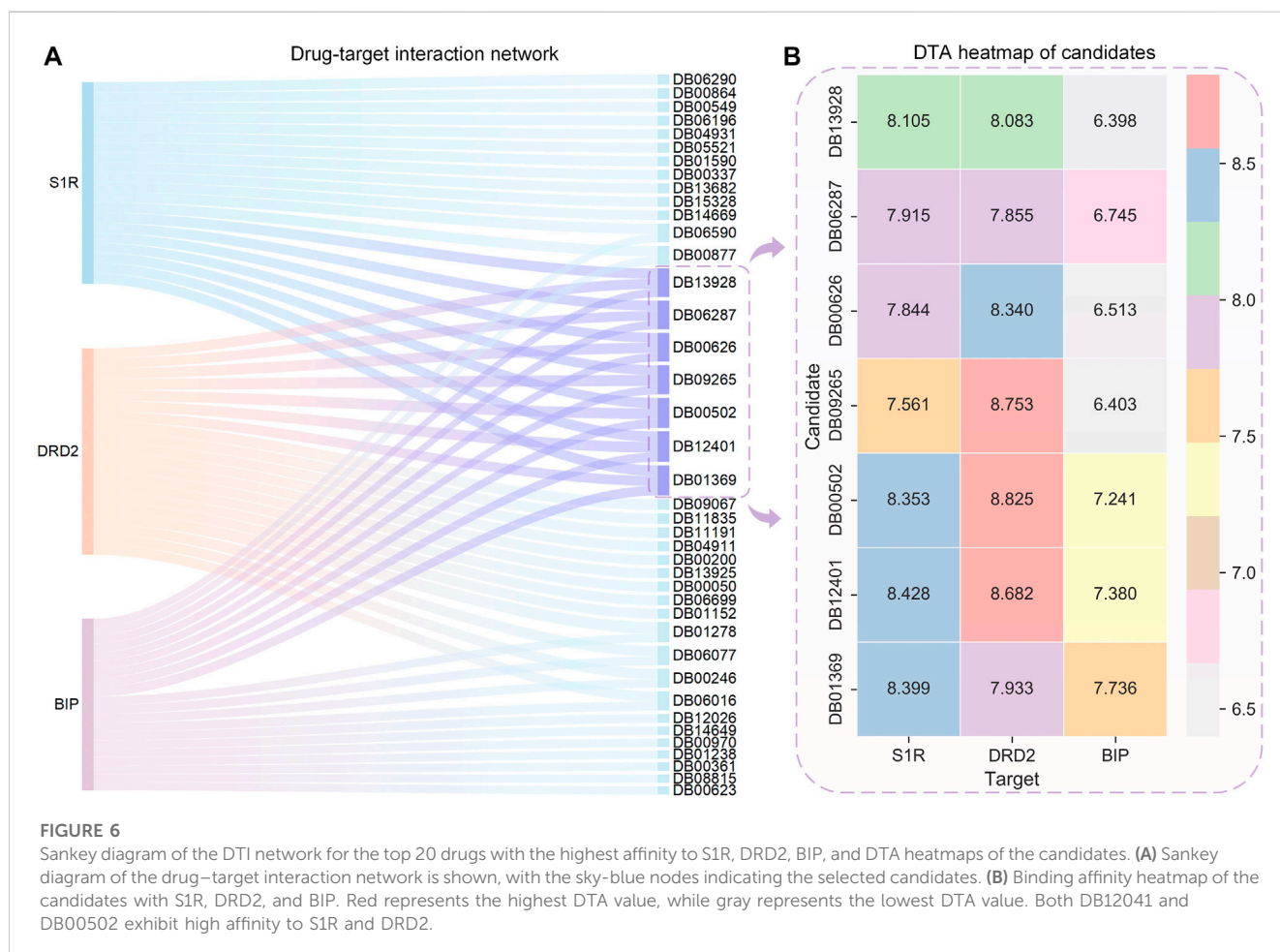


TABLE 2 Collected known drug–target pairs with high and low binding affinities.

Target	Inhibitor	Affinity	Type
S1R	Haloperidol	8.54	Ki
S1R	Donepezil	7.84	Ki
S1R	Fluvoxamine	7.44	Ki
S1R	Corticosterone	4.45	Ki
S1R	Cocaine	5.05	Ki
DRD2	Haloperidol	8.76	Ki
DRD2	Pimozide	7.93	Ki
DRD2	Amisulpride	7.90	Ki
DRD2	Procaterol	4.07	Ki
DRD2	Isoproterenol	4.32	Ki
BIP	CHEMBL462871	7.22	Kd
BIP	CHEMBL516197	4.85	Kd

were collected from the ChEMBL and BindingDB databases for validating the docked scores of the screened drugs as being higher than or comparable to the known high-affinity ligands. Conversely, known low-affinity ligands were gathered to demonstrate that the docked scores of the screened drugs are superior to them. The information on known ligands and their affinities is presented in Table 2.

First, the HNNDTA model was utilized to predict the binding affinities (pIC50) of the collected ligands to the three targets. The prediction results are shown in Table 3, where the green boxes and red boxes represent known high- and low-affinity drug–target pairs, respectively. Overall, the predicted affinities in the green boxes are higher than those in the red boxes, indicating that our model can accurately differentiate between high and low affinities among drug–target pairs. Subsequently, blind docking of ligand–protein was performed using QuickVina-W software (Hassan et al., 2017), and the docking scores are presented in Table 4. Lower docking scores indicate smaller binding energies and higher binding affinity. The docking scores in the green boxes are generally lower than those in the red boxes, suggesting the effectiveness of the docking procedure.

Our screened drugs, haloperidol and bromperidol, exhibit lower overall docking scores with the three targets compared to most other drugs. Furthermore, the docking scores of the screened drugs are comparable to those of known high-affinity ligands and significantly lower than those of known low-affinity ligands. This indicates that the HNNDTA model successfully identified high-affinity drugs suitable for multiple targets. It is worth noting that Table 4 shows that the drug pimozide has the best multi-target docking score. However, molecular docking requires manual preprocessing of 3D structures and is computationally time-consuming, making it difficult to apply to high-throughput drug target screening in network pharmacology. The HNNDTA model can expedite this process and successfully screen multi-target high-affinity drugs, even if it may represent a suboptimal solution.

TABLE 3 DTA predicting results obtained from HNNDTA are presented. The green boxes and red boxes represent known high-affinity and low-affinity drug–target pairs, respectively.

Ligand targets															
	DB00502			DB12401			DB00843			DB00176			DB04652		
	DB00907			DB01100			DB06288			DB01366			DB01064		
	CHEMBL516197			CHEMBL462871			DB15477								
S1R	7.5	7.3	6.8	7.3	7.9	6.7	6.3	6.8	6.3	6.3	6.3	6.3	5.6	6.8	5.4
DRD2	7.8	7.9	5.9	6.7	6.7	6.3	6.4	8.5	5.4	5.2	5.9	5.2	4.4	5.9	4.4
BIP	6.8	6.7	5.3	6.7	6.7	5.0	5.0	7.3	4.4	4.5	4.4	4.5	4.4	5.4	4.4

TABLE 4 The molecular docking results obtained from QuickVina-W are presented, where a lower docking score indicates weaker binding energy and stronger binding affinity. The green boxes and red boxes represent known high-affinity and low-affinity drug-target pairs, respectively.

Ligand targets	DB00502	DB12401	DB00843	DB00176	DB04652	DB00907	DB01100	DB06288	DB01366	DB01064	CHEMBL462871	CHEMBL516197	DB15477
S1R	-10.3	-10.3	-10.5	-8.3	-6.5	-8.7	-11.3	-5.4	-5.7	-6.8	-8.1	-7.5	-6.7
DRD2	-11.3	-10.7	-9.5	-7.5	-8.3	-7.3	-10.9	-7.9	-7.0	-7.2	-10.7	-6.9	-8.5
BIP	-8.4	-8.6	-8.4	-7.3	-7.4	-7.7	-10.1	-7.6	-6.6	-6.9	-9.2	-7.6	-8.3

3.4 Virtual screening of molecular docking

Small molecules have smaller molecular weights, which favor better pharmacokinetics and less toxicity. The molecular weight of antibiotics is large, and the metabolic process affects the drug's efficacy. Small molecules have good medicinal properties, such as high bioavailability, good tissue specificity, and low toxicity and side effects, and are suitable for drug research and development. Therefore, we only choose small molecules with a weight of less than 500 as lead compounds. In AutoDock Vina docking, we use the binding energy value of the best pose as the docking score and tabulate the results in Table 5. The molecules of DB09265 and DB13928 are very large, beyond the active site region of the receptor, causing errors in Vina, which indicates that the binding between the two ligands and the receptor is difficult. Since Vina uses binding energy as a docking score, a smaller score indicates tighter binding between the two molecules, which generally indicates better docking. However, when the score is positive, it means that docking is difficult to produce. Both of these conditions can indicate a docking failure. Table 5 shows that although the ligands DB01369 and DB06287 have good docking effects on DRD2 and BIP receptors, they are difficult to bind to S1R receptors. Ligands DB00502 (bromperidol) and DB12401 (haloperidol) have good binding abilities to the three receptors, and the molecular weight is less than 500, meeting the screening requirements, so they may become potential drugs for AD.

3.5 Explanatory analysis of DTA

Figure 7 shows the 2D chemical structures of haloperidol and bromperidol, and the 2D poses resulting from docking with the target S1R. As shown in Figure 7A, their chemical structures are very similar, differing only by one halogen atom: haloperidol with a Cl atom and bromperidol with a Br atom. They are both high-affinity ligands for S1R, with only slight differences. This may be caused by the different interaction distances between the halogen atoms in the bromperidol molecule and the six amino acids of S1R. As shown in Figure 7B, both drugs produced hydrogen bonds with SER34, SER99, and LEU100 amino acids of S1R, and produced $\pi - \pi$ interactions with TRP29, HIS72, LEU214, and TYR217 amino acids of S1R. The two molecules stabilize the association between them through their interaction with S1R.

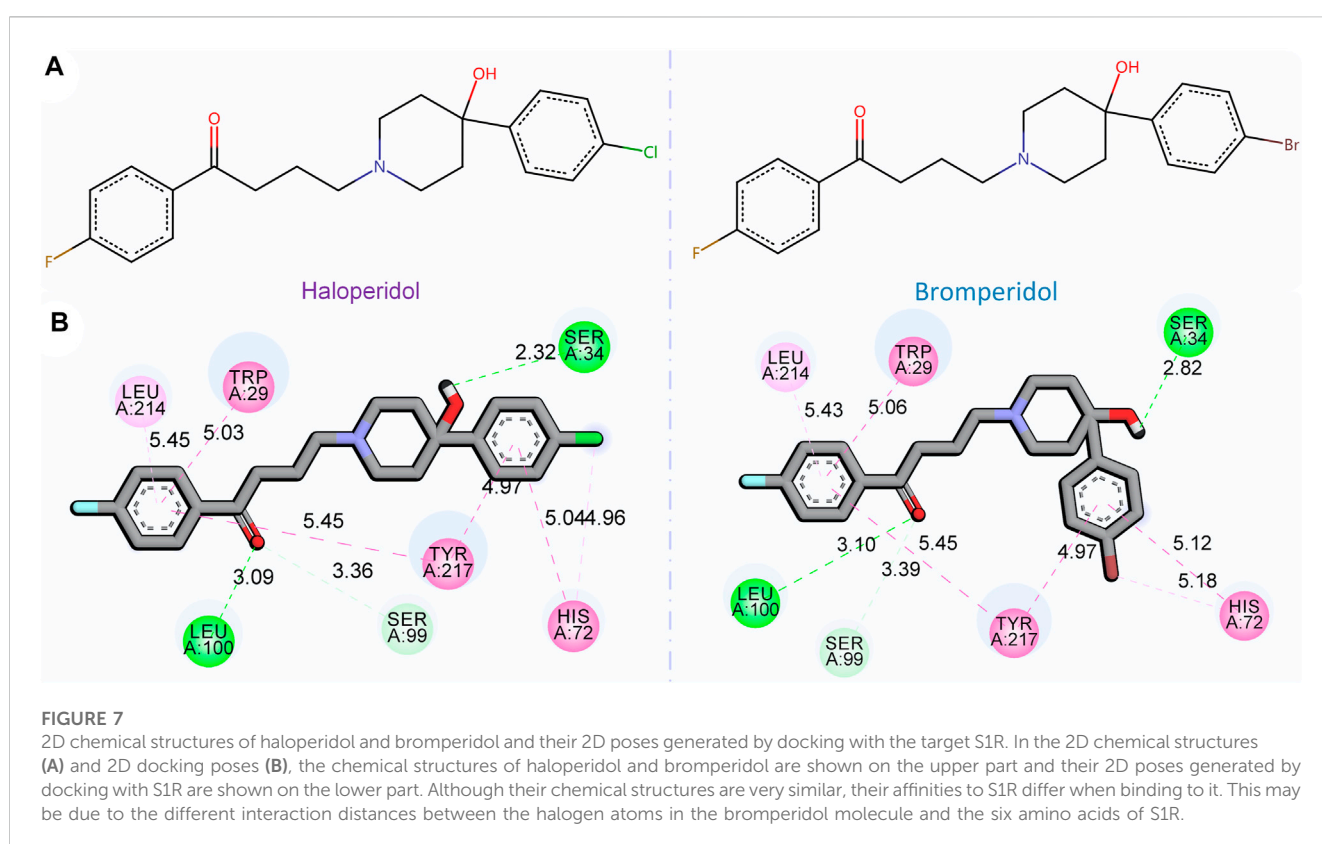
In order to further observe the docking poses of haloperidol and bromperidol with S1R, we also plotted the 3D docking simulation results, as shown in Figure 8. Both haloperidol and bromperidol dock at the S1R surface and interact with surrounding S1R amino acids. As shown in Figures 8A, B, the docking poses of haloperidol and bromperidol are very close to S1R, which is related to their similar chemical structures. They jointly participate in the stable combination with S1R and produce more interactions.

To evaluate the ADMET, of haloperidol and bromperidol, we evaluated them using the ADMETlab 2.0 tool (Xiong et al., 2021), as shown in Figure 9. The evaluation results of haloperidol and bromperidol are roughly similar, except for logD and logP, and their compound properties are distributed between the upper and lower limits. This shows that haloperidol and bromperidol have better pharmacokinetic conditions and almost no toxicity. Haloperidol is an antipsychotic drug used to treat schizophrenia and other psychotic disorders, as well as symptoms of agitation, irritability, and delirium. Bromperidol

TABLE 5 Overview of candidate compounds and their docking scores with S1R, DRD2, and BIP proteins. The docking scores were calculated using the molecular docking software application AutoDock Vina, with higher scores indicating stronger interactions.

DrugBank ID	Generic name	Summary	Docking score		
			S1R	DRD2	BIP
DB00502	Haloperidol	Antipsychotic	-8.856	-7.265	-8.585
DB00626	Bacitracin	Antibiotic	-7.979	-6.412	-6.35
DB01369	Quinupristin	Antibiotic	-	-9.455	-9.689
DB06287	Temsirolimus	Antineoplastic	-	-9.674	-9.858
DB09265	Lixisenatide	GLP-1 receptor agonist	-	-	-
DB12401	Bromperidol	Antipsychotic	-8.516	-7.031	-8.245
DB13928	Semaglutide	Peptide 1 receptor agonist	-	-	-

The bold values indicate the docking scores of the top two drugs with the highest docking scores for a specific target.

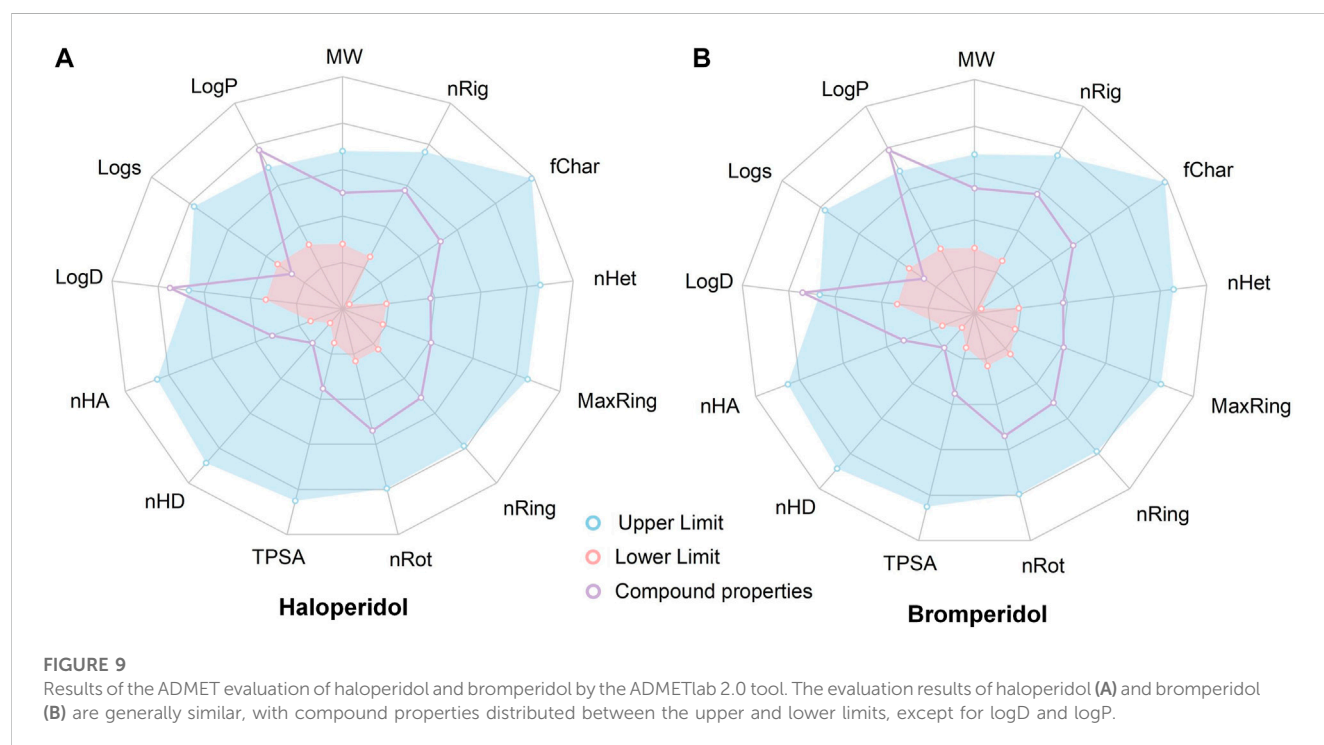
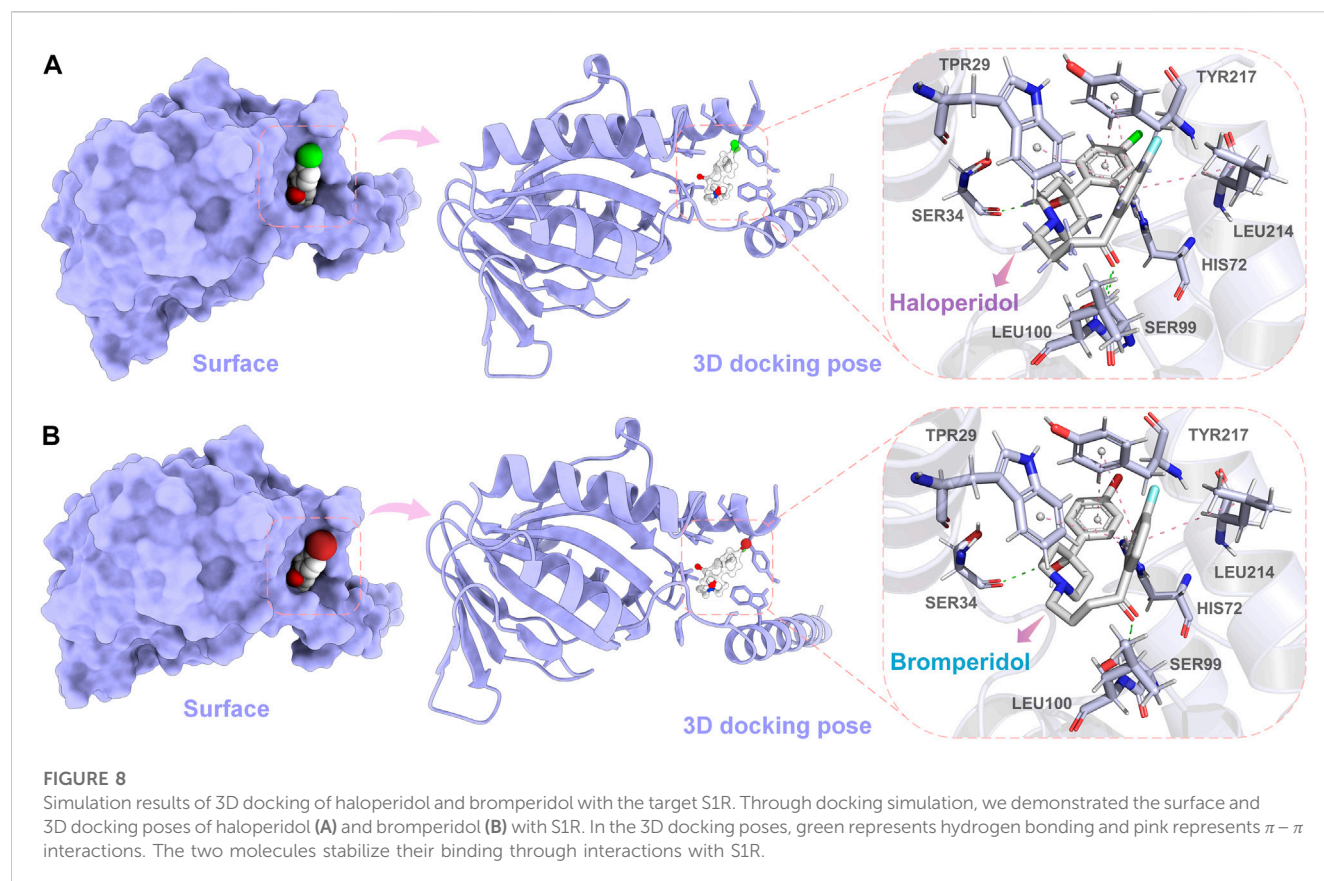


is used to treat schizophrenia and other psychotic symptoms and has been used in trials investigating the treatment of dementia, depression, schizophrenia, anxiety disorders, and psychosomatic disorders, among others. It further illustrates the accuracy of our HNNDTA screening by finding a trial that is already in the treatment of AD and, at the same time, screening a new potential drug for the treatment of AD.

4 Discussion

Alzheimer's disease is a significant age-related illness that has garnered widespread attention in society. In this article, we propose a drug-screening framework that combines network pharmacology

and hybrid neural networks to discover potential drugs for treating Alzheimer's disease. Existing evidence supports S1R as a potential therapeutic target for Alzheimer's disease. Initially, we conducted protein-protein interaction analysis using the STRING database to identify the most relevant targets associated with S1R, including DRD2 and BIP. These targets were then utilized in network pharmacology for drug screening. We developed a hybrid neural network framework to predict the binding affinity between targets and ligands, enabling the prediction of multi-target interactions for drug candidates. Benchmark testing was performed using a collection of known ligands with high and low affinity, demonstrating our model's ability to differentiate between high- and low-affinity ligands. Furthermore, our model identified two



drugs, haloperidol and bromperidol, with overall higher docking scores than other drugs, thereby validating the effectiveness of our proposed framework.

In PPI analysis, our results indicated that BIP and DRD2 have a higher combined score than other proteins related to S1R. A substantial body of evidence suggests that S1R, in combination with BIP, a

regulator of endoplasmic reticulum stress (ERS), plays a pivotal role in the ERS pathway, which is a component of cellular stress and a core mechanism underlying synaptic loss and neurodegeneration in AD pathology (Ortega-Roldan et al., 2013; Venkataraman et al., 2022). S1R-dependent neuroprotection is likely to be mediated by the regulation of the unfolded protein response (UPR) in ERS (Voronin et al., 2023). Under ERS conditions, S1R agonists promote the dissociation of S1R-BIP calcium ion-sensitive chaperone complexes, resulting in enhanced chaperone activity of BIP toward misfolded proteins and S1R binding to client protein IRE1 α . The regulatory effect of S1R agonists can increase the expression of BIP and brain-derived neurotrophic factor (BDNF) and decrease the expression of pro-inflammatory interleukin-6 (IL-6) (Hayashi and Su, 2007; Rosen et al., 2019; Zhemkov et al., 2021). Thus, S1R agonist regulation presents a viable strategy for the neuroprotective treatment of AD, aimed at reducing ERS and neuroinflammation while enhancing neural plasticity (Voronin et al., 2023).

It should be noted that the HNNTA model does not differentiate between ligands as agonists or antagonists of the targets. Unfortunately, the existing literature reports that haloperidol is an antagonist of S1R (Maurice and Su, 2009), while S1R agonists are potential drugs for treating AD. Therefore, haloperidol is not suitable for the treatment of AD. On the other hand, bromperidol, which was selected by the HNNTA model, may be the optimal candidate drug for AD treatment. The existing literature has discussed the potential of antipsychotic drugs, including bromperidol, on multiple targets related to AD (Kumar et al., 2017).

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.

References

- Barazzuol, L., Giamogante, F., and Cali, T. (2021). Mitochondria associated membranes (MAMs): Architecture and physiopathological role. *Cell Calcium* 94, 102343. doi:10.1016/j.ceca.2020.102343
- 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
- Briggs, R., Kennelly, S. P., and O'Neill, D. (2016). Drug treatments in alzheimer's disease. *Clin. Med.* 16, 247–253. doi:10.7861/clinmedicine.16-3-247
- Cantini, L., Medico, E., Fortunato, S., and Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Sci. Rep.* 5, 17386. doi:10.1038/srep17386
- Cheng, F., and Cummings, J. (2022). "Artificial intelligence in Alzheimer's drug Discovery," in *Alzheimer's disease drug development: research and development ecosystem*. Editors H. Fillit, J. Kinney, and J. Cummings (Cambridge: Cambridge University Press), 62–72. doi:10.1017/9781108975759.007
- Chou, K.-C. (2000). Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. biophysical Res. Commun.* 278, 477–483. doi:10.1006/bbrc.2000.3815
- Choudhury, C., Murugan, N. A., and Priyakumar, U. D. (2022). Structure-based drug repurposing: Traditional and advanced ai/ml-aided methods. *Drug Discov. Today* 27, 1847–1861. doi:10.1016/j.drudis.2022.03.006
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*
- Chyr, J., Gong, H., and Zhou, X. (2022). Dota: Deep learning optimal transport approach to advance drug repositioning for alzheimer's disease. *Biomolecules* 12, 196. doi:10.3390/biom12020196
- Consortium, U. (2019). Uniprot: A worldwide hub of protein knowledge. *Nucleic acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049
- Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. (2021). Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.* 61, 3891–3898. doi:10.1021/acs.jcim.1c00203
- Fan, L., Tan, L., Chen, Z., Qi, J., Nie, F., Luo, Z., et al. (2020). Haloperidol bound d2 dopamine receptor structure inspired the discovery of subtype selective ligands. *Nat. Commun.* 11, 1074. doi:10.1038/s41467-020-14884-y
- Fang, J., Zhang, P., Wang, Q., Chiang, C.-W., Zhou, Y., Hou, Y., et al. (2022). Artificial intelligence framework identifies candidate targets for drug repurposing in alzheimer's disease. *Alzheimer's Res. Ther.* 14, 7–23. doi:10.1186/s13195-021-00951-z
- Fisher, A., Bezprozvanny, I., Wu, L., Ryskamp, D. A., Bar-Ner, N., Natan, N., et al. (2015). AF710B, a novel M1/ σ 1 agonist with therapeutic efficacy in animal models of alzheimer's disease. *Neurodegener. Dis.* 16, 95–110. doi:10.1159/000440864
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic acids Res.* 40, D1100–D1107. doi:10.1093/nar/gkr777
- Harrell, F. E., Jr, Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* 247, 2543–2546. doi:10.1001/jama.1982.03320430047030
- Hasan, M. R., Paul, B. K., Ahmed, K., and Bhuyian, T. (2020). Design protein-protein interaction network and protein-drug interaction network for common cancer diseases: A bioinformatics approach. *Inf. Med. Unlocked* 18, 100311. doi:10.1016/j.imu.2020.100311

Author contributions

CY-C designed the research. XW, GC, ZL, and YY worked together to complete the experiment. XW and ZL contributed to analytic tools. GC and YY analyzed the data. XW, GC, ZL, YY, and CY-C wrote the manuscript together. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (grant No. 62176272), the Research and Development Program of Guangzhou Science and Technology Bureau (No. 2023B01J1016), the Key-Area Research and Development Program of Guangdong Province (No. 2020B1111100001), and China Medical University Hospital (DMR-112-085).

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.

- Hassan, N. M., Alhossary, A. A., Mu, Y., and Kwok, C.-K. (2017). Protein-ligand blind docking using QuickVina-W with inter-process spatio-temporal integration. *Sci. Rep.* 7, 15451. doi:10.1038/s41598-017-15571-7
- Hayashi, T., and Su, T.-P. (2007). Sigma-1 receptor chaperones at the endoplasmic reticulum interface regulate Ca^{2+} signaling and cell survival. *Cell* 131, 596–610. doi:10.1016/j.cell.2007.08.036
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735
- Hopkins, A. L. (2008). Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690. doi:10.1038/nchembio.118
- Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., and Sun, J. (2020). Deepurpose: A deep learning library for drug-target interaction prediction. *Bioinformatics* 36, 5545–5547. doi:10.1093/bioinformatics/btaa1005
- Hung, S.-Y., and Fu, W.-M. (2017). Drug candidates in clinical trials for alzheimer's disease. *J. Biomed. Sci.* 24, 47–12. doi:10.1186/s12929-017-0355-7
- Ihara, M., and Saito, S. (2020). Drug repositioning for alzheimer's disease: Finding hidden clues in old drugs. *J. Alzheimer's Dis.* 74, 1013–1028. doi:10.3233/JAD-200049
- Jiang, H., Wang, J., Cong, W., Huang, Y., Ramezani, M., Sarma, A., et al. (2022). Predicting protein-ligand docking structure with graph neural network. *J. Chem. Inf. Model.* 62, 2923–2932. doi:10.1021/acs.jcim.2c00127
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). Pubchem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109. doi:10.1093/nar/gky1033
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi:10.1145/3065386
- Kumar, S., Chowdhury, S., and Kumar, S. (2017). *In silico* repurposing of antipsychotic drugs for Alzheimer's disease. *BMC Neurosci.* 18, 76. doi:10.1186/s12868-017-0394-8
- Leal, N. S., and Martins, L. M. (2021). Mind the gap: Mitochondria and the endoplasmic reticulum in neurodegenerative diseases. *Biomedicines* 9, 227. doi:10.3390/biomedicines9020227
- Lee, W.-Y., Lee, C.-Y., Kim, Y.-S., and Kim, C.-E. (2019). The methodological trends of traditional herbal medicine employing network pharmacology. *Biomolecules* 9, 362. doi:10.3390/biom9080362
- Macias, A. T., Williamson, D. S., Allen, N., Borgognoni, J., Clay, A., Daniels, Z., et al. (2011). Adenosine-derived inhibitors of 78 kda glucose regulated protein (grp78) atpase: Insights into isoform selectivity. *J. Med. Chem.* 54, 4034–4041. doi:10.1021/jm101625x
- Malandraki-Miller, S., and Riley, P. R. (2021). Use of artificial intelligence to enhance phenotypic drug discovery. *Drug Discov. Today* 26, 887–901. doi:10.1016/j.drudis.2021.01.013
- Maurice, T., and Su, T.-P. (2009). The pharmacology of sigma-1 receptors. *Pharmacol. Ther.* 124, 195–206. doi:10.1016/j.pharmthera.2009.07.001
- Noor, F., Asif, M., Ashfaq, U. A., Qasim, M., and Tahir ul Qamar, M. (2023). Machine learning for synergistic network pharmacology: A comprehensive overview. *Briefings Bioinforma.* 24, bbad120. doi:10.1093/bib/bbad120
- Ortega-Roldan, J. L., Ossa, F., and Schnell, J. R. (2013). Characterization of the human sigma-1 receptor chaperone domain structure and binding immunoglobulin protein (bip) interactions. *J. Biol. Chem.* 288, 21448–21457. doi:10.1074/jbc.M113.450379
- Padhi, D., and Govindaraju, T. (2022). Mechanistic insights for drug repurposing and the design of hybrid drugs for alzheimer's disease. *J. Med. Chem.* 65, 7088–7105. doi:10.1021/acs.jmedchem.2c00335
- Parisi, D., Adasme, M. F., Sveshnikova, A., Bolz, S. N., Moreau, Y., and Schroeder, M. (2020). Drug repositioning or target repositioning: A structural perspective of drug-target-indication relationship for available repurposed drugs. *Comput. Struct. Biotechnol. J.* 18, 1043–1055. doi:10.1016/j.csbj.2020.04.004
- Patel, L., Shukla, T., Huang, X., Ussery, D. W., and Wang, S. (2020). Machine learning methods in drug discovery. *Molecules* 25, 5277. doi:10.3390/molecules2525277
- Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* 18, 41–58. doi:10.1038/nrd.2018.168
- Reczko, M., and Bohr, H. (1994). The def data base of sequence based protein fold class predictions. *Nucleic Acids Res.* 22, 3616–3619.
- Rodriguez, S., Hug, C., Todorov, P., Moret, N., Boswell, S. A., Evans, K., et al. (2021). Machine learning identifies candidates for drug repurposing in alzheimer's disease. *Nat. Commun.* 12, 1033. doi:10.1038/s41467-021-21330-0
- Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi:10.1021/ci100050t
- Rosen, D. A., Seki, S. M., Fernández-Castañeda, A., Beiter, R. M., Eccles, J. D., Woodfolk, J. A., et al. (2019). Modulation of the sigma-1 receptor-ire1 pathway is beneficial in preclinical models of inflammation and sepsis. *Sci. Transl. Med.* 11, eaau5266. doi:10.1126/scitranslmed.aau5266
- Ryskamp, D., Wu, L., Wu, J., Kim, D., Rammes, G., Geva, M., et al. (2019). Pridopidine stabilizes mushroom spines in mouse models of Alzheimer's disease by acting on the sigma-1 receptor. *Neurobiol. Dis.* 124, 489–504. doi:10.1016/j.nbd.2018.12.022
- Schmid, M., Wright, M., and Ziegler, A. (2016). *On the use of Harrell's C for clinical risk prediction via random survival forests*. Nature Publishing Group. ArXiv: 1507.03092 [stat].
- Schmidt, H. R., Zheng, S., Gurpinar, E., Koehl, A., Manglik, A., and Kruse, A. C. (2016). Crystal structure of the human σ_1 receptor. *Nature* 532, 527–530. doi:10.1038/nature17391
- Selvaraj, C., Chandra, I., Singh, S. K., and Abhirami, R. (2021). Artificial intelligence and machine learning approaches for drug design: Challenges and opportunities for the pharmaceutical industries. *Mol. Divers.* 126, 1–38. doi:10.1016/bs.apcsb.2021.02.001
- Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci.* 104, 4337–4341. doi:10.1073/pnas.0607879104
- Siavelis, J. C., Bourdakou, M. M., Athanasiadis, E. I., Spyrou, G. M., and Nikita, K. S. (2016). Bioinformatics methods in drug repurposing for alzheimer's disease. *Briefings Bioinforma.* 17, 322–335. doi:10.1093/bib/bbv048
- Srivastava, S., Ahmad, R., and Khare, S. K. (2021). Alzheimer's disease and its treatment by different approaches: A review. *Eur. J. Med. Chem.* 216, 113320. doi:10.1016/j.ejmech.2021.113320
- Stiefl, N., Watson, I. A., Baumann, K., and Zaliani, A. (2006). Erg: 2d pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* 46, 208–220. doi:10.1021/ci050457y
- Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., et al. (2023). The string database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 51, D638–D646. doi:10.1093/nar/gkac1000
- Thomas, S., Abraham, A., Baldwin, J., Piplani, S., and Petrovsky, N. (2022). Artificial intelligence in vaccine and drug design. *Vaccine Des. Methods Protoc.* 1, 131–146. doi:10.1007/978-1-0716-1884-4_6
- Tsuji, S., Hase, T., Yachie-Kinoshita, A., Nishino, T., Ghosh, S., Kikuchi, M., et al. (2021). Artificial intelligence-based computational framework for drug-target prioritization and inference of novel repositionable drugs for alzheimer's disease. *Alzheimer's Res. Ther.* 13, 92–15. doi:10.1186/s13195-021-00826-3
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. arXiv. doi:10.48550/arXiv.1706.03762
- Vatansever, S., Schlessinger, A., Wacker, D., Kaniskan, H. Ü., Jin, J., Zhou, M.-M., et al. (2021). Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions. *Med. Res. Rev.* 41, 1427–1473. doi:10.1002/med.21764
- Venkataraman, A. V., Mansur, A., Rizzo, G., Bishop, C., Lewis, Y., Kocagoncu, E., et al. (2022). Widespread cell stress and mitochondrial dysfunction occur in patients with early alzheimer's disease. *Sci. Transl. Med.* 14, eabk1051. doi:10.1126/scitranslmed.abk1051
- Voronin, M. V., Abramova, E. V., Verbovaya, E. R., Vakhitova, Y. V., and Seredenin, S. B. (2023). Chaperone-dependent mechanisms as a pharmacological target for neuroprotection. *Int. J. Mol. Sci.* 24, 823. doi:10.3390/ijms24010823
- Wang, J., and Dokholyan, N. V. (2022). Yuel: Improving the generalizability of structure-free compound-protein interaction prediction. *J. Chem. Inf. Model.* 62, 463–471. doi:10.1021/acs.jcim.1c01531
- Weng, T.-Y., Tsai, S.-Y. A., and Su, T.-P. (2017). Roles of sigma-1 receptors on mitochondrial functions relevant to neurodegenerative diseases. *J. Biomed. Sci.* 24, 74–14. doi:10.1186/s12929-017-0380-6
- Wilson, E. L., and Metzackopian, E. (2021). Er-Mitochondria contact sites in neurodegeneration: Genetic screening approaches to investigate novel disease mechanisms. *Cell Death Differ.* 28, 1804–1821. doi:10.1038/s41418-020-00705-8
- Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). Drugbank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906. doi:10.1093/nar/gkm958
- Wu, Y., Liu, H., Yan, J., and Hu, X. (2022). *Drug repositioning for alzheimer's disease with transfer learning*. arXiv. 10.48550/arXiv.2210.15271.
- Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., et al. (2021). Admetlab 2.0: An integrated online platform for accurate and comprehensive predictions of admet properties. *Nucleic Acids Res.* 49, W5–W14. doi:10.1093/nar/gkab255
- Yin, Z., and Wong, S. T. (2021). Artificial intelligence unifies knowledge and actions in drug repositioning. *Emerg. Top. Life Sci.* 5, 803–813. doi:10.1042/ETLS20210223
- Zhemkov, V., Geva, M., Hayden, M. R., and Bezprozvanny, I. (2021). Sigma-1 receptor (s1r) interaction with cholesterol: Mechanisms of s1r activation and its role in neurodegenerative diseases. *Int. J. Mol. Sci.* 22, 4082. doi:10.3390/ijms22084082
- Zhi, H.-Y., Zhao, L., Lee, C.-C., and Chen, C. Y.-C. (2021). A novel graph neural network methodology to investigate dihydroorotate dehydrogenase inhibitors in small cell lung cancer. *Biomolecules* 11, 477. doi:10.3390/biom11030477



OPEN ACCESS

EDITED BY

KunHong Liu,
Xiamen University, China

REVIEWED BY

Yanpeng Zhao,
Academy of Military Medical Sciences
(AMMS), China
Tan Qiong,
Xiamen University, China

*CORRESPONDENCE

Wenjie Du,
✉ duwenjie@mail.ustc.edu.cn
Yang Wang,
✉ angyan@ustc.edu.cn

[†]These authors have contributed equally
to this work

RECEIVED 04 May 2023

ACCEPTED 15 June 2023

PUBLISHED 30 June 2023

CITATION

Zhang J, Du W, Yang X, Wu D, Li J, Wang K
and Wang Y (2023), SMG-BERT:
integrating stereoscopic information and
chemical representation for molecular
property prediction.
Front. Mol. Biosci. 10:1216765.
doi: 10.3389/fmolb.2023.1216765

COPYRIGHT

© 2023 Zhang, Du, Yang, Wu, Li, Wang
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.

SMG-BERT: integrating stereoscopic information and chemical representation for molecular property prediction

Jiahui Zhang^{1,2†}, Wenjie Du^{1,2*†}, Xiaoting Yang^{2,3}, Di Wu^{1,2},
Jiahe Li^{1,2}, Kun Wang² and Yang Wang^{1,2,3*}

¹School of Software Engineering, University of Science and Technology of China, Hefei, China, ²Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, China, ³School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

Molecular property prediction is a crucial task in various fields and has recently garnered significant attention. To achieve accurate and fast prediction of molecular properties, machine learning (ML) models have been widely employed due to their superior performance compared to traditional methods by trial and error. However, most of the existing ML models that do not incorporate 3D molecular information are still in need of improvement, as they are mostly poor at differentiating stereoisomers of certain types, particularly chiral ones. Also, routine featurization methods using only incomplete features are hard to obtain explicable molecular representations. In this paper, we propose the Stereo Molecular Graph BERT (SMG-BERT) by integrating the 3D space geometric parameters, 2D topological information, and 1D SMILES string into the self-attention-based BERT model. In addition, nuclear magnetic resonance (NMR) spectroscopy results and bond dissociation energy (BDE) are integrated as extra atomic and bond features to improve the model's performance and interpretability analysis. The comprehensive integration of 1D, 2D, and 3D information could establish a unified and unambiguous molecular characterization system to distinguish conformations, such as chiral molecules. Intuitively integrated chemical information enables the model to possess interpretability that is consistent with chemical logic. Experimental results on 12 benchmark molecular datasets show that SMG-BERT consistently outperforms existing methods. At the same time, the experimental results demonstrate that SMG-BERT is generalizable and reliable.

KEYWORDS

molecular property prediction, chemical feature fusion, unambiguous molecular descriptor, molecular representation learning, molecular stereoscopic information

1 Introduction

The prediction of molecular properties is one of the fundamental tasks in chemistry (Wieder et al., 2020) and deserves special attention. Traditional computational methods, such as density functional theory (DFT) or field experiments, are time-consuming and poorly scalable with size (Chen et al., 2021). This could cause inevitable and serious moral and ethical issues with experimental testing involving animals or humans *in vivo*. Recently, Machine Learning (ML), including Deep Learning (DL), has emerged as a powerful data-

driven approach for establishing a connection between molecular structure and properties (Chen et al., 2021). ML methods can often deliver results that are comparable to DFT in terms of accuracy while being significantly faster by approximately 3–5 orders of magnitude (Hohenberg and Kohn, 1964; Kohn and Sham, 1965).

A key component/challenge in applying ML to molecular science is molecular featurization. This transforms molecular structures into machine-readable formats (Wu et al., 2018) and therefore dictates the embedded chemical information into final representations (Raghunathan and Priyakumar, 2021). Effective molecular representations are essential for a variety of molecular prediction tasks, such as property prediction (Du et al., 2023a), retrosynthesis (Segler et al., 2018; Zhang et al., 2022), generative molecular design (Moret et al., 2020), and so on (Dral and Barbatti, 2021). Current molecular representations can be categorized into three different classes: molecular fingerprints based on molecular topological substructures encoded as a sequence of bits, sequence-based representations by SMILES, and graph-based representations (Fang et al., 2022). However, current featurization methods still have certain shortcomings, as they only focus on extracting various hierarchical molecular information, which makes it challenging to thoroughly integrate the molecular information and achieve effective generalization among potential chemical compounds. In this study, one-dimensional (1D) SMILES strings, two-dimensional (2D) topological structures, and three-dimensional (3D) geometric structures are the intuitive expressions of molecular information at different levels. SMILES strings could naturally be used as input to some NLP models such as Transformer (Tetko et al., 2020; Schwaller et al., 2021) and BERT (Wang et al., 2019; Zhang et al., 2021) to reach high performance, no matter if for a molecular generation (Moret et al., 2020) or property prediction (Chen et al., 2021; Du et al., 2023a). However, these methods tend to lose the chemical context during preprocessing, as they often remove essential chemical symbols such as “#” and “()”, from the SMILES string. Moreover, only 1D information would inevitably lose adjacency information (Du et al., 2023b). The 2D topological structure is one of the most important chemical representations, which was expertly developed and has been used for centuries as a crucial carrier for the exchange, dissemination, and transmission of chemical knowledge. However, it is difficult to distinguish stereochemistry molecular features such as cis-trans isomerism, chirality, and other enantiomers only based on adjacency matrices (Stärk et al., 2021; Fang et al., 2022). Therefore, 3D information is an important and non-negligible piece of knowledge that the model needs to master to solve stereochemical problems (Chen et al., 2021; Du et al., 2023b). Each of these three modalities focuses on different aspects, and all are fundamental to molecular featurization.

On the other hand, interpretability is also an obstacle to the widespread application of deep learning models. Current ML models mainly focus on the prediction task of compound properties, but only a few ML methods are interpretable (Wang et al., 2021). Therefore, there is often a trade-off between predictive performance and the ability to interpret ML models (Rodríguez-Perez and Bajorath, 2021). Although causal analysis theories such as contrastive explanations or counterfactuals, feature perturbation (sensitivity analysis), and gradient-based methods could obtain feature importance analysis to a certain extent, interpretable results still need to be improved to match the actual chemical logic for individual explanations

(Prosperi et al., 2020; Wang et al., 2021). Attention mechanisms have been widely adopted for visualizing molecular prediction results, as they allow for intuitive visualization and human-friendly explanations (Ross et al., 2022). However, to the best of our knowledge, current attention mechanisms rarely embed basic chemical intuitions or expert prior knowledge to enhance interpretability. Chemical properties are ultimately determined by intrinsic properties (Zhang et al., 2022), and most of these are determined by the electron density and electronegativity of neighboring atoms, which could be represented by NMR chemical shifts and bond dissociation energy (BDE). Thus, we could consider them perfect candidates for ML descriptors to improve model interpretability.

In this paper, we propose a stereo molecular graph BERT (SMG-BERT) by integrating the 3D space geometric parameters, 2D adjacency information, and 1D SMILES representation into a self-attention-based BERT model. SMG-BERT could generate accurate chemical representations for various molecules, including chiral molecules, which provides assurance for precise property prediction results and expands the application scope. Meanwhile, SMG-BERT incorporates the NMR chemical shifts and bond dissociation energies (BDEs) as chemical descriptors using a transformer encoder to improve interpretability. This results in visualizations that conform to chemical logic and are more convincing. A series of experimental results show that SMG-BERT can consistently outperform previous state-of-the-art molecular property prediction models on 12 benchmark molecular datasets.

2 Methods

In this section, we describe in detail the data preprocessing process, model structure, and loss function in three parts. In the data preprocessing process, the model could obtain an input representation that consists of three components: a molecular representation z is generated solely from the atomic and NMR sequence by the embedding layer, which lacks topological information and thus can be regarded as 1D information. The bond dissociation energy matrix B , which not only provides topological information but also includes vital chemical knowledge about bond energies. Finally, the distance fraction matrix D , based on the distance matrix D^{raw} , could be regarded as 3D information. We present the implementation details of our model architecture, which is based on the transformer-encoder architecture and introduces multiple modal information of the molecules. Meanwhile, various learning tasks are presented in the pre-training phase to enhance the representation capabilities of the model.

2.1 Data preprocessing

In the pre-training process, the dataset was collected from PubChem (Kim et al., 2023). Although increasing the amount of pre-training data could potentially further improve the performance of the model, the improvement in model performance became less significant after a 480 k training size (Supplementary Figure S1). Considering the balance between training time and effect, we

randomly selected 480 k molecules (SMILES). Three preprocessing tasks were performed, including generating: (1) input representation z of the molecules (2) the bond dissociation energy matrix B , (3) and the distance fraction matrix D .

The input representation z of the molecules: we used RDKit to transform each SMILES into an atomic sequence $S_A = [A_1, A_2, \dots, A_n]$ of length n and generate the corresponding NMR sequence $S_N = [N_1, N_2, \dots, N_n]$ for the atomic sequence S_A by a DL model with six message-passing neural networks (MPNN) layers and two fully connected network layers as in our previous work (Zhang et al., 2022) (continuous NMR was transformed into discrete). Then, 80% of Atom/NMR in the two sequences were randomly selected and replaced by <M> (which stands for MASKL); 10% were replaced by another Atom/NMR one, and the rest were left unchanged. In addition, we added a global node <G> at the beginning of the sequence, which represents the global representation of the whole molecule. Finally, two independent embedding layers were used to map the two new input sequences S'_A and S'_N to a continuous input representation $z = [z_1, z_2, \dots, z_n]$:

$$z = E_A(S'_A) \parallel E_N(S'_N)$$

where E_A is the embedding layer of the atomic sequence, E_N is the embedding layer of the NMR sequence, and \parallel denotes the concatenation operation.

The bond energy matrix: we generated the bond energy matrix B by an additional DL model with four MPNN layers and two fully connected network layers according to the method in our previous work (Zhang et al., 2022), and normalized it:

$$B^{\text{norm}} = \text{Norm}(B) = \frac{B - B_{\min}}{B_{\max} - B_{\min}}$$

where B_{\max} is the maximum value of matrix B and B_{\min} is the minimum value of matrix B .

Distance fraction matrix D' : the ground state 3D structure of the molecule can be obtained by the RDKit package. Based on this, we were able to obtain the atomic distance information and generate the original distance matrix D^{raw} . Then, the distance matrix D^{raw} was transformed by a transformer encoder layer into the distance fraction matrix D .

$$D = \text{Trans}(D^{\text{raw}})$$

where D^{raw} represents the 1, 2, \dots , n -th column vector of D^{raw} , and Trans is a transformer encoder module.

2.2 Modified attention mechanism

Our model is based on the self-attention mechanism. For our task, the input representation z was first mapped onto the query matrix Q , the key matrix K , and the value matrix V using the projection matrices W_q, W_k, W_v , respectively:

$$Q = W_q z$$

$$K = W_k z$$

$$V = W_v z$$

The attention score matrix A could then be calculated from the Q, K matrix. Specifically, we computed the dot products of the query with all keys, divided each by d_k , and applied a softmax function to obtain the weights on the values.

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

where d_k is the dimension of the key.

However, the global attention score matrix, A , is difficult to optimize because it requires considering the relationships among all the atoms, resulting in a high degree of freedom. To address this problem, we introduced an adjacency matrix to constrain the global attention score matrix:

$$M = \text{Binary}(B)$$

$$A_{2d} = A \odot M + \lambda \text{Norm}(B^{\text{norm}})$$

where "Binary" is a binarization operation that transforms the bond-energy matrix into an adjacency matrix M , \odot denotes an element-wise product, and λ is a balancing hyperparameter between the mask attention score matrix and the bond-energy matrix. Here, λ is set to 0.2. The hyperparameters are provided in Supplementary Tables S1, S2.

Furthermore, to incorporate 3D information, we brought the distance matrix D into the attention score matrix to reflect the interaction strength of atoms:

$$A_{3d} = A_{2d} + D$$

Once the final correlation matrix A_{3d} is obtained, we multiplied it with the value matrix V to obtain the output sequence z :

$$z = A_{3d} V$$

In addition to the attention sub-layers, the transformer encoder layer also contains a position-wise feed-forward network:

$$r_i = \text{FFN}(z_i)$$

where r_i denotes the final output representation of the i -th atom. We wrote the representation of the whole sequence of atoms as $r = [r_1, r_2, \dots, r_n]$.

2.3 Loss function

During the pre-training stage, we aimed to increase the richness of information contained in the atomic representation sequence r . To achieve this, we propose three self-supervised learning (SSL) tasks: atomic and NMR reconstruction, bond energy prediction, and 3D information reconstruction.

Atomic and NMR reconstruction: During data preprocessing, some atoms in the atomic sequence are randomly replaced by the special token "<M>". The task of atomic reconstruction involves predicting the correct class of these masked atoms. Specifically, given the representation r_i of the masked atom, the model outputs the predicted class probability p_i after passing through the MLP and SoftMax layers.

$$p_i = \text{softmax}(\text{MLP}(r_i))$$

The cross-entropy loss is used as the loss function, which computes the difference between the predicted probability p_i and the ground truth label y_i of the masked atom:

$$\mathcal{L}_A = -\frac{1}{m} \sum_{i=1}^m y_i \log p_i$$

where m is the total number of masked atoms.

Similarly, the NMR reconstruction task is consistent with the atomic reconstruction principle, which we denoted as \mathcal{L}_N .

Bond energy prediction: The bond representation can be determined by the nodes connected at both ends. The predicted bond energy q_{ij} between the atomic representation r_i and r_j can then be obtained by running the bond representation through the MLP.

$$q_{ij} = \text{MLP}(r_i \parallel r_j)$$

where \parallel denotes the concatenation operation. Mean Squared Error (MSE) is the loss function and y_{ij} is the ground truth:

$$\mathcal{L}_B = \sum_{i=1}^n \sum_{j=1}^n (y_{ij} - q_{ij})^2$$

3D information reconstruction: To avoid the complexity of modeling direct prediction of atomic coordinates, which requires translation-rotation invariance and order invariance, we use intermediate quantities that reflect 3D information, such as interatomic distances, bond angles, and torsion angles, to predict atomic coordinates. Specifically, the atomic representation r is mapped to a new representation r' using the projection matrix W_r , with a vector length of 3 to represent the coordinates in 3D space.

$$r' = W_r r$$

The interatomic distances \hat{d} , bond angles $\hat{\theta}$, and torsion angle $\hat{\varphi}$ predicted by the model can be calculated directly:

$$\begin{aligned} \hat{d} &= \|r'_i - r'_j\|_2 \\ \hat{\theta} &= \cot^{-1} \left(\frac{r'_i \cdot r'_j}{\langle r'_i, r'_j \rangle} \right) \\ \hat{\varphi} &= \cos^{-1} \left(\frac{n_\alpha \cdot n_\beta}{\|n_\alpha\| \cdot \|n_\beta\|} \right) \end{aligned}$$

where i and j refer to two different atoms, r'_i and r'_j indicate the coordinate vectors of atoms i and j , n_α and n_β correspond to the normal vector of the α and β planes.

Finally, we used the mean squared error (MSE) as the loss function to compute the difference between the predicted values and the corresponding ground truth values for atomic distances d , bond angles θ , and torsion angles φ .

$$\mathcal{L}_{3D} = (d - \hat{d})^2 + (\theta - \hat{\theta})^2 + (\varphi - \hat{\varphi})^2$$

Loss functions: To balance the different objective functions represented by L_A , L_N , L_B , and L_{3D} , it is necessary to consider their relative importance. The σ_1 , σ_2 , σ_3 , and σ_4 are the learnable parameters as the proportion of L_A , L_N , L_B , and L_{3D} in the total loss (Kendall et al., 2017), and are optimized through backpropagation to appropriate values. This enables the model to

effectively learn from all four SSL tasks while ensuring that the different losses are appropriately weighted.

$$\mathcal{L} = \frac{1}{\sigma_1^2} L_A + \frac{1}{\sigma_2^2} L_N + \frac{1}{\sigma_3^2} L_B + \frac{1}{\sigma_4^2} L_{3D} + \log \sigma_1 + \log \sigma_2 + \log \sigma_3 + \log \sigma_4$$

2.4 Baseline model and test data sets

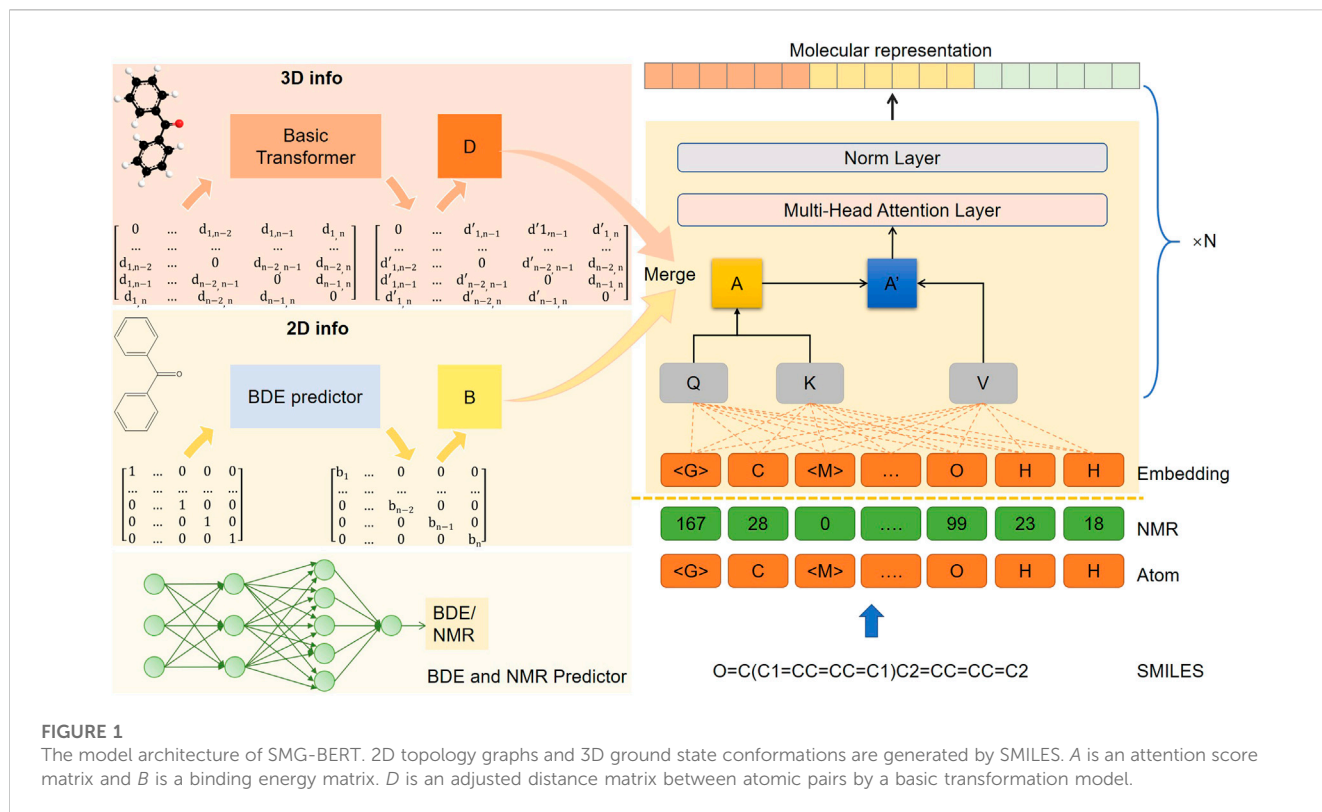
Several advanced models in recent years were selected for comparison as the baseline, namely, GAT (Veličković et al., 2017), GIN (Xu et al., 2018), D-MPNN (Yang et al., 2019), GROVER (Rong et al., 2020), GraphMVP (Liu et al., 2021a), and AttentiveFP (Xiong et al., 2019). Among them, GIN, GAT, D-MPNN, and AttentiveFP are all non-pre-training methods based on GNN. GAT introduced the attention mechanism into GNN and adaptively learned the weight of nodes. GIN was derived from the Weisfeiler-Lehman graph isomorphism test degree and exhibited almost the same representation ability as the WL test. D-MPNN utilizes messages that are associated with directed edges (bonds) rather than atom nodes. AttentiveFP presents a novel graph neural network architecture that incorporates an attention mechanism to extract nonlocal effects at the intramolecular level for molecular representation. GROVER and GraphMVP employ a pre-training process. GROVER can effectively learn rich structural and semantic information about molecules from a large volume of unlabeled molecular data by performing SSL tasks at the node, edge, and graph levels. Meanwhile, GraphMVP uses an SSL approach to achieve correspondence and consistency between 2D topological structures and 3D geometric views.

