



ADVANCES IN MOLECULAR DOCKING AND STRUCTURE-BASED MODELLING

EDITED BY: Alexandre G. De Brevern, Ramanathan Sowdhamini,
Agnel Praveen Joseph and Joseph Rebehmed
PUBLISHED IN: Frontiers in Molecular Biosciences



frontiers

Frontiers eBook Copyright Statement

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

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

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

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

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

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

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

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

ISSN 1664-8714

ISBN 978-2-88974-509-8

DOI 10.3389/978-2-88974-509-8

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

ADVANCES IN MOLECULAR DOCKING AND STRUCTURE-BASED MODELLING

Topic Editors:

Alexandre G. De Brevern, INSERM U1134 Biologie Intégrée du Globule Rouge, France

Ramanathan Sowdhamini, National Centre for Biological Sciences, India

Agnel Praveen Joseph, Science and Technology Facilities Council, United Kingdom

Joseph Rebehmed, Lebanese American University, Lebanon

Citation: De Brevern, A. G., Sowdhamini, R., Joseph, A. P., Rebehmed, J., eds. (2022). Advances in Molecular Docking and Structure-Based Modelling. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-509-8

Table of Contents

- 05 Editorial: Advances in Molecular Docking and Structure-Based Modelling**
Joseph Rebehmed, Alexandre G. de Brevern, Ramanathan Sowdhamini and Agnel Praveen Joseph
- 08 Evaluation of CONSRANK-Like Scoring Functions for Rescoring Ensembles of Protein–Protein Docking Poses**
Guillaume Launay, Masahito Ohue, Julia Prieto Santero, Yuri Matsuzaki, Cécile Hilpert, Nobuyuki Uchikoga, Takanori Hayashi and Juliette Martin
- 16 Identification of the Interactions Interference Between the PH and START Domain of CERT by Limonoid and HPA Inhibitors**
Mariem Ghoula, Axelle Le Marec, Christophe Magnan, Hervé Le Stunff and Olivier Taboureau
- 30 Analyzing In Silico the Relationship Between the Activation of the Edema Factor and Its Interaction With Calmodulin**
Irène Pitard, Damien Monet, Pierre L. Goossens, Arnaud Blondel and Thérèse E. Malliavin
- 47 Perspectives on High-Throughput Ligand/Protein Docking With Martini MD Simulations**
Paulo C. T. Souza, Vittorio Limongelli, Sangwook Wu, Siewert J. Marrink and Luca Monticelli
- 56 MD Simulations on a Well-Built Docking Model Reveal Fine Mechanical Stability and Force-Dependent Dissociation of Mac-1/GPIIb α Complex**
Xiaoyan Jiang, Xiaoxi Sun, Jiangguo Lin, Yingchen Ling, Ying Fang and Jianhua Wu
- 67 Identification of Potential Binders of Mtb Universal Stress Protein (Rv1636) Through an in silico Approach and Insights Into Compound Selection for Experimental Validation**
Sohini Chakraborti, Moubani Chakraborty, Avipsa Bose, Narayanaswamy Srinivasan and Sandhya S. Visweswariah
- 87 Structure-Guided Computational Approaches to Unravel Druggable Proteomic Landscape of Mycobacterium leprae**
Sundeep Chaitanya Vedithi, Sony Malhotra, Marta Acebrón-García-de-Eulate, Modestas Matusevicius, Pedro Henrique Monteiro Torres and Tom L. Blundell
- 98 Synthesis and Cytotoxic Activity of Novel Indole Derivatives and Their in silico Screening on Spike Glycoprotein of SARS-CoV-2**
Perumal Gobinath, Ponnusamy Packialakshmi, Kaliappillai Vijayakumar, Magda H. Abdellattif, Mohd Shahbaaz, Akbar Idhayadhulla and Radhakrishnan Surendrakumar
- 109 Cryo-EM Map–Based Model Validation Using the False Discovery Rate Approach**
Mateusz Olek and Agnel Praveen Joseph

- 125** *Structural Design and Analysis of the RHOA-ARHGEF1 Binding Mode: Challenges and Applications for Protein-Protein Interface Prediction*
Ennys Gheyouche, Matthias Bagueneau, Gervaise Loirand, Bernard Offmann and Stéphane Téletchéa
- 142** *Active-Site Models of Streptococcus pyogenes Cas9 in DNA Cleavage State*
Honghai Tang, Hui Yuan, Wenhao Du, Gan Li, Dongmei Xue and Qiang Huang
- 154** *Conformational Characterization of the Co-Activator Binding Site Revealed the Mechanism to Achieve the Bioactive State of FXR*
Anita Kumari, Lovika Mittal, Mitul Srivastava, Dharam Pal Pathak and Shailendra Asthana
- 175** *RIGI, TLR7, and TLR3 Genes Were Predicted to Have Immune Response Against Avian Influenza in Indigenous Ducks*
Aruna Pal, Abantika Pal and Pradyumna Baviskar



Editorial: Advances in Molecular Docking and Structure-Based Modelling

Joseph Rebehmed^{1*}, Alexandre G. de Brevern^{2*}, Ramanathan Sowdhamini^{3,4,5*} and Agnel Praveen Joseph^{6*}

¹Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon, ²Université de Paris and Université de la Réunion, INSERM UMR_S 1134, DSIMB, Paris, France, ³National Centre for Biological Sciences (TIFFR), Bangalore, India, ⁴Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India, ⁵Institute of Bioinformatics and Applied Biotechnology, Bengaluru, India, ⁶Scientific Computing, Science and Technology Facilities Council, Research Complex, Rutherford Appleton Laboratory, Didcot, United Kingdom

Keywords: docking, molecular dynamics simulation, structure-function relationship, structure modeling, drug design

Editorial on the Research Topic

Advances in Molecular Docking and Structure-Based Modelling

OPEN ACCESS

Edited and reviewed by:

Francesco Luigi Gervasio,
University College London,
United Kingdom

*Correspondence:

Joseph Rebehmed
joseph.rebehmed@lau.edu.lb
Alexandre G. de Brevern
alexandre.debrevern@u-paris.fr
Ramanathan Sowdhamini
mini@ncbs.res.in
Agnel Praveen Joseph
agnel-praveen.joseph@stfc.ac.uk

Specialty section:

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

Received: 20 December 2021

Accepted: 07 January 2022

Published: 27 January 2022

Citation:

Rebehmed J, de Brevern AG,
Sowdhamini R and Joseph AP (2022)
Editorial: Advances in Molecular
Docking and Structure-
Based Modelling.
Front. Mol. Biosci. 9:839826.
doi: 10.3389/fmolb.2022.839826

The three-dimensional (3D) structures of proteins form the structural framework of their functions. Having access to the structure allows scientists to better apprehend molecular details of protein functions; it is also crucial for protein engineering, e.g., to modify and optimize an enzyme for a certain biochemical reaction; or for designing new and improved drug molecules based on the structure of the target protein. Structures are also needed to investigate how proteins interact; a vast majority of the protein-protein interface residues are involved in extensive intra-protein interactions apart from inter-protein interactions (Jayashree et al., 2019).

With the increase in protein structures available in the Protein DataBank (PDB, 184,929 entries, on the 11th of December 2021, <https://www.rcsb.org>) (Berman et al., 2000), and the recent development of machine learning approaches for protein structure prediction, e.g. AlphaFold2 (Jumper et al., 2021), it is evident that protein structures will be an essential component of a large number of biological research studies. It also highlights the importance of efficient computational tools to process the structures and derive various biological interpretations.

These *in silico* methodologies are often complex, have limitations, and the results must be associated with appropriate statistical and quality measures. The objective of this Research Topic was to bring together various contributions based on cutting-edge computational methodologies; these include computational analysis of structures and complexes with developments and applications that integrate docking and molecular dynamics approaches, and complex experimental data such as cryogenic electron microscopy (cryo-EM).

Two articles underline the importance of new computational approaches for evaluating atomic models derived from experimental data or built *ab-initio*. The first work by Olek and Joseph dealt with the quality of models obtained by cryo-EM. In fact, the final atomic model is often incomplete or contains regions where atomic positions are less reliable or incorrectly built. They presented a software tool for the validation of the backbone trace of atomic models built in the cryo-EM maps. They use the false discovery rate analysis to segregate molecular signals from the background and show how this approach can properly complement current measures Olek and Joseph. Launay

et al. tackled the challenging question of scoring in protein–protein docking. They explored several ways to perform consensus-based rescoring. They showed that rescoring performs worse than the traditional physics-based evaluation but the two complement each other and can be used in combination (Launay et al.).

Classical approaches such as molecular dynamics (MD) are useful to apprehend new biological systems. In this field, Pitard et al. studied the interaction of calmodulin (CaM) with the bacterial virulence factor, Edema Factor (EF). The system is of great interest as orthosteric and allosteric ligands have been proposed to inhibit EF activity. Using state-of-the-art MD simulations, they underlined the presence of cavities at the interface between EF and CaM that could be linked to allosteric events Pitard et al.; Tang et al. combined molecular modelling and MDs to apprehend new mechanistic insights into the exciting CRISPR-Cas9 system in the DNA cleavage state. Their results provide useful guidance for engineering new CRISPR-Cas9 editing systems with improved specificity Tang et al.; Kumari et al. look at the Farnesoid X receptor (FXR) that is essential in regulating the network of genes involved in maintaining bile acid and lipid homeostasis. MDs of FXR with or without cofactors allowed a precise description of critical residue positioning during conformational changes that explain FXR activation state underlying main differences between bound and unbound forms Kumari et al.; Ghoulia et al. analysed the multi domain ceramide transfer protein (CERT) implicated in the transport of ceramide from the endoplasmic reticulum to the Golgi and plays a major role in sphingolipid metabolism. Combining docking and MD simulations, the binding affinity of 14 ligands was tested. This study suggests a novel inhibitory mechanism of CERT for limonoid compounds involving the stabilization of the sub-domain interface and could help in developing new and potentially more selective inhibitors of this transporter Ghoulia et al.

As previously mentioned, experimental 3D structures or structural models are crucial for the design of new drug molecules. Gheyouché et al. applied different approaches to model the structure of RHOA-ARHGEF1 complex and they further analysed the protein–protein interface. They refined the models using MD simulations and highlighted the importance of data-driven human inspection. The modelled RHOA-ARHGEF1 interface will be extremely useful for the design of inhibitors targeting this protein–protein interaction (PPI). Gheyouché et al. Similarly, Pal et al. look at systems of economic interest. They have characterized, using molecular docking, immune response molecules of duck protein TLR3,

TLR7, and RIGI and predicted to have potent antiviral activities against different identified strains of Avian influenza Pal et al.; Gobinath et al. combined experiments and docking approaches in COVID research. They have screened and proposed new indole derivatives on the famous spike glycoprotein of SARS-CoV-2 Gobinath et al.

At a larger-scale, Chakraborti et al. looked at the infectious pathogen with a serious global impact: *Mycobacterium tuberculosis*. There is a constant need to search for novel therapeutic strategies because of the emergence of multidrug-resistant *tuberculosis* (MDR-TB). Universal stress protein (USP, Rv1636) is a perfect target in this field. A library of 1.9 million compounds was subjected to virtual screening, which led to the selection of 2,000 hits through an enrichment process, then 22 potential binders of Rv1636 were prioritized for experimental validations where two compounds of natural origin showed promising binding affinities Chakraborti et al.; Vediti et al. looked at the proteome of *Mycobacterium leprae*. They presented a large set of computational approaches to unravel new potential druggable targets Vediti et al.

Finally, Souza et al. presented innovative approaches to perform high-throughput ligand–protein docking calculations by using coarse-grained molecular dynamics simulations, based on the most recent version of the Martini force field. Their approach, characterized by excellent computational efficiency, offers a level of accuracy comparable to all-atom simulations Souza et al.; Jiang et al. looked at the Interaction of leukocyte integrin macrophage-1 antigen (Mac-1) to platelet glycoprotein Iba (GPIba) implicated in haemostasis and inflammatory responses. They performed a series of “ramp-clamp” steered molecular dynamics (SMD) simulations and compared the results with single molecular AFM measures. The concordance in the results underlined the importance of such approach to understand the platelet–leukocyte interaction in haemostasis and inflammatory responses under mechano- and microenvironments Jiang et al.

This special issue is dedicated to the loving memory of Prof. Narayanaswamy Srinivasan who left us too soon on the third of September 2021 (Eisenhaber et al., 2021). As a passionate scientist and a wonderful human being, he is a true inspiration. May his soul rest in peace.

AUTHOR CONTRIBUTIONS

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

REFERENCES

- 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
- Eisenhaber, F., Verma, C., and Blundell, T. (2021). In Memoriam of Narayanaswamy Srinivasan (1962–2021). *Proteins*. doi:10.1002/prot.26287
- Jayashree, S., Murugavel, P., Sowdhamini, R., and Srinivasan, N. (2019). Interface Residues of Transient Protein-Protein Complexes Have Extensive Intra-protein Interactions Apart from Inter-protein Interactions. *Biol. Direct* 14, 1. doi:10.1186/s13062-019-0232-2
- 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
- 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 Rebehmed, de Brevern, Sowdhamini and Joseph. 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.



Evaluation of CONSRANK-Like Scoring Functions for Rescoring Ensembles of Protein–Protein Docking Poses

Guillaume Launay^{1†}, Masahito Ohue^{2*†}, Julia Prieto Santero¹, Yuri Matsuzaki³, Cécile Hilpert¹, Nobuyuki Uchikoga⁴, Takanori Hayashi² and Juliette Martin^{1*}

¹ CNRS, UMR 5086 Molecular Microbiology and Structural Biochemistry, University of Lyon, Lyon, France, ² Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo, Japan, ³ Tokyo Tech Academy for Leadership, Tokyo Institute of Technology, Tokyo, Japan, ⁴ Department of Network Design, School of Interdisciplinary Mathematical Sciences, Meiji University, Tokyo, Japan

OPEN ACCESS

Edited by:

Ramanathan Sowdhamini,
National Centre for Biological
Sciences, India

Reviewed by:

Stéphane Téletchéa,
Université de Nantes, France
Jeremy Esque,
Institut Biotechnologique de Toulouse
(INSA), France

*Correspondence:

Masahito Ohue
ohue@c.titech.ac.jp
Juliette Martin
juliette.martin@ibcp.fr

[†] 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: 04 May 2020

Accepted: 28 September 2020

Published: 21 October 2020

Citation:

Launay G, Ohue M,
Prieto Santero J, Matsuzaki Y,
Hilpert C, Uchikoga N, Hayashi T and
Martin J (2020) Evaluation
of CONSRANK-Like Scoring
Functions for Rescoring Ensembles
of Protein–Protein Docking Poses.
Front. Mol. Biosci. 7:559005.
doi: 10.3389/fmolb.2020.559005

Scoring is a challenging step in protein–protein docking, where typically thousands of solutions are generated. In this study, we ought to investigate the contribution of consensus-rescoring, as introduced by Oliva et al. (2013) with the CONSRANK method, where the set of solutions is used to build statistics in order to identify recurrent solutions. We explore several ways to perform consensus-based rescoring on the ZDOCK decoy set for Benchmark 4. We show that the information of the interface size is critical for successful rescoring in this context, but that consensus rescoring in itself performs less well than traditional physics-based evaluation. The results of physics-based and consensus-based rescoring are partially overlapping, supporting the use of a combination of these approaches.

Keywords: protein–protein interaction, docking, scoring, prediction, interface

INTRODUCTION

Protein–protein docking aims at predicting the structure of a complex starting from the structures of isolated components (Melquiond et al., 2012; Vakser, 2014). The CAPRI community-wide initiative allows a blind assessment of the participant methods on common data sets and evaluation criteria, offering an updated view of progress in the field since 2001 (Lensink et al., 2007, 2017; Lensink and Wodak, 2010). Protein–protein docking methods typically generate thousands of potential solutions for a particular complex. Scoring the models to discriminate near-native solutions is a known bottleneck of docking methods (Moal et al., 2013a,b; Malhotra et al., 2015). Most scoring functions are physics-based, attempting to capture the determinants underlying the stability of protein–protein complexes, e.g., shape complementary, electrostatics and desolvation potential (Dominguez et al., 2003; Cheng et al., 2007; Pierce and Weng, 2007, 2008; Moal and Bates, 2010; Ritchie and Venkatraman, 2010; Ohue et al., 2014). Knowledge-based functions, on the other hand, aim at taking advantage of the information from available structures, *via* pair potentials (Lu et al., 2003; Huang and Zou, 2008; Mezei, 2017), or multibody potentials (Khashan et al., 2012). Docking methods often use scoring functions that combine physical terms with knowledge-based

terms (Kozakov et al., 2006; Liang et al., 2009; Feliu et al., 2011; Vreven et al., 2011). More recently, evolutionary information has been successfully used for scoring (Andreani et al., 2013; Yu et al., 2017).

Another approach consists in relying on the recurrences observed in the set of solutions, i.e., consensus-based scoring. Consensus-based scoring functions seek to identify solutions with features that are the most frequent in the solution set, independently of any physics-based or evolutionary evaluation. The CONSRANK scoring function, proposed by Oliva et al. (2013, 2015), Vangone et al. (2013), and Chermak et al. (2015, 2016) has shown very good results, based on the conservation of interface contacts.

In this study, we compare several CONSRANK-like scoring functions on large sets of docking poses generated by ZDOCK, including CONSRANK. We then explore how to combine consensus-based rescoring with the native scoring function of ZDOCK.

METHODS

Docking Decoy Set

The ZDOCK3.0.2 decoy set (Pierce et al., 2011) (6 degree sampling, fixed receptor format) for Benchmark4 (Hwang et al., 2010a) was retrieved from <https://zlab.umassmed.edu/zdock/decoys.shtml>. This data set encompasses 176 protein–protein complexes, with 54,000 docking poses for each complex. For each pose, the interface C α RMSD, with respect to the bound structure, is given. A near-native docking hit is defined as a prediction with interface C α RMSD < 2.5 Å.

Consensus-Based Rescoring Schemes

Following the CONSRANK method (Oliva et al., 2013; Chermak et al., 2015), docking poses are rescored using the frequencies of interface contacts in the set of docking poses. Interface contacts are defined using a distance cut-off of 5 Å between the heavy atoms of receptor and ligand proteins.

For each contact C_{ij} between residue i from receptor and residue j from ligand the relative frequency in the decoy set is defined by:

$$S(C_{ij}) = \frac{F(C_{ij})}{N} \in [0, 1] \quad (1)$$

where $F(C_{ij})$ denotes the frequency of C_{ij} at the protein–protein interface in the set of N decoys. These relative frequencies are then averaged, i.e., normalized by the interface size, to compute the CONSRANK score of each pose P :

$$\text{CONSRANK_score}(P) = \frac{\sum_{C_{ij} \in P} S(C_{ij})}{N_{\text{cont}}(P)}, \quad (2)$$

where $N_{\text{cont}}(P)$ denotes the number of interface contacts in docking pose P .

Variations of CONSRANK Scores

First, we considered the un-normalized version of CONSRANK scores (Oliva et al., 2013), denoted as CONSRANK_U, where the

relative frequencies of interface contacts are only summed, and not averaged:

$$\text{CONSRANK_U}(P) = \sum_{C_{ij} \in P} S(C_{ij}). \quad (3)$$

Then, we implemented two other variations, by replacing relative frequencies of contacts $S(C_{ij})$ by relative frequencies of residues:

$$S(R_i) = \frac{F(R_i)}{N} \in [0, 1] \quad (4)$$

where $F(R_i)$ denotes the frequency of residue i at the protein–protein interface (distance between heavy atoms lower than 5 Å) in the set of N decoys. The two related scores are respectively defined by:

$$\text{Residue_Average}(P) = \frac{\sum_{R_i \in P} S(R_i)}{N_{\text{res}}(P)}, \quad (5)$$

$$\text{Residue_Sum}(P) = \sum_{R_i \in P} S(R_i). \quad (6)$$

where $N_{\text{res}}(P)$ denotes the number of interface residues in pose P . Here, interface residues are simply those involved in contacts at the interface.

Note that it is possible to compute the contact and residue frequencies (Eqs 1 and 4) on a given set of docking poses and then to evaluate another set of docking poses (with Eqs 2, 3, 5, and 6).

Clustering

We implemented the BSAS clustering procedure (Basic Sequential Algorithmic Scheme) (Koutroumbas and Theodoridis, 2008; Jiménez-García et al., 2018) to reduce the structural redundancy of docking poses. The principle of BSAS is the following. Docking poses are ranked according to a score in decreasing order. The pose with the highest score initiates the first clusters. The other poses are sequentially compared to already clustered poses: they are included in a cluster if they are within a given cut-off of cluster members, otherwise they initiate a new cluster. At the end of the process, the pose with the highest score in each cluster is the representative of each cluster. In order to allow a fast clustering process, we do not compute the RMSD between ligand atoms. Instead we use a distance cut-off between the centers of mass of the ligands, here set to 8 Å.

Evaluation

The top 2,000 solutions according to the ZDOCK native scoring function were rescored using the rescoring schemes detailed below. We monitored the presence of near-native docking hits (interface C α RMSD < 2.5 Å) in the top 10 solutions after re-ranking. Each protein–protein complex with a near-native docking hit in the top 10 solutions is counted as a success.

Implementation

The consensus-based rescoring functions are implemented in python code accessible on GitHub, which operates directly on

ZDOCK output files, and allows to treat rapidly thousands of docking poses (typically a few seconds for 2,000 poses, up to 1 min for 54,000 poses). In addition, the code allows to compute statistics on a given set of poses and re-score another of structures (see “Results” section). All the scripts necessary to reproduce the results shown in this article are available at: <https://github.com/MMSB-MOBI/CHOKO>.

RESULTS

In this study, we compare four consensus-based scores to identify the near-native solutions among the ensembles generated by ZDOCK. The traditional CONSRANK score (Oliva et al., 2013) is considered, as well as its un-normalized version, and two variations that consider residue statistics instead of contact statistics. We first evaluate each consensus score separately. Then, we combine these results with the native ZDOCK physics-based scoring function. Finally, we add a clustering step, to reduce structural redundancy and further improve the results.

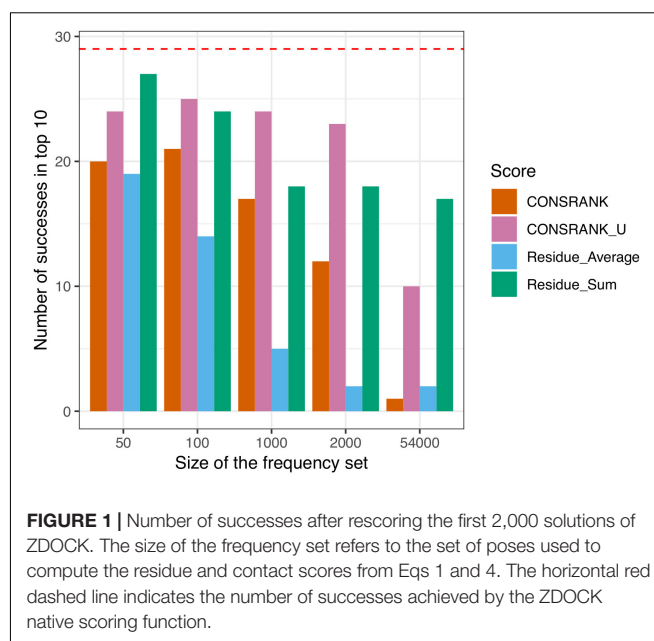
Quality of Decoys

In the initial data set of 176 protein–protein complexes, ZDOCK was able to generate at least one near-native docking hit (interface C α RMSD < 2.5 Å) in the first top 2,000 solutions for 90 protein–protein complexes. These 90 protein–protein complexes thus constitute our reference data set for the rest of the study. We explore if and how consensus-based rescoring is efficient at scoring the decoys of these 90 protein–protein complexes.

Evaluation of Different Consensus-Based Rescoring Functions

First, we compare the four versions of consensus-based rescoring functions: either contact-based [following the CONSRANK (Oliva et al., 2013) scheme] or residue-based, with or without interface size normalization. We estimate the performance by counting the number of successes, i.e., number of complexes with at least one near-native hit (interface C α RMSD < 2.5 Å) in the first 10 solutions after rescoring. We also tested the effect of varying the subset of docking poses used to compute the contact and residue frequencies (Eqs 1 and 4): we used either the first 50, 100, 1,000 or 2,000 first poses provided by ZDOCK, or the full set of 54,000 poses, referred as the frequency set. In any case, we rescored the first 2,000 poses provided by ZDOCK.

The results of this evaluation are shown in **Figure 1**. We can see that un-normalized rescoring functions (CONSRANK_U and Residue_Sum) constantly outperform the normalized rescoring functions (CONSRANK, Residue_Average). The size of the subset used to compute contact and residue frequencies (Eqs 1 and 4) has a major influence on the number of successes. Indeed, ZDOCK solutions are ranked by the ZDOCK native scoring function; hence the top of the list is, in many cases, enriched in near-native docking hits. Estimating contact and residue scores on a reduced subset of poses at the top of the list is logically more efficient. On the contrary, estimating contact and residue scores from the full list leads to a loss of information, and worsens the prediction. When frequencies were



estimated on the 54,000 poses, the number of successes was 1 for CONSRANK, 10 for CONSRANK_U, 2 for Residue_Average, and 17 for Residue_Sum. In the best settings tested here, estimating the scores on the first 50 solutions to rescore the first 2,000 solutions allows to reach a number of successes equal to 27 with the Residue_Sum scoring function, *versus* 20 for the CONSRANK scheme. It is thus possible to rescore large sets of docking poses using consensus-based scoring functions, with better performance than the commonly used CONSRANK scheme.

Combination With ZDOCK Native Scoring Function

In this section, we explore how to combine rescoring functions with the native scoring function of ZDOCK. Out of the 90 protein–protein complexes with at least one near-native docking hit in the top 2,000 solutions, the ZDOCK native scoring function identifies 29 successes, i.e., 29 complexes with at least one near-native docking hit in the top 10, see **Figure 1**. This is indeed better than the four consensus-based rescoring functions tested here. One could wonder if it is then possible to improve the initial prediction of ZDOCK using rescoring.

We first analyzed the overlap of the successful cases by ZDOCK and each rescoring function, and found that many successful cases achieved by rescoring are well predicted by the ZDOCK scoring function, see **Supplementary Figure S1**. For example, when estimating scores on the first 50 solutions for the rescoring (**Supplementary Figure S1A**), all the successes identified by the normalized rescoring functions CONSRANK and Residue_Average are included in the successes identified by ZDOCK. Un-normalized scoring functions are able to identify 1 case not included in the ZDOCK successes for CONSRANK_U, and 4 for Residue_Sum.

We then tested a combination of ZDOCK poses and rescored poses by combining the first N_1 poses of ZDOCK with the first N_2 poses after rescoring, with $N_1 + N_2 = 10$, and no redundancy. Again, we vary the subset of docking poses used to compute the contact and residue frequencies with Eqs 1 and 4 (frequency set = top 50, 100, 1,000 or 2,000 poses) and in any case, we rescore the first 2,000 solutions provided by ZDOCK using Eqs 2, 3, 5, and 6. We estimate the performance by counting the number of successes, i.e., number of complexes with at least one near-native docking hit (interface C α RMSD < 2.5 Å) in the first 10 solutions.

The results of this evaluation are shown in **Supplementary Figure S2**. Regardless of the size of the frequency set, the best combination is always obtained with the Residue_Sum scoring function. Combining the first six ZDOCK poses with the first four Residue_Sum rescored poses, and estimating the frequencies on the full set of 2,000 poses (bottom right panel in **Supplementary Figure S2**) allows to reach a number of successes equal to 32, compared to 18 with Residue_Sum alone and 29 with ZDOCK alone. This suggests the possibility to marginally improve the native results of ZDOCK by a simple combination of poses. It is interesting to note that, in this situation, the information about residues is more efficient in rescoring than the information about pairwise contacts.

Combining Clusters

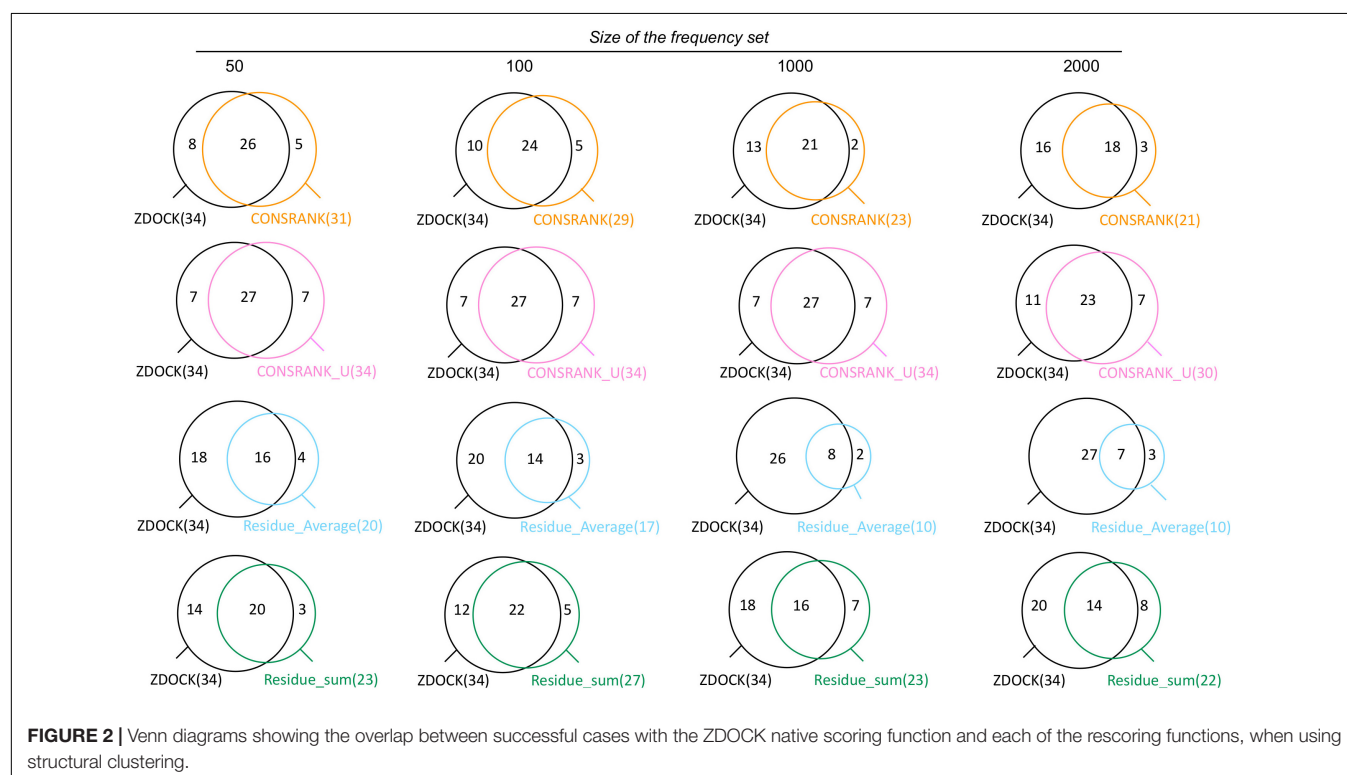
Clustering is classically used to improve the performance, by reducing the structural redundancy of docking solutions (Kozakov et al., 2005; Hwang et al., 2010b; Koukos et al., 2020). Here, we used the BSAS clustering algorithm, which takes into account the scores, to cluster poses by their ligand center of mass.

When applied to ZDOCK results, independently of rescoring, we obtained an improvement in terms of number of successes: 34 successes instead of 29, reflecting a structural redundancy of the ZDOCK set. We first analyzed the overlap between ZDOCK results and the results of each rescoring function when using structural clustering, see **Figure 2**.

As shown in **Figure 2**, the results of ZDOCK and rescoring functions are partially overlapping also after structural clustering. The rescoring functions are able to identify between 2 and 8 additional successful cases, with more additional cases brought by un-normalized scoring functions CONSRANK_U and Residue_Sum. This means that a perfect combination of ZDOCK and rescoring with no loss would reach a number of successes equal to 42.

We then explore how to combine ZDOCK results and rescoring results. We have tested a combination of clusters. On the one hand, we computed clusters from the poses ranked by their initial ZDOCK scores. On the other hand, we computed clusters from poses reordered after consensus rescoring. We then combine the representative poses of the first N_1 ZDOCK clusters, with the representative poses of the first N_2 poses after rescoring, with $N_1 + N_2 = 10$. We estimate the performance by counting the number of successes, i.e., number of complexes with at least one near-native hit (interface C α RMSD < 2.5 Å) in the first 10 solutions.

The results of this evaluation are shown in **Figure 3**. In agreement with the Venn diagram analysis, the best combination is obtained using an un-normalized rescoring function, CONSRANK_U. It constantly outperforms CONSRANK, regardless of the frequency set. When using the first 1,000



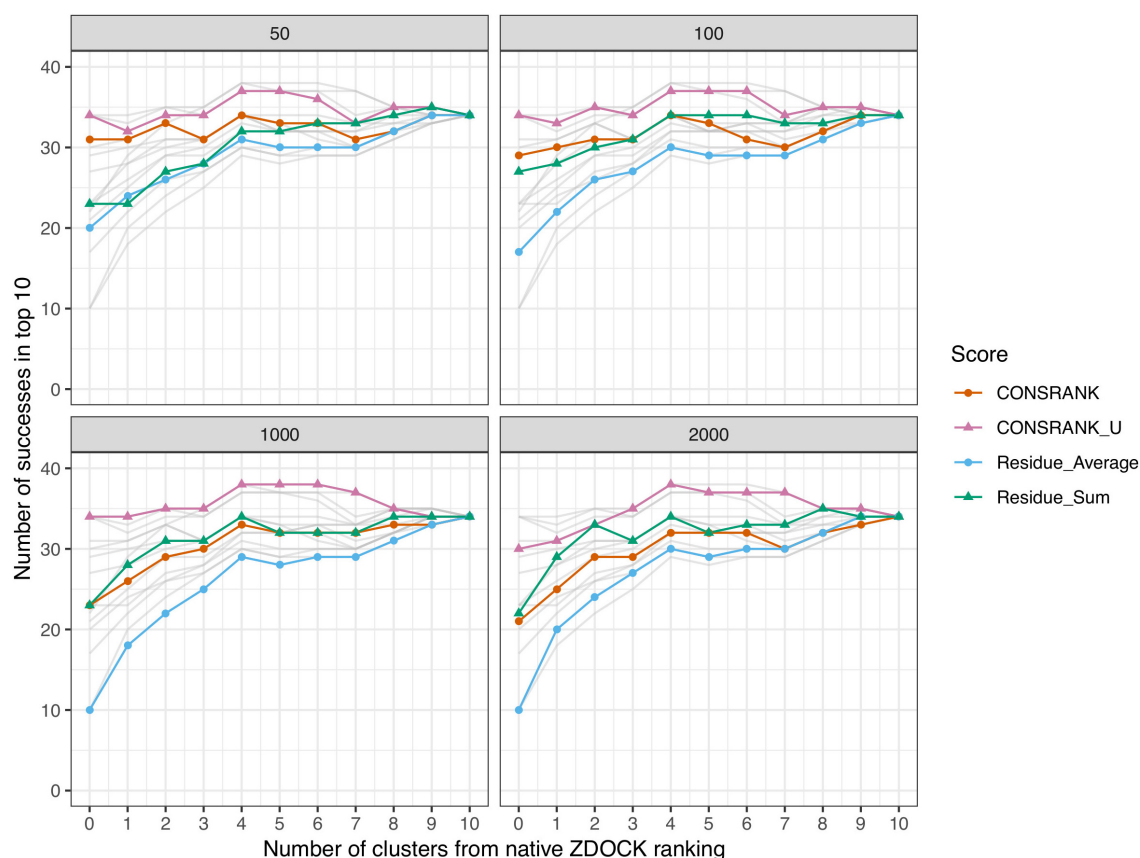


FIGURE 3 | Number of successes after combination of clusters with the ZDOCK native scoring function. Each panel corresponds to different frequency sets, i.e., sets of poses used to compute the residue and contact scores from Eqs 1 and 4. In any cases, the first 2,000 solutions of ZDOCK are rescored. Gray lines represent data from other panels for comparison.

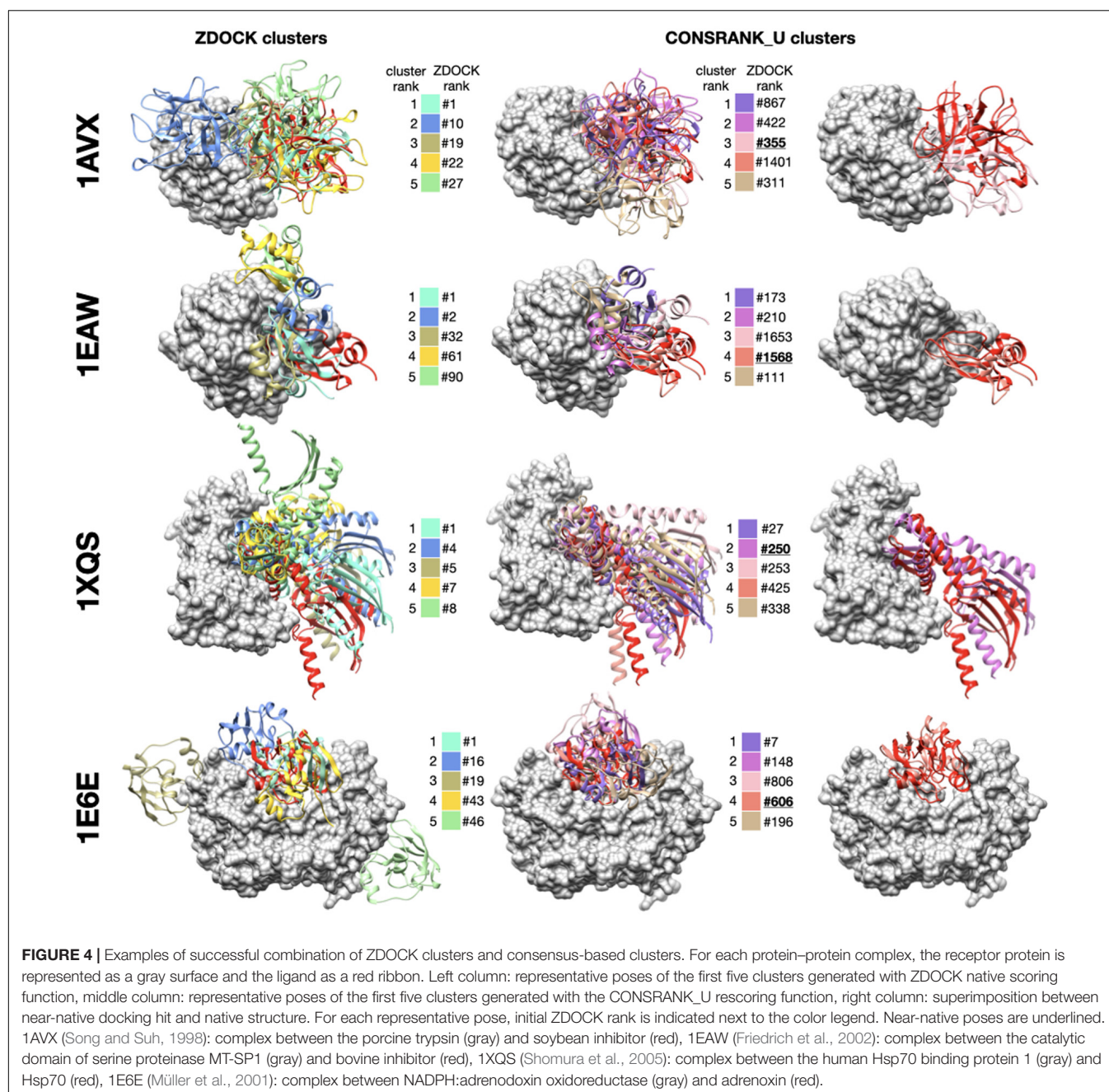
poses to estimate the frequencies (bottom left panel in **Figure 3**), the combination of five ZDOCK clusters and five CONSRANK_U clusters achieves a number of successes equal to 38, compared to 34 with ZDOCK alone and 34 with CONSRANK_U alone. This result suggests that CONSRANK-like rescoring could be used together with physics-based evaluation. Contrary to what was observed in simple pose combination (**Figure 2**), the most efficient rescoring scheme when dealing with clusters is based on contact frequencies, not residue frequencies. It seems that, after structural clustering, the information of pairwise contacts, which is more precise than residues, becomes more useful in discrimination.

Illustrative Examples

To complete this study, in this section, we present examples to illustrate the asset of consensus rescoring when used in combination with the native ZDOCK scoring function. We used the results generated using the CONSRANK_U function, with frequencies estimated on the first 1,000 poses, and combined five ZDOCK clusters and five consensus clusters. As explained in the previous section, this setting allows to reach 38 successes. We present four examples from the ZDOCK decoy set where

the use of consensus rescoring is critical in **Figure 4**. For all these protein-protein complexes, no near-native docking hit is observed in the first 10 clusters of ZDOCK (or in the first 10 poses). The use of CONSRANK_U rescoring in conjunction with clustering allows the identification of near-native docking poses in the top 10. In every case, these near-native poses do not belong to the top of the ZDOCK initial list: they are ranked 355 for 1AVX, 1568 for 1EAW, 250 for 1XQS, and 606 for 1E6E. These examples highlight the usefulness of consensus-based rescoring to rescue poses with poor initial ranks.

For the 38 successful complexes in this experiment, we systematically computed the number of near-native poses coming from ZDOCK clusters and the number of near-native poses coming from CONSRANK_U clusters. Detailed results are provided in **Supplementary Table S1** for the 38 complexes with at least one near-native pose in the top ten. In eight cases, the near-native poses were present only in ZDOCK clusters, in 10 cases, the near-native poses were present only in CONSRANK_U clusters and in the 20 remaining cases, near-native poses were present in both ZDOCK and CONSRANK_U clusters. We observed no significant bias in terms of functional category or interface size between



protein complexes that were successful only with ZDOCK or only with CONSRANK_U. This indicates that ZDOCK and CONSRANK_U results are only partially overlapping, justifying the need to combine them.