A total of 12 datasets (seven for regression and five for classification) were selected from MoleculeNet (Wu et al., 2018) and ADMETlab (Dong et al., 2018) to conduct downstream experiments. According to this benchmark (Rong et al., 2020; Liu et al., 2021b), we split these datasets with scaffolds according to the molecular substructure, as this splitting method is more challenging and better evaluates the generalization ability in out-of-distribution data. In the testing process, we randomly selected 80% of the samples as the training set, 10% as the validation set, and the remaining 10% as the test set. Five independent runs were executed for each method, and the mean and standard deviation of the metrics were reported. ROC-AUC, RMSE, and R^2 are used as evaluation indicators for classification and regression tasks, respectively.

3 Results and discussion

3.1.1 Model architecture of SMG-BERT

The architecture of our model is shown Figure 1, consisting of one embedding layer, six transformer encoder layers, and one output layer. The model processes 1D, 2D, and 3D information separately. The 1D information includes both the atomic sequence obtained from the SMILES string using the RDKit package (Landrum, 2019) and the NMR sequence generated (Zhang et al., 2022) (the predicted NMR values are discretized). Each sequence is independently masked by about 20% (as a hyperparameter) and then embedded



in a high-dimensional vector space through two separate embedding layers. For the 2D information, we introduced the bond energy result (B matrix in Figure 1) to provide differentiation information about the bond connection. The B matrix is fused into the global attention score matrix (A matrix in Figure 1) at the transformer encoder layer. As for the 3D geometric information, we calculated the interatomic distances, bond angles, and torsion angles in the ground state conformations using the RDKit package (Faber et al., 2017; Lubbers et al., 2018). The distance matrix was then processed by an additional transformer encoder module to obtain the distance fraction matrix (D matrix in Figure 1) as the final 3D information, where the farther distance could have a smaller value. These three modal inputs, along with multiple self-supervised learning tasks, which include masked atom inference and 3D geometric feature reconstruction, allow for a multimodal representation of model learning.

The resulting molecular representation would be used for downstream tasks and would adopt the fine-tuning method. Specifically, after pre-training, the atom representation of the global super-node "<G>" is the final molecular representation, with a 512-dimensional vector. This would be fed into a two-layer, fully connected network with random initialization, which yields the final prediction results. The network uses ReLU as the activation function and sets the dropout ratio to 0.1. Considering that catastrophic forgetting issues could occur as the model targets specific downstream tasks that are completely different from the pre-training process (Kirkpatrick et al., 2017), we would retain the pre-training loss as a regular term, which would maintain the chemical information and spatial characteristics learned in the pre-training process. In addition, our model is a flexible,

comprehensive feature fusion framework that supports multi-dimensional information removal and fusion. For specific downstream tasks, 3D or chemical information could be considered a super parameter, and we could dynamically adjust or increase the available input features according to the target.

3.2 Model validation results on common datasets

Table 1 shows that compared to no pre-training, the RMSE index decreased by 12.71%, while the ROC-AUC improved by 20.7% on the classification task. And R^2 increased by 5.07% in Supplementary Table S3. These results demonstrate the importance and necessity of pre-training in our strategy. Moreover, a noteworthy trend is that the smaller the dataset, such as FreeSolv and ESOL, the higher the improvement effect to some extent, which demonstrates the excellent generalization ability of the pre-trained model. Besides, Table 1 also records the prediction results and the performance of our model with several advanced models. SMG-BERT outperforms six out of eight baselines and achieves a close second in the other two (Tox21 and HIV). Specifically, in all four regression datasets, SMG-BERT achieves the SOTA results and has an overall relative improvement of 15.3% on average compared to previous SOTA results. Relatively, only 5.81% is achieved on average for the AUC-ROC score in classification tasks, which could be due to the regression tasks being more relevant to the 3D geometric information of molecules (Fang et al., 2022), such as the label of water-soluble or hydration-free energies in ESOL and FreeSolv dataset, which is

TABLE 1 Overall performance for regular regression and classification tasks.

Regression					Classification							
Dataset	ESOL	FreeSolv	Lipo	LogS	QM7	QM8	QM9	BACE	Tox21	HIV	BBBP	BBBP
NO. molecules	1,128	642	4,200	5,045	6,830	21,786	133,885	1,513	7,831	41,127	2039	2039
GIN	0.982 _(0.049)	2.023 _(0.036)	0.723 _(0.038)	1.739 _(0.123)	94.7 _(4.32)	0.0193 _(0.0011)	0.00923 _(0.00007)	0.752 _(0.027)	0.768 _(0.008)	0.727 _(0.013)	0.663 _(0.021)	0.663 _(0.021)
GAT	1.433 _(0.078)	2.317 _(0.077)	1.054 _(0.056)	1.675 _(0.166)	84.6 _(3.98)	0.0182 _(0.0009)	0.00868 _(0.00012)	0.771 _(0.015)	0.755 _(0.006)	0.746 _(0.007)	0.641 _(0.032)	0.641 _(0.032)
D-MPNN	0.988 _(0.010)	1.889 _(0.042)	0.732 _(0.053)	1.302 _(0.084)	101.6 _(4.32)	0.0201 _(0.0007)	0.01023 _(0.00004)	0.799 _(0.025)	0.750 _(0.060)	0.769 _(0.009)	0.712 _(0.024)	0.712 _(0.024)
AttentiveFP	0.865 _(0.066)	1.891 _(0.063)	0.710 _(0.012)	1.226 _(0.066)	69.3 _(3.78)	0.0204 _(0.0008)	0.00873 _(0.00004)	0.792 _(0.024)	0.765 _(0.007)	0.760 _(0.006)	0.721 _(0.017)	0.721 _(0.017)
GROVER	0.973 _(0.042)	1.826 _(0.101)	0.766 _(0.033)	1.214 _(0.032)	91.3 _(3.29)	0.0211 _(0.0014)	0.00802 _(0.00005)	0.812 _(0.016)	0.749 _(0.004)	0.701 _(0.011)	0.701 _(0.013)	0.701 _(0.013)
GraphMVP	0.947 _(0.020)	1.841 _(0.054)	0.718 _(0.033)	1.163 _(0.073)	98.4 _(4.20)	0.0208 _(0.0018)	0.00899 _(0.00007)	0.819 _(0.017)	0.772 _(0.003)	0.743 _(0.007)	0.722 _(0.016)	0.722 _(0.016)
Our method (no PT)	0.974 _(0.033)	1.893 _(0.063)	0.756 _(0.033)	1.295 _(0.033)	86.3 _(3.78)	0.0193 _(0.0012)	0.00942 _(0.00006)	0.660 _(0.039)	0.764 _(0.008)	0.710 _(0.016)	0.649 _(0.022)	0.649 _(0.022)
Our method (PT)	0.859 _(0.029)	1.616 _(0.047)	0.694 _(0.033)	1.120 _(0.052)	57.4 _(3.01)	0.0172 _(0.0008)	0.00792 _(0.00004)	0.823 _(0.012)	0.766 _(0.008)	0.758 _(0.007)	0.736 _(0.014)	0.736 _(0.014)

ROC-AUC was used for classification tasks, and RMSE was used for regression tasks, with standard deviations in brackets; PT, pre-training. Bold numbers indicate the best result. Standard deviations are in brackets. Bold numbers indicate the best result.

closely related to the molecular polarity, which is in turn the geometric symmetry concept of the 3D conformation of a molecule. Especially on the QM7, QM8, and QM9 datasets, the improvement results are more significant, reaching an average of 20.7%. The properties in these datasets are directly related to the 3D geometric information.

On the other hand, stereochemical molecules deserve our special attention because they are a rarely studied class of molecules in nature. Current DL models often overlook chiral pair discrimination, leading to inaccurate predictions (MacKenzie and Stachelek, 2021; Cho et al., 2023). Although chiral analysis is fundamental to many fields, limited datasets restrict our ability to study it. Nonetheless, we conducted a macromolecule chiral classification task to evaluate SMG-BERT's prediction and generalization capabilities. A protein-chiral ligand binding dataset was used in this case, where each enantiomer of the ligand could demonstrate significantly different binding affinities. In this dataset, a chiral pair was defined as two enantiomers measured in the same biochemical binding assay, which is a common occurrence in biochemistry referred to as a "chiral cliff" (Schneider et al., 2018) (Figure 2A). The dataset contained approximately 3,800 chiral pairs with a more complex structure that included a diverse range of atoms and elements, such as C, H, O, N, B, Br, Cl, and so on (Figure 2B).

This dataset was divided into training, validation, and test sets in a ratio of 8:1:1. As shown in Figure 2C, SMG-BERT could effectively discriminate between chiral molecules, achieving an AUC score of 0.75, which is about 12.81% higher than the other models on average. The PRC curve also shows that our model outperformed the other models (Figure 2D). Obviously, including 3D geometric information models such as GraphMVP or GROVER is better than using models based on 2D molecular graphs since the left- and right-handed versions of enantiomers have identical connectivity (Du et al., 2023b). Additionally, as we can see, without the pre-training process, the classification accuracy of the model would drop significantly, approaching 50%. This level of accuracy is virtually meaningless, given that the problem is a binary classification task. 3D information is relatively difficult to capture and is especially important in 3D-related downstream tasks. During pre-training, our model focuses on learning the complete 3D stereo geometric information of the molecules by incorporating interatomic distance, angle, and dihedral angle, which is a critical factor contributing to the superiority of our model over other models. In addition, the explicitly introduced distance information is also more conducive to the interpretability of the model and better reflects the correlation between the atoms.

3.3 Interpretability analysis

In the final phase of our study, we examined the attention matrix generated by SMG-BERT to reveal the chemical insight acquired during pre-training. We calculated the similarity between attentional scores for atoms at different levels of information integration, using the benzophenone molecule (C₁₅H₁₂O) as a case study. We also presented visualization results for several molecules.

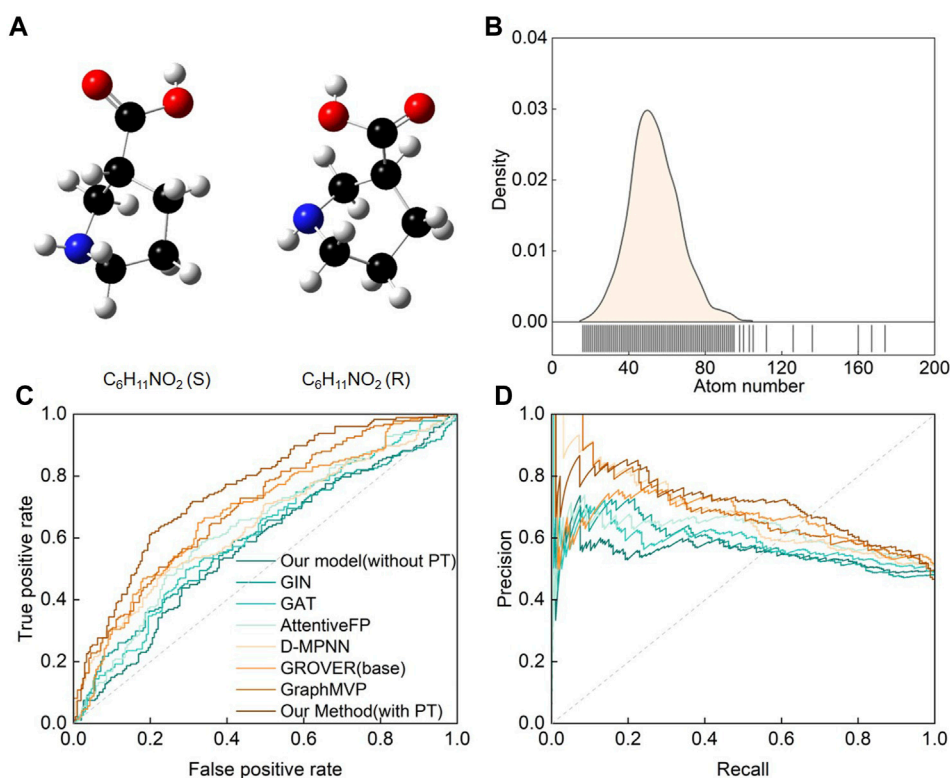


FIGURE 2

Performance of the SMG-BERT model in discriminating chiral molecules. (A) A pair of chiral molecules of L-proline and R-proline as an example. (B) Atom count distribution of the chiral molecule dataset. (C) ROC and (D) PRC curves of CFFN compared with other random classification models in discriminating enantiomeric pairs. PT: Pre-training.

Figure 3A shows that the molecular representation obtained solely from 1D SMILES string information in pre-training for the benzophenone molecule (depicted in Figure 1) is relatively vague. The similarity between different atoms is within 0.001, indicating a lack of learned explicit chemical information and atomic differences (Figure 3A). However, after incorporating 2D information, the overall correlation between atoms increased, and some regions became more pronounced (Figure 3B). Notably, the current high correlation is closely related to the adjacency matrix, especially the higher attention scores of the atoms themselves, while the correlation in other unrelated regions is relatively low. This suggests that the model initially pays sufficient attention to adjacency information, but it is still not the chemically logical result we expected. Furthermore, the addition of 3D geometric information led to significant changes in the model's attention scores, with atoms themselves receiving a score of 0 due to the 3D information matrix values, and two nearly symmetrical rectangular regions emerging (Figure 3C). This is because benzophenone has two symmetrical phenyl rings on its left and right sides with nearly identical geometric information. These findings are consistent with expectations and demonstrate that 3D information significantly enhances the model's output representation, making it more consistent with chemical spatial geometric information. After incorporating the chemical information, noticeable differences are seen in the roughly similar phenyl ring regions compared to the previous results (Figure 3D). This phenomenon could be attributed to the ketone group (C=O), as a strongly polar group, having a stronger electron cloud-attracting ability than the phenyl ring, which disrupts the original large π

bond conjugation system of the phenyl ring and re-forms a stable conjugated structure. In this case, the chemical information clearly reflects the influence of the chemical environment on the atoms, such as chemical shifts in NMR. This clearly shows that the added chemical information effectively improves the interpretability of the model and makes the results of the attention matrix more in line with chemical knowledge.

Here we present another six molecules to represent the pre-training results of SMG-BERT (Figure 3E). The model can effectively capture the weight results of different atoms and even differentiate between symmetric substructures in molecules such as benzophenone. Our results highlight the integration of spatial structure information and chemical priors in the model.

3.4 Ablation experiment

In this section, we present various ablation analyses of SMG-BERT to gain insight into its remarkable performance. To understand the impact and confirm the importance of explicit information, we performed a series of ablation analyses by removing the corresponding modal components from SMG-BERT. This new variant removes either 3D information and/or chemical information and serves as a comparison to the vanilla version. We conducted 10 random tests on eight datasets for classification and regression tasks. First, we compared the variant without chemical information in

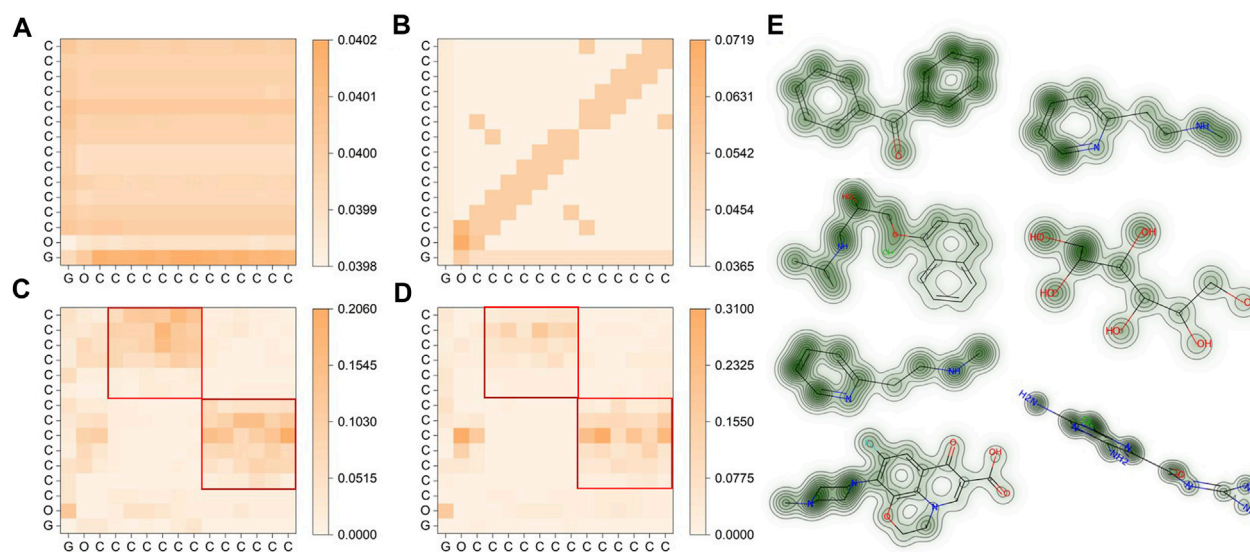


FIGURE 3

Visualization of molecular representations for benzophenone in SMG-BERT with varying degrees of features. (A) Only 1D information is considered. (B) 1D + 2D information is considered. (C) 1D + 2D + 3D geometric information is considered. (D) 1D + 2D + 3D + chemical information is considered. The red squares are the positions of the two benzene rings. (E) Attention maps for (left column, from top to bottom) benzophenone, propranolol, betahistine, ofloxacin, betahistine, hexitol, and amiloride. Greener areas represent higher weight values.

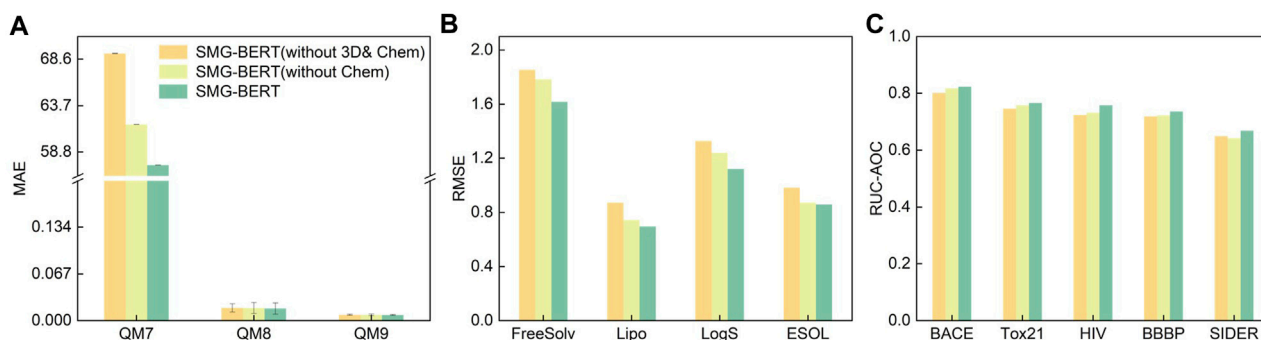


FIGURE 4

Results of the ablation experiment on regression and classification datasets. Test results on (A) QM7, QM8, and QM9 datasets; (B) on FreeSolv, Lipo, LogS, and ESOL datasets; and (C) on BACE, Tox21, HIV, BBBP, and SIDER datasets.

terms of changes in classification and regression tasks. Overall, SMG-BERT exhibited varying degrees of performance degradation after removing chemical information, especially in more challenging regression tasks where its RMSE increased by approximately 10% (Figures 4A and 4B). Conversely, removing chemical information had only a small impact on classification tasks, with a decrease of approximately 5% (Figure 4C). This demonstrates that incorporating chemical knowledge can enhance the model's expressive power and improve its performance. Furthermore, we removed 3D information on this basis (without 3D & Chem) and found that the model's results became worse, with an average increase in RMSE errors of approximately 7%. This also illustrates the effectiveness and importance of 3D information.

Explicitly adding 3D and chemical information introduces a new problem: an increase in complexity. However, with more complete guidance, unsupervised large-scale models are more likely to learn detailed molecular/atomic features and output precise molecular representations. 3D information increases the model's attention to the relationship between atoms and unbound atoms, while chemical information supplements the influence of the surrounding groups on atoms. This information can provide guidance for the model's important domain knowledge, resulting in superior performance. The ablation analysis results of the three sets of experiments undoubtedly confirm the accuracy and robustness of our model. And the importance of 3D and chemical information.

4 Conclusion

Molecular representations play an important role in determining both the performance and the interpretability of machine learning models. While most explanatory methods can be applied regardless of the features or descriptors used, the interpretability of features is critical for effective explanations. In particular, features should be both understandable and chemically intuitive whenever possible. For instance, if a specific atom or functional group strongly influences the prediction of high metabolic clearance, a medicinal chemist may consider replacing it. Thus, it is essential that key descriptors are actionable to understand the process by which a prediction is made, which can increase model transparency, facilitate the integration of expert knowledge, enable model tuning for specific applications, and uncover valuable insights, such as learned QSPR patterns.

In this study, we introduced a novel model, called stereo molecular graph BERT (SMG-BERT), which integrates a number of molecular features, including 3D spatial geometric parameters, 2D adjacency information, and 1D SMILES representation, into a self-attention-based BERT architecture. Additionally, SMG-BERT incorporates NMR chemical shifts and BDEs as chemical descriptors through a transformer encoder, which improves interpretability and results in visualizations that are chemically consistent and more compelling. As the result shows, SMG-BERT generates accurate chemical representations for various molecules, including chiral molecules, ensuring precise property prediction results and expanding the scope of applications. In contrast, our work focuses exclusively on chiral pairs, meaning that only compounds with a chiral center were considered, while chiral centers in sulfur or phosphorus were excluded. Diastereomers and atropisomers were not taken into account in this work, as diastereomers are not mirror images, and the conformation of atropisomers is typically not described in most activity databases.

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

- Chen, D., Gao, K., Nguyen, D. D., Chen, X., Jiang, Y., Wei, G. W., et al. (2021). Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat. Commun.* 12, 3521. doi:10.1038/s41467-021-23720-w
- Cho, N. H., Guerrero-Martínez, A., Ma, J., Bals, S., Kotov, N. A., Liz-Marzán, L. M., et al. (2023). Bioinspired chiral inorganic nanomaterials. *Nat. Rev. Bioeng.* 1, 88–106. doi:10.1038/s44222-022-00014-4
- Dong, J., Wang, N. N., Yao, Z. J., Zhang, L., Cheng, Y., Ouyang, D., et al. (2018). ADMETLab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J. Cheminform* 10, 29. doi:10.1186/s13321-018-0283-x
- Dral, P. O., and Barbatti, M. (2021). Molecular excited states through a machine learning lens. *Nat. Rev. Chem.* 5, 388–405. doi:10.1038/s41570-021-00278-1
- Du, W., Yang, X., Wu, D., Ma, F., Zhang, B., Bao, C., et al. (2023a). Fusing 2D and 3D molecular graphs as unambiguous molecular descriptors for conformational and chiral stereoisomers. *Brief. Bioinform* 24, bbac560. doi:10.1093/bib/bbac560
- Du, W., Yang, X., Wu, D., Ma, F., Zhang, B., Bao, C., et al. (2023b). Fusing 2d and 3d molecular graphs as unambiguous molecular descriptors for conformational and chiral stereoisomers. *Briefings Bioinforma.* 24, 1–12. doi:10.1093/bib/bbac560
- Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., et al. (2017). Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* 13, 5255–5264. doi:10.1021/acs.jctc.7b00577
- Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., et al. (2022). Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* 4, 127–134. doi:10.1038/s42256-021-00438-4
- Hohenberg, P., and Kohn, W. (1964). Inhomogeneous electron gas. *Phys. Rev.* 136, B864–B871. doi:10.1103/physrev.136.b864
- Kendall, A., Gal, Y., and Cipolla, R. (2017). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. Available at: <https://arxiv.org/abs/1705.07115> (Accessed May 19, 2017).
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2023). PubChem 2023 update. *Nucleic Acids Res.* 51 (2023), D1373–D1380. doi:10.1093/nar/gkac956
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U. S. A.* 114, 3521–3526. doi:10.1073/pnas.1611835114
- Kohn, W., and Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Phys. Rev.* 140, A1133–A1138. doi:10.1103/physrev.140.a1133

Author contributions

WD, JZ, and YW designed the research; WD, XY, DW, and JZ performed the research and analyzed the data; and WD, JZ, and YW wrote the paper. All authors contributed to the article and approved the submitted version.

Funding

This paper was partially supported by the Project of Stable Support for Youth Teams in Basic Research Field, CAS (YSBR-005), the Anhui Science Foundation for Distinguished Young Scholars (No. 1908085J24), the Natural Science Foundation of China (No. 62072427), and the Jiangsu Natural Science Foundation (No. BK20191193).

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2023.1216765/full#supplementary-material>

- Landrum, G. (2019). RDKit: Open-source cheminformatics from machine learning to chemical registration. *Abstr. Pap. Am. Chem. Soc.* 258, 15–24. doi:10.1021/ja02125a604
- Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. (2021a). Pre-training molecular graph representation with 3d geometry. Available at : <https://arxiv.org/abs/2110.07728> (Accessed October 7, 2021).
- Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. (2021b). Pre-training molecular graph representation with 3d geometry. Available at : <https://arxiv.org/abs/2110.07728> (Accessed October 7, 2021).
- Lubbers, N., Smith, J. S., and Barros, K. (2018). Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* 148, 241715. doi:10.1063/1.5011181
- MacKenzie, L. E., and Stachek, P. (2021). The twists and turns of chiral chemistry. *Nat. Chem.* 13, 521–522. doi:10.1038/s41557-021-00729-8
- Moret, M., Friedrich, L., Grisoni, F., Merk, D., and Schneider, G. (2020). Generative molecular design in low data regimes. *Nat. Mach. Intell.* 2, 171–180. doi:10.1038/s42256-020-0160-y
- Proserpi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., et al. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat. Mach. Intell.* 2, 369–375. doi:10.1038/s42256-020-0197-y
- Ragunathan, S., and Priyakumar, U. D. (2021). Molecular representations for machine learning applications in chemistry. *Int. J. Quantum Chem.* 122, e26870. doi:10.1002/qua.26870
- Rodriguez-Perez, R., and Bajorath, J. (2021). Explainable machine learning for property predictions in compound optimization. *J. Med. Chem.* 64, 17744–17752. doi:10.1021/acs.jmedchem.1c01789
- Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., et al. (2020). “Self-supervised graph transformer on large-scale molecular data,” in Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver BC Canada, December 2020, 12559–12571.33.
- Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* 4, 1256–1264. doi:10.1038/s42256-022-00580-7
- Schneider, N., Lewis, R. A., Fechner, N., and Ertl, P. (2018). Chiral cliffs: Investigating the influence of chirality on binding affinity. *ChemMedChem* 13, 1315–1324. doi:10.1002/cmdc.201700798
- Segler, M. H. S., Preuss, M., and Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610. doi:10.1038/nature25978
- Schwaller, P., Hoover, B., Jean-Louis, R., Hendrik Strobel, and Laino, Teodoro (2021). Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* 7, 1–9. doi:10.1126/sciadv.abe4166
- Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., et al. (2021). 3D infomax improves gns for molecular property prediction. Available at : <https://arxiv.org/abs/2110.04126> (Accessed October 8, 2021).
- Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. (2020). State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* 11, 5575. doi:10.1038/s41467-020-19266-y
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. available at : <https://arxiv.org/abs/1710.10903> (Accessed October 30, 2017).
- Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. (2019). “Smiles-bert,” in Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, New York City, NY, USA, September 2019, 429–436.
- Wang, S. H., Pillai, H. S., Wang, S., Achenie, L. E. K., and Xin, H. (2021). Infusing theory into deep learning for interpretable reactivity prediction. *Nat. Commun.* 12, 5288. doi:10.1038/s41467-021-25639-8
- Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., et al. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today Technol.* 37, 1–12. doi:10.1016/j.ddtec.2020.11.009
- 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
- Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., et al. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* 63, 8749–8760. doi:10.1021/acs.jmedchem.9b00959
- Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? Available at : <https://arxiv.org/abs/1810.00826> (Accessed October 1, 2018).
- 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, 3370–3388. doi:10.1021/acs.jcim.9b00237
- Zhang, B., Zhang, X., Du, W., Song, Z., Zhang, G., Zhang, G., et al. (2022). Chemistry-informed molecular graph as reaction descriptor for machine-learned retrosynthesis planning. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2212711119. doi:10.1073/pnas.2212711119
- Zhang, X. C., Wu, C. K., Yang, Z. J., Wu, Z. X., Yi, J. C., Hsieh, C. Y., et al. (2021). MG-BERT: Leveraging unsupervised atomic representation learning for molecular property prediction. *Brief. Bioinform.* 22, bbab152. doi:10.1093/bib/bbab152



OPEN ACCESS

EDITED BY

Cy Jeffries,
European Molecular Biology Laboratory
Hamburg, Germany

REVIEWED BY

Federico Forneris,
University of Pavia, Italy
Tomas Klumpler,
Masaryk University, Czechia
Natacha Rochel,
INSERM U964 Institut de Génétique et de
Biologie Moléculaire et Cellulaire (IGBMC),
France

*CORRESPONDENCE

Eric di Luccio,
✉ e.diluccio@hbio.jp
Rocco Caliandro,
✉ rocco.caliandro@cnr.it

†PRESENT ADDRESS

Eric di Luccio, Hirotsu Bio Science Inc., Tokyo,
Japan

†These authors have contributed equally to
this work

RECEIVED 21 March 2023

ACCEPTED 19 February 2024

PUBLISHED 07 March 2024

CITATION

Belviso BD, Shen Y, Carrozzini B, Morishita M,
di Luccio E and Caliandro R (2024), Structural
insights into the C-terminus of the histone-
lysine N-methyltransferase NSD3 by small-
angle X-ray scattering.
Front. Mol. Biosci. 11:1191246.
doi: 10.3389/fmolb.2024.1191246

COPYRIGHT

© 2024 Belviso, Shen, Carrozzini, Morishita, di
Luccio and Caliandro. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).
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.

Structural insights into the C-terminus of the histone-lysine N-methyltransferase NSD3 by small-angle X-ray scattering

Benny Danilo Belviso^{1†}, Yunpeng Shen^{2†}, Benedetta Carrozzini¹,
Masayo Morishita³, Eric di Luccio^{3*†} and Rocco Caliandro^{1*}

¹Institute of Crystallography, CNR, Bari, Italy, ²Department of Biotechnology, School of Biological Engineering, Henan University of Technology, Zhengzhou, Henan, China, ³Department of Genetic Engineering, School of Life Sciences, College of Natural Sciences, Kyungpook National University, Daegu, Republic of Korea

NSD3 is a member of six H3K36-specific histone lysine methyltransferases in metazoans. Its overexpression or mutation is implicated in developmental defects and oncogenesis. Aside from the well-characterized catalytic SET domain, NSD3 has multiple clinically relevant potential chromatin-binding motifs, such as the proline–tryptophan–tryptophan–proline (PWWP), the plant homeodomain (PHD), and the adjacent Cys-His-rich domain located at the C-terminus. The crystal structure of the individual domains is available, and this structural knowledge has allowed the designing of potential inhibitors, but the intrinsic flexibility of larger constructs has hindered the characterization of mutual domain conformations. Here, we report the first structural characterization of the NSD3 C-terminal region comprising the PWWP2, SET, and PHD4 domains, which has been achieved at a low resolution in solution by small-angle X-ray scattering (SAXS) data on two multiple-domain NSD3 constructs complemented with size-exclusion chromatography and advanced computational modeling. Structural models predicted by machine learning have been validated in direct space, by comparison with the SAXS-derived molecular envelope, and in reciprocal space, by reproducing the experimental SAXS profile. Selected models have been refined by SAXS-restrained molecular dynamics. This study shows how SAXS data can be used with advanced computational modeling techniques to achieve a detailed structural characterization and sheds light on how NSD3 domains are interconnected in the C-terminus.

KEYWORDS

nuclear receptor-binding SET domain protein 3, small-angle X-ray scattering, computational modeling, epigenetic cancer therapy, molecular dynamics

1 Introduction

Nuclear receptor-binding SET domain (NSD) proteins are three protein lysine methyltransferases that are predominantly mono- and di-methylate lysine 36 of histone 3 (H3K36) (Kuo et al., 2011). They are called NSD1, NSD2 (also known as WHSC1 or MMSET), and NSD3 (also known as WHSC1L1) and are critical in maintaining chromatin integrity. Their overexpression or mutation is implicated in developmental defects and oncogenesis. In addition, the dysfunction of their methylation activity results in epigenomic

aberrations, which are relevant for oncogenesis. Thus, reducing NSD activity through specific lysine-HMTase inhibitors appears promising for epigenetic cancer therapy (Vougiouklakis et al., 2015).

NSD2 is an oncoprotein that is aberrantly expressed, amplified, or somatically mutated in multiple types of cancer (Vougiouklakis et al., 2015). Notably, the t (4; 14) NSD2 translocation in multiple myeloma and the hyperactive NSD2 mutation E1099K in a subset of pediatric acute lymphoblastic leukemia result in altered chromatin methylation that drives oncogenesis (Keats et al., 2003; Jaffe et al., 2013). NSD3 is involved in several varieties of cancers as it contributes to tumorigenesis by interacting with the bromodomain-containing protein 4 (BRD4) and the bromodomain and extra-terminal (BET) protein, which are potential therapeutic targets in acute myeloid leukemia (Han et al., 2018).

NSD2 and NSD3 have multiple protein–protein interaction domains that may be clinically relevant and arranged in a conserved sequence that contains two proline–tryptophan–tryptophan–proline (PWWP) domains, which are assumed to be critical for binding to methylated H3-histone and the DNA molecule, four plant homeodomains (PHDs)—which appear essential for interactions with other methylated histones—an associated with SET (AWS) domain, a catalytic SET domain, and a post-SET domain—including a Cys-His-rich region (C5HCH) (Angrand et al., 2001).

The first PWWP domain (PWWP1) of NSD2 binds *in vitro* H3K36me₂, presumably through a conserved aromatic cage composed of three orthogonally positioned aromatic side chains (Y233, W236, and F266) that can engage in cation– π and hydrophobic interactions with the ammonium group of the methylated lysine (Qin and Min, 2014). However, the contribution of the PWWP domains and the role in histone methylation of the aromatic residues in the cage mentioned above is not established yet. For example, the F266A mutation at the aromatic cage, known to inhibit cancer proliferation, appears to affect chromatin/NSD2 binding without significantly affecting H3K36 dimethylation (Sankaran et al., 2016). Studies have revealed that AWS, SET, and post-SET domains also play a critical role in recognizing and methylating molecular targets of histones H3 and H4 *in vitro*, particularly in the case of NSD3 (Morishita et al., 2014).

High-resolution structural knowledge of individual domains from X-ray crystallography is available for NSD2 and NSD3 and has been used to design small-molecule inhibitors. The crystal structure of the SET domain supported the design and characterization of N-alkyl sinefungin derivatives for NSD2 (Tisi et al., 2016) and a norleucine-containing inhibitor peptide derived from the histone H4 sequence for NSD3 (Morrison et al., 2018). The crystal structure of the NSD2-PWWP1 enabled both the discovery of a small-molecule antagonist with a K_d of 3.4 μ M, which abrogates histones containing H3K36me₂ binding in cells (de Freitas et al., 2021), and the characterization of its interactions with methylated histone peptides and dsDNA (Zhang et al., 2021). Moreover, the crystal structure of PWWP1 of NSD3 allowed a fragment-based discovery of a potent, selective, and cellular active antagonist (Bottcher et al., 2019). Binding assay studies of the region, including the PHD closest to the C-terminus and the C5HCH motif of the NSD3, along with the crystal structure of such

regions, revealed a histone-binding specificity of the PHD domain between the three members of the NSD family (He et al., 2013). Recently, cryo-electron microscopy has made available structures of the SET domain for NSD2 and NSD3 bound to mononucleosomes (Li et al., 2021; Sato et al., 2021), thus providing molecular insights into nucleosome-based recognition and histone-modification mechanisms.

Although both NSD2 and NSD3 are attractive therapeutic targets, efforts to target their domains with small-molecule inhibitors have so far met with little success (Morishita et al., 2017; Shen et al., 2019). On the other hand, drug design initiatives targeting NSD2 and NSD3 have been severely hampered by the lack of structural knowledge about mutual interactions between domains. The high-resolution structure of NSD2 or NSD3 constructs comprising PWWP, SET, and PHD domains is still missing, likely due to the high flexibility of these proteins that make them recalcitrant to obtain good-quality crystals for the structural solution by X-ray diffraction.

Here, we present the first structural investigation of the C-terminal region of NSD3, comprising the second PWWP domain (PWWP2), the SET domain, and the PHD closest to the C-terminus (PHD4), in solution, determined by small-angle X-ray scattering (SAXS) combined with advanced computational modeling. In particular, the molecular envelope determinations from SAXS data were complemented with structural predictions based on artificial intelligence, which is in line with a recent trend in the field of SAXS data analysis (Receveur-Bréchet, 2023), and with a molecular dynamics flexible-fitting approach, which has recently proven effective even for highly flexible proteins (Belviso et al., 2022). The mutual conformation of interacting domains in solution, thus not affected by the typical artifacts due to sample preparation for X-ray diffraction and cryo-EM, i.e., crystal packing or vitrification effects, respectively, was disclosed.

2 Materials and methods

2.1 NSD3 construct expression and purification

Two constructs for the C-terminal region of the NSD3 (UniProt code Q9BZ95) protein were designed: the first including PWWP2, AWS, SET, and PostSET domains, comprising residues from 942 to 1,318, and named NSD3-PWWP2-SET, and the second including AWS, SET, PostSET, and PHD4 domains, comprising residues from 1,070 to 1,423, and named NSD3-SET-PHD4. The conformed pTYB12-NSD construct plasmids were transformed into *Escherichia coli* BL21 (DE3) cells. The culture was incubated in an LB medium containing 100 mg/L ampicillin at 37°C, 180 rpm, until OD₆₀₀ reached around 0.6. Then, 125 μ M isopropyl 1-thio-D-galactopyranoside (IPTG) was added to induce the recombinant expression of the target construct proteins for 16 h at 12°C. Cells were harvested and frozen at –80°C for 2 h minimum. The frozen cells were re-suspended and lysed (shaking) for 30 min in IMPACT buffer (500 mM NaCl, 20 mM Tris pH 8.0, and 0.1 mM EDTA) with 0.1% Triton and 10 mM phenylmethanesulfonyl fluoride (PMSF), followed by 20 cycles of sonication (2.5 min at 85 Amp) on ice. After removing the cell debris, the lysate containing CBD (chitin-binding

domain)-intein-target protein was loaded onto a chitin resin column and then flashed with 1 L IMPACT buffer with 0.1% Triton X-100 (45–60 column volumes) to remove other proteins and impurities and 0.5 L IMPACT buffer (25–30 column volumes) to remove the detergent Triton. Cleavage of the intein tag was induced by incubation in IMPACT buffer supplemented with 50 mM 2-mercaptoethanol at 4°C for 40 h. The pure target protein was eluted in 65 mL IMPACT buffer, concentrated, and washed with IMPACT buffer using 10K Amicon Ultra centrifugal filters.

2.2 SAXS measurements

Small-angle X-ray scattering (SAXS) measurements were performed at the beamline B21 of the Diamond Light Source (Didcot, UK), a beamline devoted to bioSAXS measurements and equipped with an EIGER 4M detector (Dectris) and in-line size-exclusion chromatography (SEC-SAXS). Protein samples were buffer exchanged against 0.5 M NaCl, 20 mM Tris-HCl (pH 8.5), and 5 mM DTT using an Amicon-4 Centrifugation Unit (cutoff 10 kDa) and concentrated up to 4.3 mg/mL just before data collection to avoid sample aggregation and/or degradation. The protein concentration was determined using a NanoDrop spectrophotometer Thermo 2000c. For both constructs, the extinction coefficient (ϵ) and molecular weight (MW) were calculated by the Expasy ProtParam server (Gasteiger et al., 2005) based on their sequence (Supplementary Table S1). SEC-SAXS data collections were performed at 20°C by loading 50 μ L of the sample onto a 4.6-mL high-performance Shodex 403 chromatographic column (10–700 kDa MW resolution range) connected to an Agilent 1200 HPLC system (Waters) and equilibrated with the same buffer as that used for the buffer-exchange step. Three different sample concentrations were loaded on the column (0.6, 1.6, and 3.8 mg/mL in the case of NSD3-PWWP2-SET and 1.0, 1.6, and 4.3 mg/mL in the case of NSD3-SET-PHD4), each prepared by diluting the protein stock solutions concentrated at 4.3 mg/mL. For such measurements, the integration time per frame was set to 3 s, and data were collected in the range of momentum transfer (q) from 0.0026 to 0.340 \AA^{-1} .

2.3 SAXS data analysis

Raw SAXS 2D images were processed by the DAWN processing pipeline (Wilhelm et al., 2027) to produce normalized and radially integrated SAXS curves. They were processed by SCATTER (Rambo, 2017) to yield chromatograms and R_g value estimates. Background subtraction and Guinier analysis were performed by the program PRIMUS of the ATSAS package (Manalastas-Cantos et al., 2021). The FIND_Dmax tool of SCATTER was used with the default parameters (suggested D_{max} and alpha ranges, Moore model, and usage of background information for $P(r)$ determination) to estimate the best value of the maximum momentum transfer q -value (q_{max}) to be used in data analysis (Tully et al., 2021). Original SAXS profiles were re-binned using the DATREGRID command of ATSAS to improve their signal-to-noise ratio and then to increase the q_{max} values.

The particle distance distribution function $P(r)$ was determined using GNOM (Svergun, 1992) in the q -value range from the beginning of the Guinier region to q_{max} (Supplementary Table S2). The AMBIMETER program (Petroukhov and Svergun, 2015) was used to determine the number of shape topologies compatible with the $P(r)$ curves and predict the uniqueness of the *ab initio* reconstructions.

Ab initio molecular envelope determination was performed on the best dataset for each construct, selected according to the values of q_{max} and the quality of the $P(r)$ profile. A total of 20 models of the molecular envelope were generated for each dataset using the annealing procedure in the fast mode of the DAMMIF program (Franke and Svergun, 2009). They were spatially aligned based on the normalized spatial discrepancy calculated by the SUPCOMB program (Kozin and Svergun, 2001) and subsequently averaged, bead occupancy-weighted, and volume-corrected using DAMAVER (Volkov and Svergun, 2003). Additional refinement to the SAXS data using DAMMIN/DAMSTART in the slow mode (Svergun, 1999) was performed to generate a final dummy-atom representation of the shape and volume of each protein. The protein molecular mass was estimated from SAXS data using the consensus Bayesian assessment (Hajizadeh et al., 2018) implemented in the program PRIMUS.

2.4 Homology modeling

Homology modeling was performed following two strategies using SAXS data as the lever arm to adjust the structural predictions. In the first strategy, which follows a bottom-up approach, individual domains were independently generated and assembled *a posteriori* based on the agreement with SAXS data. Homology models of the following domains/regions belonging to the C-terminal region of NSD3 were generated by the Phyre2 server (Kelley et al., 2015): the core of the PWWP2 domain (942–1,025); the link connecting domains PWWP2 and SET (1,026–1,056); the region containing AWS, SET, and postSET (1,070–1,318); the core of the SET domain (1,070–1,289); the link connecting the SET and PHD4 domains (1,290–1,310); and the PHD4 domain (1,319–1,423). These models were manually placed into molecular envelopes calculated from the SEC-SAXS datasets to obtain starting models for rigid body fitting that has been performed by SASREF (Petoukhov and Svergun, 2005). In the second strategy, a structural prediction of the whole C-terminal region from PWWP2 to PHD4 was performed, following a top-down approach that ensures compatible modeling of the NSD3-PWWP2-SET and the NSD3-SET-PHD4 constructs. In the first instance, the AlphaFold prediction about the whole NSD3 protein was downloaded from the AlphaFold protein structure database (Jumper et al., 2021), entry n. Q9BZ95, the fourth version of the model, was considered. In the second instance, ColabFold (Mirdita et al., 2022), RaptorX (Xu et al., 2021), and I-Tasser (Zheng et al., 2021) servers were used as the predictors, each supplying the five most probable structural models. They all make use of a machine learning approach; specifically, the first combines the fast homology search of MMseqs2 (Steinegger and Söding, 2017) with AlphaFold2 (Jumper et al., 2021) or RoseTTAFold (Baek et al., 2021), the second integrates deep learning and co-evolutionary analysis by means of convolutional

residual neural networks, and the third combines contact maps from deep neural network learning with fragment assembly simulations. A mixed-strategy approach was also followed, where individual domains extracted from the AlphaFold prediction were used for SAXS-based rigid body modeling performed by the program CORAL (Petoukhov et al., 2012).

The quality of structural predictions has been assessed by comparison with SAXS data: each predicted model has been split in an NSD3-PWWP2-SET and NSD3-SET-PHD4 part, which has been separately fitted with SAXS data both in reciprocal and direct space. The validation parameter of the model in reciprocal space is the χ^2 of the least-square fit with raw SAXS data, as determined by the CRYSOLO program (Svergun et al., 1995), and that in direct space is the normalized spatial discrepancy with respect to the molecular envelope determined *ab initio* from SAXS data. This latter quantifier tends to be 0 for similar objects, is less than 1 among different DAMMIF/N model reconstructions of the same SAXS dataset, and is expected to be less than 3 when comparing SAXS-derived dummy-atom models with full-atom atomic models.

2.5 SAXS-driven optimization of structural models

The best-quality homology modeling models were subjected to molecular dynamics (MD) restrained by the SAXS-derived molecular envelope using the molecular dynamics flexible fitting (MDFF) tool (Trabuco et al., 2008), which implements the fitting of flexible atomic structures into a density map. The molecular envelopes determined by SEC-SAXS data were used as reference density maps, from which external potentials were generated and added to molecular dynamics. Simulations were performed by NAMD (NANOSCALE MOLECULAR DYNAMICS) (Phillips et al., 2020), and simulated data were analyzed by VMD (visual molecular dynamics) (Humphrey et al., 1996). MD simulations were run with an explicit solvent. Long-range electrostatic interactions were treated with the particle-mesh Ewald method (Darden et al., 1993). A 1.0 nm cutoff was used for van der Waals interactions and the real-space part of the electrostatic interactions. All bond lengths were constrained with the LINCS algorithm, and the time step was set to 1 fs. MDFF simulations were run with an implicit solvent, while targeted molecular dynamics (TMD) was used to maintain the internal consistency of the PWWP, SET, and PHD4 domains with respect to their experimental structures. Both the values of the dielectric constant and the scaling factor of the MD external potential generated from the SAXS density map were fine-tuned by optimizing the *a posteriori* agreement of the MD models with SAXS data. They were finally set to 100 and 0.08, respectively.

MDFF simulations were monitored by calculating the cross-correlation coefficient (CORR) between the target density map and each frame of the MDFF trajectory and the root-mean-square deviation of the C $_{\alpha}$ atoms (RMSD) for the initial structural model. The structural models were prepared for MD by setting the histidine protonation state to that expected at the pH used in the SAXS data collection (8.5), as predicted by the H++ server (Anandakrishnan et al., 2012), by adding Zn ions guided by their

positions in the experimental models (four of them were positioned in the zinc-finger domain PHD4 and three in the SET domain) and deprotonating the closest cysteine residues to form expected S–S bonds. The metal coordination in the seven Zn sites was restrained using the NAMD extraBonds command, with a spring constant of 50 kcal/mol and a reference distance of 2.5 Å from Cys S or His N atoms.

MD trajectories were analyzed by extracting the region's NSD3-PWWP2-SET and NSD3-SET-PHD4 from each frame and separately fitting them against SAXS data.

The structural models were compared using a descriptor based on the backbone dihedral angles. It is named the protein angular value (PAV) (Liuzzi et al., 2017), which is defined as follows:

$$PAV_i = \frac{180}{\pi} \cos^{-1}(\cos(\psi_i + \phi_i)), \quad (1)$$

where ψ_i and ϕ_i are the backbone dihedral angles of the *i*th residue. The PAV values range between 0° and 180° and represent the $\psi_i + \phi_i$ values expressed in degrees. Equation 1 avoids the problem of range definition connected with the circular nature of the angular variables. PAV profiles of each structure were calculated through the script TPAD (Caliandro et al., 2012) run on VMD (Humphrey et al., 1996). PAV profiles from different structures were separately analyzed using principal component analysis (PCA) and hierarchical clustering implemented in the program RootProf (Caliandro and Belviso, 2014).

Details about SAXS samples, data collection, analysis, and 3D modeling are summarized in [Supplementary Table S2](#).

3 Results

3.1 Analysis of the SEC-SAXS data

SEC-SAXS analyzed NSD3-PWWP2-SET and NSD3-SET-PHD4 constructs at a concentration of protein loaded in the column of 0.6, 1.6, and 3.8 mg/mL for NSD3-PWWP2-SET and 1.0, 1.6, and 4.3 mg/mL for NSD3-SET-PHD4. A whitish precipitate appeared at higher protein concentrations, suggesting the onset of protein aggregation effects. SEC profiles and radius of gyration per frame (R_g) are shown in [Figures 1A and B](#). The presence of two peaks characterizes both SEC profiles, hereinafter named p1 (the peak at lower elution time) and p2 (the peak at higher elution time). SEC also shows a shoulder of p1 (at a lower elution time than the peak) for each dataset, which is particularly evident in the case of NSD3-PWWP2-SET ([Figure 1A](#)). However, a visual inspection of the R_g values suggests that only the p2 peak of both constructs is related to a homogeneous species and, therefore, is the only region of the chromatogram that is suitable for data analysis.

Frames under the p2 peak were selected for averaging using the standard deviation of the R_g values ($\sigma_{<R_g>}$ in [Supplementary Table S3](#)). For each construct and protein concentration, we chose a set of adjacent frames that minimizes $\sigma_{<R_g>}$ while keeping the number of frames as high as possible. The similarity among datasets of the same construct has been assessed by a reduced χ^2 statistic test, which showed that all datasets of the same construct are compatible with the same distribution (each pair of datasets shows a calculated *p*-value higher than a significance level $\alpha = 0.01$ in

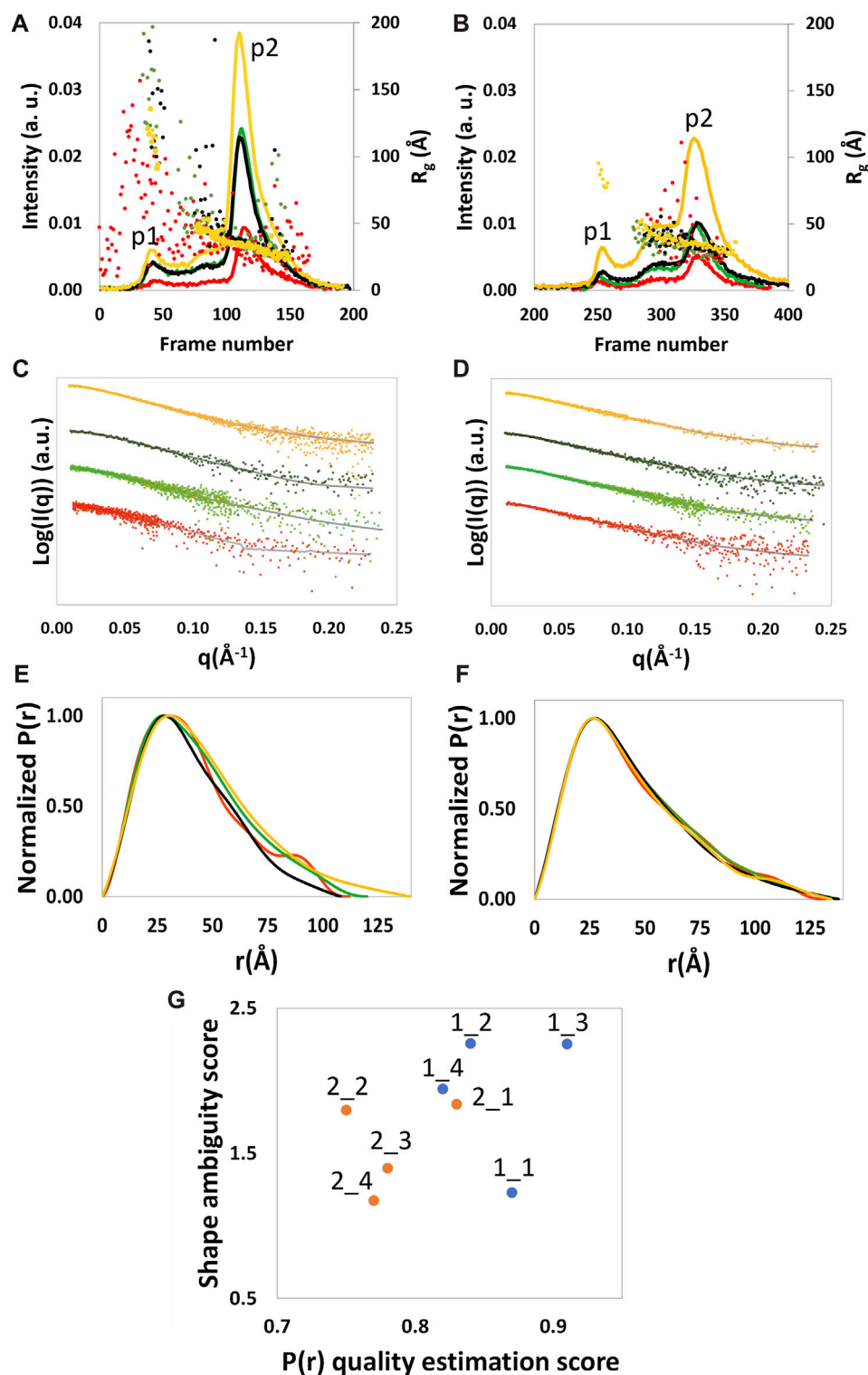


FIGURE 1

SEC profile and the R_g calculated by SCATTER for each frame (A, B), experimental (dots) and calculated from the reciprocal space fit of $P(r)$ to the data (full gray line) scattering intensity, with a scaled off set applied for presentation purposes (C, D), and $P(r)$ functions (E, F) are shown for NSD3-PWWP2-SET (first column) and NSD3-SET-PHD4 (second column) constructs. Red and yellow colors are used, respectively, for the samples at the lowest and higher protein concentrations, and green and black colors are used for the samples at 1.6 mg/mL. Correlation plot (G) between the shape ambiguity score, related to the number of shape topologies compatible with a given $P(r)$ curve (vertical axis), and the quality score of the $P(r)$ fit (horizontal axis). Points related to NSD3-PWWP2-SET and NSD3-SET-PHD4 are represented in orange and cyan color, respectively. Optimal values correspond to lower shape ambiguity (values lower than 1 correspond to potentially unique 3D reconstructions) and a higher quality score of the $P(r)$ fit (maximum value is 1). Datasets 1_3, 1_4, 2_1, and 2_3 were linearly rebinned, and the others were log-rebinned.

TABLE 1 Data and model parameters estimated for each dataset collected in the SEC-SAXS mode for NSD3-PWWP2-SET and NSD3-SET-PHD4 constructs. Protein concentration, maximum momentum transfer (q_{\max}) estimated before and after re-binning the data, radius of gyration (R_g) from Guinier analysis (reciprocal space), $P(r)$ function determination (real space), maximum inter-particle distance (D_{\max}), and molecular weight (MW) are shown.

Construct	ID	Protein concentration (mg/mL)	q_{\max} (\AA^{-1})	After re-binning				
				q_{\max} (\AA^{-1})	R_g (\AA) reciprocal space	R_g (\AA) direct space	D_{\max} (\AA)	MW (kDa)
NSD3-PWWP2-SET	1_4	3.8	0.16	0.23	36.7	36.8	139.4	53.1
	1_3	1.6	0.21	0.23	31.3	31.4	108.0	50.8
	1_2	1.6	0.13	0.24	34.1	34.2	120.0	48.7
	1_1	0.6	0.23	0.23	33.2	33.3	112.0	43.7
NSD3-SET-PHD4	2_4	4.3	0.20	0.29	36.2	36.3	134.6	42.8
	2_3	1.6	0.16	0.29	35.8	36.0	138.0	40.2
	2_2	1.6	0.14	0.30	36.7	36.8	137.0	40.2
	2_1	1.0	0.18	0.29	36.5	36.6	132.0	41.9

Supplementary Figure S1). The lowest p -values (still higher than 0.01) were found while comparing the datasets at the lowest and highest protein concentrations, suggesting a lower probability that these datasets are comparable with each other concerning the other cases.

The Guinier analysis provided R_g values (in the reciprocal space) ranging from 31 to 34 \AA for NSD3-PWWP2-SET and from 33 to 35 \AA for NSD3-SET-PHD4 (**Table 1**). Regarding the maximum momentum transfer at which SAXS data analysis can be performed (q_{\max}), it is expected that its values increase with the protein concentration as a consequence of a higher signal-to-noise ratio. However, we found a non-negligible correlation only in the case of the NSD3-SET-PHD4 construct (Pearson coefficient = 0.6) (**Table 1**).

Given the limited resolution of available data (**Table 1**), we re-binned the SAXS profiles by reducing the number of points on a linear or a log scale in q . The degree of data reduction was optimized for each dataset based on the new value of q_{\max} and the quality of the pair distance distribution function $P(r)$ obtained. An example of the dependence of q_{\max} on the number of points is given in **Supplementary Figure S2**. The re-binned profiles are shown in **Figures 1C and D**, and the corresponding $P(r)$ curves were calculated for each dataset by selecting a range from the beginning of the Guinier region to q_{\max} (**Figures 1E and F**). The related geometrical parameters (R_g direct space and D_{\max} in **Table 1**) confirmed the slightly smaller dimensions of the NSD3-PWWP2-SET construct with respect to the NSD3-SET-PHD4 one. A good agreement between real and reciprocal R_g values is present for each dataset. The molecular weight values estimated in **Table 1** are in fair agreement with those expected based on the primary sequence (42.5 and 44.5 kDa for NSD3-SET-PHD4 and NSD3-PWWP2-SET, respectively).

3.1.1 Dataset selection

Dataset selection has been performed using the quality of the $P(r)$ function determination, which was assessed by considering the quality estimation score supplied by GNOM and the shape

ambiguity score supplied by the AMBIMETER program (Petoukhov and Svergun, 2015), which is related to the number of shape topologies compatible with a given $P(r)$ curve (**Figure 1G**). Their values indicate that datasets 1_1 and 2_1 are the best ones for the NSD3-PWWP2-SET and NSD3-SET-PHD4 constructs, respectively, since their representative points in the scatter plot of **Figure 1G** are in the region of the lowest shape ambiguity and higher fit quality. In particular, dataset 1_1 has a very low shape ambiguity score (0.82), indicating a unique *ab initio* 3D reconstruction. In contrast, dataset 2_1 has a very high fit quality (0.84, the maximum is 1), indicating a reliable estimate of the pair distribution function. Both the selected datasets correspond to samples with lower protein concentrations. They have been obtained by re-binning the SAXS profiles on a log-scale to 800 points (1_1) or joining every third point (2_1). Further indications that corroborate this choice are the following: dataset 2_1 has a lower difference between direct and reciprocal R_g values, and dataset 1_1 shows the lowest difference between the estimated molecular weight (43.7 kDa) and the expected one (44.5 kDa) (**Table 1**). From **Figure 1G**, it can be noted that representative points of NSD3-SET-PHD4 have a systematically lower $P(r)$ quality estimation score than those of NSD3-PWWP2-SET.