CONCLUSION

We have implemented four variants of consensus-based rescoring functions: the CONSRANK score, the CONSRANK un-normalized score, and their equivalents based on residue frequencies and tested them on the rescoring of large sets of

docking poses of the ZDOCK benchmark. In this context, un-normalized scores that do take into account the size of the interfaces are in general more efficient than normalized scores. When used alone, consensus-based scoring functions degraded the initial performance of the physics-based ZDOCK scoring function. However, when both physics-based and consensus-based scoring functions were used in combination, we observed a marginal improvement. This calls for calibration when using consensus-based scoring functions to re-rank large sets of docking decoys, since they are, by definition, highly dependent on the docking decoy population.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

GL, JS, CH, and JM contributed to software. GL, YM, and NU contributed to investigation. MO and JM contributed to the methodology. TH contributed to the data curation. JM conceptualized and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by the “PHC Sakura” program, implemented by the French Ministry of Foreign Affairs, the French Ministry of Higher Education and Research and the Japan Society for Promotion of Science. This project has received funding from the European Union’s Horizon

2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at bioRxiv (Launay G, Ohue M, Prieto Santero J, Matsuzaki Y, Hilpert C, Uchikoga N, et al. Rescoring ensembles of protein–protein docking poses using consensus approaches. bioRxiv. 2020; 2020.04.24.059469. doi: 10.1101/2020.04.24.059469) (Launay et al., 2020). We would like to thank Robinson Martin Minard for his support during this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2020.559005/full#supplementary-material>

REFERENCES

- Andreani, J., Faure, G., and Guerois, R. (2013). InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics* 29, 1742–1749. doi: 10.1093/bioinformatics/btt260
- Cheng, T. M.-K., Blundell, T. L., and Fernandez-Recio, J. (2007). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins* 68, 503–515. doi: 10.1002/prot.21419
- Chermak, E., Donato, R. D., Lensink, M. F., Petta, A., Serra, L., Scarano, V., et al. (2016). Introducing a clustering step in a consensus approach for the scoring of protein–protein docking models. *PLoS One* 11:e0166460. doi: 10.1371/journal.pone.0166460
- Chermak, E., Petta, A., Serra, L., Vangone, A., Scarano, V., Cavallo, L., et al. (2015). CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts. *Bioinformatics* 31, 1481–1483. doi: 10.1093/bioinformatics/btu837
- Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125, 1731–1737. doi: 10.1021/ja026939x
- Feliu, E., Aloy, P., and Oliva, B. (2011). On the analysis of protein–protein interactions via knowledge-based potentials for the prediction of protein–protein docking. *Protein Sci.* 20, 529–541. doi: 10.1002/pro.585
- Friedrich, R., Fuentes-Prior, P., Ong, E., Coombs, G., Hunter, M., Oehler, R., et al. (2002). Catalytic domain structures of MT-SP1/Matriptase, a matrix-degrading transmembrane serine proteinase. *J. Biol. Chem.* 277, 2160–2168. doi: 10.1074/jbc.m109830200
- Huang, S.-Y., and Zou, X. (2008). An iterative knowledge-based scoring function for protein–protein recognition. *Proteins* 72, 557–579. doi: 10.1002/prot.21949
- Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010a). Protein–protein docking benchmark version 4.0. *Proteins* 78, 3111–3114. doi: 10.1002/prot.22830
- Hwang, H., Vreven, T., Pierce, B. G., Hung, J.-H., and Weng, Z. (2010b). Performance of ZDOCK and ZRANK in CAPRI rounds 13–19. *Proteins* 78, 3104–3110. doi: 10.1002/prot.22764
- Jiménez-García, B., Roel-Touris, J., Romero-Durana, M., Vidal, M., Jiménez-González, D., and Fernández-Recio, J. (2018). LightDock: a new multi-scale approach to protein–protein docking. *Bioinformatics* 34, 49–55. doi: 10.1093/bioinformatics/btx555
- Khashan, R., Zheng, W., and Tropsha, A. (2012). Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins* 80, 2207–2217. doi: 10.1002/prot.24110
- Koukos, P. I., Roel-Touris, J., Ambrosetti, F., Geng, C., Schaarschmidt, J., Trellet, M. E., et al. (2020). An overview of data-driven HADDOCK strategies in CAPRI rounds 38–45. *Proteins* 88, 1029–1036. doi: 10.1002/prot.25869
- Koutroumbas, K., and Theodoridis, S. (2008). *Pattern Recognition*. Amsterdam: Academic Press.
- Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* 65, 392–406. doi: 10.1002/prot.21117
- Kozakov, D., Clodfelter, K. H., Vajda, S., and Camacho, C. J. (2005). Optimal clustering for detecting near-native conformations in protein docking. *Biophys. J.* 89, 867–875. doi: 10.1529/biophysj.104.058768
- Launay, G., Ohue, M., Santero, J. P., Matsuzaki, Y., Hilpert, C., Uchikoga, N., et al. (2020). Rescoring ensembles of protein–protein docking poses using consensus approaches. *bioRxiv[Preprint]*, 2020.04.24.059469. doi: 10.1101/2020.04.24.059469
- Lensink, M. F., Méndez, R., and Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* 69, 704–718. doi: 10.1002/prot.21804
- Lensink, M. F., Velankar, S., and Wodak, S. J. (2017). Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins* 85, 359–377. doi: 10.1002/prot.25215
- Lensink, M. F., and Wodak, S. J. (2010). Docking and scoring protein interactions: CAPRI 2009. *Proteins* 78, 3073–3084. doi: 10.1002/prot.22818
- Liang, S., Meroueh, S. O., Wang, G., Qiu, C., and Zhou, Y. (2009). Consensus scoring for enriching near-native structures from protein–protein docking decoys. *Proteins* 75, 397–403. doi: 10.1002/prot.22252
- Lu, H., Lu, L., and Skolnick, J. (2003). Development of unified statistical potentials describing protein–protein interactions. *Biophys. J.* 84, 1895–1901. doi: 10.1016/s0006-3495(03)74997-2
- Malhotra, S., Mathew, O. K., and Sowdhamini, R. (2015). DOCKSCORE: a webserver for ranking protein–protein docked poses. *BMC Bioinform.* 16:127. doi: 10.1186/s12859-015-0572-6
- Melquiond, A. S. J., Karaca, E., Kastiris, P. L., and Bonvin, A. M. J. J. (2012). Next challenges in protein–protein docking: from proteome to interactome and beyond. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2, 642–651. doi: 10.1002/wcms.91

- Mezei, M. (2017). Rescore protein-protein docked ensembles with an interface contact statistics. *Proteins* 85, 235–241. doi: 10.1002/prot.25209
- Moal, I. H., and Bates, P. A. (2010). SwarmDock and the use of normal modes in protein-protein docking. *Int. J. Mol. Sci.* 11, 3623–3648. doi: 10.3390/ijms11103623
- Moal, I. H., Moretti, R., Baker, D., and Fernández-Recio, J. (2013a). Scoring functions for protein-protein interactions. *Curr. Opin. Struct. Biol.* 23, 862–867. doi: 10.1016/j.sbi.2013.06.017
- Moal, I. H., Torchala, M., Bates, P. A., and Fernández-Recio, J. (2013b). The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinform.* 14:286. doi: 10.1186/1471-2105-14-286
- Müller, J. J., Lapko, A., Bourenkov, G., Ruckpaul, K., and Heinemann, U. (2001). Adrenodoxin reductase-adrenodoxin complex structure suggests electron transfer path in steroid biosynthesis. *J. Biol. Chem.* 276, 2786–2789. doi: 10.1074/jbc.m008501200
- Ohue, M., Shimoda, T., Suzuki, S., Matsuzaki, Y., Ishida, T., and Akiyama, Y. (2014). MEGADOCK 4.0: an ultra-high-performance protein-protein docking software for heterogeneous supercomputers. *Bioinformatics* 30, 3281–3283. doi: 10.1093/bioinformatics/btu532
- Oliva, R., Chermak, E., and Cavallo, L. (2015). Analysis and ranking of protein-protein docking models using inter-residue contacts and inter-molecular contact maps. *Molecules* 20, 12045–12060. doi: 10.3390/molecules200712045
- Oliva, R., Vangone, A., and Cavallo, L. (2013). Ranking multiple docking solutions based on the conservation of inter-residue contacts. *Proteins* 81, 1571–1584. doi: 10.1002/prot.24314
- Pierce, B., and Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins* 67, 1078–1086. doi: 10.1002/prot.21373
- Pierce, B., and Weng, Z. (2008). A combination of rescoring and refinement significantly improves protein docking performance. *Proteins* 72, 270–279. doi: 10.1002/prot.21920
- Pierce, B. G., Hourai, Y., and Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One* 6:e24657. doi: 10.1371/journal.pone.0024657
- Ritchie, D. W., and Venkatraman, V. (2010). Ultra-fast FFT protein docking on graphics processors. *Bioinformatics* 26, 2398–2405. doi: 10.1093/bioinformatics/btq444
- Shomura, Y., Dragovic, Z., Chang, H.-C., Tzvetkov, N., Young, J. C., Brodsky, J. L., et al. (2005). Regulation of Hsp70 function by HspBP1: structural analysis reveals an alternate mechanism for Hsp70 nucleotide exchange. *Mol. Cell* 17, 367–379. doi: 10.1016/s1097-2765(05)01010-5
- Song, H. K., and Suh, S. W. (1998). Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from *Erythrina caffra* and tissue-type plasminogen activator11Edited by R. Huber. *J. Mol. Biol.* 275, 347–363. doi: 10.1006/jmbi.1997.1469
- Vakser, I. A. (2014). Protein-protein docking: from interaction to interactome. *Biophys. J.* 107, 1785–1793. doi: 10.1016/j.bpj.2014.08.033
- Vangone, A., Cavallo, L., and Oliva, R. (2013). Using a consensus approach based on the conservation of inter-residue contacts to rank CAPRI models. *Proteins* 81, 2210–2220. doi: 10.1002/prot.24423
- Vreven, T., Hwang, H., and Weng, Z. (2011). Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci.* 20, 1576–1586. doi: 10.1002/pro.687
- Yu, J., Andreani, J., Ochsenbein, F., and Guerois, R. (2017). Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI Rounds 28–35. *Proteins* 85, 378–390. doi: 10.1002/prot.25180

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

Copyright © 2020 Launay, Ohue, Prieto Santero, Matsuzaki, Hilpert, Uchikoga, Hayashi and Martin. 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 the Interactions Interference Between the PH and START Domain of CERT by Limonoid and HPA Inhibitors

Mariam Ghoula¹, Axelle Le Marec², Christophe Magnan², Hervé Le Stunff^{3*} and Olivier Taboureau^{1*}

¹ Université de Paris, INSERM U1133, CNRS UMR 8251, Paris, France, ² Université de Paris, BFA CNRS UMR 8251, Paris, France, ³ Université Paris Saclay, Institut des Neurosciences Paris Saclay, CNRS UMR 9197, Orsay, France

OPEN ACCESS

Edited by:

Ramanathan Sowdhamini,
National Centre for Biological
Sciences, India

Reviewed by:

Natalia Kulik,
Academy of Sciences of the Czech
Republic, Czechia
Zhili Zuo,
Chinese Academy of Sciences, China

*Correspondence:

Hervé Le Stunff
herve.le-stunff@universite-paris-
saclay.fr
Olivier Taboureau
olivier.taboureau@u-paris.fr;
olivier.taboureau@univ-paris-diderot.fr

Specialty section:

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

Received: 08 September 2020

Accepted: 04 November 2020

Published: 27 November 2020

Citation:

Ghoula M, Le Marec A,
Magnan C, Le Stunff H and
Taboureau O (2020) Identification
of the Interactions Interference
Between the PH and START Domain
of CERT by Limonoid and HPA
Inhibitors.
Front. Mol. Biosci. 7:603983.
doi: 10.3389/fmolb.2020.603983

The multi domain ceramide transfer protein (CERT) which contains the domains START and PH, is a protein that allows the transport of ceramide from the endoplasmic reticulum to the Golgi and so it plays a major role in sphingolipid metabolism. Recently, the crystal structure of the PH-START complex has been released, suggesting an inhibitory action of START to the binding of the PH domain to the Golgi apparatus and thus limiting the CERT activity. Our study presents a combination of docking and molecular dynamic simulations of N-(3-hydroxy-1-hydroxymethyl-3-phenylpropyl)alkanamides (HPA) analogs and limonoids compounds known to inhibit CERT. Through our computational study, we compared the binding affinity of 14 ligands at both domains (START and PH) and also at the START-PH interface, including several mutations known to play a role in the CERT's activity. At the difference of HPA compounds, limonoids have a stronger binding affinity for the START-PH interface. Furthermore, 2 inhibitors (HPA-12 and isogedunin) were investigated through molecular dynamic (MD) simulations. 50 ns of molecular dynamic simulations have displayed the stability of isogedunin as well as keys residues in the binding of this molecule at the interface of the PH-START complex. Therefore, this study suggests a novel inhibitory mechanism of CERT for limonoid compounds involving the stabilization of the START-PH interface. This could help to develop new and potentially more selective inhibitors of this transporter, which is a potent target in cancer therapy.

Keywords: CERT, START domain, PH domain, limonoid inhibitors, cancer therapy, Ceramide

INTRODUCTION

Sphingolipids belong to a major class of lipids in eukaryotic cells. They are not only involved in the membrane structure. They also act as important mediators in cellular signaling (Hannun and Obeid, 2008). Sphingolipid metabolism is highly regulated by various enzymes located in different subcellular compartments (i.e., endoplasmic reticulum, Golgi apparatus, plasma membrane, mitochondria, and lysosomes). This compartmentalized enzymatic network contributes largely to the cellular function of sphingolipids. Among these sphingolipids, ceramides have been shown

to play a central role in the induction of apoptosis (Hannun and Obeid, 2018) and several ceramide metabolizing enzymes have been involved in induced-apoptosis in response to a variety of agents such as cytokines, chemotherapy and radiotherapy (Reynolds et al., 2004; Nganga et al., 2018). In contrast, other sphingolipids derived from ceramide metabolism, such as sphingosine-1-phosphate (S1P), sphingomyelin and glucosylceramide, have been shown to play either proliferative or protective properties (Hannun and Obeid, 2008; Maceyka et al., 2012). Cancer cells seem to have an altered balance between pools of sphingolipids promoting tumors from those having a suppressing role (Furuya et al., 2011; Morad and Cabot, 2013) which finally favor cells proliferation/survival. The inability of cancer cells to accumulate pro-apoptotic ceramides can be the consequence of not only a defect of *de novo* biosynthesis but also of an increased degradation into S1P or transformation of ceramide into sphingomyelin/glucosylceramide (Liu et al., 2013; Yandim et al., 2013). Consequently, the inability to accumulate ceramides has been associated with insensitivity to apoptosis induced by chemotherapy and radiotherapy (Dimanche-Boitrel and Rebillard, 2013). Importantly, inhibition of ceramide metabolism into glucosylceramide has been shown to be a critical factor to restore ceramide and re-sensitizes cancer cells to chemotherapy (Morad and Cabot, 2013).

Looking at the *de novo* ceramide biosynthesis which takes place in the endoplasmic reticulum (ER), it is known that the ceramides conversion into complex sphingolipids is not only based on enzyme activities. Namely glucosylceramide and sphingomyelin synthases (Gault et al., 2010). They are also regulated by specific transport between the ER and the Golgi apparatus (Kumagai and Hanada, 2019). Indeed, *de novo* sphingomyelin biosynthesis relies on non-vesicular ceramide trafficking of ceramide mediated by the ceramide transporter protein CERT. Specifically, CERT transports ceramides from the ER to the trans-Golgi regions at the ER-Golgi membrane contact sites. The inactivation of this transporter was shown to be a cellular response to induced apoptosis in several cell lines (Charruyer et al., 2008; Chandran and Machamer, 2012). It has been found that CERT expression is higher in drug-resistant cell lines (Swanton et al., 2007) and that molecular inhibition of CERT resulted in re-sensitization of cancer cells to chemotherapy (Lee et al., 2012; Palau et al., 2018). Taken together, this suggests that pharmacological inhibition of CERT can represent a novel anti-cancer strategy by overcoming drug resistance.

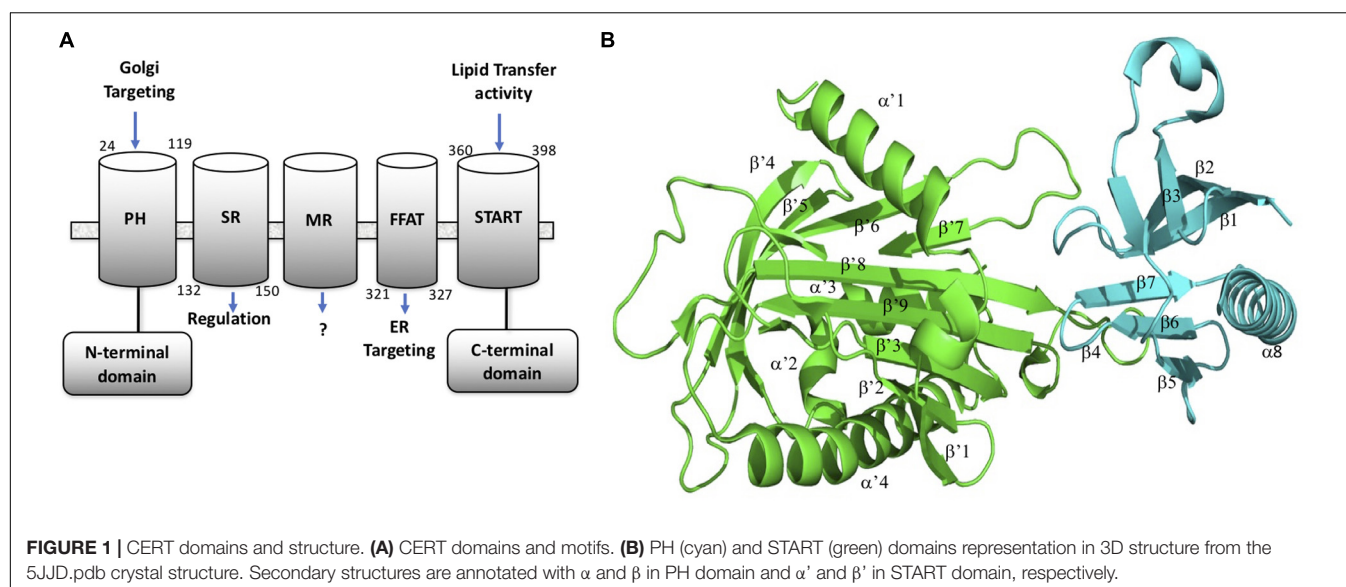
CERT is a cytosolic monomeric protein constituted of different domains involved in its function (Kumagai and Hanada, 2019). CERT has a lipid binding START (Steroidogenic Acute Regulatory protein-related lipid Transfer) domain at the C-terminus, a PH (Pleckstrin homology) domain at the N-terminus, and a FFAT (diphenylalanine in an acidic track) in the middle region. The latter domain binds the ER-localized protein, VAP-A, whereas the START domain is responsible for the binding and transport of ceramide (Hanada et al., 2003). The PH domain also plays a role in ceramide transport by binding the phosphoinositide, phosphatidyl-inositol-4-phosphate (PtdIns4P) which is abundant in the Golgi apparatus (De Matteis et al., 2005). START domain, PH domain, and FFAT motif are all required

for the full activity of CERT. CERT has also a serine-repeat (SR) motif, which decreases PtdIns4P binding and ceramide transfer activity when it is phosphorylated. It has been shown that the inhibition of PtdIns4P binding to the PH domain by hyper-phosphorylated SR motif requires the presence of the START domain (Kumagai et al., 2007). Recent crystal structure revealed that in fact the START domain interacts with PH domain at its PtdIns4P binding site (Prashek et al., 2017). Amino acid mutations that disrupt the PH/START interaction, increase ceramide-transfer activity of CERT, suggesting that this interaction plays an important role in the regulation of CERT cellular localization and ceramide transfer. The PH domain is formed of one α helix ($\alpha 8$) and seven β -sheets ($\beta 1$ – $\beta 7$). The START domain is structured with four α helix ($\alpha' 1$ – $\alpha' 4$) and nine β -sheets ($\beta' 1$ – $\beta' 9$). Representation of the CERT domains and START-PH interaction complex are summarized in **Figure 1**.

Three unrelated families of CERT inhibitors have been described up to date. The N-(3-hydroxy-1-hydroxymethyl-3-phenylpropyl)alkanamides (HPA) family, synthesized analogs of ceramide, with HPA-12 as a lead, were identified as inhibitors of ceramide trafficking at the beginning of 2000 (Yasuda et al., 2001) by specifically binding to the START domain. It was found that the amphiphilic cavity of the START domain consists of hydrophobic residues that recognize the amide and hydroxyl groups of HPAs by hydrophobic interactions (Kudo et al., 2008). Importantly, many hydrophobic interactions are conserved in both HPA and ceramide binding, supporting a competitive inhibitory effect of this compound. More recently, the virtual screening of small chemicals leads to the discovery of a new CERT inhibitor, HPCB-5, which selectively binds the START domain (Nakao et al., 2019). At the difference of HPA compounds, HPCB-5 has no apparent ceramide mimicry and can be considered as a ceramide-non-mimetic inhibitor of CERT. Another screen of a library of natural small compounds reveals that limonoids such as isogedunin selectively inhibited CERT activity (Hullin-Matsuda et al., 2012). It was shown that limonoids inhibit the CERT-mediated ceramide extraction from isolated ER membranes in order to block sphingomyelin biosynthesis. At the difference of HPA ligands, limonoids are unable to inhibit the rapid transfer of an exogenously added fluorescent short chain ceramide analog, supporting the idea of different inhibitory mechanisms of CERT mediated by HPA and limonoid families.

Up to date, there is no molecular mechanism proposed for the inhibitory effect of limonoid on CERT activity. The objective of this study is to determine the mechanism of interaction of limonoids to CERT in comparison to the binding of HPA ligands. We specifically explored the role of the new CERT regulation system where the PH and START domains interact with each other on a basal state to maintain CERT as inactive (Prashek et al., 2017). Accordingly, a computational study has been conducted to investigate the stability of isogedunin, and HPA-12 binding at the interaction interface of the two domains in order to maintain a self-inhibition.

First, a docking approach has been performed with a set of 16 ligands on the START, PH and START-PH complex system showing a preference of limonoid compounds to bind at the START-PH interaction site. Then, molecular dynamic (MD)



simulations were performed to evaluate the stability of HPA-12 and isogedunin ligand at the interface of both domains and in a presence of mutations W33A, R43A, and Y54A at the PH domain and, E494R, N495K, P535R, and E537K at the START domain. These mutations have been described to be involved in the START-PH interaction (Prashek et al., 2017) and seem to also play a role in the interaction of isogedunin. Results of these molecular modeling approaches are presented in the next section.

MATERIALS AND METHODS

Protein Preparation and Docking Protocol

In order to predict the protein-ligand interactions through molecular docking, known 3D structures of CERT were selected from the Protein Data Bank (PDB) (Burley et al., 2018)¹. For the docking protocol, the chosen structures were based on a few criteria: (a) The best possible resolution; (b) Protein domains bound to a ligand when available; (c) Protein domains containing no mutations or modified residues; (d) human protein. Accordingly, X-ray crystal structures of START domain complexed with HPA-13 ligand (PDB ID: 3H3Q), unbound PH domain (PDB ID: 4HHV), and START/PH complex (PDB ID: 5JJD) were used as protein targets in the docking protocol (Supplementary Table S1).

Proteins and Ligands Preparation for Molecular Docking

To have a good estimation of the protonated state of charged residues, each protein was protonated according to the physiological pH (pH = 7.4) using the PROPKA server (Li et al., 2005). About the ligands, ten limonoid compounds reported to inhibit CERT by Hullin-Matsuda et al.

(2012) and five (1R, 3R)-N-(3-Hydroxy-1-hydroxymethyl-3-phenylpropyl)alkanamide (HPA) analogs (Kudo et al., 2010) were used in this analysis (Supplementary Figure S1). The molecules were collected in SMILES (Simplified Molecular Input Line Entry Specification) format and converted into 3D conformation using KNIME software (Fillbrunn et al., 2017). Molecules were treated with the same physiological pH (pH 7.4) as the proteins and Gasteiger's charges were added using an Open Babel script (O'Boyle et al., 2011) before to be saved into mol2 format.

Molecular Docking Studies

To study the interaction between limonoids/HPAs and CERT domains, water molecules were removed and the interaction surfaces were identified as follow: For START and PH domain, a grid around the HPA binding site in START domain and around the sulfate (SO₄) that binds in place for ligand binding in PH domain was built (using the option "center on ligand" in AutoDock), respectively. For the START-PH complex, a grid that encompasses the interaction surface between the two domains were developed (using the option center on macromolecule). Then, docking was run with the standard AutoDock (v4.2) suit incorporated in MGL tools (v1.5.6) using Lamarckian Genetic Algorithm (Forli et al., 2016). To identify the most favorable binding site of each inhibitor into the two domains and the complex system, flexible ligand docking was performed. The input grid parameter files were modified and the grid size was adjusted with 0.375 Å grid spacing to cover the active site region of receptors (Supplementary Figure S2). Since the target-ligand poses are ranked using an energy-based scoring function with AutoDock, only the best pose conformation of docked ligand was saved, visualized and studied with PyMOL (Rigsby and Parker, 2016). Basically, the lower is the energy, the higher is the binding affinity. Hydrogen bonding interactions and distances between the different domains and ligands were visualized and measured using PyMOL. PyMOL was also considered in the development of several virtual mutations of residues known to

¹<http://www.rcsb.org/>

play an important role in the interaction of PH and START domains (Prashek et al., 2017).

Stability Evaluation by Molecular Dynamics Simulations (MD)

The MD approach was performed to study the dynamic behavior of CERT as well as the structural stability of START-PH complex with docked isogedunin and HPA-12 ligands. MD simulations were carried out using GROMACS software package (v5.1.4) (Van der Spoel et al., 2005). First, ligand-free CERT topology was prepared using “pdb2gmx” with the OPLS-AA/L all atom force field. On the other hand, PDB files of docked complexes (CERT-isogedunin and CERT-HPA-12) were separated into PDB files of protein and ligand. Then, topology of CERT was prepared using GROMOS96 43a1 force field. The ligands topology was developed using the “PRODRG” server (van Aalten et al., 2005). The box (unit cell) in which the protein was located has been defined and the system has been filled with water. The protein ligand system was kept in a cubic box, filled of waters (TIP3) and preserving a minimum distance of 10 Å between each atom of the system and walls of the box. The resulting system was solvated by a single point charge (SPC) 216 solvent model, provided in GROMACS parameters. In order to neutralize the system with a net charge of -7, counter ions of 7 NA⁺ were added. An energy minimization was performed on the system to eliminate steric clashes using the “steepest descent” method. Next, the system was equilibrated for 100 ps with 50,000 steps. For the equilibration phase, NVT (Number of particles, Volume, and Temperature) equilibration was performed for 100 ps at a temperature of 300k with a coupling constant of 0.1 ps. Once the temperature was stabilized, NPT (Number of particles, Pressure, and Temperature) was run by setting the temperature to 300k and the pressure to 1 bar. Electrostatic interactions were calculated using the Particle-Mesh Ewald method (PME). Finally, the production phase was performed for 50 ns (MD run) with a time step of 2 fs to make sure the system is stable. The MD simulation was run in triplicate on each system. The parameters of the MD simulations are described in **Supplementary Figure S3**.

MM-PBSA Analysis

The molecular mechanics Poisson Boltzmann surface area (MM-PBSA) method is the widely used method for binding free energy calculations from the snapshots of MD trajectory and *g_mmpbsa* was used for this present study (Kumari et al., 2014). It integrates functions from GROMACS and APBS² to determine the polar and non-polar contributions of the binding energy. The dielectric relative constant ϵ has been set to 2 for ligands and 80 for water (Kukic et al., 2013). In this approach, the binding free energy ΔG_{bind} between a protein and a ligand include different energy terms and could be calculated as:

$$\begin{aligned}\Delta G_{\text{bind}} &= G_{\text{complex}} - (G_{\text{protein}} + G_{\text{ligand}}) \\ &= \Delta E_{\text{MM}} - T\Delta S + \Delta G_{\text{sol}} \\ &= \Delta E_{\text{vdw}} + \Delta E_{\text{ele}} + \Delta G_{\text{PB}} + \Delta G_{\text{SA}} - T\Delta S\end{aligned}$$

²http://rashmikumari.github.io/g_mmpbsa/

ΔG_{bind} is the binding free energy. ΔE_{MM} stands for the gas-phase interaction energy, which is the sum of van der Waals energy ΔE_{vdw} and electrostatic energy ΔE_{ele} . ΔG_{sol} is the sum of polar solvation energy ΔG_{PB} and the non-polar solvation energy ΔG_{SA} . The polar solvation energy is calculated using Poisson Boltzmann (PB) approximation model, while the non-polar solvation energy is estimated by solvent accessible surface area (SASA). The entropy contribution ($-T\Delta S$) is ignored in this study because of its expensive computational demand. MM/PBSA is applied to, the 50 ns MD simulations (500 ps-spaced) of our different protein-ligand systems to estimate their free binding energies.

RESULTS AND DISCUSSION

Docking Analysis of START- PH- and START-PH Complex Ligands Interactions

In this study, fifteen natural compounds were docked into each of the three different domains, i.e., START and PH domains independently (**Table 1**), and on the START/PH interaction domain (**Table 2**). We selected the top binding pose of each molecule bound to the domains, based on the binding energies estimated by AutoDock (Onawole et al., 2018).

About the PH domain, it was reported that the SO4 ligand bound with R43, K32, Y54, and K56 (Prashek et al., 2013). In our study, most of the docked ligands do not interact simultaneously with these four residues but only with one of them. Cedrelone was the ligand with the lowest binding energy, with an estimated binding score of -7.18 kcal/mol (**Table 1**). Cedrelone formed two hydrogen bonds with Y63 and S57 located on loop $\beta 3/\beta 4$. This loop is known to contribute to the conservative PH domain pocket composition. Additional contacts including hydrophobic

TABLE 1 | Docking results of the 15 ligands on PH domain and START domain.

Ligands	PH	START
	Energy (kcal/mol)	Energy (kcal/mol)
Carapin	-6.27	-8.48
Cedrelone	-7.18	-8.62
Isogedunin	-6.85	-9.79
Khayantone	-6.71	-10.17
Khivorin	-6.61	-8.53
Limonin	-6.95	-6.26
Methyl angolensate	-6.99	-8.07
Obliquin	-5.58	-6.06
Odoratone	-6.81	-10.60
Prieuranin	-2.61	-3.28
HPA-12	-4.40	-6.57
HPA-13	-4.93	-6.85
HPA-14	-4.73	-7.11
HPA-15	-4.83	-6.71
HPA-16	-3.07	-6.79

The energy score is a binding affinity estimation between a compound and the surrounding residues.

TABLE 2 | Docking results of the 15 ligands on the START-PH interaction site wild type and on 4 CERT mutants.

Ligands	PH-START interface				
	WT	R43A	R43A/Y54A/W33A	E494R/N495K	E494R/N495K/P535R/E537K
	Energy (kcal/mol)	Energy (kcal/mol)	Energy (kcal/mol)	Energy (kcal/mol)	Energy (kcal/mol)
Isogedunin	−11.75	−12.03	−11.08	−11.31	−11.38
Cedrelone	−9	−8.91	−8.71	−8.88	−8.86
Limonin	−9.93	−9.95	−9.53	−10.24	−10.23
Khivarin	−5.91	−1.43	−0.95	0.32	−2.73
Khavanthhone	−3.74	−3.34	−2.18	−3.31	−3.13
Obliquin	−6.93	−6.93	−7.05	−6.94	−6.88
Odoratone	−7.55	−6.74	−10.56	−7.17	−6.85
Methyl angolensate	−9.65	−9.74	−9.63	−10.29	−10.21
Prieuranin	0.14	1.19	14.93	10.38	8.99
Carapin	−8.66	−8.27	−8.62	−8.6	−8.5
HPA-12	−6.99	−8.17	−7.64	−7.88	−7.48
HPA-13	−5.62	−7.46	−8.11	−7.83	−7.59
HPA-14	−6.36	−8.91	−7.21	−8.91	−7.56
HPA-15	−5.85	−6.8	−7.74	−7.25	−6.52
HPA-16	−6.45	−6.83	−6.68	−6.8	−8.4

The energy score is a binding affinity estimation between a compound and the surrounding residues. The mutants R43A, Y54A, and W33A are located in the PH domain, whereas E495R, N495K, P535R, and E537K are located on the START domain.

interactions, have been observed with residues W44 and V29 (**Figure 2**). In opposite, all the HPA's analogs bound with higher energies than the limonoids compounds (i.e., lower binding affinity) (**Table 1**). They preferably interacted with hydrophobic residues and formed hydrogen bonds with the residue R43 only. These results suggest a more suitable interaction site for limonoids compounds into the PH domain.

About the START domain, HPA-13 has been reported to bind with residues N504, E446, Y553, and Q467 (Kudo et al., 2010). From our docking analysis, HPA-13 did not recover completely the same pose as the one observed in the crystal structure (**Figure 3**). This is probably due to the long alkyl chain which gives some flexibility to the molecule. Still, interactions with residues E446, Q467, and Y553 are present.

Surprisingly, some limonoids compounds showed lower binding energies than HPA compounds suggesting that these compounds would have a higher inhibition effect than HPA ligands on the START domain (**Table 1**). For example, carapin forms hydrogen bonds with Y553, N504, Q467, and E446. Hydrophobic interactions are made with W445, T448, I523, V525, and Y576 and a salt bridge with R442 is also present (**Figure 4**).

Finally, following the study from Prashek et al. (2017) ligands were docked into the START-PH interaction domain in order to evaluate if these compounds could have an impact by stabilizing the inhibitory mechanism of interaction between these two domains. Again, limonoid compounds showed lower binding energies score to START-PH compared to the HPA ligands. Isogedunin, cedrelone, and methylangolensate have the lowest binding energies when the two domains are linked together, resulting in more favorable interactions (**Table 2**). About isogedunin, hydrogen bonds are made with N495, P497, E498 from the START domain, and with W33 and Y35 from the

PH domain. Hydrophobic interactions can be observed with Y96, P497, and L532. Furthermore, isogedunin forms a salt bridge with K32, as well as π -stacking with Y36. All these interactions seem to stabilize isogedunin at the interface of CERT. Interactions with some of these residues are found for other limonoids such as limonin, methyl angolensate, and cedrelone and represented in **Figure 5**.

Interestingly, some of the residues that participate in the binding with these ligands have recently been reported to be important in the interaction of the two domains. Residues of the START domain, such as N495, E498, and V533, appeared to form extensive hydrogen bonds, and hydrophobic or π -stacking interactions with K32, W33, Y36, and Y96 which are located at the PH domain. Therefore, we decided to run docking of our set of ligands on four mutated START-PH complexed systems, including the mutations R43A, W33A/R43A/Y54A in PH domain, E494R/N495K and E494R/N495K/P535R/E537K in START domain. For, the limonoid compounds, the mutations do not have a big impact on the binding energies in the interface of the START-PH domains which seems to be maintained in an inactive form. This is not the case for HPA compounds for which docking results reveal mutations improve the ligands' binding affinity for many HPA compounds (**Table 2**). However, the binding energies are still weaker than with some limonoid compounds such as isogedunin, limonin, or methyl angolensate. At the end of this docking study, we could conclude that the limonoid compounds bind more favorably to the interface of the START-PH domains and so could play a major role in the inhibitory activity of CERT by stabilizing this interaction. To estimate further this hypothesis, we decided to run several molecular dynamic (MD) simulations corresponding to (i) the START-PH complex without ligand and with different mutations suggested previously, (ii) the START-PH complex

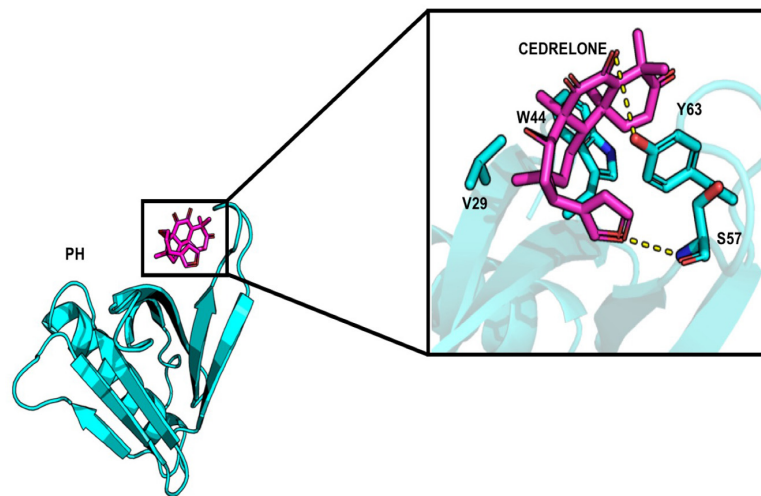


FIGURE 2 | Representation of cedrelone docked into the PH domain. Cedrelone is in magenta and the residues interacting with cedrelone are in sticks.

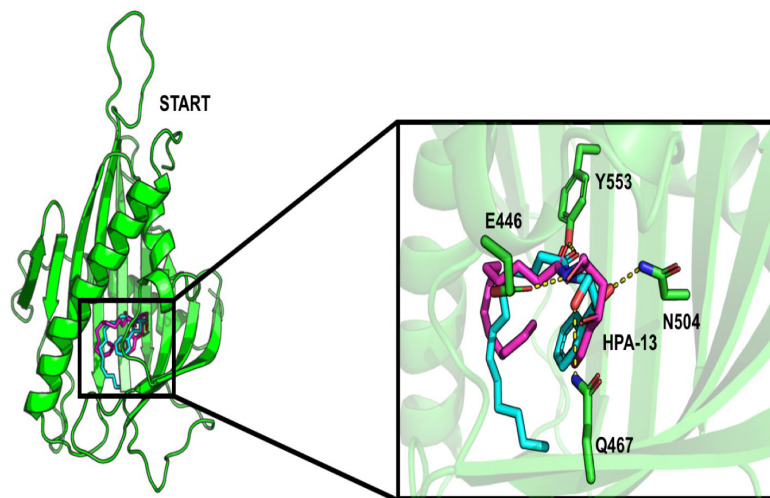


FIGURE 3 | Comparison of the interaction of HPA-13 into the START domain from the X-ray (in cyan) and the docking approach (magenta). Residues that interact with both poses are in sticks.

with the HPA-12 ligand, (iii) the START-PH complex with the isogedunin ligand, (iv) the START-PH complex with the mutation E494R/N495K and the isogedunin ligand. For the last one, the E494R/N495K mutant was reported to be important to the PH-START interaction and is also present in the binding of isogedunin. So, it is expected to observe how the ligands are stabilized during the MD simulations.

Stability Evaluation of PH-START Complex by MD Simulations

Backbone RMSD of the Wild Type and Mutated PH-START Complex

Based on the 50 ns of production from the MD simulations for the WT and the 4 CERT mutants, root-mean square deviation

(RMSD) analysis enabled the measure of average distances (in Å) of the studied systems from the corresponding starting structure over the simulation period. We have to notice that the G387 residues on the START domain was missing in the X-ray structure. Therefore, the region from the N terminal T364 to the C terminal of V386 was removed from the RMSD and RMSF analysis as it was fluctuated much more compared to the others area.

In general, the systems were relatively stable with a maximum RMSD around 4 Å. The WT and the two other CERT mutants on the PH domain (R43A and W33A/R43A/Y54A) remain stable around 2 Å along the 50 ns. These mutants seem to destabilize the interaction of the START-PH domains less (**Figure 6**) compared to the CERT mutants, E494R/N495K and E494R/N495K/P535R/E537K, on the START domain. These

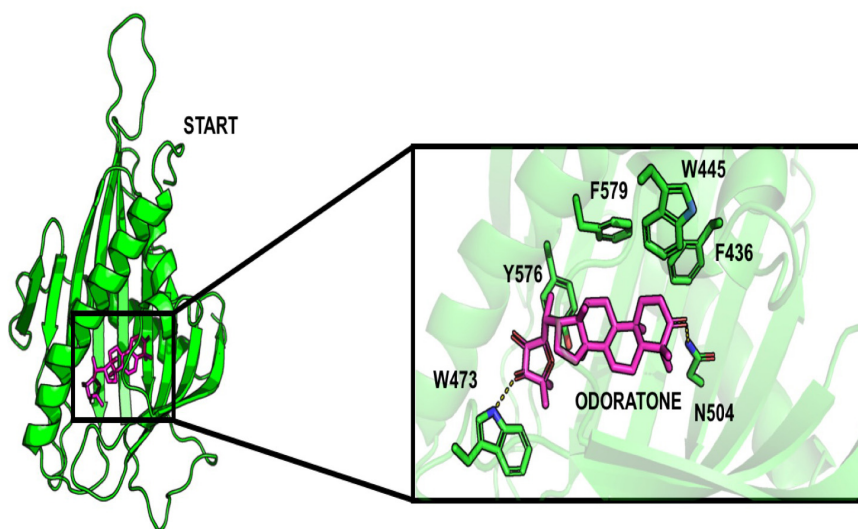


FIGURE 4 | Docking pose of odoratone into the START domain. The odoratone is in magenta and the residues of START domain surrounding the ligand are in green. The odoratone and the residues involved in the ligand binding are represented in sticks.

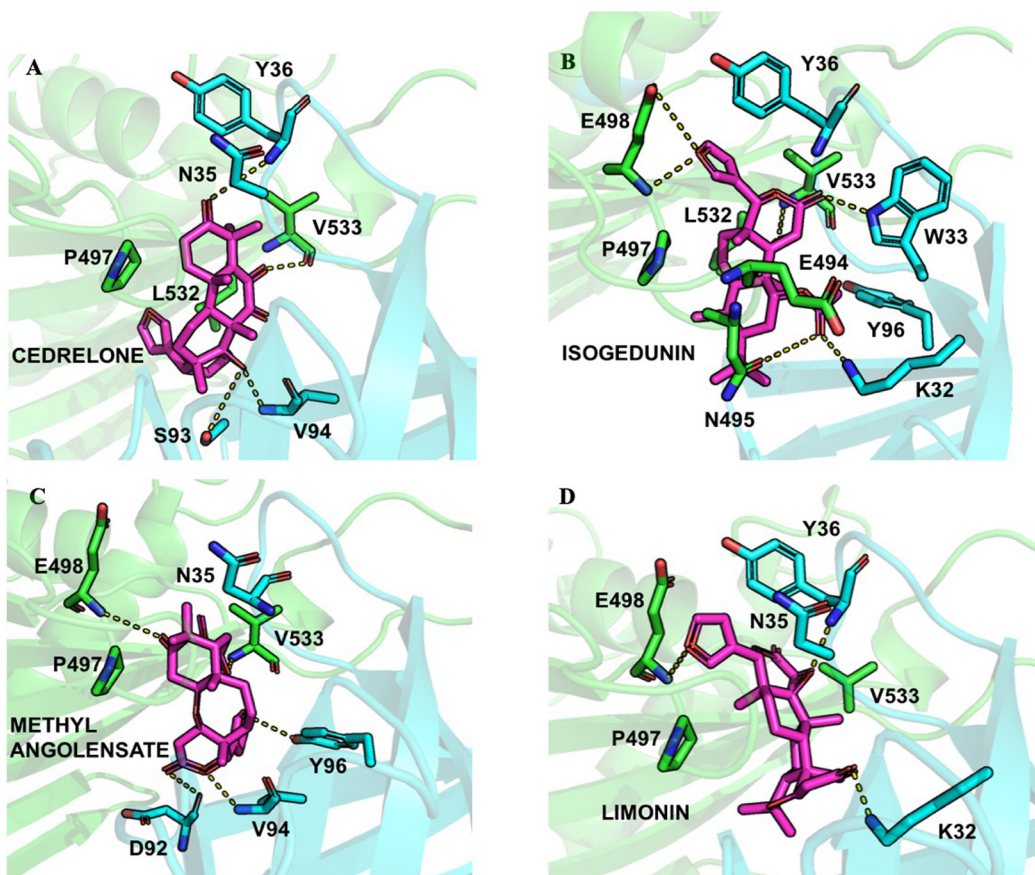
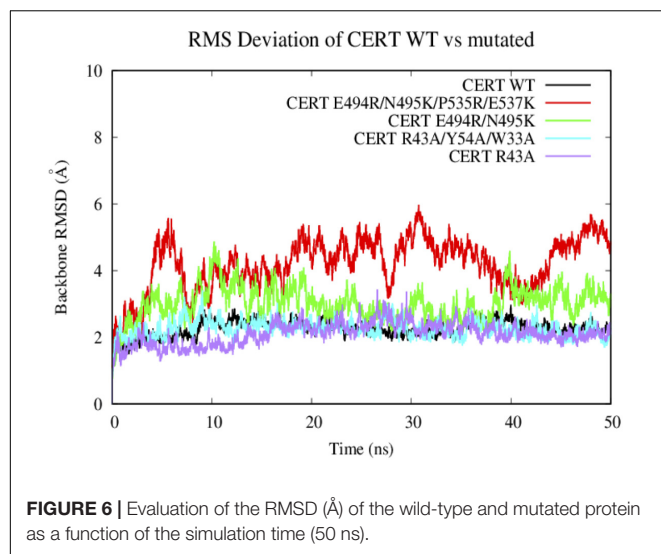


FIGURE 5 | Ligand interactions depicted between (A) Cedrelone, (B) Isogedunin, (C) Methyl angolensate, and (D) Limonin within the PH-START interface.



results suggest that the mutations on the START domain could have an impact on the stability of the START-PH interaction (Prashek et al., 2017).

C-Alpha RMSF of the Wild Type and Mutated PH-START Complex

The RMSF results are subsequently correlated with the root-mean-square fluctuation of the residues along the MD simulations (RMSF) (Figure 7). Some mutations have more or less affected the flexibility of the PH domain residues. R43A and R43A/Y54A/W33A mutations have significantly increased the flexibility of the loops between the different β -strands. For example, R43A/Y54A/W33A mutation had a greater effect on the $\beta 1/\beta 2$, $\beta 3/\beta 4$, and $\beta 5/\beta 6$ loops which are involved in the interaction surface of both domains. This means that these mutated residues destabilized the environment of the residues on the PH domain surface. However, no effect was observed with the presence of the R494R/N495K mutation. With the E494R/N495K/P535R/E537K mutation, only residues involved in loop $\beta 3/\beta 4$ (E58-D59) showed an important flexibility with an RMSF value going from 1 to about 4 Å. The START domain showed significant increase in the flexibility of the various loops. A major gain of flexibility is observed on the loop between $\alpha 1'$ and $\beta 1'/\beta 2'$. Such result was expected since this loop is partly crystallized and therefore tends to be unstable. More importantly, all mutations affected the residues of loops $\beta 6'/\beta 7'$ and $\beta 8'/\beta 9'$ involved in the CERT interface and known to be crucial for PH/START stability. Due to the E494R/N495K/P535R/E537K mutation, RMSF values of loop $\beta 6'/\beta 7'$ went from 2 to 4 Å and this probably explains the instability of the complex. These results also agree with the findings discussed in the previous section.

Analysis of the Ligands Effect on the PH-START Complex by MD Simulations

Docking approaches help to determine potential ligand binding patterns inside a protein, but such methods do not propose information on the structural stability of protein-ligand

complexes. For this reason, in the aim to reinforce the potential mode of binding of limonoid compound at the interface of the START-PH domains, MD simulations were produced (50 ns) and analyzed to evaluate the stability of isogedunin and HPA-12 in the START-PH interaction domains.

a. Backbone RMSD and C-Alpha RMSF of CERT Structure

Looking at the RMSD measured on the backbone of the two domains, the introduction of the ligands HPA-12 and isogedunin seems to have slightly disturbed the general conformation of the complex system compared to the wild type. We reached in average a RMSD around 2, 4, and 6 Å for the WT and with the introduction of isogedunin and HPA-12 ligands, respectively, (Figure 8). The increase of the RMSD could suggest several dynamic movements around the protein, notably at the interface of the START-PH domains. Isogedunin is relatively more stable compared to HPA for which the RMSD varies sensibly from one to another MD simulation and always higher than for isogedunin. This variability can be explained by the fact that HPA-12 has a very flexible and unstable alkyl chain, as compared to isogedunin that remains more static between the START-PH domains.

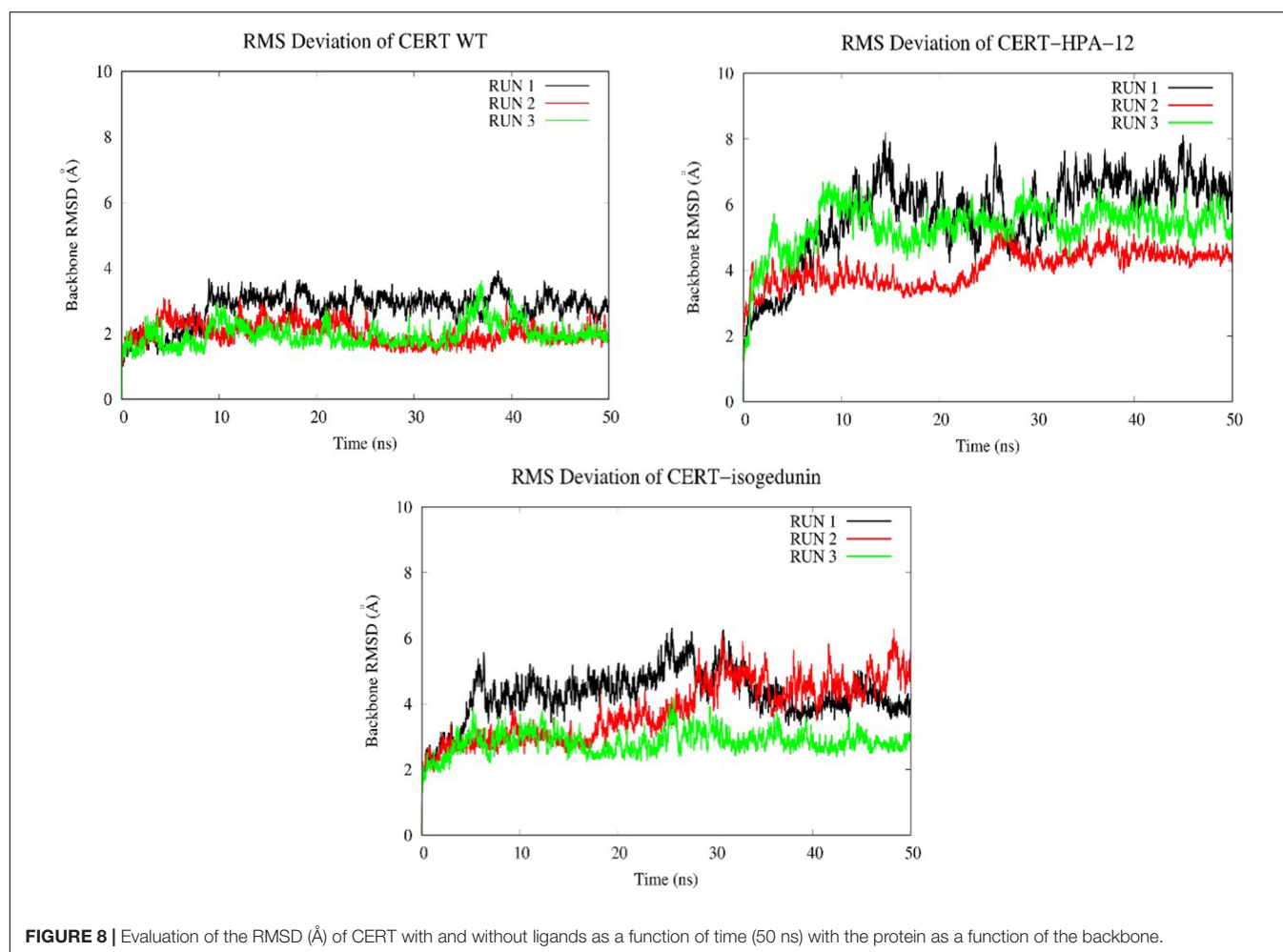
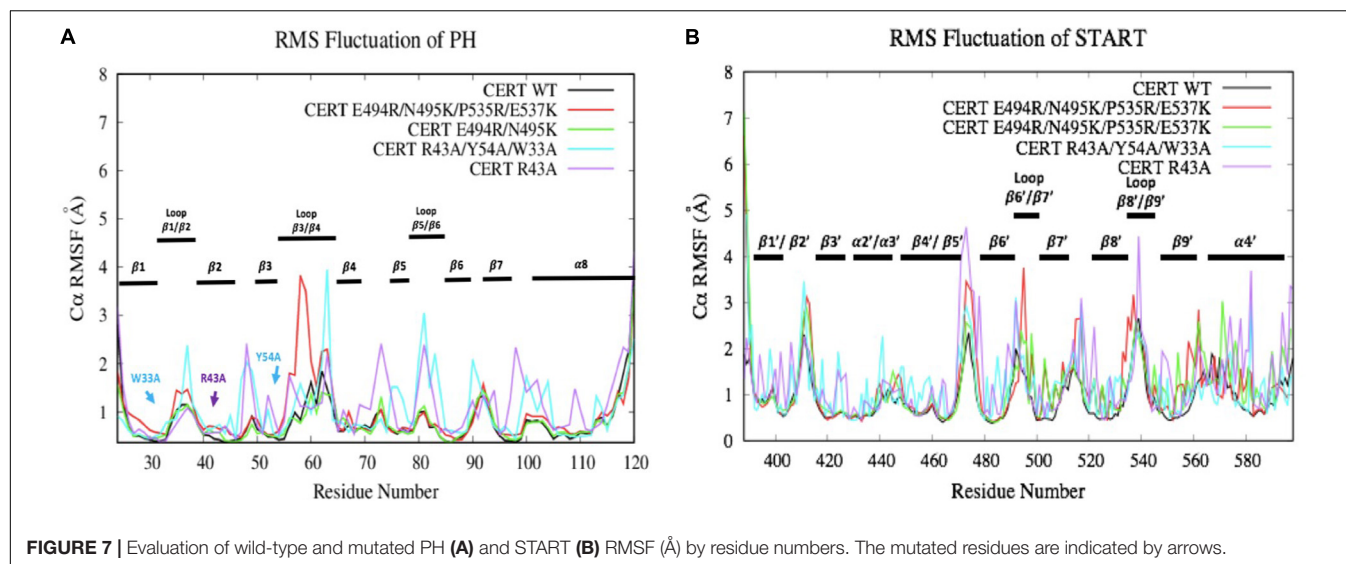
Looking at the RMSF (Figure 9) of CERT, there are bigger fluctuations in the START domain compared to the PH domain, notably the $\beta'2/\beta'3$ loops and $\beta'5$ to $\beta'9$ and $\alpha'4/\beta'8$. A similar trend is observed with isogedunin except that the fluctuations are a little higher in $\alpha 3$ and $\beta 5/\beta 6$ loops in the PH domain and $\beta'5/\beta'6$ in the START domain. In opposite, the RMSF are much higher for HPA-12, especially in the START domain suggesting that HPA-12 is more disturbing the START-PH complex than isogedunin.

To look at the impacts of the E494R/N495K mutation on the CERT-isogedunin complex, three molecular dynamics runs of 50 ns were launched following the same MD protocol (Figure 10). The three runs converged to different states of the complex. In fact, the third run seems to be more stable than the two first ones during the simulation time. However, we can clearly notice an increase of the backbone RMSD values compared to the wild type complex. Altogether, the molecular dynamics simulations of the mutated CERT-isogedunin complex reveal more variability. Therefore, we can assume the E494R/N495K mutation might destabilize the CERT-isogedunin complex and disturb their interaction over time.

To investigate the impact of the E494R/N495K mutation on the flexibility of CERT-isogedunin complex, the RMSF of both CERT domains were analyzed (Figure 11). Compared to the wild type complex, only a small increase of the RMSF values were seen. This increase was visible on the main loops involved in the CERT interaction surface. In fact, there is a slight gain of flexibility concerning the $\beta 6'/\beta 7'$ and $\beta 8'/\beta 9'$ loops of the START domain and loop $\beta 1/\beta 2$ loop of the PH domain. These observations show that the E494R/N495K mutation tends to disturb the START-PH interaction through less stable loops movements on the CERT interface.

Hydrogen Bond Analysis

Hydrogen bonds are known to play an essential role in the molecular recognition and stability of protein structures.



A greater number of interactions between the intermolecular hydrogen bond interaction results in the better stability of the protein-ligand complex. In this study, a hydrogen bond analysis

was conducted to examine the stability of docked isogedunin and HPA-12 systems. Concerning the CERT-isogedunin system, the hydrogen bond interactions reach a maximum number of 6

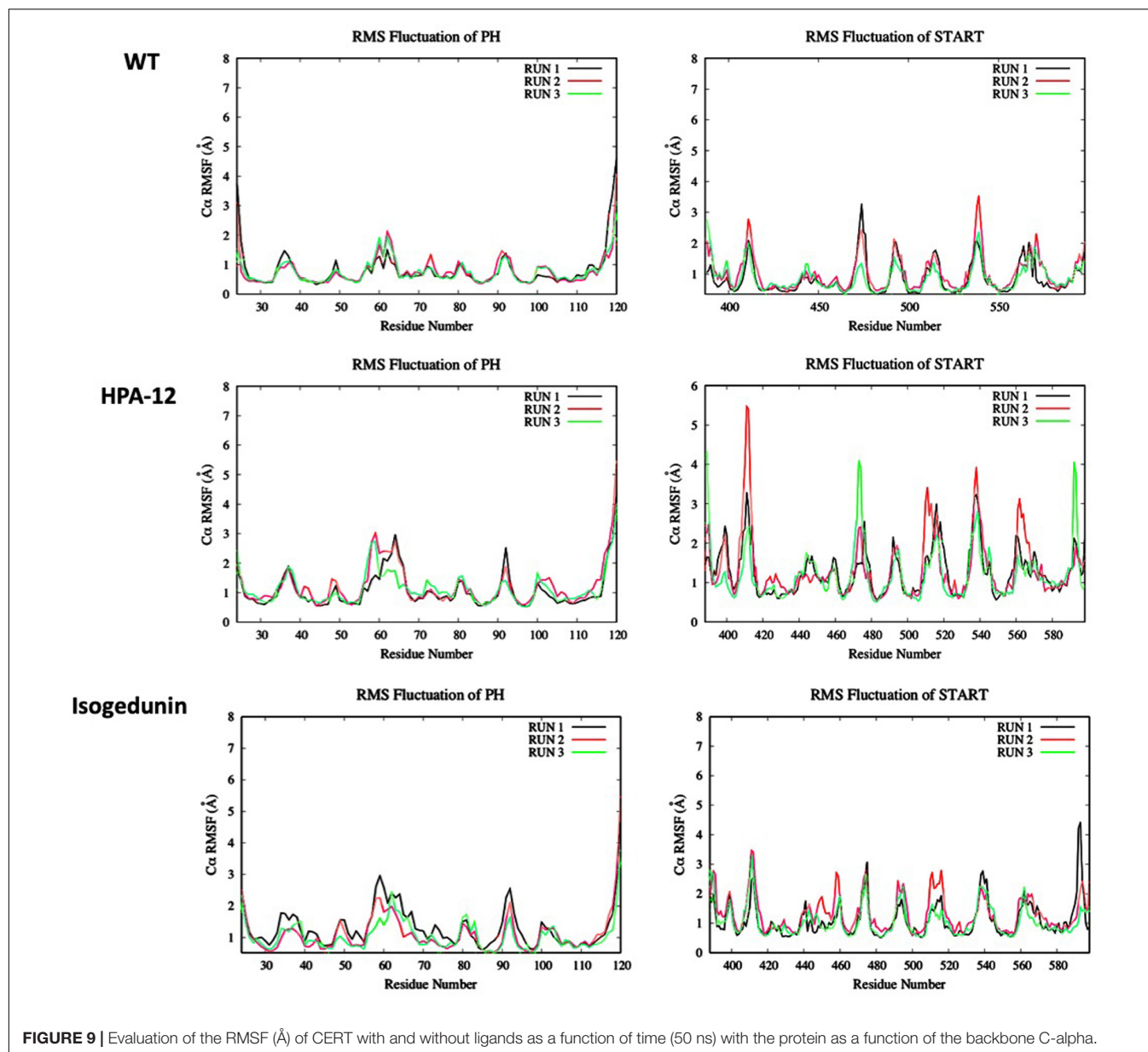


FIGURE 9 | Evaluation of the RMSF (Å) of CERT with and without ligands as a function of time (50 ns) with the protein as a function of the backbone C-alpha.

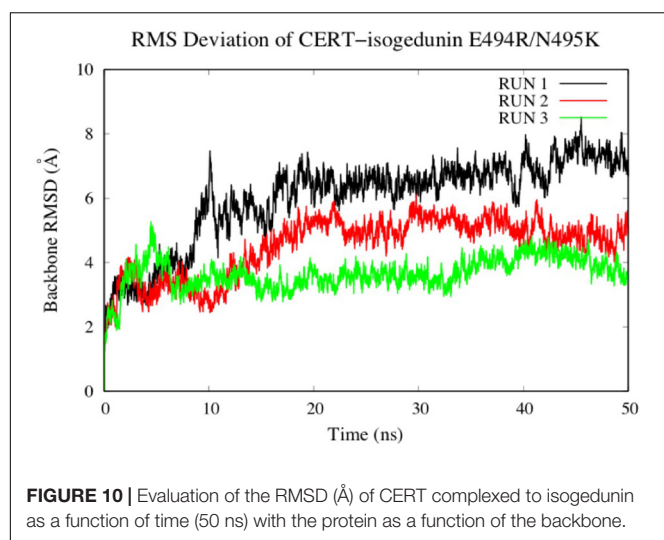
between 20 and 21 ns. Afterward, the hydrogen bonds numbers remain stable throughout the MD simulation and only fluctuate between 3 or 4 (**Figure 12A**). The hydrogen bond interactions number of CERT-HPA-12 system reaches a maximum of 6 as well but only up to 42 ns, then drops significantly to one and comes back to two at the end of the simulation. Furthermore, major interruptions dropping the number of bonds to 0 were observed from 9 to 30 ns (**Figure 12B**). Therefore, we suggest that the few H-bonding of CERT with HPA-12 may enable its disassociation.

To further investigate the stabilization of our systems, trajectories of protein-ligand complexes were analyzed. Hydrogen bonds are constantly formed with Q498 and V533 and isogedunin (**Supplementary Figure S4**). π -stacking interaction is also present between Y36 and isogedunin along the MD simulation suggesting stability of the ligand at the

interface of the START-PH domains. Nonetheless, HPA-12 remained very unstable and flexible on the CERT interaction surface (**Supplementary Figure S5**). At 10 ns, unlike isogedunin, HPA-12 did not form any hydrogen bonds with CERT. It was maintained at the interface through hydrophobic interactions instead. At 40 ns, it reaches its maximum hydrogen bonds number and interacts with K425, P535, Q537, and R85. However, these bonds quickly break and only K425 remains bonded to HPA-12. More about the contact frequency between residues and ligands are described in Supplementary Information (Contact Frequency Analysis and **Supplementary Figure S6**).

MM-PBSA Analysis

The MM-PBSA calculation of isogedunin and HPA-12 was performed using the g_mmpbsa tool (**Table 3**). The affinities

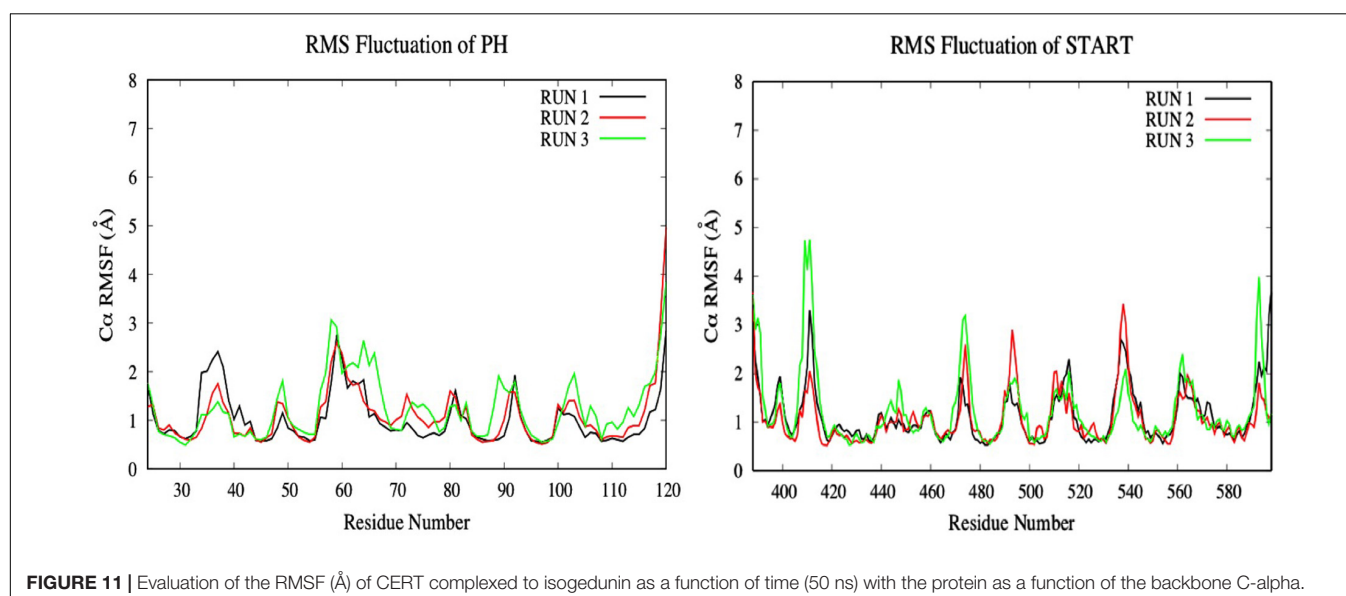


of the CERT-ligand complexes were analyzed based on the MD trajectory of each system. According to *g_mmpbsa* calculations, HPA-12 binds to CERT with a ΔG_{bind} of -168.869 kJ/mol. On the other hand, isogedunin binds to CERT with a lower free energy value of -211.753 kJ/mol. CERT-isogedunin in presence of the E494R/N495K mutation showed a ΔG_{bind} of -192.407 kJ/mol which is a little higher than the CERT-isogedunin. According to the energy composition, van der Waals energy seems to be the major force of the binding process of both complexes, with a slightly high affinity for the CERT-isogedunin complex. In fact, van der Waals energy is represented by hydrophobic interactions which play a major role to form stable complexes. The rest of the energy terms, such as the electrostatic energy or the solvent-accessible surface area (apolar) were also favorable components for the binding energy contribution. The overall binding free energies displayed that isogedunin has a

better free energy for CERT compared to HPA-12 which is in agreement with previous analyses.

DISCUSSION

In this study, the interaction between CERT and its potential inhibitors (limonoids and HPAs), as well as the effect of mutated key residues, were investigated using several computational approaches such molecular docking and MD simulations. The docking results allowed us to conclude on the greater affinity of limonoids for the START-PH complex as compared to HPAs. This specific type of compound has a better affinity at the START-PH interface. Mutations conducted on both domains confirmed that the mutated residues have no effect or increase the stability of the interaction between the START and PH domain. The MD simulations and the results of structural characteristic features such as RMSD, RMSF, and hydrogen bonding plots revealed higher stability of isogedunin at the START-PH interface due to increased hydrogen bond interactions. This may imply that this interface would be a favorable site of binding for isogedunin and could explain the assumption of a different inhibitory mechanism of CERT mediated by HPA and limonoid families. At the difference of HPA-12, limonoid compounds are unable to inhibit intracellular transport of fluorescent short chain ceramide from endoplasmic reticulum (RE) to the Golgi apparatus (Hullin-Matsuda et al., 2012). The discrepancy between these results could be explained by our data which suggests that limonoids would stabilize CERT in an inactive form (interaction between PH and START domain) unable to take in charge exogenous fluorescent ceramide (usually used to analyze CERT-mediated ceramide transport in cellulo). Limonoids inhibit the CERT-mediated extraction of endogenous ceramide from isolated membranes (Hullin-Matsuda et al., 2012), also probably by stabilizing CERT under its inactive form. In contrast, HPA-12 which preferentially bound START domain



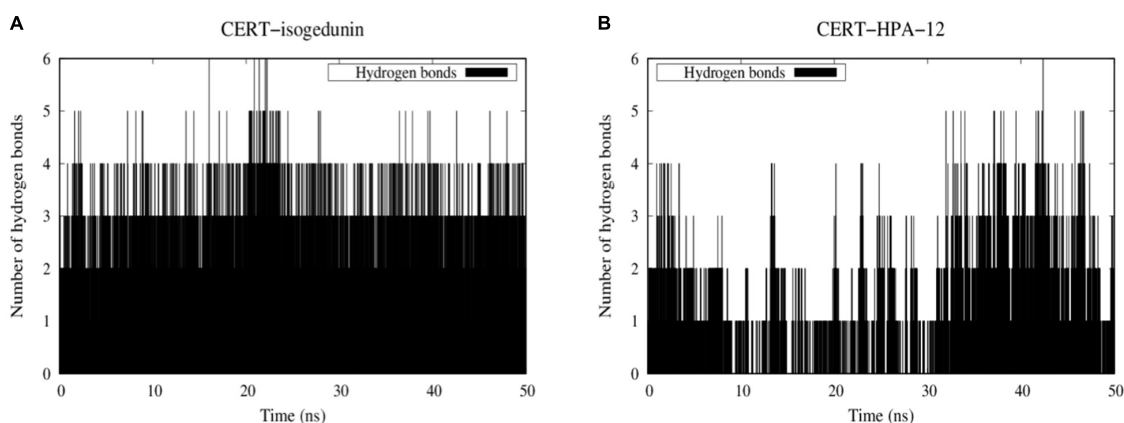


FIGURE 12 | Evolution of stability of CERT-isogedunin (A) and CERT-HPA-12 (B) complexes using a diagram showing intermolecular hydrogen interactions as a function of time (ns) with Gromacs (Van der Spoel et al., 2005).

TABLE 3 | Binding free energy of CERT-HPA-12 and CERT-isogedunin using g_mmpbsa method.

Complexes	Binding energy components (kJ/mol)				
	ΔE_{vdw}	ΔE_{elec}	ΔE_{polar}	ΔE_{apolar}	ΔG_{bind}
CERT-HPA-12	-285.45 ± 17.68	-49.03 ± 17.30	159.89 ± 23.18	-21.27 ± 1.14	-168.86 ± 21.31
CERT-isogedunin	-286.51 ± 14.28	-31.284 ± 6.63	126.28 ± 21.87	-20.24 ± 1.11	-211.75 ± 22.64
CERT-isogedunin E494R/N495K	-263.04 ± 55.69	-23.466 ± 7.47	112.26 ± 34.83	-18.17 ± 3.37	-192.41 ± 32.90

free of the PH domain, could easily prevent the binding of exogenous fluorescent ceramide. Prashek et al. have shown that the stability of PH-START domain interaction dictates CERT subcellular localization (Prashek et al., 2017). Indeed, disruption of PH-START domain interaction results in the translocation of CERT from the ER to the Golgi apparatus, mediating the transport of ceramide. We could therefore hypothesize that limonoids, by stabilizing PH-START interaction domain will stick CERT in the ER. On the other hand, HPA-12, which binds START domain, will rather favor a Golgi (or cytoplasmic) localization of CERT. Finally, interaction between PH and START domain contributes to the inhibition of CERT through its SR hyperphosphorylation (Kumagai et al., 2007). Limonoids, by stabilizing the interaction between the PH and the START domains could also favor hyperphosphorylation of the SR motif and can therefore inhibit CERT function.

CONCLUSION

In conclusion, our results demonstrate a novel mechanism of inhibition of CERT by limonoid compounds: an interfacial inhibitory mechanism. These inhibitors are thermodynamically more efficient (as evidenced in our study by comparing HPA-12 and isogedunin) and are likely to be more selective than competitive inhibitors (Pommier and Marchand, 2011). We believe that our findings will provide insights in the development of *in vitro* assays that can validate our computational study and guide for the development of limonoid analogs that

could selectively target CERT and used in novel cancer therapy strategies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

HL, CM, and OT elaborated the study. MG and OT performed the modeling analysis. All authors contributed to the writing of the manuscript.

ACKNOWLEDGMENTS

AL was granted by Cifre–Conventions Industrielles de Formation par la Recherche. This study contributes to IdEx Université de Paris ANR-18-IDEX-0001.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2020.603983/full#supplementary-material>

REFERENCES

- Burley, S. K., Berman, H. M., Christie, C., Duarte, J. M., Feng, Z., Wesbrook, J., et al. (2018). RCSB protein data bank: sustaining a libing digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci.* 27, 316–330. doi: 10.1002/pro.3331
- Chandran, S., and Machamer, C. E. (2012). Inactivation of ceramide transfer protein during pro-apoptotic stress by Golgi disassembly and caspase cleavage. *BioChem. J.* 442, 391–401. doi: 10.1042/BJ20111461
- Charruyer, A., Bell, S. M., Kawano, M., Douangpanya, S., Yen, T. Y., Macher, B. A., et al. (2008). Decreased ceramide transport protein (CERT) function alters sphingomyelin production following UVB irradiation. *J. Biol. Chem.* 283, 16682–16692. doi: 10.1074/jbc.M800799200
- De Matteis, M. A., Di Campli, A., and Godi, A. (2005). The role of the phosphoinositides at the Golgi complex. *Biochim. Biophys. Acta* 1744, 396–405. doi: 10.1016/j.bbamcr.2005.04.013
- Dimanche-Boitrel, M. T., and Rebillard, A. (2013). Sphingolipids and response to chemotherapy. *Handb. Exp. Pharmacol.* 216, 73–91. doi: 10.1007/978-3-7091-1511-4_4
- Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., and Berthhold, M. R. (2017). KNIME for reproducible cross-domain analysis of life science data. *J. Biotechnol.* 261, 149–156. doi: 10.1016/j.jbiotec.2017.07.028
- Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., and Olson, A. J. (2016). Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* 11, 905–919. doi: 10.1038/nprot.2016.051
- Furuya, H., Shimizu, Y., and Kawamori, T. (2011). Sphingolipids in cancer. *Cancer Metastasis Rev.* 30, 567–576. doi: 10.1007/s10555-011-9304-1
- Gault, C. R., Obeid, L. M., and Hannun, Y. A. (2010). An overview of sphingolipid metabolism: from synthesis to breakdown. *Adv. Exp. Med. Biol.* 688, 1–23. doi: 10.1007/978-1-4419-6741-1_1
- Hanada, K., Kumagai, K., Yasuda, S., Miura, Y., Kawano, M., Fukasawa, M., et al. (2003). Molecular machinery for non-vesicular trafficking of ceramide. *Nature* 426, 803–809. doi: 10.1038/nature02188
- Hannun, Y. A., and Obeid, L. M. (2008). Principles of bioactive lipid signaling: lessons from sphingolipids. *Nat. Rev. Mol. Cell. Biol.* 9, 139–150. doi: 10.1038/nrm2329
- Hannun, Y. A., and Obeid, L. M. (2018). Sphingolipids and their metabolism in physiology and disease. *Nat. Rev. Mol. Cell. Biol.* 19, 175–191. doi: 10.1038/nrm.2017.107
- Hullin-Matsuda, F., Tomishige, N., Sakai, S., Ishitsuka, R., Ishii, K., Makino, A., et al. (2012). Limonoid compounds inhibit sphingomyelin biosynthesis by preventing CERT protein-dependent extraction of ceramides from the endoplasmic reticulum. *J. Biol. Chem.* 287, 24397–24411. doi: 10.1074/jbc.M112.344432
- Kudo, N., Kumagai, K., Matsubara, R., Kobayashi, S., Hanada, K., Wakatsuki, S., et al. (2010). Crystal structures of the CERT START domain with inhibitors provide insights into the mechanism of ceramide transfer. *J. Mol. Biol.* 396, 245–251. doi: 10.1016/j.jmb.2009.12.029
- Kudo, N., Kumagai, K., Tomishige, N., Yamaji, T., Wakatsuki, S., Nishijima, M., et al. (2008). Structural basis for specific lipid recognition by CERT responsible for nonvesicular trafficking of ceramide. *Proc. Natl. Acad. Sci. U.S.A.* 105, 488–493. doi: 10.1073/pnas.0709191105
- Kukic, P., Farrell, D., McIntosh, L. P., Garcia-Moreno, E. B., Jensen, K. S., Toleikis, Z., et al. (2013). Protein dielectric constants determined from NMR chemical shift perturbations. *J. Am. Chem. Soc.* 135, 16968–16976. doi: 10.1021/ja406995j
- Kumagai, K., and Hanada, K. (2019). Structure, functions and regulation of CERT, a lipid-transfer protein for the delivery of ceramide at the ER-Golgi membrane contact sites. *FEBS Lett.* 593, 2366–2377. doi: 10.1002/1873-3468.13511
- Kumagai, K., Kawano, M., Shinkai-Ouchi, F., Nishijima, M., and Hanada, K. (2007). Interorganelle trafficking of ceramide is regulated by phosphorylation-dependent cooperativity between the PH and START domains of CERT. *J. Biol. Chem.* 282, 17758–17766. doi: 10.1074/jbc.M702291200
- Kumari, R., Kumar, R., Open Source Drug Discovery Consortium, and Lynn, A. (2014). g_mmpbsa—a GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.* 54, 1951–1962. doi: 10.1021/ci500020m
- Lee, A. J. X., Roylance, R., Sander, J., Gorman, P., Endesfelder, D., Kschischo, M., et al. (2012). CERT depletion predicts chemotherapy benefit and mediates cytotoxic and polyploid-specific cancer cell death through autophagy induction. *J. Pathol.* 226, 482–494. doi: 10.1002/path.2998
- Li, H., Robertson, A. D., and Jensen, J. H. (2005). Very fast empirical prediction and rationalization of protein pKa values. *Proteins* 61, 704–721. doi: 10.1002/prot.20660
- Liu, Y. Y., Hill, R. A., and Li, Y. T. (2013). Ceramide glycosylation catalyzed by glucosylceramide synthase and cancer drug resistance. *Adv. Cancer Res.* 117, 59–89. doi: 10.1016/B978-0-12-394274-6.00003-0
- Maceyka, M., Harikumar, K. B., Milstien, S., and Spiegel, S. (2012). Sphingosine-1-phosphate signaling and its role in disease. *Trends Cell Biol.* 22, 50–60. doi: 10.1016/j.tcb.2011.09.003
- Morad, S. A., and Cabot, M. C. (2013). Ceramide-orchestrated signaling in cancer cells. *Nat. Rev. Cancer* 13, 51–65. doi: 10.1038/nrc3398
- Nakao, N., Ueno, M., Sakai, S., Egawa, D., Hanzawa, H., Hawasaki, S., et al. (2019). Natural ligand-nonmimetic inhibitors of the lipid-transfer protein CERT. *Commun. Chem.* 2:20. doi: 10.1038/s42004-019-0118-3
- Nganga, R., Oleinik, N., and Ogretmen, B. (2018). Mechanisms of ceramide-dependent cancer cell death. *Adv. Cancer Res.* 140, 1–25. doi: 10.1016/bs.acr.2018.04.007
- O'Boyle, N. M., Banck, M., Craig, A. J., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: an open chemical toolbox. *J. Cheminform.* 3:33. doi: 10.1186/1758-2946-3-33
- Onawole, A. T., Kolapo, T. U., Sulaiman, K. O., and Adegoke, R. O. (2018). Structure based virtual screening of the Ebola virus trimeric glycoprotein using consensus scoring. *Comput. Biol. Chem.* 72, 170–180. doi: 10.1016/j.compbiolchem.2017.11.006
- Palau, V. E., Chakraborty, K., Wann, D., Lightner, J., Hilton, K., Brannon, M., et al. (2018). γ -Tocotrienol induces apoptosis in pancreatic cancer cells by upregulation of ceramide synthesis and modulation of sphingolipid transport. *BMC Cancer* 18:564. doi: 10.1186/s12885-018-4462-y
- Pommier, Y., and Marchand, C. (2011). Interfacial inhibitors: targeting macromolecular complexes. *Nat. Rev. Drug Discov.* 11, 25–36. doi: 10.1038/nrd3404
- Prashek, J., Bouyain, S., Fu, M., Li, Y., Berkes, D., and Yao, X. (2017). Interaction between the PH and START domains of ceramide transfer protein competes with phosphatidylinositol 4-phosphate binding by the PH domain. *J. Biol. Chem.* 292, 14217–14228. doi: 10.1074/jbc.M117.780007
- Prashek, J., Truong, T., and Yao, X. (2013). Crystal structure of the pleckstrin homology domain from the ceramide transfer protein. Implications for conformational change upon ligand binding. *PLoS One* 8:e79590. doi: 10.1371/journal.pone.0079590
- Reynolds, C. P., Maurer, B. J., and Kolesnick, R. N. (2004). Ceramide synthesis and metabolism as a target for cancer therapy. *Cancer Lett.* 206, 169–180. doi: 10.1016/j.canlet.2003.08.034
- Rigsby, R. E., and Parker, A. B. (2016). Using the pymol application to reinforce visual understanding of protein structure. *Biochem. Mol. Biol. Educ.* 44, 433–437. doi: 10.1002/bmb.20966
- Swanton, C., Marani, M., Pardo, O., Warne, P. H., Kelly, G., Sahai, E., et al. (2007). Regulators of mitotic arrest and ceramide metabolism are determinants of sensitivity to paclitaxel and other chemotherapeutic drugs. *Cancer Cell.* 11, 498–512. doi: 10.1016/j.ccr.2007.04.011
- van Aalten, D. M., Bywater, R., Findlay, J. B., Hendlich, M., Hooft, R. W., and Vriend, G. (2005). Prodrgr, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comput. Aided Mol. Des.* 10, 255–262. doi: 10.1007/BF00355047

- Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005). GROMACS: fast, flexible and free. *J. Comput. Chem.* 26, 1701–1718. doi: 10.1002/jcc.20291
- Yandim, M. K., Apohan, E., and Baran, Y. (2013). Therapeutic potential of targeting ceramide/glucosylceramide pathway in cancer. *Cancer Chemother. Pharmacol.* 71, 13–20. doi: 10.1007/s00280-012-1984-x
- Yasuda, S., Kitagawa, H., Ueno, M., Ishitani, H., Fukasawa, M., Nishijima, M., et al. (2001). A novel inhibitor of ceramide trafficking from the endoplasmic reticulum to the site of sphingomyelin synthesis. *J. Biol. Chem.* 276, 43994–44002. doi: 10.1074/jbc.M104884200

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

Copyright © 2020 Ghoula, Le Marec, Magnan, Le Stunff and Taboureau. 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.



Analyzing *In Silico* the Relationship Between the Activation of the Edema Factor and Its Interaction With Calmodulin

Irène Pitard^{1,2,3}, Damien Monet^{1,2,3}, Pierre L. Goossens⁴, Arnaud Blondel^{1,2} and Thérèse E. Malliavin^{1,2*}

¹ Unité de Bioinformatique Structurale, Institut Pasteur and CNRS UMR 3528, Paris, France, ² Center de Bioinformatique, Biostatistique et Biologie Intégrative, Institut Pasteur and CNRS USR 3756, Paris, France, ³ Ecole Doctorale Université Paris Sorbonne, Paris, France, ⁴ Institut Pasteur, rue du Dr Roux, Unité Yersinia, Paris, France

OPEN ACCESS

Edited by:

Agnel Praveen Joseph,
Science and Technology Facilities
Council, United Kingdom

Reviewed by:

Sony Malhotra,
Birkbeck, University of London,
United Kingdom
Igor N. Berezovsky,
Bioinformatics Institute (A*STAR),
Singapore

*Correspondence:

Thérèse E. Malliavin
therese.malliavin@pasteur.fr

Specialty section:

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

Received: 23 July 2020

Accepted: 02 November 2020

Published: 04 December 2020

Citation:

Pitard I, Monet D, Goossens PL,
Blondel A and Malliavin TE (2020)
Analyzing *In Silico* the Relationship
Between the Activation of the Edema
Factor and Its Interaction With
Calmodulin.
Front. Mol. Biosci. 7:586544.
doi: 10.3389/fmolb.2020.586544

Molecular dynamics (MD) simulations have been recorded on the complex between the edema factor (EF) of *Bacillus anthracis* and calmodulin (CaM), starting from a structure with the orthosteric inhibitor adefovir bound in the EF catalytic site. The starting structure has been destabilized by alternately suppressing different co-factors, such as adefovir ligand or ions, revealing several long-distance correlations between the conformation of CaM, the geometry of the CaM/EF interface, the enzymatic site and the overall organization of the complex. An allosteric communication between CaM/EF interface and the EF catalytic site, highlighted by these correlations, was confirmed by several bioinformatics approaches from the literature. A network of hydrogen bonds and stacking interactions extending from the helix V of CaM, and the residues of the switches A, B and C, and connecting to catalytic site residues, is a plausible candidate for the mediation of allosteric communication. The greatest variability in volume between the different MD conditions was also found for cavities present at the EF/CaM interface and in the EF catalytic site. The similarity between the predictions from literature and the volume variability might introduce the volume variability as new descriptor of allostery.

Keywords: protein-protein interaction, *Bacillus anthracis*, virulence factor, Cavity detection, allostery

INTRODUCTION

As the interactions between proteins are essential in all biological processes, the modulation of these interactions with small ligands is a promising direction (Morelli et al., 2011; Tuffery and Derreumaux, 2012; Huang, 2014; Zhang et al., 2014; Aguirre et al., 2015; Kuenemann et al., 2015; Fischer et al., 2016; Shin et al., 2017). During the last decade, virtual screening has experienced a turning point where interest has widened from protein active sites to cryptic sites (Beglov et al., 2018; Vajda et al., 2018) not visible in the isolated protein conformation but formed upon ligand binding. These cryptic sites have been proposed to be detected by analysis of protein structures (Kozakov et al., 2015; Cimermanic et al., 2016), by mixed-solvent MD simulations (Ghanakota et al., 2019; Martinez-Rosell et al., 2020) or by biased MD simulations (Comitani and Gervasio, 2018; Sun et al., 2020). Inhibition of protein-protein interactions, search for new cryptic sites and targeting protein function in an allosteric way are three closely related goals.

Targeting allostery (Tscharmer, 2016; Deredge et al., 2017; Feng et al., 2018; Ni et al., 2019) is particularly suitable in the case of inhibition of protein-protein interactions, because the conformational changes in which allosteric communication plays an important role when establishing the interaction, may create or modify cavities (Goodey and Benkovic, 2008).