3.1.2 Molecular envelope determination

The molecular envelopes determined for each dataset are shown in **Figure 2** for each dataset. They have a similar elongated shape for both constructs, apart from datasets 1_2 and 1_3, for which the superposition of the 20 envelopes calculated by DAMMIF was not optimal, which is in agreement with the fact that these datasets have the highest shape ambiguity scores (**Figure 1G**).

The selected SAXS data relative to the NSD3-PWWP2-SET and NSD3-SET-PHD4 samples (datasets 1_1 and 2_1, respectively) have been deposited in the SASBDB database (Kikhney et al., 2020) in entries n. SASDNL8 and SASDNK8, respectively. All individual models and fits of the molecular envelope are available in these entries as additional information.

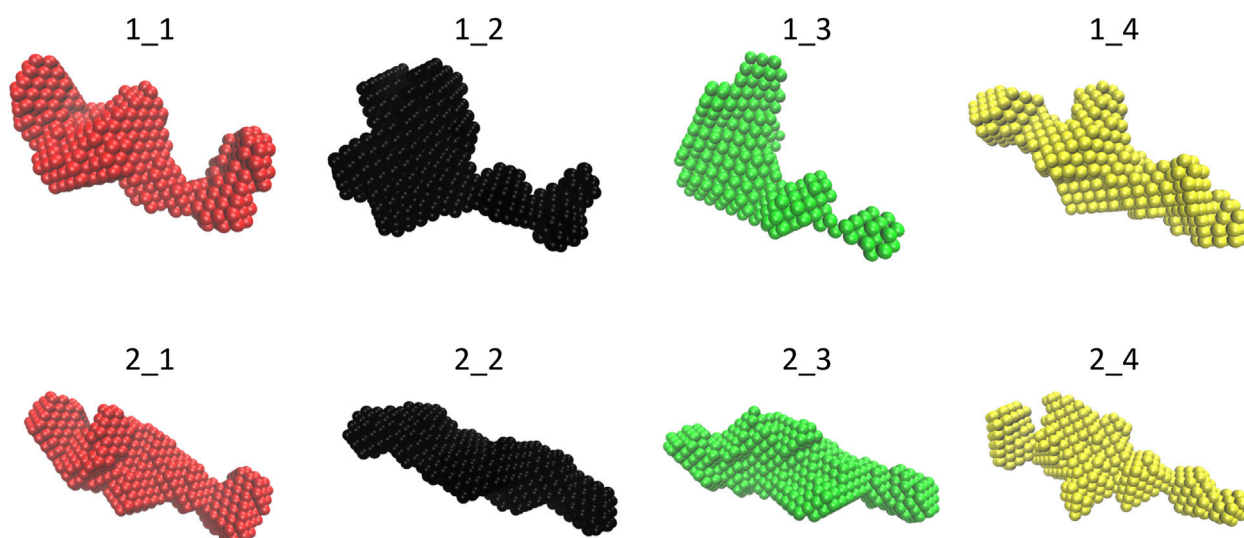


FIGURE 2
Final molecular envelope models for SEC-SAXS datasets of NSD3-PWWP2-SET (first row) and NSD3-SET-PHD4 (second row) constructs. The color code is the same as of [Figure 1](#).

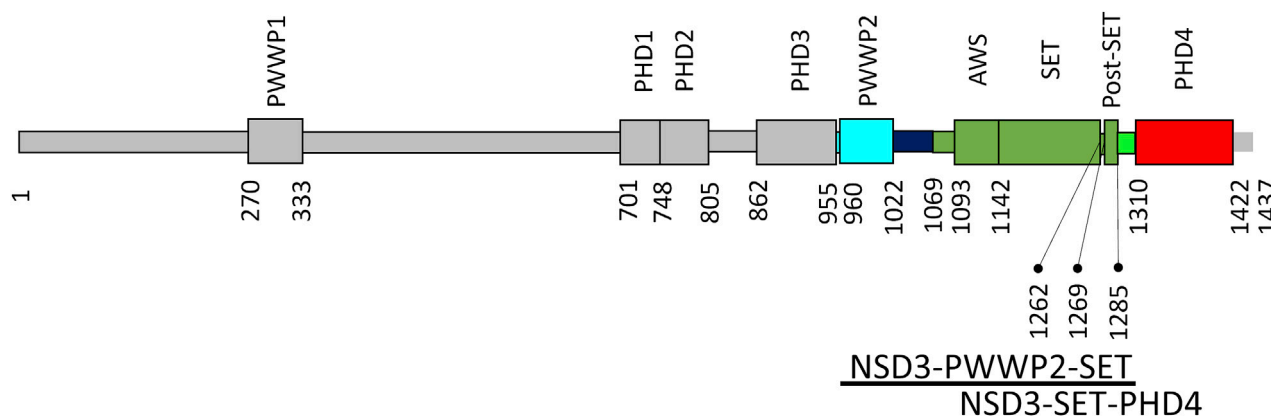


FIGURE 3
NSD3 domain organization. Domains are shown as rectangles, and those of interest for this work are represented in cyan (PWWP2), green (AWS, SET, and postSET), and red (PHD4). The numbers of residues delimiting the domains are reported together with the range of residues covered by the two constructs under investigation (bottom).

3.2 Structural modeling

3.2.1 Homology modeling and validation

[Figure 3](#) shows the domain organization of the whole NSD3 protein. In such a figure, the regions used in the homology modeling processes exploited in this work are colored in cyan (PWWP2), green (AWS, SET, and postSET), and red (PHD4). The models produced by the top-down modeling strategy (the one based on the entire sequence from PWWP2 to PHD4) are shown in [Supplementary Figures S3A–D](#). Peculiar differences can be observed among the models as follows: the AlphaFold model shows the highest content of secondary structure elements ([Supplementary Figure S3A](#)), the ColabFold

models constantly maintain the orientation of the PWWP2-SET and SET-PHD4 linkers concerning the SET domain ([Supplementary Figure S3B](#)), the I-Tasser models show a compact arrangement of individual domains and their linkers ([Supplementary Figure S3D](#)), and RaptorX provides a large variability in the orientation of PWWP and PHD4 domains with respect to the SET domain ([Supplementary Figure S3C](#)).

In the case of the bottom-up strategy, individual domains generated by Phyre2 ([Supplementary Figure S4](#)) have been used to build NSD3 models able to fit the envelopes of selected SAXS datasets, i.e. 1_1, related to NSD3-PWWP-SET injected at 0.6 mg/mL, and 2_1, related to NSD3-SET-PHD4 injected at 1.0 mg/L. Although such a strategy allows using SAXS data from an early stage,

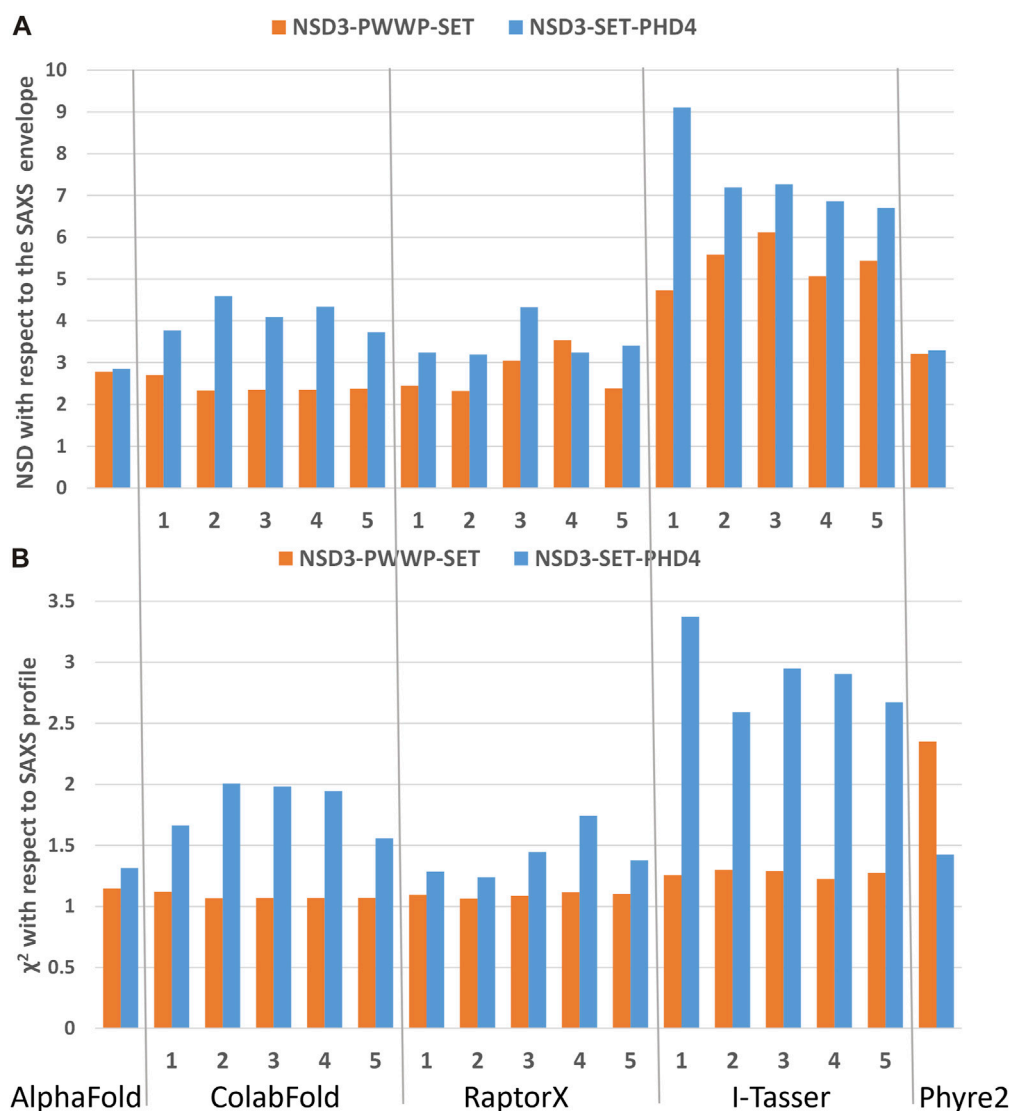


FIGURE 4

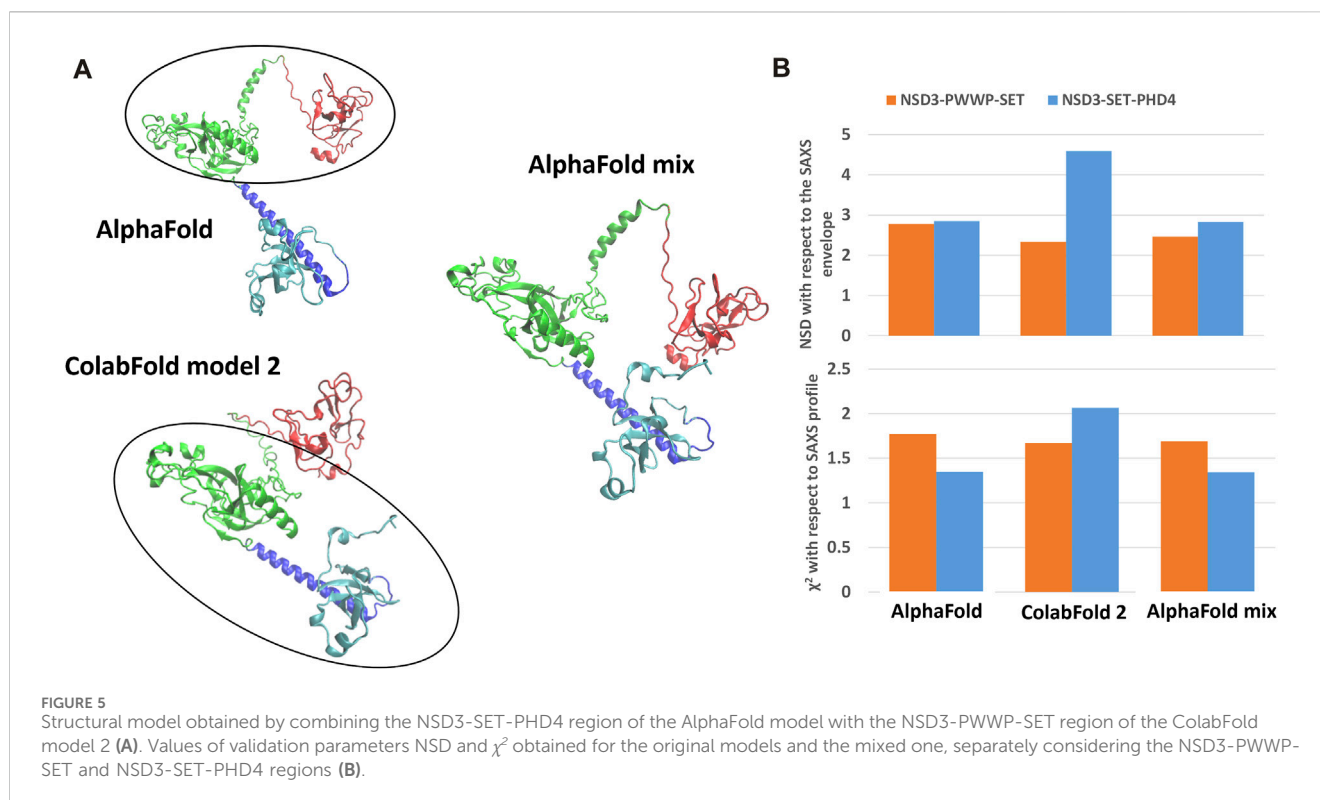
Validation of the homology models by means of SAXS data on the dataset 1_1 (NSD3-PWWP2-SET injected at 0.6 mg/mL) and 2_1 (NSD3-SET-PHD4 injected at 1.0 mg/mL) in reciprocal (A) and direct (B) spaces. Predictions of web servers AlphaFold, ColabFold, RaptorX, I-Tasser, and Phyre2 have been assessed by fitting them with SAXS data in reciprocal space (A) and by measuring their normalized spatial discrepancy (NSD) with respect to the corresponding SAXS molecular envelopes in direct space (B). The validation parameters NSD and χ^2 obtained for each generated model are shown.

it has the drawback that it does not guarantee the overlap between the common region between the two NSD3 constructs, as occurred in the case of top-down modeling (Supplementary Figure S3E).

The best predictions resulting from such modeling processes (those showing the lowest χ^2 against SAXS data and normalized spatial discrepancy values against the SAXS envelope) for both constructs have been obtained for the AlphaFold model (Figure 4). Second, there are the first two models generated by RaptorX, which mainly differ in how the linkers are structured and in the plane in which they interact (they are rotated by about 90°, as shown in Supplementary Figure S3C). The compact configuration of the I-Tasser models is a systematic disagreement with SAXS. Considering the two constructs in Figure 4 separately, it is worth noting that NSD3-SET-PHD4 has the lowest χ^2 and NSD scores for

AlphaFold, while NSD3-PWWP2-SET is best modeled by the ColabFold model 2. Based on this evidence, we have created an AlphaFold mixed model by combining the best regions from the two models, considering the common region among the two constructs as the lever arm for the superposition (Figure 5A). Notably, this operation brings the PWWP and PHD4 domains close to each other, although they were far away in the two starting models. As expected, the validation parameters of the so-obtained mixed model are improved with respect to the original models (Figure 5B).

A further approach to generate an atomistic model of the NSD3 C-terminal involved the use of CORAL to place individual domains, as predicted by AlphaFold, guided by the agreement with the SAXS profile. This procedure is heavily influenced by the choice of even loose restraints about contacting residues. The best model



obtained by combining the results of the procedure applied separately to the NSD3-PWWP2-SET and NSD3-SET-PHD4 regions is shown in [Supplementary Figure S5](#) together with the related validation parameters.

3.2.2 Optimization of the best models against SAXS data

The best homology model was refined against SAXS data by making them flexible through molecular dynamics (MD). Experimental data were included in the simulation using the technique known as molecular dynamics flexible fitting (MDFF), where the MD is restrained by the experimental molecular envelope, which represents an additional potential that drives the simulation. An additional restraint from high-resolution data from X-ray diffraction or NMR was introduced using the targeted molecular dynamics approach, which was applied to the PWWP, SET, and PHD4 domains, considering their respective experimental structures as targets. The SAXS restraints were not applied separately to NSD3-PWWP2-SET and NSD3-SET-PHD4 regions since this would have led to final models of the two regions that are not compatible with each other and would have involved performing MD on partial models, leading to approximate results. Instead, the SAXS restraints were applied by overlapping them on the initial conformation of the homology model. In this way, the two experimental envelopes of the two constructs were combined to form a unique restraint that can be used for local optimization of the whole homology model driven by MD.

The MDFF procedure was applied to the AlphaFold mixed model, which showed the best validation parameters among the full-atom models generated. For comparison, it was also applied to the AlphaFold model and the RaptorX model 1 (the latter was

preferred to the RaptorX model 2, which shows a similar mutual positioning of the PWWP and PHD4 domains because it holds a more structured linker between PWWP and SET). Instead, it was not possible to apply the MDFF procedure to the CORAL model due to the incomplete modeling of their linkers.

Results of the MDFF optimization of the AlphaFold mixed model are reported in [Figure 6](#), where the model conformations before and after the MDFF run are shown together with the experimental molecular envelopes applied as a restraint during the simulation. The initial model partially covered by the envelope ([Figure 6A](#)) is well-fitted within it at the end of the simulation ([Figure 6B](#)), where the biggest variations concern the linker between SET and PHD4. As a result, the cross-correlation coefficient between the experimental and calculated envelopes (CORR) and the mean C_α deviation with respect to the initial model (RMSD) both increase during the MDFF run ([Supplementary Figures S6A and B](#)). Considering the NSD3-PWWP2-SET and NSD3-SET-PHD4 regions separately, it can be found that the first slightly decreases its size, while the second increases it by about 0.5 Å ([Supplementary Figures S6C and D](#)). The direction of these changes is consistent with the information given by the experimental assessment of the radius of gyration ([Table 1](#)), since the R_g of the NSD3-PWWP2-SET region of the AlphaFold model (35.0 Å) is above its SAXS-derived value in direct space (33.3 Å), while the contrary occurs for the R_g of the NSD3-SET-PHD4 region (32.9 Å of AlphaFold model *versus* 36.6 Å for the experimental value). In the reciprocal space, the initial and final models, considered separately for the two regions, produce different calculated SAXS profiles ([Figures 6C and D](#)). The *a posteriori* assessment of the agreement between the experimental and calculated SAXS profiles as a function of the simulation time

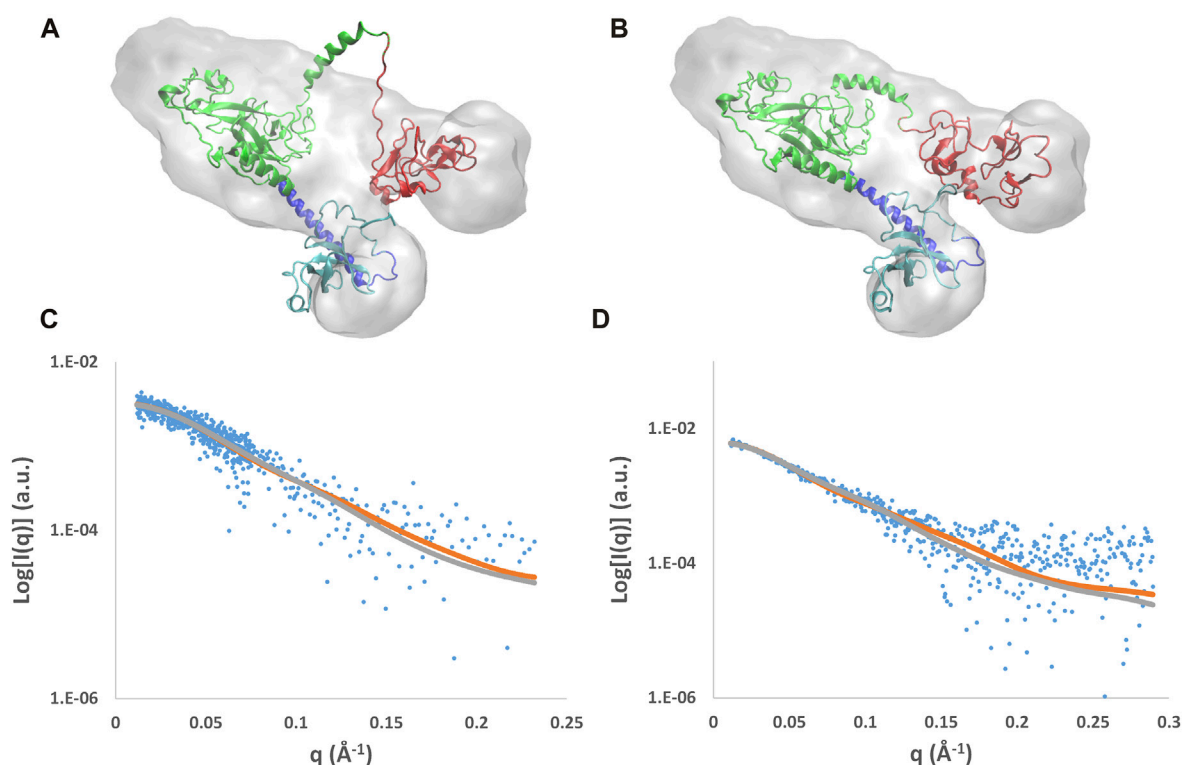


FIGURE 6

Results of the MDFF optimization applied to the AlphaFold mix model. Initial (A) and final (B) models superposed to the molecular envelope calculated from SAXS data and their fit with SAXS profiles for the NSD3-PWWP2-SET (C) and NSD3-SET-PHD4 (D) regions. The molecular envelope is shown as the transparent gray surface, and the models are shown in graphical representation, with the following color code: PWWP2 (cyan), PWWP2-SET linker (blue), SET (green), and PHD4 (red). Observed SAXS profiles (blue dots) and those calculated before (gray line) and after (brown line) application of MDFF are shown.

(Supplementary Figures S6C and D) indicates that the simulation rapidly converges toward best models, reaching χ^2 values of 1.07 for NSD3-PWWP2-SET and 1.23 for NSD3-SET-PHD4.

Analogous results are obtained by applying the MDFF optimization to the AlphaFold model (Supplementary Figure S7), although a higher value of χ^2 (1.74) is reached for the NSD3-PWWP2-SET region with respect to the AlphaFold mix model. Instead, in the MDFF optimization of the RaptorX model 1, a better fit of the model in the direct space does not turn into an overall improvement of the model in the reciprocal space. In particular, the NSD3-SET-PHD4 region has an opposite behavior with respect to the previous cases as it decreases its radius of gyration while increasing the χ^2 of the fit (Supplementary Figure S8).

3.2.3 Comparative analysis of the generated models

A comparative analysis of the structural solutions obtained was performed by considering the structural diversity, as measured by the residue-by-residue backbone dihedral angles, and the agreement of the model with SAXS data, which was assessed in the direct space by the normalized structural discrepancy with the *ab initio* molecular envelope and in the reciprocal space by the χ^2 of the fit with the SAXS profile. This analysis, detailed in Supplementary Material (Supplementary Figures S9, S10), indicates that the structural variations introduced by MDFF are not covered by other homology modeling tools and that the solution obtained by

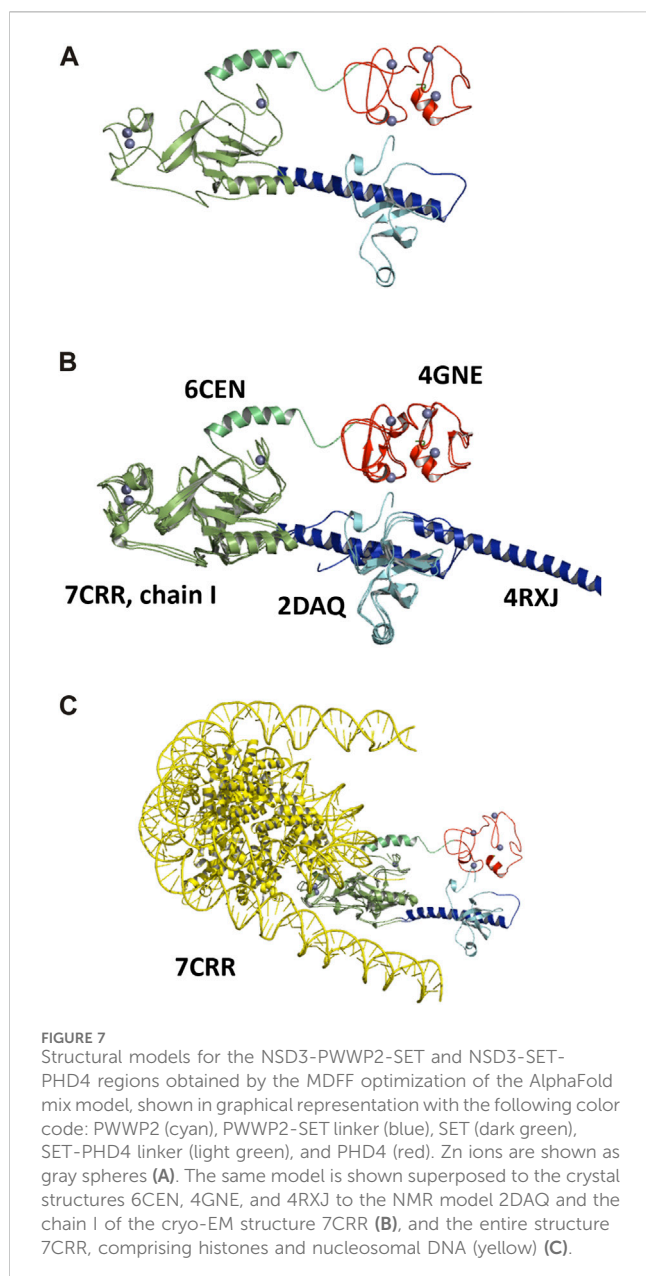
MDFF on the AlphaFold mix model is the best one since the agreement with SAXS data is improved in both the NSD3-PWWP2-SET and NSD3-SET-PHD4 regions. The resulting model shows better agreement with SAXS data than those generated by AlphaFold, Raptor X (model 1), or even CORAL.

The AlphaFold-derived models optimized by MDFF relative to the selected NSD3-PWWP2-SET and NSD3-SET-PHD4 samples have been deposited in the SASBDB entries n. SASDNL8 and SASDNK8, respectively.

3.2.4 Analysis of the full-length C-terminal model

The added-value of this structural investigation is to supply a complete characterization of the NSD3 C-terminal region comprising the PWWP2, SET, and PHD4 domains. The most plausible model, i.e., the one obtained by the MDFF refinement applied to the AlphaFold mix model, is given in Figure 7A and confirms the presence of a fully structured linker between PWWP and SET and a partially structured linker between SET and PHD4, where an α -helix is present in the residues ranging from 1,292 to 1,311.

The superposition of this model with the known structural models of individual NSD3 C-terminal domains is shown in Figure 7B. The SET domain characterized in this study is in a good overlap with that from the crystal structure with the PDB code 6CEN (RMSD = 0.9 \AA over 217 aligned residues) and from the cryo-EM structure 7CRR (RMSD = 1.6 \AA over 240 aligned residues); the PWWP2 domain is in fair overlap with those of the NMR model



2DAQ (RMSD = 1.0 Å over 72 aligned residues) and the crystal structure 4RXJ (RMSD = 0.9 Å over 73 aligned residues), while the PHD4 domain overlaps with the crystal structure 4GNE (RMSD = 0.9 Å over 95 aligned residues). However, none of the existing experimental structures can cover the full-length PWWP2-SET-PHD4 segment, so the mutual arrangement of individual domains can only be inferred by using the SAXS-derived structural model. It is interesting to note that the α -helix connecting the PWWP2 and SET domains actually adopts two opposing directions in the 2DAQ and 4RXJ models, so our investigation resolves this controversy by indicating 2DAQ as the model that best fits the actual conformation adopted by the helix when the full C-terminal region is considered.

The superposition of our SAXS-derived model with the cryo-EM structure 7CRR, comprising the NSD3 AWS, SET, and POST-SET domains interacting with the H3, H4, H2A, and H2B histone and the nucleosomal DNA, is shown in Figure 7C. We observe that no

clashes occur between the two structures, i.e., the NSD3 C-terminal reconstructed by SAXS data is fully compatible with the high-resolution structure of the NSD3 catalytic core bound to mononucleosome. In particular, we note that alternative conformations of the NSD3-PWWP2-SET and NSD3-SET-PHD4 constructs, for example, those assumed by the CORAL model (Supplementary Figure S5A), would not be compatible with the cryo-EM structure due to clashes with the histone proteins bound to NSD3s. Thus, the proximity of the PWWP2 and PHD4 domains, a peculiar feature of the SAXS-derived model, is in line with the function performed by the protein. We can envisage that the presence of mononucleosomes could induce a conformational change of the NSD3 C-terminal that leads the PWWP2 and PHD4 domains to interact with the DNA.

4 Discussion

Several crystal structures of individual C-terminal domains of NSD3 are present in the Protein Data Bank. However, no structural information is available about the C-terminal region from PWWP2 to PHD4, despite many efforts to crystallize such a region. Here, we performed a structural investigation at a low resolution (>20 Å) of such a region using the SAXS technique coupled with size-exclusion chromatography and complemented by advanced computational modeling.

Two constructs whose sequences overlap for 247 residues were considered: one covering the region from PWWP2 to SET and the other related to the region from SET to PHD4 (Figure 3). Datasets obtained by measuring at different concentrations were selected based on two quality parameters: the shape ambiguity of their molecular envelope and the quality of the $P(r)$ fit of their SAXS profile.

Homology modeling was performed using state-of-the-art procedures that strongly rely on machine learning approaches to predict the three-dimensional structure of the full-length NSD3 C-terminal region comprising the region from PWWP2 to PHD4. SAXS data on the individual constructs were then used for model validation and refinement. This top-down strategy has proven more effective than the bottom-up approach of building separate models of the two constructs driven by SAXS data and trying to put them together to form a full-length model.

Model validation was performed in direct and reciprocal space using the following two quality metrics: the normalized spatial discrepancy between the atomic model and the molecular envelope, and the agreement between calculated and observed SAXS profiles. This dual-space approach improved the sensitivity of the SAXS data, benchmarked the predicting tools adopted, and allowed the selection of the full-atom model of the NSD3 C-terminal that was in best agreement with the SAXS data. This model, obtained as a combination of two different models generated by AlphaFold, predicts closely spaced PWWP and PHD4 domains, a feature that is shared by two other well-scored models (RaptorX 1 and 2).

Model refinement was carried out on the full-length homology models by adopting molecular dynamics (MD) to introduce flexibility based on *a priori* physicochemical knowledge in the context of a complex fitting procedure. The SAXS-derived molecular envelope and experimental structural knowledge about

individual domains were then introduced as restraints in MD. This flexible fitting approach, called MDFF, improved not only the agreement with SAXS data in direct space, ensuring better coverage of the *ab initio* molecular envelope, but also the agreement in reciprocal space, as verified by a *posteriori* fit of the SAXS profile, with those calculated from the MD frames.

A comparative analysis of the MDFF results was carried out by considering (i) the minimum spatial discrepancy with the SAXS-derived molecular envelope in direct space, (ii) the agreement between observed and calculated SAXS profiles in reciprocal space, and (iii) the mutual orientation of individual residues allowed to select of the best models for the NSD3-PWWP2-SET and NSD3-SET-PHD4 constructs and build a consistent model of the NSD3 C-terminal region that sheds light into the mutual arrangement of the PWWP2, SET, and PHD4 domains. Alternative generated models predicting different mutual orientations of PWWP2 and PHD4 domains were ruled out by this analysis, thus enforcing the evidence that these models are closely spaced, thus interacting with each other in solution. Known crystallographic, NMR, and cryo-EM structures of the PWWP2, SET, and PHD4 NSD3 domains cannot be located relative to each other without using this new SAXS-derived structural knowledge. Moreover, the structural model of the NSD3 C-terminal obtained here is compatible with the binding of NSD3 to mononucleosomes.

This study discloses the mutual arrangement of the PWWP2, SET, and PHD4 domains in the NSD3 C-terminal, which is not accessible by high-resolution structural techniques due to the intrinsic flexibility of this protein region. Such results could provide implications for the mechanism of functional diversity of NSD proteins and the underexplored biological function of the PWWP2 domain.

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 at:
<https://www.sasbdb.org/data/SASDNL8/>, SASDNL8.

Author contributions

EL designed and coordinated research. MM and YS prepared the protein samples. EL and YS generated the homology models. BB and RC performed the SAXS experiments. BB, BC, and RC analyzed

data. RC performed the molecular dynamics simulations. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by a joint project of 2016–2017 between the National Research Foundation of Korea (NRF) and the National Research Council of Italy (CNR) entitled “Static and dynamic crystallographic investigations for developing specific and selective inhibitors for the epigenetic therapy of cancers”.

Acknowledgments

The authors wish to thank the Diamond Light Source for access to beamline B21 (Proposal No. MX15832, beamline session 4, and No. MX21741, beamline session 12).

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2024.1191246/full#supplementary-material>

References

- Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012). H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* 40, W537–W541. doi:10.1093/nar/gks375
- Angrand, P. O., Apiou, F., Stewart, A. F., Dutrillaux, B., Losson, R., and Chambon, P. (2001). NSD3, a new SET domain-containing gene, maps to 8p12 and is amplified in human breast cancer cell lines. *Genomics* 74, 79–88. doi:10.1006/geno.2001.6524
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754
- Belviso, B. D., Mangiatordi, G. F., Alberga, D., Mangini, V., Carrozzini, B., and Caliendo, R. (2022). Structural characterization of the full-length anti-CD20 antibody Rituximab. *Front. Mol. Biosci.* 9, 823174. doi:10.3389/fmolb.2022.823174
- Bottcher, J., Dilworth, D., Reiser, U., Neumuller, R. A., Schleicher, M., Petronczki, M., et al. (2019). Fragment-based discovery of a chemical probe for the PWWP1 domain of NSD3. *Nat. Chem. Biol.* 15, 822–829. doi:10.1038/s41589-019-0310-x
- Caliandro, R., and Belviso, D. B. (2014). RootProf: software for multivariate analysis of unidimensional profiles. *J. Appl. Cryst.* 47, 1087–1096. doi:10.1107/S1600576714005895
- Caliandro, R., Rossetti, G., and Carloni, P. (2012). Local fluctuations and conformational transitions in proteins. *J. Chem. Theory Comput.* 8, 4775–4785. doi:10.1021/ct300610y

- Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397
- de Freitas, R. F., Liu, Y., Szweczyk, M. M., Mehta, N., Li, F., McLeod, D., et al. (2021). Discovery of small-molecule antagonists of the PWWP domain of NSD2. *J. Med. Chem.* 64, 1584–1592. doi:10.1021/acs.jmedchem.0c01768
- Franke, D., and Svergun, D. I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* 42 (2), 342–346. doi:10.1107/S0021889809000338
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). “Protein identification and analysis tools on the ExPASy server,” in *The proteomics protocols handbook*. Editor J. M. Walker (New Jersey, United States: Humana Press), 571–607. doi:10.1385/1-59259-890-0:571
- Hajizadeh, N. R., Franke, D., Jeffries, C. M., and Svergun, D. I. (2018). Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data. *Sci. Rep.* 8, 7204. doi:10.1038/s41598-018-25355-2
- Han, X., Piao, L., Zhuang, Q., Yuan, X., Liu, Z., and He, X. (2018). The role of histone lysine methyltransferase NSD3 in cancer. *Oncotargets Ther.* 11, 3847–3852. doi:10.2147/OTT.S166006
- He, C., Li, F., Zhang, J., Wu, J., and Shi, Y. (2013). The methyltransferase NSD3 has chromatin-binding motifs, PHD5-C5HCH, that are distinct from other NSD (nuclear receptor SET domain) family members in their histone H3 recognition. *J. Biol. Chem.* 288, 4692–4703. doi:10.1074/jbc.M112.426148
- Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD - visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi:10.1016/0263-7855(96)00018-5
- Jaffe, J. D., Wang, Y., Chan, H. M., Zhang, J., Huether, R., Kryukov, G. V., et al. (2013). Global chromatin profiling reveals NSD2 mutations in pediatric acute lymphoblastic leukemia. *Nat. Genet.* 45, 1386–1391. doi:10.1038/ng.2777
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nat* 596, 583–589. doi:10.1038/s41586-021-03819-2
- Keats, J. J., Reiman, T., Maxwell, C. A., Taylor, B. J., Larratt, L. M., Mant, M. J., et al. (2003). In multiple myeloma, t(4;14)(p16;q32) is an adverse prognostic factor irrespective of FGFR3 expression. *Blood* 101, 1520–1529. doi:10.1182/blood-2002-06-1675
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The PyMol web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi:10.1038/nprot.2015.053
- Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M., and Svergun, D. I. (2020). SASBDB: towards an automatically curated and validated repository for biological scattering data. *Protein Sci.* 29, 66–75. doi:10.1002/pro.3731
- Kozin, M. B., and Svergun, D. I. (2001). Automated matching of high- and low-resolution structural models. *J. Appl. Cryst.* 34, 33–41. doi:10.1107/S0021889800014126
- Kuo, A. J., Cheung, P., Chen, K., Zee, B. M., Kioi, M., Lauring, J., et al. (2011). NSD2 links dimethylation of histone H3 at lysine 36 to oncogenic programming. *Mol. Cell* 44, 609–620. doi:10.1016/j.molcel.2011.08.042
- Li, W., Tian, W., Yuan, G., Deng, P., Sengupta, D., Cheng, Z., et al. (2021). Molecular basis of nucleosomal H3K36 methylation by NSD methyltransferases. *Nat* 590, 498–503. doi:10.1038/s41586-020-03069-8
- Liu, V. C., Mirabelli, V., Cimmarusti, M. T., Haidukowski, M., Leslie, J. F., Logrieco, A. F., et al. (2017). Enniatin and beauvericin biosynthesis in *Fusarium* species: production profiles and structural determinant prediction. *Toxins* 9, 45. doi:10.3390/toxins9020045
- Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., et al. (2021). ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Cryst.* 54, 343–355. doi:10.1107/S1600576720013412
- Mirdita, M., Schütz, K., Moriawaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi:10.1038/s41592-022-01488-1
- Morishita, M., Mevius, D., and di Luccio, E. (2014). *In vitro* histone lysine methylation by NSD1, NSD2/MMSET/WHSC1 and NSD3/WHSC1L. *BMC Struct. Biol.* 14, 25. doi:10.1186/s12900-014-0025-x
- Morishita, M., Mevius, D. E. H. F., Shen, Y., Zhao, S., and di Luccio, E. (2017). BIX-01294 inhibits oncoproteins NSD1, NSD2 and NSD3. *Med. Chem. Res.* 26, 2038–2047. doi:10.1007/s00044-017-1909-7
- Morrison, M. J., Boriack-Sjodin, P. A., Swinger, K. K., Wigle, T. J., Sadalge, D., Kuntz, K. W., et al. (2018). Identification of a peptide inhibitor for the histone methyltransferase WHSC1. *PLoS One* 13, e0197082. doi:10.1371/journal.pone.0197082
- Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., et al. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Cryst.* 45, 342–350. doi:10.1107/S0021889812007662
- Petoukhov, M. V., and Svergun, D. I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* 89, 1237–1250. doi:10.1529/biophysj.105.064154
- Petoukhov, M. V., and Svergun, D. I. (2015). Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Cryst. D* 71, 1051–1058. doi:10.1107/S1399004715002576
- Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* 153, 044130. doi:10.1063/5.0014475
- Qin, S., and Min, J. (2014). Structure and function of the nucleosome-binding PWWP domain. *Trends biochem. Sci.* 39, 536–547. doi:10.1016/j.tibs.2014.09.001
- Rambo, R. P. (2017). ScÅtter a java based graphical user interface for the processing and analysis of SAXS data. Available at: <https://bl1231.als.lbl.gov/scatter/>.
- Receveur-Bréchet, V. (2023). AlphaFold, small-angle X-ray scattering and ensemble modelling: a winning combination for intrinsically disordered proteins. *J. Appl. Cryst.* 56, 1313–1314. doi:10.1107/S1600576723008403
- Sankaran, S. M., Wilkinson, A. W., Elias, J. E., and Gozani, O. (2016). A PWWP domain of histone-lysine N-methyltransferase NSD2 binds to dimethylated lys-36 of histone H3 and regulates NSD2 function at chromatin. *J. Biol. Chem.* 291, 8465–8474. doi:10.1074/jbc.M116.720748
- Sato, K., Kumar, A., Hamada, K., Okada, C., Oguni, A., Machiyama, A., et al. (2021). Structural basis of the regulation of the normal and oncogenic methylation of nucleosomal histone H3 Lys36 by NSD2. *Nat. Commun.* 12, 6605. doi:10.1038/s41467-021-26913-5
- Shen, Y., Morishita, M., Lee, D., Kim, S., Lee, T., Mevius, D. E. H. F., et al. (2019). Identification of LEM-14 inhibitor of the oncoprotein NSD2. *Biochem. Biophys. Res. Commun.* 508, 102–108. doi:10.1016/j.bbrc.2018.11.037
- Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *J. Nat. Biotechnol.* 35, 1026–1028. doi:10.1038/nbt.3988
- Svergun, D., Barberato, C., and Koch, M. H. J. (1995). CRYSOLE – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* 28, 768–773. doi:10.1107/S0021889895007047
- Svergun, D. I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Cryst.* 25, 495–503. doi:10.1107/S0021889892001663
- Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* 76, 2879–2886. doi:10.1016/S0006-3495(99)77443-6
- Tisi, D., Chiarparin, E., Tamanini, E., Pathuri, P., Coyle, J. E., Hold, A., et al. (2016). Structure of the epigenetic oncoprotein MMSET and inhibition by N-alkyl sinefungin derivatives. *ACS Chem. Biol.* 11, 3093–3105. doi:10.1021/acschembio.6b00308
- Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16, 673–683. doi:10.1016/j.str.2008.03.005
- Tully, M. D., Tarbouriech, N., Rambo, R. P., and Hutin, S. (2021). Analysis of SEC-SAXS data via EFA deconvolution and Scatter. *J. Vis. Exp.* 167. doi:10.3791/61578
- Volkov, V. V., and Svergun, D. I. (2003). Uniqueness of *ab initio* shape determination in small-angle scattering. *J. Appl. Cryst.* 36, 860–864. doi:10.1107/S0021889803000268
- Vougiouklakis, T., Hamamoto, R., Nakamura, Y., and Saloura, V. (2015). The NSD family of protein methyltransferases in human cancer. *Epigenomics* 7, 863–874. doi:10.2217/epi.15.32
- Wilhelm, H., Ashton, A. W., Chang, P. C. Y., Chater, P. A., Day, S. J., Drakopoulos, M., et al. (2017). Processing two-dimensional X-ray diffraction and small-angle scattering data in DAWN 2. *J. Appl. Cryst.* 50, 959–966. doi:10.1107/S1600576717004708
- Xu, J., McPartlon, M., and Li, J. (2021). Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat. Mach. Intell.* 3, 601–609. doi:10.1038/s42256-021-00348-5
- Zhang, M., Yang, Y., Zhou, M., Dong, A., Yan, X., Loppnau, P., et al. (2021). Histone and DNA binding ability studies of the NSD subfamily of PWWP domains. *Biochem. Biophys. Res. Commun.* 569, 199–206. doi:10.1016/j.bbrc.2021.07.017
- Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E. W., and Zhang, Y. (2021). Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Rep. Methods* 1, 100014. doi:10.1016/j.crmeth.2021.100014



OPEN ACCESS

EDITED BY

Guo-Wei Wei,
Michigan State University, United States

REVIEWED BY

Nenad Filipovic,
University of Kragujevac, Serbia
Ali Kouhi,
Tehran University of Medical Sciences, Iran
Menglun Wang,
United States Food and Drug Administration,
United States

*CORRESPONDENCE

Qiyuan Li,
✉ qiyuan.li@xmu.edu.cn
Chengfu Cai,
✉ ysc96@126.com

[†]These authors share first authorship

RECEIVED 30 June 2023

ACCEPTED 27 December 2023

PUBLISHED 21 March 2024

CITATION

Qi F, You Z, Guo J, Hong Y, Wu X, Zhang D, Li Q
and Cai C (2024), An automatic diagnosis model
of otitis media with high accuracy rate using
transfer learning.
Front. Mol. Biosci. 10:1250596.
doi: 10.3389/fmolb.2023.1250596

COPYRIGHT

© 2024 Qi, You, Guo, Hong, Wu, Zhang, Li and
Cai. 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.

An automatic diagnosis model of otitis media with high accuracy rate using transfer learning

Fangyu Qi^{1,2,3†}, Zhiyu You^{2,3†}, Jiayang Guo^{2,3}, Yongjun Hong⁴,
Xiaolong Wu^{2,3}, Dongdong Zhang², Qiyuan Li^{2,3*} and
Chengfu Cai^{5*}

¹Department of Anesthesiology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen, China,

²School of Medicine, Xiamen University, Xiamen, China, ³National Institute for Data Science in Health and
Medicine, Xiamen University, Xiamen, China, ⁴Zhongshan Hospital Affiliated to Xiamen University, Xiamen,
China, ⁵College of Otorhinolaryngology Head and Neck Surgery, Xiamen Haicang Hospital, Xiamen, China

Introduction: Chronic Suppurative Otitis Media (CSOM) and Middle Ear Cholesteatoma are two common chronic otitis media diseases that often cause confusion among physicians due to their similar location and shape in clinical CT images of the internal auditory canal. In this study, we utilized the transfer learning method combined with CT scans of the internal auditory canal to achieve accurate lesion segmentation and automatic diagnosis for patients with CSOM and middle ear cholesteatoma.

Methods: We collected 1019 CT scan images and utilized the nnUnet skeleton model along with coarse grained focal segmentation labeling to pre-train on the above CT images for focal segmentation. We then fine-tuned the pre-training model for the downstream three-classification diagnosis task.

Results: Our proposed algorithm model achieved a classification accuracy of 92.33% for CSOM and middle ear cholesteatoma, which is approximately 5% higher than the benchmark model. Moreover, our upstream segmentation task training resulted in a mean Intersection of Union (mIoU) of 0.569.

Discussion: Our results demonstrate that using coarse-grained contour boundary labeling can significantly enhance the accuracy of downstream classification tasks. The combination of deep learning and automatic diagnosis of CSOM and internal auditory canal CT images of middle ear cholesteatoma exhibits high sensitivity and specificity.

KEYWORDS

chronic suppurative otitis media (CSOM), middle ear cholesteatoma, CT images, computer-aided diagnosis (CAD), transfer learning (TL)

Introduction

Otitis media is a prevalent ear disease that affects a significant portion of the global population, with an estimated 65 to 350 million individuals affected worldwide ([World Health Organization, 2004](#)). In developing countries, the prevalence of Chronic Suppurative Otitis Media (CSOM) ranges from 0.4% to 33.3% ([Kaur et al., 2017](#)). Our study mainly focuses on non-invasive temporal bone CT images, in order to help clinicians quickly get a relatively accurate preliminary diagnosis and lay the foundation for further judgment of whether patients need surgical treatment.

Otitis media is classified into three categories: acute otitis media, chronic otitis media (COM), and middle ear cholesteatoma. Chronic otitis media typically exhibits more pronounced pathological changes on CT images due to its protracted course, whereas acute otitis media does not usually display this characteristic. Consequently, it is often recommended that patients with chronic otitis media undergo internal auditory canal CT scan to assess their condition. Chronic otitis media is further divided into two subcategories: chronic non-suppurative otitis media and chronic suppurative otitis media (CSOM) (Schilder et al., 2017). CSOM and Middle Ear Cholesteatoma are two typical otitis media diseases that are diagnosed primarily through temporal bone CT scans (Fukudome et al., 2013; Lustig et al., 2018). CSOM typically occurs following improper treatment of acute otitis media, often resulting in tympanic membrane perforation and persistent middle ear purulence (Ahmad et al., 2022). Middle ear cholesteatoma, on the other hand, is the pathological outcome of abnormal accumulation of keratin squamous epithelium, primarily composed of keratinized, exfoliated epithelium. It often accumulates in the middle ear, with a tendency to erode the ossicular chain, tympanic wall, and/or mastoid area (Sun et al., 2011; Jang et al., 2014).

Clinicians usually identify CSOM and middle ear cholesteatoma through CT scanning of the internal auditory canal. However, CT reports of these two types both show erosion and/or loss of the ossicular chain with diffuse abnormal soft tissue shadow (Madabhushi and Lee, 2016). Theoretically, the two types differ in bone erosion margins and soft tissue shadow contours: The soft tissue shadow of cholesteatoma has a smooth, clear outline, while that of CSOM lacks a clear outline and is often accompanied by pus accumulation. In addition, the edge of the bone erosion caused by CSOM is serrated, while bone destruction caused by cholesteatoma is frequently surrounded by a ring of sclerosis. Therefore, our group proposed using deep learning to differentiate between CSOM and cholesteatoma to achieve more accurate clinical diagnoses based on the theoretical differences between these two diseases (Kempainen et al., 1999; Yorgancılar et al., 2013).

Deep learning techniques have seen widespread use in the medical field in recent years, enabling the extraction of key features from patients to facilitate predictive modeling (Elfiky et al., 2018). Transfer learning is a deep learning technique that involves leveraging knowledge gained from solving one problem to address another related problem. It is particularly useful when the amount of labeled data for the target task is limited. By leveraging transfer learning, a pre-trained model developed for one task can be fine-tuned and adapted for another task with different data but similar features. This approach enables the model to benefit from the knowledge learned by the pre-trained model on a larger dataset and adapt it to the new task by making only minor adjustments to the model's architecture or parameters. A bunch of researches have uncovered that superiority of transfer learning over traditional strategy. S. Deepak implement brain tumor classification using deep CNN features via transfer learning (Deepak and Ameer, 2019). For tuberculosis detection, a VGGNet based model had been proposed combining transfer learning (Ahsan et al., 2019). Dube S presented an automatic content-based image retrieval system for brain tumors on contrast-enhanced MRI (Dube et al., 2006).

Given the small morphological differences between various types of otitis media, the challenge of manual identification, and the unclear contour of lesions, we sought to establish a transfer learning framework by integrating coarse-grained labeled contour information as pre-trained data and employing the CNN model

skeleton to extract high-level features from internal auditory canal CT images. Specifically, the deep representation of images obtained through pre-training was utilized to accurately classify CSOM, middle ear cholesteatoma, and normal samples in downstream tasks.

In conclusion, this paper's key contributions are:

1. We propose a transfer learning-based framework that utilizes coarsely annotated segmentation data as input for pretraining the model. The proposed model effectively extracts implicit information from the data and can subsequently be used for classification prediction.
2. We propose an end-to-end learning model that can effectively improve the accuracy of middle ear infection classification prediction. Our proposed model outperforms non-pretrained models in all metrics.
3. In the field of deep learning combined with medicine, we look forward to replacing the heavy and repetitive manual labeling task with more mature machine automated labeling.
4. Our combination of otological diseases and computer learning can increase the coverage of related research and provide more precise and diversified help for clinicians in diagnosis and treatment.

Materials and methods

Data acquisition

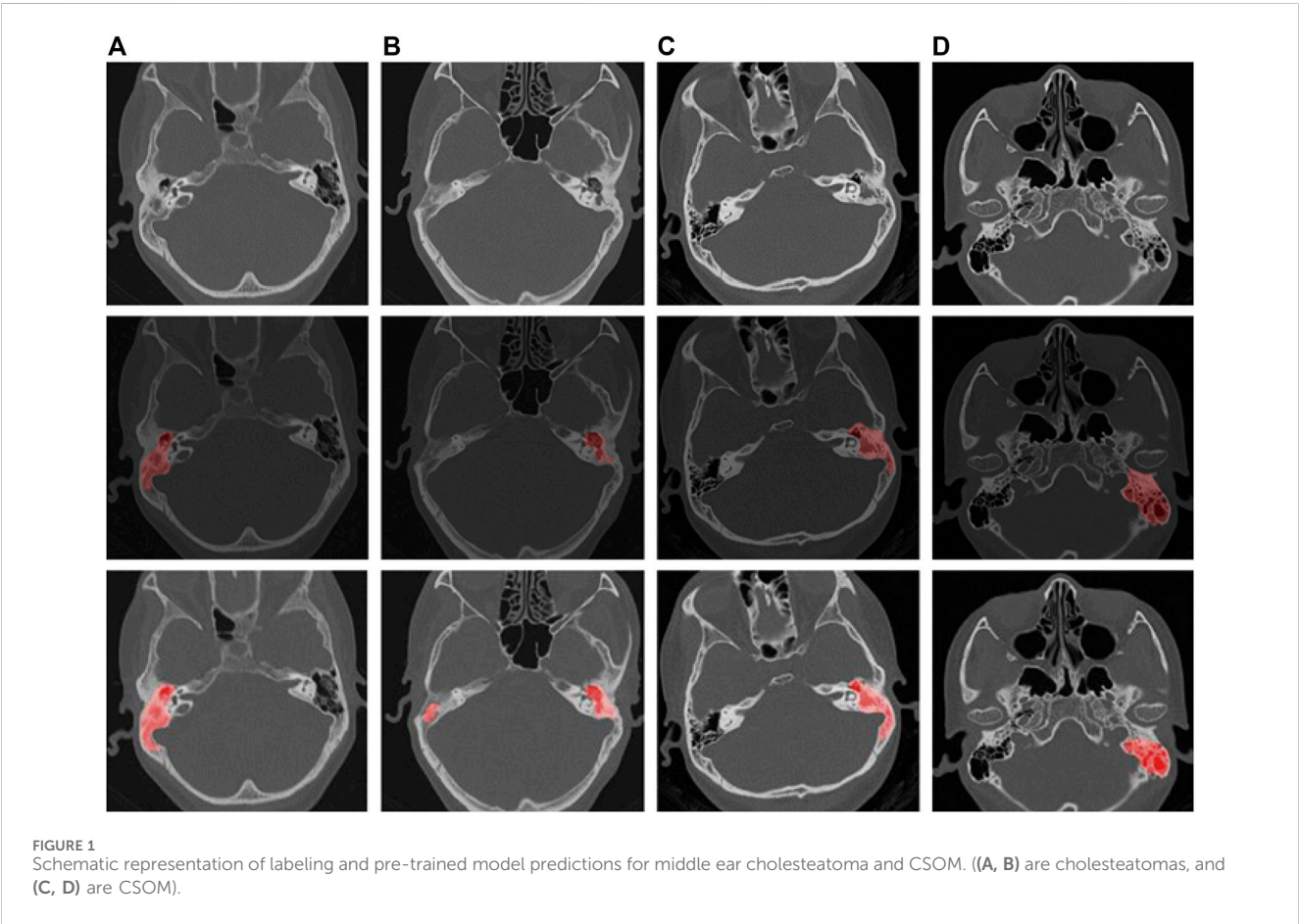
We conducted a retrospective study at Zhongshan Hospital Affiliated to Xiamen University to investigate patients diagnosed with otitis media and middle ear cholesteatoma from 2012 to 2021. This study was approved by the ethics committee and informed consent was waived due to the retrospective nature of the study. The inclusion criteria for the study were based on pathology or medical history, ear examination, audiogram, and imaging examination of the surgical side of the ear. We referred to the previous medical records of the hospital to obtain specific diagnosis results. A total of 6,967 axial high-resolution CT images of temporal bone were collected from 180 patients, including 27 female patients with middle ear cholesteatoma, 31 male patients with middle ear cholesteatoma; 50 female patients and 51 male patients with chronic otitis media, including 40 females and 39 males with CSOM, and Chronic otitis media with effusion, including 10 females and 12 males. Besides, there were 15 normal controls (6 females and 9 males), and 2 children with middle ear cholesteatoma, 2 children with CSOM (less than 10 years old), and 2 normal controls. Except for children, the patients collected were between 30 and 80 years of age. All the patients information has been shown in Table 1.

The CT features of chronic otitis media with effusion (COME) are often very similar to those of chronic suppurative otitis media (CSOM), with both conditions typically presenting with varying degrees of effusion in mastoid cells. As such, we excluded a total of 412 axial CT data from 22 patients diagnosed with COME. For each patient, we selected approximately 10–20 CT images that showed well-defined lesions. Ultimately, we used 410 CSOM CT images, 398 middle ear cholesteatoma CT images, and 211 normal CT control images for our analysis.

TABLE 1 The collected patients information.

	Female (30–80 years old)	Male (30–80 years old)	Pediatric patients (<10 years old)
Middle Ear Cholesteatoma	27	31	2
CSOM	40	39	2
Chronic Otitis Media with Effusion	10	12	0
Normal	6	9	2
Total	83	91	6

Data sources: Zhongshan Hospital Affiliated to Xiamen University.
Note: numbers represent the quantity of patients.



CT scanner settings

Temporal bone CT is derived from GE LightSpeed 64-row volume CT. Its detector is the core technology of multi-slice spiral CT. The detector arrangement adopts 64×0.625 mm detector unit to ensure the maximum coverage of 40 mm/circle at present, and at the same time, it can also perform sub-millimeter thick scanning in any mode. The isotropic resolution is up to 0.30 mm, which ensures a large range of volume acquisition and high resolution acquisition. The CT imaging parameters used were as follows: CT collimator 128×1.0 mm, field of view 220×220 mm, matrix size $1,024 \times 1,024$, voltage 120 kV, current 240 mAs, and axial CT slice number 30–50 per scan.

Data marking

The CT findings of chronic suppurative otitis media are often difficult to distinguish from middle ear cholesteatoma. To accurately identify middle ear cholesteatomas, we marked local or isolated cholesteatomas in the erosion area of the incudostapedial joint or hammer-incus joint in the rotation plane of the middle tip of the cochlea. We also highlighted areas of bone destruction within the tympanic sinus, epitympanic region, or mastoid process in other levels. Additionally, we marked any irregular soft tissue shadows with smooth edges on any plane. In contrast, when marking CT images of suppurative otitis media, we identified the soft tissue shadow around the auricle in the tympanic cavity, the sclerotic

hyperplasia part of the mastoid, and the bone with uniform density and serrated edge in the tympanic sinus at the vestibular level. These characteristics helped distinguish them from the sclerosing ring that is formed by the compression of a cholesteatoma. At the apical spiral layer of the cochlea, we marked the hammer-incus joint of the ossified epitympanic, the “ice cream cone-like structure,” and the serrated bone around the mastoid cavity and sinus. At the bottom spiral level of the cochlea, we marked the thickened mucosa on the promontory surface. Finally, at the mastoid level, we marked the erosion of the mastoid bone and the thickening of the mucous membrane caused by suppurative effusion (Gomaa et al., 2013; Zelikovich, 2004). We provide visual examples of typical lesion markers and predictions for these two diseases in Figure 1.

The original data was stored in the Dicom format, which we converted into PNG image data using MicroDicom software. This step allowed us to separate the patient’s personal information from the image, thereby ensuring patient privacy.

To mark the lesion area on each image, we enlisted the help of a team consisting of five professional otolaryngologists and two radiologists. They used polygonal markers on LabelMe software to eliminate background interference and generate unified coordinates for each lesion area.

Data pre-processing

To improve the robustness of our model, we applied various image data augmentation and processing techniques during the training process. Specifically, we randomly transformed the input images by performing horizontal and vertical translations, flipping, rotation, slight scaling, and adjustments to hue, contrast, and numerical values.

In order to balance the size of our training model and the time required for training, we scaled all images to a uniform size of 224×224 using bilinear interpolation. This allowed us to efficiently process and train on a large dataset of images while still maintaining a high level of accuracy and performance in our final model.