Allostery, discovered in the early days of molecular biology (Monod et al., 1965; Koshland et al., 1966) has since then, evolved from a model of discrete protein conformations to a more continuous description of the free energy landscape of proteins (Goodey and Benkovic, 2008; Liu and Nussinov, 2016; Wodak et al., 2019). Consequently, the allostery is nowadays quite often described as being closely connected to the variations of equilibrium between protein conformations. Networks of protein residues have been pointed out to be involved in the transmission of these equilibrium variations (Gur et al., 2013; van den Bedem et al., 2013; Raman et al., 2016). The evolution of allostery description has allowed the emergence of so-called allosteric ligands (Goodey and Benkovic, 2008; Nussinov and Tsai, 2013), for which the binding to the protein has a long-range influence on the protein conformation via residue networks. In this regard, a pioneering approach has been based on a detailed mechanistic simulation of functional motions (Laine et al., 2010a). The allosteric ligands and the associated allosteric pockets attract a lot of attention, because they expand the range of pockets and ligands that could be exploited in effector design by virtual screening studies (Zhang and Nussinov, 2019). In addition, the diversity of allosteric phenomena can be utilized to limit the emergence of resistance mutations in target proteins, making carefully selected allosteric ligands more robust to the appearance of resistance in pathogens.

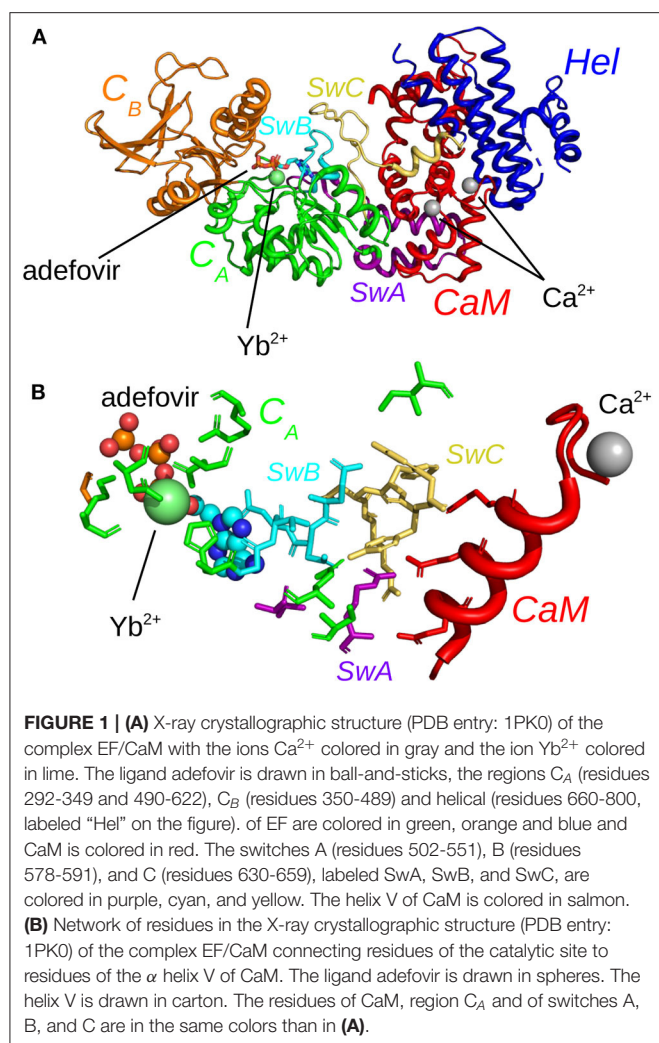
As a result, numerous methods have been developed for allosteric pocket detection in protein structures (Panjkovich and Daura, 2014; Xu et al., 2018; Ghode et al., 2020; Tan et al., 2020). In particular, it was shown (Ma et al., 2016) that most of the known allosteric site motions show high correlations with corresponding orthosteric site motions, whereas other surface cavities did not. Many of the prediction methods are based on a description of protein structures as an elastic network in which each residue is replaced by the atom α and the structure deformations are modeled through a set of springs between these atoms (Panjkovich and Daura, 2014; Guarnera and Berezovsky, 2019). The structure-based statistical mechanical model of allostery (SBSMMA) (Guarnera and Berezovsky, 2019) has been introduced which permits to calculate the free energy variation due to allosteric communication.

In the present work, we investigate the use of molecular dynamics (MD) simulations and bioinformatics approaches to analyze protein-protein interactions, variability of cavities and allostery for one example of protein-protein interaction corresponding to the activation of a virulence factor. The Edema Factor (EF) of *Bacillus anthracis* is activated in the cytoplasm of the host cell by interacting with the ubiquitous protein calmodulin (CaM). This interaction depends on the level of Ca^{2+} loaded by CaM, the C terminal lobe of CaM (C-CaM) displaying the highest affinity for ions Ca^{2+} during the interaction with EF (Ulmer et al., 2003). The EF/CaM

complex has been extensively studied by structural biology and biophysical techniques (Drum et al., 2000, 2001, 2002; Shen et al., 2002, 2004a, 2005; Ulmer et al., 2003; Guo et al., 2004, 2008) as well as by molecular modeling (Laine et al., 2008, 2009, 2010a,b, 2012; Martínez et al., 2009). This complex (**Figure 1A**) represents a very good example of an interaction with induced conformational selection for both partners. Indeed, free CaM in solution displays a very heterogeneous set of conformations, with wide range of relative re-orientations of N terminal (N-CaM) and C terminal (C-CaM) lobes (Bertini et al., 2004; Anthi et al., 2011), whereas CaM in complex with EF is blocked in an extended conformation. Similarly, the inactive state and the activated state of EF display largely different conformations. The helical region (residues 660-800) is moved apart from the C^A (residues 292-349 and 490-622) region to allow CaM insertion. A large conformational reorganization of switches A (residues 502-551, purple), B (residues 578-591, cyan), and C (residues 630-659, yellow) also takes place and the catalytic site is reshaped in its active organization (**Figure 1A**). Strikingly, in switch C two strings β and a connection loop present in the structure of isolated EF are converted to a α helix in the structure of the EF/CaM complex.

In the literature, both orthosteric and allosteric ligands have been proposed to inhibit EF activity. Several orthosteric inhibitors, binding to the catalytic site, have been discovered (Soelaiman et al., 2003; Shen et al., 2004b; Chen et al., 2009; Taha et al., 2009, 2012; Geduhn et al., 2011). Among them, the ligand adefovir (Shen et al., 2004b) was found by X-ray crystallography to bind in the catalytic site in the presence of an Yb^{3+} ion coordinated by adefovir as well as by protein residues. On the other hand, the compound 10506-2A has been shown to be an IPPI (inhibitor of protein-protein interaction) and to bind close to the EF helical regions (Lee et al., 2004). Thiophen ureoacid ligands have been discovered following virtual screening on the pocket SABC, formed by residues from the three switches A, B, and C (Laine et al., 2010a). Since it is believed that they do not bind to the enzymatic site of EF, compounds 10506-2A and thiophene ureoacids must by definition bind to an allosteric site.

Here, we propose the following approach to detect protein regions which should be targeted using an allosteric approach to inhibit the activity of EF. Starting from the X-ray crystallographic structure of EF/CaM complex bound to the orthosteric inhibitor adefovir (Shen et al., 2004b), we destabilized it by removing alternatively several co-factors: the ion Mg^{2+} present in the catalytic site, the ions Ca^{2+} loaded by CaM or the ligand adefovir. The analysis of the trajectories made it possible to detect a network of hydrogen bonds and stacking interactions, connecting the EF catalytic site and the EF/CaM interface and showing a strong destabilization when the co-factors are removed from the EF/CaM structure. In addition, cavities present in the EF/CaM complex have been tracked along all MD trajectories, and cavity volume variability has been proposed as a method of detecting allosteric pockets. The results obtained by this approach on EF/CaM have been confirmed by analyzes (Xu et al., 2018; Guarnera and Berezovsky, 2019). Therefore, the network of hydrogen bonds and stacking interactions connecting the EF catalytic site and the EF/CaM interface could be considered a



plausible candidate for the allosteric communication path within the complex structure.

MATERIALS AND METHODS

Preparation of the Systems for MD Simulations

The X-ray crystallographic complex of EF with the inhibitor adefovir (Shen et al., 2004b) (PDB entry: 1PK0, **Figure 1A**) served as the starting point of the MD trajectories. The protein chain was analyzed using Molprobiy (Chen et al., 2010) (molprobiy.biochem.duke.edu), in order to add hydrogen atoms and to select the sidechains orientations optimizing the network of hydrogen bonds. The ion Yb^{3+} present in the catalytic site, was replaced by a physiologically compatible ion Mg^{2+} .

The files to perform MD simulations were prepared with the CHARMM GUI interface (www.charmm-gui.org) (Lee et al., 2016; Jo et al., 2017). The chains A and D of the structure 1PK0 were neutralized using potassium ions and solvated with water molecules (**Table 1**). The loop (residues 675-695) located between

helices L and M (Drum et al., 2002) in the helical domain is disordered and not visible in the initial 1PK0 structure. This missing loop was added to the structure using the CHARMM GUI interface. The force field CHARMM36 (MacKerell et al., 1998, 2004; Best et al., 2012) and the TIP3P water model (Jorgensen et al., 1983) were used to model the physical interactions. The parameters for ligand adefovir were obtained using the CHARMM GUI interface (www.charmm-gui.org), with the Ligand Reader and Modeler tool (Kim et al., 2017). Six different systems have been prepared with different molecular compositions, starting from the structure 1PK0 then by removing various co-factors: ions Mg^{2+} , Ca^{2+} and adefovir (**Table 1**).

Recording MD Trajectories

The MD trajectories were recorded using NAMD 2.13 (Phillips et al., 2005) (www.ks.uiuc.edu/Research/namd/). The simulations were performed in the NPT ensemble. A cutoff of 12 Å and a switching distance of 10 Å were defined for non-bonded interactions, while long-range electrostatic interactions were calculated with the Particle Mesh Ewald (PME) protocol (Darden et al., 1993). The RATTLE algorithm (Ryckaert et al., 1977; Andersen, 1983) was used to keep all covalent bonds involving hydrogens rigid, enabling a time step of 2 fs. Atomic coordinates were saved every 10 picoseconds. At the beginning of each trajectory, the system was first minimized for 10,000 steps, then heated up gradually from 0 to 300 K in 300,000 integration steps. Then, the system was equilibrated for 50,000 steps. For each of six various conditions (**Table 1**), two independent trajectories of 200 ns were recorded (named the replicas R1 and R2) and corresponding to a total simulation duration of 2.4 μs .

Analysis of MD Trajectories

The root-mean-square deviations (RMSD, Å) of atomic coordinates, their root-mean-square fluctuations (RMSE, Å), as well as distance and angle analysis between atoms along the recorded trajectories were performed using cpptraj (Roe and TE Cheatham, 2013). Angles between CaM α helix axes were calculated using python scripts based on the python MDAnalysis library (Michaud-Agrawal et al., 2011; Gowers et al., 2016), the helices being defined as CaM regions including residues 8-19 (helix I), 31-37 (helix II), 46-53 (helix III), 66-73 (helix IV), 83-92 (helix V), 103-110 (helix VI), 119-127 (helix VII), 139-145 (helix VIII). The axis of an helix spanning residues n to p is defined as a segment connecting the geometric centers of the atoms Ca^n and Ca^{n+2} and of the atoms Ca^p and Ca^{p+2} .

The solvent accessible surfaces of residues along the trajectory were calculated using a python script based on the python MDAnalysis library (Michaud-Agrawal et al., 2011; Gowers et al., 2016) and the software FreeSASA (Mitternacht, 2016). The EF catalytic site surface was defined as the sum of solvent accessible surfaces of EF residues H351, K353, S354, K372, R329, K346, L348, D491, D493, H577, G578, T579, D582, N583, E588, F586, and T548. The surface of the SABC pocket was defined as the sum of the solvent accessible surfaces of EF residues A496, P499, I538, E539, P542, S544, S550, W552, Q553, T579, Q581, L625, Y626, Y627, N629, and N709. The surface of hydrophobic patches

TABLE 1 | Systems composition.

System	Solute	Number of TIP3P waters	Neutralizing K ⁺ ions	Total number of atoms
EF_ade_Mg_CaM_Ca	EF, CaM, Mg ²⁺ , 2 Ca ²⁺ , adefovir	68,219	9	215,183
EF_ade_CaM_Ca	EF, CaM, 2 Ca ²⁺ , adefovir	68,220	11	215,187
EF_ade_Mg_CaM	EF, CaM, Mg ²⁺ , adefovir	69,773	13	219,847
EF_ade_CaM	EF, CaM, adefovir	69,795	15	219,914
EF_CaM_Ca	EF, CaM, 2 Ca ²⁺	69,849	11	220,034
EF_CaM	EF, CaM	69,785	15	219,844

are defined according to Yang et al. (2004). The N-CaM patch is formed by the N-CaM residues A10, F12, A15, L18, F19, L32, M36, L39, M51, V55, M71, M72, and M76. The C-CaM patch is formed by the C-CaM residues I85, A88, V91, F92, L105, M109, L112, L116, M124, F141, M144, M145, and A147. The accessible surface was calculated using the Lee-Richards algorithm (Lee and Richards, 1971) with a probe radius of 1.4 Å.

Cavities of the EF/CaM complex were detected using the software *mkgridXf* (Monet et al., 2019). The cavities are delimited between two surfaces: the inner and out surfaces determined by rolling probes of radii 1.4 and 8 Å on the protein atoms. In addition, allosteric pockets were predicted using the Web server of the PARS approach (Panjkovich and Daura, 2014) (bioinf.uab.cat/cgi-bin/pars-cgi/pars.pl).

RESULTS

The Removal of Co-factors Destabilizes the Organization of the EF/CaM Complex

The evolution of coordinate root-mean-square deviations (RMSD, Å) with respect to the initial conformations of chains A (EF) and D (CaM) in the structure 1PK0 calculated along the trajectories, displays quite similar trends, with a plateau attained in most of the cases, after 50 ns, and located between 4 and 6 Å (Figure 2). The reproducibility of the replicas of a given trajectory has been further analyzed by realizing a clustering on each replica. This clustering was performed using a self-organizing map approach, described in the **Supplementary Material**. RMSD coordinates were calculated between the representative conformations extracted from clustering by comparing them two by two (Supplementary Figure 1). The average RMSD values between replicas of the same trajectory (yellow cells) are larger than the average RMSD values within each replicas (gray cells), but are mostly smaller than the average RMSD between different trajectories. Interestingly, EF_ade_Mg_CaM_Ca, which corresponds to the full system, displays the smallest RMSD of 3.9 Å between the replicas.

The coordinate root-mean-square fluctuations (RMSE, Å) (Figure 3) display similar profiles for all trajectories. Unsurprisingly, a certain variability is observed for the peak of fluctuations on the loop (residues 675–695) missing in the initial X-ray crystallographic structure and modeled when preparing

the system for molecular dynamics simulation, as described in section 2. Globally, the N-CaM region fluctuates more than the C-CaM region, loaded with Ca²⁺ ions.

Interestingly, the removal of ions does not increase much the mobility provided that the ligand adefovir is still present (trajectory EF_ade_CaM). By contrast, the removal of one type of ion in the presence of the ligand (trajectories EF_ade_CaM_Ca and EF_ade_Mg_CaM) or of the ions and the ligand (trajectory EF_CaM) increases the internal mobility of EF and produces a shift between the replicas. The trajectory EF_CaM_Ca, corresponding to the active toxin ready to interact with adefovir, has an internal mobility similar to that of EF_ade_Mg_CaM_Ca. The similarity of internal mobility for these two EF/CaM complexes mirrors their functional correspondence.

The global shape of the EF/CaM complex was analyzed by monitoring the gyration radius of the complex as well as the bending angle of the central α helix of CaM defined as the angle between the axes of helices IV and V (Figure 4). The gyration radius mainly samples values in the 19–22 Å range, with the exception of EF_ade_Mg_CaM in which the gyration radius vary in the 18–21 Å range (Figure 4, x-axis). Removal of Magnesium (EF_ade_CaM_Ca) or of Calcium (EF_ade_Mg_CaM) ions in the presence of adefovir induces an increase in the range of variations. On the other hand, the removal of the ligand makes it possible to maintain a narrow distribution of values, but shifts the radius of gyration to smaller values, around 19 Å. Thus both ligand and ions have an influence on the complex expansion.

The bending of the central α helix of CaM, monitored as the angle between axes of α helices IV and V (Figure 4, y-axis), is relatively stable around 130° for the full system EF_ade_Mg_CaM_Ca as well as for EF_CaM_Ca and EF_CaM. But, the systems EF_ade_CaM_Ca and EF_ade_Mg_CaM, in which the adefovir is present and only one ion type is conserved show large drifts of angle toward smaller angles down to 100° or larger angles up to 140°. Furthermore, the 2D contour plot describing the joint probability distribution of the gyration radius and the bending angle (Figure 4) reveals a strong correlation between the variations of these two parameters for systems EF_ade_CaM_Ca and EF_ade_Mg_CaM.

The drifts in the bending angle of the central α helix is a sign of a destabilization of the EF/CaM complex. Indeed, CaM displays in the EF/CaM complex a relatively unusual extended conformation (Yamniuk and Vogel, 2004). It was highlighted in Laine et al. (2008) using normal mode analysis of different

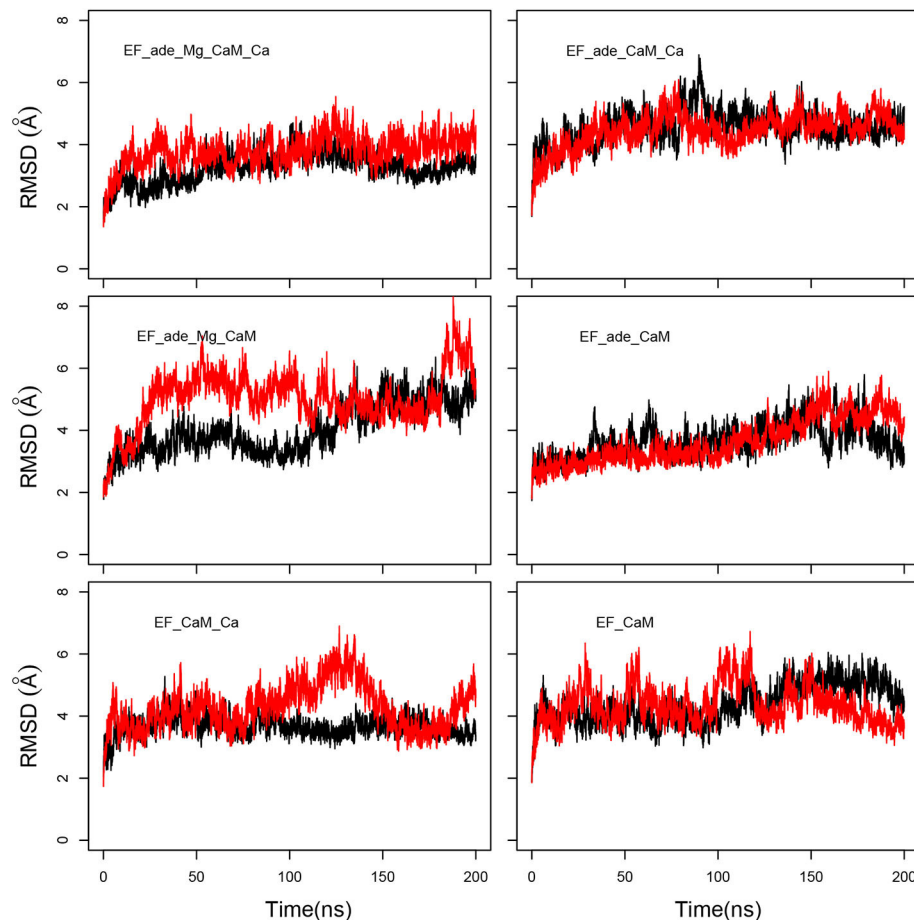


FIGURE 2 | Coordinate root-mean-square deviations (RMSD: Å) of the backbone heavy atoms of EF and CaM with respect to the PDB structure 1PK0, calculated along all trajectories. For each trajectory, the black and red curves correspond to the two replicas R1 and R2.

trajectories recorded on EF/CaM that an exact fit of the extended conformation of CaM to the EF structure is required for the complex stability. Indeed, in the EF/CaM complex, the extended structure of CaM is used to push away the helical domain from the remaining part of EF. The variation of bending angle in CaM central helix has a direct influence on the extension of the conformation of CaM and therefore on its adjustment to the activated EF. On the other hand, the variation of gyration radius of the whole complex corresponds to a major perturbation and perturbs the catalytic site and catalytic activity of EF. Thus the correlation observed between the gyration radius and the bending angle (**Figure 4**) proves that the fit of extended CaM to the complex EF/CaM has an influence on the EF function.

Overall, the removal of ions induces perturbations in the EF/CaM complex, which displays important conformational changes highlighted by variations of the gyration radius. These perturbations are visible through the atomic fluctuations as well as the correlated variations of gyration radius and CaM central α helix bending. The bending of the central helix α of CaM, due to the local environment, is related to the radius of gyration,

describing the general shape of the complex, thus highlighting a long-distance effect.

CaM Conformation Inside the EF/CaM Complex Conserves Features of the Isolated CaM

The CaM conformation will be analyzed and compared to the literature (Crivici and Ikura, 1995) information in order to assess the fitting of CaM to the interaction with EF. CaM is an extremely flexible protein (Bertini et al., 2004; Anthis et al., 2011), its conformation being strongly modulated by the loading of Calcium ions (Finn B.E. et al., 1995; Zhang et al., 1995; Komeiji et al., 2002). This allows CaM to bind various target proteins and peptides involved in the signaling processes (Ikura et al., 1992).

In CaM, the EF-hand domains are helix-loop-helix motifs responsible for Calcium binding. The Calcium ions bound to the C-CaM lobe, are coordinated by the carbonyl oxygen of residue Y99 and side chain carbonyl groups of residues D93, D95, and E104 in the EF-hand 3, and by the carbonyl oxygen of residue

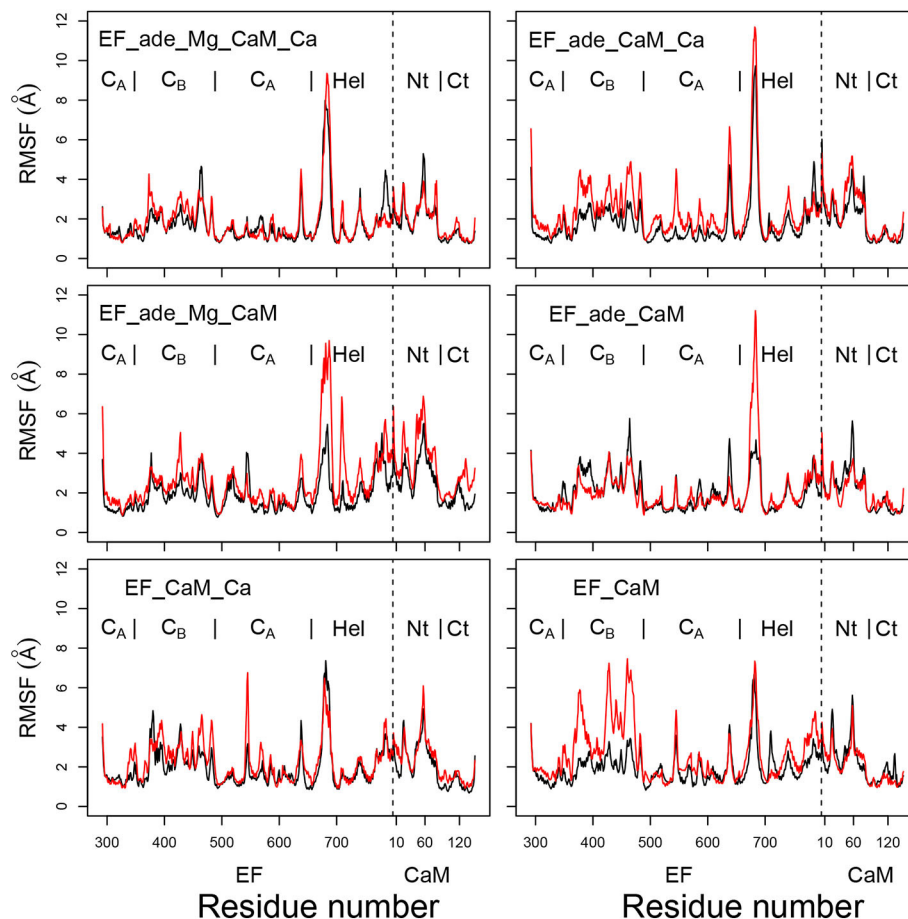


FIGURE 3 | Coordinate root-mean-square fluctuations (RMSF: Å) of the backbone heavy atoms of the complex EF/CaM, each frame being superimposed on the chains A and D of the PDB entry 1PK0. For each trajectory, the black and red curves correspond to replicas R1 and R2. The N terminal residue of CaM is indicated by a vertical dashed line, the EF and CaM regions are labeled on the top of each plot. The N-CaM and C-CaM regions are labeled Nt and Ct for sake of clarity.

Q135 and side chain carbonyl groups of residues D131, D133, and E140 in the EF-hand 4. The average coordination distances reveal that the ions keep similar coordination geometry along all trajectories (**Supplementary Table 1**). Side chain atoms O δ from residues D93, D95 and residues D131, D133 show weaker coordination than the other coordinated residues.

It is well-known from the literature that the presence of Calcium ions has a strong influence on the conformation of the isolated CaM. The angles between the α helices of EF hands in CaM increase, as well as the accessible surfaces of hydrophobic patches, upon Calcium loading (Finn B.E. et al., 1995; Zhang et al., 1995; Yamniuk and Vogel, 2004). In addition, in the absence of Calcium ions, the central α helix is more disordered (Barbato et al., 1992; Komeiji et al., 2002). These variations of CaM conformation in the presence of Calcium ions allows a better interaction of CaM with peptides involved in Calcium signaling (Ikura et al., 1992; Crivici and Ikura, 1995).

Using as definition of the hydrophobic patches the residues previously listed in section 2, the surfaces of hydrophobic patches of N-CaM and C-CaM (**Figure 5**) display quite different trends.

The C-CaM patch, corresponding to EF-hands loaded with ions Ca^{2+} , displays profiles mostly concentrated around 200 Å², with some replica displaying few jumps up to 900 Å². By contrast, the N-CaM patch shows much more diversity spanning a range of 200–600 Å². Similar trends have been observed for accessible surfaces of methionines in previous simulations of the literature (Yang et al., 2004), with a cumulative exposed surface of N-CaM Met of about 26 Å² in apo-form and 88 Å² with Calcium, whereas the C-CaM methionines displayed cumulative exposed surfaces of 45 Å² in apo-form and 124 Å² with Calcium (Table 4 of Yang et al., 2004). The paradoxical behavior of the C-CaM patch which is blocked at a smaller accessible surface value than the N-CaM patch can be explained by the presence of the C_A region of EF which blocks the motions of C-CaM, as C-CaM is inserted between helical domain and C_A. At the contrary, one side of N-CaM interacts with the helical domain, whereas the opposite side of N-CaM is free, which allows more mobility of the N-CaM patch. The variations of hydrophobic patches in the complex EF/CaM are not similar to those described for the isolated CaM. Indeed, in the isolated

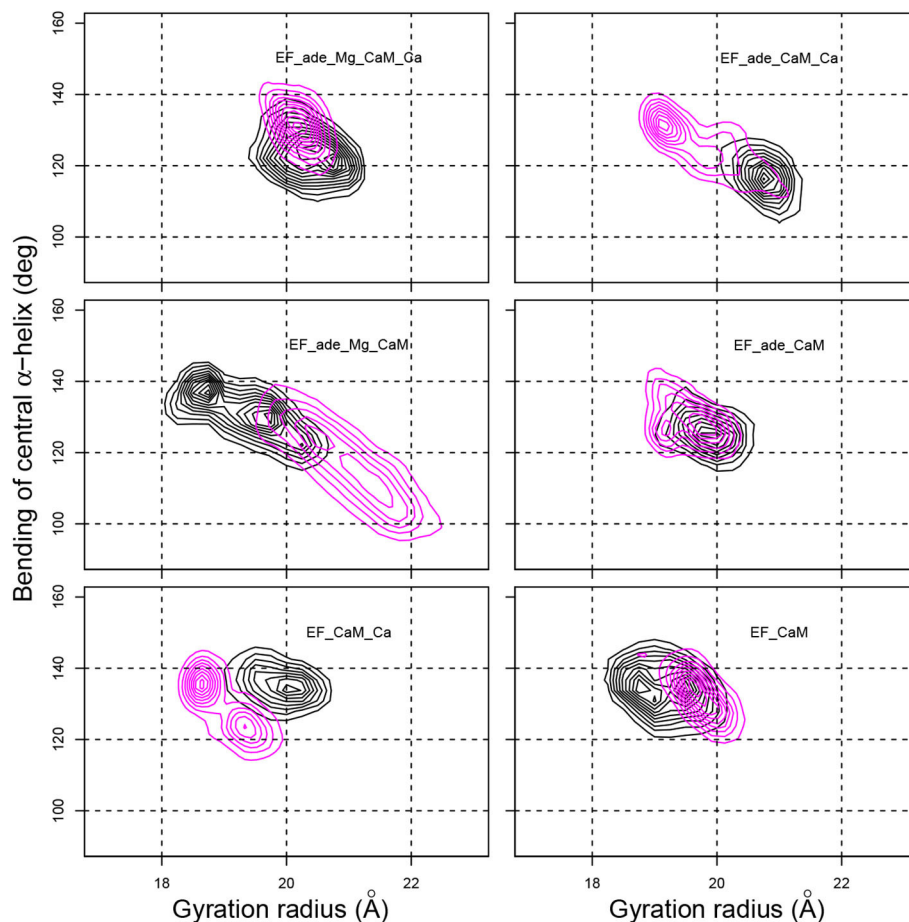


FIGURE 4 | Contour plots describing the variation of the gyration radius (Å) with respect to the bending angle of the CaM central α -helix (deg) determined as the angles between axes of α helices IV (residues 66-73) and V (residues 83-92). The contour lines describe the joint probability distribution of these two parameters. For each trajectory, the black and magenta contours correspond to replicas R1 and R2.

CaM, the variations of hydrophobic patches favor the CaM interaction with signaling peptides, but this aspect is not relevant in the case of EF/CaM interaction.

In isolated CaM, EF hands show a trend to open when CaM is loaded with calcium (Finn B.E. et al., 1995; Zhang et al., 1995). Similarly, in previous MD simulations of the EF/CaM complex (Laine et al., 2008), the C-CaM EF-hands, loaded with Ca^{2+} , display more open α helices than the N-CaM EF-hands. Such behavior is also observed in the present simulations (**Figure 6**). In the present work, the EF-hands 3 and 4, located in C-CaM loaded with Ca^{2+} ions, display quite stable values around, respectively 80 and 90°, corresponding to open conformations. The EF-hand 4 fluctuates slightly more than the EF-hand 3, specially for some trajectories in which Calcium or Magnesium ions are absent (trajectories EF_ade_CaM and EF_ade_Mg_CaM). By contrast, the angles of EF-hands 1 and 2, located in N-CaM, display much wider variations among trajectories. The angle of EF-hand 1 is located in the 40–80° range for most of the trajectories, corresponding to conformations of the hand oscillating between closed and semi-open configurations. The

angle of EF-hand 2 display the largest variations in the range 20–80°. For EF_ade_Mg_CaM trajectories (green curves), the EF-hand 2 explores open conformations with angles larger than 80° and, for one replica of EF_ade_Mg_CaM_Ca (black curves), EF_ade_CaM (yellow curves), and EF_CaM (cyan curves), displays equilibrium between closed and semi-open conformations. These EF-hands are located in the N-CaM lobe, which is not loaded with Calcium ions and interacts also in a much less intricate way with EF as only one side of N-CaM interacts with EF helical domain. For these two reasons, the angles between the two α helices are much more variable between the different conditions of simulations as well as between replicas of a given condition. In addition, the EF-hand 2 shows greater variability than the EF-hand 1, because it is farther away from the helical domain.

The overview of the CaM conformations reveals that the EF hands of N-CaM, which are not loaded with calcium ions, show much more heterogeneity in conformations, and populate conformations corresponding to closed and semi-open EF hand configurations. Otherwise, the overall conformations of CaM as

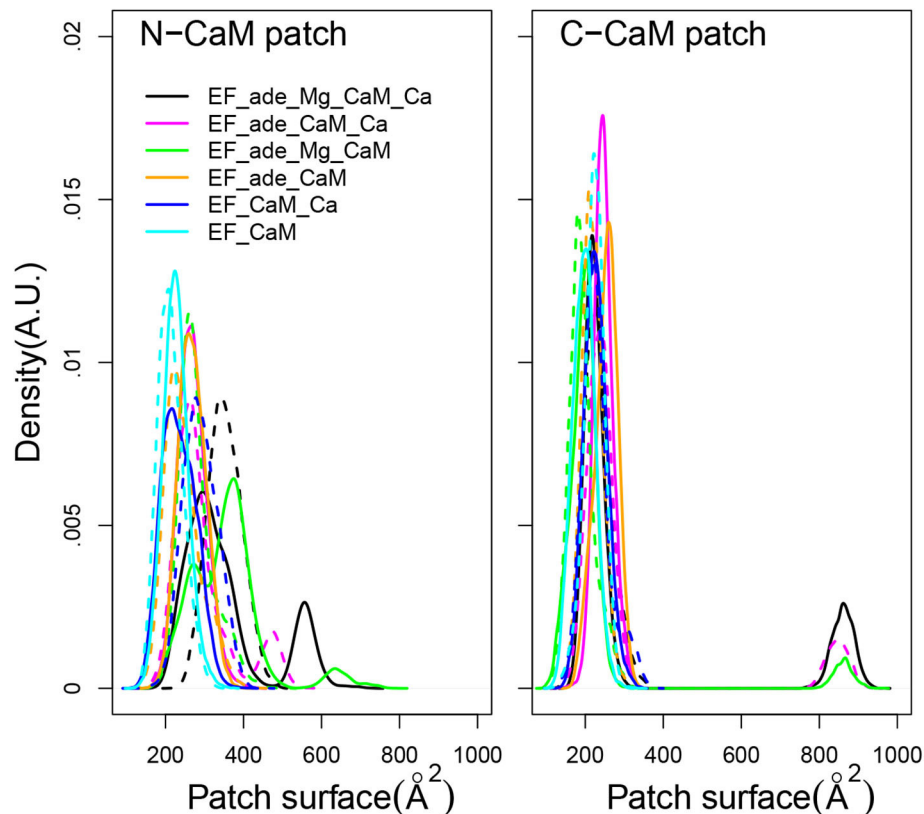


FIGURE 5 | Distribution of accessible surface (\AA^2) of the hydrophobic patch of N-CaM and C-CaM. The N-CaM patch is defined by the residues A10, F12, A15, L18, F19, L32, M36, L39, M51, V55, M71, M72, M76. The C-CaM patch is defined by the residues I85, A88, V91, F92, L105, M109, L112, L116, M124, F141, M144, M145, A147. For each trajectory, the plain and dashed curves correspond to replicas R1 and R2.

well as the coordination of calcium are not strikingly modified among the trajectories recorded here.

A Network of Amino-Acid Interactions Connect the EF Catalytic Site With CaM

The EF activity has been investigated according to the accessible surface of the catalytic site, calculated as the sum of solvent accessible surfaces of residues H351, K353, S354, K372, R329, K346, L348, D491, D493, H577, G578, T579, D582, N583, E588, F586, and T548. This accessible surface varies in the 400–1,200 \AA^2 range (Figure 7, left). This range of values agrees with the average catalytic surfaces of EF previously observed in MD studies (Laine et al., 2008), as: 928 \AA^2 in the complex with 2 Calcium-loaded CaM, 866 \AA^2 in the complex with the 4 Calcium-loaded CaM and 501 \AA^2 in the complex with the apo CaM. Noticeably, in the presence of adefovir (black, yellow, magenta, and green curves) the surfaces are smaller around 400–800 \AA^2 . This corresponds to a more closed catalytic site, in agreement with the inhibitory effect of adefovir. Removal of ion Mg^{2+} or removal of ions Ca^{2+} and Mg^{2+} in the presence of adefovir (magenta and yellow curves) induces a certain shift toward larger values, but the largest shifts toward the 800–1,200 \AA^2 range, is observed if adefovir is removed (cyan and blue curves). The trajectories EF_CaM_Ca

(blue curves) corresponding to the activated EF display the most open catalytic site, which supports the use of accessible surface as an estimator of the EF catalytic activity.

The two following aspects concerning the interface between CaM and EF have been analyzed: (i) the accessible surface of the pocket SABC, (ii) the variation of interactions along a residue network spanning from the catalytic site to CaM.

The SABC pocket, previously used for the virtual screening having conducted to the discovery of thiophen ureoacids (Laine et al., 2010a), is formed by residues, A496, P499, I538, E539, P542, S544, S550, W552, Q553, T579, Q581, L625, Y626, Y627, N629, and N709, belonging to the three switches A, B, and C (Drum et al., 2002). The accessible surface of the SABC pocket (Figure 7, right) varies in a much smaller range of 200–800 \AA^2 than the catalytic pocket. Although the two pockets are defined by a similar number of residues, large differences are nonetheless observed for pocket SABC, sign of significant reorganizations in this region, resulting in some cases in the disappearance of a large accessible surface.

Initial analysis of the X-ray crystallographic structure 1PK0 (Shen et al., 2004b) has revealed a network of hydrogen bonds connecting adefovir and residues from catalytic site with EF residues at the EF/CaM interface and residues from CaM

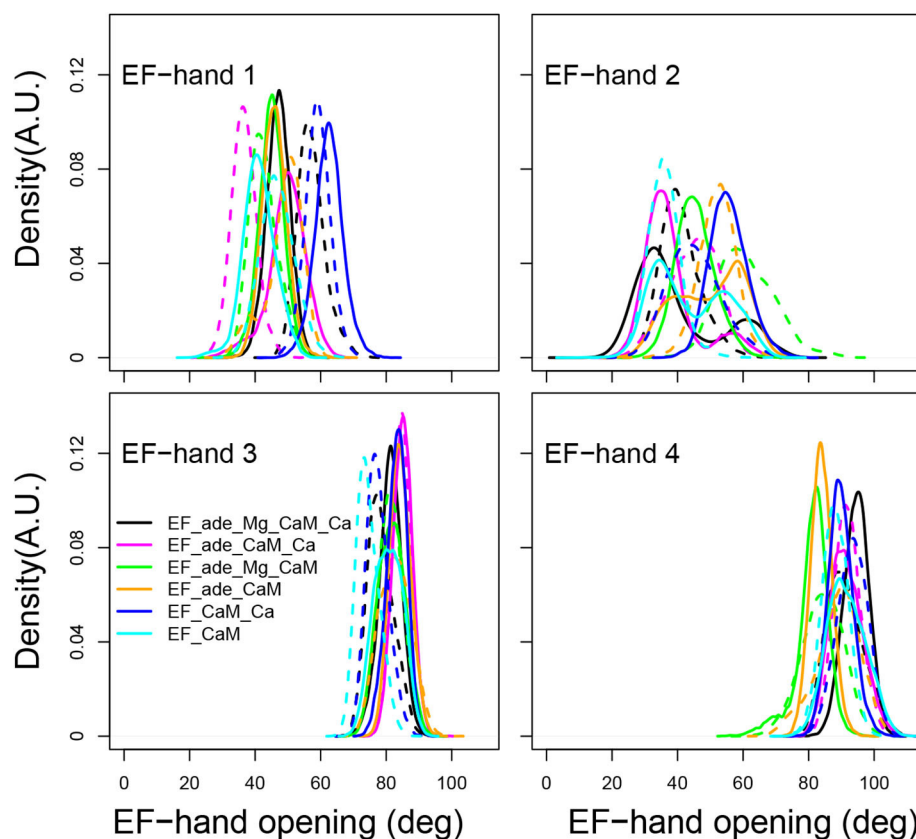


FIGURE 6 | Angle of the EF-hands (deg) calculated between axes of α helices including residues 8-19 (helix I), 31-37 (helix II), 46-53 (helix III), 66-73 (helix IV), 83-92 (helix V), 103-110 (helix VI), 119-127 (helix VII), 139-145 (helix VIII). The angle of EF-hand 1 is the angle between helices I and II. The angle of EF-hand 2 is the angle between helices III and IV. The angle of EF-hand 3 is the angle between helices V and VI. The angle of EF-hand 4 is the angle between helices VII and VIII.

(Figure 1B). This network displays hydrogen bond and stacking interactions which have been monitored along trajectories (Supplementary Table 2). The contacts involve the residues T519, T548, Q553, G578, D582, N583 starting from the catalytic site and expanding to CaM. These residues are located in the switches A (residues 502-551), B (residues 578-591), and C (residues 630-659) (Figure 1B). Interestingly, in the X-ray structures of EF (Drum et al., 2002), these switches undergo a major reorganization between the inactive state and the active state of EF.

Overall, the proportion of formed interactions strongly decreases as soon as ions are removed from the system. In particular, the interactions involving ion Mg^{2+} are strongly reduced. In the initial X-ray crystallographic structure (PDB entry: 1PK0), the ion Yb^{3+} is penta-coordinated by three atoms from adefovir (O1-EMA, P3-EMA, O2-EMA) and two atoms from EF (N ϵ 2-H577, O-Y492). In the MD trajectories, only the contact between Mg^{2+} and N ϵ 2-H577 is stable along the trajectories EF_ade_CaM_Ca and EF_ade_Mg_CaM (Supplementary Table 2) and only three contacts are still present in the trajectory EF_ade_Mg_CaM_Ca: the ones with O2-ade and N ϵ 2-H577 at a significant level and the one with O-Y492 at a negligible frequency. This loss of contacts could be due to the

reduction of the charge and of the van der Waals radius between ions Yb^{3+} and Mg^{2+} .

Among the interactions between adefovir and protein present in the structure 1PK0, only three (hydrogen bonds ade-H6/O-T548, ade-H7/O-T579, and stacking ade/N583) are still present along the trajectory EF_ade_Mg_CaM_Ca. Among them, the stacking between the indole part of adefovir and the aliphatic part of N583 is the only interaction significantly present along all trajectories (Supplementary Table 2). Nevertheless, as soon as the ions are removed from the system, the interaction frequency is reduced in one replica of EF_ade_Mg_CaM (B), EF_ade_CaM_Ca (C), and in all replicas of EF_ade_CaM (D). The adefovir hydrogen bonds conserved with protein mainly involve protons of the amine groups on the indole part. The contacts involving the phosphate group P1 in the structure 1PK0 are completely lost along all trajectories. This destabilization of the adefovir/protein contact and Magnesium contacts as soon as the ion Yb^{3+} is replaced by the ion Mg^{2+} could support an artifactual character of the structure 1PK0, in which the stability of adefovir in the catalytic site was obtained by the presence of the non-biological ion Yb^{3+} .