Model architecture and training strategy

We utilized the nnUnet (Isensee et al., 2021) architecture as the foundation for our deep learning model to extract critical features from CT images. This model has demonstrated exceptional performance in various medical image segmentation tasks. Our model consists of two branches: coarse-grained segmentation task and exact classification task.

In our workflow, we first pre-trained a model for lesion segmentation, which includes the nnUnet skeleton and a pixel-level prediction head that outputs three classification results for each pixel: CSOM, middle ear cholesteatoma, or normal samples. On the back of the above-mentioned process, our goal was to acquire a well-trained backbone that could extract underlying information containing pixel-level features, which would then be fine-tuned for picture-level classification. We trained this model using the gradient descent algorithm until convergence was achieved.

The nnUNet model is composed of an encoder and a decoder. The encoder reduces the image size layer by layer while capturing features of varying granularity from different images. It consists of seven layers, each containing {1, 3, 4, 6, 6, 6, 6} blocks, with each block containing two convolutional layers, two activation layers, and two normalization layers. Successive layers are directly connected with a pooling layer, which reduces the image size by half. The first layer of the encoder contains 32 features, and the number of features in each subsequent layer doubles but does not exceed the maximum number of features, which is 512.

The decoder has six layers, each consisting of {2, 2, 2, 2, 2, 2} blocks. These layers use linear interpolation upsampling to increase the image size. The encoder and decoder are connected using residual layer hopping. The decoder outputs the hidden variables of the image as inputs to both the pixel-level projection head and the image-level projection head for further processing.

Once the model was successfully trained, we extracted the hidden variables before the pixel-level prediction head of the image as inputs for the downstream classification prediction head. This allowed us to efficiently classify images with high accuracy by leveraging the previously extracted features.

We employed a five-fold cross-validation approach to train and evaluate our model (see Figure 2). Given that a series of adjacent CT images from the same subject tend to exhibit strong similarities, we took care to avoid overfitting due to data leakage. Specifically, during the training process, we randomly divided the dataset into a training set and validation set at a ratio of 4:1, based on the subject’s name. This ensured that CT images from the same subject were assigned to either the training or validation set, but not both.

Our models were compiled using Python 3.8, trained with PyTorch version 1.10, and accelerated with Nvidia A100 high-performance GPUs. During the training process, we set the maximum training epoch to 500 epochs, with a training batch size of 8 samples. We used Adam as the model optimizer, with an initial learning rate of 0.001. The dynamic learning rate decreased gradually with each increase in training batch until it reached $10e-5$. For pre-training optimization, we utilized the cross-entropy of each pixel classification of the image. Downstream training utilized the cross-entropy of the image classification as the loss function. Our specific workflow is depicted in Figure 3 for further visualization.

(The pre-training phase uses the pixel-level labels of route A to train the CNN, and the pixel predictor is responsible for output the category of each pixel. When performing the downstream picture classification task, according to route B, the pre-trained model is used to fine-tune the neural network through the category classifier, and the final picture prediction result is output.)

Result

In the upstream segmentation task, our deep learning model achieved a mean Intersection of Union (mIoU) index of 0.5376, indicating excellent performance in accurately removing background noise. Subsequently, we employed this well-performing model for the downstream fine-tuning step, where we aimed to classify otitis media into three distinct categories. Our model achieved a micro-f1 index of 92.33%, a significant improvement of 4.83% compared to the benchmark model.

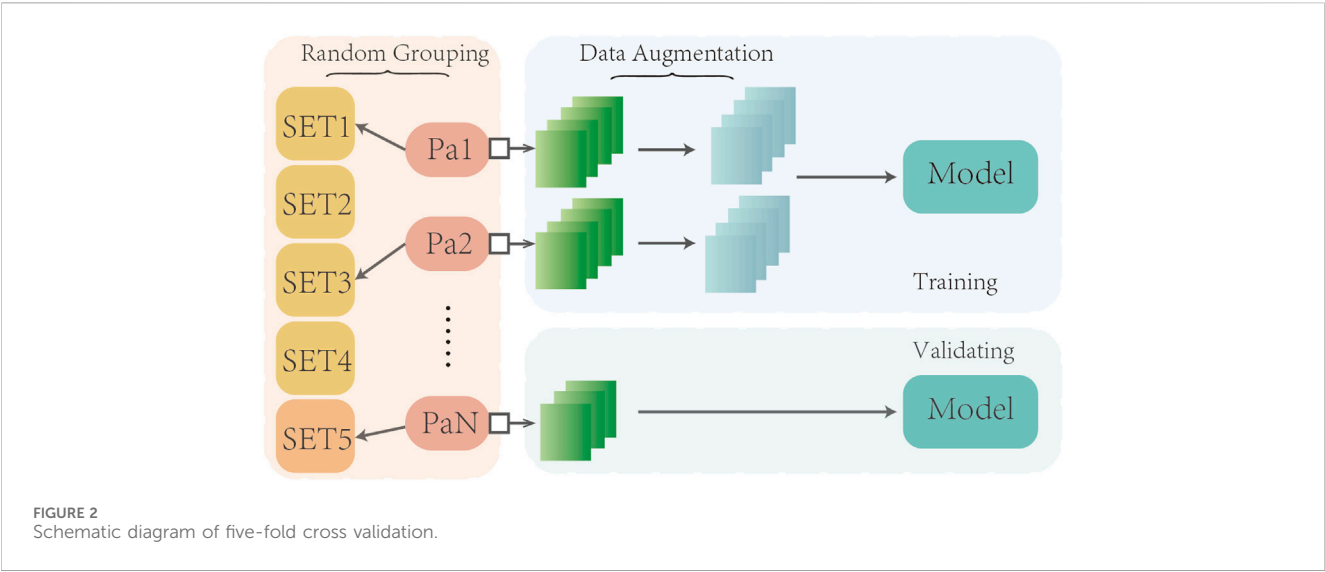
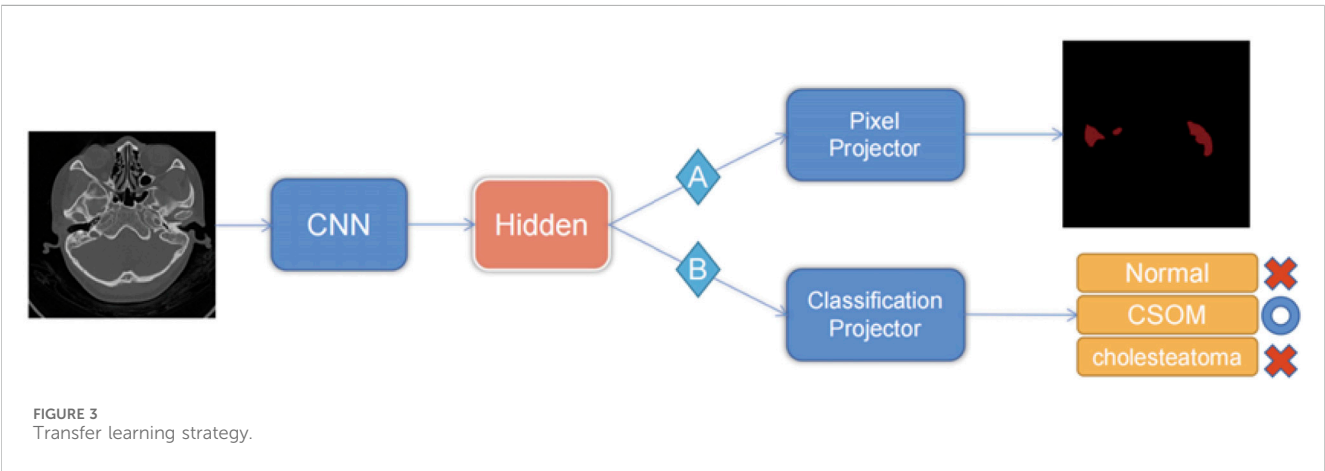


TABLE 2 Comparison of results.

		Accuracy	Auc
nnUnet	Normal vs. the others	0.8732 ± 0.1694	0.9558 ± 0.0454
	CSOM vs. the others	0.8619 ± 0.0776	0.9695 ± 0.0212
	Cholesteatoma vs. the others	0.9013 ± 0.0394	0.9555 ± 0.0242
p_nnUnet	Normal vs. the others	0.8873 ± 0.1593	0.9530 ± 0.0526
	CSOM vs. the others	0.9434 ± 0.0343	0.9916 ± 0.0069
	Cholesteatoma vs. the others	0.9290 ± 0.0170	0.9622 ± 0.0316

p represents the fine-tuning results after using the pre-trained model. The above results are the means after five-fold cross-validation. The mIoU index is used to describe the average ratio of intersection and union of all pixel categories in the image segmentation task. In this experiment, the background normal tissue categories were removed to obtain more accurate prediction results. mIoU is described as follows.



On the other hand, the pre-trained model exhibits an overall area under the receiver curve of 0.9689, which is slightly higher than that of the benchmark model which reach 0.9603. As is depicted in Figure 4, it can be observed that the performance in distinguishing chronic suppurative otitis media (CSOM) is the best, by a margin of 8.15%, as is showed in Table 2. These results indicate that the pre-trained model has a superior ability to accurately classify CSOM cases compared to the benchmark model. These results highlight the potential of deep learning technology in medical image analysis and its ability to significantly improve diagnostic accuracy and treatment outcomes.

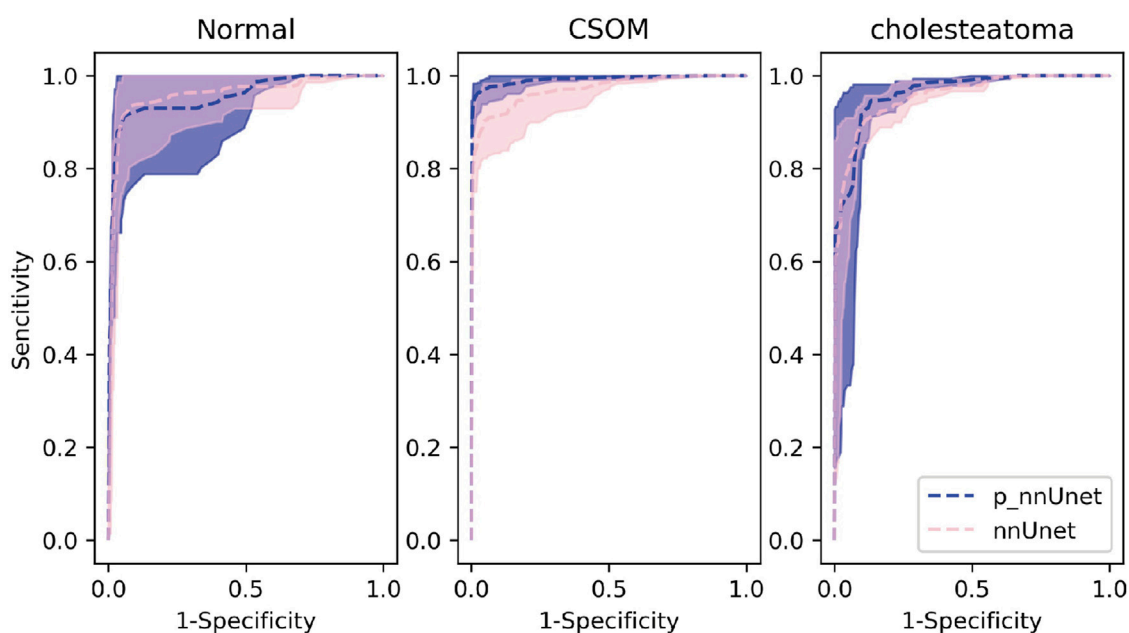


FIGURE 4
Receiver curves for each category of the model.

TABLE 3 Diagnostic accuracy results of manual diagnostic group.

The clinical experts ($n = 3$)/Real diagnose ($n = 3,039$)	COSM ($n = 1,182$)	Cholesteatoma ($n = 1,224$)	Normal ($n = 633$)
CSOM	948 (80.20%)	489 (39.95%)	48 (7.58%)
Cholesteatoma	222 (18.78%)	708 (57.84%)	9 (1.42%)
Normal	12 (1.01%)	27 (2.20%)	576 (90.99%)

The horizontal represents the actual diagnosis results, while the vertical represents the manual diagnosis results.

$$mIoU = \frac{1}{|D|} \sum_{i \in D} \frac{TP_i}{TN_i + FP_i + FP_i}$$

Note: In the above equation, FN is the false negative class, FP is the false positive class, TP is the true class, and D represents the CSOM and cholesteatoma set.

On the other hand, we demonstrate the disparity in accuracy between AI models and manual diagnosis by clinicians. We intentionally randomized and combined a total of 1013 CT images of middle ear cholesteatoma, CSOM, and normal control images, while effectively concealing the actual diagnostic labels. These images were then distributed to both the manual diagnosis group and the model diagnosis group for a double-blind evaluation of diagnostic accuracy. The comprehensive test results are shown in Tables 3, 4, and it can be observed that CSOM and cholesteatoma exhibit a high misdiagnosis rate.

Our experimental results show that in the otitis media classification task, the use of contour boundary labeling can well improve the accuracy of downstream classification tasks, and the area under the receiver operating characteristic curve is better than that of the non-pre-trained model shown in Figure 4. These results indicate that the predictive power of our model on this task has the possibility of real-world application.

(p-represents the fine-tuning results after using the pre-trained model. In the multi-classification ROC curve, the positive samples belong to a particular category while the negative samples belong to all other categories combined. Based on this distinction, the true positive rate and false positive rate have been accurately calculated).

Discussion

Otitis media is characterized by a prolonged course of illness, high incidence, easy recurrence, conductive deafness, and potentially fatal intracranial infection (Otten and Grote, 1990; Hutz et al., 2018). A case analysis conducted in a public hospital in the United States revealed that the incidence of postoperative complications associated with complex chronic otitis media with middle ear cholesteatoma was similar to that observed in developing regions (Greenberg and Manolidis, 2001). As a result, early diagnosis, intervention measures, and clinical management of this disease are especially crucial, regardless of whether one resides in developed or developing regions. In our study, we utilized CT images of CSOM and middle ear cholesteatoma labeled by

TABLE 4 Diagnostic accuracy results of model diagnosis group.

The AI model/Real diagnose (<i>n</i> = 1,013)	COSM (<i>n</i> = 394)	Cholesteatoma (<i>n</i> = 408)	Normal (<i>n</i> = 211)
COSM	336 (85.28%)	16 (3.92%)	7 (9.52%)
Cholesteatoma	44 (11.17%)	336 (82.35%)	2 (0.95%)
Normal	14 (3.55%)	20 (4.90%)	188 (89.52%)

The horizontal represents the actual diagnosis results, while the vertical depicts the diagnostic outcomes of the deep learning model.

medical experts as the training set for our algorithmic model. Our algorithm model accurately predicted unlabeled CT images with a high degree of precision, achieving excellent agreement between predicted lesion types and actual clinical findings.

So far, CT scan and Endoscopy of ear, as the classical methods for the diagnosis of various types of otitis media, are still the latest diagnostic methods (Gomaa et al., 2013; Zelikovich, 2004). The golden standard for the diagnosis of CSOM and middle ear cholesteatoma is intraoperative histopathological examination. However, it takes a long time to make a preliminary diagnosis of one single patient. Despite efforts to reduce the prevalence of Chronic Suppurative Otitis Media (CSOM) in underdeveloped areas, clinical diagnosis, treatment, and prognosis of the disease remain suboptimal. Recent epidemiological investigations have shown that CSOM has shifted to a population dominated by adults, despite a decrease in overall prevalence during the past 2 decades (Orji et al., 2016). Additionally, an Australian survey highlighted the high incidence of CSOM and middle ear cholesteatoma among impoverished individuals and the need for early diagnosis (Benson and Mwanri, 2012).

Deep learning has emerged as a valuable tool in various medical fields. A substantial amount of research on deep learning applied to clinical datasets, using high-quality medical examination images, has showcased its efficacy in defining patient categories, identifying and locating lesions, and other relevant tasks (Wang et al., 2020). With our transfer learning model, medical researchers can avoid the time-consuming and resource-intensive process of training models from scratch, while also benefiting from the wealth of knowledge captured in existing non-medical datasets. In the field of computational vision, pre-trained models have become a commonly used tool in many applications, particularly in addressing medical imaging challenges. These challenges can arise from imaging modalities such as X-ray, Magnetic Resonance Imaging (MRI), CT scan, and Ultrasound data. Many works have demonstrated the potential of pre-trained models to improve diagnostic accuracy, reduce processing time, and assist in the development of automated diagnosis systems. Our transfer learning model could also be used in other diseases which need CT scan or endoscope or any examinations that take images as the method to diagnose. Once the medical examination images are too similar to find the differences, our model could give several suggestions in differential diagnosis based on the previous history image labels.

Among the different types of otitis media, there are varying methods for diagnosis and treatment. However, the CT image features tend to be similar across these types, which can pose challenges for clinicians in terms of differential diagnosis. Such challenges can lead to delays in proper treatment, and potentially result in errors or overmedication. Moreover, the COVID-19

patients were found to have relationship with Otitis media (Choi et al., 2022), they demand to be diagnosed earlier than before, as otitis media always intend to recurrence and even cause Sensorineural-hearing-loss (Xia et al., 2022). As a result, achieving rapid differential diagnosis for otitis media is crucial to ensure optimal patient outcomes, in this situation, an efficient diagnosis can be given using our transfer learning model.

However, our transfer learning models have shown some limitations in classifying certain ear diseases. For instance, when differentiating between secretory otitis media and suppurative otitis media, deep learning models tend to confuse the two because their CT scans are very similar. This could be attributed to inadequate data sets. As such, there is a need for more comprehensive and diverse medical data to improve the accuracy of diagnostic models used to differentiate between various ear diseases. Moreover, our model showed significant differences in lesion information extraction. For instance, some predicted lesions would perform fewer or more lesions compared to those marked by medical experts. Also, in images with unclear lesions, there were discrepancies in identifying the lesion. For example, the images of the mastoid layer of chronic suppurative otitis media often have varying degrees of mucosal thickening due to chronic inflammation, while the images of the mastoid layer of chronic secretory otitis media show fluid levels caused by chronic effusion. These conditions are quite similar, with only slight differences in the contour of the mucosal within the mastoid bone. In most cases, our transfer learning networks could detect and label prominent lesions such as large soft tissue shadow of middle ear cholesteatoma, eroded bone structure surrounded by soft tissue shadow, and eroded bone structure of chronic suppurative otitis media. Unfortunately, the CT images of chronic otitis media with effusion (COME) do not have the typical erosive features of CSOM and middle ear cholesteatoma. Therefore, our research group excluded the CT images of chronic secretory otitis media and focused solely on collecting CT images of middle ear cholesteatoma and chronic suppurative otitis media as the objects of our study. In the future, when the amount of data collection is large enough, we will continue to promote the application of new migration models to this type of classification project.

What's more, due to the limited clinical applications of deep learning and the laborious, time-consuming nature of acquiring supervised data such as lesion regions, our research group aims to identify alternative weakly supervised signals for model transfer learning pre-training. "Human-in-the-loop" is an effective interactive mode between doctors and models, which can provide weakly supervised signals and ensure continuous learning of the model. This approach also represents a practical scenario for the clinical application of the model. This can help reduce manual labeling costs and improve overall prediction performance. In the

future, we hope to replace the repetitive, cumbersome task of manual labeling with more advanced machine automated labeling techniques in the field of deep learning combined with medicine.

Regarding the types of otological diseases combined with deep learning models, there are few studies on using detection results of acoustic immittance, acoustic reflex, and pure tone hearing threshold to achieve accurate predictions, prognosis, and treatment. Additionally, while otitis media and vertigo have received significant research attention, other diseases such as otosclerosis, ear tumors, and sudden neurotropic hearing loss remain understudied. Future research on combining otological diseases with computer learning may increase the coverage of relevant studies and provide clinicians with more precise and diverse tools for diagnosis and treatment.

Data availability statement

The datasets presented in this article are not readily available because The data is not publicly available. Requests to access the datasets should be directed to moondancer122@163.com.

Ethics statement

The studies involving humans were approved by the Scientific Research Sub-Committee of Medical Ethics Committee of Zhongshan Hospital Affiliated to Xiamen University. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because The data collected were CT images of historical medical records, which were non-invasive medical examination images. Written informed consent was not obtained from the minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article because The data collected were historical medical record CT images, which were

non-invasive medical examination images. All patients had signed waivers of informed consent with the ethics committee of the hospital, and the identity information of all patients was concealed.

Author contributions

QF performed the major jobs including literature search and data marking. YZ provided model training. All authors contributed to the article and approved the submitted version.

Funding

The project was supported by the Natural Science Foundation of Fujian Province of China (No. 2022J05006), Natural Science Foundation of Fujian Science and Technology (No. 2020J02060), Key Medical and Health Project of Xiamen Science and Technology Bureau (No. 3502Z20204009), Key Research and Development (Digital Twin) Program of Ningbo City (No. 2023Z219) and the Fundamental Research Funds for the Chinese Central Universities (No. 0070ZK1096 to JG).

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.

References

- Ahmad, R. U., Ashraf, M. F., Qureshi, M. A., Shehryar, M., Tareen, H. K., and Ashraf, M. A. (2022). Chronic Suppurative Otitis Media leading to cerebellar brain abscess, still a problem in 21st century: a case report. *Ann. Med. Surg. (Lond.)* 80, 104256. doi:10.1016/j.amsu.2022.104256
- Ahsan, M., Gomes, R., and Denton, A. (2019). "Application of a Convolutional Neural Network using transfer learning for tuberculosis detection," in 2019 IEEE International Conference on Electro Information Technology (EIT), USA, 20-22 May 2019 (IEEE). doi:10.1109/EIT.2019.8833768
- Benson, J., and Mwanri, L. (2012). Chronic suppurative otitis media and cholesteatoma in Australia's refugee population. *Aust. Fam. Physician* 41 (12), 978–980. doi:10.3316/informit.998418740667479
- Choi, S. Y., Yon, D. K., Choi, Y. S., Lee, J., Park, K. H., Lee, Y. J., et al. (2022). The impact of the COVID-19 pandemic on otitis media. *Viruses* 14 (11), 2457. doi:10.3390/v14112457
- Deepak, S., and Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* 111, 103345. doi:10.1016/j.combiomed.2019.103345
- Dube, S., Elsadon, S., Cloughesy, T. F., and Sinha, U. Content based image retrieval for MR image studies of brain tumors[J]. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2006, 1(1): 3337–3340. doi:10.1109/IEMBS.2006.260262
- Elfiky, A. A., Pany, M. J., Parikh, R. B., and Obermeyer, Z. (2018). Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. *JAMA Netw. open* 1 (3), e180926. doi:10.1001/jamanetworkopen.2018.0926
- Fukudome, S., Wang, C., Hamajima, Y., Ye, S., Zheng, Y., Narita, N., et al. (2013). Regulation of the angiogenesis of acquired middle ear cholesteatomas by inhibitor of DNA binding transcription factor. *JAMA Otolaryngology-Head Neck Surg.* 139 (3), 273–278. doi:10.1001/jamaoto.2013.1750
- Gomaa, M. A., Abdel Karim, A. R., Abdel Ghany, H. S., Elhiny, A. A., and Sadek, A. A. (2013). Evaluation of temporal bone cholesteatoma and the correlation between high resolution computed tomography and surgical finding. *Clin. Med. Insights Ear Nose Throat* 6, 21–28. Published 2013 Jul 23. doi:10.4137/CMENT.S10681
- Greenberg, J. S., and Manolidis, S. (2001). High incidence of complications encountered in chronic otitis media surgery in a U.S. metropolitan public hospital. *Otolaryngol. Head. Neck Surg.* 125 (6), 623–627. doi:10.1067/mhn.2001.120230
- Hutz, M. J., Moore, D. M., and Hotaling, A. J. Neurological complications of acute and chronic otitis media. *Curr. Neurology Neurosci. Rep.*, 2018, 18(3): 11–17. doi:10.1007/s11910-018-0817-7

- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. methods* 18 (2), 203–211. doi:10.1038/s41592-020-01008-z
- Jang, C. H., Choi, Y. H., Jeon, E. S., Yang, H. C., and Cho, Y. B. (2014). Extradural granulation complicated by chronic suppurative otitis media with cholesteatoma. *vivo* 28 (4), 651–655.
- Kaur, I., Goyal, J. P., and Singh, D. (2017). Prevalence of chronic suppurative otitis media in school going children of patiala district of Punjab, India. *J. Evol. Med. Dent. Sci.* 75, 5402–5407. doi:10.14260/Jemds/2017/1171
- Kemppainen, H. O., Puhakka, H. J., Laippala, P. J., Sipilä, M. M., Manninen, M. P., and Karma, P. H. (1999). Epidemiology and aetiology of middle ear cholesteatoma. *Acta oto-laryngologica* 119 (5), 568–572. doi:10.1080/00016489950180801
- Lustig, L. R., Limb, C. J., Baden, R., et al. (2018). *Chronic otitis media, cholesteatoma, and mastoiditis in adults*. MA: UpToDate Waltham. (citirano 145 2019).
- Madabhushi, A., and Lee, G. (2016). Image analysis and machine learning in digital pathology: challenges and opportunities. *Med. image Anal.* 33, 170–175. doi:10.1016/j.media.2016.06.037
- Orji, F. T., Ukaegbe, O., Alex-Okoro, J., Ofoegbu, V. C., and Okorafor, I. J. (2016). The changing epidemiological and complications profile of chronic suppurative otitis media in a developing country after two decades. *Eur. Arch. Otorhinolaryngol.* 273 (9), 2461–2466. Epub 2015 Nov 26. PMID: 26611685. doi:10.1007/s00405-015-3840-1
- Otten, F. W. A., and Grote, J. J. (1990). Otitis media with effusion and chronic upper respiratory tract infection in children: a randomized, placebo-controlled clinical study. *Laryngoscope* 100 (6), 627–633. doi:10.1288/00005537-199006000-00014
- Schilder, A. G. M., Marom, T., Bhutta, M. F., Casselbrant, M. L., Coates, H., Gisselsson-Solén, M., et al. (2017). Panel 7: otitis media: treatment and complications. *Otolaryngology-Head Neck Surg.* 156 (4_Suppl. 1), S88–S105. doi:10.1177/0194599816633697
- Sun, X. W., Zhang, J. J., Ding, Y. P., Dou, F. f., Zhang, H. b., Gong, K. b., et al. (2011). Efficacy of high-resolution CT in differential diagnosis of chronic suppurative otitis media and cholesteatoma otitis media by soft-tissue shadows. *Zhonghua er bi yan hou tou jing wai ke za zhi= Chin. J. Otorhinolaryngology Head Neck Surg.* 46 (5), 388–392. (Language: Chinese).
- Wang, Y. M., Li, Y., Cheng, Y. S., He, Z. Y., Yang, J. M., Xu, J. H., et al. Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography. *Ear Hear.*, 2020, 41(3): 669–677. doi:10.1097/AUD.0000000000000794
- World Health Organization (2004). *Chronic suppurative otitis media: burden of illness and management options*.
- Xia, A., Thai, A., Cao, Z., Chen, X., Chen, J., Bacacao, B., et al. (2022). Chronic suppurative otitis media causes macrophage-associated sensorineural hearing loss. *J. Neuroinflammation* 19 (1), 224. doi:10.1186/s12974-022-02585-w
- Yorgancılar, E., Yıldırım, M., Gun, R., Bakır, S., Tekin, R., Gocmez, C., et al. (2013). Complications of chronic suppurative otitis media: a retrospective review. *Eur. Archives Oto-rhino-laryngology* 270 (1), 69–76. doi:10.1007/s00405-012-1924-8
- Zelikovich, E. I., Vozmozhnosti, K. T., and Visochnoi Kosti, V. (2004). diagnostike khronicheskogo gnoynogo srednego otita i ego oslozhnenii [Potentialities of temporal bone CT in the diagnosis of chronic purulent otitis media and its complications]. *Vestn. Rentgenol. Radiol.* 1, 15–22. (Language: Russian).



OPEN ACCESS

EDITED BY

Huiyong Sun,
China Pharmaceutical University, China

REVIEWED BY

Padhmanand Sudhakar,
Kumaraguru College of Technology, India
Jianzhong Chen,
Shandong Jiaotong University, China

*CORRESPONDENCE

Marius Reto Bigler,
✉ mariusreto.bigler@insel.ch

RECEIVED 30 December 2023

ACCEPTED 04 April 2024

PUBLISHED 01 May 2024

CITATION

Bigler MR and Baum O (2024), Deep learning-based classification of the capillary ultrastructure in human skeletal muscles. *Front. Mol. Biosci.* 11:1363384. doi: 10.3389/fmolb.2024.1363384

COPYRIGHT

© 2024 Bigler and Baum. 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.

Deep learning-based classification of the capillary ultrastructure in human skeletal muscles

Marius Reto Bigler^{1*} and Oliver Baum²

¹Department of Cardiology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland,

²Institut für Physiologie, Charité–Universitätsmedizin Berlin, Berlin, Germany

Background: Capillary ultrastructure in human skeletal muscles is dynamic and prone to alterations in response to many stimuli, e.g., systemic pathologies such as diabetes mellitus and arterial hypertension. Using transmission electron microscopy (TEM) images, several studies have been conducted to quantify the capillary ultrastructure by means of morphometry. Deep learning techniques like convolutional neural networks (CNNs) are utilized to extract data-driven characteristics and to recognize patterns. Hence, the aim of this study was to train a CNN to identify morphometric patterns that differ between capillaries in muscle biopsies of healthy participants and patients with systemic pathologies for the purpose of hypothesis generation.

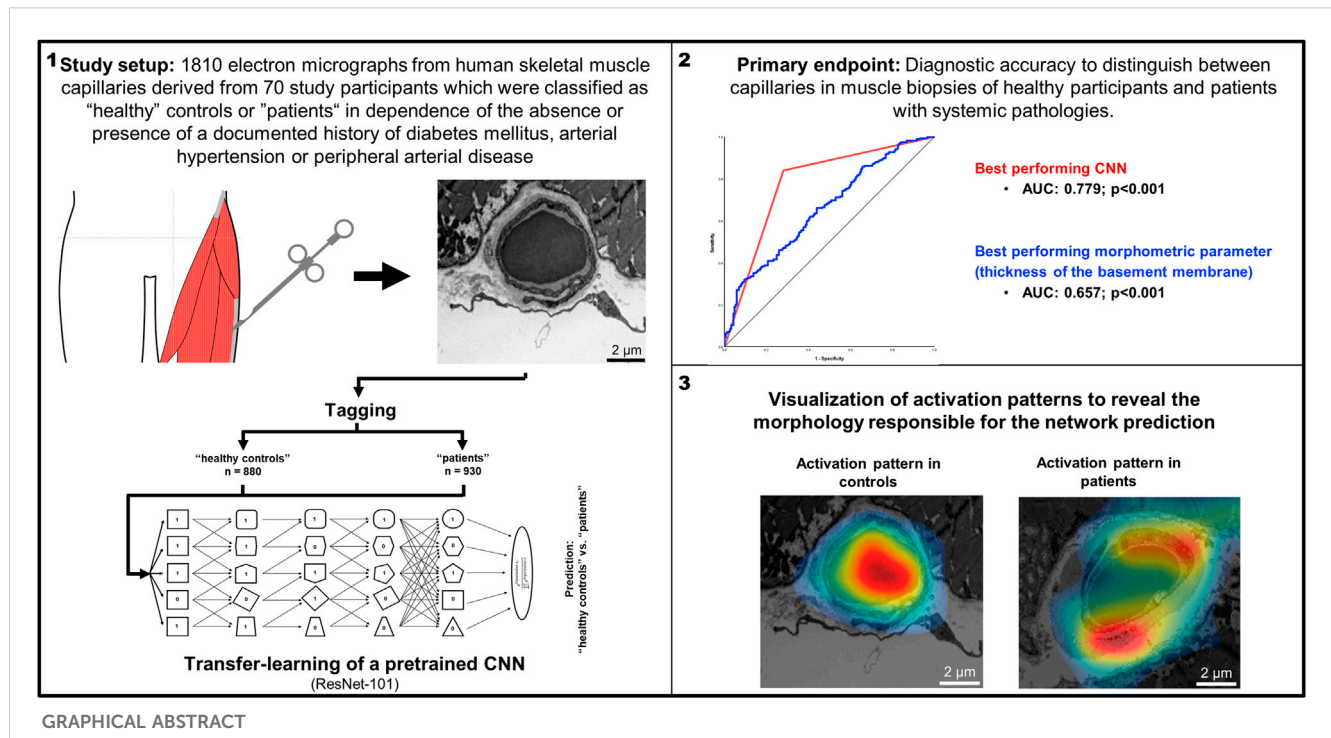
Methods: In this retrospective study we used 1810 electron micrographs from human skeletal muscle capillaries derived from 70 study participants which were classified as “healthy” controls or “patients” in dependence of the absence or presence of a documented history of diabetes mellitus, arterial hypertension or peripheral arterial disease. Using these micrographs, a pre-trained open-access CNN (ResNet101) was trained to discriminate between micrographs of capillaries of the two groups. The CNN with the highest diagnostic accuracies during training were subsequently compared with manual quantitative analysis of the capillary ultrastructure to distinguish between “healthy” controls and patients.

Results: Using classification into controls or patients as allocation reference, receiver-operating-characteristics (ROC)-analysis of manually obtained BM thickness showed the best diagnostic accuracy of all morphometric indicators (area under the ROC-curve (AUC): 0.657 ± 0.050). The best performing CNN demonstrated a diagnostic accuracy of 79% (sensitivity 93%, specificity 92%). DeLong-Test of the ROC-curves showed a significant difference ($p < 0.001$) between the AUC of the best performing CNN and the BM thickness. The underlying morphology responsible for the network prediction focuses mainly on debridement of pericytes.

Conclusion: The hypothesis-generating approach using pretrained CNN distinguishes between capillaries depicted on electron micrographs of “healthy” controls and participants with a systemic pathology more accurately than by commonly used morphometric analysis.

KEYWORDS

capillaries, skeletal muscle, transmission electron microscopy, convolutional neural networks, deep learning



Introduction

Capillaries are the sections of the vascular system with the most narrow diameter (Tuma et al., 2011). They branch from arterioles to meander through the tissues and then drain into collecting venules and ensuing veins. According to the law of Hagen-Poiseuille, which states that the blood flow velocity is proportional to the fourth power of the vessel radius, this transition from the arterioles into the capillary network is accompanied by a significant reduction in the velocity of the blood flow. Of note, as the transition into smaller vessels results in a significant increase in the overall diameter of the arterial vascular system, the total blood flow remains constant, i.e., the cardiac output. The reduction of the blood flow velocity in the capillary system ensures that the red blood cells release ample oxygen amounts to supply the surrounding tissues during their microcirculation passage and, in addition, facilitates the essentially balanced exchange of energy substrates and metabolic end products between the vascular system and the tissue.

As most clearly visualized using transmission electron microscopy (TEM), capillaries are of simple structure. Endothelial cells (ECs) close together as the vessel wall in such a way that a capillary lumen is formed. The abluminal surface of the ECs is covered by a continuous basement membrane (BM) mainly consisting of collagen type IV and other extracellular matrix (ECM) components such as laminin, heparan-sulphate proteoglycans (HSPGs) and nidogen/entactin (Kalluri, 2003). Pericytes (PC) are embedded in this BM and wrap their protrusions abuminally around the ECs. These contractile cells may influence the capillary blood flow in many tissues and communicate with the underlying ECs to influence the functional integrity of the capillaries (Armulik et al., 2011; Yamazaki and Mukoyama, 2018).

The capillary phenotype is dynamic. Inflation of the ECs volume during ischemia highlights the structural versatility of capillaries (Egginton and Hudlická, 1999). Furthermore, the thickness of the peri-capillary BM in human skeletal muscles increases in common cardiovascular diseases such as peripheral arterial disease (PAD), diabetes mellitus or arterial hypertension (Baum et al., 2020), but decreases in response to physical activity (Williamson et al., 1996). Strikingly, the BM thickening is accompanied by significant changes in the pathophysiology of the capillaries (Baum and Bigler, 2016).

Sophisticated methodological approaches have been developed in recent years that significantly improved the ultrastructural analysis by means of TEM been applied for more than 50 years. However and despite some simplifications (e.g., tablet-based image analysis (TBIA) (Bigler et al., 2016)), the quantitative evaluation of the images is still largely manually performed, posing a challenge for the morphometric processing of large amounts of data. In addition, the morphometry rules stipulate that the morphological features to be assessed are defined in advance, which means that changes in the capillary structure related to the pathophysiology could remain undetected during the analysis due to a selection bias. In contrast, deep learning methods such as convolutional neural networks (CNN) are not affected by this selection bias. Instead, the algorithm tries to find patterns in data sets to solve a pre-defined task without observer guidance (LeCun et al., 2015; Goodfellow et al., 2016).

We hypothesized that a deep learning-based approach with transfer learning of open-available, pre-trained CNN allows the identification of morphometric patterns that differ between capillaries in muscle biopsies of healthy participants and patients with systemic pathologies for the purpose of hypothesis generation. Thus, the aim of the study was to train a CNN and subsequently visualize its activation patterns to demonstrate the triggering

morphology for the network prediction. In a second step, we compared the results obtained applying the deep learning-based approach with data based on classic morphometry (i.e., the conventional method).

Methods

Study participants and muscle biopsies

For this retrospective study electron micrographs of capillaries were used, that were taken by transmission electron microscopy on biopsies of the vastus lateralis muscle (VL). The biopsies were derived from human participants of five studies conducted at the Department of Anatomy, University of Bern (Rosler et al., 1986; Suter et al., 1995), the University of Copenhagen (Nyberg et al., 2012; Winding et al., 2018), or the University of the sunshine Coast, Australia (Walker et al., 2016). Written informed consent was obtained in each case prior to the study beginning. In all investigations, the criteria and ethical guidelines for treatment of human participants conform to the principles outlined in the Declaration of Helsinki were fulfilled. Each study protocol was approved by the local ethics committee responsible for supervision at the time of study execution, as described earlier (Rosler et al., 1986; Suter et al., 1995; Nyberg et al., 2012; Hoier et al., 2013; Walker et al., 2016; Winding et al., 2018).

The VL muscle biopsies were taken by authorized medical practitioners using Bergstroem needles after local subcutaneous analgesia and immediately fixed in 6.25% (v/v) glutaraldehyde buffered with 0.1 M sodium cacodylate-HCl (pH 7.4) to be stored at 4°C until analysis. Ultrathin sections of the muscle biopsies were prepared and subjected to TEM analysis to record electron micrographs, as previously described in detail (Baum et al., 2020).

For this analysis, participants were classified as “healthy” controls or ‘patients’ in dependence of the absence or presence of a documented history of diabetes mellitus, PAD or arterial hypertension. Application of these criteria resulted in 42 controls and 28 patients, providing a total of 1810 electron micrographs of capillary profiles. In the patient group, 9 patients had a documented history of arterial hypertension, 10 patients had diabetes mellitus and 9 patients had clinically relevant PAD.

Capillary morphometry

Study parameters were adopted from the original studies, i.e., lumen radius (in nm), thickness of the endothelial cell (in nm), thickness of the BM (in nm) as well as capillary radius (in nm). Furthermore, all study parameters including pericyte cells were calculated as fraction of the capillary area (in %) (Baum and Bigler, 2016).

General principle of the applied deep-learning based method

The construction and training of a complex CNN architecture requires a large dataset and training over a considerable period of time, even for the establishment of general pattern recognition. To streamline this process, we employed openly accessible pre-trained

CNN models that were trained using an extensive dataset from the ImageNet Large Scale Visual Recognition Challenge (Deng, 2009; ILSVRC; <http://www.image-net.org/challenges/LSVRC/>). Consequently, this approach enables the utilization of a complex network architecture even with a limited dataset. However, a drawback of this method is the pre-defined input layer, which requires data adjustments such as resizing to match the selected networks. In a next step, the output layers of these pretrained CNN are replaced to fit the new task. After the training phase during which the CNN learns to perform the new task, its performance is evaluated using a new dataset. Subsequently, in the final step, the CNNs exhibiting the highest performance are subjected to additional analysis in order to visualize the specific regions within the images that contribute to the CNN’s decision-making process (i.e., the trigger morphology).

Computational hardware

Network training was simultaneously performed on two computers (Intel® Core™ i7-7700 CPU@3.60GHz, 8GB RAM respectively Intel® Core™ i7-8550U CPU@1.80GHz, 8GB RAM) using customized software (written in Matlab R2019b and R2020a).

Randomization, image allocation and preparation

Initially, randomization on participant level into training, validation and examination data (75% respectively 25% (validation + examination) as recommended by Goodfellow et al. (Goodfellow et al., 2016)) was performed using a random number vector to avoid overfitting of single participants, resulting in 52 participants in the training group, 13 in the validation group and five in the examination group. Electron micrographs (Figure 1 upper panel) were then plotted in Matlab, saved as jpg-images with predefined image size (224 × 224 × 3 pixels, Figure 1 lower panel) and stored in group-specific folders (880 control and 930 pathologic images).

Prior to each training iteration, all training images were randomly shuffled and processed by adding data noise to prevent overfitting (Goodfellow et al., 2016; Trask, 2020). Therefore, the images were randomly rotated in a range between ±45° and translocated ±10 pixels in each direction.

Selection and preparation of the pretrained convolutional neural networks

For this study, we applied ResNet101, a 101 convolutional layer deep CNN developed by He et al. (He et al., 2015). ResNet101 uses a special residual learning framework allowing the training of a deeper and thus more accurate network compared to other network architectures (i.e., GoogLeNet (Szegedy et al., 2014)) in terms of diagnostic accuracy.

To prepare for the transfer learning process, the last three layers of the networks responsible for the network prediction had to be replaced for the new task, i.e., classification of electron micrographs into either the control or the pathologic group. In addition, a

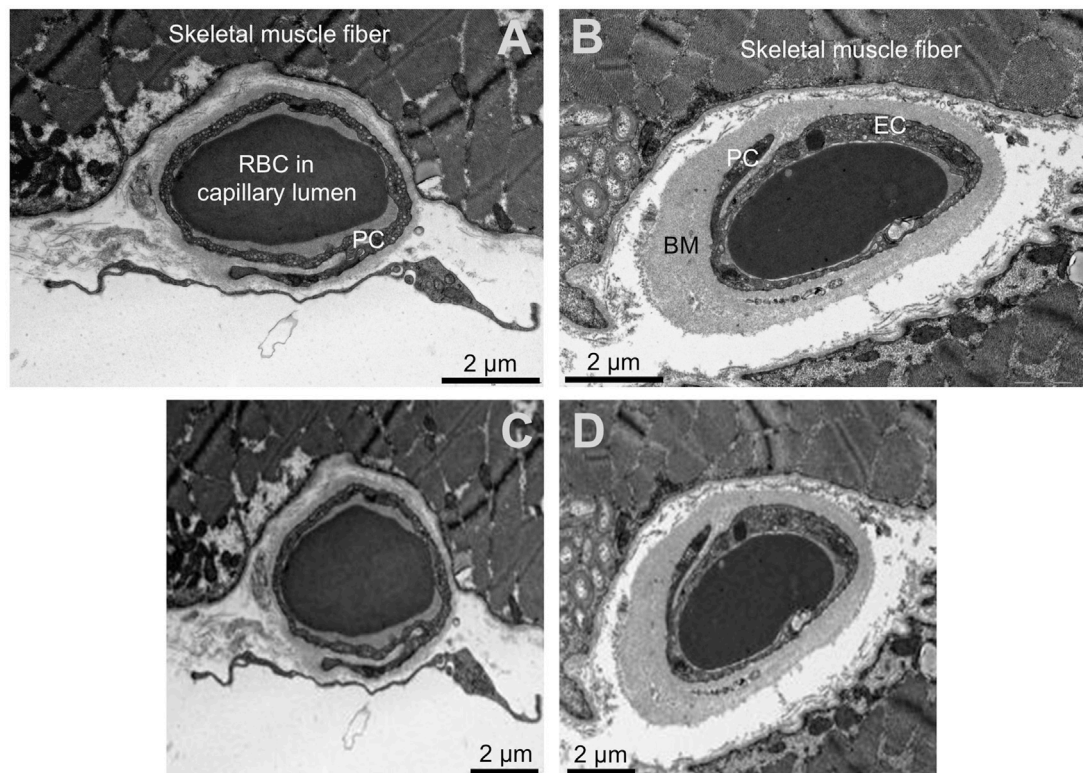


FIGURE 1
Input data for the neural networks. Upper panels (A, B): Original transmission electron microscopy images of the human skeletal muscle capillaries. (A) capillary of a healthy participant, (B) capillary of a patient with diabetes mellitus. Lower panels (C, D): Input images for the convolutional neuronal networks after required adjustment of the dimension. Please note the loss of resolution. The scale was not available for the networks and later added for better visualization. RBC = red blood cell PC = pericytes, EC = endothelial cell, BM = basement membrane.

dropout layer was added to prevent the network from overfitting (Srivastava et al., 2014). The remaining layers accounting for pattern recognition and feature extraction were not changed. General learning rate was chosen low while the new layers received a learning rate weight factor of 10 (i.e., 10-fold the normal learning rate) to improve and accelerate their training process.

Training of the neural networks

Hyperparameter optimization of the transfer learning process was performed for three parameters, i.e., learning rate, dropout probability and minibatch size, using Bayesian optimization technique (Bengio, 2000) and the adaptive moment estimation learning rate algorithm, ADAM (Diederik and Ba, 2014)). Validation of the network performance was performed every ten iterations. Further, a preliminary termination term was added to the algorithm, which terminated the training process when the validation and the training performances diverged twenty times in a row.

Network performance analysis

Networks performing above the arbitrary threshold of 60% classification accuracy (i.e. (true positive + true negative)/(true positive + true negative + false positive + false negative)) on the

validation data during the training process were stored for in-depth evaluation with determination of diagnostic accuracy on each subset (i.e., the validation data and the examination data) as well as the combined data sets. Based on the results of this evaluation, the best three networks were further evaluated with class activation mapping (CAM) (Zhou et al., 2015; Selvaraju et al., 2017), i.e., parametric visualization of their activation patterns to find the morphology responsible for the network prediction using ten characteristic electron micrographs (Supplementary Figure S2). When multiple networks showed similar performance, the network with the smallest discrepancy between validation and examination data was selected.

Statistical analysis

Two study groups (“healthy” and “patients” based on the above mentioned classification in controls or patients in dependence of the absence or presence of a documented history of diabetes mellitus, PAD or arterial hypertension) were formed. Between-group comparison of continuous study parameters was performed by an unpaired Student’s t-test. Network performance was analyzed by determination of classification accuracy (i.e., correct classified images/all images) using a 4-field matrix and calculation of sensitivity, specificity and F1-score (harmonic mean of sensitivity and positive predictive value). Nonparametric receiver operating

TABLE 1 Study parameters.

	Controls	Patients	<i>p</i> -value
Overall, n	880	930	-
Lumen radius (nm)	1,589 ± 421	1,452 ± 460	<i>p</i> < 0.001
Thickness of the endothelium (nm)	421 ± 303	481 ± 290	<i>p</i> < 0.001
Thickness of the basement membrane (nm)	218 ± 69	308 ± 118	<i>p</i> < 0.001
Capillary radius (nm)	2,406 ± 356	2,429 ± 440	<i>p</i> = 0.226
Training data	691	656	-
Lumen radius (nm)	1,589 ± 427	1,443 ± 464	<i>p</i> < 0.001
Thickness of the endothelium (nm)	423 ± 317	485 ± 294	<i>p</i> < 0.001
Thickness of the basement membrane (nm)	215 ± 67	319 ± 123	<i>p</i> < 0.001
Capillary radius (nm)	2,400 ± 361	2,442 ± 422	<i>p</i> = 0.047
Validation data	159	231	-
Lumen radius (nm)	1,576 ± 407	1,459 ± 459	<i>p</i> = 0.012
Thickness of the endothelium (nm)	419 ± 248	456 ± 276	<i>p</i> = 0.190
Thickness of the basement membrane (nm)	232 ± 79	287 ± 98	<i>p</i> < 0.001
Capillary radius (nm)	2,425 ± 342	2,388 ± 497	<i>p</i> = 0.411
Examination data	30	43	-
Lumen radius (nm)	1,672 ± 352	1,558 ± 393	<i>p</i> = 0.212
Thickness of the endothelium (nm)	375 ± 228	524 ± 290	<i>p</i> = 0.021
Thickness of the basement membrane (nm)	211 ± 50	252 ± 99	<i>p</i> = 0.037
Capillary radius (nm)	2,452 ± 304	2,449 ± 377	<i>p</i> = 0.971
Validation + Examination data	189	274	-
Lumen radius (nm)	1,591 ± 400	1,477 ± 449	<i>p</i> = 0.006
Thickness of the endothelium (nm)	412 ± 245	469 ± 279	<i>p</i> = 0.030
Thickness of the basement membrane (nm)	229 ± 75	281 ± 99	<i>p</i> < 0.001
Capillary radius (nm)	2,429 ± 336	2,397 ± 480	<i>p</i> = 0.428

characteristics (ROC) analysis was performed for accuracy assessment of differentiating between electron micrographs of controls or patients by manually obtained study parameters (continuous) and the CNN prediction (dichotomous). Comparison of the area under the ROC curves was performed using the DeLong-Test.

Statistical significance was defined at a *p*-level of <0.05. Continuous variables are given as mean ± standard deviation. All analyses were performed using SPSS version 25 (IBM Statistics, Armonk, New York) or MedCalc for Windows, version 19.1 (MedCalc Software, Ostend, Belgium).

Results

1810 electron micrographs from 70 participants were included in the study, among these 880 micrographs were derived from muscle

biopsies of 42 healthy control subjects and 930 from those of 28 patients. 1,347 were used for the training and 463 electron micrographs for the performance evaluation of the CNN (Table 1). Most of the participants were male (69%) with a mean age of 49.2 years (range 23–75 years). Of note, participants included in the patient group were significantly older than participants in the control group (57.6 years *versus* 43.9 years, *p* < 0.001).

Descriptive statistics

Descriptive statistics of the study parameters grouped according to the classification “healthy” and “patients” are presented in Table 1; Figure 2 (respectively Supplementary Table S1; Supplementary Figure S1 for the fraction values). Overall, endothelial cell thickness and BM thickness were significantly different between the groups in each data set.

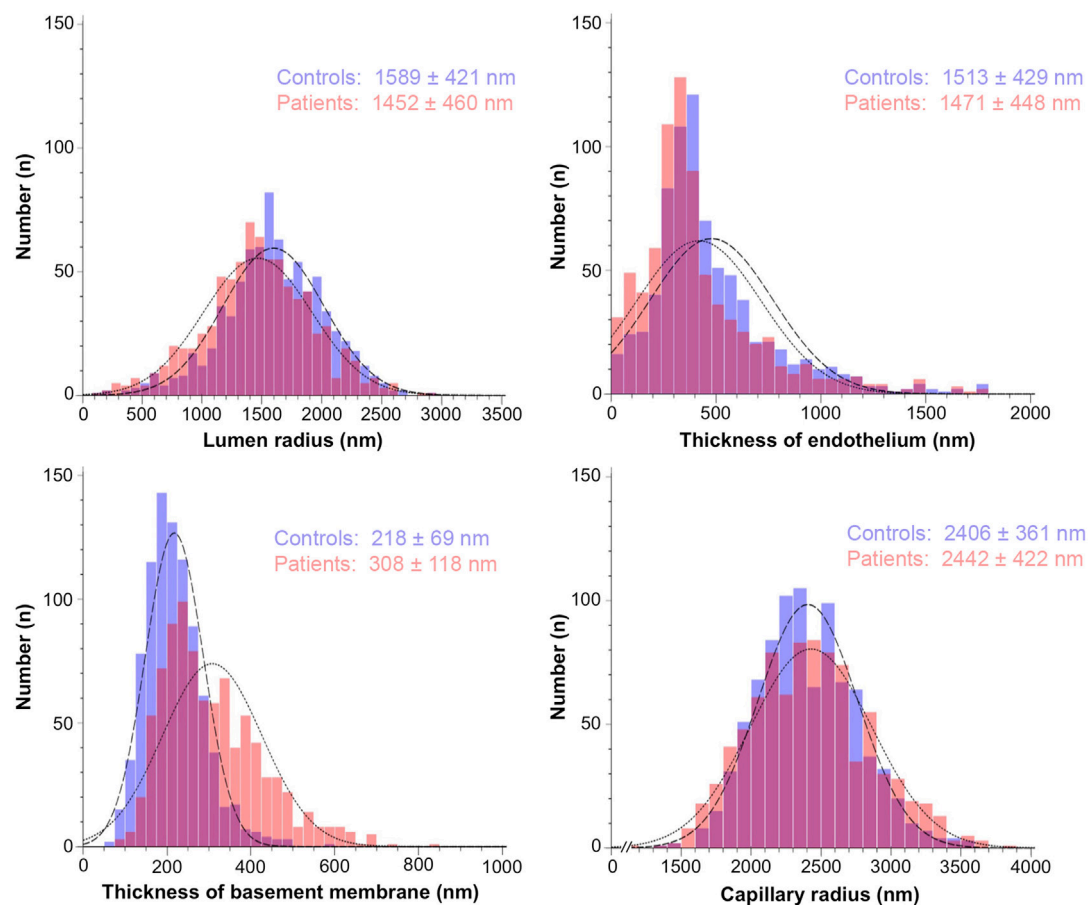


FIGURE 2

Histograms representing the frequency distribution of the study parameters grouped by the absence or presence of systemic pathologies. The morphometric values for the capillary structure of 880 electron micrographs from control participants (blue bars) and 930 electron micrographs from patients (red bars) were taken from the original studies listed in Materials/Methods. They were determined using tablet-based image analysis and represent mean ± standard deviation.

Receiver-operating characteristic curves

Complete data

Using classification into controls or patients as allocation reference, receiver-operating-characteristics (ROC) analysis of the lumen radius showed an area under the ROC-curve of 0.592 ± 0.027 ($p < 0.001$; Figure 3). AUC for endothelial thickness was 0.588 ± 0.027 ($p < 0.001$), for BM thickness 0.743 ± 0.023 ($p < 0.001$) and for the capillary radius 0.511 ± 0.027 ($p = 0.419$).

DeLong-Test of the ROC-curves showed a significant difference of AUC for BM thickness in comparison to all other parameters ($p \leq 0.0001$). There was no significant difference between the AUCs of the lumen radius and endothelial thickness ($p = 0.844$), but a significant difference of these parameters and capillary radius ($p = 0.013$ respectively $p \leq 0.001$).

Validation and examination data

Using classification into controls or patients as allocation reference, ROC analysis of the lumen radius showed an area

under the ROC-curve of 0.580 ± 0.054 ($p = 0.004$). AUC for endothelial thickness was 0.574 ± 0.055 ($p = 0.009$), for BM thickness 0.657 ± 0.050 ($p < 0.001$) and for the capillary radius 0.532 ± 0.053 ($p = 0.235$).

Regarding the optimum cut-off of the study parameters, a lumen radius of 1,332 nm distinguished best between control and patient, sensitivity 36%, specificity 80%. The best cut-off point for endothelial thickness was 368 nm (sensitivity 58%, specificity 88%), for BM-thickness 314 nm (sensitivity 32%, specificity 91%) and for capillary radius 2,182 nm (sensitivity 36%, specificity 80%). Of note, lumen and capillary radius decreased with presence of pathologies. Thus, the optimum cut-off points for these parameters were inversely set (i.e., pathologic below 1332nm respectively 2,182 nm). Using these thresholds obtained in the validation and examination data, diagnostic accuracy was calculated to allow a comparison with the CNN (Table 2). Of note, due to missing data for all but BM data, different (lower) diagnostic accuracies are shown than presented in the ROC analysis. Further, absolute numbers of the study parameters are dependent on biopsy fixation and storage and are not generally representative.

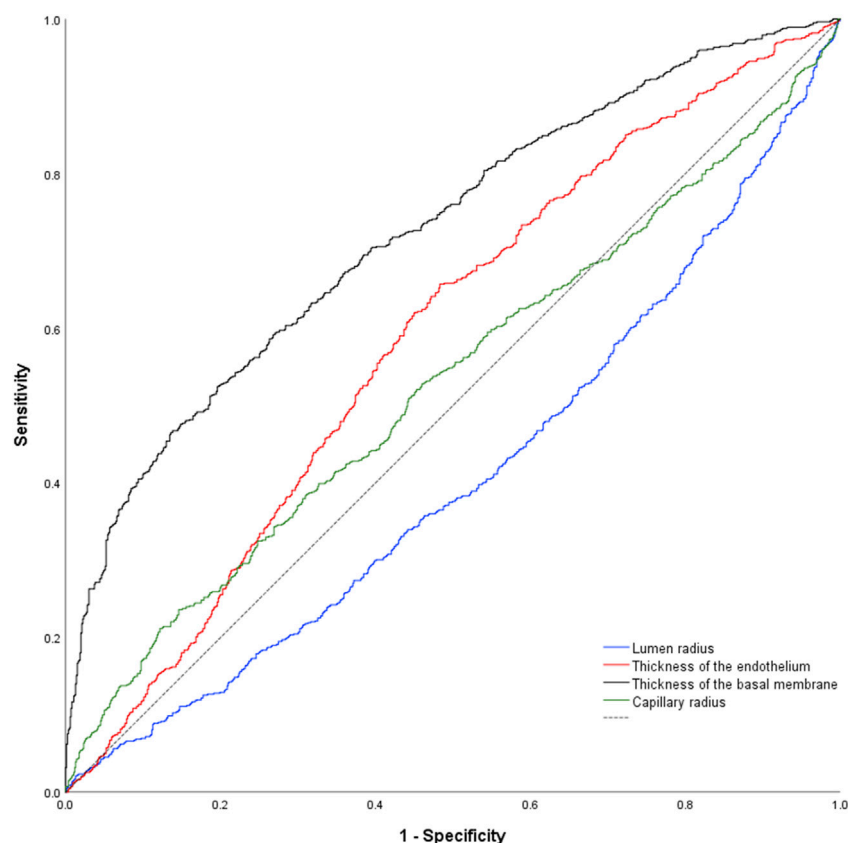


FIGURE 3

Nonparametric receiver-operating characteristic curve of the study parameters using the complete data set. Of note, all parameters but capillary radius were higher in samples of pathologies. As consequence, the data set for capillary radius is below the reference line (dashed black line).

Study parameter and network performance

Prediction of the study parameters and the three best performing networks including their accuracy, sensitivity and specificity are presented in Table 2. Using participant allocation as reference for the ROC analysis, the three best-performing networks showed a diagnostic accuracy of 79% (RN1: sensitivity 93%, specificity 92%, AUC 0.779 ± 0.023), 77% (RN2: sensitivity 88%, specificity 96, AUC 0.745 ± 0.024) and 75% (RN5: sensitivity 89%, specificity 95%, AUC 0.725 ± 0.025 ; Figure 4). By visualization of the activation patterns on ten characteristic electron micrographs, it could be shown that the underlying morphology responsible for the network prediction focuses primarily on debridement of pericytes and to a lesser extent on the structure of the endothelium. These network activation patterns are depicted in Figure 5, and in detail in Supplementary Figure S2.

Comparison of study parameter and network performance

Based on the performance of the different morphologic parameters as well as missing data for lumen radius, endothelial

thickness and capillary radius, only a comparison with BM thickness was performed.

DeLong-Test of the ROC-curves (Figure 4) showed significant difference of AUCs between BM thickness and the three networks (RN1: $p < 0.001$; RN2: $p = 0.002$; RN5: $p = 0.015$). Further, there was a significant difference between the AUCs of RN1 and RN5 ($p = 0.003$).