The hydrogen bond and stacking interactions involving EF and CaM residues are reduced between trajectories

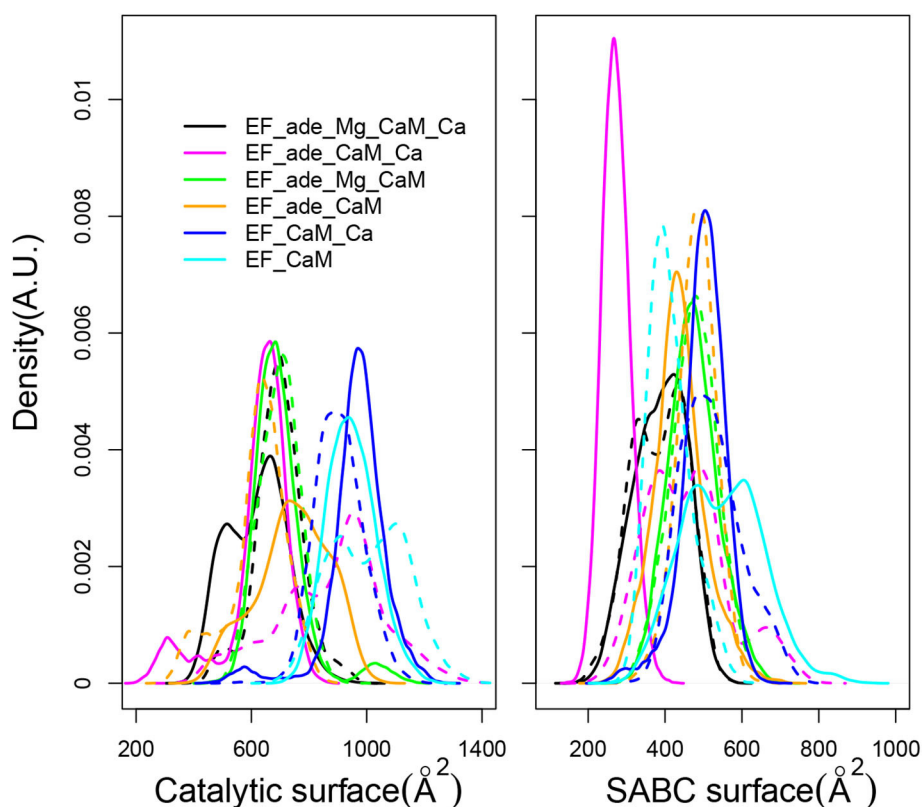


FIGURE 7 | (Left) Distribution of accessible catalytic surface (\AA^2) of EF, calculated as the sum of the accessible surfaces of residues H351, K353, S354, K372, R329, K346, L348, D491, D493, H577, G578, T579, D582, N583, E588, F586, T548 of EF. **(Right)** Distribution of accessible surface (\AA^2) of the pocket SABC, calculated as the sum of the accessible surfaces of residues A496, P499, I538, E539, P542, S544, S550, W552, Q553, T579, Q581, L625, Y626, Y627, N629, and N709 of EF. For each trajectory, the plain and dashed curves correspond to replicas R1 and R2.

EF_ade_Mg_CaM_Ca (A) and EF_ade_Mg_CaM (B) and between trajectories EF_ade_CaM_Ca (C) and EF_ade_CaM (D), when the ions Ca^{2+} are removed from the system. Overall, the removal of Ca^{2+} ions has more influence to reduce the contact stability as the removal of Mg^{2+} . Indeed, six interactions decrease under a formation percentage of 10% if Ca^{2+} ions are removed, whereas no interaction decreases below this percentage in the absence of Mg^{2+} . Interestingly, this influence is visible for hydrogen bonds established between the side chain guanidino group of R630 from EF and the side chain carboxyl groups of residues E84 and E87 from CaM, for the hydrogen bond between the sidechains of R540 from EF and the carboxyl groups of E87 from CaM, and also for the stacking between F628 from EF and R90 from CaM. Noticeably, since the CaM residues E84, E87, and R90 are located at the C terminal part of the α helix V, just before the EF-hand 3 in C-CaM, the presence or absence of Ca^{2+} ions in this lobe has a direct effect on the interaction EF/CaM.

Analysis of the EF/CaM interface, as well as the interactions within the catalytic site, highlight the influence of the variation in the composition of the system on the accessible surface of the catalytic site, impacting the enzymatic activity. In addition, a network of interactions between EF and CaM residues detected in the initial X-ray crystallographic structure, undergo strong

destabilization when co-factors are removed: this network could be considered as a plausible communication path for allosteric correlation between CaM and the EF catalytic site.

Analysis of Cavities Deformation to Detect Allosteric Pockets

In this section, we investigate the possibility to use the cavity tracking implemented in the software *mkgridXf* (Monet et al., 2019) along EF MD trajectories in order to predict allosteric sites in the EF/CaM complex. This approach is motivated by the following reasons. As already quoted in the introduction, several approaches for prediction of allosteric sites are based on a measure of the deformation of the protein described by an elastic network (Panjkovich and Daura, 2014; Guarnera and Berezovsky, 2016) or described by a normal mode perturbation (Greener and Sternberg, 2015). Also, a recent analysis of a large set of protein structures containing ligands showed (Alfayate et al., 2019) that the binding sites of allosteric ligands display larger deformations. Similarly, several bioinformatics approaches predict allosteric pockets as the ones on which ligand binding induces the largest variations in protein structures (Panjkovich and Daura, 2014; Guarnera and Berezovsky, 2019). Allosterism has been thus

repeatedly associated to larger local or global deformability. Beside, cavity tracking of *mkgridXf* (Monet et al., 2019) along MD trajectories made possible to correlate deformations of individual cavities to the principal components of the proteins global motions (Desdouts et al., 2015). The association of the observation made by Desdouts et al. (2015) with the literature approaches conducted us to investigate in the present work on the reliability of analyzing the cavity deformation to predict allosteric sites.

A systematic analysis and tracking of the cavities present in the complex was performed along trajectories using *mkgridXf* (Desdouts et al., 2015; Monet et al., 2019). The cavities were determined by rolling probes as described in section 2. Each cavity was tracked along MD trajectories using a description based on a consensus list of protein atoms delineating the cavities (Monet et al., 2019).

From each trajectory, only proteins EF and CaM have been kept, water molecules, ions and adefovir being removed. One frame every 40 was kept over the time interval 120–200ns of the trajectories EF_ade_CaM_Ca, EF_ade_CaM, EF_ade_Mg_CaM_Ca, EF_ade_Mg_CaM, EF_CaM_Ca and EF_CaM, and then concatenated. The two trajectory replicas series were then analyzed independently with *mkgridXf* in order to probe the reproducibility of the cavity analyses.

The deformations of protein cavities have been monitored through the variations of their volumes. The volumes of each cavity averaged along each trajectory are plotted as points along the cavity index, the points being colored according to the trajectory (**Figure 8**). For most of the cavities, the volumes are smaller than 100\AA^3 for all trajectories. Few of them (indexes #5, #64, #83, #97, #130, #136, #140) display larger volumes up to $2,500\text{\AA}^3$ as well as large variations among the various trajectory conditions. The consensus residues defining each of these cavities have been determined using a cutoff of 2.5 on the residue score and are listed in **Supplementary Table 3**.

For both replicas, the cavity #140, located inside the catalytic site, is among the most variable cavities. Other very variable cavities are located at the interface between different CaM and EF regions. The CaM regions are the EF-hand 1 (#5, #130), the EF-hand 3 (#136), the EF-hand 4 (#97), and the central α -helix (#64). The EF regions are the helical region (#5, #64, #83, #97, #130) and the region C^A (#136). Two cavities located at the interface between the EF-hands and the region C^A (#97 and #136) display smaller volumes. According to the rationale exposed at the beginning of the section and which we explore on the EF/CaM complex, the “variable cavities” cited above and located at the interface between CaM and different EF regions are allosteric cavity candidates.

In order to probe this proposition, we compared the approach proposed here to current approaches in the literature. The initial conformation of the EF/CaM complex was thus processed with different web servers able to predict druggable or allosteric pockets (**Figure 9**). The servers PARS (Panjkovich and Daura, 2014), DeepSite (Jiménez et al., 2017), FTMap (Kozakov et al., 2015), POCASA (Yu et al., 2010), and CavityPlus (Xu et al., 2018) were used. Globally, the *mkgridXf* cavities 140 and 64, located respectively in the catalytic site and at the EF/CaM

interface, correspond to predicted druggable pockets. The cavity 130, located in the helical domain, is present in some of the predicted sets of druggable pockets. Then, the server CavityPlus allows to identify the location of allosteric sites, using the CorrSite method (Ma et al., 2016). Using the catalytic site as the orthosteric site, the corresponding allosteric pockets predicted by CavityPlus correspond to the *mkgridXf* cavities 64, 130, and 83, located at the EF/CaM interface and in the helical domain of EF. To summarize, the set of *mkgridXf* cavities with large volume variation contain druggable pockets, and three of the *mkgridXf* cavities are predicted to be allosteric pockets with respect to the catalytic site. This last result supports the existence of an allosteric communication between the EF/CaM interface and the catalytic site.

As described in the introduction, a theoretical frame for detection of allostery has been developed during the last 5 years: the structure-based statistical mechanical model of allostery (SBSMMA) (Guarnera and Berezovsky, 2019), based on a description of the protein as an elastic network between $C\alpha$. The effect of ligand is implicitly modeled through an additional energetic term added to the harmonic energy of the elastic network. The initial EF/CaM complex conformation has been processed on the AlloSigMA server (allosigma.bii.a-star.edu.sg/home) (Tan et al., 2020) implementing the SBSMMA model. The $\Delta h_i^{(m\uparrow)}$ energy values, describing the allosteric communication between protein residues have been plotted for all residues involved in the *mkgridXf* cavities 5, 64, 83, 97, 130, 136, and 140 (**Figure 10**). The probe residues sample the cavity residues (filled triangles) and, for all cavities except 140, positive $\Delta h_i^{(m\uparrow)}$ values describing a free energy change are observed for several residues located in the catalytic site (filled rectangle). The AlloSigMA results thus agree with an allosteric communication between *mkgridXf* cavities 5, 64, 83, 97, 130, 136, and the catalytic site.

The allosteric communication demonstrated by the variability of cavity volumes as well as by the energetics of the elastic network, is also supported by the variation of the network of interactions described in the previous section between the catalytic site and the helix V of CaM. This network is a plausible candidate for a communication path within the EF/CaM structure.

The cavities present in the EF/CaM complex were tracked along MD trajectories recorded with various perturbation conditions related to functional aspects of EF activity. In that way, the EF/CaM interface has been pointed out as a region containing pockets allosteric with respect to the orthosteric catalytic site. The targeting of these pockets by virtual screening has higher chances to conduct to the discovery of allosteric ligands.

DISCUSSION

In the present work, MD trajectories were recorded from the 1PK0 crystallographic structure of the EF/CaM complex, in the presence of various sets of co-factors: ions Ca^{2+} and Mg^{2+} and ligand adefovir.

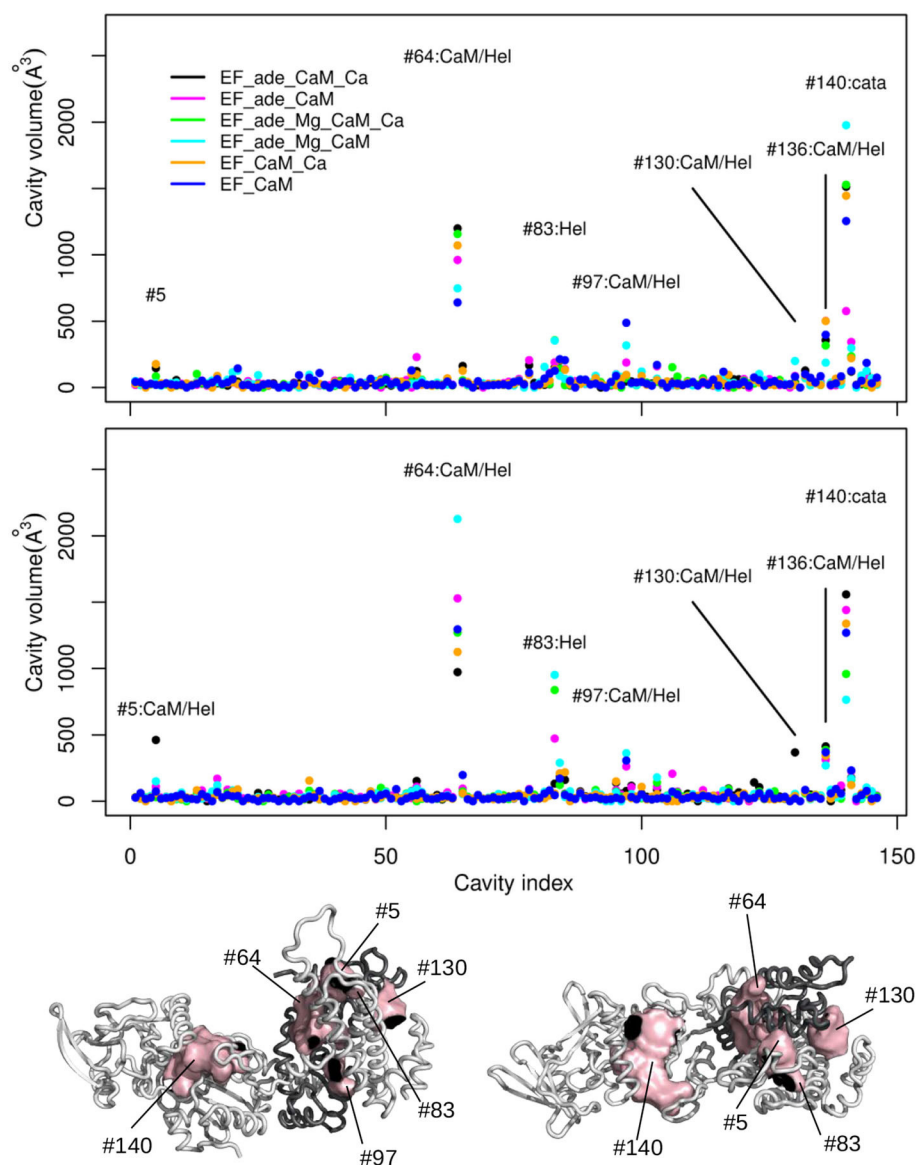


FIGURE 8 | (Top) Averaged volumes (\AA^3) of the cavities detected by *mkgridXf* plotted along the cavity index. The two plots correspond to the two replicas series of the trajectories. The points are colored according to the trajectory on which the volume was averaged: EF_ade_CaM_Ca (black), EF_ade_CaM (magenta), EF_ade_Mg_CaM_Ca (green), EF_ade_Mg_CaM (cyan), EF_CaM_Ca (orange), and EF_CaM (blue). The cavities for which at least one volume larger than 250 \AA^3 has been observed, are labeled with the cavity number and annotated according to the cavity location: Hel, helical domain; cata, catalytic site; Hel/CaM, interaction interface between helical domain and CaM. **(Bottom)** Opposite views of the complex EF/CaM with CaM colored in dark gray. The cavities labeled on the plots are drawn in surfaces and colored in pink. CaM is colored in dark gray.

The main finding from the comparison of trajectories is that the removal of ions has a strong effect on the conformations of the complex, at local and global levels. Indeed, the removal of Mg^{2+} ion destabilizes the interactions between EF and adefovir, but also affects contacts between EF and CaM. Similarly, the removal of Ca^{2+} ions destabilizes the interaction of EF with CaM, but also the geometry of the catalytic site and the EF/adeovir interactions even in the presence of ion Mg^{2+} . This distant influence of ions agrees with the existence of a network of hydrogen bonds and stacking interactions which connects the catalytic pocket with

the EF/CaM interface and which is destabilized if co-factors are removed. Noticeably a similar network of hydrogen bonds has been observed and validated using MD and mutagenesis (Selwa et al., 2014) in the adenylyl cyclase (AC) toxin from *Bordetella pertussis*.

Another observation is that the replacement of the ion Yb^{3+} , observed in the initial X-ray crystallographic structure 1PK0 (Shen et al., 2004b), by the more biologically relevant Mg^{2+} ion induces a destabilization of numerous contacts between adefovir, ion and residues of the catalytic pocket. Consequently,

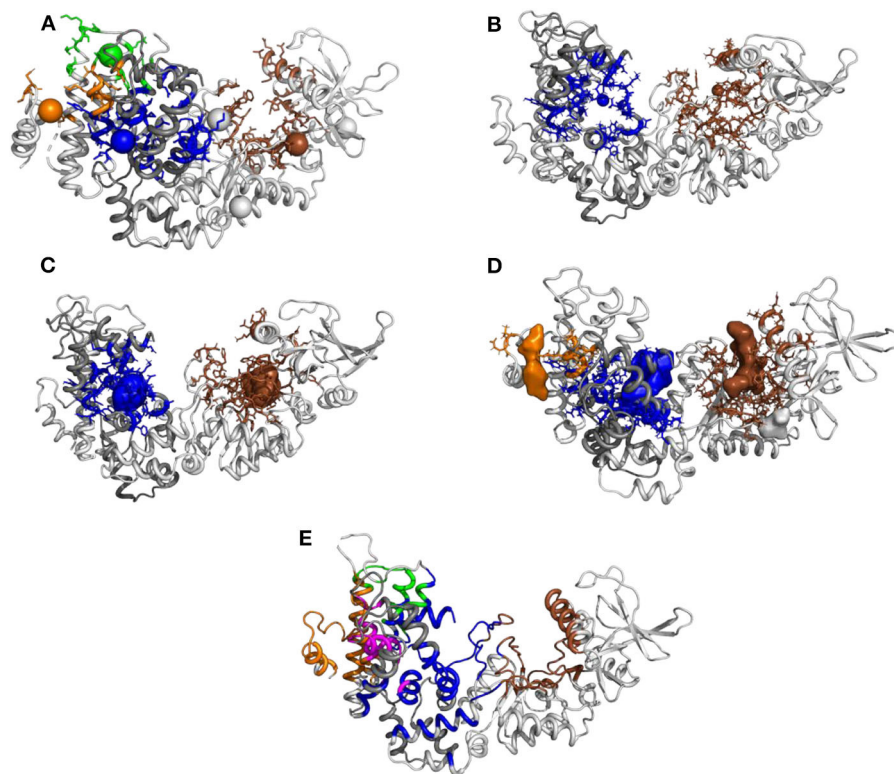


FIGURE 9 | Prediction of pockets druggability or allosterity on the initial EF/CaM structure 1PK0 using several methods: **(A)** PARS (bioinf.uab.cat/cgi-bin/pars-cgi/pars.pl), prediction of the druggable pockets, three of them correspond to *mkgridXf* cavities: 64 (blue), 130 (orange), and 140 (brown). **(B)** DeepSite (www.playmolecule.com/deepsite), prediction of two druggable pockets, displaying scores of 0.999 and corresponding to *mkgridXf* cavities: 64 (blue) and 140 (brown). **(C)** FTMap (ftsites.bu.edu), prediction of two druggable pockets, corresponding to *mkgridXf* cavities: 64 (blue) and 140 (brown). **(D)** POCASA (altair.sci.hokudai.ac.jp/g6/Research/POCASA_e.html), prediction of five druggable pockets, two of them corresponding to the *mkgridXf* cavity 140 (brown), and two others to the *mkgridXf* cavities: 64 (blue) and 130 (orange). **(E)** CavityPlus (www.pkumdl.cn:8000/cavityplus/index.php), prediction of the allosteric pockets influenced by the orthosteric pocket located in the catalytic site (brown). Four allosteric pockets are predicted with the corresponding scores: Cavity1 (3.36), Cavity5 (1.50), Cavity10 (0.75), Cavity4 (0.70). These cavities correspond to the *mkgridXf* cavities: 64 (blue), 130 (orange/magenta), 83 (green).

the establishment of interactions due to the presence of Yb^{3+} ion could have enforced the binding of adefovir inhibitor to the catalytic site or induced the specific conformation of adefovir in the site. Previous computational analyses (Martínez et al., 2009, 2011) already highlighted the artifactual character of some ions observed in the catalytic site of various EF X-ray crystallographic structures.

The analysis of CaM conformations in the EF/CaM revealed that the removal of ions Ca^{2+} induces an unfitting of the conformation of CaM to its position in the complex as it is visible by the CaM central helix bending. The angles of the EF hands also show greater variations in N-CaM than in C-CaM, which could be related to a weaker interaction between EF and N-CaM.

The tracking of *mkgridXf* cavities in the EF/CaM complex along MD trajectories revealed large variations of cavity volumes in two regions: (i) the catalytic site and (ii) the interface between EF and CaM. The analysis of initial EF/CaM complex structure in the frame of the model SBSMMA (Guarnera and Berezovsky, 2019) has shown that the residues located in the *mkgridXf* cavities at the interface between EF and CaM, display allosteric

communication with residues in the catalytic site. The procedure followed in the present manuscript has several points in common with the structure-based statistical mechanical model of allostery (SBSMMA). Indeed, SBSMMA states that the allostery can be quantitatively demonstrated using a model of elastic network which is deformed when the protein undergoes transition between unperturbed and perturbed states. In the present work, the unperturbed state is the MD trajectories recorded on the EF/CaM complex in presence of all co-factors, and the perturbed states are the MD trajectories recorded when one or more co-factors have been removed. In the frame of SBSMMA, the deformable cavities located at the interface between CaM and EF qualify them as being related in an allosteric way to the catalytic site and thus to the catalytic function of EF. Consequently, ligands designed to bind such cavities could have an allosteric effect on the catalytic activity of EF. In that respect, one should note that an inhibitor of EF, the compound 10506-2A, has been claimed to bind to the helical region (Lee et al., 2004).

During the last decade, many approaches have been developed for detecting pockets susceptible to bind allosteric ligands, and

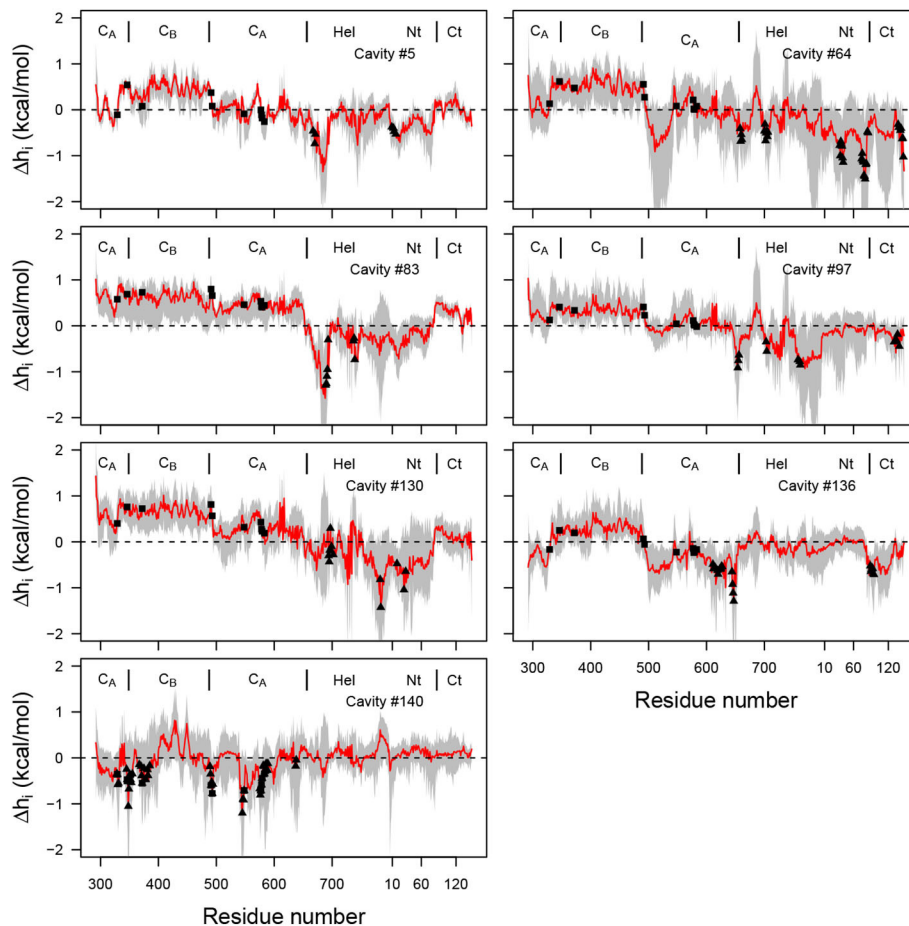


FIGURE 10 | Plot of free energy $\Delta h_i^{(m)}$ (kcal/mol) profile, where the probe residue m samples the residues defining each of the selected cavities (Supplementary Table 3), and i samples all residues of the complex (x-axis). The average free energy profile is plotted in red with range of values drawn in gray. The residues defining the cavity (Supplementary Table 3) are shown with filled triangles, and the residues 329, 346, 372, 577, 491, 493, 548, 578, 579, 583 of the catalytic site are shown with filled rectangles. The dashed line corresponds to the zero energy value.

allosteric paths through protein structures (Daily and Gray, 2009; Mitternacht and Berezovsky, 2011; Bowman and Geissler, 2012; Panjkovich and Daura, 2014; Greener and Sternberg, 2015; Clarke et al., 2016; Guarnera and Berezovsky, 2016; Pflieger et al., 2017; Song et al., 2017; Huang et al., 2018; Abrusan and Marsh, 2019). These approaches are mostly based on a graph description of protein structures. The graphs are then analyzed either from the point of view of protein rigidity and graph theory, or from a more physical point of view of normal mode or elastic network analysis. In the present analysis, we decided to focus on the protein cavities. The relationship found here between the variability of *mkgridXf* cavity volumes and protein long distance communication is not surprising since such correlation has been observed previously (Desdouts et al., 2015) between *mkgridXf* cavities deformation and protein functional motions.

The proposition of using the volume variability of cavities for detection of allosteric communication has been only used here

on the EF/CaM system. It is obvious that the application on a unique system does not constitute a general proof of concept. Nevertheless, use of cavities tracking along MD trajectories presents some advantages with respect to methods based on the modeling of protein via an elastic interaction network (Panjkovich and Daura, 2014; Guarnera and Berezovsky, 2016). Indeed, by contrast with the network where only one atom (generally $\text{C}\alpha$) per residues is included in the calculation, the cavity calculation and tracking take information about all atoms into account, as well as their mutual interactions and their interaction with the solvent and co-factors. Moreover, the model for internal dynamics of the complex is more realistic than the quadratic energy surface of the elastic network model. Finally, perturbing the system by removing co-factors, as ions, highly involved in the EF function, makes the observation of protein deformation more specifically related to the inhibition of EF function. However, all these improvements are accessible at the cost of a very intense computational effort.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

IP, PG, AB, and TM designed the study. PG, AB, and TM wrote the manuscript. IP, DM, and TM recorded the trajectories and performed the analyses. All authors contributed to the article and approved the submitted version.

REFERENCES

- Abrusan, G., and Marsh, J. A. (2019). Ligand-binding-site structure shapes allosteric signal transduction and the evolution of allostery in protein complexes. *Mol. Biol. Evol.* 36, 1711–1727. doi: 10.1093/molbev/msz093
- Aguirre, C., Cala, O., and Krimm, I. (2015). Overview of probing protein-ligand interactions using NMR. *Curr. Protoc. Protein Sci.* 81, 1–17. doi: 10.1002/0471140864.ps1718s81
- Alfayate, A., Rodriguez Caceres, C., Gomes Dos Santos, H., and Bastolla, U. (2019). Predicted dynamical couplings of protein residues characterize catalysis, transport and allostery. *Bioinformatics* 35, 4971–4978. doi: 10.1093/bioinformatics/btz301
- Andersen, H. (1983). Rattle: a “velocity” version of the shake algorithm for molecular dynamics calculations. *J. Comp. Phys.* 52, 24–34. doi: 10.1016/0021-9991(83)90014-1
- Anthis, N. J., Doucleff, M., and Clore, G. M. (2011). Transient, sparsely populated compact states of apo and calcium-loaded calmodulin probed by paramagnetic relaxation enhancement: interplay of conformational selection and induced fit. *J. Am. Chem. Soc.* 133, 18966–18974. doi: 10.1021/ja2082813
- Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W., and Bax, A. (1992). Backbone dynamics of calmodulin studied by ¹⁵N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. *Biochemistry* 31, 5269–5278. doi: 10.1021/bi00138a005
- Beglov, D., Hall, D. R., Wakefield, A. E., Luo, L., Allen, K. N., Kozakov, D., et al. (2018). Exploring the structural origins of cryptic sites on proteins. *Proc. Natl. Acad. Sci. U.S.A.* 115, E3416–E3425. doi: 10.1073/pnas.1711490115
- Bertini, I., Del Bianco, C., Gelis, I., Katsaros, N., Luchinat, C., Parigi, G., et al. (2004). Experimentally exploring the conformational space sampled by domain reorientation in calmodulin. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6841–6846. doi: 10.1073/pnas.0308641101
- Best, R., Zhu, X., Shim, J., Lopes, P., Mittal, J., Feig, M., et al. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi1 and chi2 dihedral angles. *J. Chem. Theor. Comput.* 8, 3257–3273. doi: 10.1021/ct300400x
- Bowman, G. R., and Geissler, P. L. (2012). Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11681–11686. doi: 10.1073/pnas.1209309109
- Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., et al. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66(Pt 1), 12–21. doi: 10.1107/S0907444909042073
- Chen, Z., Moayeri, M., Zhao, H., Crown, D., Leppla, S. H., and Purcell, R. H. (2009). Potent neutralization of anthrax edema toxin by a humanized monoclonal antibody that competes with calmodulin for edema factor binding. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13487–13492. doi: 10.1073/pnas.0906581106
- Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., et al. (2016). CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* 428, 709–719. doi: 10.1016/j.jmb.2016.01.029

ACKNOWLEDGMENTS

IP thanks DGA for Ph.D. financial support (grant 2017033) and Ecole Doctorale ED515 complexité du vivant. Authors thank CNRS and Institut Pasteur for recurrent funding. This work was also supported by the INCEPTION project ANR-16-CONV-0005.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2020.586544/full#supplementary-material>

- Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R. W. F., Chen, J., et al. (2016). Identifying allosteric hotspots with dynamics: application to inter- and intra-species conservation. *Structure* 24, 826–837. doi: 10.1016/j.str.2016.03.008
- Comitani, F., and Gervasio, F. L. (2018). Exploring cryptic pockets formation in targets of pharmaceutical interest with SWISH. *J. Chem. Theory Comput.* 14, 3321–3331. doi: 10.1021/acs.jctc.8b00263
- Crivici, A., and Ikura, M. (1995). Molecular and structural basis of target recognition by calmodulin. *Annu. Rev. Biophys. Biomol. Struct.* 24, 85–116. doi: 10.1146/annurev.bb.24.060195.000505
- Daily, M. D., and Gray, J. J. (2009). Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput. Biol.* 5:e1000293. doi: 10.1371/journal.pcbi.1000293
- Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald and an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 3684–3690. doi: 10.1063/1.464397
- Deredge, D. J., Huang, W., Hui, C., Matsumura, H., Yue, Z., Moënné-Loccoz, P., et al. (2017). Ligand-induced allostery in the interaction of the *Pseudomonas aeruginosa* heme binding protein with heme oxygenase. *Proc. Natl. Acad. Sci. U.S.A.* 114, 3421–3426. doi: 10.1073/pnas.1606931114
- Desdouits, N., Nilges, M., and Blondel, A. (2015). Principal Component Analysis reveals correlation of cavities evolution and functional motions in proteins. *J. Mol. Graph. Model.* 55, 13–24. doi: 10.1016/j.jmgm.2014.10.011
- Drum, C., Shen, Y., Rice, P., Bohm, A., and Tang, W. (2001). Crystallization and preliminary X-ray study of the edema factor exotoxin adenyl cyclase domain from *Bacillus anthracis* in the presence of its activator, calmodulin. *Acta Crystallogr. D Biol. Crystallogr.* 57, 1881–1884. doi: 10.1107/S0907444901014937
- Drum, C., Yan, S., Bard, J., Shen, Y., Lu, D., Soelaiman, S., et al. (2002). Structural basis for the activation of anthrax adenyl cyclase exotoxin by calmodulin. *Nature* 415, 396–402. doi: 10.1038/415396a
- Drum, C., Yan, S., Sarac, R., Mabuchi, Y., Beckingham, K., Bohm, A., et al. (2000). An extended conformation of calmodulin induces interactions between the structural domains of adenyl cyclase from *Bacillus anthracis* to promote catalysis. *J. Biol. Chem.* 275, 36334–36340. doi: 10.1074/jbc.M004778200
- Feng, C., Roy, A., and Post, C. B. (2018). Entropic allostery dominates the phosphorylation-dependent regulation of Syk tyrosine kinase release from immunoreceptor tyrosine-based activation motifs. *Protein Sci.* 27, 1780–1796. doi: 10.1002/pro.3489
- Finn, B. E., Evenäs, J., Drakenberg, T., Waltho, J. P., Thulin, E., and Forsén, S. (1995). Calcium-induced structural changes and domain autonomy in calmodulin. *Nat. Struct. Biol.* 2, 777–783. doi: 10.1038/nsb0995-777
- Fischer, E. S., Park, E., Eck, M. J., and Thomä, N. H. (2016). SPLINTS: small-molecule protein ligand interface stabilizers. *Curr. Opin. Struct. Biol.* 37, 115–122. doi: 10.1016/j.sbi.2016.01.004
- Geduhn, J., Dove, S., Shen, Y., Tang, W. J., König, B., and Seifert, R. (2011). Bis-halogen-anthraniloyl-substituted nucleoside 5'-triphosphates as potent and selective inhibitors of *Bordetella pertussis* adenyl cyclase toxin. *J. Pharmacol. Exp. Ther.* 336, 104–115. doi: 10.1124/jpet.110.174219

- Ghanakota, P., DasGupta, D., and Carlson, H. A. (2019). Free energies and entropies of binding sites identified by MixMD cosolvent simulations. *J. Chem. Inf. Model.* 59, 2035–2045. doi: 10.1021/acs.jcim.8b00925
- Ghosh, A., Gross, L., Tee, W., Guarnera, E., Berezovsky, I., Biondi, R., et al. (2020). Synergistic allostery in multiligand-protein interactions. *Biophys. J.* 119, 1833–1848. doi: 10.1016/j.bpj.2020.09.019
- Goodey, N. M., and Benkovic, S. J. (2008). Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.* 4, 474–482. doi: 10.1038/nchembio.98
- Gowers, R., Linke, M., Barnoud, J., Reddy, T., Melo, M., Seyler, S., et al. (2016). “MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations,” in *Proceedings of the 15th Python in Science Conference, Vol. 32* (Austin, TX), 102–109. doi: 10.25080/Majora-629e541a-00e
- Greener, J. G., and Sternberg, M. J. (2015). AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics* 16:335. doi: 10.1186/s12859-015-0771-1
- Guarnera, E., and Berezovsky, I. N. (2016). Structure-based statistical mechanical model accounts for the causality and energetics of allosteric communication. *PLoS Comput. Biol.* 12:e1004678. doi: 10.1371/journal.pcbi.1004678
- Guarnera, E., and Berezovsky, I. N. (2019). Toward comprehensive allosteric control over protein activity. *Structure* 27, 866–878. doi: 10.1016/j.str.2019.01.014
- Guo, Q., Jureller, J., Warren, J., Solomaha, E., Florián, J., and Tang, W. (2008). Protein-protein docking and analysis reveal that two homologous bacterial adenyl cyclase toxins interact with calmodulin differently. *J. Biol. Chem.* 283, 23836–23845. doi: 10.1074/jbc.M802168200
- Guo, Q., Shen, Y., Zhukovskaya, N. L., Florian, J., and Tang, W. J. (2004). Structural and kinetic analyses of the interaction of anthrax adenyl cyclase toxin with reaction products cAMP and pyrophosphate. *J. Biol. Chem.* 279, 29427–29435. doi: 10.1074/jbc.M402689200
- Gur, M., Madura, J. D., and Bahar, I. (2013). Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase. *Biophys. J.* 105, 1643–1652. doi: 10.1016/j.bpj.2013.07.058
- Huang, M., Song, K., Liu, X., Lu, S., Shen, Q., Wang, R., et al. (2018). AlloFinder: a strategy for allosteric modulator discovery and allosterome analyses. *Nucleic Acids Res.* 46, W451–W458. doi: 10.1093/nar/gky374
- Huang, S. Y. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov. Today* 19, 1081–1096. doi: 10.1016/j.drudis.2014.02.005
- Ikura, M., Clore, G. M., Gronenborn, A. M., Zhu, G., Klee, C. B., and Bax, A. (1992). Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science* 256, 632–638. doi: 10.1126/science.1585175
- Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., and De Fabritiis, G. (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* 33, 3036–3042. doi: 10.1093/bioinformatics/btx350
- Jo, S., Cheng, X., Lee, J., Kim, S., Park, S. J., Patel, D. S., et al. (2017). CHARMM-GUI 10 years for biomolecular modeling and simulation. *J. Comput. Chem.* 38, 1114–1124. doi: 10.1002/jcc.24660
- Jorgensen, W., Chandrasekhar, J., Madura, J., Impey, R., and Klein, M. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi: 10.1063/1.445869
- Kim, S., Lee, J., Jo, S., Brooks, C. L., Lee, H. S., and Im, W. (2017). CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules. *J. Comput. Chem.* 38, 1879–1886. doi: 10.1002/jcc.24829
- Komeiji, Y., Ueno, Y., and Uebayasi, M. (2002). Molecular dynamics simulations revealed Ca(2+)-dependent conformational change of Calmodulin. *FEBS Lett.* 521, 133–139. doi: 10.1016/S0014-5793(02)02853-3
- Koshland, D. E., Némethy, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 5, 365–385. doi: 10.1021/bi00865a047
- Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* 10, 733–755. doi: 10.1038/nprot.2015.043
- Kuenemann, M. A., Sperandio, O., Labbé, C. M., Lagorce, D., Miteva, M. A., and Villoutreix, B. O. (2015). *In silico* design of low molecular weight protein-protein interaction inhibitors: overall concept and recent advances. *Prog. Biophys. Mol. Biol.* 119, 20–32. doi: 10.1016/j.pbmolbio.2015.02.006
- Laine, E., Blondel, A., and Malliavin, T. (2009). Dynamics and energetics: a consensus analysis of the impact of calcium on EF-CaM protein complex. *Biophys. J.* 96, 1249–1263. doi: 10.1016/j.bpj.2008.10.055
- Laine, E., Goncalves, C., Karst, J., Lesnard, A., Rault, S., Tang, W., et al. (2010a). Use of allostery to identify inhibitors of calmodulin- induced activation of *Bacillus anthracis* Edema Factor. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11277–11282. doi: 10.1073/pnas.0914611107
- Laine, E., Ladant, L. M. D., Malliavin, T., and Blondel, A. (2012). Molecular motions as a drug target: mechanistic simulations of anthrax toxin edema factor function led to the discovery of novel allosteric inhibitors. *Toxins* 4, 580–604. doi: 10.3390/toxins4080580
- Laine, E., Martínez, L., Blondel, A., and Malliavin, T. (2010b). Activation of the edema factor of *Bacillus anthracis* by calmodulin: evidence of an interplay between the EF-calmodulin interaction and calcium binding. *Biophys. J.* 99, 2264–2272. doi: 10.1016/j.bpj.2010.07.044
- Laine, E., Yoneda, J., Blondel, A., and Malliavin, T. (2008). The conformational plasticity of calmodulin upon calcium complexation gives a model of its interaction with the oedema factor of *Bacillus anthracis*. *Proteins* 71, 1813–1829. doi: 10.1002/prot.21862
- Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400. doi: 10.1016/0022-2836(71)90324-X
- Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., et al. (2016). CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.* 12, 405–413. doi: 10.1021/acs.jctc.5b00935
- Lee, Y. S., Bergson, P., He, W. S., Mrksich, M., and Tang, W. J. (2004). Discovery of a small molecule that inhibits the interaction of anthrax edema factor with its cellular activator, calmodulin. *Chem. Biol.* 11, 1139–1146. doi: 10.1016/j.chembiol.2004.05.020
- Liu, J., and Nussinov, R. (2016). Allostery: an overview of its history, concepts, methods, and applications. *PLoS Comput. Biol.* 12:e1004966. doi: 10.1371/journal.pcbi.1004966
- Ma, X., Meng, H., and Lai, L. (2016). Motions of allosteric and orthosteric ligand-binding sites in proteins are highly correlated. *J. Chem. Inf. Model.* 56, 1725–1733. doi: 10.1021/acs.jcim.6b00039
- MacKerell, A., Bashford, D., Bellott, M., Dunbrack, R., Evanseck, J., Field, M., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102, 3586–3616. doi: 10.1021/jp973084f
- MacKerell, A., Feig, M., and Brooks, C. (2004). Extending the treatment of backbone energetics in protein force fields and limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comp. Chem.* 25, 1400–1415. doi: 10.1002/jcc.20065
- Martinez, L., Laine, E., Malliavin, T., Nilges, M., and Blondel, A. (2009). ATP conformations and ion binding modes in the active site of anthrax edema factor: a computational analysis. *Proteins* 77, 971–983. doi: 10.1002/prot.22523
- Martinez, L., Malliavin, T., and Blondel, A. (2011). Mechanism of reactant and product dissociation from the anthrax edema factor: a locally enhanced sampling and steered molecular dynamics study. *Proteins* 79, 1649–1661. doi: 10.1002/prot.22991
- Martinez-Rosell, G., Lovera, S., Sands, Z. A., and De Fabritiis, G. (2020). PlayMolecule crypticscout: predicting protein cryptic sites using mixed-solvent molecular simulations. *J. Chem. Inf. Model.* 60, 2314–2324. doi: 10.1021/acs.jcim.9b01209
- Michaud-Agrawal, N., Denning, E., Woolf, T., and Beckstein, O. (2011). MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* 32, 2319–2327. doi: 10.1002/jcc.21787
- Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculation. *F1000Research* 5:189. doi: 10.12688/f1000research.7931.1
- Mitternacht, S., and Berezovsky, I. N. (2011). Binding leverage as a molecular basis for allosteric regulation. *PLoS Comput. Biol.* 7:e1002148. doi: 10.1371/journal.pcbi.1002148
- Monet, D., Desdoutis, N., Nilges, M., and Blondel, A. (2019). mkgridXf: consistent identification of plausible binding sites despite the elusive nature of cavities