Discussion

In the present project, we used a deep learning-based approach with transfer learning of open-available pre-trained CNN to identify morphometric patterns that differ between capillaries in skeletal muscle biopsies of healthy participants and patients with systemic pathologies. Our most relevant findings were: 1. Electron micrographs of skeletal muscle capillaries from healthy controls and participants with a systemic pathology are more accurately distinguishable by CNN than by commonly used morphometric analysis. 2. The underlying morphology responsible for the network prediction focuses primarily on debridement of pericytes and to a lesser extent on the structure of the endothelium.

TABLE 2 Prediction and performance of the study parameters and the networks.

Data	Validation Data				Examination Data				Validation + Examination Data			
Parameter	True	Normal	Pathologic	Accuracy	True	Normal	Pathologic	Accuracy	True	Normal	Pathologic	Accuracy
	Predicted				Predicted				Predicted			
Lumen radius Cut-off: ≤1332nm	Normal	122	120	55.90	Normal	27	33	50.68	Normal	149	153	55.01
	Pathologic	37	77		Pathologic	3	10		Pathologic	40	87	
EC-Thickness Cut-off: ≥368nm	Normal	86	82	56.64	Normal	22	17	65.75	Normal	108	99	57.98
	Pathologic	72	113		Pathologic	8	26		Pathologic	80	139	
BM-Thickness Cut-off: ≥314 nm	Normal	141	155	55.64	Normal	30	31	57.53	Normal	171	186	55.94
	Pathologic	18	76		Pathologic	0	12		Pathologic	18	88	
Capillary radius Cut-off: ≤2182 nm	Normal	127	144	54.87	Normal	24	32	47.95	Normal	151	176	53.78
	Pathologic	32	87		Pathologic	6	11		Pathologic	38	98	
ResNet1:L4.4e-5_D0.69_M8	Normal	106	22	80.77	Normal	30	22	69.86	Normal	136	44	79.05
	Pathologic	53	209		Pathologic	0	21		Pathologic	53	230	
ResNet2:L4.8e-5_D0.32_M13	Normal	92	16	78.72	Normal	26	21	65.75	Normal	118	37	76.67
	Pathologic	67	215		Pathologic	4	22		Pathologic	71	237	
ResNet5:L1e-4DO51M31	Normal	82	13	76.92	Normal	25	19	67.12	Normal	107	32	75.38
	Pathologic	77	218		Pathologic	5	24		Pathologic	82	242	

Order according to accuracy. L = learning rate, D = dropout rate, M = minibatch size.
Of note, there were missing data for lumen radius, EC-Thickness and Capillary radius resulting in a smaller total number of cases.

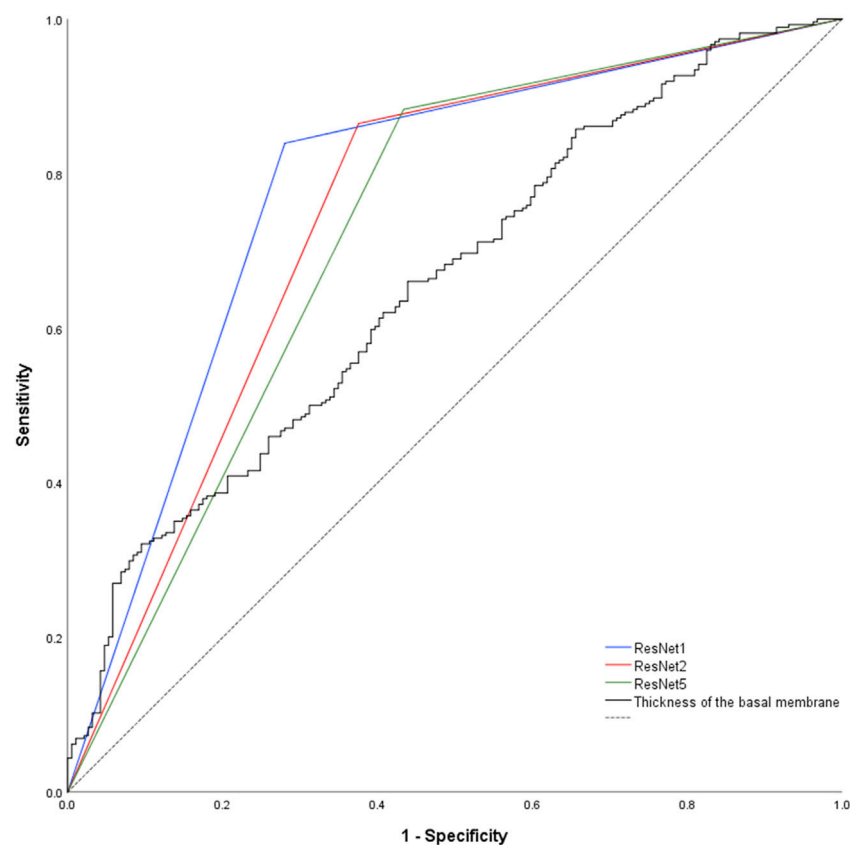


FIGURE 4

Nonparametric receiver-operating characteristic curve of the basement membrane thickness and the network predictions using the validation and examination data. Of note, network prediction provides a dichotomous output ("healthy control" respectively "patient"), resulting in a triangular ROC-curve. Hence, there is only one combination of sensitivity and specificity possible for each CNN. Dashed black line = reference line.

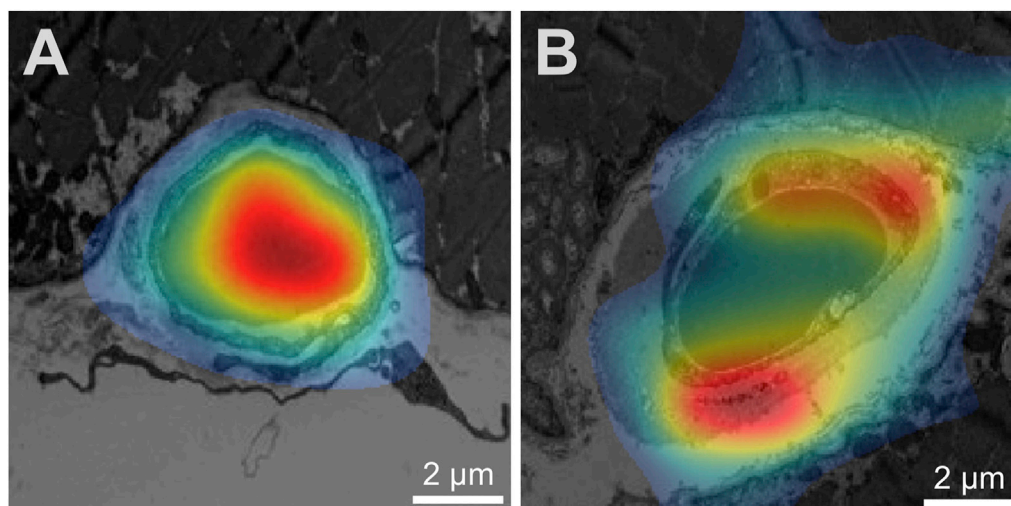


FIGURE 5

Visualization of network activation patterns of the best performing CNN (A) Capillary of a healthy participant, (B) capillary of a patient. Red regions contributed most to the network class prediction. Hence, the electron micrograph of the patient was recognized by the debridement of the pericyte (lower red region) and the thickness of the endothelium cell (upper red region). Of note, the scale was later added for better visualization.

Classification of electron micrographs by established morphometric parameters

Ever since the initial observation that certain pathologies are associated with morphological changes of the peri-capillary BM in skeletal muscles, there have been fundamental discussions regarding the methodology for the quantitative determination of this entity. Originally, the scientific debate was driven by methodological approaches developed by two research groups. Siperstein et al. (Siperstein et al., 1968) determined the capillary basement membrane thickness (CBMT) by calculation the mean of 20 measurements of the distances between the abluminal EC surface and the endomysium that do not intersect a PC profile ("20-line measurement"). On the other hand, Williamson et al. (Williamson et al., 1969) preferred the measurement of the CBMT at the two sites of the capillary profile where the BM appeared smallest ("two-minimum-point technique"). However, both morphometric methods are characterized by a time-intensive nature and exhibit specific technical limitations as previously discussed (Baum and Bigler, 2016). Therefore, given the technological advancements, a novel tablet-based image analysis (TBIA) methodology was developed to facilitate the precise quantitative assessment of CBMT (Bigler et al., 2016). Application of this approach allowed not only accurate and reproducible analysis of the CBMT, but also the assessment of numerous other structural indicators simultaneously during the same analysis. As a result, our study group could not only confirm the direct correlation between hypertension, diabetes mellitus, PAD or age with CBMT (Bigler et al., 2016), but also corroborate the favorable impact of physical exercise on CBMT with a partial reduction (Baum and Bigler, 2016).

Application of deep-learning on morphometric data/electron micrographs

An increasing number of research groups have applied deep-learning based methodologies in basic science. There, its application has spanned a wide spectrum, encompassing the identification of gold nanoparticles in TEM images of tumor cells (Kaphle et al., 2023), deep-learning assisted segmentation of atomic structures (Sadre et al., 2021), and translational research involving the correlation of deep learning-based kidney histomorphometry with patient data (Ginley et al., 2023). The wide array of applications underscores the versatility of this approach. However, to the best of our knowledge, this study presents the first application of a pretrained CNN-approach on TEM-images of the capillary ultrastructure in human skeletal muscles.

Comparison of CNN and established morphometric parameters

The primary finding of this study is that transfer learning of a pretrained CNN is accurate for allocating electron micrographs of human skeletal muscle capillaries to healthy controls or participants with a systemic pathology. Noteworthy, its diagnostic accuracy for this allocation is higher than the methods previously used and established morphometric indicators for the evaluation of

capillary ultrastructure. Using parametric visualization of the activation patterns, we could demonstrate that CNN focuses on distinctive features of the capillary ultrastructure, in particular debridement of pericytes.

Our findings are in agreement with the current hypothesis on the etiology of capillary BM thickening according to Tilton et al. (Tilton et al., 1981) and Vracko et al. (Vracko and Benditt, 1970). Based on their observation of widespread cellular debris within the thickened BM, they independently proposed a disturbed and incomplete turnover of cells associated with the capillaries including apoptosis and replacement of the degenerated cells by new pericyte precursor cells, which then differentiate and generate a new BM layer. Consequently, the inadequately regulated turnover of PCs results in an accumulation of BM material during each cycle. Of note, this hypothesis would provide an explanation for the frequently observed lamellar structure of the BM in capillary profiles of diabetic patients (Baum and Bigler, 2016), akin to growth rings of trees.

Despite this established hypothesis, a comprehensive quantitative assessment of cellular debris and its correlation with BM thickness have yet to be conducted. Hence and in light of the recent reaffirmation of this pathophysiological explanation by the present study, further quantitative analysis with focus on this phenomenon are required.

Limitations

The present study has several limitations. First, the categorization of the study participants into the distinct groups "healthy" and "patients" introduced heterogeneity into the data. Based on this dichotomy with its potential complexities, it is conceivable that there exists the potential for undetected arterial hypertension or incipient diabetes in the former group with already initiated microvascular changes. Conversely, pathological cases exhibiting optimal medical management may result in minimal pathophysiological alternations. In summary, variability could lead to a considerable degree of overlap between the groups. Nevertheless, due to the retrospective nature of the study design, the adjustment for these factors was not feasible.

Second, age was the only variable not accounted for by the study design excluding relevant co-morbidities during enrollment in the original studies. However, in a preliminary study by our research team, we could demonstrate that most markers of capillary ultrastructure exhibit only non-significant changes ($p > 0.05$) with age, except for the basement membrane thickness. This exception was attributed rather to an increase in age-related comorbidities (such as hypertension and diabetes), than to the aging process itself (Bigler et al., 2016).

Third, the utilization of transfer learning of a pretrained CNN facilitated the implementation of networks with a high capacity for small data. However, this advantage came at the cost of predefined inputs. In the presented study, this constraint led to a notable reduction in input dimensions and consequently, image resolution. As a result, it is conceivable that nuanced morphological patterns may have escaped detection by the network.

Fourth, the applied strategy with transfer-learning of a pretrained CNN without further adjustments to the output is

insufficient for the development of a diagnostic model. Therefore, analysis of the probabilities rather than the binary output, inclusion of the morphometric features as covariates as well as cross-validation would be required to gain prediction stability. Nonetheless, given that the aim of our study is to generate hypotheses that necessitate subsequent validation, the study design offers significant benefits in terms of its simplicity and ease of application.

Last, application of CNN results in a “complicated interconnected hierarchical representations of the training data to produce its predictions” (Lundervold and Lundervold, 2019). Thus, interpretation of these predictions remains intricate, even with the assistance of class activation maps, which provide insight into the general distinction procedure. In this study, the CNN exhibited an astonishing diagnostic accuracy, surpassing that of conventional morphometric parameters. Notwithstanding these results, the depicted activation maps demonstrated a diffuse activation pattern leading to indistinct predictions, which complicates the interpretation even further. However, the CNN’s performance remained consistent across various datasets and, importantly, the results substantiate a biologically plausible underlying pathophysiological mechanism.

Implications

In this study, the morphometric patterns employed by the CNN for distinguishing between TEM images of capillaries in muscle biopsies from healthy participants and patients with systemic pathologies were innovative, yet rooted in a plausible pathophysiological mechanism. This underscores the feasibility of a hypothesis-generating process using transfer learning of pretrained CNN on a small data set employing single CPU computers. Of note, this approach does not replace the conventional scientific method and further studies, i.e., the quantitative analysis of pericyte debridements across different pathologies, are required to validate the presented findings. However, the study highlights the feasibility of the proposed approach, making it applicable to a diverse range of scientific problems.

Conclusion

The presented hypothesis-generating approach using pretrained CNN distinguishes electron micrographs of healthy controls and participants with a systemic pathology more accurately than established morphometric analysis. Of note, in addressing this task, the CNN primarily concentrates on debridements of pericytes and thus, a biological plausible mechanism. Hence, demonstrating the feasibility of the hypothesis-generating approach in pretrained CNN on a small data set. However, further quantitative and prospective analyses are required to validate these findings.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

This retrospective study used data (TEM images of biopsies) from studies involving humans. The biopsies were derived from human participants of five studies conducted at the Department of Anatomy, University of Bern (Rosler et al., 1986; Suter et al., 1995), the University of Copenhagen (Nyberg et al., 2012; Winding et al., 2018), or the University of the sunshine Coast, Australia (Walker et al., 2016). Written informed consent was obtained in each case prior to the study beginning. In all investigations, the criteria and ethical guidelines for treatment of human participants conform to the principles outlined in the Declaration of Helsinki were fulfilled. Each study protocol was approved by the local ethics committee responsible for supervision at the time of study execution, as described earlier (Rosler et al., 1986; Suter et al., 1995; Nyberg et al., 2012; Hoier et al., 2013; Walker et al., 2016; Winding et al., 2018).

Author contributions

MB: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review and editing. OB: Data curation, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2024.1363384/abstract#supplementary-material>

References

- Armulik, A., Genové, G., and Betsholtz, C. (2011). Pericytes: developmental, physiological, and pathological perspectives, problems, and promises. *Dev. Cell* 21, 193–215. doi:10.1016/j.devcel.2011.07.001
- Baum, O., Bernd, J., Becker, S., Odriozola, A., Zuber, B., Tschanz, S. A., et al. (2020). Structural microangiopathies in skeletal muscle related to systemic vascular pathologies in humans. *Front. Physiol.* 11, 28. doi:10.3389/fphys.2020.00028
- Baum, O., and Bigler, M. (2016). Pericapillary basement membrane thickening in human skeletal muscles. *Am. J. physiology. Heart circulatory physiology* 311, H654–H666. doi:10.1152/ajpheart.00048.2016
- Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural Comput.* 12, 1889–1900. doi:10.1162/089976600300015187
- Bigler, M., Koutsantonis, D., Odriozola, A., Halm, S., Tschanz, S. A., Zakrzewicz, A., et al. (2016). Morphometry of skeletal muscle capillaries: the relationship between capillary ultrastructure and ageing in humans. *Acta Physiol.* 218, 98–111. doi:10.1111/apha.12709
- Diederik, P. K., and Ba, J. (2014). Adam: a method for stochastic optimization. *CoRR*. doi:10.48550/arXiv.1412.6980
- Egginton, S., and Hudlická, O. (1999). Early changes in performance, blood flow and capillary fine structure in rat fast muscles induced by electrical stimulation. *J. physiology* 515 (Pt 1), 265–275. doi:10.1111/j.1469-7793.1999.265ad.x
- Gunley, B., Zee, J., Mimar, S., Paul, A. S., Jain, S., Rodrigues, L., et al. (2023). Correlating deep learning-based automated reference kidney histomorphometry with patient demographics and creatinine. *bioRxiv* 4, 1726–1737. doi:10.34067/KID.000000000000299
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*. abs/1512.03385. doi:10.48550/arXiv.1512.03385
- Hoier, B., Walker, M., Passos, M., Walker, P. J., Green, A., Bangsbo, J., et al. (2013). Angiogenic response to passive movement and active exercise in individuals with peripheral arterial disease. *J. Appl. Physiol.* (1985) 115, 1777–1787. doi:10.1152/japplphysiol.00979.2013
- Kalluri, R. (2003). Basement membranes: structure, assembly and role in tumour angiogenesis. *Nat. Rev. Cancer* 3, 422–433. doi:10.1038/nrc1094
- Kaphle, A., Jayarathna, S., Moktan, H., Aliru, M., Raghuram, S., Krishnan, S., et al. (2023). Deep learning-based TEM image analysis for fully automated detection of gold nanoparticles internalized within tumor cell. *Microsc. Microanal.* 29, 1474–1487. doi:10.1093/micmic/ozad066
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539
- Lundervold, A. S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Z. für Med. Phys.* 29, 102–127. doi:10.1016/j.zemedi.2018.11.002
- Nyberg, M., Jensen, L. G., Thaning, P., Hellsten, Y., and Mortensen, S. P. (2012). Role of nitric oxide and prostanoids in the regulation of leg blood flow and blood pressure in humans with essential hypertension: effect of high-intensity aerobic training. *J. physiology* 590, 1481–1494. doi:10.1113/jphysiol.2011.225136
- Rosler, K., Conley, K. E., Howald, H., Gerber, C., and Hoppeler, H. (1986). Specificity of leg power changes to velocities used in bicycle endurance training. *J. Appl. physiology (Bethesda, Md, 1985)* 61, 30–36. doi:10.1152/jappl.1986.61.1.30
- Sadre, R., Ophus, C., Butko, A., and Weber, G. H. (2021). Deep learning segmentation of complex features in atomic-resolution phase-contrast transmission electron microscopy images. *Microsc. Microanal.* 27, 804–814. doi:10.1017/S1431927621000167
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). “Grad-CAM: visual explanations from deep networks via gradient-based localization,” in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 Oct. 2017 (IEEE), 618–626. doi:10.1109/ICCV.2017.74
- Siperstein, M. D., Unger, R. H., and Madison, L. L. (1968). Studies of muscle capillary basement membranes in normal subjects, diabetic, and prediabetic patients. *J. Clin. investigation* 47, 1973–1999. doi:10.1172/JCI105886
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). *J. Mach. Learn. Res.* 15, 1929–1958. doi:10.48550/arXiv.1409.4842
- Suter, E., Hoppeler, H., Claassen, H., Billeter, R., Aebi, U., Horber, F., et al. (1995). Ultrastructural modification of human skeletal muscle tissue with 6-month moderate-intensity exercise training. *Int. J. Sports Med.* 16, 160–166. doi:10.1055/s-2007-972985
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2014). Going deeper with convolutions. *CoRR*. abs/1409.4842.
- Tilton, R. G., Hoffmann, P. L., Kilo, C., and Williamson, J. R. (1981). Pericyte degeneration and basement membrane thickening in skeletal muscle capillaries of human diabetics. *Diabetes* 30, 326–334. doi:10.2337/diab.30.4.326
- Trask, A. W. (2020). *Neuronale Netze und Deep Learning kapieren*. mitp.
- Tuma, R. F., Duran, W. N., and Ley, K. (2011). *Microcirculation*. Academic Press.
- Vracko, R., and Benditt, E. P. (1970). Capillary basal lamina thickening. Its relationship to endothelial cell death and replacement. *J. Cell Biol.* 47, 281–285. doi:10.1083/jcb.47.1.281
- Walker, M. A., Hoier, B., Walker, P. J., Schulze, K., Bangsbo, J., Hellsten, Y., et al. (2016). Vasoactive enzymes and blood flow responses to passive and active exercise in peripheral arterial disease. *Atherosclerosis* 246, 98–105. doi:10.1016/j.atherosclerosis.2015.12.029
- Williamson, J. R., Hoffmann, P. L., Kohrt, W. M., Spina, R. J., Coggan, A. R., and Holloszy, O. (1996). Endurance exercise training decreases capillary basement membrane width in older nondiabetic and diabetic adults. *J. Appl. physiology (Bethesda, Md, 1985)* 80, 747–753. doi:10.1152/jappl.1996.80.3.747
- Williamson, J. R., Vogler, N. J., and Kilo, C. (1969). Estimation of vascular basement membrane thickness. Theoretical and practical considerations. *Diabetes* 18, 567–578. doi:10.2337/diab.18.8.567
- Winding, K. M., Munch, G. W., Iepsen, U. W., Van Hall, G., Pedersen, B. K., and Mortensen, S. P. (2018). The effect on glycaemic control of low-volume high-intensity interval training versus endurance training in individuals with type 2 diabetes. *Diabetes, Obes. metabolism* 20, 1131–1139. doi:10.1111/dom.13198
- Yamazaki, T., and Mukoyama, Y. S. (2018). Tissue specific origin, development, and pathological perspectives of pericytes. *Front. Cardiovasc. Med.* 5, 78. doi:10.3389/fcvm.2018.00078
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Learning deep features for discriminative localization. *CoRR*. abs/1512.04150. doi:10.48550/arXiv.1512.04150



OPEN ACCESS

EDITED BY

Huiyong Sun,
China Pharmaceutical University, China

REVIEWED BY

Pramodkumar Pyarelal Gupta,
Padmashree Dr. D.Y. Patil University, India
Noureddin Saleh,
Schrodinger, United States

*CORRESPONDENCE

Zhongheng Zhang,
✉ zhongheng_zhang@zju.edu.com
Tao Shen,
✉ 11718228@zju.edu.cn

RECEIVED 19 April 2024

ACCEPTED 23 May 2024

PUBLISHED 11 June 2024

CITATION

Zeng Q, Hu H, Huang Z, Guo A, Lu S, Tong W,
Zhang Z and Shen T (2024), Active and machine
learning-enhanced discovery of new
FGFR3 inhibitor, Rhapontin, through virtual
screening of receptor structures and anti-
cancer activity assessment.
Front. Mol. Biosci. 11:1413214.
doi: 10.3389/fmolb.2024.1413214

COPYRIGHT

© 2024 Zeng, Hu, Huang, Guo, Lu, Tong, Zhang
and Shen. 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.

Active and machine learning-enhanced discovery of new FGFR3 inhibitor, Rhapontin, through virtual screening of receptor structures and anti-cancer activity assessment

Qingxin Zeng¹, Haichuan Hu¹, Zhengwei Huang¹, Aotian Guo¹,
Sheng Lu¹, Wenbin Tong², Zhongheng Zhang^{3*} and Tao Shen^{1*}

¹Department of Thoracic Surgery, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China, ²Department of Thoracic Surgery, Longyou County People's Hospital, Hangzhou, China, ³Department of Emergency Medicine, Sir Run Run Shaw Hospital, Hangzhou, China

Introduction: This study bridges traditional remedies and modern pharmacology by exploring the synergy between natural compounds and Ceritinib in treating Non-Small Cell Lung Cancer (NSCLC), aiming to enhance efficacy and reduce toxicities.

Methods: Using a combined approach of computational analysis, machine learning, and experimental procedures, we identified and analyzed PD173074, Isoquercitrin, and Rhapontin as potential inhibitors of fibroblast growth factor receptor 3 (FGFR3). Machine learning algorithms guided the initial selection, followed by Quantitative Structure-Activity Relationship (QSAR) modeling and molecular dynamics simulations to evaluate the interaction dynamics and stability of Rhapontin. Physicochemical assessments further verified its drug-like properties and specificity.

Results: Our experiments demonstrate that Rhapontin, when combined with Ceritinib, significantly suppresses tumor activity in NSCLC while sparing healthy cells. The molecular simulations and physicochemical evaluations confirm Rhapontin's stability and favorable interaction with FGFR3, highlighting its potential as an effective adjunct in NSCLC therapy.

Discussion: The integration of natural compounds with established cancer therapies offers a promising avenue for enhancing treatment outcomes in NSCLC. By combining the ancient wisdom of natural remedies with the precision of modern science, this study contributes to evolving cancer treatment paradigms, potentially mitigating the side effects associated with current therapies.

KEYWORDS

NSCLC, FGFR3 inhibitor, AI-derived drug discovery, Rhapontin, biology evaluation

1 Introduction

Lung cancer, in its various forms, poses a severe global health challenge. Accounting for approximately 85% of all lung cancer cases, non-small cell lung cancer (NSCLC) reigns as the most prevalent form of this malignancy (Sung et al., 2021). As per recent global statistics, NSCLC maintains a distressingly high mortality rate, cementing its position as one of the leading contributors to cancer-related deaths worldwide (Siegel et al., 2024). Concurrently, the incidence rate for NSCLC continues on an upward trend. Despite significant strides in diagnostic technologies and therapeutic methodologies, the prognosis for NSCLC remains bleak, with a 5-year survival rate barely reaching the 20% threshold (de Groot et al., 2018). The persistence of this grim statistic highlights the urgency for development of more effective therapeutic strategies in the battle against NSCLC.

A potentially promising approach in ameliorating treatment efficacy involves sensitizing cancer cells to extant therapeutic agents. Amidst the plethora of targets, the spotlight has recently shifted towards the Fibroblast Growth Factor Receptor 3 (FGFR3) (Turner and Grose, 2010). FGFR3, part of the larger fibroblast growth factor receptor family, has been implicated in numerous cellular processes, including cell proliferation, survival, and differentiation (Wesche et al., 2011). Recent studies suggest that modulation of FGFR3 activity could potentially augment the effectiveness of existing treatments, such as Ceritinib—an ALK (Anaplastic Lymphoma Kinase) inhibitor (Tang et al., 2008; Zhang et al., 2013). Despite this promising insight, therapeutic combinations incorporating FGFR3 inhibition to augment sensitivity of NSCLC cells to Ceritinib remain largely unexplored.

Historically, the drug development process has heavily leaned on experimental methodologies. Whilst these approaches have their merit, they come saddled with a suite of limitations. For instance, these traditional strategies often prove to be labor-intensive and time-consuming (Munos, 2009). Moreover, their applicability to high-throughput screening is limited, thereby underscoring the need for more efficient methodologies (Paul et al., 2010). Enter the realm of computational biology, which offers a more expedient alternative to traditional strategies. With the ability to conduct *in silico* screenings of expansive compound libraries, computational approaches promise significant savings in terms of both time and resources (Ekins et al., 2007; Green, 2008). These techniques enable prediction of interactions between small molecules and protein targets, thus providing preliminary insights into the potential efficacy and toxicity of candidate inhibitors. Complementing this, molecular dynamics (MD) simulations furnish a more granular understanding of the behaviour of protein-ligand complexes over time, thereby enhancing our grasp of the binding process (Dror et al., 2012; Arnittali et al., 2019).

Herein, we propose a melding of virtual screening and MD simulations as an integrative approach to identifying prospective FGFR3 inhibitors. Our overarching goal is to enhance the sensitivity of NSCLC cells to Ceritinib, offering a potentially viable strategy to circumvent the common therapeutic resistance observed in NSCLC. This innovative methodology presents a novel angle to the design of inhibitors, potentially paving the way for breakthrough combination therapies for NSCLC (Brown and Toker, 2015; Colmegna et al., 2018). By boosting the efficacy of treatment regimens, such therapeutic strategies have the potential to significantly enhance the prognosis for NSCLC patients, impacting a large patient population worldwide.

2 Method

2.1 Structure relaxation

In the pursuit of FGFR3 inhibitors, structure-based computational methodologies were employed, utilizing the FGFR3 crystal structure (PDB code: 6LVM) in complex with Pyrimidine Derivative 37b was selected as the receptor protein (Kuriwaki et al., 2020). All molecular dynamics (MD) simulations presented in this study were conducted using the GROMACS 23.1 package (<https://www.gromacs.org/>). The AMBER 99SB-ILDN (Lindorff-Larsen et al., 2010) and explicit solvation were employed, and each system was placed in a rectangular box of SPC water molecules with a minimum distance of 10Å between any solute atom and the edges of the periodic box. Counter ions were added to neutralize the total charge of the system. The system underwent an energy minimization process using the steepest descent method, with the maximum set to 1000.0 kJmol⁻¹nm⁻¹. Subsequently, the system was equilibrated in two steps: 1) canonical ensemble (NVT, 1ns) and 2) isothermal-isobaric ensemble (NPT, 1ns). Following equilibration, the MD simulations were run for 500ns. To ensure numerical stability, all bonds involving hydrogen atoms were constrained using the default linear constraint solver algorithm (LINCS) (Hess, 2008). The Vrescale thermostat and Parrinello–Rahman barostat were utilized with the temperature set at 300 K and pressure at 1.0bar, with time constants of 0.1 and 2ps, respectively. The Particle-Mesh Ewald (PME) method was employed to handle long-range interactions, and a 10Å cutoff was utilized for van der Waals interactions (Darden et al., 1993). The time step was set to 2 fs, and a snapshot was collected every 1.0 ps. The free energy landscape (Malmstrom et al., 2015) was obtained by means of covariance matrix construction and principal component analysis (PCA) (Campitelli et al., 2021) to explore the local conformational landscape and return to a local energy minimum.

2.2 Protein preparation

The Schrödinger Protein Preparation Wizard was employed to meticulously prepare the complex, involving various steps such as adding missing hydrogen atoms, correcting metal ionization states, enumerating bond orders in HET groups, determining ligand protonation states and associated energy penalties, optimizing histidine residues' protonation states, rectifying potentially transposed heavy atoms, optimizing the protein's hydrogen bond network, and performing a restrained minimization. The binding region within the 3D receptor structure, where the Pyrimidine Derivative 37b binds, was identified as the screening ligands' target site, and a corresponding grid was created.

2.3 Active learning based virtual screening

Active Learning Glide will generate a receptor grid from a prepared protein and prepare the TargetMol Natural Compound Library, which contains approximately 190,000 compounds. All of these compounds are available for purchase. It will also dock a subset

of these ligands using Glide SP (Friesner et al., 2004). Active Learning workflows train a machine learning (ML) model on physics-based data, such as FEP+ (Wang et al., 2015) predicted affinities or Glide docking scores, iteratively sampled from a full library using Schrödinger's deep-learning powered QSAR platform, DeepAutoQSAR (<https://www.schrodinger.com/science-articles/benchmark-study-deepautoqsar-chemprop-and-deeppurpose-admet-subset-therapeutic-data>). 3 iterative training rounds were set. After all the ligands have been screened using the last model, a selection of the top ligands will then be docked using Glide SP.

2.4 Machine learning principles using AutoQSAR

AutoQSAR is a machine-learning algorithm provided by the Schrödinger suite that builds and applies QSAR models through automation (Dixon et al., 2016). In order to build a predictive model, AutoQSAR takes the one-, two-, and three-dimensional structural data of a molecule along with a IC_{50} property to be modeled as an input. It will then compute the fingerprints and descriptors, using machine-learning statistical methods to create a predictive QSAR model. The process utilizes multiple regression algorithms, including optimal subset multiple linear regression (MLR), partial least squares regression (PLS), kernel-based least squares regression (KPLS), and principal component regression (PCR), to construct numerical models. The predictive accuracy of the model is evaluated using various parameters such as ranking score, root mean square error (RMSE), standard deviation (SD), Q^2 , and R^2 values (de Oliveira and Katekawa, 2018). It is worth mentioning that the present analysis utilizes a series of Pyrimidine Derivative 37b (Kuriwaki et al., 2020) and some clinically oriented medicines from Drugbank for predictive model development.

2.5 Binding pose metadynamics

The metadynamics simulations employed a hill height of 0.05 kcal/mol and a width of 0.02 Å. RMSD calculations were performed by considering a distance of 3 Å between protein residues and ligands. Prior to the metadynamics simulations, the system underwent preparation in an SPC water box, followed by energy minimization, constraint application, and a gradual temperature increase to 300 K. The last 0.5 nanoseconds of an unbiased MD simulation served as the reference for the subsequent metadynamics protocol.

Three BPMD scores, namely, PoseScore, PersScore, and CompScore, were utilized to assess the stability of ligand binding. PoseScore represented the average RMSD from the ligand's initial pose, where a steeper increase indicated instability in ligand binding. A PoseScore below 2 Å was considered indicative of a stable ligand-protein complex (Fusani et al., 2020). PersScore quantified the persistence of hydrogen bonds (HB) during the metadynamics simulations, with higher values indicating greater stability. Finally, the CompScore, a composite score, was obtained by linearly combining the PoseScore and PersScore (Jin et al., 2023). Lower CompScore values were associated with more stable ligand-protein complexes.

2.6 Physicochemical property and medicinal chemistry property prediction

The most promising compounds, identified through structure-based virtual screening, underwent further evaluation using ADMETlab 2.0 (Xiong et al., 2021). The analysis aimed to provide valuable insights into the compounds' pharmacokinetic properties, bioavailability, and overall suitability as potential drug candidates.

2.7 Molecular dynamic simulation of desmond

In the initial phase, all-atom molecular dynamics (MD) simulations were conducted using the Desmond module of the Schrödinger software package. The simulations were performed within Maestro, starting with docked complexes that were placed in a cubic water box with a buffer distance of 10 Å. The systems were solvated with SPC water models, and a 0.15 M NaCl salt concentration was introduced for physiological relevance. To maintain system neutrality, additional Na^+ and Cl^- ions were included. Long-range electrostatic interactions were computed using the particle-mesh Ewald method, while short-range van der Waals and Coulomb interactions were cutoff at 9.0 Å.

Following solvation, the systems underwent minimization and equilibration using the default Desmond protocol in Maestro. This involved restrained simulations in both the NVT (constant number of particles, volume, and temperature) and NPT (constant number of particles, pressure, and temperature) ensembles. After equilibration, a 100 ns MD simulation was performed in the NPT ensemble with periodic boundary conditions. The OPLS4 force field was employed to describe interatomic interactions. The temperature was maintained at 300 K using the Nosé-Hoover chain thermostat, and the pressure was kept at 1 atm using the Martyna-Tobias-Klein barostat method.

2.8 Cell culture

Two distinct cell lines were employed for the experimentation: A549 cells, characterized as an adenocarcinoma human lung epithelial cell line, and BEAS-2B cells, identified as a human bronchial epithelial cell line. These cell lines were sourced from iCell Bioscience Inc. Located in Shanghai, China. Both A549 and BEAS-2B cell lines have been authenticated using short tandem repeat (STR) analysis. A549 cells were nurtured using Ham's F-12K (Kaighn's) medium, while BEAS-2B cells were cultivated in Dulbecco modified Eagle's medium (DMEM). In both cases, the culture mediums were supplemented with 10% exosome-depleted fetal bovine serum (EXO-FBS-50A-1) from System Biosciences, Palo Alto, CA, to eliminate potential interference from bovine exosomes. Additionally, a 1% penicillin-streptomycin solution (Tianhang Biotechnology, Hangzhou, China) was added. The cells were incubated under controlled conditions at 37°C within a 5% CO_2 atmosphere (Exosomes of A549 Cells Induced Migration, Invasion, and EMT of BEAS-2B Cells Related to let-7c-5p and miR-181b-5p).

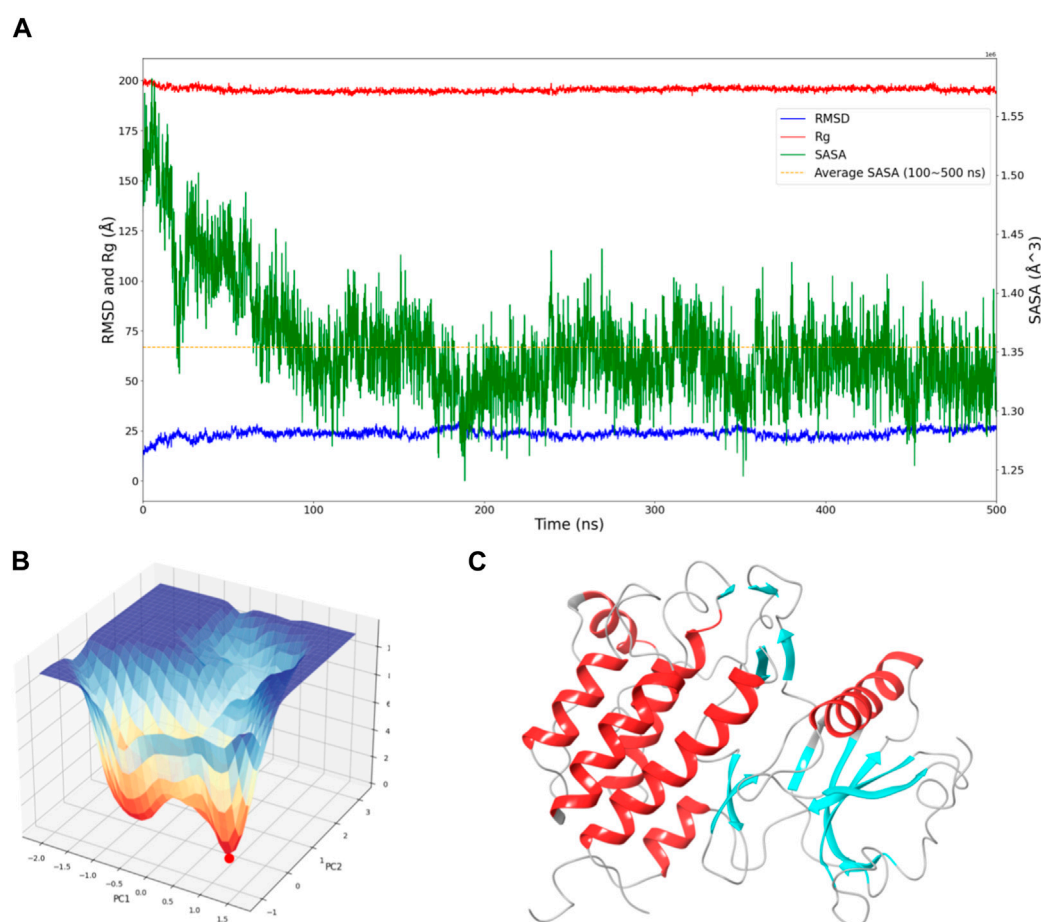


FIGURE 1
Dynamic Behavior and Free Energy Landscape of FGFR3. **(A)** The graph displays the time-dependent dynamics of FGFR3, including RMSD and Rg shown on the left y-axis, and SASA shown on the right y-axis. The dashed line represents the average SASA value after a sharp decrease. **(B)** The 3D free energy landscape of FGFR3 is depicted, with the energy minima indicated by the red dot. A 2D projection of the landscape provides an overview of the conformational space explored by FGFR3. **(C)** Resting state of FGFR3.

2.9 Cell viability detected by CCK8

After co-cultured for 24 h, cell proliferation was detected with CCK8 detection kit. Each well was incubated with 10 μ L CCK8 detection reagent at 37°C for 2 h. The OD value of each well was detected with the microplate reader at 450 nm wavelength to calculate cell viability.

3 Results

3.1 Relaxation of FGFR3 structure

During the virtual screening process, the identification of compounds with the closest and most stable interactions with the target is crucial to selecting potential drug candidates. Molecular dynamics simulations of the target's lowest energy conformation offer valuable insights into compounds with favorable binding affinities, providing crucial guidance for subsequent experimental screenings. To achieve this, a 500 ns molecular dynamics simulation was performed to explore FGFR3's lowest energy conformation after releasing the

Pyrimidine Derivative 37b, ensuring comprehensive sampling and equilibrium attainment for subsequent pocket-based virtual screenings.

To assess the convergence of the simulation, RMSD, Rg, and SASA of FGFR3 were calculated. As shown in Figure 1A, during the 500 ns simulation, both the RMSD and Rg of FGFR3 exhibited minimal fluctuations, indicating an early attainment of stability. Regarding Figure 1A, the high RMSD observed likely results from significant conformational changes in the FGFR3 protein following the removal of the ligand from its binding site. While SASA showed some dynamic changes, it oscillated around the average value after 100 ns, suggesting a continuous periodic thermal motion of FGFR3 rather than a lack of equilibrium. Based on these parameters, the system was considered to reach equilibrium and achieve thorough sampling of FGFR3 after releasing the Pyrimidine Derivative 37b.

Subsequently, Gibbs free energy was statistically analyzed during the simulation, and a free energy landscape was constructed using the first and second eigenvectors, as shown in the Figure 1B. Three energy basins were identified, with the highest energy basin corresponding to the state when FGFR3 was bound to the Pyrimidine Derivative 37b, and the lowest energy points

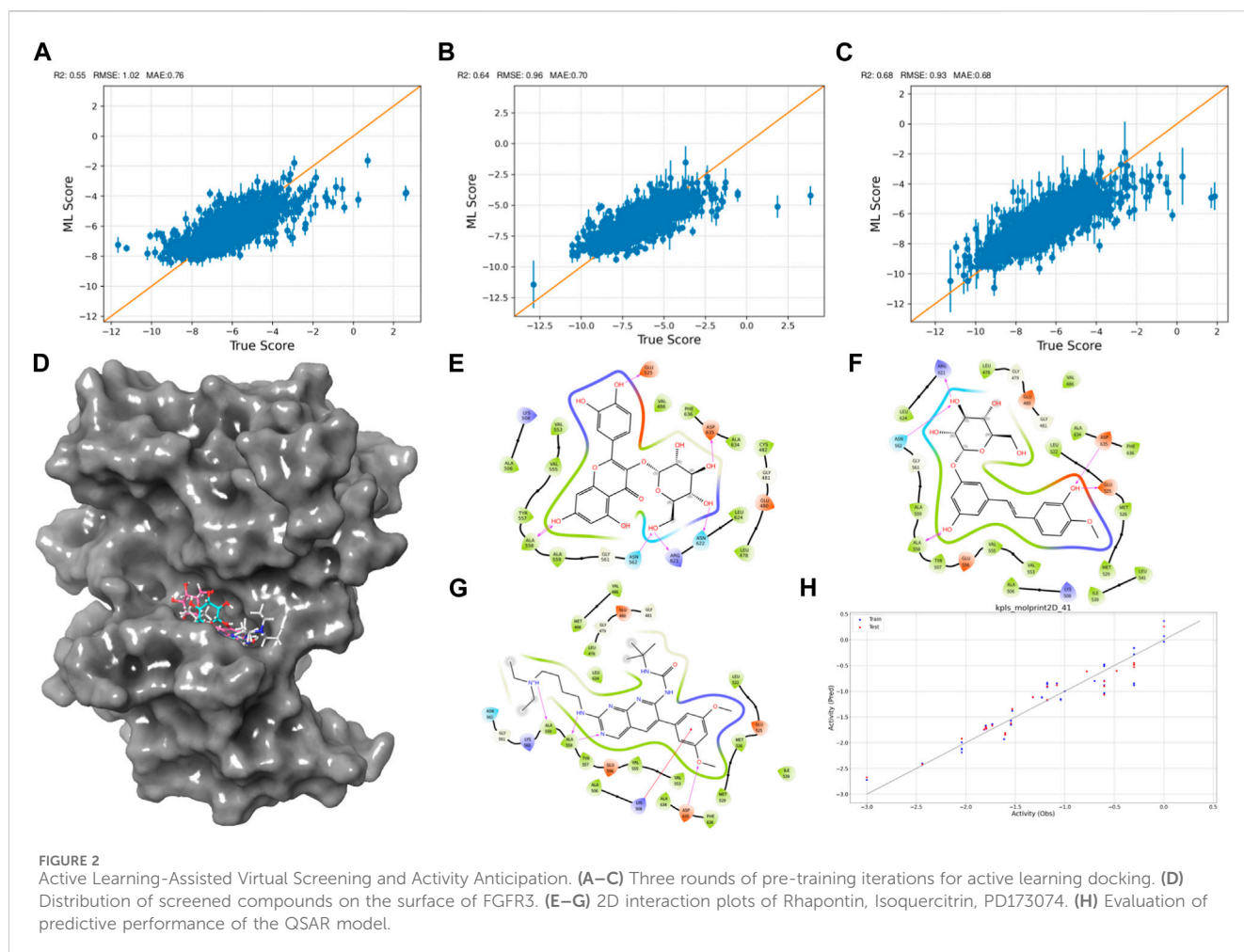


FIGURE 2

Active Learning-Assisted Virtual Screening and Activity Anticipation. (A–C) Three rounds of pre-training iterations for active learning docking. (D) Distribution of screened compounds on the surface of FGFR3. (E–G) 2D interaction plots of Rhapontin, Isoquercitrin, PD173074. (H) Evaluation of predictive performance of the QSAR model.

distributed in the remaining two smaller-volume basins. The transition state connecting these two states was determined. Notably, the lowest energy point emerged at 175 ns and remained stable until 500 ns, smoothly connecting the initial and final states. Based on this, we concluded that the 500 ns simulation successfully sampled FGFR3 after releasing the Pyrimidine Derivative 37b and the lowest energy point in the free energy landscape represented the resting state of FGFR3, as shown in Figure 1C. Building upon this information, subsequent pocket-based virtual screenings will be conducted.

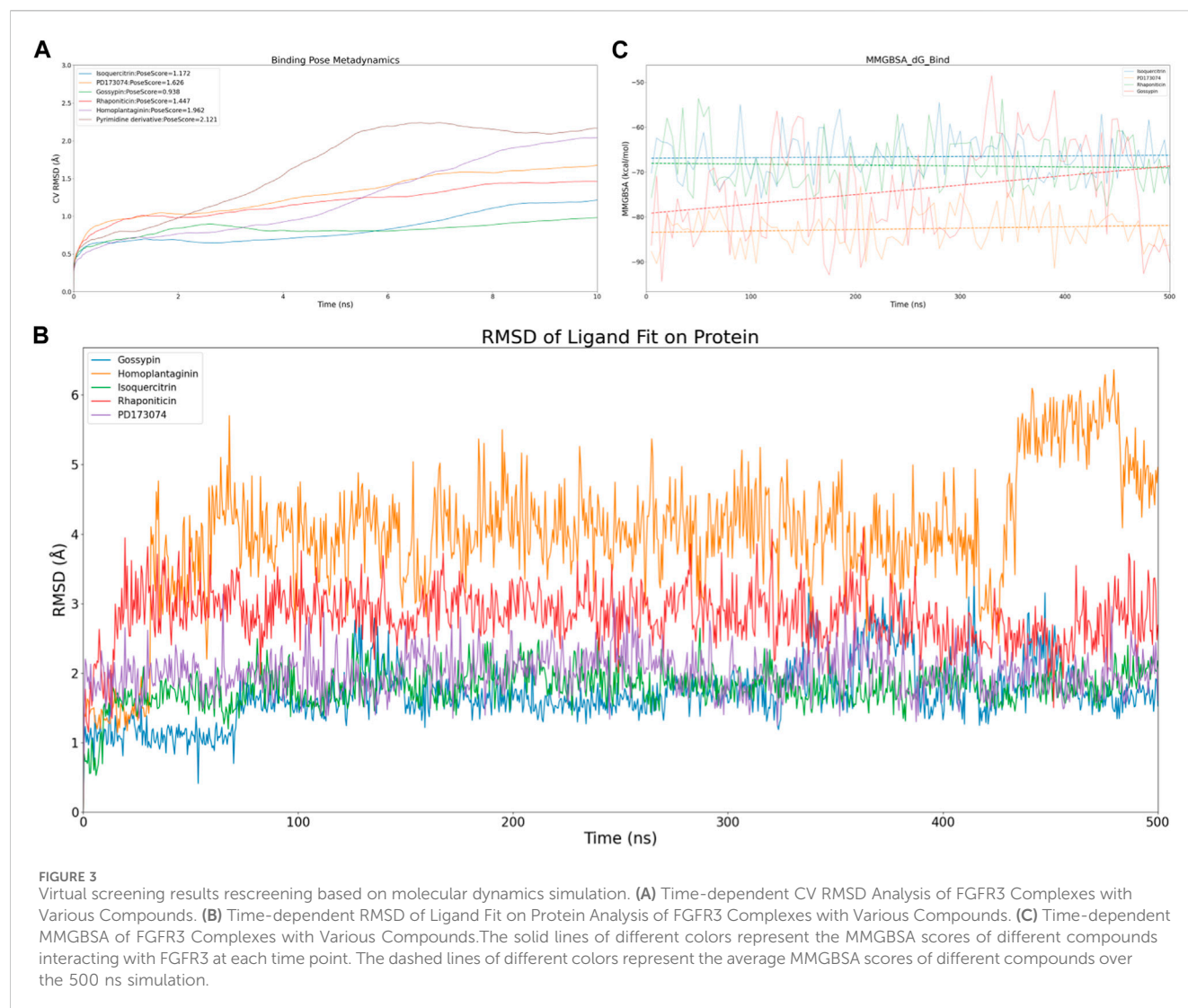
3.2 Virtual compound screening and activity forecasting through active learning

Machine learning and deep learning have revolutionized drug discovery by powering applications such as structure-based virtual screening, efficiently sifting through compound libraries to identify potential hits, and activity prediction models leveraging molecular features to accurately estimate compound bioactivity.

As the iterations progressed, the models consistently exhibited improved performance, as shown in Figures 2A–C. In the initial iteration, the coefficient of determination (R^2) value was 0.55, accompanied by root mean square error (RMSE) and mean

absolute error (MAE) values of 1.02 and 0.7, respectively. Notably, the second iteration displayed an enhanced R^2 of 0.64, alongside reduced RMSE (0.96) and MAE (0.7) values, indicating improved model accuracy. The third iteration showed the most significant advancement, achieving an R^2 value of 0.68. Additionally, the RMSE decreased to 0.93, and the MAE reached 0.68, suggesting an increasingly precise parameter prediction.

Through meticulous analysis, incorporating a comprehensive evaluation of Docking Score, State Penalty, Ligand Strain Energy, and MMGBSA ΔG Bind, three compounds—PD173074 (Lamont et al., 2011), Isoquercitrin (Valentova et al., 2014), and Rhapontin—emerged as promising candidates. A detailed list of scores is presented in Table 1. PD173074 demonstrated exceptional binding affinity with a Docking Score of -10.3 , and exhibited optimal receptor conformation with a State Penalty of 0, alongside a favorable Ligand Strain Energy of 2.1 kcal/mol and a MMGBSA ΔG Bind of -56.5 kcal/mol. Isoquercitrin also presented a strong case, with a Docking Score of -11.1 , State Penalty of 0, Ligand Strain Energy of 9.1 kcal/mol, and a MMGBSA ΔG Bind of -94.4 kcal/mol. Rhapontin, while displaying a slightly higher Docking Score of -10.2 , maintained a State Penalty of 0, a Ligand Strain Energy of 5.0 kcal/mol, and a MMGBSA ΔG Bind of -68.6 kcal/mol, ensuring its position as a candidate of interest. These stringent criteria ensured the selection of compounds not just with strong binding affinities, but



also with optimal receptor conformations and stability, providing a solid foundation for the subsequent stages of our analysis and future experimental validation.

Then, we explored interactions between Isoquercitrin, PD173074, and Rhapontin with FGFR3 residues, as shown in **Figures 2D–G**. Crucial binding residues, such as Lys-508 (ATP-binding site) and Asp-617 (active site), were identified. Isoquercitrin engaged FGFR3 residues Ala-558, Ala-559, Lys-508, and Asp-635, indicating potential modulation of the ATP-binding pocket and its vicinity. PD173074s interactions encompassed Arg-621, Asn-562, Asp-635, Glu-525, and Ala-558, pointing to involvement with the active site and neighboring domains. Rhapontin's interactions spanned Glu-525, Asp-635, Asn-622, Arg-621, Asn-562, and Ala-558, showcasing its adaptable binding capacity across critical regions.

Our Quantitative Structure-Activity Relationship (QSAR) modeling efforts yielded compelling results, as shown in **Figure 2H**. The training set exhibited a Q^2 of 0.2402 and an R^2 of 0.8980, confirming the model's ability to capture intricate activity relationships within the dataset. During external validation, the testing set demonstrated an RMSE of 0.2171 and a Q^2 of 0.9069, attesting to the model's robustness. Furthermore, the model

demonstrated predictive prowess by estimating IC_{50} values for Isoquercitrin, Rhapontin, and PD173074. The calculated values—18.45 nM, 17.46 nM, and 11.67 nM—underscore the model's potential to anticipate compound activities across different chemical entities. The notably close alignment between the predicted IC_{50} for PD Compound and experimental IC_{50} in the RT112 cell line targeting FGFR3 (Lamont et al., 2011) bolsters the model's theoretical reliability.

3.3 Investigating binding mode and stability based on md simulation analysis for potential binding candidates

The results obtained from virtual screening required validation through molecular dynamics simulations to assess their dynamic behavior and interaction stability within the biological system, providing crucial theoretical guidance for further confirmation of potential drug candidates' efficacy and safety in drug development.

To efficiently assess the stability of ligands in solution, we employed binding pose metadynamics (BPMD) as an enhanced sampling technique. By applying bias in the metadynamics

TABLE 1 Binding characteristics of tested compounds.

Name	Docking score	Glide ligand efficiency	MMGBSA dG bind	Lig strain energy
Forsythoside A	−16.884	−0.384	−49.03	47.778
Apigenin 7-O-(2G-rhamnosyl)gentiobioside	−15.204	−0.292	−75.5	18.267
Vitexin -4''-O-glucoside	−15.169	−0.361	−59.14	13.998
Kuromanin chloride	−14.644	−0.458	−56.76	22.869
Xylopentaose	−14.321	−0.311	−54.68	18.536
Neoeriocitrin	−14.184	−0.338	−69.72	21.13
Pectolinarin	−14.158	−0.322	−62.19	28.427
Isoquercitrin	−13.789	−0.418	−60.55	9.111
Gossypin	−13.72	−0.404	−72.69	6.723
Plantainoside D	−13.697	−0.304	−72.94	19.812
Neohesperidin	−13.535	−0.315	−79.03	15.221
Neodiosmin	−13.367	−0.311	−75.18	14.731
YKL-05-099	−13.297	−0.309	−80.06	11.416
Desmopressin	−13.228	−0.179	−64.58	32.245
Rhaponitacin	−13.204	−0.44	−69.41	4.781
Luteolin-7-glucuronide	−13.175	−0.399	−37.65	11.544
Homoplantagin	−13.049	−0.395	−52.66	7.762
PD173074	−12.992	−0.342	−96.31	4.589
Didymine	−12.788	−0.304	−44.01	28.096
Pyrimidine derivative	−14.134	−0.267	−108.12	9.293

simulation, ligand poses that exhibited instability were likely to be rarely occupied in the energy landscape, thereby exerting minimal influence on the overall binding affinity. We performed ten sets of BPMD simulations for the five compounds, with Pyrimidine Derivative 37b as a reference. The results, as shown in [Figure 3A](#), indicated that the CV RMSD values remained below 2.5 Å for all five compounds, whereas only Pyrimidine Derivative 37b's PoseScore exceeded 2 Å, suggesting that the remaining five compounds possessed stronger and more stable interactions with FGFR3 ([Allegra et al., 2021](#)). For a detailed list of scores, refer to [Table 2](#).

Despite conducting ten sets of simulations, the BPMD simulation time remained relatively short. Subsequently, we performed classical molecular dynamics simulations for the five compounds for an extended period of 500 ns, employing Ligand Fit on Protein RMSD and MMGBSA as reference values to evaluate the complex from both conformational and energetic perspectives. Ligand Fit on Protein RMSD represents the RMSD of a ligand when the protein-ligand complex is first aligned on the protein backbone of the reference, and then the RMSD of the ligand heavy atoms is measured. If the observed values are significantly larger than the RMSD of the protein, it suggests that the ligand may have diffused away from its initial binding site.

First, we evaluated the conformational changes, as shown in [Figure 3B](#), which indicated that the five compounds exhibited a similar trend of achieving preliminary stability within the first 100 ns of the simulation. However, after 400 ns,

Homoplantagin showed a noticeable increase in RMSD, implying a potential time-dependence in its binding to FGFR3, and a possibility of off-target effects. Subsequently, we assessed the energetic aspects, focusing on the four remaining compounds since Homoplantagin displayed potential off-target behavior. The MMGBSA results, as shown in [Figure 3C](#), displayed significant fluctuations. To facilitate result analysis, we plotted the trendlines of four groups of MMGBSA values over time. The results revealed a clear upward trend for Gossypin, indicating a continuous decrease in binding energy between Gossypin and the receptor. This suggested that as the conformational adjustments continued, the binding energy between Gossypin and FGFR3 may decrease further, possibly leading to Gossypin dissociation from the binding pocket, implying the possibility of off-target effects. In summary, Isoquercitrin, Rhaponitacin, and PD173074 demonstrated a high potential to act as FGFR3 inhibitors, both from the conformational and energetic perspectives. Consequently, these three compounds were chosen for further interaction analysis.

3.4 Physicochemical parameters, medicinal chemistry parameters, and selectivity analysis

The physicochemical parameters and medicinal chemistry parameters of the compounds provided essential initial

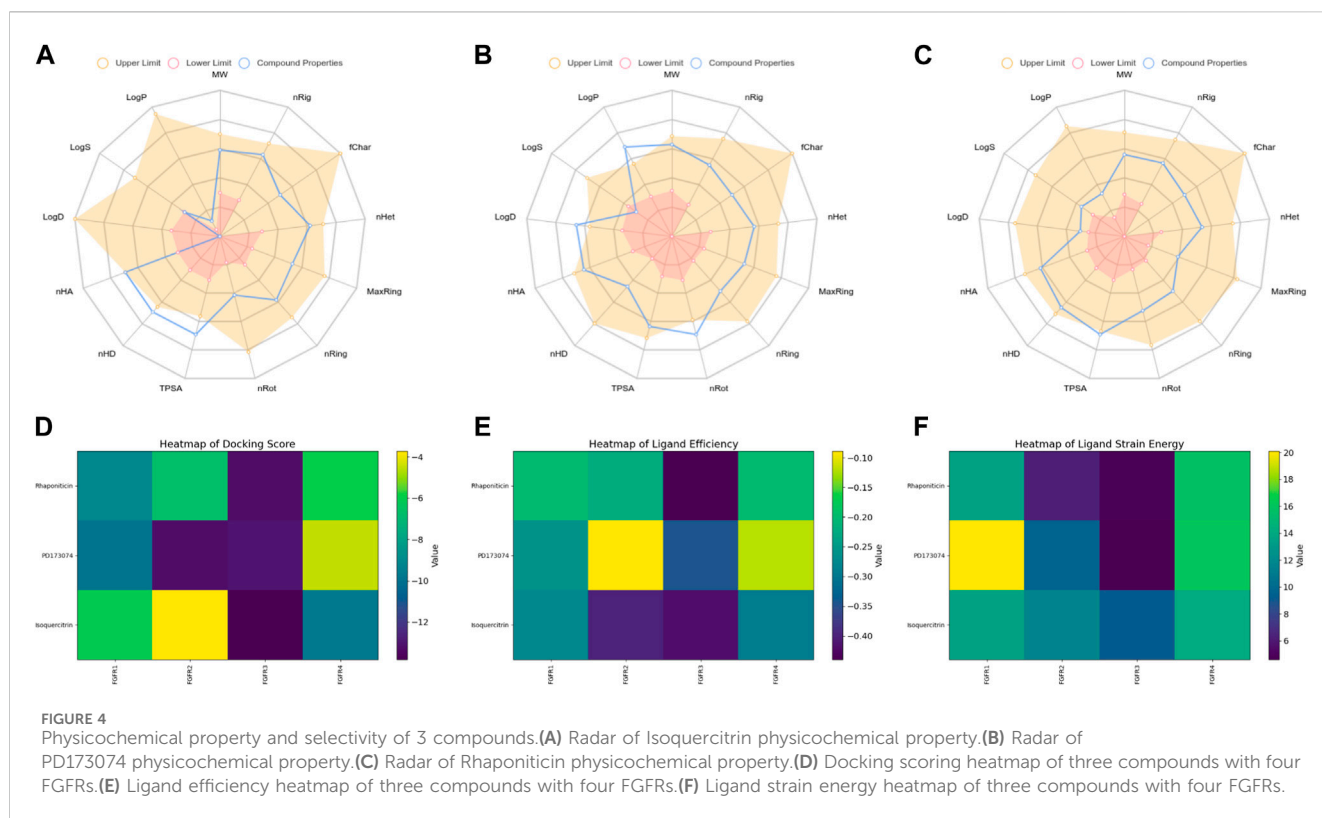


TABLE 2 Dynamic interaction scores for tested compounds.

Compound name	PersScore	PoseScore	CompScore
Isoquercitrin	0.577	1.172	-3.828
PD173074	0.641	1.626	-3.374
Gossypin	0.348	0.938	-4.062
Rhaponticin	0.43	1.447	-3.553
Homoplantagin	0.265	1.962	-3.038
Pyrimidine derivative	0.752	2.121	-1.639

evaluations for drug development, aiding in the screening of potentially drug-like compounds. The selectivity analysis also contributed to identifying potential advantageous targets and guiding subsequent drug optimization and development, thereby increasing the likelihood of successful drug development.

The radar plot in Figures 4A–C illustrates the analysis of physicochemical parameters, such as MW, TPSA, and LogP, for the investigated compounds. Rhaponticin was the only compound that fell within the specified threshold range. The medicinal chemistry studies presented in Table 3 showed that Rhaponticin exhibited a higher QED (Quantitative Estimate of Drug-likeness) (Kosugi and Ohue, 2021) value, indicating its potential as a drug-like molecule, adhering to general drug development guidelines. Additionally, its low PAINS Alter value suggested a lower risk of being a promiscuous compound, making it more suitable for drug development. Moreover, the higher SA Score of Rhaponticin indicated relatively facile synthesis, which facilitated further research.

Furthermore, the higher proportion of sp³-hybridized carbon atoms (Wei et al., 2020) in Rhaponticin suggested its potential for enhanced drug activity. Notably, the receptor selectivity analyses, including docking scoring, ligand efficiency, and ligand strain energy, demonstrated that Rhaponticin exhibited exceptional selectivity against FGFR3, as depicted in Figures 4E,F.

In contrast, PD173074, while displaying some drug-like characteristics, exhibited a relatively lower QED value (Lipinski, 2004). Although it met the criteria of the Pfizer Rule, its performance might not be as effective as Rhaponticin in certain aspects. The smaller Molar Refractivity (MCE-18) (Ivanenkov et al., 2019) value of PD173074 suggested a smaller molecular volume, potentially affecting interactions within the biological system. Despite having a PAINS Alter value of 0, indicating a lower probability of being a promiscuous compound, further investigation was still warranted.

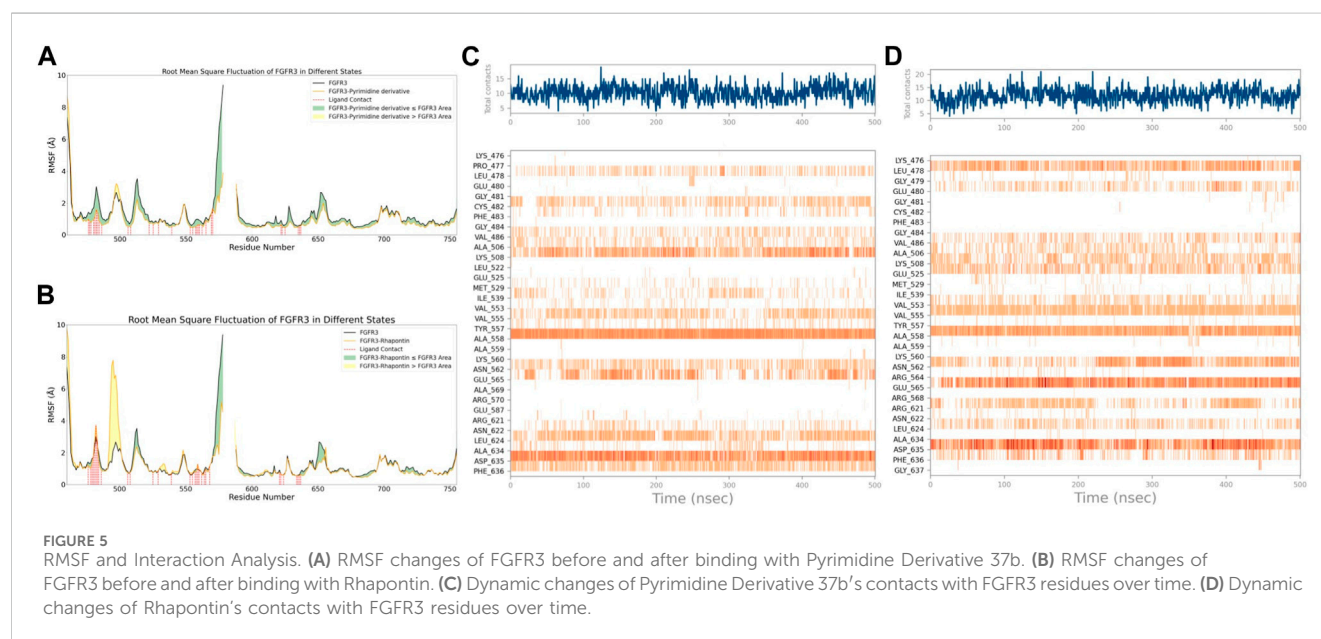
Regarding Isoquercetin, its lower QED value indicated the necessity for further optimization. While satisfying the Pfizer Rule, its PAINS Alter value of 1 implied potential promiscuity, demanding additional evaluation. Isoquercetin's higher SA Score indicated relatively facile synthesis, but its larger MCE-18 value suggested it might occupy a larger volume during interactions.

3.5 Exploring ligand binding effects on FGFR3 flexibility and interactions based on RMSF and interaction analysis

Through molecular dynamics simulations, studying the interactions between receptors and ligands provides in-depth insights into the binding modes and dynamic processes of drugs with their target receptors. This valuable information supports

TABLE 3 Medicinal chemistry of 3 compounds.

Compound name	Rhapontin	PD173074	Isoquercetin
QED	0.366	0.347	0.229
SA Score	3.763	3.634	4.008
FSP ³	0.333	0.5	0.286
MCE-18	67.143	22	91
PAINS Alter	0	0	1
Lipinski Rule	Accepted	Accepted	Rejected
Pfizer Rule	Accepted	Accepted	Accepted



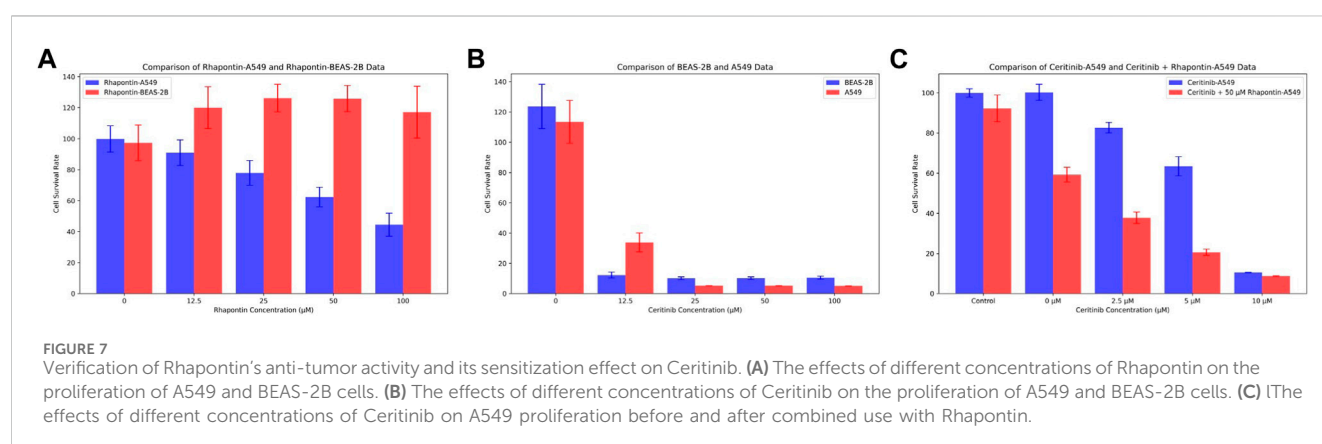
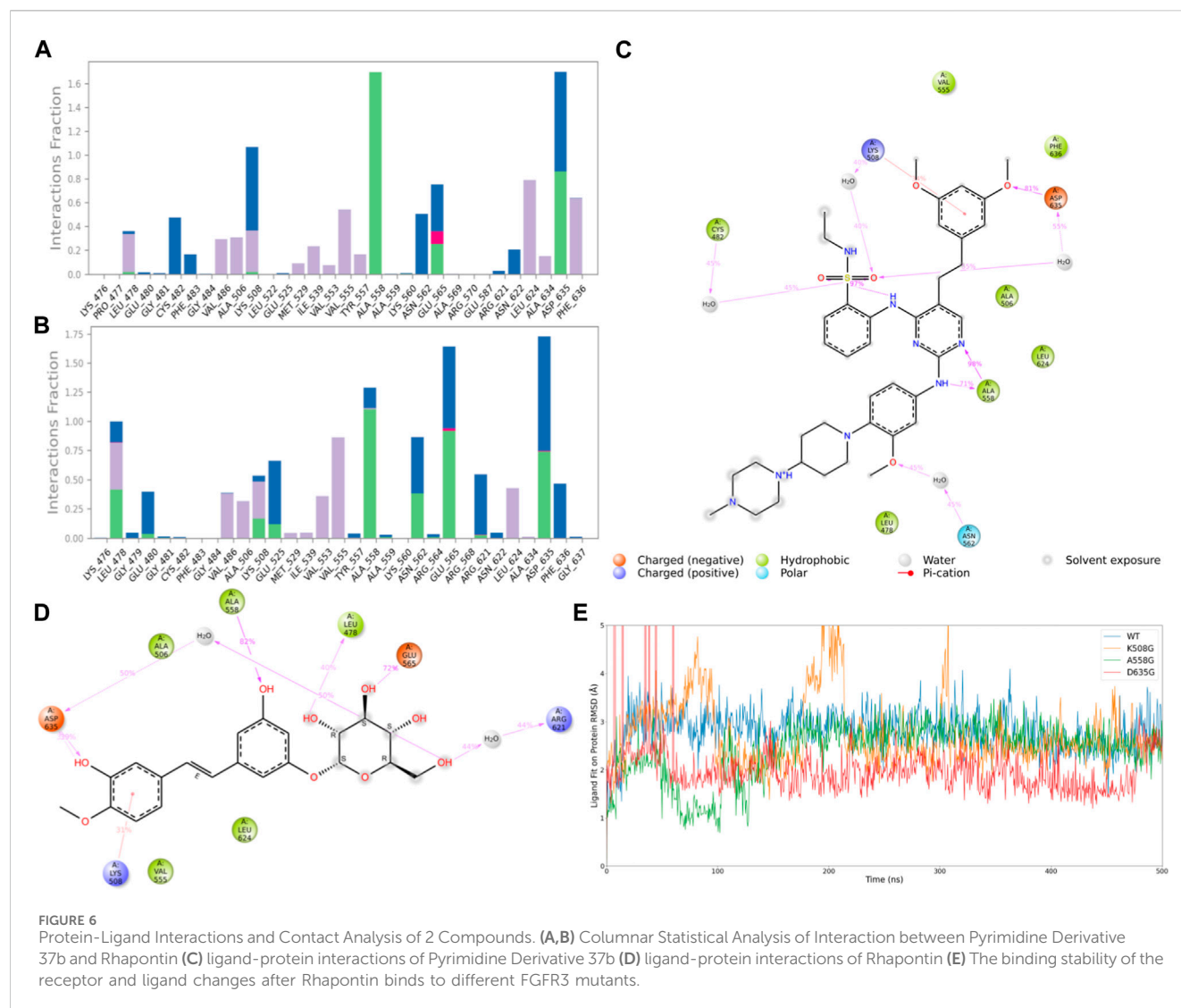
subsequent drug development, aiding in the optimization and improvement of drug molecules to enhance their affinity and selectivity towards target receptors, thus improving drug efficacy and safety, and providing scientific foundations for drug development.

With this purpose in mind, we first used the RMSF of FGFR3 in its apo state as a baseline to observe the similarities and differences in the effects of Pyrimidine Derivative 37b and Rhapontin on FGFR3, as shown in **Figures 5A,B**. Comparatively, the main differences between the two ligands were observed in two peptide segments. Firstly, in the region of 491–500, both Pyrimidine Derivative 37b and Rhapontin increased the flexibility to varying degrees, with Rhapontin causing a significantly greater effect. Secondly, in the region of 600–640, Pyrimidine Derivative 37b did not exhibit any significant influence, while Rhapontin slightly increased the flexibility in this area. Apart from these differences, both ligands showed minimal distinctions in their overall impact on FGFR3 residues and their contact frequency with FGFR3 residues.

Regarding the contact situation with residues, as shown in **Figures 5C,D**, the overall pattern was quite similar, but the average contact frequency in the simulation was higher for Rhapontin than for Pyrimidine Derivative 37b. For a detailed

comparison of the binding conformations of Rhapontin before and after simulation, please refer to **Supplementary Figure S1**. However, this did not appear to be due to unreasonable conformations of the compounds inside the binding pocket but rather an increased contact frequency with certain residues, such as Glu-565 and Asp-635, which showed higher interaction frequencies than Rhapontin. The upregulation of RMSF in the region of 491–500 might not be directly related to changes in contact frequency with Rhapontin, as neither Pyrimidine Derivative 37b nor Rhapontin directly contacts these residues. A plausible explanation could be that Rhapontin does not contact residues 482 and 483, indirectly relieving the restrictions on this peptide segment.

Subsequently, after classifying and statistically analyzing the interactions between compounds and individual residues, we selected residues with interaction frequencies exceeding 30% and depicted the interaction details between these residues and the compounds, as shown in **Figures 6A–D**. The interaction statistics showed consistency with the differences mentioned earlier, where the number of interacting residues with Rhapontin was fewer than with Pyrimidine Derivative



37b, but the interaction frequencies were slightly higher. Furthermore, in the interaction detail plots, it was observed that three key residues in the FGFR3 pharmacophore relationship, Lys-508, Ala-558, and Asp-635, were reproduced in the interaction details with Rhapontin. To further validate the

importance of these three sites in Rhapontin binding, we conducted dynamic simulations with the three sites mutated to Gly and assessed their impact on the binding between the two, as shown in Figure 6E. It was evident that all three mutations significantly affected the binding between Rhapontin and FGFR3,

demonstrating the importance of these residues in Rhapontin binding.

3.6 Biological evaluation of rhapontin through CCK-8 assay

In consideration of the limitations of our previous theoretical analyses, we conducted further experimental validations on Rhapontin. Initially, we subjected both BEAS-2B and A549 cell lines to varying concentrations of Rhapontin and Ceritinib (0–100 μM), as depicted in [Figures 7A,B](#). As concentrations escalated, Rhapontin demonstrated a concentration-dependent proliferation inhibitory effect on A549 cells, with an IC_{50} of 62 μM . Despite its IC_{50} being significantly higher than that of Ceritinib, Rhapontin exhibited minimal impact on the proliferation of BEAS-2B cells. Conversely, Ceritinib exhibited a notable proliferation inhibitory effect on normal cells, including instances of substantial cytotoxicity.

Considering the potential of FGFR3 inhibitors to sensitize Ceritinib, we subsequently employed a reduced concentration (50 μM) of Rhapontin in combination with varying doses of Ceritinib (0–10 μM , adjusted from previous concentrations). The outcomes, as illustrated in [Figure 7C](#), indeed displayed an enhanced tumor inhibitory effect to a certain extent when distinct concentrations of Ceritinib were co-administered with 50 μM Rhapontin. This co-administration led to a heightened sensitization of A549 cells to Ceritinib.

4 Discussion

The present study presents a systematic exploration aimed at augmenting the efficacy of Ceritinib, a prominent FGFR3 inhibitor, via the integration of natural compound-derived alternatives. Our investigation embraces a multidimensional approach, employing active learning derived virtual screen ([Ma et al., 2009](#)), deep learning derived QSAR modeling ([Matsuzaka and Uesawa, 2023](#)), molecular dynamics simulations ([Duay et al., 2023](#)), and biological assays to dissect the mechanisms underlying the potential synergy between Ceritinib and the identified natural compounds.

The selection of natural compounds as potential drug candidates draws attention to their inherent structural diversity and recognized pharmacological safety ([Zhang et al., 2023](#)). Natural products have, over the years, emerged as a wellspring of bioactive molecules, often possessing unique chemical scaffolds and physiological properties ([Safranko et al., 2023](#)). Notably, the prospect of leveraging certain natural compounds as nutraceutical agents underscores their compatibility with biological systems and augments the overall therapeutic potential ([Wang and Wang, 2021](#)).

Rhapontin, one of the highlighted natural compounds, presents intriguing prospects despite its moderate inhibitory activity in comparison to Pyrimidine Derivative 37b. This finding resonates with the broader paradigm of molecular design, urging for meticulous structural optimization to fine-tune both binding interactions and inhibitory potency ([Azimian and Dastmalchi, 2023](#)). The journey toward harnessing Rhapontin's full potential entails a systematic exploration of its structural landscape, with a

focus on judicious modifications to enhance its binding interactions.

The combined application of Rhapontin and Ceritinib, while not achieving the zenith of efficacy exhibited by certain established combination therapies, merits profound scrutiny ([Krol et al., 2023](#)). The nuanced response could stem from intricate intracellular interactions, wherein Rhapontin's engagement with alternative molecular targets competes with its interaction with FGFR3 ([Ho et al., 2014](#); [Cascetta et al., 2022](#)). This observation augments the need for a rigorous dissection of these competitive binding events, necessitating an iterative process of targeted compound engineering ([Ho et al., 2014](#)).

5 Conclusion

This study underscores the potential of natural compound-derived FGFR3 inhibitors to anti-cancer and sensitize Ceritinib. The utilization of natural compounds not only diversifies the drug discovery landscape but also accentuates their potential as bioactive agents with intrinsic safety profiles. Rhapontin's modest inhibitory activity, coupled with its structural attributes, calls for a deeper exploration to unlock its latent potential. The observed synergy between Ceritinib and Rhapontin, albeit nuanced, underscores the intricate cellular dynamics that govern combination therapies. As we continue to unravel the complexities of molecular interactions, strategic compound engineering offers a promising avenue to enhance therapeutic outcomes and guide the evolution of precision medicine paradigms.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

QZ: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing—original draft. HH: Data curation, Formal Analysis, Validation, Visualization, Writing—original draft. ZH: Investigation, Methodology, Writing—original draft. AG: Validation, Writing—original draft. SL: Visualization, Writing—original draft. WT: Investigation, Writing—original draft. ZZ: Writing—review and editing. TS: Writing—review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The work was supported by the Hengrui Pharmaceutical Oncology Research Project: Y202044260 to QZ; the National Natural Science Foundation of China (Grant Number: 82103305) to TS.

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2024.1413214/full#supplementary-material>

References

- Allegra, M., Tutone, M., Tesoriere, L., Attanzio, A., Culetta, G., and Almerico, A. M. (2021). Evaluation of the IKK β binding of indicaxanthin by induced-fit docking, binding pose metadynamics, and molecular dynamics. *Front. Pharmacol.* 12, 701568. doi:10.3389/fphar.2021.701568
- Arnittali, M., Rissanou, A. N., and Harmandaris, V. (2019). "Structure of biomolecules through molecular dynamics simulations," in 8th International Young Scientists Conference on Computational Science (YSC), 2019, Jun 24–28 2019 (Heraklion, Greece: Univ Amsterdam), 69–78.
- Azimian, F., and Dastmalchi, S. (2023). Recent advances in structural modification strategies for lead optimization of tyrosine kinase inhibitors to explore novel anticancer agents. *Curr. Med. Chem.* 30, 2734–2761. doi:10.2174/0929867329666220920092908
- Brown, K. K., and Tokar, A. (2015). The phosphoinositide 3-kinase pathway and therapy resistance in cancer. *F1000prime Rep.* 7, 13. doi:10.12703/P7-13
- Campitelli, P., Swint-Kruse, L., and Ozkan, S. B. (2021). Substitutions at nonconserved rheostat positions modulate function by rewiring long-range, dynamic interactions. *Mol. Biol. Evol.* 38, 201–214. doi:10.1093/molbev/msaa202
- Cascetta, P., Marinello, A., Lazzari, C., Gregorc, V., Planchard, D., Bianco, R., et al. (2022). KRAS in NSCLC: state of the art and future perspectives. *Cancers* 14, 5430. doi:10.3390/cancers14215430
- Colmegna, B., Morosi, L., and D'Incalci, M. (2018). "Molecular and pharmacological mechanisms of drug resistance: an evolving paradigm," in *Mechanisms of drug resistance in cancer therapy*. Editors M. Mandala and E. Romano 1–12.
- Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald - an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397
- De Groot, P. M., Wu, C. C., Carter, B. W., and Munden, R. F. (2018). The epidemiology of lung cancer. *Transl. Lung Cancer Res.* 7, 220–233. doi:10.21037/tlcr.2018.05.06
- De Oliveira, M. T., and Katekawa, E. (2018). On the virtues of automated quantitative structure-activity relationship: the new kid on the block. *Future Med. Chem.* 10, 335–342. doi:10.4155/fmc-2017-0170
- Dixon, S. L., Duan, J., Smith, E., Von Bargen, C. D., Sherman, W., and Repasky, M. P. (2016). AutoQSAR: an automated machine learning tool for best-practice quantitative structure-activity relationship modeling. *Future Med. Chem.* 8, 1825–1839. doi:10.4155/fmc-2016-0093
- Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophysics* 41, 429–452. doi:10.1146/annurev-biophys-042910-155245
- Duay, S. S., Yap, R. C. Y., Gaitano, A. L., Santos, J. A. A., and Macalino, S. J. Y. (2023). Roles of virtual screening and molecular dynamics simulations in discovering and understanding antimalarial drugs. *Int. J. Mol. Sci.* 24, 9289. doi:10.3390/ijms24119289
- Ekins, S., Mestres, J., and Testa, B. (2007). *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* 152, 9–20. doi:10.1038/sj.bjp.0707305
- 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
- Fusani, L., Palmer, D. S., Somers, D. O., and Wall, I. D. (2020). Exploring ligand stability in protein crystal structures using binding pose metadynamics. *J. Chem. Inf. Model.* 60, 1528–1539. doi:10.1021/acs.jcim.9b00843
- Green, D. V. S. (2008). Virtual screening of chemical libraries for drug discovery. *Expert Opin. Drug Discov.* 3, 1011–1026. doi:10.1517/17460441.3.9.1011
- Hess, B. (2008). P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* 4, 116–122. doi:10.1021/ct700200b
- Ho, H. K., Yeo, A. H. L., Kang, T. S., and Chua, B. T. (2014). Current strategies for inhibiting FGFR activities in clinical applications: opportunities, challenges and toxicological considerations. *Drug Discov. Today* 19, 51–62. doi:10.1016/j.drudis.2013.07.021
- Ivanenkov, Y. A., Zagribelnyy, B. A., and Aladinskiy, V. A. (2019). Are we opening the door to a new era of medicinal chemistry or being collapsed to a chemical singularity? *J. Med. Chem.* 62, 10026–10043. doi:10.1021/acs.jmedchem.9b00004
- Jin, H., Wu, C., Su, R., Sun, T., Li, X., and Guo, C. (2023). Identifying dopamine D3 receptor ligands through virtual screening and exploring the binding modes of hit compounds. *Molecules* 28, 527. doi:10.3390/molecules28020527
- Kosugi, T., and Ohue, M. (2021). Quantitative estimate index for early-stage screening of compounds targeting protein-protein interactions. *Int. J. Mol. Sci.* 22, 10925. doi:10.3390/ijms222010925
- Krol, K., Mazur, A., Stachyra-Strawa, P., and Grzybowska-Szatowska, L. (2023). Non-small cell lung cancer treatment with molecularly targeted therapy and concurrent radiotherapy-A review. *Int. J. Mol. Sci.* 24, 5858. doi:10.3390/ijms24065858
- Kuriwaki, I., Kameda, M., Hisamichi, H., Kikuchi, S., Iikubo, S., Kawamoto, Y., et al. (2020). Structure-based drug design of 1,3,5-triazine and pyrimidine derivatives as novel FGFR3 inhibitors with high selectivity over VEGFR2. *Bioorg. Med. Chem.* 28, 115453. doi:10.1016/j.bmc.2020.115453
- Lamont, F. R., Tomlinson, D. C., Cooper, P. A., Shnyder, S. D., Chester, J. D., and Knowles, M. A. (2011). Small molecule FGF receptor inhibitors block FGFR-dependent urothelial carcinoma growth *in vitro* and *in vivo*. *Br. J. Cancer* 104, 75–82. doi:10.1038/sj.bjc.6606016
- Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Funct. Bioinforma.* 78, 1950–1958. doi:10.1002/prot.22711
- Lipinski, C. A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. today. Technol.* 1, 337–341. doi:10.1016/j.ddtec.2004.11.007
- Ma, X. H., Jia, J., Zhu, F., Xue, Y., Li, Z. R., and Chen, Y. Z. (2009). Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Comb. Chem. High Throughput Screen.* 12, 344–357. doi:10.2174/138620709788167944
- Malmstrom, R. D., Kornev, A. P., Taylor, S. S., and Amaro, R. E. (2015). Allosteric through the computational microscope: cAMP activation of a canonical signalling domain. *Nat. Commun.* 6, 7588. doi:10.1038/ncomms8588
- Matsuzaka, Y., and Uesawa, Y. (2023). Ensemble learning, deep learning-based and molecular descriptor-based quantitative structure-activity relationships. *Molecules* 28, 2410. doi:10.3390/molecules28052410
- Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* 8, 959–968. doi:10.1038/nrd2961
- Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* 9, 203–214. doi:10.1038/nrd3078
- Safranko, S., Subaric, D., Jerkovic, I., and Jokic, S. (2023). Citrus by-products as a valuable source of biologically active compounds with promising pharmaceutical, biological and biomedical potential. *Pharm. Basel, Switz.* 16, 1081. doi:10.3390/ph16081081
- Siegel, R. L., Giaquinto, A. N., and Jemal, A. (2024). Cancer statistics, 2024. *CA a cancer J. Clin.* 74, 12–49. doi:10.3322/caac.21820
- 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-a Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

- Tang, Z., Du, R., Jiang, S., Wu, C., Barkauskas, D. S., Richey, J., et al. (2008). Dual MET-EGFR combinatorial inhibition against T790M-EGFR-mediated erlotinib-resistant lung cancer. *Br. J. Cancer* 99, 911–922. doi:10.1038/sj.bjc.6604559
- Turner, N., and Grose, R. (2010). Fibroblast growth factor signalling: from development to cancer. *Nat. Rev. Cancer* 10, 116–129. doi:10.1038/nrc2780
- Valentova, K., Vrba, J., Bancirova, M., Ulrichova, J., and Kren, V. (2014). Isoquercitrin: pharmacology, toxicology, and metabolism. *Food Chem. Toxicol.* 68, 267–282. doi:10.1016/j.fct.2014.03.018
- Wang, L., and Wang, W. (2021). Safety and efficacy of anaplastic lymphoma kinase tyrosine kinase inhibitors in non-small cell lung cancer (Review). *Oncol. Rep.* 45, 13–28. doi:10.3892/or.2020.7851
- Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* 137, 2695–2703. doi:10.1021/ja512751q
- Wei, W., Cherukupalli, S., Jing, L., Liu, X., and Zhan, P. (2020). Fsp(3): a new parameter for drug-likeness. *Drug Discov. Today* 25, 1839–1845. doi:10.1016/j.drudis.2020.07.017
- Wesche, J., Haglund, K., and Haugsten, E. M. (2011). Fibroblast growth factors and their receptors in cancer. *Biochem. J.* 437, 199–213. doi:10.1042/BJ20101603
- Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., et al. (2021). ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.* 49, W5–W14. doi:10.1093/nar/gkab255
- Zhang, J., Tao, S., Hou, G., Zhao, F., Meng, Q., and Tan, S. (2023). Phytochemistry, nutritional composition, health benefits and future prospects of Passiflora: a review. *Food Chem.* 428, 136825. doi:10.1016/j.foodchem.2023.136825
- Zhang, J., Zhang, L., Su, X., Li, M., Xie, L., Malchers, F., et al. (2013). Translating the therapeutic potential of AZD4547 in FGFR1-amplified non-small cell lung cancer through the use of patient-derived tumor xenograft models (vol 18, pg 6658, 2012). *Clin. Cancer Res.* 19, 3714. doi:10.1158/1078-0432.CCR-12-2694



OPEN ACCESS

EDITED BY

Annalisa Pastore,
King's College London, United Kingdom

REVIEWED BY

Chang Liu,
Biogen Idec, United States
Jie E. Yang,
University of Wisconsin-Madison,
United States

*CORRESPONDENCE

Jun Yang,
✉ yangjun2009@hmc.edu.cn
Linjie Chen,
✉ chenlinjie@hmc.edu.cn

[†]These authors have contributed equally
to this work

RECEIVED 09 April 2024

ACCEPTED 15 July 2024

PUBLISHED 30 July 2024

CITATION

Zhang H, Lan J, Wang H, Lu R, Zhang N, He X,
Yang J and Chen L (2024), AlphaFold2 in
biomedical research: facilitating the
development of diagnostic strategies for
disease.
Front. Mol. Biosci. 11:1414916.
doi: 10.3389/fmolb.2024.1414916

COPYRIGHT

© 2024 Zhang, Lan, Wang, Lu, Zhang, He,
Yang and Chen. 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.

AlphaFold2 in biomedical research: facilitating the development of diagnostic strategies for disease

Hong Zhang^{1†}, Jiajing Lan^{1†}, Huijie Wang¹, Ruijie Lu¹,
Nanqi Zhang¹, Xiaobai He^{1,2}, Jun Yang^{1*} and Linjie Chen^{1,3*}

¹School of Laboratory Medicine, Hangzhou Medical College, Hangzhou, China, ²Key Laboratory of Biomarkers and In Vitro Diagnosis Translation of Zhejiang Province, Hangzhou, China, ³Zhejiang Engineering Research Centre for Key Technology of Diagnostic Testing, Hangzhou, China

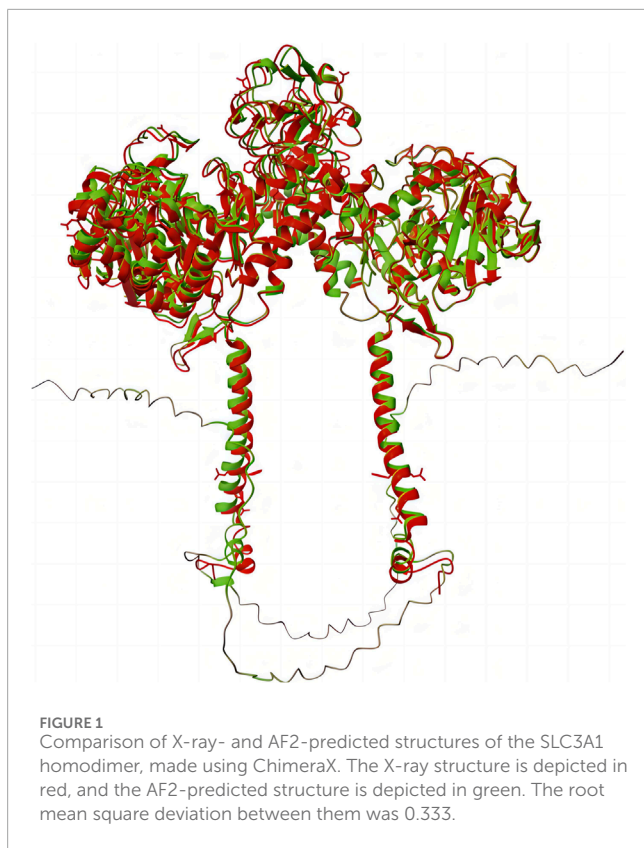
Proteins, as the primary executors of physiological activity, serve as a key factor in disease diagnosis and treatment. Research into their structures, functions, and interactions is essential to better understand disease mechanisms and potential therapies. DeepMind's AlphaFold2, a deep-learning protein structure prediction model, has proven to be remarkably accurate, and it is widely employed in various aspects of diagnostic research, such as the study of disease biomarkers, microorganism pathogenicity, antigen-antibody structures, and missense mutations. Thus, AlphaFold2 serves as an exceptional tool to bridge fundamental protein research with breakthroughs in disease diagnosis, developments in diagnostic strategies, and the design of novel therapeutic approaches and enhancements in precision medicine. This review outlines the architecture, highlights, and limitations of AlphaFold2, placing particular emphasis on its applications within diagnostic research grounded in disciplines such as immunology, biochemistry, molecular biology, and microbiology.

KEYWORDS

AlphaFold2, deep learning, protein structure prediction, structural biology, disease diagnosis

1 Introduction

AlphaFold2 (AF2), developed by DeepMind, is a modeling method that harnesses the cutting-edge technologies of artificial intelligence and deep learning for predicting protein structures with extremely high prediction accuracy ([Figure 1](#)). Rooted in the principle of co-evolution within protein structures, AF2 integrates novel deep learning approaches through the deployment of a suite of trained deep neural network models based on MSA-Transformer, a classical neural network model. These models can generate three-dimensional protein structures with atomic-level precision, informed by both specific amino acid sequence data and information from homologous proteins and multiple sequence alignments (MSAs) ([Jumper et al., 2021](#); [Yang et al., 2023](#)). Its outstanding performance at the international CASP14 protein structure prediction competition showcased a significant breakthrough in both speed and accuracy, leading to its decisive triumph ([Kryshtafovych et al., 2021](#)). The success of AF2 relies on the accumulation of experimental data on protein structures and the comprehensive research conducted on protein structure prediction. Additionally, the active development



community surrounding AF2 ensures a constant influx of fresh talent into the AF2 series, including updates and derivative versions.

Proteins play a vital role in physiological processes, and alterations in the structure and function of specific proteins can lead to distinct diseases. Detecting changes in these specific proteins serves as a crucial diagnostic indicator. Proteins are also essential players in the biological functions of pathogenic microorganisms, simultaneously driving disease and influencing treatment strategies. Furthermore, proteins with strong antigenicity not only act as antigens but also serve as potential targets and essential tools in disease diagnosis. Clinical serum antibody detection is one of the many diverse applications of these proteins. The application of specific proteins in disease diagnosis relies on comprehensive research into their unique functions and disease-related changes, involving multiple fields of biology, such as immunology, biochemistry, molecular biology, and microbiology. Since the release of AF2, it has been widely used in various protein research areas. For these studies, numerous excellent reviews have thoroughly explained AF2's multifaceted functions in biological and medical research, demonstrating its superior performance in predicting protein structures, analyzing mutations, and predicting catalytic and binding sites (Bongirwar and Mokhadde, 2022; Paiva et al., 2022; Bertoline et al., 2023). In the meantime, a lot of studies have demonstrated AF2's robust and exceptional capabilities in investigating disease-related protein structures, functions, interactions, and proteomics. Consequently, research findings utilizing AF2 not only facilitate the development of

diagnostic tools and therapeutic drugs, including antibodies and antigens, but also advance our understanding of protein structures, functions, and mutations related to diseases. This helps better understand the impact of specific proteins on the onset and progression of disease, leading to the development of novel disease indicators, targets, detection tools, and treatments (Figure 2). However, there is a lack of comprehensive reviews on AF2's research in the field of disease diagnosis.

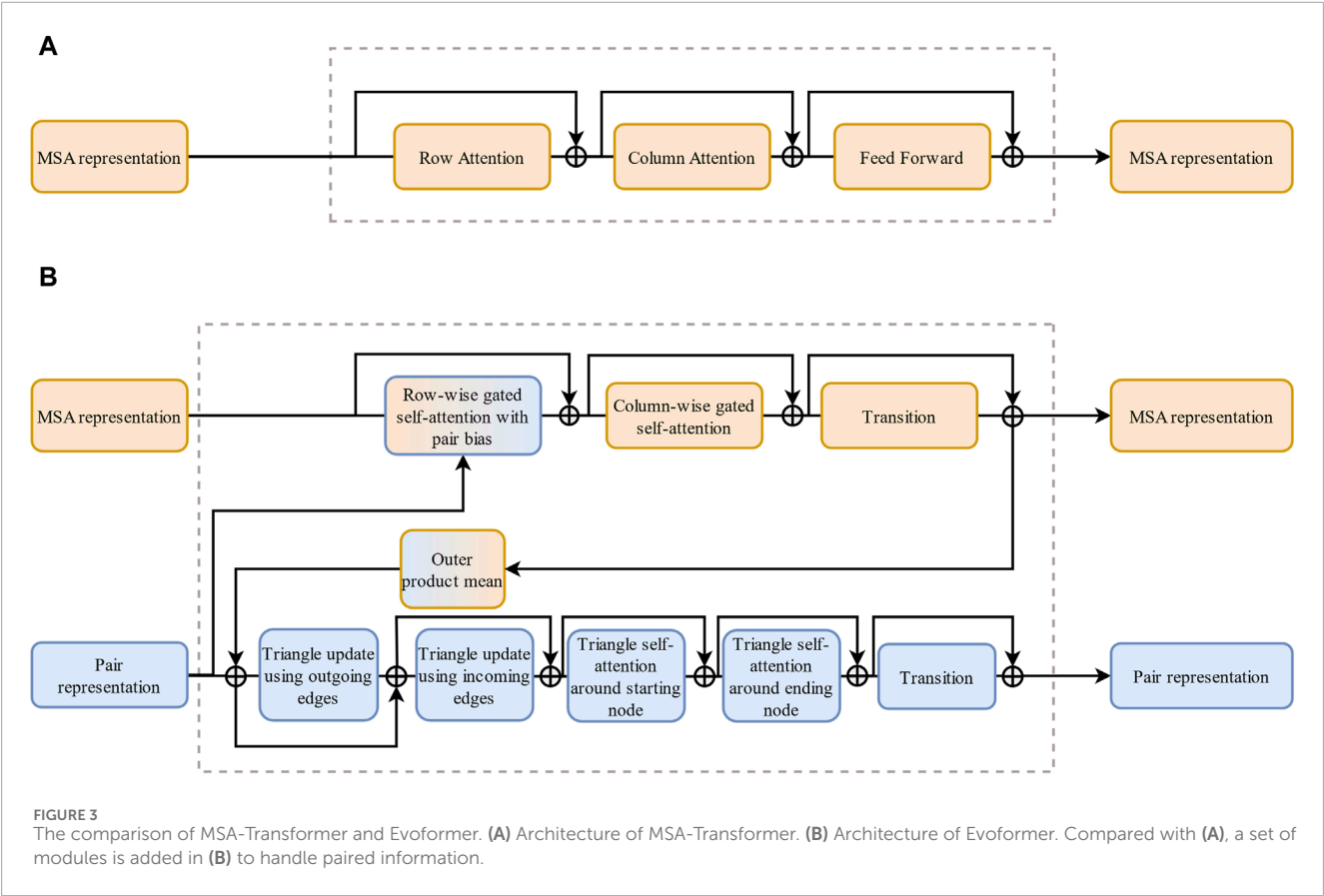
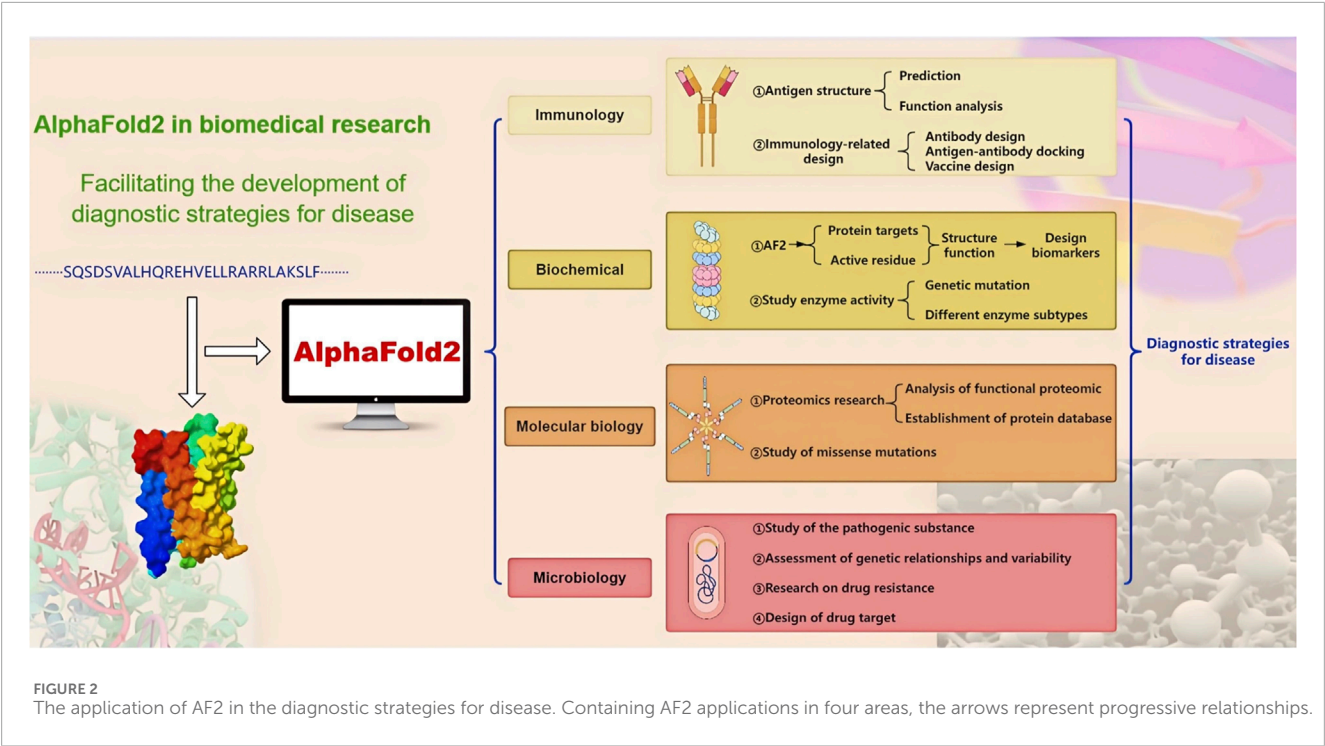
This review aims to comprehensively examine the model architecture, key features, and limitations of AF2. It performs a deep investigation into the extensive applications of AF2 in protein-related research across several disciplines. Finally, this paper briefly touches upon the prospective future development of AF2 and discusses the promotion of basic biological research using AF2 in disease diagnosis.

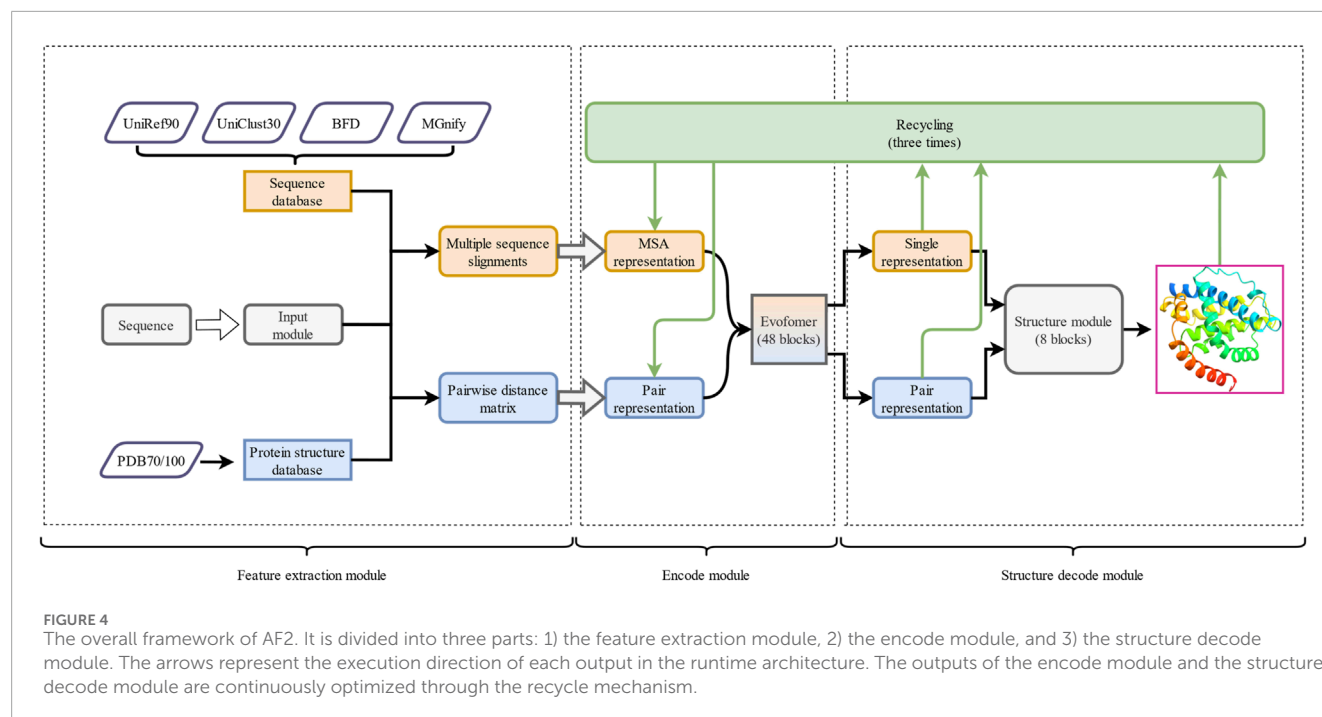
2 AF2

2.1 The model structure of AF2

AF2 is DeepMind's foremost protein structure prediction method, distinguished by its utilization of the innovative neural network architecture known as Evoformer. Inspired by the MSA-Transformer (Figure 3) (Vaswani et al., 2017; Rao et al., 2021), Evoformer combines evolutionary mechanisms, physical principles, and geometric constraints inherent in protein structures to yield exceptional protein structure predictions. Evoformer, comprising two sets of MSA-Transformer-based structures, captures information from MSAs and features related to structural constraints between amino acid residues. This dual-focus approach significantly enhances the prediction quality.

At the core of AF2 lies the application of structural information embedded in protein co-evolution (Pazos and Valencia, 2008; Ashenberg and Laub, 2013). MSA is a bioinformatics technique used to align three or more biological sequences, such as proteins, DNA, or RNA. The objective of MSA is to identify regions of similarity that suggest functional, structural, or evolutionary relationships among the sequences. This method arranges the sequences so that homologous residues, which are derived from a common ancestor, are aligned in columns (Prjibelski et al., 2019). By doing so, MSA can uncover crucial information, including conserved sequences and mutation events like point mutations, insertions, and deletions, and can also help infer phylogenetic relationships. MSAs of AF2 sequences are used to extract conservation and covariation information from protein sequences exhibiting co-evolutionary relationships with the target proteins. By integrating this valuable information with structural constraints between amino acid residues, AF2 achieves high-precision and efficient predictions of the target protein's structure (Yang et al., 2023). Moreover, AF2 incorporates various optimization techniques, such as specific loss functions (Jumper et al., 2021) (e.g., frame point alignment error loss, auxiliary loss and violation loss), a recycling mechanism, self-distillation (Xie et al., 2020), and self-accuracy estimation (Jumper et al., 2021), and other methods to enhance the predictive performance of the model.





The comprehensive architecture of AF2 (Tunyasuvunakool et al., 2021; Yang et al., 2023), outlined in Figure 4, comprises three main modules: a feature extraction module, an encoder module, and a structure decoding module. The input module initiates a search for sequences homologous to the template in the sequence database and performs MSA, which reveals similarity and co-evolution information between the protein sequences and is crucial for accurate protein structure predictions. Simultaneously, the input module checks for homologous sequences with known three-dimensional structures and constructs a pairwise distance matrix in the protein structure database to depict the spatial distance between each pair of amino acids. The input module then generates MSA representations and pair representations, which capture co-evolution information and structural constraint features, respectively. The generated MSA pairwise representations are fed into the encode module, which is composed of Evoformer and infers both spatial and evolutionary relationships between proteins using the collected co-evolution information. In the final module, the structure decode module, the output of the encode module is converted into the three-dimensional structure of the target protein. The encoding module and the structure decoding module continuously optimize the predicted structure through the recycling mechanism (Jumper et al., 2021; Yang et al., 2023).

2.2 Highlights and limitations of AF2

AF2 utilizes various deep learning training methods combined with efficient search algorithms to collect information from protein sequences and structural data, resulting in more accurate predictions of unknown protein structures.

2.2.1 The neural network architecture adopted by AF2

AF2 uses the Evoformer to learn features of protein sequences and structures from different perspectives. The Evoformer consists of two sets of MSA-Transformer-based modules, which operate on the original MSA and pairwise information and combines a gated mechanism and an attention mechanism to dynamically adjust the network's output based on the input information (Makkuva et al., 2020). The MSA row-wise gated self-attention mechanism enables the model to capture long-range dependencies in amino acid sequences and protein structures, while the MSA column-wise gated self-attention mechanism allows for element exchange between different species. The Evoformer also learns the geometric constraints inside protein molecules through a triangular self-attention mechanism. The structure decoding module is based on methods similar to MSA-Transformer, encoding residue geometry into a directed reference frame in three-dimensional space (Jumper et al., 2021). AF2 also allows the model to update and optimize its output several times throughout the recycling mechanism to achieve better convergence and stability.

2.2.2 Databases and search algorithms adopted by AF2

AF2 utilizes sequence data from excellent protein sequence databases such as MGnify, Uniclust30, Uniref90, and the Big Fantastic Database, which helps it construct high-quality MSAs (Suzek et al., 2015; Mirdita et al., 2017; Mitchell et al., 2020). The protein structure data are derived from widely recognized databases, including Protein Data Bank (PDB) and PDB70/100 (Steinegger et al., 2019). Such a large amount of amino acid sequence and structure data enables deep learning neural networks to explore various dependencies between protein sequences and

structures (Yang et al., 2023), helping to improve the accuracy of AF2 prediction results. AF2 also uses several algorithms, including JackHMMER (Johnson et al., 2010), HHblits (Remmert et al., 2011), and HHSearch (Steinegger et al., 2019), to significantly improve the search efficiency.

2.2.3 The training methods adopted by AF2

The training set of AlphaFold2 consists of 75% self-distilled data and 25% known structures from the PDB. Self-distillation is a popular method of knowledge distillation that involves the student model learning from the teacher model, thereby enhancing the model's performance and efficiency. It avoids the complexity and time costs associated with the independent training and optimization of the teacher model in traditional knowledge distillation (Zhang et al., 2019; Xie et al., 2020). During the self-distillation training phase of AF2, the model is initially trained with data from the PDB and then predicts the structures of approximately 350,000 protein sequences in the Uniclust database. These predicted structures are used as data for subsequent training, with the model being retrained on a small subset of random samples in each training cycle. To improve the model's generalization ability and predictive accuracy, the training data is enhanced through a series of data augmentation processes, including random filtering, MSA preprocessing, and amino acid cropping. Such methods allow the model to make more effective use of limited data and enhance its capability to handle different protein domains and diverse MSA data (Jumper et al., 2021).

2.2.4 The robust AF2 development community

AF2 also boasts a thriving development ecosystem, with DeepMind and researchers in related fields continually updating and expanding on it to meet their investigative needs. For instance, Evans et al. modified AF2 to facilitate predictions of multi-chain complexes, dubbing this enhanced model AlphaFold-Multimer (Evans et al., 2022). Gao et al. built upon AF2 to devise a system, AF2Complex, capable of predicting direct physical interactions between multi-protein assemblies without requiring paired MSA input (Gao et al., 2022). Wayment-Steele et al. employed sequence clustering of protein sequences based on similarity and subsequently applied AF2 to each cluster to predict alternative conformations, a methodology they termed AF-Cluster (Wayment-Steele et al., 2023). Recently, in collaboration with Isomorphic Labs, DeepMind unveiled the latest iteration of AlphaFold, AlphaFold3 (AF3), which, beyond predicting protein-protein interfaces, is capable of forecasting interactions between proteins and nucleic acids and proteins and small molecule ligands, as well as those between antigens and antibodies (Abramson et al., 2024). However, DeepMind is not releasing the AF3 as open source. The multidimensional advancements surrounding AF2 showcase its immense potential across various scientific disciplines.

2.2.5 However, AF2 lacks sufficient predictive ability for the fine structure of proteins

A study by He et al. showed that there are significant differences between the AF2-predicted structures and experimental structures in many aspects, such as the assembly of extracellular and transmembrane domains, the shape of ligand-binding pockets, and the conformation of the transduction binding

interface (He et al., 2023). The predicted structure and relative positioning of each domain in AF2 exhibit uncertainty, regardless of the confidence level. This uncertainty can be attributed to several factors (Akdel et al., 2022). One such factor is the presence of indecipherable protein disorder regions in the X-ray data used for AF2 training, which results in the generation of low-confidence, disordered segments in AF2 predictions. Another factor is that some highly confident structural domains are connected by flexible links, leading to errors in the relative positioning of the domains. This uncertainty introduces the possibility of inaccurate results or identifications in structural similarity, structure of pockets, mutational effects, or model construction. These findings underscore the highlight of experimental research in protein structure analysis and emphasize the need for manual inspection and correction of AF2-predicted structures with experimental data. Consequently, the integration of experimental data and artificial intelligence has emerged as a potential solution to addressing these challenges.

During our usage, we observed that AF2 failed to simulate the natural conformation of the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein that “pops out” due to enzymatic cleavage, regardless of whether or not a custom template was provided. We speculate that this limitation may stem from the development of AF2 based on protein structures in aqueous solutions, which are unable to replicate the effects of environmental conditions such as solvent conditions, pH, and ion strength on protein structure (Rey et al., 2023).

2.3 Other methods of protein structure prediction

Before the advent of AF2, the first generation of AlphaFold (AF1) had already made significant strides in the field of protein structure prediction by employing deep learning to forecast the distances between protein residues. AF1 constructed a potential of mean force based on these distances, which allowed for the creation of highly accurate protein structures without complex sampling procedures (Senior et al., 2020). Subsequently, AF2 has built upon these achievements by incorporating new neural network architectures and training methodologies. By integrating evolutionary, physical, and geometric insights into protein structures, AF2 has notably increased the precision of predictions, achieving atomic-level accuracy even for proteins without known homologous structures. In addition to AlphaFold, this section will introduce four other protein structure prediction models: Rosetta, RoseTTAFold All-Atom, ESMFold, and RGN2, each with its own distinctive features and strengths (Table 1).

2.3.1 Rosetta

Rosetta (Rohl et al., 2004) is a classical *de novo* protein structure prediction method based on fragment assembly, developed by the Baker Lab at the University of Washington, which has had a long-standing impact and wide application in the field of protein structure prediction. The core principle of Rosetta relies on an energy function that utilizes information from fragments of known protein structures, assembling these fragments through Monte Carlo strategies to simulate the natural folding process of proteins,

TABLE 1 Features, advantages and limitations of 5 different protein structure prediction models.

Method	Feature	Advantages	Limitations
AlphaFold2	A neural network architecture combining attention mechanisms and evolutionary information	1. High accuracy in protein structure prediction 2. Continuous updates and development	1. High computational resources requirements 2. Homologous sequence dependence 3. Lack of fine structure prediction ability
Rosetta	Uses energy functions with fragments, Monte Carlo strategies	High computational efficiency with low search space	1. Limited exploration for intricate topology proteins 2. Low-resolution energy functions
RoseTTAFold All-Atom	Merges sequence-based representations of biopolymers with atomic graph representations of small molecules and covalent modifications	Prediction of proteins, nucleic acids, small molecules, metals, covalent modifications	1. Average accuracy 2. Small training datasets
ESMFold	Utilizes protein language model with training parameters instead of MSA.	1. Faster prediction speed. 2. Efficient exploration of large-scale protein structure space	1. Limited prediction accuracy 2. Less effective with complex structures
RGN2	Uses AminoBERT language model and recurrent geometric network	Prediction of orphan and de novo-designed protein structures	1. Poor prediction with sufficient sequence homologs 2. Hard to predict beta-sheet structures 3. Limited to local dependencies between Ca atoms

thereby generating conformations close to the native state. This approach ingeniously transforms the continuous conformational space optimization problem into a discrete fragment combination optimization problem, effectively reducing the search space and enhancing computational efficiency.

Nonetheless, Rosetta is accompanied by several drawbacks (Simkovic et al., 2017; Kuenze and Meiler, 2019). 1) When dealing with proteins of high molecular weight or those possessing intricate topologies, the conformational search strategy based on fragment assembly may fall short in thoroughly exploring the complete conformational space. Consequently, this limitation can lead to the omission of the globally optimal solution. 2) Employing low-resolution energy functions, while enhancing computational tractability, inadvertently compromises the precision in depicting detailed interactions.

2.3.2 RoseTTAFold All-Atom

RoseTTAFold All-Atom (RFAA) is a deep learning network that extends the capabilities of conventional protein structure prediction (Krishna et al., 2024; Marchal, 2024). It incorporates the ability to simulate complete biological assemblies, encompassing proteins, nucleic acids, small molecules, metals, and covalent modifications. RFAA merges sequence-based representations of biopolymers with atomic graph representations of small molecules and covalent modifications to predict the three-dimensional structures of these biological assemblies. This enables RFAA to predict the structure of biomolecules more comprehensively, not limited to pure protein systems alone. In terms of protein structure prediction accuracy, RFAA is on par with AF2.

While RFAA has an immediate effect in protein-small molecule binding design and complex biomolecular assembly modeling, its accuracy still needs to be further improved (Krishna et al., 2024).

The RFAA's training set is relatively small, so larger training datasets are needed to improve prediction accuracy for novel protein-small molecule complexes.

2.3.3 ESMFold

ESMFold (Lin et al., 2023; Meng et al., 2023) is a protein structure prediction method built upon pretrained language models capable of directly generating atomic-level three-dimensional spatial structures from a single protein sequence, eliminating the need for multiple sequence alignments or external modeling programs. It employs the extensive pretraining of the ESM-2 protein language model, currently the largest with 15 billion training parameters (Lin et al., 2023), as a replacement for MSA. The predictive performance of ESMFold in terms of structure improved with both the size of the language model and the comprehension of the protein sequence, which exhibited a negative correlation with perplexity. Notably, the prediction speed of ESMFold is one order of magnitude faster than that of MSA-based methods, enabling efficient exploration of large-scale protein structure space.