- and grooves in protein dynamics. *J. Chem. Inf. Model.* 59, 3506–3518. doi: 10.1021/acs.jcim.9b00103
- Monod, J., Wyman, J., and Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* 12, 88–118. doi: 10.1016/S0022-2836(65)80285-6
- Morelli, X., Bourgeois, R., and Roche, P. (2011). Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P21). *Curr. Opin. Chem. Biol.* 15, 475–481. doi: 10.1016/j.cbpa.2011.05.024
- Ni, D., Liu, N., and Sheng, C. (2019). Allosteric modulators of protein-protein interactions (PPIs). *Adv. Exp. Med. Biol.* 1163, 313–334. doi: 10.1007/978-981-13-8719-7_13
- Nussinov, R., and Tsai, C. J. (2013). Allostery in disease and in drug discovery. *Cell* 153, 293–305. doi: 10.1016/j.cell.2013.03.034
- Panjikovich, A., and Daura, X. (2014). PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics* 30, 1314–1315. doi: 10.1093/bioinformatics/btu002
- Pfleger, C., Mingos, A., Boehm, M., McClendon, C. L., Torella, R., and Gohlke, H. (2017). Ensemble- and rigidity theory-based perturbation approach to analyze dynamic allostery. *J. Chem. Theory Comput.* 13, 6343–6357. doi: 10.1021/acs.jctc.7b00529
- Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802. doi: 10.1002/jcc.20289
- Raman, A. S., White, K. I., and Ranganathan, R. (2016). Origins of allostery and evolvability in proteins: a case study. *Cell* 166, 468–480. doi: 10.1016/j.cell.2016.05.047
- Roe, D., and TE Cheatham, I. (2013). PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* 9, 3084–3095. doi: 10.1021/ct400341p
- Ryckaert, J., Ciccotti, G., and Berendsen, H. (1977). Numerical integration of the cartesian equations of motion of a system with constraints and Molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341. doi: 10.1016/0021-9991(77)90098-5
- Selwa, E., Davi, M., Chenal, A., Sotomayor-Pérez, A., Ladant, D., and Malliavin, T. (2014). Allosteric activation of *Bordetella pertussis* adenylyl cyclase by calmodulin: molecular dynamics and mutagenesis studies. *J. Biol. Chem.* 289, 21131–21141. doi: 10.1074/jbc.M113.530410
- Shen, Y., Guo, Q., Zhukovskaya, N. L., Drum, C. L., Bohm, A., and Tang, W. J. (2004a). Structure of anthrax edema factor-calmodulin-adenosine 5'-(alpha,beta-methylene)-triphosphate complex reveals an alternative mode of ATP binding to the catalytic site. *Biochem. Biophys. Res. Commun.* 317, 309–314. doi: 10.1016/j.bbrc.2004.03.046
- Shen, Y., Lee, Y., Soelaiman, S., Bergson, P., Lu, D., Chen, A., et al. (2002). Physiological calcium concentrations regulate calmodulin binding and catalysis of adenylyl cyclase exotoxins. *EMBO J.* 21, 6721–6732. doi: 10.1093/emboj/cdf681
- Shen, Y., Zhukovskaya, N. L., Guo, Q., Florian, J., and Tang, W. J. (2005). Calcium-independent calmodulin binding and two-metal-ion catalytic mechanism of anthrax edema factor. *EMBO J.* 24, 929–941. doi: 10.1038/sj.emboj.7600574
- Shen, Y., Zhukovskaya, N. L., Zimmer, M. I., Soelaiman, S., Bergson, P., Wang, C. R., et al. (2004b). Selective inhibition of anthrax edema factor by adefovir, a drug for chronic hepatitis B virus infection. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3242–3247. doi: 10.1073/pnas.030652101
- Shin, W. H., Christoffer, C. W., and Kihara, D. (2017). *In silico* structure-based approaches to discover protein-protein interaction-targeting drugs. *Methods* 131, 22–32. doi: 10.1016/j.ymeth.2017.08.006
- Soelaiman, S., Wei, B. Q., Bergson, P., Lee, Y. S., Shen, Y., Mrksich, M., et al. (2003). Structure-based inhibitor discovery against adenylyl cyclase toxins from pathogenic bacteria that cause anthrax and whooping cough. *J. Biol. Chem.* 278, 25990–25997. doi: 10.1074/jbc.M301232200
- Song, K., Liu, X., Huang, W., Lu, S., Shen, Q., Zhang, L., et al. (2017). Improved method for the identification and validation of allosteric sites. *J. Chem. Inf. Model.* 57, 2358–2363. doi: 10.1021/acs.jcim.7b00014
- Sun, Z., Wakefield, A. E., Kolossvary, I., Beglov, D., and Vajda, S. (2020). Structure-based analysis of cryptic-site opening. *Structure* 28, 223–235. doi: 10.1016/j.str.2019.11.007
- Taha, H., Dove, S., Geduhn, J., König, B., Shen, Y., Tang, W. J., et al. (2012). Inhibition of the adenylyl cyclase toxin, edema factor, from *Bacillus anthracis* by a series of 18 mono- and bis-(M)ANT-substituted nucleoside 5'-triphosphates. *Naunyn Schmiedeberg's Arch. Pharmacol.* 385, 57–68. doi: 10.1007/s00210-011-0688-9
- Taha, H. M., Schmidt, J., Gottle, M., Suryanarayana, S., Shen, Y., Tang, W. J., et al. (2009). Molecular analysis of the interaction of anthrax adenylyl cyclase toxin, edema factor, with 2'(3')-O-(N-(methyl)anthraniloyle)-substituted purine and pyrimidine nucleotides. *Mol. Pharmacol.* 75, 693–703. doi: 10.1124/mol.108.052340
- Tan, Z. W., Guarnera, E., Tee, W. V., and Berezovsky, I. N. (2020). AlloSigMA 2: paving the way to designing allosteric effectors and to exploring allosteric effects of mutations. *Nucleic Acids Res.* 48, W116–W124. doi: 10.1093/nar/gkaa338
- Tschammer, N. (2016). Allosteric modulators of the class A G protein coupled receptors. *Adv. Exp. Med. Biol.* 917, 185–207. doi: 10.1007/978-3-319-32805-8_9
- Tuffery, P., and Derreumaux, P. (2012). Flexibility and binding affinity in protein-ligand, protein-protein and multi-component protein interactions: limitations of current computational approaches. *J. R. Soc. Interface* 9, 20–33. doi: 10.1098/rsif.2011.0584
- Ulmer, T., Soelaiman, S., Li, S., Klee, C., Tang, W., and Bax, A. (2003). Calcium dependence of the interaction between calmodulin and anthrax edema factor. *J. Biol. Chem.* 278, 29261–29266. doi: 10.1074/jbc.M302837200
- Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M., and Whitty, A. (2018). Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.* 44, 1–8. doi: 10.1016/j.cbpa.2018.05.003
- van den Bedem, H., Bhabha, G., Yang, K., Wright, P. E., and Fraser, J. S. (2013). Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat. Methods* 10, 896–902. doi: 10.1038/nmeth.2592
- Wodak, S. J., Paci, E., Dokholyan, N. V., Berezovsky, I. N., Horovitz, A., Li, J., et al. (2019). Allostery in its many disguises: from theory to applications. *Structure* 27, 566–578. doi: 10.1016/j.str.2019.01.003
- Xu, Y., Wang, S., Hu, Q., Gao, S., Ma, X., Zhang, W., et al. (2018). CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res.* 46, W374–W379. doi: 10.1093/nar/gky380
- Yamniuk, A. P., and Vogel, H. J. (2004). Calmodulin's flexibility allows for promiscuity in its interactions with target proteins and peptides. *Mol. Biotechnol.* 27, 33–57. doi: 10.1385/MB:27:1:33
- Yang, C., Jas, G., and Kuczer, K. (2004). Structure, dynamics and interaction with kinase targets: computer simulations of calmodulin. *Biochim. Biophys. Acta* 1697, 289–300. doi: 10.1016/j.bbapap.2003.11.032
- Yu, J., Zhou, Y., Tanaka, I., and Yao, M. (2010). Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* 26, 46–52. doi: 10.1093/bioinformatics/btp599
- Zhang, J., and Nussinov, R. (ed.). (2019). "Protein allostery in drug discovery," in *Advances in Experimental Medicine and Biology*, eds Zhang and Nussinov (Springer). doi: 10.1007/978-981-13-8719-7. Available online at: <https://www.springer.com/gp/book/9789811387180>
- Zhang, M., Tanaka, T., and Ikura, M. (1995). Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. *Nat. Struct. Biol.* 2, 758–767. doi: 10.1038/nsb0995-758
- Zhang, X., Betzi, S., Morelli, X., and Roche, P. (2014). Focused chemical libraries-design and enrichment: an example of protein-protein interaction chemical space. *Future Med. Chem.* 6, 1291–1307. doi: 10.4155/fmc.14.57

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

Copyright © 2020 Pitard, Monet, Goossens, Blondel and Malliavin. 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.



Perspectives on High-Throughput Ligand/Protein Docking With Martini MD Simulations

Paulo C. T. Souza^{1,2,3*}, Vittorio Limongelli^{4,5*}, Sangwook Wu^{2,6*}, Siewert J. Marrink^{1*} and Luca Monticelli^{3*}

¹ Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, Groningen, Netherlands, ² PharmCADD, Busan, South Korea, ³ Molecular Microbiology and Structural Biochemistry (MMSB, UMR 5086), CNRS, University of Lyon, Lyon, France, ⁴ Faculty of Biomedical Sciences, Institute of Computational Science, Università della Svizzera Italiana (USI), Lugano, Switzerland, ⁵ Department of Pharmacy, University of Naples "Federico II", Naples, Italy, ⁶ Department of Physics, Pukyong National University, Busan, South Korea

OPEN ACCESS

Edited by:

Agnel Praveen Joseph,
Science and Technology Facilities
Council, United Kingdom

Reviewed by:

Valeria Losasso,
United Kingdom Research
and Innovation, United Kingdom
Sophie Sacquin-Mora,
UPR 9080 Laboratoire de Biochimie
Théorique (LBT), France

*Correspondence:

Paulo C. T. Souza
paulo.telles-de-souza@ibcp.fr
Sangwook Wu
s.wu@pharmacadd.com
Vittorio Limongelli
vittoriolimongelli@gmail.com
Siewert J. Marrink
s.j.marrink@rug.nl
Luca Monticelli
luca.monticelli@inserm.fr

Specialty section:

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

Received: 22 January 2021

Accepted: 05 March 2021

Published: 29 March 2021

Citation:

Souza PCT, Limongelli V, Wu S,
Marrink SJ and Monticelli L (2021)
Perspectives on High-Throughput
Ligand/Protein Docking With Martini
MD Simulations.
Front. Mol. Biosci. 8:657222.
doi: 10.3389/fmolb.2021.657222

Molecular docking is central to rational drug design. Current docking techniques suffer, however, from limitations in protein flexibility and solvation models and by the use of simplified scoring functions. All-atom molecular dynamics simulations, on the other hand, feature a realistic representation of protein flexibility and solvent, but require knowledge of the binding site. Recently we showed that coarse-grained molecular dynamics simulations, based on the most recent version of the Martini force field, can be used to predict protein/ligand binding sites and pathways, without requiring any *a priori* information, and offer a level of accuracy approaching all-atom simulations. Given the excellent computational efficiency of Martini, this opens the way to high-throughput drug screening based on dynamic docking pipelines. In this opinion article, we sketch the roadmap to achieve this goal.

Keywords: molecular dynamics, coarse-grain, ligand-protein, protein-protein interaction, Martini, dynamic docking, high-throughput screening, drug design

INTRODUCTION

Structure-based drug design has been extensively used by pharmaceutical companies and academic research groups to reduce the cost and time necessary for the discovery of new drugs. The approach relies on the knowledge of the atomistic structure of the biological target, obtained by experiments (e.g., X-ray crystallography, NMR spectroscopy, cryo-electron microscopy) or modeling (e.g., based on homology). Standard pipelines often start with *in silico* docking experiments, used for virtual screening of thousands of compounds or molecular fragments (Sliwoski et al., 2014; Leelananda and Lindert, 2016; Duarte et al., 2019). After a significant reduction of the chemical space, a selected group of molecules can be optimized by all-atom (AA) molecular dynamics (MD) based simulations (Jorgensen and Thomas, 2008; Jorgensen, 2009; Vivo et al., 2016; Limongelli, 2020). AA MD simulations can be used not only to improve the prediction of the binding pose and affinity, but also to get insight into (un)binding rates and pathways (Dror et al., 2011; Shan et al., 2011; Limongelli et al., 2013; Tiwary et al., 2015; Copeland, 2016; Bruce et al., 2018). As a third step, further selection can be performed considering predictions of absorption, distribution, metabolism, excretion and toxicity (ADMET) (Van De Waterbeemd and Gifford, 2003; Cheng et al., 2013). The obtained lead compounds need to be validated by *in vitro* assays and structurally improved – lead

optimization – to achieve drug candidates, which are tested in animal models and eventually enter clinical trials before final approval. Despite the rapid advances in computer-aided drug discovery methods, the limitations of such approaches are still major. Docking assays remain to date the first option in the drug discovery pipeline thanks to their capability of “virtually” testing thousands of molecules in a short time. However, docking accuracy is poor due to limitations in simplified energy (“scoring”) functions, sampling ligand and protein flexibility (Grinter and Zou, 2014), and representation of the environment – crucial in hydrated binding pockets and in transmembrane proteins, that represent a large fraction of the pharmaceutically relevant protein targets. AA MD simulations can tackle these limitations, but they are still computationally prohibitively expensive (Durrant and McCammon, 2011; Liu et al., 2018; Miller et al., 2020), due to relatively long time scales of conformational dynamics in proteins. Moreover, predictions of dissociation pathways and rates are extremely challenging, and require high performance computing and enhanced sampling techniques (Limongelli et al., 2013; Casasnovas et al., 2017; Brotzakis et al., 2019; Schuetz et al., 2019). Peptide and protein design for biopharmaceutical applications have similar pitfalls, with the current approaches reasonably successful in predicting protein structures (Hutson, 2019; Callaway, 2020) and rigid-body protein-protein interactions (Siebenmorgen and Zacharias, 2020) but with limitations in the design of conformational changes (Feldmeier and Höcker, 2013; Yang and Lai, 2017; Perkel, 2019; D’Annessa et al., 2020).

Coarse-grained (CG) modeling is a computationally cheaper alternative to high-resolution atomistic approaches (Ingólfsson et al., 2014; Kmiecik et al., 2016), as it reduces the computational cost by grouping atoms into effective interaction sites. Numerous CG models have been developed during the past two decades, with different levels of coarsening and different mathematical representations. CG models have been successfully applied to study a large range of processes in biology (Yen et al., 2018; Bruininks et al., 2020; Lucendo et al., 2020) and materials science (Casalini et al., 2019; Alessandri et al., 2020; Li et al., 2020; Vazquez-Salazar et al., 2020). Applications such as structure-based drug design are particularly challenging for CG modeling because of the severe requirements: (1) high chemical specificity (i.e., allowing to distinguish most chemical groups); (2) capability to represent all possible components of the system (proteins, cofactors, nucleic acids, drug candidates, waters, lipids, etc.) in a coherent way; (3) realistic representation of conformational flexibility of each molecule in the system; and (4) accurate thermodynamics and kinetics of binding. Currently, none of the CG force fields available fulfills all the requirements above, but the Martini CG force field fulfills at least some (Marrink et al., 2007; Marrink and Tieleman, 2013), as it allows modeling all main biomolecules (Monticelli et al., 2008; López et al., 2009; de Jong et al., 2013, 2015; Uusitalo et al., 2015, 2017; Wassenaar et al., 2015) with relatively high chemical specificity, and proteins may still retain reasonable conformational flexibility (Periole et al., 2009; Melo et al., 2017; Poma et al., 2017). As in AA MD simulations, most of the details of the environment can be included in

Martini CG simulations, for instance an explicit solvent model or a complex bilayer composition (Ingólfsson et al., 2015; Marrink et al., 2019).

Although Martini-based CG MD simulations have been used to study a wide range of biomolecular processes, examples of protein–ligand binding are still scarce (Negami et al., 2014, 2020; Delort et al., 2017; Ferré et al., 2019; Jiang and Zhang, 2019; Dandekar and Mondal, 2020). Studies of protein–protein interactions are more common, although usually restricted to membrane environments (Baaden and Marrink, 2013; Castillo et al., 2013; Lelimosin et al., 2016; Sun et al., 2020). In some cases, binding of lipids to sites deeply buried inside the protein can be obtained by brute force Martini MD (Arnarez et al., 2013; Van Eerden et al., 2017; Corradi et al., 2019). Overall, some limiting factors hampered the use of Martini in small-molecule and protein design: (1) chemical specificity to reproduce the broad chemical space of drugs; (2) the thermodynamics of ligand–protein and protein–protein interactions are generally overestimated (Stark et al., 2013; Javanainen et al., 2017; Alessandri et al., 2019); and (3) introduction of conformational flexibility in proteins requires case-by-case optimization (Negami et al., 2020; Ahalawat and Mondal, 2021). A new version of the Martini force field, named Martini 3 (Souza et al., 2021), partly solves these issues: it can represent a broader variety of chemical compounds, and it features improved molecular packing and optimized molecular interactions (along with specific interactions mimicking H-bonding and electronic polarizability). Recently, Martini 3 was successfully applied to a range of protein–ligand system examples, from the well-characterized T4 lysozyme to members of the GPCR family and nuclear receptors to a variety of enzymes (Souza et al., 2020). In addition, combination of Martini 3 and Gō-like potentials can substantially improve the modeling of protein flexibility (Poma et al., 2017; Souza et al., 2019). Combined, these new features open the possibility of computer-aided drug design based on CG models.

In this perspective, we sketch a possible roadmap for a drug design pipeline using Martini, where no *a priori* information about the target pocket is necessary. Competition between ligands for different pockets and environments can be included in the screening. Protein flexibility can be incorporated to a certain degree, allowing the possible discovery of cryptic (hidden) pockets (Kuzmanic et al., 2020). Ligand (un)binding pathways are accessible via enhanced sampling techniques (Raniolo and Limongelli, 2020), and enable for the first time the possibility of a “dynamic” drug screening based not only on ligand binding modes, but also on kinetically relevant states – that is considering binding affinity and dissociation rates (i.e., drug residence time). The next sections detail the key steps of this pipeline.

LIGAND DATABASES: COARSE-GRAINING THE LIGANDS

The very first step to develop a Martini drug design pipeline is to create curated and validated databases containing hundreds

to thousands of small-molecule models. This CG database needs to include molecular moieties usually found in drugs, such as halogens, heterocycles, and sulfamides. Alternatively, the databases of low-molecular-weight molecules (~ 150 Da) can also be created for fragment-based drug discovery campaigns (Rognan, 2012). Parameters for molecules/fragments of pharmaceutical interest need to be validated by comparison between CG, AA and, if available, experimental data for a subset of relevant target systems. Once validated, all the models will be made available via the open-access Martini Database (MAD) web server¹. The initial CG databases are also the foundation to develop and calibrate automatic tools to generate parameters for new CG models. Such automatic tools should perform AA to CG mapping [as performed by auto-martini (Bereau and Kremer, 2015)], bead assignment (i.e., the choice of the CG interaction parameters), and determination of the bonded parameters [as PyCGTOOL (Graham et al., 2017) or Swarm-CG (Empereur-mot et al., 2020)], allowing further coverage of chemical space. The creation of accurate databases and integration of automatic tools is currently one of the main bottlenecks hampering high-throughput screening with Martini.

VIRTUAL SCREENING: MARTINI DYNAMIC DOCKING

Virtual screening is the core of the drug design pipeline, and usually relies on docking algorithms. The use of Martini CG models will enable a new approach: dynamic docking with no *a priori* knowledge of the binding pocket in the target structure. The concept here is to sample protein–ligand interactions with CG MD simulations, which is around 300 to 1,000 times faster than atomistic MD (Souza et al., 2020). A practical example of such speed up can be given for propranolol binding to β_2 adrenergic receptor, which has been simulated in atomistic (Dror et al., 2011) and coarse-grained (Souza et al., 2020) resolution. Atomistic simulations showed one binding event every 11.9 μ s, which for a single simulation would take 84 days of computing time (using the 4 CPUs and 1 GPU in a computer/conditions described in the performance tests of Souza et al., 2020). The same system in CG simulations showed roughly the same number of binding events per μ s (considering a normalization based in the different concentration of ligands), and would take 2 to 7 h of computing time, on the same hardware. We remark that a fair comparison between coarse-grained and atomistic simulation time is not trivial, since this should consider the different simulation conditions (e.g., ligand concentration) and parameters (Souza et al., 2020).

Multiple strategies are possible to accelerate sampling even more, with different computational costs and different levels of sophistication. Unbiased MD simulations could be applied in certain cases, to obtain not only binding poses but also estimates of binding affinities, as recently demonstrated for

T4 lysozyme (Souza et al., 2020). However, for a general approach to virtual screening, faster methods are necessary. One possibility is to combine CG models with enhanced sampling techniques that do not depend on prior knowledge of the binding pathways. Examples are Gaussian accelerated molecular dynamics (GaMD) (Miao et al., 2015; Pang et al., 2017), and Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) (Wang et al., 2013; Luitz and Zacharias, 2014). Computational performance can be straightforwardly increased by optimizing ligand concentration, to increase the probability of binding. The approach was already tested with atomistic simulations in a variety of systems (Dror et al., 2011; Shan et al., 2011; Decherchi et al., 2015; Schneider et al., 2016; Mondal et al., 2018). To avoid ligand aggregation, artificial repulsive interactions among ligands may be used (Shan et al., 2011). Similar strategies are also extensively used in so-called mixed-solvent (or co-solvent) approaches, where high concentrations of fragments are used to identify and stabilize cryptic pockets (Guvench and MacKerell, 2009; Bakan et al., 2012; Schmidt et al., 2019; Kuzmanic et al., 2020). Another idea is to only use isolated beads as probes representing chemical groups or fragments, to predict the chemical topology in pockets and generate pharmacophore models (Michelarakis et al., 2018; Michelarakis, 2019). The combination of CG models, enhanced sampling, and ligand/fragment concentration strategies will allow simulations of competitive binding assays.

An advantage of Martini dynamic docking approach is the improved representation of protein flexibility via Gō-like potentials (Poma et al., 2017; Souza et al., 2019). Although some docking strategies can also include protein flexibility (Amaro et al., 2018; Evangelista Falcon et al., 2019), they usually depend on prior sampling of the protein conformational space, followed by docking in a specific chosen pocket. In the strategy proposed here, no *a priori* selection of the binding pocket is needed. Both induced-fit and conformational-selection mechanisms are included in MD simulations, as recently demonstrated (Souza et al., 2020); however, accuracy will depend on the quality of the protein CG model.

Another major advantage is the possibility to include complex environments, such as multicomponent membranes, crowded protein solutions, or other relevant *in vivo*-like conditions, allowing more realistic predictions. Competition with the environment may be relevant for proper interpretation of ligand biological activity. For instance, lipid membrane composition may affect kinetic rates and (un)binding constants in GPCRs (Vauquelin, 2010; Sykes et al., 2014, 2019; Yuan et al., 2018) by altering ligand partitioning to the membrane where the target protein is located. Atomistic MD simulations of such complex systems are computationally very costly, while they are already within reach with Martini (Marrink et al., 2019).

Combining “standard” docking algorithms with Martini provides a computationally cheap alternative to all-atom docking. As recently demonstrated by HADDOCK (Honorato et al., 2019; Roel-Touris et al., 2019), docking with Martini can be one order of magnitude faster than atomistic docking. This would allow to routinely explore very large ligand datasets (Lyu et al., 2019) or even to use massive docking with grids covering the whole

¹mad.ibcp.fr

the protein, or exploring multiple proteins/conformations at the same time. However, common problems of docking approaches (mentioned above) still would be present; probably Martini MD approaches represent a better compromise between accuracy and computational performance.

LEAD OPTIMIZATION: BACKMAPPING AND COARSE GRAINING IN CHEMICAL SPACE

Accurate predictions of ligand binding poses and affinities are key aspects for lead optimization (Jorgensen, 2009; Vivo et al., 2016). In atomistic pipelines, MD simulations can be used as a post-processing tool to validate and/or refine the binding poses from docking (Vivo et al., 2016). After this first check, more rigorous estimates of ligand binding affinities can be achieved by free energy perturbation (FEP) or thermodynamics integration (TI) (Jorgensen and Thomas, 2008; Jorgensen, 2009) – methods based on conversion of one ligand to another, allowing to add or replace substituents, in order to optimize ligand-protein interactions. In a Martini drug design pipeline (step 3A of **Figure 1**), one could simply convert the CG representation to all-atom (“backmapping” procedure) to verify and refine the CG docking poses. Currently, the most reliable approach for backmapping is the geometric projection implemented in *Backward* (Wassenaar et al., 2014). The main disadvantage is the need for mapping files for each ligand. After obtaining the atomistic structures, any MD-based simulations can be straightforwardly used. Careful equilibration is necessary to allow relaxation of the system, in particular, the water molecules may need to fill small cavities in pockets not accessible to CG water. One possibility is to model buried water molecules or ions using smaller beads, as previously showcased (Souza et al., 2020). Such difficulties are also common in standard docking approaches, as they usually do not include water molecules.

An alternative possibility for lead optimization in Martini would be to reverse the order of the steps, performing first a preliminary set of FEP/TI calculations at the Martini CG level. Such approach would allow to explore a broader portion of the chemical space. On top of the default computational efficiency of CG models, additional speed up could be obtained. First, given the smoother potential surface, the replacement of one bead for another (representing different chemical groups) could be performed in less FEP/TI windows. Additionally, as Martini CG beads generally represent more than one chemical fragment (Menichetti et al., 2019; Bereau, 2020), the exploration of chemical space increases computational efficiency by an additional factor 10^3 – 10^4 (Menichetti et al., 2019; Bereau, 2020) thanks to the reduction in the size of chemical space. Each bead of the CG model can be transformed into different chemical groups, for instance by using different mapping files for each bead in the *Backward* code (Wassenaar et al., 2014). With this alternative lead optimization approach, backmapping would be performed as the last step, to increase accuracy of the predictions.

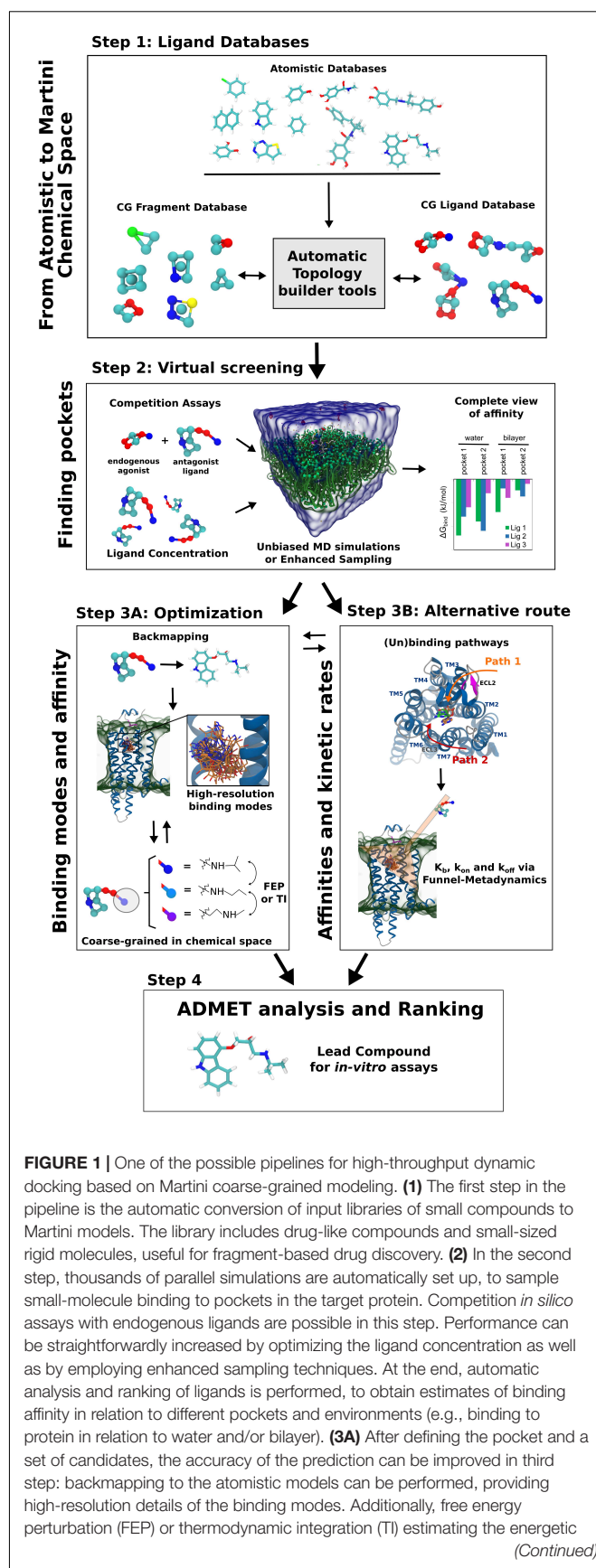


FIGURE 1 | cost of converting certain chemical groups into others can allow further optimization of the molecular structure. Here, coarse-graining in the chemical space is possible, as Martini CG moieties can represent more than one chemical fragment at the same time. **(3B)** An alternative or complementary third step, based on binding affinity and kinetics, is also considered here. Analysis of trajectories obtained in step 2 can help to identify the drug (un)binding pathways, which can be used in methods as Funnel-Metadynamics to provide lowest energy binding modes and dissociation rates k_{off} (drug residence time) states determining. **(4)** The combined analysis of steps II and III can be used for predictions of activity, which in combination with ADMET predictions leads to the final rankings and selection of the lead compounds for *in vitro* assays. Part of the figure is adapted from Souza et al. (2020).

ALTERNATIVE ROUTE: LIGAND BINDING PATHWAYS, BINDING AFFINITIES, AND KINETIC RATES

Drug discovery is historically focused on the elucidation and optimization of the ligand binding mode and binding affinity. However, *in vivo* drug activity is quantitatively correlated to the drug residence time – i.e., dissociation constant rate k_{off} – more than binding affinity K_b (Copeland et al., 2006). The idea of integrating kinetic data in drug screening has been around since the beginning of 2000s (Limongelli, 2020; Nunes-Alves et al., 2020). However, ligand binding kinetics is determined by rare events, crossing ephemeral, high-energy states, elusive to both experiments, and computations (Copeland, 2016). The recent proof of concept with Martini 3 (Souza et al., 2020) opens the possibility of including information on ligand binding pathways in drug design pipelines (step 3B of **Figure 1**). The data coming from unbiased CG MD simulations should be integrated in a rigorous theoretical framework. One possibility would be to use Markov state models (Husic and Pande, 2018) based on Martini dynamic docking screening (step 2 of in **Figure 1**). The method has proven useful in atomistic ligand binding simulation studies (Buch et al., 2011) but it shows difficulties in defining the macrostates of the process, the choice of lag-time, and the sampling necessary to ensure statistical significance. An attractive strategy is to combine CG MD with Funnel-Metadynamics (Limongelli et al., 2013; Raniolo and Limongelli, 2020) that has emerged as a powerful method to reproduce binding mechanisms in ligand/protein and ligand/DNA complexes, identify crystallographic binding modes and predict binding free energies (Troussicot et al., 2015; Comitani et al., 2016; Moraca et al., 2017; Saleh et al., 2017; Yuan et al., 2018; D'Annessa et al., 2019). During FM simulations, the whole drug binding mechanism is reproduced, from the fully solvated state to the final binding mode, allowing to disclose important aspects of the binding process such as (i) the presence of alternative binding modes; (ii) the role of the solvent; and (iii) the kinetically relevant states (Tiwary et al., 2015; Brotzakis et al., 2019; Raniolo and Limongelli, 2020). CG-FM allows quantitative predictions of k_{off} and K_b , ligand binding modes, and rate determining steps (**Figure 1**). This advance will represent a paradigm shift in drug design, as medicinal chemists would optimize the structure of drug candidates not only based on the

static representation of the ligand binding mode, but also on the structures of kinetically relevant states. We point out that the reduction of friction from the missing atomistic degrees of freedom speeds up CG dynamics and affects kinetic estimates. However, estimating trends may be useful enough for ligand screening, while realistic kinetics rates might be recovered from estimates of the friction reduction (Español and Zúñiga, 2011).

FURTHER CONSIDERATIONS AND DISCUSSION

We described a new vision of high-throughput drug screening based on Martini CG models. Although most of the recent efforts in new drug design approaches focused on artificial intelligence (AI), the development of new methods covering gaps in standard approaches is equally important. Machine learning and other AI approaches have great advantages when tackling problems with enough experimental data to be used as training dataset. In situations where this is not the case, physics-based approaches (such as CG molecular dynamics) can perform better. In particular, structural databases of transmembrane proteins are still limited. The same is also true for databases that include dynamic information, which can be important to elucidate hidden allosteric pockets, to properly model fit-induced ligand binding process or to determine ligand association/dissociation pathways. More than complementary, AI and physics-based approaches can be combined, with CG MD simulations being used for the training of AI models or for the further refinement of AI predictions.

The proposed Martini drug design workflow (**Figure 1**) could be applied in full, or specific modules could be adapted in more traditional virtual screening campaigns. Screening of drugs based on ligand binding pathways and dissociation rates is currently out of reach for all-atom descriptions, due to the prohibitively high computational cost. Flexible proteins in complex environments are also too costly for all-atom docking approaches. Martini greatly reduces the computational costs of MD, while offering reasonable accuracy and structural detail. Accuracy will be further improved with the implementation of polarizable models (Yesylevskyy et al., 2010; de Jong et al., 2013; Michalowsky et al., 2017, 2018; Khan et al., 2020). Additionally, protonation state changes and pH effects can be included with Titratable Martini approaches (Grünwald et al., 2020). Also within reach is the design of epitopes and nucleic acids, useful for rational vaccine development (Kulp and Schief, 2013; Hodgson, 2020; Norman et al., 2020). In this context, even CG MD simulations may be overly expensive, as the approach demands scanning of protein-protein and protein-nucleic acid interfaces. Here, combination with standard docking is already a reality, as recently implemented in HADDOCK (Honorato et al., 2019; Roel-Touris et al., 2019). Overall, we believe dynamic docking with CG models has great innovation potential, both in academic and private sectors, and we hope this Perspective will contribute to motivate the modeling community to expand the efforts in this area.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

PCTS wrote the first draft of the manuscript and prepared the figure. All authors contributed to the conception of the perspective article, manuscript revision, read, and approved the submitted version.

REFERENCES

- Ahalawat, N., and Mondal, J. (2021). An appraisal of computer simulation approaches in elucidating biomolecular recognition pathways. *J. Phys. Chem. Lett.* 12, 633–641. doi: 10.1021/acs.jpclett.0c02785
- Alessandri, R., Sami, S., Barnoud, J., Vries, A. H., Marrink, S. J., and Havenith, R. W. A. (2020). Resolving donor–acceptor interfaces and charge carrier energy levels of organic semiconductors with polar side chains. *Adv. Funct. Mater.* 30:2004799. doi: 10.1002/adfm.202004799
- Alessandri, R., Souza, P. C. T., Thallmair, S., Melo, M. N., de Vries, A. H., and Marrink, S. J. (2019). Pitfalls of the Martini model. *J. Chem. Theory Comput.* 15, 5448–5460. doi: 10.1021/acs.jctc.9b00473
- Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., Andrew McCammon, J., Miao, Y., et al. (2018). Ensemble docking in drug discovery. *Biophys. J.* 114, 2271–2278. doi: 10.1016/j.bpj.2018.02.038
- Arnarez, C., Mazat, J.-P., Elezgaray, J., Marrink, S.-J., and Periole, X. (2013). Evidence for cardiolipin binding sites on the membrane-exposed surface of the cytochrome bc1. *J. Am. Chem. Soc.* 135, 3112–3120. doi: 10.1021/ja310577u
- Baaden, M., and Marrink, S. J. (2013). Coarse-grain modelling of protein–protein interactions. *Curr. Opin. Struct. Biol.* 23, 878–886. doi: 10.1016/j.sbi.2013.09.004
- Bakan, A., Nevins, N., Lakdawala, A. S., and Bahar, I. (2012). Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules. *J. Chem. Theory Comput.* 8, 2435–2447. doi: 10.1021/ct300117j
- Bereau, T. (2020). *Computational Compound Screening of Biomolecules and Soft Materials by Molecular Simulations*. Available online at: <http://arxiv.org/abs/2010.03298> (accessed November 20, 2020).
- Bereau, T., and Kremer, K. (2015). Automated parametrization of the coarse-grained Martini force field for small organic molecules. *J. Chem. Theory Comput.* 11, 2783–2791. doi: 10.1021/acs.jctc.5b00056
- Brotzak, Z. F., Faidon Brotzak, Z., Limongelli, V., and Parrinello, M. (2019). Accelerating the calculation of protein–ligand binding free energy and residence times using dynamically optimized collective variables. *J. Chem. Theory Comput.* 15, 743–750. doi: 10.1021/acs.jctc.8b00934
- Bruce, N. J., Ganotra, G. K., Kokh, D. B., Kashif Sadiq, S., and Wade, R. C. (2018). New approaches for computing ligand–receptor binding kinetics. *Curr. Opin. Struct. Biol.* 49, 1–10. doi: 10.1016/j.sbi.2017.10.001
- Bruininks, B. M., Souza, P. C., Ingolfsson, H., and Marrink, S. J. (2020). A molecular view on the escape of lipoplex DNA from the endosome. *eLife* 9:e52012. doi: 10.7554/eLife.52012
- Buch, I., Giorgino, T., and De Fabritiis, G. (2011). Complete reconstruction of an enzyme–inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10184–10189. doi: 10.1073/pnas.1103547108
- Callaway, E. (2020). “It will change everything”: DeepMind’s AI makes gigantic leap in solving protein structures. *Nature* 588, 203–204. doi: 10.1038/d41586-020-03348-4
- Casalini, T., Limongelli, V., Schmutz, M., Som, C., Jordan, O., Wick, P., et al. (2019). Molecular modeling for nanomaterial–biology interactions: opportunities, challenges, and perspectives. *Front. Bioeng. Biotechnol.* 7:268. doi: 10.3389/fbioe.2019.00268
- Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P., and Parrinello, M. (2017). Unbinding kinetics of a p38 MAP kinase Type II inhibitor from metadynamics simulations. *J. Am. Chem. Soc.* 139, 4780–4788. doi: 10.1021/jacs.6b12950
- Castillo, N., Monticelli, L., Barnoud, J., and Tieleman, D. P. (2013). Free energy of WALP23 dimer association in DMPC, DPPC, and DOPC bilayers. *Chem. Phys. Lipids* 169, 95–105. doi: 10.1016/j.chemphyslip.2013.02.001
- Cheng, F., Li, W., Liu, G., and Tang, Y. (2013). In silico ADMET prediction: recent advances, current challenges and future trends. *Curr. Top. Med. Chem.* 13, 1273–1289. doi: 10.2174/15680266113139990033
- Comitani, F., Limongelli, V., and Molteni, C. (2016). The free energy landscape of GABA binding to a pentameric ligand-gated ion channel and its disruption by mutations. *J. Chem. Theory Comput.* 12, 3398–3406. doi: 10.1021/acs.jctc.6b00303
- Copeland, R. A. (2016). The drug–target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* 15, 87–95. doi: 10.1038/nrd.2015.18
- Copeland, R. A., Pompliano, D. L., and Meek, T. D. (2006). Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* 5, 730–739. doi: 10.1038/nrd2082
- Corradi, V., Sejdiu, B. I., Mesa-Gallosio, H., Abdizadeh, H., Noskov, S. Y., Marrink, S. J., et al. (2019). Emerging diversity in lipid–protein interactions. *Chem. Rev.* 119, 5775–5848. doi: 10.1021/acs.chemrev.8b00451
- Dandekar, B. R., and Mondal, J. (2020). Capturing protein–ligand recognition pathways in coarse-grained simulation. *J. Phys. Chem. Lett.* 11, 5302–5311. doi: 10.1021/acs.jpclett.0c01683
- D’Annessa, I., Di Leva, F. S., La Teana, A., Novellino, E., Limongelli, V., and Di Marino, D. (2020). Bioinformatics and biosimulations as toolbox for peptides and peptidomimetics design: where are we? *Front. Mol. Biosci.* 7:66. doi: 10.3389/fmolb.2020.00066
- D’Annessa, I., Raniolo, S., Limongelli, V., Di Marino, D., and Colombo, G. (2019). Ligand binding, unbinding, and allosteric effects: deciphering small-molecule modulation of HSP90. *J. Chem. Theory Comput.* 15, 6368–6381. doi: 10.1021/acs.jctc.9b00319
- de Jong, D. H., Liguori, N., van den Berg, T., Arnarez, C., Periole, X., and Marrink, S. J. (2015). Atomistic and coarse grain topologies for the cofactors associated with the photosystem II core complex. *J. Phys. Chem. B* 119, 7791–7803. doi: 10.1021/acs.jpcc.5b00809
- de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., et al. (2013). Improved parameters for the Martini coarse-grained protein force field. *J. Chem. Theory Comput.* 9, 687–697. doi: 10.1021/ct300646g
- Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W., and Cavalli, A. (2015). The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nat. Commun.* 6:6155. doi: 10.1038/ncomms7155
- Delort, B., Renault, P., Charlier, L., Raussin, F., Martinez, J., and Floquet, N. (2017). Coarse-grained prediction of peptide binding to G-protein coupled receptors. *J. Chem. Inf. Model.* 57, 562–571. doi: 10.1021/acs.jcim.6b00503
- Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., et al. (2011). Pathway and mechanism of drug binding to G-protein-coupled