However, ESMFold is not without its challenges (Lin et al., 2023). 1) The accuracy of ESMFold predictions shows a negative correlation with the perplexity of the sequence, implying difficulty in inferring the structure when the language model struggles to comprehend the sequence. 2) Currently, there is a disparity in the prediction ability of more intricate structures, such as multiple chains or complexes, compared with that of AF2. Further refinement and optimization of ESMFold are required to bridge this gap.

2.3.4 RGN2

RGN2 (Chowdhury et al., 2022) represents an innovative approach to protein structure prediction that utilizes language models and deep learning to directly generate three-dimensional

structures from a single protein sequence, eliminating the need for multiple sequence alignments or external modeling programs. The method incorporates AminoBERT, a protein language model, along with a recurrent geometric network to forecast the local geometry of each residue. AminoBERT, employing a Transformer-based architecture, captures latent structural information from unaligned protein sequences. The recurrent geometric network predicts the local geometry of each residue using a rotation matrix, ensuring rotational and translational invariance and avoiding unrealistic torsion angles. Notably, RGN2 excels in predicting the structures of orphan and de novo-designed proteins, which traditionally poses challenges for MSA-based methods.

Nevertheless, RGN2 exhibits certain limitations: 1) When applied to proteins with sufficient sequence homologs to generate multiple sequence alignments (MSAs), RGN2 underperforms compared to AF2 which utilizes MSA for protein structure prediction. 2) Challenges persist for RGN2 in accurately predicting beta-sheet structures from single sequences, particularly for orphan and designed proteins. 3) RGN2 primarily predicts local dependencies between C α atoms and does not directly consider arbitrary pairwise dependencies across the entire protein structure.

3 The application of AF2 in the diagnostic strategies for disease

3.1 AF2 in antigen research and design of immunological tool in disease diagnosis

In clinical practice, the immunological assays based on antigen-antibody interactions are an important method for identifying pathogenic agents. Utilizing known antibodies or antigens, we can detect their counterparts in test samples. Concurrently, exploring the structure, functionality, and mutations of pathogenic antigens deepens our comprehension of pathogen traits, supports vaccine creation, and aids in identifying receptors that bind to pathogens. This comprehensive strategy in immunology significantly improves diagnostic accuracy. Presently, AF2 has been effectively employed in a variety of research and design endeavors related to immunology. In this section, we will explore its applications within this field.

3.1.1 Antigen structure prediction and function analysis

AF2 is widely employed in the prediction of antigenic structures in pathogens, analysis of the structural and functional characteristics of antigen proteins, and assessment of the impact of antigenic variations. For example, Hu et al. utilized AF2 to predict the novel fold of the rotavirus glycan-binding domain, which was confirmed through X-ray crystallography (Hu et al., 2022). Veit et al. (2022) used AF2 to predict the structure of the Gp5/M dimer of porcine respiratory and reproductive syndrome virus (PRRSV) and analyzed the heterogeneity of PRRSV Gp5 signal peptide cleavage sites. Both Fang et al. and Yang et al. employed AF2 to predict the structure of the S protein of SARS-CoV-2 and its Omicron variant (Yang et al., 2021; Fang et al., 2023). They investigated the impact of mutations in the S protein on its binding arrangement and affinity to the ACE2 receptor. Yang et al. used AF2 to create a high-precision structural model (pLDDT>70) and compared it with experimental

data, considering the root mean square deviation (RMSD) values and amino acid charge properties. The results indicated that the Omicron variant affects the interaction between the RBD region of the S protein and ACE2 without altering the interaction site. Additionally, Fang et al. utilized ColabFold (Mirdita et al., 2022), a protein-protein complex prediction model based on AF2, to analyze the S protein complex with two co-receptors, AXL and LDLRAD3. Based on the predictive complex model, they found that the binding modes of AXL and LDLRAD3 are different: AXL binds to the NTD region of S protein, while LDLRAD3 binds to the RBD region of S protein, and there are competitive binding sites with ACE2. These findings align with their experimental results.

3.1.2 AF2 in immunology-related design

There are numerous applications for immunological study in disease diagnosis and prevention, including pathogen detection, antibody level measurement, immune cell analysis, and vaccine development. These tests require antibodies that can react immunologically with the target antigen, as well as auxiliary anti-antibodies as detection tools. AF2 and its derivative models possess the capability to predict protein structures and protein-protein docking. Numerous experiments have demonstrated that AF2 can accurately predict vaccine and antibody structures, as well as optimize antibody-antigen complexes. This highlights its potential in designing tools for immunological detection.

3.1.2.1 Antibody design

Antibodies serve as critical tools for immunological detection. Their ability to bind to antigens largely relies on the topological complementarity between the variable domain of antibodies and the spatial structure of antigen epitopes (Graham et al., 2019). Therefore, accurate identification of the antibody structure and a precise understanding of the antibody-antigen (Ab-Ag) interface (i.e., the antibody epitope) are essential for antibody design (Sela-Culang et al., 2013; Guest et al., 2021; Hummer et al., 2022). Due to the superior performance of AF2 in predicting protein structures, it has been used in several studies to predict antibody structures and epitopes.

There are two major obstacles in predicting antibody structures: 1) determining the relative orientation of the heavy chain (Vh) and light chain (Vl) domains and 2) predicting the complementary determining regions (CDRs), especially highly variable and conformationally diverse CDR-H3 loop structures (Jaszczyszyn et al., 2023). Polonsky et al. achieved highly accurate predictions of 50% of the positions within the Fab region of 222 antibodies using AF2, with an average TM-score of 0.83 for individual Vh and Vl (Polonsky et al., 2023). This not only implies identical folding but also signifies very close proximity between the predicted and native structures. Ruffolo et al. tested the performance of AF2 and AlphaFold-Multimer in antibody structure prediction (Evans et al., 2022; Ruffolo et al., 2023) and found that AlphaFold-Multimer can accurately predict the backbone structure of antibodies, the relative orientation of Vh and Vl, and the CDR loop structure. For the relative orientation of Vh and Vl, they calculated the orientation coordinate distance (OCD) (Marze et al., 2016) of the predicted models to determine the accuracy of the relative orientation between Vh and Vl in the predicted models. The results indicate that the Fv (variable region of antibody) structure

predicted by AlphaFold-Multimer has an OCD of 4.18, which is within one standard deviation of the native structure. Moreover, AlphaFold-Multimer demonstrated sub-angstrom accuracy in predicting the CDR1 and CDR2 loop structures, and for CDR3, it exhibited greater prediction accuracy and novel predicted structures compared to many other models, demonstrating superior performance in predicting antibodies such as the PDB identifier 7N3G. AF2 performs best in predicting the CDR structures of nanobodies, as it considers various structural arrangements during the training process, giving it an advantage in predicting the secondary structures of nanobodies. Although both AlphaFold-Multimer and AF2 can predict the structure of antibody CDRs, their ability to predict CDR-H3 loop structures is still insufficient. Continuous updates to AF2 may improve this issue in the future.

The advancement of deep learning methods has allowed researchers to work toward enhancing the accuracy of antibody epitope prediction through the integration of models that combine sequences and structures and incorporate both local and global features (Zeng et al., 2023). Researchers have leveraged AF2's remarkable monomer protein structure prediction capabilities to forecast antibody epitopes, utilizing the predicted antibody structure model as input data for the prediction system (Desta et al., 2023a; Desta et al., 2023b). Desta et al. devised a method for antibody epitope prediction known as PIPER-Map (Desta et al., 2023b). This approach utilizes AF2 to anticipate antibody structures and employs the docking program PIPER, which is based on fast Fourier Transform (FFT), to perform docking between the antibody models and antigens. The docking results are subsequently ranked for analysis. Studies have shown that this method predicts antibody epitope structures with excellent accuracy, with the AF2 predictions comparable to those based on existing antibody crystal structures. In addition, Desta et al. reviewed the advanced antibody epitope localization software ServerClusPro AbEMap Web Server and investigated the effectiveness of predicting antibody epitopes using the AF2 prediction model as input (Desta et al., 2023a). The results indicated that the antibody epitope predictions generated by AF2 were similar to those generated based on established antibody structure templates, with improved predictive power for partial antibody epitopes such as PDB ID 2W9D compared to X-ray structures. Notably, the performance of AF2 for antibody epitope prediction using existing antibody templates was inferior to that achieved without utilizing antibody templates for prediction.

3.1.2.2 Optimization of antigen-antibody docking models

Antigen-antibody binding serves as the foundation for immunoassays and holds significant value in medical and immunological research. However, the current challenge lies in achieving effective antigen-antibody docking, and a universal solution to this problem remains elusive (Hogues et al., 2018). Despite these obstacles, AF2 can make robust predictions of protein-protein binding, and it has been successfully used to predict structural aspects of antigen-antibody docking and assess the outcomes of the predictions.

In a study by Yin et al., the ability of AF2 to predict antigen-antibody docking was scrutinized using over 400 nonredundant antigen-antibody complexes (Yin and Pierce, 2024). Their findings indicated that the their than-latest version of AlphaFold, v.2.3, has a higher prediction success rate compared to the

previous version, v.2.2. Additionally, the updated AlphaFold demonstrated increased efficacy in predicting nano antigen-antibody docking, underscoring the potential of AF2 in identifying antigen-antibody docking structures. This research emphasized that the accuracy of AF2 can be improved by optimizing the framework or model, enhancing sequence information within the MSA, and establishing a positive correlation between subunit prediction accuracy and the success rate of antigen-antibody interaction predictions. Consequently, the modification of AF2's architecture, particularly the structural module, holds promise for augmenting prediction accuracy by integrating contemporary factors (Abanades et al., 2023; Ruffolo et al., 2023) that enhance antibody prediction precision, potentially refining AF2's overall predictive capabilities.

Gaudreault et al. (2023) used AF2 to augment the predictive accuracy of antigen-antibody docking structures, refining the expected docking models and improving early success rates. They employed standardized pLDDT and pTMscore ($Z_{pTMscore}$ and Z_{pLDDT}) to compute a composite score, the AF2Composite score, which measures the confidence levels associated with these docking models (Eq. 1). The experimental results demonstrate the practicality, simplicity, and efficacy of this scoring method, which is free from the constraints of a specific physical methodology and remains uninfluenced by any subjective biases introduced during training or calibration. Notably, the correlation between the score and the experimentally observed docking structure strengthened with increasing quality of the predicted docking models. For instance, when $R^2 < 0.4$ (indicating poor mutual correlation between pLDDT and pTMscore), the correlation is significant only for models of acceptable quality. For models exhibiting superior prediction quality, the score proves instrumental in elevating the ranking of true positives within the predictive structure, thereby enhancing the discriminatory ability of these prediction models in the negative/positive classification of antibody-antigen docking.

$$AF2_{Composite} = Z_{pLDDT} + Z_{pTMscore} \quad (1)$$

3.1.3 Vaccine design

The vaccine development for respiratory syncytial virus (RSV) has demonstrated the importance of structure-based vaccine design (Graham et al., 2019). Using AF2 to predict protein structures could aid in structure-based design, potentially overcoming difficulties faced in previous vaccine development work.

Currently, various antibodies targeting the hemagglutinin (HA) stem region have been identified as neutralizing antibodies against influenza B virus (IBV). Therefore, vaccines designed based on HA can broadly prevent IBV infection. Zheng et al. used AF2 to design a hemagglutinin stem cell vaccine specific to IBV, named "B60-Stem-8071" (Zeng et al., 2022). They used AF2 to predict the vaccine's structure and screened for vaccine sequences that could correctly fold and maintain the natural conformation of the HA stem region in prokaryotic systems. Additionally, to enhance the stability of the HA stem region structure and improve the immune response against HA vaccine *in vivo*, they rationalized and engineered the epitope linker of the neutralizing antibody CR8071 using AF2, connecting the optimized structure to the vaccine, allowing it to target the CR8071 epitope.

3.2 AF2 in biochemical studies

3.2.1 Development of auxiliary protein targets and biomarkers

Proteins that perform crucial functions in vital life processes, such as enzymes, receptors, and ion channels, serve as significant targets for biochemical detection and drug therapy. While protein-protein interaction has been identified as a new path to discover protein targets (Liu et al., 2024), the structure and function of novel proteins are often difficult to determine. Studying protein targets with AF2 can not only predict the interaction between proteins to find protein targets but can also improve the understanding of protein structure and function, accelerate drug design, and contribute to advances in biology and medicine. Gómez-Marín et al. used AF2 to predict the structure and interaction domain of high mobility group 20A (HMG20A) and PHD Finger Protein 14 (PHF14) and found that they form a stable nuclear complex through coiled-coil domain interactions, identifying them as potential protein targets (Gómez-Marín et al., 2022). It can affect important biological processes, such as epithelial–mesenchymal transition and the TGF and Hippo signaling pathways.

Transmembrane proteins are recognized as significant targets in drug design. Hegedűs et al. reported that AF2 can accurately predict the structure of transmembrane proteins, highlighting the usefulness of AF2 in transmembrane protein studies (Hegedűs et al., 2022). This study provides valuable information for research into the ability of transmembrane proteins to correct structural errors, discover new conformational states, and simulate kinetic processes. Loring et al. used AF2 to predict the structures of different subtypes of resistance to inhibitors of cholinesterase 3 (RIC-3) (Loring, 2022). Based on these predicted structures, they analyzed how RIC-3 interacts with the $\alpha 7$ nicotinic receptor ($\alpha 7$ nAChR) subunits and promotes the folding and assembly of the $\alpha 7$ nAChR into the final conformation and subsequently proposed two possible models for the interaction between RIC-3 and $\alpha 7$ nAChR.

The function of these critical proteins often relies on their essential active residues. When the structures of these residues change, it can lead to alterations in protein function and concentration, which frequently preludes the onset of disease. Several studies have utilized AF2 to gain insight into protein function, uncover protein interactions, and identify crucial protein active sites, contributing to the advancement of disease diagnosis. Freeman et al. (2023) used AF2 to construct a structural model of the nuclease Ankyrin Repeat and LEM Domain Containing 1 (ANKLE1) and analyze its key active residues. The results indicated that the mutation of each of these residues impaired enzyme activity. ATG8/LC3 is the key protein involved in the autophagic process, and the ATG8-interacting motif/LC3-interacting region (AIM/LIR) facilitates the binding of ATG8 to autophagy cargo receptors and adaptors (Fracchiolla et al., 2017). Ibrahim et al. used AlphaFold-Multimer to analyze the spatial structure of the ATG8/LC3 protein family and accurately predicted the pockets formed by both typical and atypical AIM/LIR within the family (Ibrahim et al., 2023). The functions and effects of these pockets in the autophagy pathway were further analyzed in this way. They also utilized three pathogen virulence factors to demonstrate that AlphaFold-Multimer could effectively identify motifs from a variety of AIMS that bind ATG8.

Proteins can serve as molecular biomarkers and are frequently utilized for early disease screening, diagnosis, prognosis assessment, individualized treatment plan formulation, and prediction of adverse drug reactions (Aronson and Ferner, 2017). The development and screening of characteristic molecular biomarkers are crucial for determining the specificity and accuracy of molecular disease diagnosis (Molinski et al., 2020). Proteins with specific modifications during disease development, along with their crucial active residues, can serve as biomarkers of disease. Consequently, AF2's ability to investigate protein targets and their associated residues could significantly contribute to biomarker development. Zhuo et al. used next-generation sequencing (NGS) to determine the amino acid sequences of the immunoglobulin and T-cell receptor V-(D)-J region in bone marrow samples of 47 children with precursor B-cell acute lymphoblastic leukemia (pre-B-ALL), and they used AF2 to predict the protein structure based on the results (Zhuo et al., 2023). They extracted the immunoglobulin heavy chain gene (IGH) CDR3 consensus sequence with rod-shaped α -helix structure similarity from the predicted protein structure as an IGH rod-shaped tracker. They further validated the predictive value of the IGH rod tracker using published IGH data from an additional 203 children with pre-B-ALL. They found that the prognosis for children who tested positive for NGS-IGH was poorer than that of those who tested negative, and they also found that the protein structure encoded by the IGH CDR3 was consistent across all NGS-IGH (+) samples. These findings suggested that the sequence could serve as a marker for monitoring minimal residual disease in children during treatment.

3.2.2 Characterization of effect of mutation on enzyme activity and the difference of enzyme activity among different subtypes

Enzyme activity and enzyme metabolites are two crucial markers in biochemical detection. Alterations in either can signify changes in associated physiological indicators and the onset of related diseases. AF2 has been widely used to study the effects of structural differences and variations in enzyme activity and enzyme metabolites, providing a basis for biochemical detection and mechanism interpretation of enzymes involved in vital activities. Aminolevulinic acid synthase (ALAS), a key regulator of catalytic heme synthesis during the initial steps of key enzymes (Taylor and Brown, 2022; Freeman et al., 2023), can carry a mutation in the extended C-terminus of the erythroid isoform (ALAS2) that impacts its ability to efficiently catalyze heme synthesis, resulting in increased risk of X-linked protoporphyria. Hunter et al. used AF2 to study the structural differences among various ALAS variants, as well as the mechanism by which the C-terminal extension of ALAS controls the rate of porphyrin synthesis (Hunter and Ferreira, 2022). They predicted the structure of six mammalian ALAS subtypes and compared the predicted structure of ALAS1 with that of ALAS2. They found that the CXXC motif and the heme regulatory motifs (HRM) 4 and 5, which extend the C-terminus of ALAS, regulate ALAS activity. Their analysis of the ALAS1 structure revealed that the CXXC motif forms disulfide bonds in its oxidized state, causing HRM4 and HRM5 to fold and thereby preventing their inhibitory effect. The CXXC motif is reduced to expose HRM4 and HRM5, inhibiting excessive heme synthesis. Furthermore, the different positions of HRM4 and HRM5 in ALAS2 compared to those in

ALAS1 prevent the closure of HRM4 and HRM5 at the extended C-terminus, resulting in the inability of the cellular redox state to regulate excessive hemoglobin concentrations.

Wiedemann-Steiner syndrome (WDSTS) is a neurodevelopmental disorder caused by *de novo* mutation of lysine methyltransferase 2A (KMT2A, a multidomain histone methyltransferase) (Jones et al., 2012). Reynisdottir et al. (2022) reported that the onset of WDSTS was closely related to the loss of the ability to recognize and bind unmethylated CpG in the CXXC domain of KMT2A due to variation. They used AF2 to predict the structure of various variations in the CXXC domains and established a high-precision classification scheme for the effects of these variations. All possible missense variations in the CXXC domain were predicted, and the variants were classified into three types based on the predicted results: no effect, damage to DNA binding, or non-folding of the domain. This allowed for the accurate determination of potential pathogenicity and effects on function that the missense variations in the CXXC domain have, thereby providing a reference resource for disease diagnosis.

3.3 AF2 in molecular biology studies

3.3.1 Proteomic research

Proteomic research involves the qualitative and quantitative study of proteins with the aim of understanding the mechanisms by which they carry out their physiological activities and exploring disease process and pathogenicity to guide diagnosis and novel drug development (Hanash, 2003; Aslam et al., 2017). Technological advancements have allowed proteomics to play a pivotal role in disease diagnosis. By comparing protein expression and functional changes between control and case groups, researchers can study specific protein characteristics associated with disease. This aids in early disease diagnosis and prognostic monitoring while also allowing for the analysis of individual protein variations to inform personalized diagnosis and medical treatment. However, due to the dynamic range and large scale of the proteome, traditional mass spectrometry methods still face challenges in terms of data acquisition and verification. With its strong data processing and mining capabilities (Zhang et al., 2014), AF2 is able to predict the three-dimensional structure of single-chain proteins as well as of protein complexes, making it particularly useful for proteomic studies.

3.3.1.1 The function of AF2 as a proteomic tool

Functional proteomics is the study of protein-to-protein or protein-to-nucleic acid interactions in a specific time and space, focusing on a functional subgroup of proteins within a cell. AF2 has been widely used in functional proteomics research due to its excellent predictive speed and accuracy, enabling large-scale research and cluster analysis of protein functions (Huang et al., 2023). By searching for proteins containing the Z-DNA/Z-RNA binding protein (Zα) domain in the AF2 predictive structure database, Bartas et al. identified 185 proteins that may bind to Z-DNA/Z-RNA and play an important role in a variety of cellular processes (Bartas et al., 2022). Huang et al. (Huang et al., 2023) selected 15 genes with a length greater than 100 bp from the deaminase family, predicted their structures, and compared

them with those in the AF2 database. Based on the comparison results, a similarity matrix was generated, and a structure tree was constructed to perform a cluster analysis on the deaminase family to elucidate the structural and functional differences among different deaminases within the family. Al-Masri et al. (2023) analyzed known protein kinase structures in the AlphaFold protein structure database to predict the specific structures of several protein kinases, subsequently using Smina to perform molecular docking on protein kinase crystals matching the protein kinase structure to evaluate the effectiveness of AF2 in virtual filtering. The results show that AF2 can effectively simulate kinase active sites that are highly characteristic of conformational states, providing a foundation for the study of protein kinase pathogenicity and the development of new drugs based on kinase active sites.

3.3.1.2 Establishment of protein database

AF2 provides a high-quality and efficient method for generating and analyzing large-scale protein structure databases, which is crucial in proteomic research (Domon and Aebersold, 2006; Fremdling et al., 2022). The construction of a protein information database is an essential step that significantly increases the speed of protein identification and the development of mass spectrometers. AF2 can be used to construct large-scale protein structure databases, providing rich and reliable protein structure resources for proteomic research and facilitating the establishment of relevant datasets for mass spectrometers. Varadi et al. created a comprehensive, open access database of high-accuracy protein structure predictions (Varadi et al., 2022). AlphaFold database contains a considerable number of high-accuracy protein structure prediction models, offering valuable resources for biological research. Hekkelman et al. used small molecules and ions in experimentally determined protein structures to “transplant” the protein model in the AlphaFold protein structure database, thereby establishing the AlphaFill database (Hekkelman et al., 2023). The database contains 12,029,789 “transplant” results of 995,411 AF2 models, providing relevant validation indicators and visual interfaces, enriching model information in the AlphaFold database, and offering researchers clues to new protein function hypotheses. Consequently, AF2 can deliver high-quality and efficient generation and analysis methods for the construction of large-scale protein structure databases, providing more possibilities for proteomic research and mass spectrometer development.

3.3.2 AF2 in the study of missense mutation on protein structure and function

Missense mutations can serve as biomarkers in clinical molecular biology tests. These mutations may alter the amino acid sequence and structure of proteins, thereby affecting their function and pathogenicity. Many studies have utilized AF2 to predict and compare the structures of normal and mutated proteins, thereby revealing the mechanisms and effects of missense mutations.

Wang et al. (2023) reported a novel mutation in the lysosomal membrane structural protein (LAMP2) gene and used AF2 to predict the three-dimensional structures of wild-type and mutant LAMP2. They found that the mutant LAMP2 is composed of only six amino acids and that it is unable to form functional peptides or proteins, confirming that LAMP2 deficiency is caused by this mutation. The LMNA gene encodes the lamin A/C protein,

which is involved in the construction of nuclear membranes, and mutation of LMNA results in a series of lamin diseases. Chang et al. (2023) used AF2 to predict the spatial structure of the lamin A/C mutant protein and found an interruption in the alpha-helix region. They used this protein structure to visualize the impact of the mutation on protein morphology and interaction compared to the wild-type protein. Finally, they used AF2's predictions to elucidate the mutation's pathogenicity at the protein level, revealing the function of different protein domains and potential therapeutic targets.

Despite great progress in AF2's ability to predict the structure of mutant proteins, some researchers have pointed out the limitations of AF2 in predicting the impact of missense mutations on protein stability. Buel et al. emphasized these limitations by comparing AF2-predicted models of wild-type and mutant structures of three protein domains to the experimentally determined structures of the wild-type proteins (Buel and Walters, 2022). This comparison revealed that the predicted models did not accurately reflect the structural changes and functional losses induced by the mutations. To address this issue, researchers have developed AF2 prediction models to deduce the structure and stability of proteins after mutation. For example, Iqbal et al. developed a predictive model, protein stability (PROST), that can estimate the changes in protein stability caused by single-point missense mutations (Iqbal et al., 2022). In two blind test datasets, PROST outperformed the other models in terms of predictive performance, achieving the highest Pearson's correlation coefficient and the lowest root mean squared error. This indicates that PROST has good accuracy and can serve as an important tool in the prediction of the three-dimensional structure of mutant proteins. Cheng et al. developed a model based on AF2, called AlphaMissense (Cheng et al., 2023), that was fine-tuned based on AlphaFold 2.3.0 using human and primate variant frequency data as weak labels and avoiding circularity arising from the use of manual annotations. AlphaMissense can simulate all possible single amino acid mutations and can distinguish 89% of missense variants as likely pathogenic or likely benign.

AF2 not only predicts the structural changes in proteins resulting from missense mutations but also analyzes the impact of these changes on protein function. It generates various models to predict the stability changes caused by missense mutations and the likelihood of pathogenicity. AF2 therefore plays a crucial role in missense mutation research—it can not only explain the pathogenic mechanisms of missense variants but can also identify missense mutations with potential clinical significance, providing biomarkers for disease diagnosis.

3.4 AF2 in pathogenic microbiology research

Pathogenic microorganisms play a crucial role in laboratory disease diagnosis. Factors such as biological characteristics, drug resistance, and variant typing all affect the pathogenicity of microorganisms, the symptoms of disease, and the effect of drug treatment. For example, the major resistance mechanism in MRSA is via the acquisition of the gene *mecA*, which encodes the protein PBP2a. *MecA*, however, has a significantly low affinity for β -lactam, which makes all currently available β -lactam drugs largely

ineffective for the treatment of MRSA (Peacock and Paterson, 2015). The key proteins involved in the pathogenic process of microorganisms are also important targets for drug development and screening.

Traditional laboratory diagnostic methods for pathogenic microorganisms (Rajapaksha et al., 2019) include culture and isolation, biochemical and serological detection, and immunological and nucleic acid assays. However, these methods have significant limitations, such as extended diagnostic time, low detection rate, inability to fully interact with *in vivo* infections, and inability to culture certain microorganisms. The advancement of cutting-edge biological theories and technologies, such as mass spectrometry (Schubert and Kostrzewa, 2017) and molecular diagnostics (Lai and Stayton, 2015; Visconti et al., 2017; Yasemin et al., 2019), coupled with the progress of artificial intelligence (Jumper et al., 2021; Tunyasuvunakool et al., 2021), makes it possible to examine clinical pathogenic microorganisms based on studies of the structure, function, and distribution of microbial proteins. To date, many studies have used AF2 to determine the pathogenicity, microbial resistance, and potential drug targets of microorganisms.

3.4.1 Study of pathogenic substances

Considering its direct impact on clinical manifestations and disease progression, studying the pathogenicity of microorganisms is key to revealing the core pathogenic mechanisms and promoting the identification and targeted treatment of pathogens. Use of AF2 in the in-depth analysis of the structural and functional properties of these key proteins that considers the composition of a variety of pathogenic proteins and biomolecules is driving the rapid development of the detection and treatment of pathogenic microorganisms. With the assistance of AlphaFold-Multimer, Le et al., 2023 predicted the structural model of the outer membrane lipoprotein Tle3, its cognate immune protein Tli3, and their immune complexes of adhesively invasive *Escherichia coli* (AIEC) and optimized the model through molecular replacement. They found that a β -lamellia stacking region in the C-terminal extension domain of Tli3 intercalates into the active cleave of Tle3, suggesting that Tli3 physically blocks Tle3 from contacting its substrate and thereby inhibits its phospholipase A1 activity. They used similar methods to predict the mode of interaction between Tle3 and VgrG, a protein constituting the spinous process of the type VI secretion system, and found a potentially specific interaction between the N-terminal loop of Tle3 and the C-terminal transthyretin-containing domain of VgrG. This provided vital structural and biochemical information for understanding the function and mechanism of type VI secretion system effectors and immune proteins in AIEC, which is of great significance for revealing the pathogenesis of AIEC and identifying new therapeutic targets. These findings will aid in the development of new anti-AIEC drugs or diagnostic reagents, thereby enhancing the efficiency and accuracy of clinical microbiology.

3.4.2 Assessment of genetic relationships and variability

AF2 has been used to analyze the differences in the protein structures of various strains or phages and evaluate their

genetic relationships and variability. This has proven beneficial for the classification and identification of different species of microorganisms, providing a reference for epidemiological surveillance and control. Goulet et al. employed AF2 to predict three-dimensional models of the components of the adhesion apparatus of two bacteriophage types, OE33PA (Jaomanjaka et al., 2018) and Vinitor162 (Philippe et al., 2020), that infect *Oenococcus oeni* (Goulet and Cambillau, 2021). Based on the known architecture of the phage adhesion apparatus, a topological model was reconstructed. OE33PA possesses an evolved distal tail protein (Dit) (Veesler et al., 2010) and an exotic receptor-binding protein (RBP), composed of two domains similar to the RBPs of different phages, and forms a chimeric structure. By contrast, Vinitor162 has a long tail-associated lysozyme protein (Tal) that is rich in carbohydrate-binding modules (CBMs). This finding suggests distinct infection mechanisms between OE33PA and Vinitor162: OE33PA employs a dual binding strategy involving its Dit-CBM and RBP head domain to engage receptors on the host cell wall for entry, whereas Vinitor162 utilizes a multipoint attachment mode through its Tal-CBM and RBD to infect host cells by interacting with receptors on the host cell wall. Monzon et al. used AF2 to predict structures lacking known adhesion domains in more than 6,500 credible fibrillar adnexins and identified 24 potential novel families of adhesion protein domains, 15 of which showed structural similarity to known adhesion domains. This contributes to the discovery of novel bacterial interaction mechanisms (Monzon and Bateman, 2022).

3.4.3 Research on drug resistance

AF2 can be used to predict mutation-induced changes in the protein structure of microorganisms as well as unreported protein structures, thereby assisting in the analysis of microbial resistance mechanisms. Multidrug-resistant *Acinetobacter baumannii* (*A. baumannii*) is one of the leading pathogenic causes of severe nosocomial infections. *A. baumannii* CipA has been identified as a plasminogen-binding and complement-inhibitory protein that plays a significant role in its immune evasion process. The use of AF2 in the structural prediction of CipA aptly explained the results obtained from several CipA variants (Ries et al., 2022). According to the structural prediction of AF2, replacing the glutamic acid (E) at position 360 with a proline (P) will induce a significant structural change in the C-terminal region of the DUF4377 domain, and the hydrogen bond pairing of the adjacent β -fold is completely lost. This change greatly inhibits the ability of CipA to interact with complement factor I, which will provide potential targets for new therapeutic interventions.

Some researchers have also successfully predicted unresolved structures using AF2. Willems et al. used the AF2 algorithm to solve domain structures that were not resolved in the previously reported *Plasmodium falciparum* Chloroquine Resistance Transporter (PfCRT) protein 7G8 isoform cryo-EM structure (Willems et al., 2023). When cryo-EM was used to analyze the 7G8 isoform of PfCRT, many of the N- and C-termini, as well as the cytosolically disposed “loop 2” connecting TM helices 2 and 3, were not resolved, presumably due to masking by bound F’ (ab) (a type of incomplete F (ab) fragment) used in solving the cryo-EM structure and/or the intrinsic flexibility of these regions. They then performed energy minimization through Monte Carlo molecular

dynamics simulations, revealing additional structures for the previously unresolved N- and C-termini. These results are crucial for understanding the structure and function of PfCRT, the mechanism of chloroquine resistance, and the development of novel second-tier drug therapies active against chloroquine-resistant malaria. The above examples further highlight the significant application and value of AF2 in biomedical research.

3.4.4 Design of drug targets

In recent years, the emergence of drug-resistant strains has gradually diminished the therapeutic effect of antibiotics on pathogenic microbial infections (Davies and Davies, 2010). To address these new challenges in anti-infection treatment and drug screening, some studies have employed AF2 to study the structure and function of proteins related to pathogenic microorganisms. This has advanced research on potential drug targets, the development of antibacterial drugs, and the screening of drugs and antimicrobial peptides. Madi-Moussa et al. used AF2 to predict the structure of Lacticaseicin 30, a rare gram-positive bacteriocin that inhibits gram-negative bacteria (Madi-Moussa et al., 2022). They found that it primarily consists in the five helical segments and contains regions and amino acids involved in anti-gram-negative activity. By studying the antimicrobial activity of a series of shortened variants or those containing point mutations in the five helical segments, they mapped these regions and the amino acids involved in inhibition. These experiments showed that at least two helical segments of the N-terminal region are required for Lacticaseicin 30 inhibition of gram-negative bacteria, which will aid in the design of additional Lacticaseicin 30 variants as potential drugs treatments of gram-negative bacterial infection. Alotaibi et al. screened a series of drug target proteins against *Vibrio* by gene alignment and used AF2 to predict the three-dimensional structure of 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, a drug target protein (Alotaibi et al., 2023). Furthermore, some effective inhibitors were identified through virtual screening (Panwar et al., 2024) and molecular docking studies (Pinzi and Rastelli, 2019), and their binding stability with target proteins was verified using molecular dynamic simulations (Koirala et al., 2024).

4 Conclusion and future perspectives

Proteins play a crucial role in disease diagnosis. They serve as diagnostic indicators and detection tools, contributing to accurate diagnosis, disease prevention, and personalized medicine. AF2, a deep learning-based protein structure prediction model, achieves remarkable accuracy and rapid protein structure predictions through its unique principles and architecture. As such, it has applications in diverse areas of protein research (Yang et al., 2023).

AF2 significantly contributes to disease diagnosis by predicting antibody structures for immunological tests and vaccines, verifying antigen-antibody affinity, and aiding in diagnostic tool design. It's used to predict structures of disease-related proteins, enhancing our understanding of their structural, functional, and activity changes. These insights form the basis for improving diagnosis, prevention, and treatment. AF2 also analyzes key enzyme variations

during disease progression, establishing diagnostic criteria. It supports proteomic data analysis, database creation, and research. Additionally, AF2 assesses missense variation impacts, aiding in biomarker design. It studies pathogenic substances' functions, drug resistance, and classification by microorganisms, aiding in accurate infection diagnosis and drug target development.

AF2 is highly adaptable and presents with unlimited potential for extensive application in several biological fields. Various prediction models based on AF2 with expanded functions have emerged, examples of which include AlphaFold-Multimer (Evans et al., 2022; Yin et al., 2022; Ibrahim et al., 2023), AF2Complex (Gao et al., 2022), ColabFold (Mirdita et al., 2022), and AlphaMissense (Cheng et al., 2023). Future versions of AlphaFold may prioritize the optimization and refinement of its architecture to enhance its predictive ability and broaden its functionality (Abramson et al., 2024), for example, with revamped diffusion-based architecture, AF3 has transcended the capabilities of its predecessor by not only predicting protein structures with higher fidelity but also accurately modeling a diverse array of biomolecular complexes. However, it is important to note that AF3 is currently not available as an open-source tool. With the continuous development and in-depth research of AF2 and its derivatives, they are expected to provide broader assistance in theoretical research and direct application in disease diagnosis in the future, becoming more powerful and effective tools for disease diagnosis. We have some ideas, for instance, AF2 can predict the structures and binding interfaces of antigens and antibodies, making it invaluable to the design of immunological assays and detection tools. It can also be utilized to reverse design corresponding antibodies or antigens with high affinity based on the predicted structures. Using AF2's reverse network, protein sequences corresponding to the designed structures can be predicted (Goverde et al., 2023). Moreover, AF2 can integrate with sequencing technologies to not only detect pathogenic genes but also predict the pathogenic potential of mutations and their impact on biological activities.

During the paper-writing process, we encountered numerous applications of deep learning predictive models such as IgFold (Ruffolo et al., 2023), DeepAb (Ruffolo et al., 2022), and ImmuneBuilder (Abanades et al., 2023). These examples underline the evolving landscape of disease diagnosis, where deep learning models, driven by artificial intelligence, have the potential to facilitate the design of swift and convenient research methodologies.

References

- Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. (2023). ImmuneBuilder: deep-Learning models for predicting the structures of immune proteins. *Commun. Biol.* 6 (1), 575. doi:10.1038/s42003-023-04927-7
- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500. doi:10.1038/s41586-024-07487-w
- Akdel, M., Pires, D. E. V., Pardo, E. P., Janes, J., Zalevsky, A. O., Meszaros, B., et al. (2022). A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* 29 (11), 1056–1067. doi:10.1038/s41594-022-00849-w
- Al-Masri, C., Trozzi, F., Lin, S. H., Tran, O., Sahni, N., Patek, M., et al. (2023). Investigating the conformational landscape of AlphaFold2-predicted protein kinase structures. *Bioinform. Adv.* 3 (1), vbad129. doi:10.1093/bioadv/vbad129
- Alotaibi, B. S., Ajmal, A., Hakami, M. A., Mahmood, A., Wadood, A., and Hu, J. (2023). New drug target identification in *Vibrio vulnificus* by subtractive genome analysis and their inhibitors through molecular docking and molecular dynamics simulations. *Heliyon* 9 (7), e17650. doi:10.1016/j.heliyon.2023.e17650
- Aronson, J. K., and Ferner, R. E. (2017). Biomarkers-A general review. *Curr. Protoc. Pharmacol.* 76, 1–9. doi:10.1002/cpph.19
- Ashenberg, O., and Laub, M. T. (2013). Using analyses of amino Acid coevolution to understand protein structure and function. *Methods Enzymol.* 523, 191–212. doi:10.1016/B978-0-12-394292-0.00009-6
- Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., and Rasool, M. H. (2017). Proteomics: technologies and their applications. *J. Chromatogr. Sci.* 55 (2), 182–196. doi:10.1093/chromsci/bmw167

Author contributions

HZ: Project administration, Writing—original draft, Writing—review and editing. JL: Supervision, Writing—original draft, Writing—review and editing. HW: Writing—original draft, Writing—review and editing. RL: Writing—original draft. NZ: Supervision, Writing—original draft. XH: Conceptualization, Writing—review and editing. JY: Writing—review and editing. LC: Conceptualization, Project administration, Writing—review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was financially supported by Zhejiang Provincial Natural Science Foundation of China under Grant LQ21H200006, Zhejiang Provincial Medical and Health Science and Technology Programs (2021KY132 and 2023KY650), and the Basic Research Fund of Hangzhou Medical College (KYZD202010).

Acknowledgments

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

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.

- Bartas, M., Slychko, K., Brazda, V., Cerven, J., Beaudoin, C. A., Blundell, T. L., et al. (2022). Searching for new Z-DNA/Z-RNA binding proteins based on structural similarity to experimentally validated *za* domain. *Int. J. Mol. Sci.* 23 (2), 768. doi:10.3390/ijms23020768
- Bertoline, L. M. F., Lima, A. N., Krieger, J. E., and Teixeira, S. K. (2023). Before and after AlphaFold2: an overview of protein structure prediction. *Front. Bioinform.* 3, 1120370. doi:10.3389/fbinf.2023.1120370
- Bongirwar, V., and Mokhadde, A. S. (2022). Different methods, techniques and their limitations in protein structure prediction: a review. *Prog. Biophys. Mol. Biol.* 173, 72–82. doi:10.1016/j.pbiomolbio.2022.05.002
- Buel, G. R., and Walters, K. J. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* 29 (1), 1–2. doi:10.1038/s41594-021-00714-2
- Chang, L., Huang, R., Chen, J., Li, G., Shi, G., Xu, B., et al. (2023). An alpha-helix variant p.Arg156Pro in LMNA as a cause of hereditary dilated cardiomyopathy: genetics and bioinformatics exploration. *BMC Med. Genomics* 16 (1), 229. doi:10.1186/s12920-023-01661-1
- Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulyte, A., Applebaum, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381 (6664), eadg7492. doi:10.1126/science.adg7492
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* 40 (11), 1617–1623. doi:10.1038/s41587-022-01432-w
- Davies, J., and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* 74 (3), 417–433. doi:10.1128/MMBR.00016-10
- Destá, I. T., Kotelnikov, S., Jones, G., Ghani, U., Abyzov, M., Kholodov, Y., et al. (2023a). The ClusPro AbEMap web server for the prediction of antibody epitopes. *Nat. Protoc.* 18 (6), 1814–1840. doi:10.1038/s41596-023-00826-7
- Destá, I. T., Kotelnikov, S., Jones, G., Ghani, U., Abyzov, M., Kholodov, Y., et al. (2023b). Mapping of antibody epitopes based on docking and homology modeling. *Proteins* 91 (2), 171–182. doi:10.1002/prot.26420
- Domon, B., and Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science* 312 (5771), 212–217. doi:10.1126/science.1124619
- Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2022). Protein complex prediction with AlphaFold-Multimer. bioRxiv. Available at: <https://www.biorxiv.org/content/10.1101/2021.10.04.463034v1> (Accessed March 27, 2024).
- Fang, Q., He, X., Zheng, X., Fu, Y., Fu, T., Luo, J., et al. (2023). Verifying AXL and putative proteins as SARS-CoV-2 receptors by DnaE intein-based rapid cell-cell fusion assay. *J. Med. Virol.* 95 (7), e28953. doi:10.1002/jmv.28953
- Fracchiolla, D., Sawa-Makarska, J., and Martens, S. (2017). Beyond Atg8 binding: the role of AIM/LIR motifs in autophagy. *Autophagy* 13 (5), 978–979. doi:10.1080/15548627.2016.1277311
- Freeman, A. D. J., Declais, A. C., Wilson, T. J., and Lilley, D. M. J. (2023). Biochemical and mechanistic analysis of the cleavage of branched DNA by human ANKLE1. *Nucleic Acids Res.* 51 (11), 5743–5754. doi:10.1093/nar/gkad416
- Fremdling, P., Esser, T. K., Saha, B., Makarov, A. A., Fort, K. L., Reinhardt-Szyba, M., et al. (2022). A preparative mass spectrometer to deposit intact large native protein complexes. *ACS Nano* 16 (9), 14443–14455. doi:10.1021/acsnano.2c04831
- Gao, M., Nakajima An, D., Parks, J. M., and Skolnick, J. (2022). AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* 13 (1), 1744. doi:10.1038/s41467-022-29394-2
- Gaudreault, F., Corbeil, C. R., and Sulea, T. (2023). Enhanced antibody-antigen structure prediction from molecular docking using AlphaFold2. *Sci. Rep.* 13 (1), 15107. doi:10.1038/s41598-023-42090-5
- Gómez-Marín, E., Posavec-Marjanović, M., Zarzuela, L., Basurto-Cayuela, L., Guerrero-Martínez, J. A., Arribas, G., et al. (2022). The high mobility group protein HMG20A cooperates with the histone reader PHF14 to modulate TGFβ and Hippo pathways. *Nucleic Acids Res.* 50 (17), 9838–9857. doi:10.1093/nar/gkac766
- Goulet, A., and Cambillau, C. (2021). Structure and topology prediction of phage adhesion devices using AlphaFold2: the case of two *Oenococcus oeni* phages. *Microorganisms* 9 (10), 2151. doi:10.3390/microorganisms9102151
- Goverde, C. A., Wolf, B., Khakzad, H., Rosset, S., and Correia, B. E. (2023). *De novo* protein design by inversion of the AlphaFold structure prediction network. *Protein Sci.* 32 (6), e4653. doi:10.1002/pro.4653
- Graham, B. S., Gilman, M. S. A., and McLellan, J. S. (2019). Structure-based vaccine antigen design. *Annu. Rev. Med.* 70, 91–104. doi:10.1146/annurev-med-121217-094234
- Guest, J. D., Vreven, T., Zhou, J., Moal, I., Jeliazkov, J. R., Gray, J. J., et al. (2021). An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure* 29 (6), 606–621.e5. doi:10.1016/j.str.2021.01.005
- Hanash, S. (2003). Disease proteomics. *Nature* 422 (6928), 226–232. doi:10.1038/nature01514
- He, X. H., You, C. Z., Jiang, H. L., Jiang, Y., Xu, H. E., and Cheng, X. (2023). AlphaFold2 versus experimental structures: evaluation on G protein-coupled receptors. *Acta Pharmacol. Sin.* 44 (1), 1–7. doi:10.1038/s41401-022-00938-y
- Hegedűs, T., Geisler, M., Lukács, G. L., and Farkas, B. (2022). Ins and outs of AlphaFold2 transmembrane protein structure predictions. *Cell Mol. Life Sci.* 79 (1), 73. doi:10.1007/s00018-021-04112-1
- Hekkelman, M. L., de Vries, I., Joosten, R. P., and Perrakis, A. (2023). AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* 20 (2), 205–213. doi:10.1038/s41592-022-01685-y
- Hogues, H., Gaudreault, F., Corbeil, C. R., Deprez, C., Sulea, T., and Purisima, E. O. (2018). ProPOSE: direct exhaustive protein-protein docking with side chain flexibility. *J. Chem. Theory Comput.* 14 (9), 4938–4947. doi:10.1021/acs.jctc.8b00225
- Hu, L., Salmen, W., Sankaran, B., Lasanajak, Y., Smith, D. F., Crawford, S. E., et al. (2022). Novel fold of rotavirus glycan-binding domain predicted by AlphaFold2 and determined by X-ray crystallography. *Commun. Biol.* 5 (1), 419. doi:10.1038/s42003-022-03357-1
- Huang, J., Lin, Q., Fei, H., He, Z., Xu, H., Li, Y., et al. (2023). Discovery of deaminase functions by structure-based protein clustering. *Cell* 186 (15), 3182–3195.e14. doi:10.1016/j.cell.2023.05.041
- Hummer, A. M., Abanades, B., and Deane, C. M. (2022). Advances in computational structure-based antibody design. *Curr. Opin. Struct. Biol.* 74, 102379. doi:10.1016/j.sbi.2022.102379
- Hunter, G. A., and Ferreira, G. C. (2022). An extended C-terminus, the possible culprit for differential regulation of 5-aminolevulinate synthase isoforms. *Front. Mol. Biosci.* 9, 920668. doi:10.3389/fmolb.2022.920668
- Ibrahim, T., Khandare, V., Mirkin, F. G., Tumas, Y., Bubeck, D., and Bozkurt, T. O. (2023). AlphaFold2-multimer guided high-accuracy prediction of typical and atypical ATG8-binding motifs. *PLoS Biol.* 21 (2), e3001962. doi:10.1371/journal.pbio.3001962
- Iqbal, S., Ge, F., Li, F., Akutsu, T., Zheng, Y., Gasser, R. B., et al. (2022). PROST: AlphaFold2-aware sequence-based predictor to estimate protein stability changes upon missense mutations. *J. Chem. Inf. Model* 62 (17), 4270–4282. doi:10.1021/acs.jcim.2c00799
- Jaomanjaka, F., Claisse, O., Philippe, C., and Le Marrec, C. (2018). Complete genome sequence of lytic *Oenococcus oeni* bacteriophage OE33PA. *Microbiol. Resour. Annu* 7 (6), e00818-18. doi:10.1128/MRA.00818-18
- Jaszczyzyn, I., Bielska, W., Gawłowski, T., Dudzic, P., Satlawa, T., Konczak, J., et al. (2023). Structural modeling of antibody variable regions using deep learning-progress and perspectives on drug discovery. *Front. Mol. Biosci.* 10, 1214424. doi:10.3389/fmolb.2023.1214424
- Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinform.* 11, 431. doi:10.1186/1471-2105-11-431
- Jones, W. D., Dafou, D., McEntagart, M., Woollard, W. J., Elmslie, F. V., Holder-Espinasse, M., et al. (2012). *De novo* mutations in MLL cause Wiedemann-Steiner syndrome. *Am. J. Hum. Genet.* 91 (2), 358–364. doi:10.1016/j.ajhg.2012.06.008
- 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
- Koirala, K., Joshi, K., Adediwura, V., Wang, J., Do, H., and Miao, Y. (2024). “Accelerating molecular dynamics simulations for drug discovery,” in *Computational drug discovery and design*. Editors M. Gore, and U. B. Jagtap (New York, NY: Springer US), 187–202.
- Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., et al. (2024). Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* 384 (6693), ead12528. doi:10.1126/science.ad12528
- Kryshtafovich, A., Schwede, T., Topf, M., Fidelis, K., and Moulton, J. (2021). Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* 89 (12), 1607–1617. doi:10.1002/prot.26237
- Kuenze, G., and Meiler, J. (2019). Protein structure prediction using sparse NOE and RDC restraints with Rosetta in CASP13. *Proteins* 87 (12), 1341–1350. doi:10.1002/prot.25769
- Lai, J. J., and Stayton, P. S. (2015). Improving lateral-flow immunoassay (LFIA) diagnostics via biomarker enrichment for mHealth. *Methods Mol. Biol.* 1256, 71–84. doi:10.1007/978-1-4939-2172-0_5
- Le, T. T. H., Kellenberger, C., Boyer, M., Santucci, P., Flaugnatti, N., Cascales, E., et al. (2023). Activity and crystal structure of the adherent-invasive *Escherichia coli* tle3/tli3 T6SS effector/immunity complex determined using an AlphaFold2 predicted model. *Int. J. Mol. Sci.* 24 (2), 1740. doi:10.3390/ijms24021740
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379 (6637), 1123–1130. doi:10.1126/science.adc2574
- Liu, J. X., Zhang, X., Huang, Y. Q., Hao, G. F., and Yang, G. F. (2024). Multi-level bioinformatics resources support drug target discovery of protein-protein interactions. *Drug Discov. Today* 29 (5), 103979. doi:10.1016/j.drudis.2024.103979

- Loring, R. H. (2022). Speculation on how RIC-3 and other chaperones facilitate $\alpha 7$ nicotinic receptor folding and assembly. *Molecules* 27 (14), 4527. doi:10.3390/molecules27144527
- Madi-Moussa, D., Deracinois, B., Teiar, R., Li, Y., Mihasan, M., Flahaut, C., et al. (2022). Structure of Lactacasein 30 and its engineered variants revealed an interplay between the N-terminal and C-terminal regions in the activity against gram-negative bacteria. *Pharmaceutics* 14 (9), 1921. doi:10.3390/pharmaceutics14091921
- Makkuva, A., Oh, S., Kannan, S., and Viswanath, P. (2020). "Learning in gated neural networks," in *International conference on artificial intelligence and statistics* (Vienna, Austria: PMLR), 3338–3348.
- Marchal, I. (2024). RoseTTAFold expands to all-atom for biomolecular prediction and design. *Nat. Biotechnol.* 42 (4), 571. doi:10.1038/s41587-024-02211-5
- Marze, N. A., Lyskov, S., and Gray, J. J. (2016). Improved prediction of antibody VL-VH orientation. *Protein Eng. Des. Sel.* 29 (10), 409–418. doi:10.1093/protein/gzw013
- Meng, Q., Guo, F., and Tang, J. (2023). Improved structure-related prediction for insufficient homologous proteins using MSA enhancement and pre-trained language model. *Brief. Bioinform.* 24 (4), bbad217. doi:10.1093/bib/bbad217
- Mirdita, M., Schütze, K., Moriawaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* 19 (6), 679–682. doi:10.1038/s41592-022-01488-1
- Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Soding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* 45 (D1), D170–D176. doi:10.1093/nar/gkw1081
- Mitchell, A. L., Almeida, A., Beracocha, M., Boland, M., Burgin, J., Cochrane, G., et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* 48 (D1), D570–D578. doi:10.1093/nar/gkz1035
- Molinski, J., Tadimety, A., Burklund, A., and Zhang, J. X. J. (2020). Scalable signature-based molecular diagnostics through on-chip biomarker profiling coupled with machine learning. *Ann. Biomed. Eng.* 48 (10), 2377–2399. doi:10.1007/s10439-020-02593-y
- Monzon, V., and Bateman, A. (2022). Large-scale discovery of microbial fibrillar adhesins and identification of novel members of adhesive domain families. *J. Bacteriol.* 204 (6), e0010722. doi:10.1128/jb.00107-22
- Paiva, V. A., Gomes, I. S., Monteiro, C. R., Mendonça, M. V., Martins, P. M., Santana, C. A., et al. (2022). Protein structural bioinformatics: an overview. *Comput. Biol. Med.* 147, 105695. doi:10.1016/j.cmpbiomed.2022.105695
- Panwar, U., Murali, A., Khan, M. A., Selvaraj, C., and Singh, S. K. (2024). "Virtual screening process: a guide in modern drug designing," in *Computational drug discovery and design*. Editors M. Gore, and U. B. Jagtap (New York, NY: Springer US), 21–31.
- Pazos, F., and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. *EMBO J.* 27 (20), 2648–2655. doi:10.1038/emboj.2008.189
- Peacock, S. J., and Paterson, G. K. (2015). Mechanisms of methicillin resistance in *Staphylococcus aureus*. *Annu. Rev. Biochem.* 84, 577–601. doi:10.1146/annurev-biochem-060614-034516
- Philippe, C., Chaib, A., Jaomanjaka, F., Claisse, O., Lucas, P. M., Samot, J., et al. (2020). Characterization of the first virulent phage infecting *Oenococcus oeni*, the queen of the cellars. *Front. Microbiol.* 11, 596541. doi:10.3389/fmicb.2020.596541
- Pinzi, L., and Rastelli, G. (2019). Molecular docking: shifting paradigms in drug discovery. *Int. J. Mol. Sci.* 20 (18), 4331. doi:10.3390/ijms20184331
- Polonsky, K., Pupko, T., and Freund, N. T. (2023). Evaluation of the ability of AlphaFold to predict the three-dimensional structures of antibodies and epitopes. *J. Immunol.* 211 (10), 1578–1588. doi:10.4049/jimmunol.2300150
- Prijbelski, A. D., Korobeynikov, A. I., and Lapidus, A. L. (2019). "Sequence analysis," in *Encyclopedia of bioinformatics and computational biology*. Editors S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Academic Press), 292–322.
- Rajapaksha, P., Elbourne, A., Gangadoo, S., Brown, R., Cozzolino, D., and Chapman, J. (2019). A review of methods for the detection of pathogenic microorganisms. *Analyst* 144 (2), 396–411. doi:10.1039/c8an01488d
- Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., et al. (2021). "MSA transformer," in *Proceedings of the 38th international conference on machine learning*. Editors M. Marina, and Z. Tong (Vienna, Austria: PMLR), 8844–8856.
- Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9 (2), 173–175. doi:10.1038/nmeth.1818
- Rey, J., Murail, S., de Vries, S., Derreumaux, P., and Tuffery, P. (2023). PEP-FOLD4: a pH-dependent force field for peptide structure prediction in aqueous solution. *Nucleic Acids Res.* 51 (W1), W432–W437. doi:10.1093/nar/gkad376
- Reynisdottir, T., Anderson, K. J., Boukas, L., and Björnsson, H. T. (2022). Missense variants causing Wiedemann-Steiner syndrome preferentially occur in the KMT2A-CXXC domain and are accurately classified using AlphaFold2. *PLoS Genet.* 18 (6), e1010278. doi:10.1371/journal.pgen.1010278
- Ries, J. I., Hess, M., Nouri, N., Wichelhaus, T. A., Gottig, S., Falcone, F. H., et al. (2022). CipA mediates complement resistance of *Acinetobacter baumannii* by formation of a factor I-dependent quadripartite assemblage. *Front. Immunol.* 13, 942482. doi:10.3389/fimmu.2022.942482
- Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004). "Protein structure prediction using Rosetta," in *Methods in enzymology* (Academic Press), 66–93.
- Ruffolo, J. A., Chu, L. S., Mahajan, S. P., and Gray, J. J. (2023). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat. Commun.* 14 (1), 2389. doi:10.1038/s41467-023-38063-x
- Ruffolo, J. A., Sulam, J., and Gray, J. J. (2022). Antibody structure prediction using interpretable deep learning. *Patterns (N Y)* 3 (2), 100406. doi:10.1016/j.patter.2021.100406
- Schubert, S., and Kostrzewa, M. (2017). MALDI-TOF MS in the microbiology laboratory: current trends. *Curr. Issues Mol. Biol.* 23, 17–20. doi:10.21775/cimb.023.017
- Sela-Culang, I., Kunik, V., and Ofra, Y. (2013). The structural basis of antibody-antigen recognition. *Front. Immunol.* 4, 302. doi:10.3389/fimmu.2013.00302
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577 (7792), 706–710. doi:10.1038/s41586-019-1923-7
- Simkovic, F., Ovchinnikov, S., Baker, D., and Rigden, D. J. (2017). Applications of contact predictions to structural biology. *IUCr* 4 (Pt 3), 291–300. doi:10.1107/s205252517005115
- Steinegger, M., Meier, M., Mirdita, M., Vohringer, H., Haunsberger, S. J., and Soding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinforma.* 20 (1), 473. doi:10.1186/s12859-019-3019-7
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31 (6), 926–932. doi:10.1093/bioinformatics/btu739
- Taylor, J. L., and Brown, B. L. (2022). Structural basis for dysregulation of aminolevulinic acid synthase in human disease. *J. Biol. Chem.* 298 (3), 101643. doi:10.1016/j.jbc.2022.101643
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596 (7873), 590–596. doi:10.1038/s41586-021-03828-1
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50 (D1), D439–D444. doi:10.1093/nar/gkab1061
- Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Neural information processing systems* (Long Beach, California, USA: Curran Associates Inc), 6000–6010.
- Veesler, D., Robin, G., Lichiere, J., Auzat, I., Tavares, P., Bron, P., et al. (2010). Crystal structure of bacteriophage SPP1 distal tail protein (gp19.1): a baseplate hub paradigm in gram-positive infecting phages. *J. Biol. Chem.* 285 (47), 36666–36673. doi:10.1074/jbc.M110.157529
- Veit, M., Gadalla, M. R., and Zhang, M. (2022). Using AlphaFold2 to predict the structure of the Gp5/M dimer of porcine respiratory and reproductive syndrome virus. *Int. J. Mol. Sci.* 23 (21), 13209. doi:10.3390/ijms232113209
- Visconti, V., Brunetti, G., Giordano, A., and Raponi, G. (2017). RT-PCR for the diagnosis of *Clostridium difficile* infection: the final answer has yet to come. *J. Clin. Pathol.* 70 (12), 1090–1091. doi:10.1136/jclinpath-2017-204523
- Wang, Y., Bai, M., Zhang, P., Peng, Y., Chen, Z., He, Z., et al. (2023). Identification and functional analysis of a novel *de novo* missense mutation located in the initiation codon of LAMP2 associated with early onset female Danon disease. *Mol. Genet. Genomic Med.* 11 (9), e2216. doi:10.1002/mgg3.2216
- Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Homberger, M., et al. (2023). Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* 625, 832–839. doi:10.1038/s41586-023-06832-9
- Willems, A., Kalaw, A., Ecer, A., Kotwal, A., Roepe, L. D., and Roepe, P. D. (2023). Structures of Plasmodium falciparum chloroquine resistance transporter (PfCRT) isoforms and their interactions with chloroquine. *Biochemistry* 62 (5), 1093–1110. doi:10.1021/acs.biochem.2c00669
- Xie, Q., Luong, M. T., Hovy, E., and Le, Q. V. (2020). "Self-training with noisy student improves ImageNet classification," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020, 10684–10695.
- Yang, Q., Syed, A. A. S., Fahira, A., and Shi, Y. (2021). Structural analysis of the SARS-CoV-2 Omicron variant proteins. *Res. (Wash D C)* 2021, 9769586. doi:10.34133/2021/9769586
- Yang, Z., Zeng, X., Zhao, Y., and Chen, R. (2023). AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct. Target Ther.* 8 (1), 115. doi:10.1038/s41392-023-01381-z
- Yasemin, A., Ahmad, S., Afzal, S., Ullah, A., and Sheed, A. (2019). Evaluation of GeneXpert MTB/RIF assay for detection of pulmonary tuberculosis on sputum samples. *J. Coll. Physicians Surg. Pak* 29 (1), 66–69. doi:10.29271/jcpsp.2019.01.66

- Yin, R., Feng, B. Y., Varshney, A., and Pierce, B. G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci.* 31 (8), e4379. doi:10.1002/pro.4379
- Yin, R., and Pierce, B. G. (2024). Evaluation of AlphaFold antibody-antigen modeling with implications for improving predictive accuracy. *Protein Sci.* 33 (1), e4865. doi:10.1002/pro.4865
- Zeng, D., Xin, J., Yang, K., Guo, S., Wang, Q., Gao, Y., et al. (2022). A hemagglutinin stem vaccine designed rationally by AlphaFold2 confers broad protection against influenza B infection. *Viruses* 14 (6), 1305. doi:10.3390/v14061305
- Zeng, X., Bai, G., Sun, C., and Ma, B. (2023). Recent progress in antibody epitope prediction. *Antibodies (Basel)* 12 (3), 52. doi:10.3390/antib12030052
- Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., and Ma, K. (2019). "Be your own teacher: improve the performance of convolutional neural networks via self distillation," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October 2019 - 02 November 2019, 3712–3721.
- Zhang, Z., Wu, S., Stenoien, D. L., and Pasa-Tolic, L. (2014). High-throughput proteomics. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* 7, 427–454. doi:10.1146/annurev-anchem-071213-020216
- Zhuo, Z., Wang, Q., Li, C., Zhang, L., Zhang, L., You, R., et al. (2023). IGH rod-like tracer: an AlphaFold2 structural similarity extraction-based predictive biomarker for MRD monitoring in pre-B-ALL. *iScience* 26 (7), 107107. doi:10.1016/j.isci.2023.107107



OPEN ACCESS

EDITED BY

Joshua S. Chappie,
Agricultural Research Service (USDA),
United States

REVIEWED BY

Yao Zhang,
Michigan State University, United States
Ruby Sharma,
Albert Einstein College of Medicine,
United States

*CORRESPONDENCE

N. V. Petukhova,
✉ petuhovanv@1spbmgmu.ru

RECEIVED 30 May 2024

ACCEPTED 17 September 2024

PUBLISHED 03 October 2024

CITATION

Bug DS, Moiseev IS, Porozov YB and
Petukhova NV (2024) Shedding light on the
DICER1 mutational spectrum of uncertain
significance in malignant neoplasms.
Front. Mol. Biosci. 11:1441180.
doi: 10.3389/fmolb.2024.1441180

COPYRIGHT

© 2024 Bug, Moiseev, Porozov and
Petukhova. 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.

Shedding light on the *DICER1* mutational spectrum of uncertain significance in malignant neoplasms

D. S. Bug¹, I. S. Moiseev², Yu. B. Porozov^{3,4} and N. V. Petukhova^{1*}

¹Bioinformatics Research Center, Pavlov First Saint Petersburg Medical State University, St. Petersburg, Russia, ²R. M. Gorbacheva Scientific Research Institute of Pediatric Hematology and Transplantation, Pavlov First Saint Petersburg State Medical University, St. Petersburg, Russia, ³St. Petersburg School of Physics, Mathematics, and Computer Science, HSE University, Saint Petersburg, Russia, ⁴Advitam Laboratory, Belgrade, Serbia

The Dicer protein is an indispensable player in such fundamental cell pathways as miRNA biogenesis and regulation of protein expression in a cell. Most recently, both germline and somatic mutations in *DICER1* have been identified in diverse types of cancers, which suggests Dicer mutations can lead to cancer progression. In addition to well-known hotspot mutations in RNAase III domains, *DICER1* is characterized by a wide spectrum of variants in all the functional domains; most are of uncertain significance and unstated clinical effects. Moreover, various new somatic *DICER1* mutations continuously appear in cancer genome sequencing. The latest contemporary methods of variant effect prediction utilize machine learning algorithms on bulk data, yielding suboptimal correlation with biological data. Consequently, such analysis should be conducted based on the functional and structural characteristics of each protein, using a well-grounded targeted dataset rather than relying on large amounts of unsupervised data. Domains are the functional and evolutionary units of a protein; the analysis of the whole protein should be based on separate and independent examinations of each domain by their evolutionary reconstruction. Dicer represents a hallmark example of a multidomain protein, and we confirmed the phylogenetic multidomain approach being beneficial for the clinical effect prediction of Dicer variants. Because Dicer was suggested to have a putative role in hematological malignancies, we examined variants of *DICER1* occurring outside the well-known hotspots of the RNase III domain in this type of cancer using phylogenetic reconstruction of individual domain history. Examined substitutions might disrupt the Dicer function, which was demonstrated by molecular dynamic simulation, where distinct structural alterations were observed for each mutation. Our approach can be utilized to study other multidomain proteins and to improve clinical effect evaluation.

KEYWORDS

Dicer1, variant of uncertain significance, variant effect prediction, gene evolution, oncology, molecular dynamics

1 Introduction

Dicer1 is a double-stranded RNA (dsRNA) endoribonuclease playing a central role in short dsRNA-mediated post-transcriptional gene splicing. It is responsible for cleaving naturally occurring long dsRNAs and short hairpin pre-microRNAs (miRNA) into 21–23-nucleotide-long fragments with a two-nucleotide 3′ overhang, producing short interfering RNAs (siRNA) and mature microRNAs (miRNAs) (Ha and Kim, 2014; Yang and Lai, 2011; Foulkes et al., 2014). These small RNAs serve as guides that direct the RNA-induced silencing complex (RISC) to complementary RNAs for its degradation or translation prevention. Gene silencing mediated by siRNAs (RNA interference) controls the degradation of exogenous RNA along with the elimination of transcripts from mobile and repetitive DNA elements triggered by endogenous loci that affect gene expression and genome organization (Wilson and Doudna, 2013; Okamura and Lai, 2008). Thus, Dicer1 plays a key role in the overall protein translational control within the canonical miRNA biogenesis pathway (Fabian and Sonenberg, 2012).

Advances in understanding the genetic and molecular functions of Dicer1 have led to new insights into its role in cancer progression (Robertson et al., 2018; Caroleo et al., 2021; Vedanayagam et al., 2019). Mutations in the *DICER1* gene were associated with a predisposition to multiple cancer types—the *DICER1* syndrome—which is characterized by disrupted miRNA biogenesis and processing with subsequent disruption in the control of gene expression (Hill et al., 2009). Missense mutations associated with *DICER1* syndrome were reported in various types of tumors: endocrine tumors, pleuropulmonary blastoma, cystic nephroma, rhabdomyosarcoma, multinodular goiter, thyroid cancer, ovarian Sertoli–Leydig cell tumor, neuroblastoma, and other neoplasias (Robertson et al., 2018). More than four thousand *DICER1* variants are available in the ClinVar database, which makes it the 19th most frequently mutated gene according to this database. Nearly half of the reported variants (2140) have unknown clinical effects, and the overwhelming majority of these are represented by missense mutations (Vogelstein et al., 2013).

Recent studies highlight the significance of miRNA biogenesis genes in hematological malignancies that are under mutational pressure during tumor progression. In particular, the downregulated expression of *DICER1* was revealed in mesenchymal stem cells (MSCs) from myelodysplastic syndrome patients (Santamaría et al., 2012). Furthermore, selective deletion of the *DICER1* gene in murine mesenchymal osteoprogenitors induces markedly disordered hematopoiesis with several MDS features, indicating the crucial role of this gene in mesenchymal “stroma” as a primary regulator of tissue function (Raaijmakers et al., 2010). Recent analysis of MDS clinical data revealed the high mutational burden in both miRNA processing genes and their association with common MDS mutations (Moiseev et al., 2021). Therefore, functional classification of variants that are currently listed as variants of uncertain significance is critically important for a fundamental understanding of *DICER1* functions as well as its role in cancer and utility in clinical diagnostics.

In this study, we evaluated the evolutionary history of Dicer1 and presented a multiple sequence alignment of Dicer1 orthologs

TABLE 1 Dicer domains.

Name (annotation rule)	Start	End
Helicase ATP-binding domain (PRU00541)	51	227
Helicase C-terminal domain (PRU00542)	433	602
Dicer double-stranded RNA-binding fold domain (PRU00657)	630	722
PAZ (PRU00142)	891	1,042
RNase III (PRU00177)	1,276	1,403
RNase III (PRU00177)	1,666	1,824
Double-stranded RNA-binding domain (PRU00266)	1,849	1,914

suitable for the interpretation of variants observed in this gene. We also show that some evolutionarily intolerable variants negatively affect the structural stability of Dicer1.

2 Materials and methods

2.1 Homology study

We carried out a BLAST search of the human Dicer protein (isoform 1, accession number NP_001258211.1) against the NCBI RefSeq protein database (Altschul et al., 1990; O’Leary et al., 2016). The resulting hits were sorted by E-value, and the first 1,387 sequences, consisting of Dicer1 proteins, a known outgroup—insect Dicer2, and a number of similar proteins were aligned using the MAFFT algorithm v7 (Katoh et al., 2002). The maximum-likelihood tree was inferred from the acquired multiple sequence alignment (MSA) using iqTree utility v2 (Minh et al., 2020) with the LG + R10 model resolved by ModelFinder (Kalyanamoorthy et al., 2017). Branch support was assessed with ultrafast bootstrap approximation [UFBoot (Minh et al., 2013; Hoang et al., 2017), 1,000 replicates]. We selected Dicer1 proteins from the tree, omitting Dicer2 paralogs, and generated a full-sequence MSA using MAFFT. Sequences with ambiguous amino acids were removed from the MSA, and misaligned amino acids were masked manually by observing the proximities of insertions and deletions in aligned sequences.

2.2 MSA refinement

Domain coordinates were obtained from PROSITE (Table 1) (Sigrist et al., 2013). Based on these coordinates, Dicer1 MSA was split into MSAs of its domains and non-domain subsequences, including interdomain, initial, and terminal sections that do not belong to any domain. All 15 subsequent MSAs were realigned by MAFFT, and then erroneous and incomplete sequences were discarded. Finally, the full-length Dicer1 MSA was assembled.

2.3 Selection of mutations for analysis

The missense mutations of *DICER1* in hematological malignancies were obtained from the COSMIC database (<https://cancer.sanger.ac.uk/cosmic>) (Tate et al., 2018) by filtering the variants in hematological and lymphoid tissues. Variants located in Dicer1 domains but not in RNase III were analyzed.

2.4 Protein structure modeling

All stages of protein modeling and analytical calculations were performed using the Schrödinger molecular modeling suite (version 2021-1) (Schrödinger, LLC, New York, NY, 2021). A Dicer full-length 3D-structure PDB ID AF-Q9UPY3-F1 predicted by AlphaFold (Jumper et al., 2021) was selected from the UniProt database (UniProt IDs Q9UPY3) (<https://www.uniprot.org/>). To ensure the AlphaFold structure was reliable and accurate, we performed the topological similarity analysis by TM-score calculation (Xu and Zhang, 2010) with the experimental Dicer structure: the TM-score was 0.8053 compared with 5ZAK for the Dicer model (Liu et al., 2018). The quality of the Dicer structure was tested and preprocessed in the Protein Preparation Wizard (PPW) (Madhavi Sastry et al., 2013). Detected problems and additional loop refinement were resolved in the Prime package (Jacobson et al., 2004). No problems were reported in the processed protein structure.

2.5 Molecular dynamics (MD) simulations

MD simulations were performed using the Desmond package (Bowers et al., 2006). The MD system was set up in “System Builder” in Maestro as follows: the TIP3P water model (Jorgensen et al., 1983) was used to simulate water molecules; the buffer distance in the orthorhombic box was set at 10 Å; a recalculated amount of Na⁺/Cl⁻ ions were added to balance the system charge and placed randomly to neutralize the solvated system; additional salt was appended for final concentration 0.15 M in order to simulate physiological conditions.

Molecular dynamic simulations were conducted with the periodic boundary conditions in the NPT ensemble class using OPLS3e force field parameters (Harder et al., 2015; Roos et al., 2019). The temperature and pressure were kept at 300 K and 1 atmospheric pressure, respectively, using Nosé–Hoover temperature coupling and isotropic scaling (Nosé, 1984). The model system was relaxed before simulations using Maestro’s default relaxation protocol, which includes two stages of minimization (restrained and unrestrained), followed by four stages of MD runs with gradually diminishing restraints. MD simulations were carried out with 100 ns and 300 ns runs and recording the trajectory configurations obtained at 50 ps intervals.

2.6 Protein site-specific mutagenesis

Initially, the preprocessed and refined structure of wild-type Dicer was relaxed by MD simulation for 100 ns in order to obtain

the relaxed system with minimized energy. The recorded trajectories were clustered, and the total energies of the representative structures were calculated in Prime (selected parameters VSGV and OPLS3e). The structure with the lowest energy was employed in further long MD simulations and protein mutagenesis. Specific mutations were introduced into the structure by the 3D Builder Panel in Maestro, and side-chain rotamers were refined. The local structure around the inserted mutation was minimized; the 10 amino acids loop around the introduced mutation was refined in the Prime package, followed by side-chain prediction to locate an appropriate residue conformation. The quality of the mutated model was validated in PPW as previously (Section 2.4), and given Dicer, mutated structures were subjected to 300-ns MD simulation.

2.7 Analysis of the MD simulation

The MD trajectory files were investigated by using simulation quality analysis (SQA) and simulation event analysis (SEA) along with simulation interaction diagram (SID) programs available with the Desmond module: root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), total intra-molecular hydrogen bonds (Hbonds Intra), radius of gyration (Rgyr), and secondary structure elements (SSE) were calculated and visualized. The recorded trajectories were clustered, and the total energies of the representative structures were calculated in Prime (options VSGV and OPLS3e). Additionally, the H-bonds formed by mutated residue with the whole protein molecule were computed by `analyse_trajectory_ppi.py` script and SEA, and the interactions were compared with the WT structure. To characterize the local changes induced by mutation, the region radius of 10 Å around the introduced residue was analyzed by calculating local RMSD, H-bonds, Rgyr, and surface area (the clustered structures with minimal total energy were used to measure the surface area in 10 Å radius of mutated residue).

3 Results

3.1 Obtaining variants of uncertain significance in *DICER1*

We examined ClinVar (Landrum et al., 2017), a public archive of human genetic variants, to identify known and predicted pathogenic and benign amino acid substitutions in *DICER1*, as well as missense variants of uncertain significance (VUSs). In total, we found 2002 variants, and more than 91% of them are VUS (accessed March 2022). We found 45 variants annotated as intolerant (11 likely pathogenic and 34 pathogenic) and 44 variants annotated as tolerant (36 likely benign and eight benign) (Figure 1). Importantly, only six hotspot positions in Dicer1 have been reported: E1705, D1709, D1713, G1809, D1810, and E1813 (Chen et al., 2018; Klein et al., 2014).

VUSs present a substantial challenge in the clinical context (Federici and Soddu, 2020), and current efforts by the scientific community focus on developing easily applicable methods for their classification. Widely used variant effect prediction tools (CADD, PROVEAN, SIFT, PolyPhen, SNAP, PhD-SNP, and MAPP) were

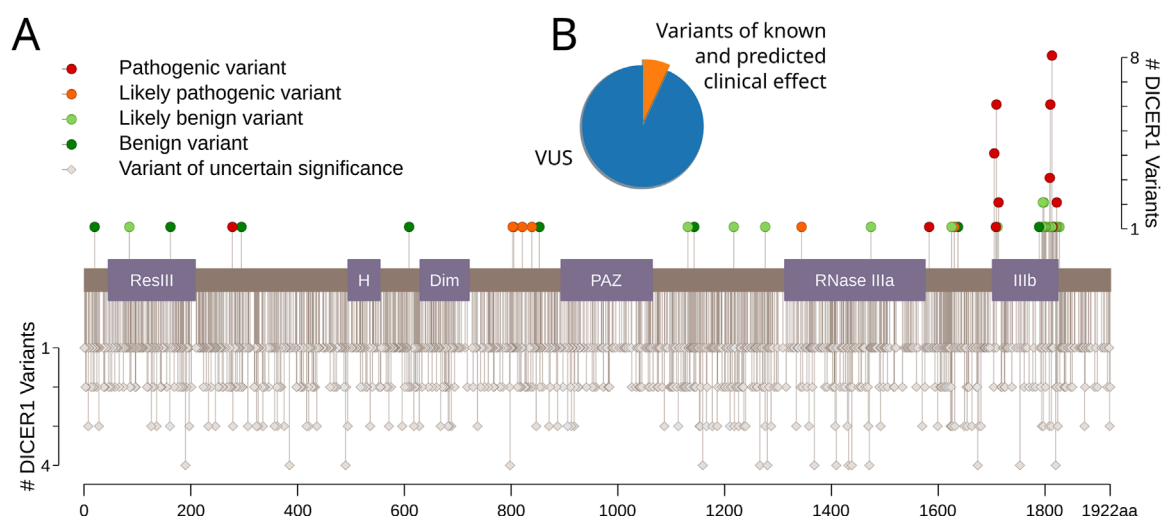


FIGURE 1
Dicer1 variants with known and predicted clinical effects. **(A)** The lollipop plot of Dicer1 variants with different clinical effects from ClinVar (accessed December 2021). Protein domains are indicated as follows: ResIII: type III restriction enzyme, res subunit; H: Helicase conserved C-terminal domain; Dim: Dicer dimerization domain; PAZ: PAZ domain; RNase IIIa: ribonuclease IIIa domain; IIIb: ribonuclease IIIb domain. **(B)** Pie chart showing the distribution of variants with known and predicted clinical effects versus VUS.

applied to identify missense mutations that are assumed to lead to *DICER1*-associated cancers. Surprisingly, one of the latest prediction models, EVE, did not provide resolution for mutations past position 1789, which leaves unresolved substitutions at several known hotspots, such as G1809, D1810, and E1813 (Kock et al., 2019). As for other tools, even in cases of known pathogenic mutations, their expected accuracy levels ranged from 65% to 80% (Thusberg et al., 2011). This low accuracy is primarily due to misalignments and the inclusion of low-quality sequences, paralogs, and remote homologs that are not functionally equivalent.

To overcome problems associated with the use of automated variant predictors, we constructed our own datasets of well-defined Dicer1 orthologs based on its evolutionary history and domain architecture and used these datasets to derive a risk map for *DICER1*-associated cancer.

3.2 Constructing the dataset

After sponges diverged from the main animal branch, but before the cnidarian split, *DICER1* was duplicated, resulting in two paralogs, *DICER1* and *DICER2*, (Mukherjee et al., 2012). The roles of these paralogs are different. Dicer1 functions in miRNA-based gene regulation (Welker et al., 2011), whereas Dicer2 is responsible for antiviral immunity (Kolaczowski et al., 2010). As Dicer2 was subsequently lost in some metazoans, including vertebrates, Dicer1 gained some of its functions (Hammond, 2005). For clarification purposes, we will use the asterisk to label such a multifunctional *DICER1* * gene and its Dicer1 * protein where necessary.

To establish the precise evolutionary history of Dicer, we first collected its homologs by carrying out a BLAST search initiated with the human Dicer protein (isoform 1, accession number NP_001258211.1) against the NCBI RefSeq protein database (Altschul et al., 1990; O'Leary et al., 2016). The resulting hits were

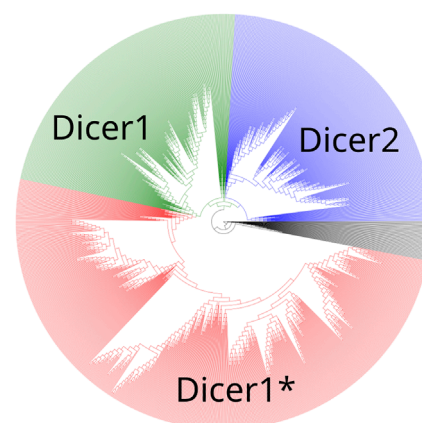


FIGURE 2
A maximum-likelihood phylogenetic tree of the Dicer group. Dicer1, Dicer2, and Dicer1 * subclades are demonstrated.

sorted by E-value, and the first 1,387 sequences, consisting of Dicer1 proteins, a known outgroup—insect Dicer2, and a number of similar proteins were aligned using the MAFFT algorithm v7 (Katoh et al., 2002). The maximum-likelihood tree was inferred from the acquired MSA using iqTree utility v2 (Minh et al., 2020) (Figure 2, Supplementary File S1) with the LG + R10 model resolved by ModelFinder (Kalyaanamoorthy et al., 2017). Branch support was assessed with ultrafast bootstrap approximation [UFBoot (Minh et al., 2013; Hoang et al., 2017), 1,000 replicates].

A maximum-likelihood phylogenetic tree showed two distinct clades corresponding to Dicer1 and Dicer2, and all Dicer1 * sequences formed a distinct subclade within the Dicer1 group (Figure 2).

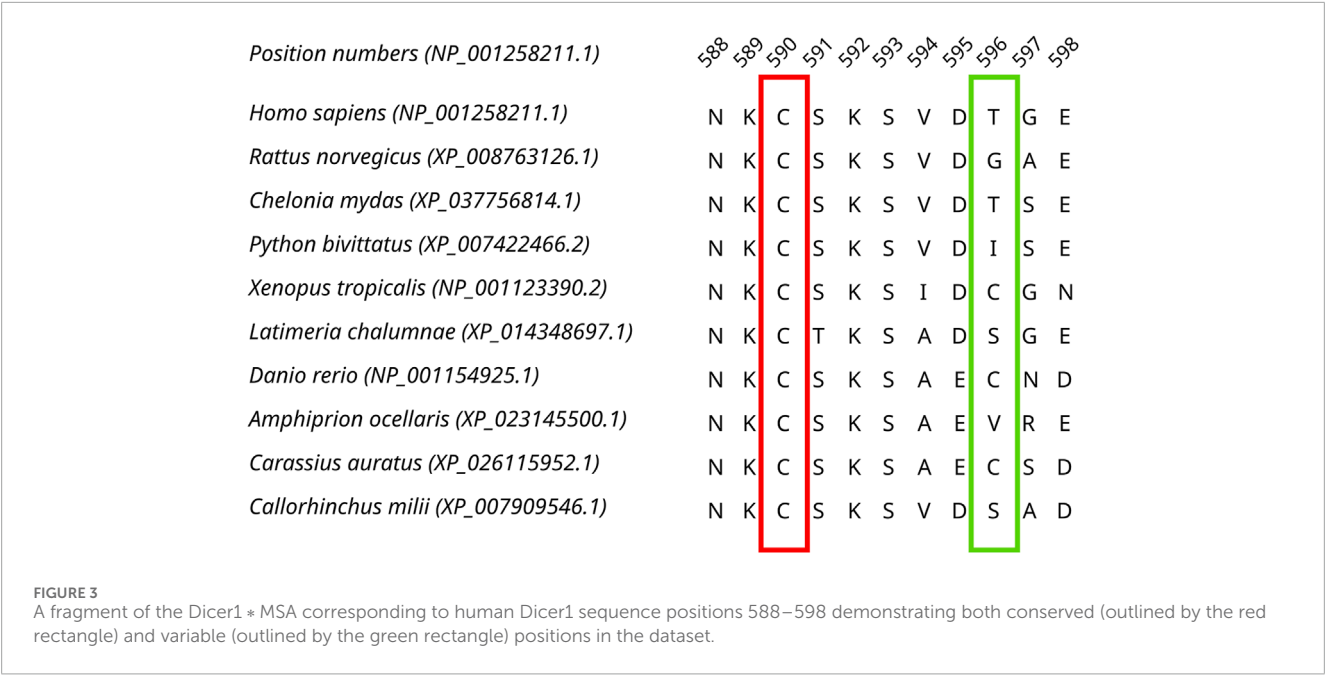


TABLE 2 Examples of variant assessment before and after domain realignment.

Variant	Known clinical effect	Tolerant substitutions	
		Before the realignment	After the realignment
E1705K	Pathogenic	E, K, N	E
E1813K	Pathogenic	E, K, L, V	E
D1822V	Pathogenic	D, C, V, K, Y	D

Sequences from the Dicer1 * sub-clade were aligned, and by identifying both invariant and highly variable positions in the MSA (Figure 3), we concluded that there was enough time for orthologs to diverge.

Next, we inspected the alignment and noticed misalignments in some Dicer1 domains. To mitigate this problem, we split the MSA of full-length protein sequences into subsequences corresponding to human Dicer1 domain coordinates and realigned sequences of each domain separately. Erroneous and incomplete sequences were removed from domain MSAs. After the realignment, we reassembled the full-length Dicer1 MSA (termed “final MSA”), resulting in a reduction in the number of misaligned regions and improving predictions according to known clinical effects for some positions (Table 2).

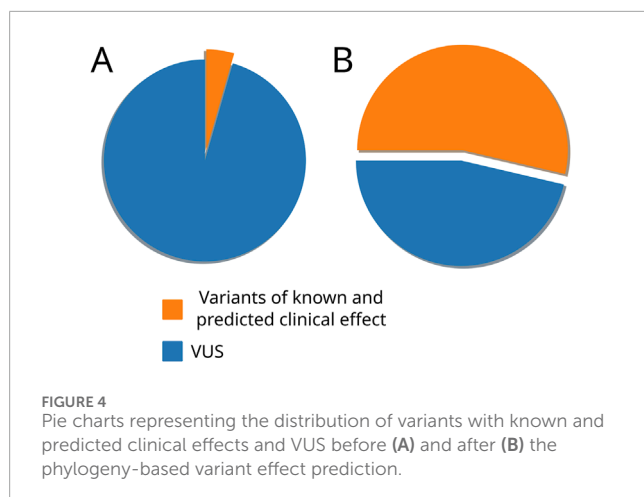
3.3 Variant effect interpretation

We collected a total of 1834 unique missense VUSs from the ClinVar database, and their positions were examined in the final MSA. We adopted the following straightforward reasoning to evaluate variants, similar to a previously reported protocol (Adebali et al., 2016): if a variant occurs in an invariant position or

if it is not seen in a highly conserved position of the final MSA, then it is intolerant. If a variant exists in any of the final MSA sequences, then it is evolutionarily allowable or tolerant. We also ensured that only single substitutions serve as evidence for benignity, and if each substitution in an examined position is accompanied by another one in an adjoining position, then the tested variant is uninterpretable. This approach allowed us to assign 485 variants as tolerant and 588 variants as intolerant and thus potentially damaging substitutions (Figure 4).

We also used the SAVER algorithm (Adebali et al., 2016) to evaluate variants against the final MSA, and it confirmed 1,067 of our 1,073 predictions. Satisfactorily, known *DICER1* hotspot mutations were evolutionarily intolerable in our final MSA and consequently were predicted as damaging (Table 2), thus providing a positive control for our analysis (Supplementary File S2).

Producing a high-quality final MSA of Dicer1 orthologs distinguishes our approach from automated variant predicting bioinformatics tools. For example, in our final MSA, position M1808 is invariable; therefore, any variant in this position is evolutionarily intolerant and thus damaging. It is worth noting that M1808 is adjacent to three known Dicer1 hotspots, G1809, D1810, and E1813, which further reinforces its potential significance. However, automated tools provide conflicting and erroneous assignments



for a documented VUS in this position: M1808L (dbSNP ID: rs763241498), is predicted to be “possibly damaging” by PolyPhen2 (which is a less confident prediction than “probably damaging”), “tolerated” by SIFT, and “neutral” by PROVEAN, whereas EVE did not provide any interpretation of this variant. These erroneous assignments result from “noisy” MSAs used by these tools. For example, we have identified several paralogs (Dicer2 sequences) in some of these MSAs (Supplementary Figure S1).

3.4 Assessment of selected *DICER1* mutations in hematological malignancies

Advances in understanding the genetic and molecular functions of Dicer1 have opened new horizons into its role in cancer progression with questions that remain unanswered (Robertson et al., 2018; Caroleo et al., 2021). We made sure all known Dicer1 hotspots were completely conserved in the MSA and turned to less-studied cases. Recent studies highlight the significance of miRNA biogenesis genes in hematological malignancies that are under mutational pressure during tumor progression, and their disruption can alter the cellular proliferation through miRNA regulation. Therefore, the investigation of mutations' pathogenicity in the context of oncohematology might shed light on the functional importance of these proteins and the mutations acquired under tumor evolution.

To demonstrate the validity of our approach, we selected four VUS that are located within functional domains of Dicer1 but outside known hot spots: Y124H (COSMIC database identifier: COSV100601713), located in the Helicase ATP-binding domain, I445S (COSV58619533) and F508C (COSV58616328), located in the Helicase domain C-terminus, and T993R (COSV58617548), located in the PAZ domain. In addition to assessing the evolutionary tolerability of each variant, we performed molecular dynamics (MD) simulations of mutated Dicer1 proteins to evaluate potential structural alterations caused by these mutations.

All four selected variants were found to be evolutionarily intolerable by our approach. None of these specific substitutions were found in the multiple alignments of (i) Dicer1 orthologs

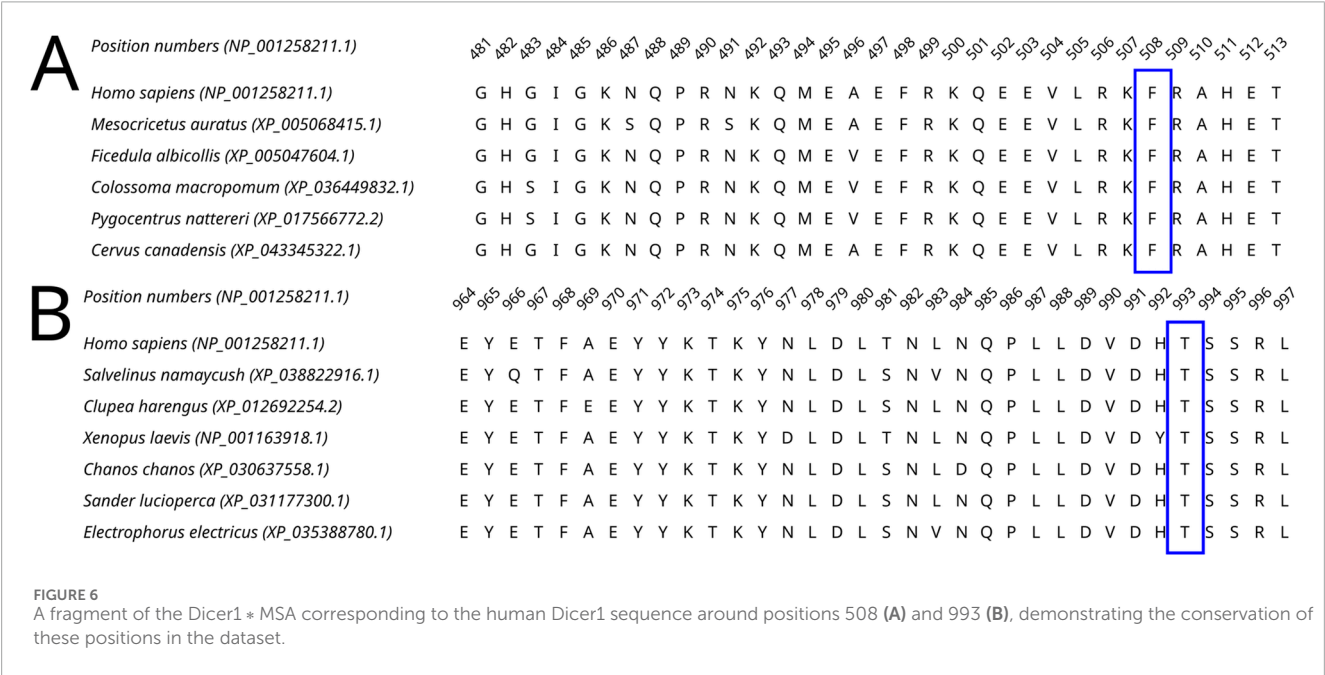
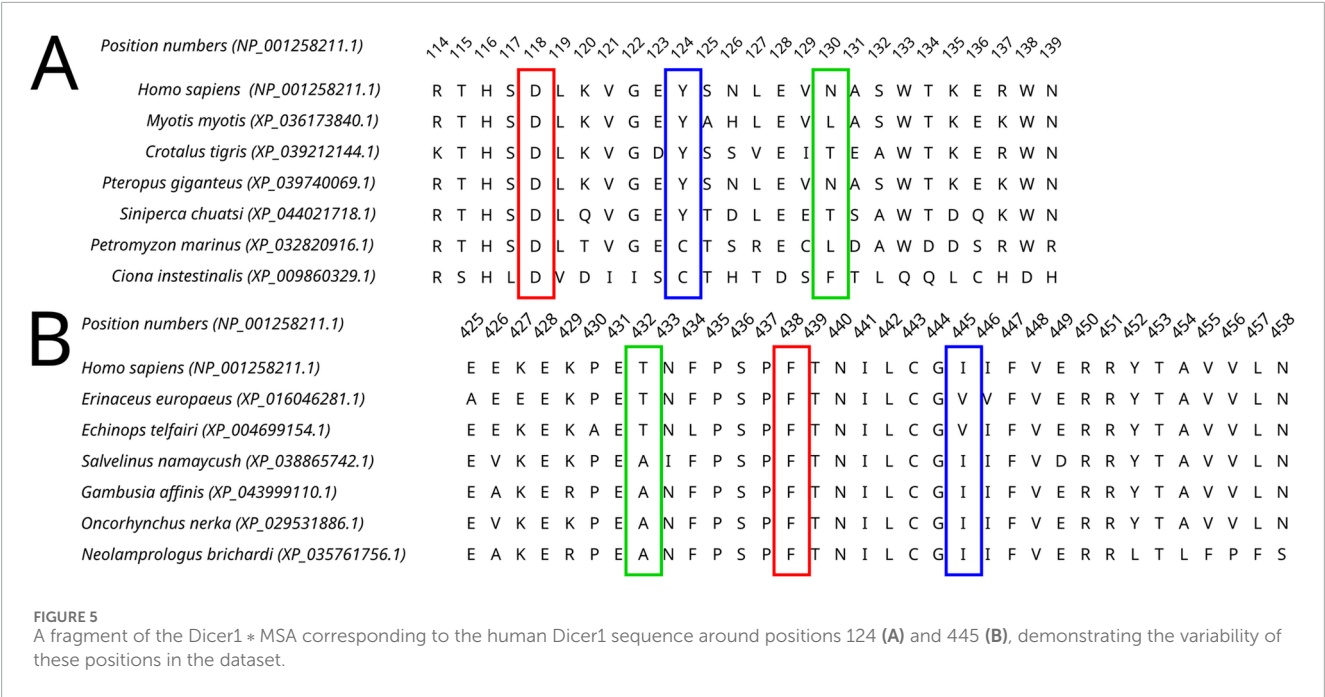
that emerged after the last duplication event, leading to Dicer subfunctionalization (MSA1) or (ii) all identified Dicer orthologs (MSA2) (Figure 2). Two positions, corresponding to Y124 and I445, were variable. In MSA1, a single substitution in position 124 was found—Y124C in the Dicer1 sequence from *Petromyzon marinus* (Figure 5); however, no instances of Y124H were detected in either MSA. Thus, we interpret this variant as evolutionarily intolerable. Similarly, several instances of I445V substitution were detected in MSA1 (Figure 5), but there were no instances of I445S substitutions in either MSA. Consequently, this variant was also considered evolutionarily intolerant. The other two positions, F508 and T993, were invariable; therefore, reported VUSs F508C and T993R are evolutionarily intolerable (Supplementary Table S1).

MD simulations showed relative stability of all four mutated Dicer proteins compared to the wild-type protein (Supplementary Figure S2; Supplementary Table S2). The variants Y124 and I445S did not show significant bond alterations, which was demonstrated by the relative stability of structural elements during MD simulation (Supplementary Figures S3,S4).

F and T are strongly conserved in the 508th and 993rd positions, respectively, by analyzed MSA, and other substitutions are evidently prohibited by evolution (Figure 6). These positions are also invariable in the majority of Dicer1 * sequences, which underscores the importance of their conservation for the functionality of Dicer1 homologs in general. Neither F508C nor T993R is ever seen among Dicer1 homologous sequences, including Dicer2. The detailed damaging effect of these Dicer variants was confirmed by MD simulation. RMSD fluctuations of F508C and T993R are roughly 30% higher than wild-type protein, in particular for T993R, which triggers a more destabilized area; both the F508 and T993R regions are characterized by a significantly increased radius of gyration, indicating the loss of local compactness and more pronounced conformational changes (Figures 7, 8; Supplementary Table S3). Moreover, significant bond alterations were observed for F508 and T993R variants (Supplementary Figure S3). In particular, both induce the loss of five H-bonds within the considered 10 Å region. The spectrum of the most frequent interactions of F508 consists of H-bonds with V504, H511, C443, G444, and L505 that are responsible for α -helix and β -sheet interposition. All these interactions were completely lost for C508, and the set of occurring bonds through the MD run was totally different. The differences led to severe structural changes: the α -helix containing residue 508 was partially disbanded along with loss of interactions with β -sheet; the whole local region was deformed with lower inter-compactness (Figure 7). Similar severe structural changes were characterized for T993R substitution: T993 and R993 have only R944 as a one-H-bond donor in common; therefore, the most frequent and stable interactions of T993 with W1048 and H856 were lost for the R993 mutant. Such a loss of an essential H-bond with W1048 leads to a significant distance increase between the corresponding α -helix and β -sheet, entire region deformation, and destabilization (Figure 8).

4 Discussion

The *DICER1* gene and its mutations draw interest from the carcinogenesis perspective as a crucial and irreplaceable player in



miRNA and the siRNA biogenesis gene, while cancer pathogenesis is widely characterized by the dysfunction of the miRNA spectrum (Vedanayagam et al., 2019; Foulkes et al., 2014). Indeed, both germline and somatic mutations in DICER1 were identified in diverse types of cancer (Hill et al., 2009; Witkowski et al., 2013; Seki et al., 2014; Wu et al., 2018; Chen et al., 2015). We have analyzed DICER1 variants available in the ClinVar database and found that 91% of registered variants are of unknown clinical significance. Among them, only six cancer-associated Dicer1 hotspots have been reported previously (Vedanayagam et al., 2019). In this case, the classification of the majority of DICER1 variants and prediction of their clinical effects would benefit the comprehension of the DICER1 role in tumorigenesis.

We applied widely used bioinformatic tools to evaluate the clinical effects of the mutations (CADD, PROVEAN, SIFT, PolyPhen, SNAP, PhD-SNP, and MAPP): unfortunately, the

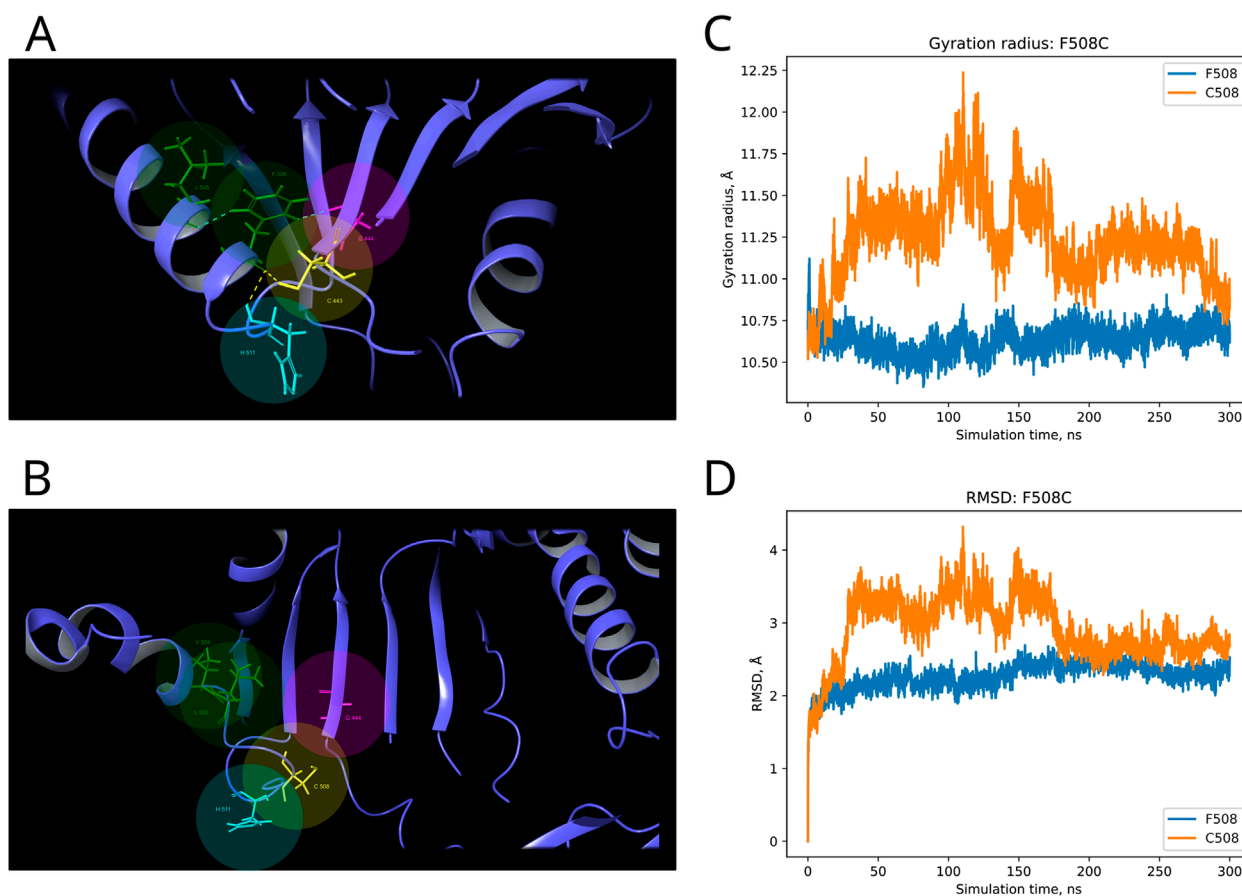


FIGURE 7

Structural alterations of Dicer1 variant F508C. (A) Interactions formed by wild-type amino acid F508. (B) Interactions formed by mutation C508. Amino acids taking part in bond formation are marked by spheres. H-bonds are indicated by dashed yellow lines, and aromatic H-bonds are indicated by dashed blue lines. Protein secondary structural elements (α -helices, β -strands, and disordered loops) are shown in blue by cartoon representation. The radius of gyration (C) and RMSD (D) fluctuations of the 10Å region around the wild-type amino acid and corresponding mutation through a 300-ns MD simulation.

expected accuracy for even well-known *DICER1* hotspot mutations did not exceed 60%–80%. After applying a comparative genomic approach, these tools produced several issues and incorrect predictions, which are basically the result of faulty MSA. Most of the errors occur from the inclusion of low-quality sequences and paralogs in the analytic dataset. Therefore, we advocate for precise and individual dataset construction for each protein of interest based on its evolutionary history and domain architecture. For this purpose, we reconstructed *DICER1* evolution and divided two paralogs, Dicer1 and Dicer2, which, in addition to their sequence homology, are functionally different proteins (Welker et al., 2011; Kolaczowski et al., 2010). Moreover, the last major evolutionary event in the history of *DICER1* homologs was the loss of *DICER2* (Mukherjee et al., 2012), and it is essential to take only Dicer1 sequences from proteomes without Dicer2. We inspected and refined the final MSA for the interpretation of Dicer1 variants. First, the MSA dataset was validated on the well-known protein hotspots. Our “straightforward” prediction approach was based on the total conservation of the position of interest and its neighboring positions corresponding to the

human Dicer1 sequence, which means intolerance for substitution. If the position is changed along with its neighbors, we consider such situations as uncertain because the change of local context could compensate for the impact of the substitution on the functionality of the whole protein and, furthermore, on clinical significance. Thus, our approach allowed us to determine the potential significance of 1,073 variants: among them, 485 were tolerant, and 588 were intolerant. In addition, we thoroughly analyzed those variants whose predictions were not consistent with the automated tools’ predictions. Several pieces of evidence were demonstrated for such conflicting variants (e.g., M1808L), which are close to several well-known “hotspots.” This example clearly explains the issues in MSA of automated programs and consequent false predictions.

Moreover, our obtained MSA was applied for analysis of those *DICER1* variants that occurred in cancer where the role of this gene is of particular interest. Recent studies showed the potential *DICER1* involvement in hematological malignancies (Santamaría et al., 2012; Raaijmakers et al., 2010; Moiseev et al., 2021). Therefore, the variants with unknown significance were

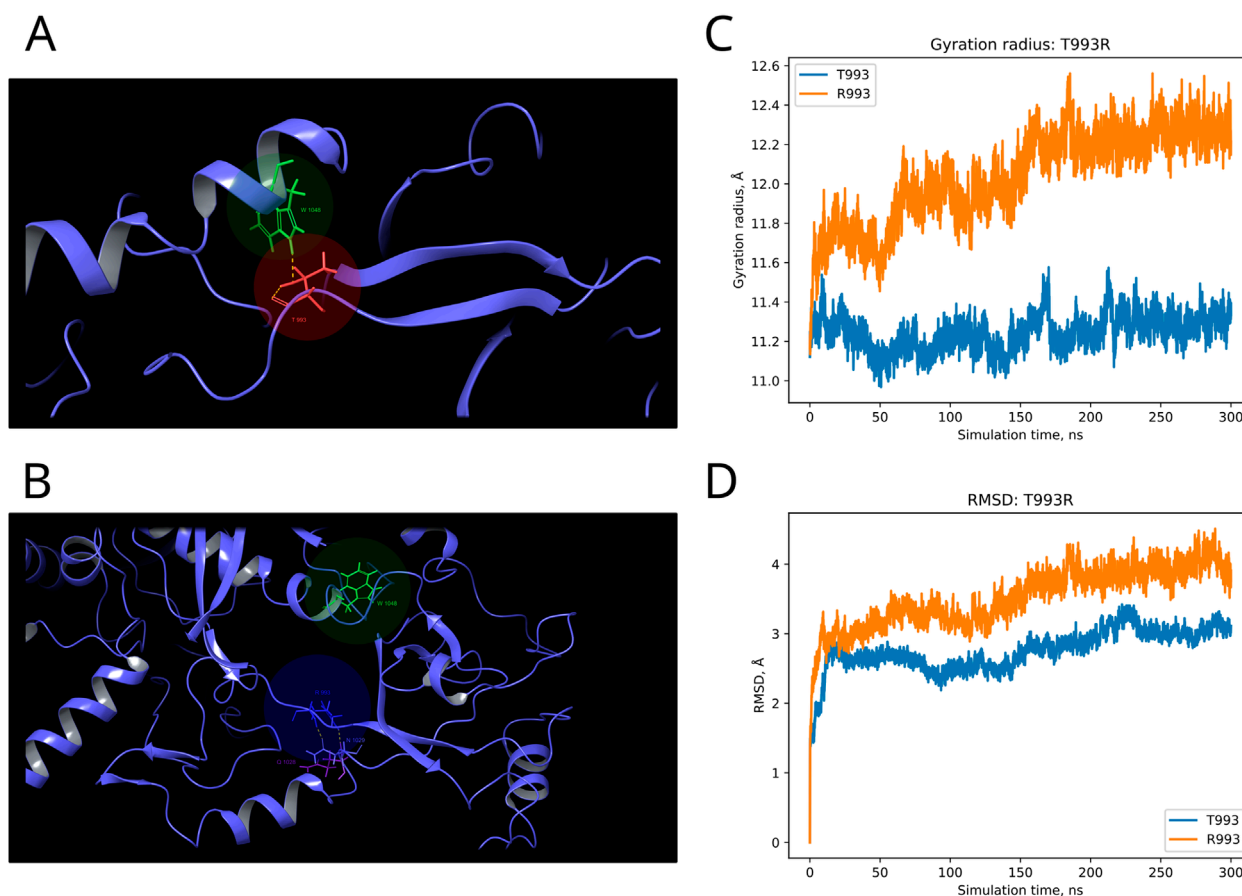


FIGURE 8

Structural alterations of Dicer1 variant T993R. (A) Interactions formed by wild-type amino acid T993. (B) Interactions formed by mutation R993. Amino acids taking part in bond formation are marked by spheres. H-bonds are indicated by dashed yellow lines, and aromatic H-bonds are indicated by dashed blue lines. Protein secondary structural elements (α -helices, β -strands, and disordered loops) are shown in blue by cartoon representation. The radius of gyration (C) and RMSD (D) fluctuations of the 10Å region around the wild-type amino acid and corresponding mutation through a 300-ns MD simulation.

analyzed using our method in order to evaluate the potential effect on cancer progression. Dicer1 missense mutations that occurred in functional domains (Y124H (Helicase ATP-binding), I445S and F508C (Helicase C-terminal), and T993R (PAZ)) were analyzed by MSA. The assessment by comparative genomics was additionally supported by the evaluation of these variants by *in silico* site-specific mutagenesis and molecular dynamics simulation. In particular, the analysis of variants Y124H and I445S (both in the Helicase domain) demonstrated some variability of these protein positions compared to F508C (Helicase C-terminal) and T993R (PAZ), which were strongly conserved, and other substitutions are evidently prohibited by evolution. The results obtained by the MSA analysis were in compliance with those of the molecular dynamics simulation, which showed the structural consequences of mutations: namely, significant structural alterations in the Dicer1 mutated with F508C and T993R substitutions. In these cases, the key interactions were lost, which led to protein local region destabilization. F508C dramatically altered the mutual proximity of secondary structural elements within the C-terminal Helicase domain; T993R disrupted the interactions of the PAZ domain with

both interdomain regions that, in turn, affect PAZ positioning between adjacent domains (Dicer dsRNA-binding fold and RNAase III). All these events are the distinct basis for protein dysfunction and/or dysregulation.

To summarize, in addition to the well-known *DICER1* tumor predisposition syndrome (González et al., 2021), the potential oncogenic role of this gene is being studied and discussed in other malignant diseases (Robertson et al., 2018). Our work was dedicated to investigating and clarifying the effect of the mutational spectrum across the whole protein sequence and marked as uncertain significance on the basis of the combination of in-depth gene evolution reconstruction and molecular modeling of mutational structural–functional consequences. Our analysis revealed the effect of newly occurring “non-hotspot” gene variants accompanying tumorigenesis progression in the example of hematological malignancies. Our study further expands our overall understanding of *DICER1* potential in neoplastic development. In the future, it could be valuable to expand such analysis to other oncology-associated genes and their inconclusive variants to develop the flexible methodology of variant evaluation in order to examine their

potential effect with an appropriate set of instruments that could be adjusted individually for each marker.

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

DB: conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing–original draft, and writing–review and editing. IM: conceptualization, formal analysis, funding acquisition, project administration, resources, writing–original draft, and writing–review and editing. YP: conceptualization, formal analysis, resources, software, writing–original draft, and writing–review and editing. NP: conceptualization, formal analysis, investigation, methodology, supervision, validation, writing–original draft, and writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The following funding is acknowledged: Russian Science Foundation (grant No. 23-15-00327).

References

- Adebali, O., Reznik, A. O., Ory, D. S., and Zhulin, I. B. (2016). Establishing the precise evolutionary history of a gene improves prediction of disease-causing missense mutations. *Genet. Med.* 18, 1029–1036. doi:10.1038/gim.2015.208
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2
- 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. *IEEE Xplore*, 43. doi:10.1109/SC.2006.54
- Caroleo, A. M., De Ioris, M. A., Boccutto, L., Alessi, I., Del Baldo, G., Cacchione, A., et al. (2021). DICER1 syndrome and cancer predisposition: from a rare pediatric tumor to lifetime risk. *Front. Oncol.* 10, 614541. doi:10.3389/fonc.2020.614541
- Chen, J., Wang, Y., McMonechy, M. K., Anglesio, M. S., Yang, W., Senz, J., et al. (2015). Recurrent DICER1 hotspot mutations in endometrial tumours and their impact on microRNA biogenesis. *J. Pathology* 237, 215–225. doi:10.1002/path.4569
- Chen, K. S., Stuart, S. H., Stroup, E. K., Shukla, A. S., Wang, J., Rajaram, V., et al. (2018). Distinct DICER1 hotspot mutations identify bilateral tumors as separate events. *JCO Precis. Oncol.* 2, 1–9. doi:10.1200/po.17.00113
- Fabian, M. R., and Sonenberg, N. (2012). The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat. Struct. Mol. Biol.* 19, 586–593. doi:10.1038/nsmb.2296
- Federici, G., and Soddu, S. (2020). Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *J. Exp. and Clin. Cancer Res.* 39, 46. doi:10.1186/s13046-020-01554-6
- Foulkes, W. D., Priest, J. R., and Duchaine, T. F. (2014). DICER1: mutations, microRNAs and mechanisms. *Nat. Rev. Cancer* 14, 662–672. doi:10.1038/nrc3802
- González, I. A., Stewart, D. R., Schultz, K. A. P., Field, A. P., Hill, D. A., and Dehner, L. P. (2021). DICER1 tumor predisposition syndrome: an evolving story initiated with the pleuropulmonary blastoma. *Mod. Pathol.* 35, 4–22. doi:10.1038/s41379-021-00905-8
- Ha, M., and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* 15, 509–524. doi:10.1038/nrm3838
- Hammond, S. M. (2005). Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett.* 579, 5822–5829. doi:10.1016/j.febslet.2005.08.079
- Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., et al. (2015). OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* 12, 281–296. doi:10.1021/acs.jctc.5b00864
- Hill, D. A., Ivanovich, J., Priest, J. R., Gurnett, C. A., Dehner, L. P., Desruisseau, D., et al. (2009). DICER1 mutations in familial pleuropulmonary blastoma. *Science* 325, 965. doi:10.1126/science.1174334
- Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2017). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi:10.1093/molbev/msx281
- Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., 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
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869
- 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
- Kalyanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi:10.1038/nmeth.4285
- Katoh, K., Misawa, K., Kuma, K. i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gk436

Acknowledgments

We thank Prof. Igor Zhulin (Department of Microbiology, Ohio State University) for the expertise, conceptual guidance, and help in writing and reviewing the manuscript.

Conflict of interest

Author YP was employed by Advitam Laboratory.

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2024.1441180/full#supplementary-material>

- Klein, S., Lee, H., Ghahremani, S., Kempert, P., Ischander, M., Teitell, M. A., et al. (2014). Expanding the phenotype of mutations in DICER1: mosaic missense mutations in the RNase IIIb domain of DICER1 cause GLOW syndrome. *J. Med. Genet.* 51, 294–302. doi:10.1136/jmedgenet-2013-101943
- Kock, L., Wu, M. K., and Foulkes, W. D. (2019). Ten years of DICER1 mutations: provenance, distribution, and associated phenotypes. *Hum. Mutat.* 40, 1939–1953. doi:10.1002/humu.23877
- Kolaczowski, B., Hupalo, D. N., and Kern, A. D. (2010). Recurrent adaptation in RNA interference genes across the Drosophila phylogeny. *Mol. Biol. Evol.* 28, 1033–1042. doi:10.1093/molbev/msq284
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2017). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. doi:10.1093/nar/gkx1153
- Liu, Z., Wang, J., Cheng, H., Ke, X., Sun, L., Zhang, Q. C., et al. (2018). Cryo-EM structure of human dicer and its complexes with a pre-miRNA substrate. *Cell* 173, 1191–1203. doi:10.1016/j.cell.2018.03.080
- Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Computer-Aided Mol. Des.* 27, 221–234. doi:10.1007/s10822-013-9644-8
- Minh, B. Q., Nguyen, M. A. T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195. doi:10.1093/molbev/mst024
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi:10.1093/molbev/msaa015
- Moiseev, I. S., Tcvetkov, N. Y., Barkhatov, I. M., Barabanshikova, M. V., Bug, D. S., Petuhova, N. V., et al. (2021). High mutation burden in the checkpoint and micro-RNA processing genes in myelodysplastic syndrome. *PLOS ONE* 16, e0248430. doi:10.1371/journal.pone.0248430
- Mukherjee, K., Campos, H., and Kolaczowski, B. (2012). Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. *Mol. Biol. Evol.* 30, 627–641. doi:10.1093/molbev/mss263
- Nosé, S. (1984). A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* 81, 511–519. doi:10.1063/1.447334
- Okamura, K., and Lai, E. C. (2008). Endogenous small interfering RNAs in animals. *Nat. Rev. Mol. Cell Biol.* 9, 673–678. doi:10.1038/nrm2479
- O’Leary, N. A., Wright, M. W., Brister, R. B., Cufio, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucl. Acids Res.* 44, D733–D745. doi:10.1093/nar/gkv1189
- Raaijmakers, M. H. G. P., Mukherjee, S., Guo, S., Zhang, S., Kobayashi, T., Schoonmaker, J. A., et al. (2010). Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. *Nature* 464, 852–857. doi:10.1038/nature08851
- Robertson, J. C., Jorczyk, C. L., and Oxford, J. T. (2018). DICER1 syndrome: DICER1 mutations in rare cancers. *Cancers* 10, 143. doi:10.3390/cancers10050143
- Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J. M., Lu, C., et al. (2019). OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* 15, 1863–1874. doi:10.1021/acs.jctc.8b01026
- Santamaría, C., Muntión, S., Rosón, B., Blanco, B., López-Villar, O., Carrancio, S., et al. (2012). Impaired expression of DICER, DROSHA, SBDS and some microRNAs in mesenchymal stromal cells from myelodysplastic syndrome patients. *Haematologica* 97, 1218–1224. doi:10.3324/haematol.2011.054437
- Seki, M., Yoshida, K., Shiraishi, Y., Shimamura, T., Sato, Y., Nishimura, R., et al. (2012). Biallelic DICER1 mutations in sporadic pleuropulmonary blastoma. *Cancer Res.* 74, 2742–2749. doi:10.1158/0008-5472.can-13-2470
- Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., et al. (2013). New and continuing developments at PROSITE. *Nucleic acids Res.* 41, D344–D347. doi:10.1093/nar/gks1067
- Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2018). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi:10.1093/nar/gky1015
- Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* 32, 358–368. doi:10.1002/humu.21445
- Vedanayagam, J., Chatila, W. K., Aksoy, B. A., Majumdar, S., Skanderup, A. J., Demir, E., et al. (2019). Cancer-associated mutations in DICER1 RNase IIIa and IIIb domains exert similar effects on miRNA biogenesis. *Nat. Commun.* 10, 3682. doi:10.1038/s41467-019-11610-1
- Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi:10.1126/science.1235122
- Welker, N., Maity, T. S., Ye, X., Joseph Aruscavage, P., Krauchuk, A. A., Liu, Q., et al. (2011). Dicer’s helicase domain discriminates dsRNA termini to promote an altered reaction mode. *Mol. cell* 41, 589–599. doi:10.1016/j.molcel.2011.02.005
- Wilson, R. C., and Doudna, J. A. (2013). Molecular mechanisms of RNA interference. *Annu. Rev. Biophys.* 42, 217–239. doi:10.1146/annurev-biophys-083012-130404
- Witkowski, L., Mattina, J., Schönberger, S., Murray, M. J., Huntsman, D. G., Reis-Filho, J. S., et al. (2013). DICER1 hotspot mutations in non-epithelial gonadal tumours. *Br. J. Cancer* 109, 2744–2750. doi:10.1038/bjc.2013.637
- Wu, M. K., Vujanic, G. M., Fahiminiya, S., Watanabe, N., Thorner, P. S., O’Sullivan, M. J., et al. (2018). Anaplastic sarcomas of the kidney are characterized by DICER1 mutations. *Mod. Pathol.* 31, 169–178. doi:10.1038/modpathol.2017.100
- Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26, 889–895. doi:10.1093/bioinformatics/btq066
- Yang, J.-S., and Lai, E. C. (2011). Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. *Mol. Cell* 43, 892–903. doi:10.1016/j.molcel.2011.07.024

Frontiers in Molecular Biosciences

Explores biological processes in living organisms
on a molecular scale

Focuses on the molecular mechanisms
underpinning and regulating biological processes
in organisms across all branches of life.

Discover the latest Research Topics

[See more](#) →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact



Frontiers in Molecular Biosciences