- receptors. *Proc. Natl. Acad. Sci. U. S. A.* 108, 13118–13123. doi: 10.1073/pnas.1104614108
- Duarte, Y., Márquez-Miranda, V., Miossec, M. J., and González-Nilo, F. (2019). Integration of target discovery, drug discovery and drug delivery: a review on computational strategies. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.* 11:e1554. doi: 10.1002/wnan.1554
- Durrant, J. D., and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biol.* 9:71. doi: 10.1186/1741-7007-9-71
- Empereur-mot, C., Pesce, L., Bochicchio, D., Perego, C., and Pavan, G. M. (2020). Swarm-CG: automatic parametrization of bonded terms in coarse-grained models of simple to complex molecules via fuzzy self-tuning particle swarm optimization. *ACS Omega* 5, 32823–32843. doi: 10.26434/chemrxiv.12613427.v2
- Español, P., and Zúñiga, I. (2011). Obtaining fully dynamic coarse-grained models from MD. *Phys. Chem. Chem. Phys.* 13, 10538–10545. doi: 10.1039/c0cp02826f
- Evangelista Falcon, W., Ellingson, S. R., Smith, J. C., and Baudry, J. (2019). Ensemble docking in drug discovery: how many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding? *J. Phys. Chem. B* 123, 5189–5195. doi: 10.1021/acs.jpcc.8b11491
- Feldmeier, K., and Höcker, B. (2013). Computational protein design of ligand binding and catalysis. *Curr. Opin. Chem. Biol.* 17, 929–933. doi: 10.1016/j.cbpa.2013.10.002
- Ferré, G., Louet, M., Saurel, O., Delort, B., Czaplicki, G., M'Kadmi, C., et al. (2019). Structure and dynamics of G protein-coupled receptor-bound ghrelin reveal the critical role of the octanoyl chain. *Proc. Natl. Acad. Sci. U. S. A.* 116, 17525–17530. doi: 10.1073/pnas.1905105116
- Graham, J. A., Essex, J. W., and Khalid, S. (2017). PyCGTOOL: automated generation of coarse-grained molecular dynamics models from atomistic trajectories. *J. Chem. Inf. Model.* 57, 650–656. doi: 10.1021/acs.jcim.7b00096
- Grinter, S., and Zou, X. (2014). Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. *Molecules* 19, 10150–10176. doi: 10.3390/molecules190710150
- Grünewald, F., Souza, P. C. T., Abdizadeh, H., Barnoud, J., de Vries, A. H., and Marrink, S. J. (2020). Titratable Martini model for constant pH simulations. *J. Chem. Phys.* 153:024118. doi: 10.1063/5.0014258
- Guvench, O., and MacKerell, A. D. Jr. (2009). Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput. Biol.* 5:e1000435. doi: 10.1371/journal.pcbi.1000435
- Hodgson, J. (2020). The pandemic pipeline. *Nat. Biotechnol.* 38, 523–532. doi: 10.1038/d41587-020-00005-z
- Honorato, R. V., Roel-Touris, J., and Alexandre, M. J. (2019). Martini-Based Protein-DNA Coarse-Grained HADDOCKing. *Front. Mol. Biosci.* 6:102. doi: 10.3389/fmolb.2019.00102
- Husic, B. E., and Pande, V. S. (2018). Markov state models: from an art to a science. *J. Am. Chem. Soc.* 140, 2386–2396. doi: 10.1021/jacs.7b12191
- Hutson, M. (2019). AI protein-folding algorithms solve structures faster than ever. *Nature* doi: 10.1038/d41586-019-01357-6
- Ingólfsson, H. I., Lopez, C. A., Uusitalo, J. J., de Jong, D. H., Gopal, S. M., Periole, X., et al. (2014). The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4, 225–248. doi: 10.1002/wcms.1169
- Ingólfsson, H. I., Tieleman, P., and Marrink, S. (2015). Lipid organization of the plasma membrane. *Biophys. J.* 108:358a. doi: 10.1016/j.bpj.2014.11.1962
- Javanainen, M., Martinez-Seara, H., and Vattulainen, I. (2017). Excessive aggregation of membrane proteins in the Martini model. *PLoS One* 12:e0187936. doi: 10.1371/journal.pone.0187936
- Jiang, Z., and Zhang, H. (2019). Molecular mechanism of S1P binding and activation of the S1P1 receptor. *J. Chem. Inf. Model.* 59, 4402–4412. doi: 10.1021/acs.jcim.9b00642
- Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Acc. Chem. Res.* 42, 724–733. doi: 10.1021/ar800236t
- Jorgensen, W. L., and Thomas, L. L. (2008). Perspective on free-energy perturbation calculations for chemical equilibria. *J. Chem. Theory Comput.* 4, 869–876. doi: 10.1021/ct800011m
- Khan, H. M., Souza, P. C. T., Thallmair, S., Barnoud, J., de Vries, A. H., Marrink, S. J., et al. (2020). Capturing choline-aromatics cation- π interactions in the Martini force field. *J. Chem. Theory Comput.* 16, 2550–2560. doi: 10.1021/acs.jctc.9b01194
- Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chem. Rev.* 116, 7898–7936. doi: 10.1021/acs.chemrev.6b00163
- Kulp, D. W., and Schief, W. R. (2013). Advances in structure-based vaccine design. *Curr. Opin. Virol.* 3, 322–331. doi: 10.1016/j.coviro.2013.05.010
- Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J., and Gervasio, F. L. (2020). Investigating cryptic binding sites by molecular dynamics simulations. *Acc. Chem. Res.* 53, 654–661. doi: 10.1021/acs.accounts.9b00613
- Leelananda, S. P., and Lindert, S. (2016). Computational methods in drug discovery. *Beilstein J. Org. Chem.* 12, 2694–2718. doi: 10.3762/bjoc.12.267
- Limousin, M., Limongelli, V., and Sansom, M. S. P. (2016). Conformational changes in the epidermal growth factor receptor: role of the transmembrane domain investigated by coarse-grained metadynamics free energy calculations. *J. Am. Chem. Soc.* 138, 10611–10622. doi: 10.1021/jacs.6b05602
- Li, C., Iscen, A., Sai, H., Sato, K., Sather, N. A., Chin, S. M., et al. (2020). Supramolecular-covalent hybrid polymers for light-activated mechanical actuation. *Nat. Mater.* 19, 900–909. doi: 10.1038/s41563-020-0707-7
- Limongelli, V. (2020). Ligand binding free energy and kinetics calculation in 2020. *WIREs Comput. Mol. Sci.* 10:e1455. doi: 10.1002/wcms.1455
- Limongelli, V., Bonomi, M., and Parrinello, M. (2013). Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. U. S. A.* 110, 6358–6363. doi: 10.1073/pnas.1303186110
- Liu, X., Shi, D., Zhou, S., Liu, H., Liu, H., and Yao, X. (2018). Molecular dynamics simulations and novel drug discovery. *Expert Opin. Drug Discov.* 13, 23–37. doi: 10.1080/17460441.2018.1403419
- López, C. A., Rzepiela, A. J., de Vries, A. H., Dijkhuizen, L., Hünenberger, P. H., and Marrink, S. J. (2009). Martini coarse-grained force field: extension to carbohydrates. *J. Chem. Theory Comput.* 5, 3195–3210. doi: 10.1021/ct900313w
- Lucendo, E., Sancho, M., Lolicato, F., Javanainen, M., Kulig, W., Leiva, D., et al. (2020). Mcl-1 and Bok transmembrane domains: unexpected players in the modulation of apoptosis. *Proc. Natl. Acad. Sci. U. S. A.* 117, 27980–27988. doi: 10.1073/pnas.200885117
- Luitz, M. P., and Zacharias, M. (2014). Protein-ligand docking using hamiltonian replica exchange simulations with soft core potentials. *J. Chem. Inf. Model.* 54, 1669–1675. doi: 10.1021/ci500296f
- Lyu, J., Wang, S., Balus, 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
- Marrink, S. J., Corradi, V., Souza, P. C. T., Ingólfsson, H. I., Peter Tieleman, D., and Sansom, M. S. P. (2019). Computational modeling of realistic cell membranes. *Chem. Rev.* 119, 6184–6226. doi: 10.1021/acs.chemrev.8b00460
- Marrink, S. J., Jelger Risselada, H., Yefimov, S., Peter Tieleman, D., and de Vries, A. H. (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* 111, 7812–7824. doi: 10.1021/jp071097f
- Marrink, S. J., and Tieleman, D. P. (2013). Perspective on the Martini model. *Chem. Soc. Rev.* 42:6801–6822. doi: 10.1039/c3cs60093a
- Melo, M. N., Arnarez, C., Sikkema, H., Kumar, N., Walko, M., Berendsen, H. J. C., et al. (2017). High-throughput simulations reveal membrane-mediated effects of alcohols on MscL gating. *J. Am. Chem. Soc.* 139, 2664–2671. doi: 10.1021/jacs.6b11091
- Menichetti, R., Kanekal, K. H., and Bereau, T. (2019). Drug-membrane permeability across chemical space. *ACS Cent. Sci.* 5, 290–298. doi: 10.1021/acscentsci.8b00718
- Miao, Y., Feher, V. A., and McCammon, J. A. (2015). Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory Comput.* 11, 3584–3595. doi: 10.1021/acs.jctc.5b00436
- Michalowsky, J., Schäfer, L. V., Holm, C., and Smiatek, J. (2017). A refined polarizable water model for the coarse-grained MARTINI force field with long-range electrostatic interactions. *J. Chem. Phys.* 146:054501. doi: 10.1063/1.4974833
- Michalowsky, J., Zeman, J., Holm, C., and Smiatek, J. (2018). A polarizable MARTINI model for monovalent ions in aqueous solution. *J. Chem. Phys.* 149:163319. doi: 10.1063/1.5028354
- Michalarakis, N. (2019). *Towards Dynamic Pharmacophore Models Through the Use of Coarse Grained Molecular Dynamic Simulations, Dissertation.* Oxford: University of Oxford.

- Michelarakis, N., Sands, Z. A., Sansom, M. S. P., and Stansfeld, P. J. (2018). Towards dynamic pharmacophore models by coarse grained molecular dynamics. *Biophys. J.* 114:558a. doi: 10.1016/j.bpj.2017.11.3050
- Miller, E., Murphy, R., Sindhikara, D., Borrelli, K., Grisewood, M., Ranalli, F., et al. (2020). A reliable and accurate solution to the induced fit docking problem for protein-ligand binding. *ChemRxiv* [Preprint]. doi: 10.26434/chemrxiv.11983845.v2
- Mondal, J., Ahalawat, N., Pandit, S., Kay, L. E., and Vallurupalli, P. (2018). Atomic resolution mechanism of ligand binding to a solvent inaccessible cavity in T4 lysozyme. *PLoS Comput. Biol.* 14:e1006180. doi: 10.1371/journal.pcbi.1006180
- Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Peter Tieleman, D., and Marrink, S.-J. (2008). The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* 4, 819–834. doi: 10.1021/ct700324x
- Moraca, F., Amato, J., Ortuso, F., Artese, A., Pagano, B., Novellino, E., et al. (2017). Ligand binding to telomeric G-quadruplex DNA investigated by funnel-metadynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* 114, E2136–E2145. doi: 10.1073/pnas.1612627114
- Negami, T., Shimizu, K., and Terada, T. (2014). Coarse-grained molecular dynamics simulations of protein-ligand binding. *J. Comput. Chem.* 35, 1835–1845. doi: 10.1002/jcc.23693
- Negami, T., Shimizu, K., and Terada, T. (2020). Coarse-grained molecular dynamics simulation of protein conformational change coupled to ligand binding. *Chem. Phys. Lett.* 742:137144. doi: 10.1016/j.cplett.2020.137144
- Norman, R. A., Ambrosetti, F., Bonvin, A. M. J. J., Colwell, L. J., Kelm, S., Kumar, S., et al. (2020). Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief. Bioinform.* 21, 1549–1567. doi: 10.1093/bib/bbz095
- Nunes-Alves, A., Kokh, D. B., and Wade, R. C. (2020). Recent progress in molecular simulation methods for drug binding kinetics. *Curr. Opin. Struct. Biol.* 64, 126–133. doi: 10.1016/j.sbi.2020.06.022
- Pang, Y. T., Miao, Y., Wang, Y., and McCammon, J. A. (2017). Gaussian accelerated molecular dynamics in NAMD. *J. Chem. Theory Comput.* 13, 9–19. doi: 10.1021/acs.jctc.6b00931
- Periole, X., Cavalli, M., Marrink, S.-J., and Ceruso, M. A. (2009). Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition. *J. Chem. Theory Comput.* 5, 2531–2543. doi: 10.1021/ct9002114
- Perkel, J. M. (2019). The computational protein designers. *Nature* 571, 585–587. doi: 10.1038/d41586-019-02251-x
- Poma, A. B., Cieplak, M., and Theodorakis, P. E. (2017). Combining the MARTINI and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins. *J. Chem. Theory Comput.* 13, 1366–1374. doi: 10.1021/acs.jctc.6b00986
- Raniolo, S., and Limongelli, V. (2020). Ligand binding free-energy calculations with funnel metadynamics. *Nat. Protoc.* 15, 2837–2866. doi: 10.1038/s41596-020-0342-4
- Roel-Touris, J., Don, C. G., Honorato, R. V., João, P. G. L., and Bonyin, A. M. J. (2019). Less is more: coarse-grained integrative modeling of large biomolecular assemblies with HADDOCK. *J. Chem. Theory Comput.* 15, 6358–6367. doi: 10.1021/acs.jctc.9b00310
- Rognan, D. (2012). Fragment-based approaches and computer-aided drug discovery. *Top. Curr. Chem.* 317, 201–222. doi: 10.1007/128_2011_182
- Saleh, N., Ibrahim, P., Saladino, G., Gervasio, F. L., and Clark, T. (2017). An efficient metadynamics-based protocol to model the binding affinity and the transition state ensemble of G-protein-coupled receptor ligands. *J. Chem. Inf. Model.* 57, 1210–1217. doi: 10.1021/acs.jcim.6b00772
- Schmidt, D., Boehm, M., McClendon, C. L., Torella, R., and Gohlke, H. (2019). Cosolvent-enhanced sampling and unbiased identification of cryptic pockets suitable for structure-based drug design. *J. Chem. Theory Comput.* 15, 3331–3343. doi: 10.1021/acs.jctc.8b01295
- Schneider, S., Provati, D., and Filizola, M. (2016). How oliceridine (TRV-130) binds and stabilizes a μ -opioid receptor conformational state that selectively triggers G protein signaling pathways. *Biochemistry* 55, 6456–6466. doi: 10.1021/acs.biochem.6b00948
- Schuetz, D. A., Bernetti, M., Bertazzo, M., Musil, D., Eggenweiler, H.-M., Recanatini, M., et al. (2019). Predicting residence time and drug unbinding pathway through scaled molecular dynamics. *J. Chem. Inf. Model.* 59, 535–549. doi: 10.1021/acs.jcim.8b00614
- Shan, Y., Kim, E. T., Eastwood, M. P., Dror, R. O., Seeliger, M. A., and Shaw, D. E. (2011). How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* 133, 9181–9183. doi: 10.1021/ja202726y
- Siebenmorgen, T., and Zacharias, M. (2020). Computational prediction of protein-protein binding affinities. *WIREs Comput. Mol. Sci.* 10:e1448. doi: 10.1002/wcms.1448
- Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacol. Rev.* 66, 334–395. doi: 10.1124/pr.112.007336
- Souza, P. C. T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grunewald, F., et al. (2021). Martini 3: a general purpose force field for coarse-grain molecular dynamics. *Nat. Methods*. doi: 10.1038/s41592-021-01098-3
- Souza, P. C. T., Thallmair, S., Conflitti, P., Ramírez-Palacios, C., Alessandri, R., Raniolo, S., et al. (2020). Protein-ligand binding with the coarse-grained Martini model. *Nat. Commun.* 11:3714. doi: 10.1038/s41467-020-17437-5
- Souza, P. C. T., Thallmair, S., Marrink, S. J., and Mera-Adasme, R. (2019). An allosteric pathway in copper, zinc superoxide dismutase unravels the molecular mechanism of the G93A amyotrophic lateral sclerosis-linked mutation. *J. Phys. Chem. Lett.* 10, 7740–7744. doi: 10.1021/acs.jpclett.9b02868
- Stark, A. C., Andrews, C. T., and Elcock, A. H. (2013). Toward optimized potential functions for protein-protein interactions in aqueous solutions: osmotic second virial coefficient calculations using the MARTINI coarse-grained force field. *J. Chem. Theory Comput.* 9, 4176–4185. doi: 10.1021/ct400008p
- Sun, F., Schroer, C. F. E., Palacios, C. R., Xu, L., Luo, S.-Z., and Marrink, S. J. (2020). Molecular mechanism for bidirectional regulation of CD44 for lipid raft affiliation by palmitoylations and PIP2. *PLoS Comput. Biol.* 16:e1007777. doi: 10.1371/journal.pcbi.1007777
- Sykes, D. A., Parry, C., Reilly, J., Wright, P., Fairhurst, R. A., and Charlton, S. J. (2014). Observed drug-receptor association rates are governed by membrane affinity: the importance of establishing “micro-pharmacokinetic/pharmacodynamic relationships” at the β 2-adrenoceptor. *Mol. Pharmacol.* 85, 608–617. doi: 10.1124/mol.113.090209
- Sykes, D. A., Stoddart, L. A., Kilpatrick, L. E., and Hill, S. J. (2019). Binding kinetics of ligands acting at GPCRs. *Mol. Cell. Endocrinol.* 485, 9–19. doi: 10.1016/j.mce.2019.01.018
- Tiwary, P., Limongelli, V., Salvalaglio, M., and Parrinello, M. (2015). Kinetics of protein-ligand unbinding: predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. U.S.A.* 112, E386–E391. doi: 10.1073/pnas.1424461112
- Troussicot, L., Guillièrre, F., Limongelli, V., Walker, O., and Lancelin, J.-M. (2015). Funnel-metadynamics and solution NMR to estimate protein-ligand affinities. *J. Am. Chem. Soc.* 137, 1273–1281. doi: 10.1021/ja511336z
- Uusitalo, J. J., Ingólfsson, H. I., Akhshi, P., Peter Tieleman, D., and Marrink, S. J. (2015). Martini coarse-grained force field: extension to DNA. *J. Chem. Theory Comput.* 11, 3932–3945. doi: 10.1021/acs.jctc.5b00286
- Uusitalo, J. J., Ingólfsson, H. I., Marrink, S. J., and Faustino, I. (2017). Martini coarse-grained force field: extension to RNA. *Biophys. J.* 113, 246–256. doi: 10.1016/j.bpj.2017.05.043
- Van De Waterbeemd, H., and Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2, 192–204. doi: 10.1038/nrd1032
- Van Eerden, F. J., Melo, M. N., Frederix, P. W. J. M., Periole, X., and Marrink, S. J. (2017). Exchange pathways of plastoquinone and plastoquinol in the photosystem II complex. *Nat. Commun.* 8:15214. doi: 10.1038/ncomms15214
- Vauquelin, G. (2010). Rebinding: or why drugs may act longer in vivo than expected from their in vitro target residence time. *Expert Opin. Drug Discov.* 5, 927–941. doi: 10.1517/17460441.2010.512037
- Vazquez-Salazar, L. I., Selle, M., de Vries, A. H., Marrink, S. J., and Souza, P. C. T. (2020). Martini coarse-grained models of imidazolium-based ionic liquids: from nanostructural organization to liquid-liquid extraction. *Green Chem.* 22, 7376–7386. doi: 10.1039/d0gc01823f
- Vivo, M. D., De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. (2016). Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* 59, 4035–4061. doi: 10.1021/acs.jmedchem.5b01684
- Wang, K., Chodera, J. D., Yang, Y., and Shirts, M. R. (2013). Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics. *J. Comput. Aided Mol. Des.* 27, 989–1007. doi: 10.1007/s10822-013-9689-8

- Wassenaar, T. A., Ingólfsson, H. I., Böckmann, R. A., Peter Tieleman, D., and Marrink, S. J. (2015). Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations. *J. Chem. Theory Comput.* 11, 2144–2155. doi: 10.1021/acs.jctc.5b00209
- Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J., and Tieleman, D. P. (2014). Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* 10, 676–690. doi: 10.1021/ct400617g
- Yang, W., and Lai, L. (2017). Computational design of ligand-binding proteins. *Curr. Opin. Struct. Biol.* 45, 67–73. doi: 10.1016/j.sbi.2016.11.021
- Yen, H.-Y., Hoi, K. K., Liko, I., Hedger, G., Horrell, M. R., Song, W., et al. (2018). PtdIns(4,5)P2 stabilizes active states of GPCRs and enhances selectivity of G-protein coupling. *Nature* 559, 423–427. doi: 10.1038/s41586-018-0325-6
- Yesylevskyy, S. O., Schäfer, L. V., Sengupta, D., and Marrink, S. J. (2010). Polarizable water model for the coarse-grained Martini force field. *PLoS Comput. Biol.* 6:e1000810. doi: 10.1371/journal.pcbi.1000810
- Yuan, X., Raniolo, S., Limongelli, V., and Xu, Y. (2018). The molecular mechanism underlying ligand binding to the membrane-embedded site of a G-protein-coupled receptor. *J. Chem. Theory Comput.* 14, 2761–2770. doi: 10.1021/acs.jctc.8b00046

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

Copyright © 2021 Souza, Limongelli, Wu, Marrink and Monticelli. 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.



MD Simulations on a Well-Built Docking Model Reveal Fine Mechanical Stability and Force-Dependent Dissociation of Mac-1/GPIIb α Complex

Xiaoyan Jiang¹, Xiaoxi Sun¹, Jiangguo Lin², Yingchen Ling¹, Ying Fang^{1*} and Jianhua Wu^{1*}

¹Institute of Biomechanics/School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China, ²Research Department of Medical Sciences, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China

OPEN ACCESS

Edited by:

Agnel Praveen Joseph,
Science and Technology Facilities
Council, United Kingdom

Reviewed by:

Matteo Degiacomi,
Durham University, United Kingdom
Jinan Wang,
University of Kansas, United States

*Correspondence:

Jianhua Wu
wujianhua@scut.edu.cn
Ying Fang
yfang@scut.edu.cn

Specialty section:

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

Received: 06 December 2020

Accepted: 12 February 2021

Published: 22 April 2021

Citation:

Jiang X, Sun X, Lin J, Ling Y, Fang Y
and Wu J (2021) MD Simulations on a
Well-Built Docking Model Reveal Fine
Mechanical Stability and Force-
Dependent Dissociation of
Mac-1/GPIIb α Complex.
Front. Mol. Biosci. 8:638396.
doi: 10.3389/fmolb.2021.638396

Interaction of leukocyte integrin macrophage-1 antigen (Mac-1) to platelet glycoprotein IIb α (GPIIb α) is critical for platelet-leukocyte crosstalk in hemostasis and inflammatory responses to vessel injuries under hemodynamic environments. The mechano-regulation and its molecular basis for binding of Mac-1 to GPIIb α remain unclear, mainly coming from the lack of crystal structure of the Mac-1/GPIIb α complex. We herein built a Mac-1/GPIIb α complex model through a novel computer strategy, which included a flexible molecular docking and system equilibrium followed by a “force-ramp + snapback” molecular dynamics (MD) simulation. With this model, a series of “ramp-clamp” steered molecular dynamics (SMD) simulations were performed to examine the GPIIb α -Mac-1 interaction under various loads. The results demonstrated that the complex was mechano-stable for both the high rupture force (>250 pN) at a pulling velocity of 3 Å/ns and the conformational conservation under various constant tensile forces (≤ 75 pN); a catch-slip bond transition was predicted through the dissociation probability, examined with single molecular AFM measurements, reflected by the interaction energy and the interface H-bond number, and related to the force-induced allostery of the complex; besides the mutation-identified residues D222 and R218, the residues were also dominant in the binding of Mac-1 to GPIIb α . This study recommended a valid computer strategy for building a likely wild-type docking model of a complex, provided a novel insight into the mechanical regulation mechanism and its molecular basis for the interaction of Mac-1 with GPIIb α , and would be helpful for understanding the platelet-leukocyte interaction in hemostasis and inflammatory responses under mechano-microenvironments.

Keywords: Mac-1, GPIIb α , molecular dynamics simulation, structure-function relation, leukocyte-platelet interaction

INTRODUCTION

Interaction of platelet glycoprotein receptor I b α (GPIIb) with $\alpha_M\beta_2$ (Mac-1) integrin mediates crosstalk between platelets and leukocytes in hemostasis and inflammatory responses to the vessel damages (Diamond et al., 1995). In platelet recruitment and thrombus formation, the circulating platelets first tether to, then roll on, and lastly adhered at the injured vessel sites, accompanying with platelet activation (Rahman and Hlady, 2019). The activated platelets recruit and activate leukocytes to prevent infection caused by an injury (Von Hundelshausen and Weber, 2007; Karshovska et al., 2013; Schrottmaier et al., 2015). Binding of P-selectin on activated platelet to P-selectin glycoprotein ligand 1 (PSGL-1) mediates adhesion of leukocytes to the platelets and further induces activation of Mac-1 integrin on leukocytes. The activated Mac-1 integrin enhances firm adhesion of leukocytes to platelets through binding with GPIIb (Diacovo et al., 1996; McEver and Cummings, 1997). The interaction of Mac-1 to GPIIb may be force dependent, especially under pathological flow environments (Kruss et al., 2013).

Mac-1 integrin, a heterodimeric protein, includes α and β subunits, regulates leukocyte functions, including crawling (Phillipson et al., 2006; Sumagin et al., 2010), chemotaxis (McDonald et al., 2010), survival (Whitlock et al., 2000), apoptosis (Coxon et al., 1996), and neutrophil extracellular trap (NET) formation (Behnen et al., 2014; Silva et al., 2020). The major binding site locates at the inserted (I) domain of the α_M subunit (Diamond et al., 1993), like the A1 domain of von Willebrand factor (VWF) (Sadler et al., 1985). GPIIb is composed of an N-terminal, a transmembrane, and an intracellular domain. The N-terminal domain exhibits a narrow and curved shape, and eight leucine-rich repeats (LRRs) constitute the central region of the molecule, belonging to the typical LRR protein (Kobe and Kajava, 2001). Binding of N-terminal region (F201–G268) of GPIIb to the I-domain of Mac-1 induces stable association of platelets with neutrophils (Simon et al., 2000; Wang et al., 2005), while deletion or inhibition of either Mac-1 or GPIIb suppresses interaction of platelets with neutrophils or monocytes under inflammatory conditions (Hidalgo et al., 2009). Mutation experiments suggest that the four residues, such as T211, T213, R216, and R244 on Mac-1, locate at the binding site of the complex, and so do the residues R218, R222, and N223 on GPIIb (Simon et al., 2000; Ehlers et al., 2003; Wang et al., 2005; Morgan et al., 2019). Mutations of the residues T213 and R216 on Mac-1 cause delay of thrombosis after carotid and cremaster muscle microvascular injury (Ehlers et al., 2003), suggesting promotion of GPIIb–Mac-1 interaction to thrombosis.

Studies *in vitro* have shown that increasing wall shear stress in the range from 0.1 to 5 dyn/cm² reduces the attachments of neutrophils on GPIIb-coated substrates (Kruss et al., 2013), but less knowledge is about mechano-regulation on the interaction of Mac-1 with GPIIb. Lack of crystal structure of the complex makes the molecular basis of Mac-1/GPIIb interaction unclear, despite the main crystallized structures of Mac-1 and GPIIb have respectively been solved (Ehlers et al., 2003; Li and Emsley, 2013).

Molecular docking is demonstrated to be a powerful tool in building a computer model of ligand–receptor complex for various adhesive molecular systems, while a very likely bad thermo- and mechano-stability make the docking model unreliable (Zeng et al., 2013). However, AFM measurements reveal successfully various force-dependent ligand–receptor interactions of adhesive (Lee et al., 2016; Li et al., 2018) and so do the molecular dynamics (MD) simulations (Fang et al., 2012; Fang et al., 2018).

We herein built a model of the Mac-1/GPIIb complex through a novel computer strategy, in which a flexible molecular docking follows a “force-ramp + snapback” MD simulation (Methods and Materials) (Smith et al., 2005; Torchala et al., 2013). This model was predicted to be a likely wild-type one for its thermo- and mechano-stability and used to examine the mechano-regulation mechanism and its molecular basis of interaction of Mac-1 to GPIIb by running a series of “ramp-clamp” steered molecular dynamics (SMD) simulations. A biphasic force-dependent dissociation of Mac-1 from GPIIb was predicted by MD simulations, examined through AFM measurements, and demonstrated to be relative to force-induced allostery of the complex. The present computer strategy for optimizing the docking model with the treatment of “force-ramp + snapback” SMD simulation might be served as a novel powerful tool in building a likely wild-type docking model of complex for various adhesive molecular systems. Besides, the present study provides a novel insight into the mechano-regulation mechanism and its molecular basis for the interaction of Mac-1 to GPIIb, and further is helpful to understand the effects of force on platelet–leukocyte crosstalk in hemostasis and inflammatory responses under flows.

MATERIALS AND METHODS

AFM Bond Lifetime Measurement Experiments

Recombinant human integrin $\alpha_M\beta_2$ (Mac-1) and recombinant human CD42b (GPIIb) were purchased from R&D Systems. Anti-6 \times His tag antibody was purchased from Abcam (Supplement Material). MnCl₂ and BSA were purchased from Sigma-Aldrich. To measure the interaction of Mac-1–GPIIb, a cantilever tip (MLCT; Bruker AFM Probes) was incubated with 30 μ l of 15 μ g/ml Mac-1 overnight at 4°C. 30 μ l of GPIIb (15 μ g/ml) was adsorbed on a small spot on a petri dish overnight at 4°C. After rinsing with PBS, the tip of the cantilever was incubated with HBSS containing 2% BSA and 1 mM Mn²⁺ for an hour at room temperature to obtain activated Mac-1. After rinsing with PBS, the petri dish was incubated for 30 min at room temperature with HBSS containing 2% BSA to block nonspecific adhesion. Notably, Mac-1 and GPIIb were also immobilized on a cantilever tip and a petri dish surface by the anti-6 \times His tag antibody capturing to examine the effect of molecule orientations on interactions (Supplement Material). During each measurement cycle, a petri dish was driven by the piezoelectric translator to contact with a cantilever tip to reach the set-point (0.5 V), and then immediately retracted slightly and held close to the tip for

0.5 s to allow bond formation and retracted along the z direction at a speed of 200 nm/s. A feedback system was applied in the experiments. During the retraction, if a tensile force was detected (adhesion) and reached the preset level, the retraction would stop to clamp the force at that level until the tensile force broke and further retracted to its initial position. If no tensile force was detected (no adhesion) or a tensile force did not reach the preset level, the petri dish was directly retracted to its initial position. The number of adhesion events and bond lifetimes at desired forces were measured from the force–time curves.

Molecular Docking

Flexible docking of I-domain of Mac-1 (residues 131–A317; PDB code 1JLM) to GPIIb (residues 1–267; PDB code 1P9A) was performed with SWARMDOCK server web (version 15.04.01) (<https://bmm.crick.ac.uk/~svc-bmm-swarmdock/submit.cgi>) (Jain, 2003). In docking, seven residues, four (T211, T213, K244, and R216) on Mac-1 and others (R218, R222, and N223) on GPIIb, were designated as binding site residues because of the mutation data (Simon et al., 2000; Wang et al., 2005). The N- and C-terminal in either of Mac-1 and GPIIb was set to be neutral. All docking results (444 complex structures) were grouped into ten clusters, in which each was defined as an ensemble of at least two complex models with ligand interface C_{α} RMSD <6 Å. The docking model with the lowest binding energy and the specified essential residues participating in the receptor–ligand interaction was considered as the best one. Each complex model was visually inspected by visual molecular dynamics (VMD), and only one was selected as the best model by the following criteria: the N- and C-terminal of Mac-1 could not be bound with the LRR domain of GPIIb because of the binding of the Mac-1 legs to both the N- and the C-terminus, and the model had not only the largest number of designated interface residues but also the lowest SWARMDOCK score. The best complex model, the so-called Model I, was selected from docking results and used for subsequent analysis.

System Setup and Equilibrium

We herein used two software packages, VMD for visualization and modeling (Humphrey et al., 1996) and the NAMD 2.13 program for molecular dynamics simulations (Phillips et al., 2005). The Model I was solvated with TIP3P water molecules in a rectangular box (6.54 nm \times 11.6 nm \times 8.4 nm). The system was neutralized by adding 118 Na^+ and 126 Cl^- (150 mM concentration) to mimic the actual physiological environment and consisted of 103,164 atoms. MD simulations were performed with periodic boundary condition and 2 fs time step as well as the CHARMM27 all-atom force field (MacKerell et al., 1998), along with cMAP correction for backbone, particle mesh Ewald (PME) algorithm for electrostatic interaction, a 12 Å cutoff for electrostatic, and Van der Waals interaction. All bonds were restrained using SHAKE to allow the time step of 2 fs. The system was energy minimized first for 15,000 steps with heavy or non-hydrogen protein atoms being fixed, and then for another 15,000 steps with all atoms free. The energy-minimized systems were heated gradually from 0 to 310 K in 0.1 ns first, and then equilibrated once for 100 ns with pressure and temperature

control. The temperature was held at 310°K using Langevin dynamics, and the pressure was held at 1 atmosphere by the Langevin piston method. The equilibrated structure of Model I with better thermal stabilization was used as the initial conformation for the subsequent steered molecular dynamics (SMD) simulations (**Supplementary Figure S1**).

Steered Molecular Dynamics Simulation

The SMD simulations in “force-ramp,” “force-ramp + snapback,” and “ramp + clamp” modes were performed for testing the mechanical strength, optimizing the structure of Model I, and the mechano-regulated structure–function interaction of the complex of Mac-1 with GPIIb, respectively. In the force-ramp MD simulation, the N-terminal C_{α} atom of GPIIb was fixed, and the C-terminal C_{α} atom of Mac-1 was pulled with constant pulling velocity (3 nm/ns) along the line between the steered and fixed atom (**Supplementary Figure S1C**) (Simon, 2012). The dummy atom and the steered atom were linked by the virtual spring with a spring constant of 13.89 pN/Å. The rupture force of the complex was read from the peak in the force–time pattern simulated with the force-ramp mode and used to scale the mechanical strength of the complex.

It is assumed that a rational docking model for the Mac-1/GPIIb complex should have both, a better thermal stabilization and stronger mechanical strength. To making Model I be more rational, we here developed a computer strategy of docking model optimizing (DMO) *via* the so-called force-ramp + snapback SMD simulations because the poor mechanical strength of the complex Model I was demonstrated by its low complex rupture forces. In a run with the “force-ramp + snapback” mode, an SMD simulation of 5 ns with constant pulling velocity (3 nm/ns) was performed first, then the system was mechanically unloaded but followed with equilibrium of 100 ns or 40 ns for the snapback complex. Through MD simulation with one or several mechanically loading–unloading cycles, as described above, the Model I might be remodeled and optimized as a more rational model, named Model II or Model III, for its better structural stabilization.

The so-called ramp-clamp SMD simulations, a force-clamp MD simulation followed a force-ramp one, were performed thrice on the equilibrated system with the Model II of the complex to examine the force-induced unbinding and conformation changing of the GPIIb bound with Mac-1. For each simulation, the complex was first pulled until the tensile force arrived at a given value, such as 25, 50, or 75 pN, and then, the SMD simulation was transformed from the force-ramp mode to a force-clamp one, at which the complex was stretched with the given constant tensile force for the following 40 ns.

Data Analysis for MD Simulations

All analyses were treated with VMD tools. The C_{α} root mean square deviation (RMSD) and the solvent accessible surface area (SASA) (with a 1.4 Å probe radius) were used to characterize the conformational change and the hydrophobic core exposure, respectively. The binding energy, consisting of van der Waals energy and electrostatic energy, was calculated through the

NAMD energy plugin in VMD. A hydrogen bond was formed if the donor–acceptor distance and the donor–hydrogen–acceptor angle were less than 3.5 Å and 30°, respectively. A salt bridge was built up once the distance between any of the oxygen atoms of acidic residues (Asp or Glu) and the nitrogen atoms of basic residues (Lys or Arg) was within 4 Å. An occupancy (or survival rate) of an H-bond or a salt bridge was scaled with the percentage of bond survival time in the simulation period. As a reflection of the mechanical strength of receptor–ligand complex (Grubmüller et al., 1996), the rupture force was read from the maximum of the force spectrum in a force-ramp run with constant pulling velocity. All visual inspections and molecular images were completed by using VMD. The formation or breakage of each hydrogen bond on the binding site was assumed to be an independent event not related to other bonds.

As a scale for the residue–residue interactions across binding site, p_{ij} , the probability of the i th ligand residue binding with the j th receptor residue, was evaluated by the following equation:

$$p_{ij} = 1 - \prod_{l=1}^{M_{ij}} (1 - \omega_{ij,l}), \quad i = 1, 2, \dots, M_L; j = 1, 2, \dots, M_R; \\ l = 0, 1, \dots, M_{ij}, \quad (1)$$

where $\omega_{ij,l}$ was the survival ratio of the l th H-bond between the i th ligand residue and the j th receptor residue, M_{ij} (≥ 0) denoted the numbers of H-bonds between the i th ligand residue and the j th receptor residue, and M_L (≥ 1) and M_R (≥ 1) were, respectively, the total numbers of ligand and receptor residues involved in binding. Thus, $P_{j,L}$ (the probabilities of the j th ligand residue binding to the receptor) and $P_{j,R}$ (the probabilities of the j th receptor residue binding to the ligand) were, respectively, estimated by the following equations:

$$P_{j,L} = 1 - \prod_{i=1}^{M_R} (1 - p_{ji}) \quad (2)$$

and

$$P_{j,R} = 1 - \prod_{i=1}^{M_L} (1 - p_{ij}). \quad (3)$$

Furthermore, P_D , the dissociation of ligand from receptor, could be estimated by the following equation:

$$P_D = 1 - \prod_{j=1}^{M_L} (1 - P_{j,L}) = 1 - \prod_{j=1}^{M_R} (1 - P_{j,R}). \quad (4)$$

And, the mechano-regulation factor or the normalized complex dissociation f_D was the ratio of P_D at tensile force of f_0 and of P_D at zero tensile force, that was given by the following equation:

$$f_D = P_D \left| \frac{f = f_0}{P_D} \right|_{f=0}, \quad (5)$$

where f expressed the tensile force on the complex and f_0 was a given tensile force. Regardless of the geometrical and timescale effects on complex dissociation P_D , it was

expected that f_D should be comparable with our AFM experiment data.

RESULTS

A Likely Wild-Type Model of Mac-1–GPIIbα Complex Was Well Built Up Through Molecular Docking With Treatment of the “Force-Ramp + Snapback” MD Simulations

To gain a likely wild-type conformation for the complex of Mac-1 with GPIIbα, we built three structural models (Models I, II, and III) for complex of Mac-1 with GPIIbα, through SWARMDOCK program (Torchala et al., 2013) with and without a DMO treatment (Materials and Methods), respectively. With the lowest SWARMDOCK energy score and the most mutation-identified residues in the binding site, the Model I was picked out from 444 poses generated by docking of Mac-1 to ligand-free GPIIbα and equilibrated by performing a system equilibrium of 100 ns or 40 ns along with the same protocol of energy minimization (see Material and Methods). Models II and III (Figure 1A) were built up, respectively, by remodeling Model I with the so-called force-ramp + snapback SMD simulation of 105 ns or 45 ns, in which 5 ns was spent for the SMD simulation with a pulling velocity of 3 nm/ns, and the other 100 ns or 40 ns was contributed to a system re-equilibration for the unloaded or snapped-back complex (Materials and Methods). Models I, II, and III should be equilibrated because the time courses of the total energy, and the root mean square deviation (RMSD) of C $_{\alpha}$ -atoms fluctuated on their respective stable levels with small relative derivations (Supplementary Figure S2A,B,E). Among the all fourteen observed H-bonds across the complex interface in Model II (Table 1; Supplementary Table S1), the first seven existed also in Model I, but others did not (Figures 1B,C); and except the 7th bond, the other six in Model I had lower occupancies than those in Models II and III. It suggested that the missed or undervalued H-bonding events on the interface in Model I might emerge and be valuation-rational through treatment with “force-ramp + snapback” SMD simulation.

The mean C $_{\alpha}$ -RMSD, binding energy (E), the interface H-bond number (N_{HB}), and the interface buried SASA for complex in equilibrium of 40 ns showed that the C $_{\alpha}$ -RMSD value climbed from 2 Å to a quasi-plateau of 6 Å for the Model I but remained almost at a low level of 2 Å for Models II and III (Figure 2A; Supplementary Figure S2B), suggesting a higher thermo-stabilization of Models II and III in comparison with Model I (Figure 2A); Models II and III rather than Model I should be energy favorable because the binding energies (-398 ± 53 kcal/mol, -369 ± 50 kcal/mol) in Models II and III were far lower than that (-293 ± 75 kcal/mol) in Model I (Figure 2B); the mean number of H-bonds on the binding site over a simulation time of 40 ns for Models I, II, and III were 4.8, 6.9, and 5.8 (Figure 2C), respectively, showing a stable linkage between Mac-1 and GPIIbα for Models II and III rather than Model I; and the mean interface buried SASA was read to be 730 Å² for Model I, 840 Å² for Model II, and 800 Å² for Model III (Figure 2D),

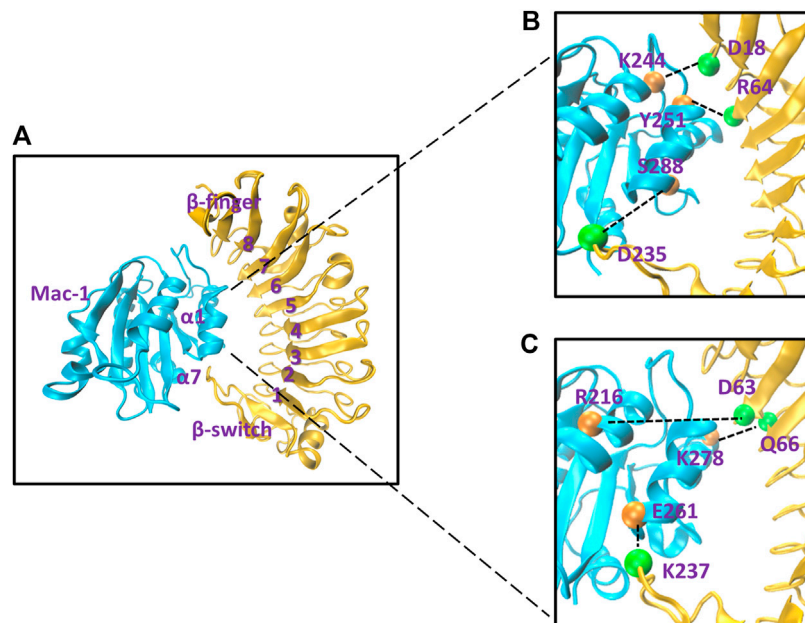


FIGURE 1 | The likely wild-type molecular docking model and some representative involved hydrogen bonds on the binding site of the Mac-1/GPIIbα complex. **(A)** The snapshot of the likely wild-type molecular docking model, which was built up by treating Model I of the Mac-1/GPIIbα complex with a “ramp-snapback” MD simulation (Material and Methods) and shown in new cartoon diagram. Mac-1 and GPIIbα were colored with cyan and orange, respectively. The hydrophobic pocket of GPIIbα consists of six β sheets (from β1 to β6), seven α helices (from α1 to α7), and loops to link any two adjacent α or β structures. **(B)** The three intrinsic hydrogen bonds, which were contributed by the residue pairs such as K244-D18 and Y51-R64 as well as S288-D235 and detected from either of Model I and II. **(C)** The three newly formed hydrogen bonds, which were the linkers between the other three residues (R216-D63, K278-Q66, and E261-237) and occurred just at Model II. The hydrogen bonds at the complex interface were shown as dashed black lines and labeled on the structure in VDW mode. The orange spheres represent the residue on Mac-1 and the green spheres represent the residue on GPIIbα. The labels here see reference (Kobe and Kajava, 2001).

TABLE 1 | Hydrogen bonds on the binding site of the complex.

No	Residue pair		Occupancy		
	Mac-1	GPIIbα	Model I	Model II	Model III
1	K244	D18	0.39	0.77	0.74
2	E282	K19	0.62	0.66	0.67
3	S288	D235	0.23	0.52	0.57
4	E242	K19	0.35	0.44	0.43
5	E252	S39	0.31	0.44	0.53
6	Y251	R64	0.17	0.34	0.53
7	K278	E40	0.67	0.32	0.28
8	E261	K237	0	0.54	0.59
9	R216	D63	0	0.31	0.25
10	K278	Q66	0	0.27	0.23
11	D259	K231	0	0.23	0.15
12	K278	R64	0	0.16	0.12
13	H294	K231	0	0.16	0.2
14	K289	K231	0	0.16	0.18

meaning the closer contact between Mac-1 and GPIIbα in Models II and III than that in Model I. The dissociation probabilities (f_D) of complex were estimated to be 0.02, 0.0005, and 0.0009 for Models I, II, and III (Figure 2E), showing that the Mac-1 affinity to GPIIbα for the Model I was down estimated and could be restored to the quasi-actual level by a DMO treatment based on a “force-ramp + snapback” MD simulation. These results demonstrated that in comparison with Model I, Models II and

III were more energy-rational and more thermostable. And, we further performed the so called force-ramp SMD simulations thrice with constant pull velocity (3 nm/s) for Models I, II, and III (Materials and Methods) to evaluate the mechanical strength of the models. The force-time curves exhibited that the rupture force of the complex was 150 pN about for Model I but 300 pN about for Models II and III, suggesting a high mechano-strength in Models II and III rather than in Model I (Figure 2F; Supplementary Figure S2E). Under pulling with a constant velocity of 3 Å/ns, the Mac-1/GPIIbα complex remained structure-stable under a pulling force <250 pN for Models II and III or 100 pN for Model I (Supplementary Figure S2E; Supplementary Videos S1, S2). These results demonstrated that in comparison with Model I, Models II and III were more energy-rational, more thermo- and mechano-stable in modeling the Mac-1/GPIIbα complex. For these reasons, Model II was regarded as the likely wild-type model of the Mac-1/GPIIbα complex and used as an initial conformation for the subsequent “ramp-clamp” SMD simulations.

Dissociation of the Stretched Mac-1–GPIIbα Complex Was Biphasic Force Dependent

To examine the mechano-regulation on the interaction of Mac-1 with GPIIbα, we performed a series of “ramp-clamp” SMD simulations of 40 ns thrice with Model II under constant

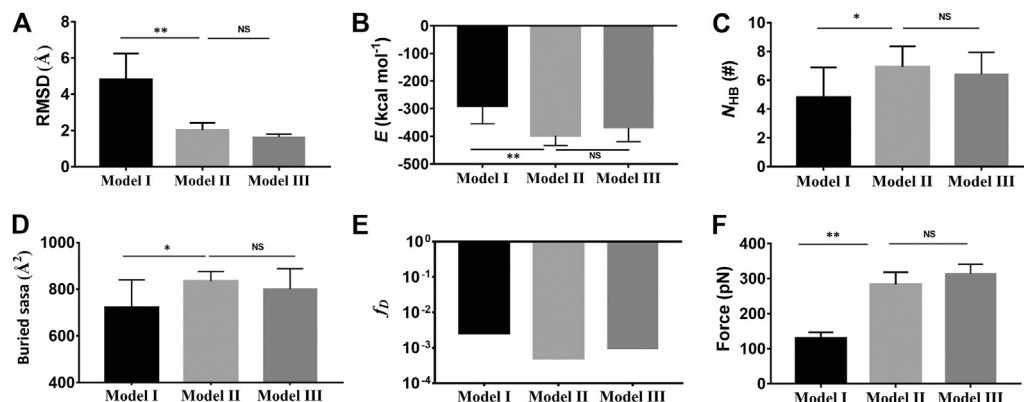


FIGURE 2 | Comparison of the complex characters in Model II (light gray) and Model III (dark gray) with those in Model I (black). **(A)** The mean C_{α} -RMSD, **(B)** the mean binding energy E (-293 ± 75 kcal/mol for Model I, -398 ± 53 kcal/mol for the II, and -369 ± 50 kcal/mol for Model III), **(C)** the mean interface H-bond number N_{HB} (4.8 ± 2 for Model I, 6.9 ± 1.4 for Model II, and 5.8 ± 1.5 for Model III), **(D)** the mean buried SASA (730 \AA^2 for Model I, 840 \AA^2 for Model II, and 800 \AA^2 for Model III), and **(E)** the dissociation probability f_D for complex in 40 ns equilibrium. **(F)** The mean rupture force (120 pN about for Model I, 300 pN for Model II, and 313 pN for Model III) in “force-ramp” SMD simulations thrice on complex with a velocity of 3 \AA/ns . The p -values of the unpaired two-tailed Student’s t test were shown to indicate the statistical difference significance (**** $p < 0.0001$), or lack thereof. Data were shown with mean \pm S.D.

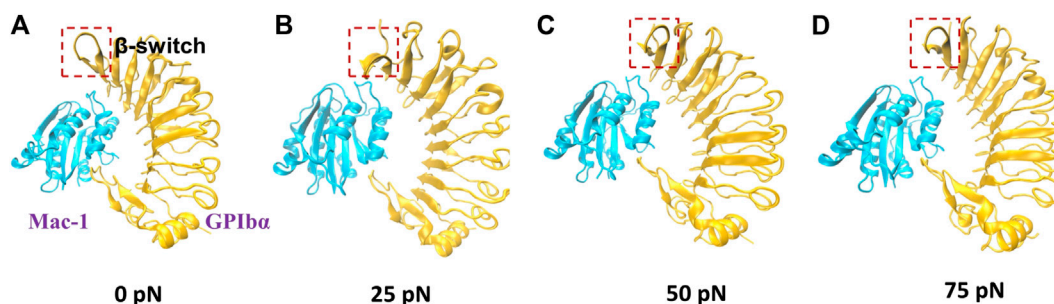


FIGURE 3 | The conformations of Mac-1/GPIIb α complex (Model II; new cartoon) after 40 ns “clamp-force” SMD simulation with tensile forces 0 **(A)**, 25 **(B)**, 50 **(C)**, and 75 pN **(D)**. Mac-1 and GPIIb α were shown as cyan and orange, respectively. The main conformation changes of β -switch, which were marked with a red box, where shown in **(A–D)**.

tensile forces of 0, 25, 50, and 75 pN (Materials and Method). Just a bit force-induced conformational change of complex is shown in **Figure 3**, meaning that Model II was reliable for its fine mechanical stability. And, the mechanical stability of the complex was also demonstrated by the very slight tension-induced increasing of the C_{α} -RMSD of the complex (**Figure 4A**) and distance between the pulled- and fixed-atom (**Figure 4B**), while the sampled structural space was regarded as quasi-perfect because the H-bond number obeyed the Gaussian distribution (**Figure 4C**), meaning that the complex conformations sampled within a simulation time of 40 ns under each given constant tensile force were enough in gaining information of the structure–function relation of the complex.

The interaction energies, the buried SASA, and the H-bonds (or salt bridges) on the binding site for the complex under constant tensile forces were sampled through the “ramp-clamp” SMD simulations with Model II (Materials and Method). Plots of the mean interaction energy (E), the mean buried SASA, and the mean number of the H-bonds (N_{HB}) (or

salt bridges) on the binding site over 40 ns for three runs against tensile force exhibited that E decreased first and then increased with F , and the force threshold occurred at 25 pN (**Figure 4D**), demonstrating a biphasic force-dependent energy preference for the stretched GPIIb α –Mac-1 complex; on the contrary, N_{HB} increased first and then decreased with F (**Figure 4E**), illustrating a transition from force-enhanced to force-weakened linkage between GPIIb α and Mac-1; as a result, f_D , the normalized complex dissociation probability, decreased first and then increased F (**Figure 4F**), suggesting a catch-slip bond transition in Mac-1 dissociation from GPIIb α . All the transition points for E , N_{HB} , f_D , and the mean buried SASA occurred at a tensile force of 25 pN, as it should be. These results were in keeping with our single molecular AFM measurement data, which exhibited a catch-slip bond transition with a force threshold of about 31 pN (**Figure 5**; **Supplementary Figure S3**) for interaction between Mac-1 with GPIIb α . The catch-slip bond transition had been measured by AFM and BFP as well as flow chamber experiments for various adhesive molecule systems,

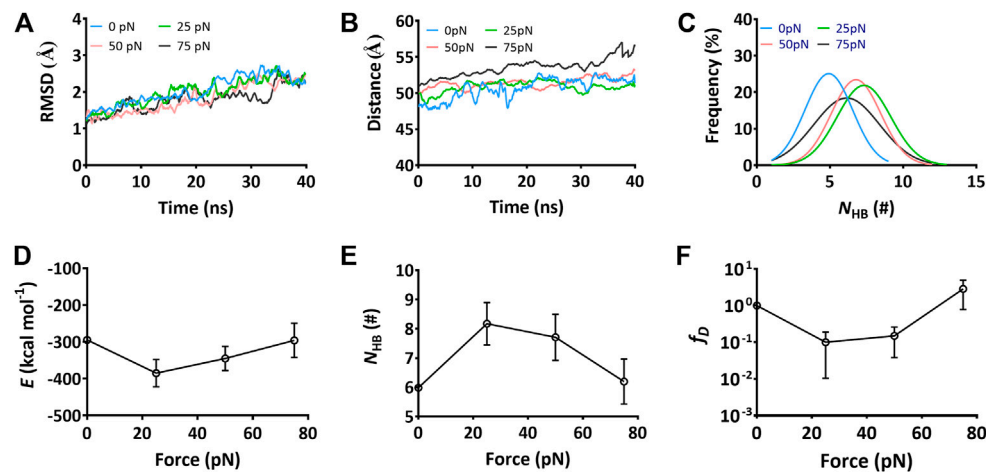


FIGURE 4 | Variations of the structural stability and interface interaction of the complex versus constant tensile force. Data were read from the thrice 40-ns “force-clamp” SMD simulations on Model II of the Mac-1–GPIIbα complex. **(A)** The representative time courses of the C_{α} -RMSD of complex at tensile forces of 0 (blue), 25 (green), 50 (red), and 75 pN (black). The C_{α} -RMSD fluctuated in a range from 1.0 to 2.5 Å for each tensile force. **(B)** The distance–time plot at tensile forces of 0, 25, 50, and 75 pN. The distance between the pulled- and fixed-atom fluctuated with an amplitude <5 Å around a plateau for each tensile force. **(C)** Gaussian fitting of the N_{HB} frequencies from thrice 40-ns runs at various tensile forces. **(D)** and **(E)** The variations of the mean binding energy E and the mean H-bond number N_{HB} on binding site over 40 ns for three runs versus the tensile force. **(F)** The normalized dissociation probability f_D of complex under various tensile forces. Pearson correlation coefficients for E , N_{HB} , and f_D are -0.832 , 0.879 , and -0.987 if $0 \leq \text{force} \leq 25$ pN but take values of 0.595 , -0.766 , and 0.749 as $25 \text{ pN} < \text{force} \leq 75$ pN, respectively, with $p < 0.05$, statistically demonstrating the dependences of E , N_{HB} , and f_D on the tensile force. The data in **(D)**, **(E)**, and **(F)** were shown as the mean \pm SD.

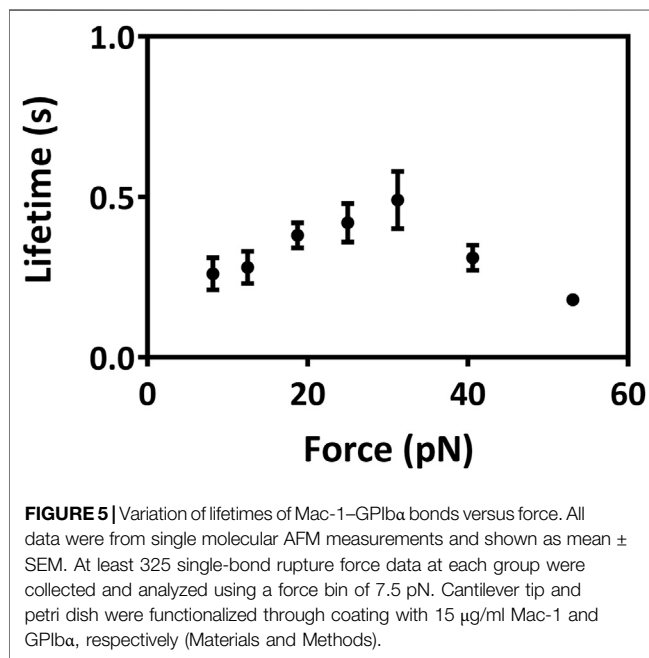


FIGURE 5 | Variation of lifetimes of Mac-1–GPIIbα bonds versus force. All data were from single molecular AFM measurements and shown as mean \pm SEM. At least 325 single-bond rupture force data at each group were collected and analyzed using a force bin of 7.5 pN. Cantilever tip and petri dish were functionalized through coating with 15 $\mu\text{g}/\text{ml}$ Mac-1 and GPIIbα, respectively (Materials and Methods).

such as von Willebrand factor with GPIIbα (Yago et al., 2008), ADMAMTS13 (Wu et al., 2010), PSGL-1 with P-, E-, and L-selectins as well as the PSGL-1-actin cytoskeleton linker protein ezrin/radixin/myosin (ERM) (Marshall et al., 2003; Yago et al., 2004; Li et al., 2016), and so on.

It signed that the computer strategy with “ramp-clamp” SMD simulation was practicable in examining the mechano-regulation on receptor–ligand interactions, as done in our previous work for the

interaction of PSGL-1 with ERM (Feng et al., 2020) or Kindlin 2 with β_3 integrin (Zhang et al., 2020) and Model II, a docking model treated with “force-ramp + snapback” SMD simulation, was suitable in studying the structure–function relation for the complex of Mac-1 with GPIIbα.

Force-Induced Allostery in Mac-1 Dissociation From GPIIbα

To scale the force-induced allostery of the Mac-1 bound with GPIIbα, we herein measured θ (Figure 6A), the mean angle between $\alpha 1$ and $\alpha 7$ helix of the ligated Mac-1 over 40 ns simulation time thrice under each tensile force. The angle θ increased remarkably first and then decreased with F (Figure 6B), was correlative negatively to the normalized complex dissociation probability f_D (Figure 4F) and the interaction energy E (Figure 4D) but positively to the H-bond number N_{HB} (Figure 4E) and the mean buried SASA (Figure 6C), suggesting that the force-induced allostery of the ligated Mac-1 might be responsible for the catch-slip bond transition in the interaction of Mac-1 with GPIIbα. An observation for the $\alpha 7$ helix of Mac-1 exhibited a descent of Mac-1 affinity to GPIIbα due to the downward change of the $\alpha 7$ helix (Figure 6A).

Besides, we measured L_{MB} , the distance from the C_{α} atom of the residue D235 in the β -switch to the mass center of GPIIbα (Figure 6D), to scale the deviation of the β -switch from its neighbor subdomains under various tensile forces. Plots of L_{MB} against tensile force (Figure 6F) said that increasing tensile force made L_{MB} lengthened significantly first and then shortened slightly, and the turning point occurred at the tensile force of 25 pN too, demonstrating that a limit on Mac-1 dissociation from GPIIbα might be provided through the β -switch deviating from GPIIbα body. Together with the force-induced allostery of the

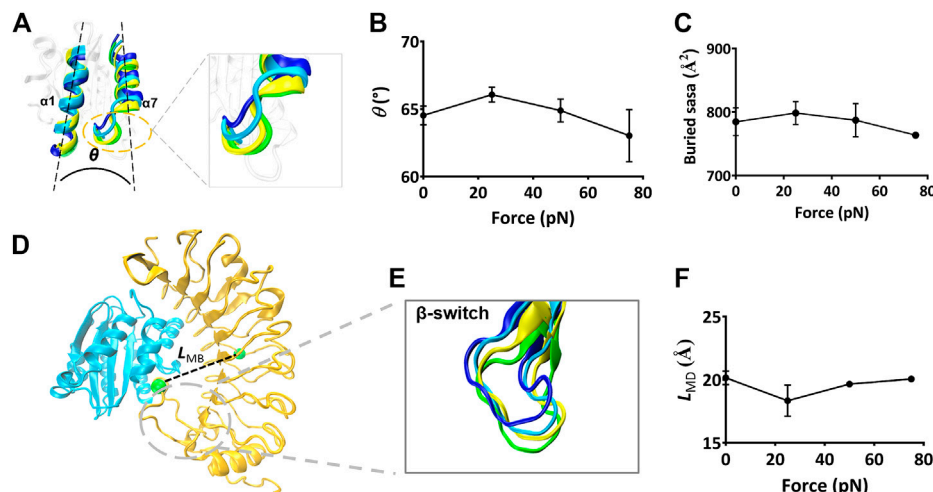


FIGURE 6 | The force-induced conformational change of the Mac-1-GPIIbα complex. **(A)** The cross angle θ between $\alpha 7$ - and $\alpha 1$ -helix of the ligated Mac-1. The four representative conformations of $\alpha 1$ - and $\alpha 7$ -helix under various tensile forces of 0, 25, 50, and 75 pN were colored in blue, cyan, yellow, and green, respectively. The superposition of these different conformations showed a force-induced down-movement of $\alpha 7$ helix tail. **(B)** The plot of θ against tensile force. θ was averaged over the simulation time of 40 ns for three runs. The mean values are 64.5, 66.5, 64.4, and 62.3 degrees, respectively. **(C)** Variation of the mean buried SASA over the three 40-ns runs versus tensile force. **(D)** The distance L_{MB} (dashed black line) between the two green spheres, which were located respectively at the centroid and the D235 residue in the bound GPIIbα. **(E)** Superposition of four representative conformations of β -switch (residues 227–241) of GPIIbα under various tensile force of 0 (blue), 25 (cyan), 50 (yellow), and 75 pN (green). A force-induced down-movement of the β -switch was shown in the structural superposition. **(F)** Variation of L_{MB} versus the tensile force. Pearson correlation coefficients for θ , SASA, and L_{MB} are 0.852, 0.408, and -0.735 if $0 \leq \text{force} \leq 25$ pN but take values of -0.606 , -0.612 , and 0.885 as $25 \text{ pN} < \text{force} \leq 75$ pN, respectively, with $p < 0.05$, statistically demonstrating the dependences of E , N_{HB} , and f_D on the tensile force. The data in **(B)**, **(C)**, and **(F)** were shown as the mean \pm SD.

TABLE 2 | H-bonds (with occupancies in top 11) on the binding site of the complex under various tensile forces.

No	Residue Pair		Occupancy			
	Mac 1	GPIIbα	0 (pN)	25 (pN)	50 (pN)	75 (pN)
1	E252	K37	0.08	0.19 \pm 0.11	0.43 \pm 0.15	0.03 \pm 0.009
2	E252	S39	0.44	0.67 \pm 0.12	0.83 \pm 0.03	0.39 \pm 0.18
3	E252	R64	0.10	0.74 \pm 0.03	0.48 \pm 0.23	0.49 \pm 0.20
4	E261	K237	0.54	0.58 \pm 0.01	0.57 \pm 0.03	0.39 \pm 0.19
5	K278	E40	0.32	0.62 \pm 0.04	0.55 \pm 0.08	0.56 \pm 0.007
6	S288	D235	0.52	0.79 \pm 0.08	0.77 \pm 0.04	0.62 \pm 0.05
7	K244	D18	0.77	0.61 \pm 0.14	0.66 \pm 0.03	0.58 \pm 0.14
8	Y251	R64	0.34	0.14 \pm 0.03	0.28 \pm 0.08	0.16 \pm 0.09
9	D259	K231	0.23	0.48 \pm 0.03	0.25 \pm 0.12	0.33 \pm 0.16
10	E243	K19	0.44	0.37 \pm 0.08	0.24 \pm 0.14	0.19 \pm 0.08
11	E282	K19	0.66	0.64 \pm 0.06	0.47 \pm 0.1	0.48 \pm 0.13

ligated Mac-1 (Figures 6A,B), the force-mediated deviation of the β -switch (Figure 6E) also might be responsible for the force-dependent mean buried SASA of the binding sites and the dissociation of Mac-1 from GPIIbα.

The Key Residues in the Biphasic Force-Dependent Mac-1-GPIIbα Interaction

To reveal the structural basis of the biphasic force-dependent Mac-1-GPIIbα interaction mentioned above, we examined the H-bonding interactions on the binding site through “force-clamp” SMD simulation of 40 ns thrice under tensile forces of

0, 25, 50, and 75 pN, and evaluated the probabilities for unbinding of either the residues in Mac-1 from GPIIbα or the residues in GPIIbα from Mac-1 (Materials and Methods). The variation of the H-bond occupancy versus tensile force (Table 2) demonstrated that the residue–residue interactions on the binding site were mechano-sensitive, and increasing tensile force might make H-bonding occurred, broken, strong, or weak, exhibiting a diversity for the H-bonds in response to the tensile force. Of all detected H-bonds, those with mean occupancies >0.20 had eleven members (Table 2; Supplementary Table S2) which could be clustered into four groups in responding to the tensile force with modes of the “slip-bond type,” the “catch-slip bond” type, the “slip-catch-slip bond,” and the “catch-slip-catch bond” type, respectively. The first group was contributed by K19 on GPIIbα with its two partners E243 and E282 on Mac-1; the second was consisted of those such as D235 on GPIIbα paired with S288, K278, and E261 on Mac-1 paired with their respective partners E40 and K237 on GPIIbα, as well as E252 on Mac-1 with its three partners K37, S39, and R64 on GPIIbα; the third included those of K244 and Y251 on Mac-1 paired with their respective partners, D18 and R64 on GPIIbα; and the fourth was contributed only by D259 on Mac-1 paired with K231 on GPIIbα (Table 2). These suggested that the force-induced changes of conformation and function for either Mac-1 or GPIIbα in complex should be mediated by the cooperative interaction of the H-bonds in responding to the tensile force with different modes.

Based on the contributions on strong interface H-bonding with high occupancies ($>50\%$), the six residues, such as K244,

E252, E261, K278, E282, and S288 on Mac-1, might be responsible for the force-dependent interaction of Mac-1 with GPIIb/IIIa, despite just K244 in these residues was mutation-identified (Simon et al., 2000; Wang et al., 2005). Other three mutation-identified residues, such as T211 and T213 as well as R216 on Mac-1 (Simon et al., 2000; Wang et al., 2005), did not emerge from the above six key interface residues because of either nothing for T211 and T213 or less contribution for R216 to H-bonding on the complex interface. This inconsistency of the mutation-experimental data and the computational results might come from the timescale effects on milliseconds MD simulations in predicting the receptor–ligand interaction in a period of 1 s or its tenth (Schwantes et al., 2014).

DISCUSSION

It was believed that MD simulation might predict druggable binding site on complex interface, while providing a valuable dynamics insight to receptor–ligand interaction mechanism. However, a near-native docking model should be required for a meaningful MD simulation, under the lack of solved structural data of the complex. A rational assumption said that Mac-1/GPIIb/IIIa complex should be thermo- and mechano-stable, like the GPIIb/IIIa–VWF complex because of the structural similarity between the two complexes. However, it is still a technical challenge to make a complex docking model near-native. We herein proposed a novel computer strategy to make the Model I (the docking model of Mac-1/GPIIb/IIIa complex) more near-native or rational. This strategy for structural improvement included a system equilibrium followed a “force-ramp + snapback” SMD simulation of 105 ns, in which 5 ns was spent for SMD simulation with a constant pulling velocity of 3 nm/ns and the other 100 ns for a system re-equilibration for the unloaded or snapped-back complex (Materials and Methods). The interface H-bonds in Model II were more and stronger than those in Model I, saying that the present computer strategy with a treatment of “force-ramp + snapback” MD simulation might be feasible for improving a docking model, and meaning that the barrier in the transition from a nonnative complex conformational model to a near-native one might overcome through adding a mechano-perturbation on the complex. However, the effectiveness and efficiency of our “force-ramp + snapback” methodology relies on a suitable choice of fixed and pulled atoms, as well as pulling velocity and direction. We have shown that given a rational choice of these parameters, our methodology can improve the quality of a modeled dimer. To conclusively demonstrate the general applicability of our methodology, tests on multiple docked models of different dimers should be carried out.

As a key event in hemostasis and inflammatory responses to vessel injuries, the crosstalk between platelet and neutrophil would be mediated by GPIIb/IIIa–Mac-1 interaction in hemodynamic environments. Lack of crystal structural data led to less knowledge on the mechano-regulation and its molecular basis on GPIIb/IIIa–Mac-1 interaction under shear stress conditions, despite those mutation-identified residues, such as T211, T213, K244, and R216 on Mac-1, were

demonstrated to be crucial for binding of Mac-1 to GPIIb/IIIa (Simon et al., 2000; Wang et al., 2005). In mediating adhesion and accumulation of circulating platelets, GPIIb/IIIa–vWF interaction was governed by a catch-slip bond mechanism, saying a force-induced prolongation of bond lifetime for complex under loads below a force threshold (Da et al., 2014). This catch-slip bond was also observed herein not only from AFM measurements for GPIIb/IIIa–Mac-1 interaction but also from a series of “ramp-clamp” mode SMD simulations with Model II of the GPIIb/IIIa–Mac-1 complex under various tensile forces (Figure 4). It might come from the structural similarity between Mac-1 I-domain and VWF-A1 domain with the major binding site for GPIIb/IIIa (Diamond et al., 1993). This better consistency of the experimental data and computational predictions might provide another support to Model II of the GPIIb/IIIa–Mac-1 complex, despite that the catch-slip bond transition occurred at 31 pN in AFM experiments but 25 pN in the “ramp-clamp” SMD simulations. The catch-slip bond phenomenon in GPIIb/IIIa–Mac-1 interaction was observed in various adhesive molecular systems, such as selectins with PSGL-1 (Marshall et al., 2003), β_2 integrin ($\alpha_L\beta_2$, $\alpha_M\beta_2$) with ICAM-1 (Kong et al., 2009; Chen et al., 2011), and VWF-A1 with GPIIb/IIIa (Yago et al., 2008; Lining et al., 2015).

The force-induced allostery of the mutually constrained Mac-1 and GPIIb/IIIa was stable for a given tensile force (Figure 3) and might synergize beneficially to induce the “catch-slip bond” phenomenon in the interaction of Mac-1 and GPIIb/IIIa. We obtained that the catch bond in the interaction of Mac-1 to GPIIb/IIIa might be derived from an increasing flexibility of the α_M domain α_1 helix and a force-induced downward movement of the α_M domain α_7 helix of the bound Mac-1, similar to the α_7 helix shifting downward and the outward movement of the α_1 helix in the force-induced conformational transition of the ICAM-1-bound LFA-1 (Chen et al., 2011). The force-induced change of angle between α_1 and α_7 helix came from the swing of α_7 tail spiral (Figures 6A,B), suggesting that α_7 helix was responsible for the affinity of the bound Mac-1. The force-induced change of the GPIIb/IIIa-binding pocket (the β -switch) might regulate the GPIIb/IIIa affinity to Mac-1 (Figures 6D–F), in consistency with the interaction of VWF A1 domain with GPIIb/IIIa.

Usually, MD simulation results at the atom level were not comparable to the single molecular measurement data. The barriers might mainly come from the timescale effects on MD simulation results in predicting receptor–ligand interactions, due to that affinity change and conformation evolution of adhesive molecules would undergo a period far longer than the simulation timescale from nanoseconds to milliseconds (Schwantes et al., 2014). These timescale effects might be overcome through a suitable computer strategy such as the “ramp-clamp” SMD simulation, as shown in the better consistency of AFM experimental data with MD simulation results. With Model II of the Mac-1/GPIIb/IIIa complex, the identified residues D222 and R218 on Mac-1 were predicted to be the key, showing the rationality of Model II and the availability of the present computer strategy. However, the random feature and the

initial-state dependence of conformational evolution might lead to fail in detecting the identified residues (T211 and T213) on Mac-1 and N223 on GPIIb herein, but enough simulations in parallel might be beneficial in locating this residue.

In conclusion, a rational docking model (Model II) for the Mac-1/GPIIb complex was built herein through the present computer strategy, and shown to be thermo- and mechano-stable. A slip-catch bond transition phenomenon was observed not only from the “ramp-clamp” SMD simulations with Model II under various tensile forces but also from AFM experiments. The force-enhanced interaction of Mac-1 to GPIIb under force below a force threshold might be required for stable crosstalk between platelets and neutrophils in mechano-microenvironments around the injured vessel sides. The present work provided not only an effective computer strategy to build a likely wild-type model of Mac-1 bound to GPIIb but also a novel insight into the mechano-regulation mechanism and its molecular structure basis for Mac-1–GPIIb interaction and should be helpful for understanding the force-dependent platelet–leukocyte interactions in hemostasis and inflammatory responses under flows.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

REFERENCES

- Behnen, M., Leschczyk, C., Moller, S., Batel, T., Klinger, M., Solbach, W., et al. (2014). Immobilized immune complexes induce neutrophil extracellular trap release by human neutrophil granulocytes via FcγRIIIB and Mac-1. *J. Immunol.* 193 (4), 1954–1965. doi:10.4049/jimmunol.1400478
- Chen, W., Lou, J., and Zhu, C. (2011). Forcing switch from short- to intermediate- and long-lived states of the alphaA domain generates LFA-1/ICAM-1 catch bonds. *J. Biol. Chem.* 285 (46), 35967–35978. doi:10.1074/jbc.M110.155770
- Coxon, A., Rieu, P., Barkalow, F. J., Askari, S., Sharpe, A. H., von Andrian, U. H., Arnaout, M. A., et al. (1996). A novel role for the beta 2 integrin CD11b/CD18 in neutrophil apoptosis: a homeostatic mechanism in inflammation. *Immunity* 5 (6), 653–666. doi:10.1016/s1074-7613(00)80278-2
- Da, Q., Behymer, M., Correa, J. I., Vijayan, K. V., and Cruz, M. A. (2014). Platelet adhesion involves a novel interaction between vimentin and von Willebrand factor under high shear stress. *Blood* 123 (17), 2715–2721. doi:10.1182/blood-2013-10-530428
- Diacovo, T. G., Roth, S. J., Buccola, J. M., Bainton, D. F., and Springer, T. A. (1996). Neutrophil rolling, arrest, and transmigration across activated, surface-adherent platelets via sequential action of P-selectin and the beta 2-integrin CD11b/CD18. *Blood* 88 (1), 146–157. doi:10.1182/blood.v88.1.146.bloodjournal881146
- Diamond, M. S., Alon, R., Parkos, C. A., Quinn, M. T., and Springer, T. A. (1995). Heparin is an adhesive ligand for the leukocyte integrin Mac-1 (CD11b/CD18). *J. Cell Biol.* 130 (6), 1473–1482. doi:10.1083/jcb.130.6.1473
- Diamond, M. S., Garcia-Aguilar, J., Bickford, J. K., Bickford, J. K., Corbi, A. L., and Springer, T. A. (1993). The I domain is a major recognition site on the leukocyte integrin Mac-1 (CD11b/CD18) for four distinct adhesion ligands. *J. Cell Biol.* 120 (4), 1031–1043. doi:10.1083/jcb.120.4.1031
- Ehlers, R., Ustinov, V., Chen, Z., Zhang, X., Rao, R., Luscinskas, F. W., et al. (2003). Targeting platelet-leukocyte interactions: identification of the integrin Mac-1 binding site for the platelet counter receptor glycoprotein Ibalpha. *J. Exp. Med.* 198 (7), 1077–1088. doi:10.1084/jem.20022181
- Fang, X., Fang, Y., Liu, L., Liu, G., and Wu, J. (2012). Mapping paratope on antithrombotic antibody 6B₄ to epitope on platelet glycoprotein Ibalpha via

AUTHOR CONTRIBUTIONS

JW and YF designed this research; XJ overall performed the research and analyzed the data; XS and YL partly performed the research and analyzed the data; XJ, XS, JL, YF, and JW wrote this paper.

FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC) Grant 11672109 (to YF), 11432006 (to JW), 11702100 (YL), and 31771012 (JL).

ACKNOWLEDGMENTS

The authors thank Xuebin Xie for his help in data treatment of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.638396/full#supplementary-material>.

molecular dynamic simulations. *PLoS One* 7 (7), e42263. doi:10.1371/journal.pone.0042263

- Fang, X., Lin, J., Fang, Y., and Wu, J. (2018). Prediction of spacer-α6 complex: a novel insight into binding of ADAMTS13 with A2 domain of von Willebrand factor under forces. *Sci. Rep.* 8 (1), 5791. doi:10.1038/s41598-018-24212-6
- Feng, J., Zhang, Y., Li, Q., Fang, Y., and Wu, J. (2020). Biphasic force-regulated phosphorylation site exposure and unligation of ERM bound with PSGL-1: a novel insight into PSGL-1 signaling via steered molecular dynamics simulations. *Int. J. Mol. Sci.* 21 (19), 7064. doi:10.3390/ijms21197064
- Grubmüller, H., Heymann, B., and Tavan, P. (1996). Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* 271 (5251), 997–999. doi:10.1126/science.271.5251.997
- Hidalgo, A., Chang, J., Jang, J. E., Peired, A. J., Chiang, E. Y., and Frenette, P. S. (2009). Heterotypic interactions enabled by polarized neutrophil microdomains mediate thromboinflammatory injury. *Nat. Med.* 15 (4), 384–391. doi:10.1038/nm.1939
- Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graphics* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5
- Jain, A. N. (2003). Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* 46 (4), 499–511. doi:10.1021/jm020406h
- Karshovska, E., Weber, C., and Hundelshausen, P. (2013). Platelet chemokines in health and disease. *Thromb. Haemost.* 110 (11), 894–902. doi:10.1160/TH13-04-0341
- Kobe, B., and Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* 11 (6), 725–732. doi:10.1016/s0959-440x(01)00266-4
- Kong, F., García, A. J., Mould, A. P., Humphries, M. J., and Zhu, C. (2009). Demonstration of catch bonds between an integrin and its ligand. *J. Cell Biol.* 185 (7), 1275–1284. doi:10.1083/jcb.200810002
- Kruss, S., Erpenbeck, L., Amschler, K., Mundinger, T. A., Boehm, H., Helms, H. J., et al. (2013). Adhesion maturation of neutrophils on nanoscopically presented platelet glycoprotein Iba. *ACS Nano* 7 (11), 9984–9996. doi:10.1021/nn403923h
- Lee, C. Y., Lou, J., Wen, K. K., McKane, M., Eskin, S. G., Rubenstein, P. A., et al. (2016). Regulation of actin catch-slip bonds with a RhoA-formin module. *Sci. Rep.* 6, 35058–42322. doi:10.1038/srep35058

- Li, N., Yang, H., Wang, M., Lü, S., Zhang, Y., and Long, M. (2018). Ligand-specific binding forces of LFA-1 and Mac-1 in neutrophil adhesion and crawling. *Mol. Biol. Cell* 29 (4), 408–418. doi:10.1091/mbc.E16-12-0827
- Li, Q., Wayman, A., Lin, J., Fang, Y., Zhu, C., and Wu, J. (2016). Flow-enhanced stability of rolling adhesion through E-selectin. *Biophys. J.* 111 (4), 686–699. doi:10.1016/j.bpj.2016.07.014
- Li, R., and Emsley, J. (2013). The organizing principle of the platelet glycoprotein Ib-IX-V complex. *J. Thromb. Haemost.* 11 (4), 605–614. doi:10.1111/jth.12144
- Lining, J., Yunfeng, C., Fangyuan, Z., Hang, L., Cruz, M. A., and Cheng, Z. (2015). Von Willebrand factor-A1 domain binds platelet glycoprotein Iba in multiple states with distinctive force-dependent dissociation kinetics. *Thromb. Res.* 136 (3), 606–612. doi:10.1016/j.thromres.2015.06.019
- MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102 (18), 3586–3616. doi:10.1021/jp973084f
- Marshall, B. T., Long, M., Piper, J. W., Yago, T., McEver, R. P., and Zhu, C. (2003). Direct observation of catch bonds involving cell-adhesion molecules. *Nature* 423 (6936), 190–193. doi:10.1038/nature01605 Epub 2003/05/09.
- McDonald, B., Pittman, K., Menezes, B. G., Hirota, S. A., Slaba, I., Waterhouse, C. C. M., et al. (2010). Intravascular danger signals guide neutrophils to sites of sterile inflammation. *Science* 330 (6002), 362–366. doi:10.1126/science.1195491
- McEver, R. P., and Cummings, R. D. (1997). Perspectives series: cell adhesion in vascular biology. Role of PSGL-1 binding to selectins in leukocyte recruitment. *J. Clin. Invest.* 100 (11 Suppl. 1), 485–491. doi:10.1172/JCI119556
- Morgan, J., Saleem, M., Ng, R., Armstrong, C., Wong, S. S., Caulton, S. G., et al. (2019). Structural basis of the leukocyte integrin Mac-1 I-domain interactions with the platelet glycoprotein Ib. *Blood Adv.* 3 (9), 1450–1459. doi:10.1182/bloodadvances.2018027011
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., and Tajkhorshid, E. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26 (16), 1781–1802. doi:10.1002/jcc.20289
- Phillipson, M., Heit, B., Colarusso, P., Liu, L., Ballantyne, C. M., and Kubes, P. (2006). Intraluminal crawling of neutrophils to emigration sites: a molecularly distinct process from adhesion in the recruitment cascade. *J. Exp. Med.* 203 (12), 2569–2575. doi:10.1084/jem.20060925
- Rahman, S. M., and Hlady, V. (2019). Downstream platelet adhesion and activation under highly elevated upstream shear forces. *Acta Biomater.* 91, 135–143. doi:10.1016/j.actbio.2019.04.028
- Sadler, J. E., Shelton-Inloes, B. B., Sorace, J. M., Harlan, J. M., Titani, K., and Davie, E. W. (1985). Cloning and characterization of two cDNAs coding for human von Willebrand factor. *Proc. Natl. Acad. Sci. U.S.A.* 82 (19), 6394–6398. doi:10.1073/pnas.82.19.6394
- Schrottmaier, W. C., Kral, J. B., Badrnya, S., and Assinger, A. (2015). Aspirin and P2Y12 Inhibitors in platelet-mediated activation of neutrophils and monocytes. *Thromb. Haemost.* 114 (3), 478–489. doi:10.1160/TH14-11-0943
- Schwantes, C. R., McGibbon, R. T., and Pande, V. S. (2014). Perspective: markov models for long-timescale biomolecular dynamics. *J. Chem. Phys.* 141 (9), 090901. doi:10.1063/1.4895044
- Silva, J. C., Rodrigues, N. C., Thompson-Souza, G. A., Muniz, V. S., Neves, J. S., and Figueiredo, R. T. (2020). Mac-1 triggers neutrophil DNA extracellular trap formation to *Aspergillus fumigatus* independently of PAD4 histone citrullination. *J. Leukoc. Biol.* 107 (1), 69–83. doi:10.1002/JLB.4A0119-009RR
- Simon, D. I., Chen, Z., Xu, H., Li, C. Q., Dong, J. F., McIntire, L. V., et al. (2000). Platelet glycoprotein Ibalph is a counterreceptor for the leukocyte integrin Mac-1 (CD11b/CD18). *J. Exp. Med.* 192 (2), 193–204. doi:10.1084/jem.192.2.193
- Simon, D. I. (2012). Inflammation and vascular injury: basic discovery to drug development. *Circ. J.* 76 (8), 1811–1818. doi:10.1253/circj.cj-12-0801
- Smith, G. R., Fitzjohn, P. W., Page, C. S., and Bates, P. A. (2005). Incorporation of flexibility into rigid-body docking: applications in rounds 3–5 of CAPRI. *Proteins* 60 (2), 263. doi:10.1002/prot.20568
- Sumagin, R., Prizant, H., Lomakina, E., Waugh, R. E., and Sarelius, I. H. (2010). LFA-1 and Mac-1 define characteristically different intraluminal crawling and emigration patterns for monocytes and neutrophils *in situ*. *J. Immunol.* 185 (11), 7057–7066. doi:10.4049/jimmunol.1001638
- Torchala, M., Moal, I. H., Chaleil, R. A. G., and Fernandez-Recio, J., and Bates, P. A. (2013). SwarmDock: a server for flexible protein-protein docking. *Bioinformatics* 29 (6), 807–809. doi:10.1093/bioinformatics/btt038
- Von Hundelshausen, P., and Weber, C. (2007). Platelets as immune cells: bridging inflammation and cardiovascular disease. *Circ. Res.* 100 (1), 27–40. doi:10.1161/01.RES.0000252802.25497.b7
- Wang, Y., Sakuma, M., Chen, Z., Ustinov, V., Shi, C., Croce, K., et al. (2005). Leukocyte engagement of platelet glycoprotein Ibalph via the integrin Mac-1 is critical for the biological response to vascular injury. *Circulation* 112 (19), 2993–3000. doi:10.1161/CIRCULATIONAHA.105.571315
- Whitlock, B. B., Gardai, S., Fadok, V., Bratton, V., Henson, D., and Henson, P. M. (2000). Differential roles for alpha(M)beta(2) integrin clustering or activation in the control of apoptosis via regulation of akt and ERK survival mechanisms. *J. Cell Biol.* 151 (6), 1305–1320. doi:10.1083/jcb.151.6.1305
- Wu, T., Lin, J., Cruz, M. A., Dong, J. F., and Zhu, C. (2010). Force-induced cleavage of single VWFA₁A₂A₃ tridomains by ADAMTS-13. *Blood* 115 (2), 370–378. doi:10.1182/blood-2009-03-210369
- Yago, T., Lou, J., Wu, T., Yang, J., Miner, J. J., Coburn, L., et al. (2008). Platelet glycoprotein Ibalph forms catch bonds with human WT vWF but not with type 2B von Willebrand disease vWF. *J. Clin. Invest.* 118 (9), 3195–3207. doi:10.1172/JCI35754
- Yago, T., Wu, J., Wey, C. D., Klopocki, A. G., Zhu, C., and McEver, R. P. (2004). Catch bonds govern adhesion through L-selectin at threshold shear. *J. Cell Biol.* 166 (6), 913–923. doi:10.1083/jcb.200403144
- Zeng, B., Gan, Q., Kafafi, Z. H., and Bartoli, F. J. (2013). Polymeric photovoltaics with various metallic plasmonic nanostructures. *J. Appl. Phys.* 113 (6), 659. doi:10.1063/1.4790504
- Zhang, Y., Lin, Z., Fang, Y., and Wu, J. (2020). Prediction of catch-slip bond transition of kindlin2/β3 integrin via steered molecular dynamics simulation. *J. Chem. Inf. Model.* 60 (10), 5132–5141. doi:10.1021/acs.jcim.0c00837

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

Copyright © 2021 Jiang, Sun, Lin, Ling, Fang and Wu. 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 Potential Binders of *Mtb* Universal Stress Protein (Rv1636) Through an *in silico* Approach and Insights Into Compound Selection for Experimental Validation

Sohini Chakraborti¹, Moubani Chakraborty², Avipsa Bose², Narayanaswamy Srinivasan^{1*} and Sandhya S. Visweswariah^{2*}

OPEN ACCESS

Edited by:

Joseph Rebehmed,
Lebanese American
University, Lebanon

Reviewed by:

Matteo Salvalaglio,
University College London,
United Kingdom
Stéphane Téletchéa,
Université de Nantes, France

*Correspondence:

Narayanaswamy Srinivasan
ns@iisc.ac.in
Sandhya S. Visweswariah
sandhya@iisc.ac.in

Specialty section:

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

Received: 26 August 2020

Accepted: 01 March 2021

Published: 03 May 2021

Citation:

Chakraborti S, Chakraborty M,
Bose A, Srinivasan N and
Visweswariah SS (2021) Identification
of Potential Binders of *Mtb* Universal
Stress Protein (Rv1636) Through an
in silico Approach and Insights Into
Compound Selection for Experimental
Validation.
Front. Mol. Biosci. 8:599221.
doi: 10.3389/fmolb.2021.599221

¹ Molecular Biophysics Unit, Indian Institute of Science, Bengaluru, India, ² Department of Molecular Reproduction, Development and Genetics, Indian Institute of Science, Bengaluru, India

Millions of deaths caused by *Mycobacterium tuberculosis* (*Mtb*) are reported worldwide every year. Treatment of tuberculosis (TB) involves the use of multiple antibiotics over a prolonged period. However, the emergence of resistance leading to multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) is the most challenging aspect of TB treatment. Therefore, there is a constant need to search for novel therapeutic strategies that could tackle the growing problem of drug resistance. One such strategy could be perturbing the functions of novel targets in *Mtb*, such as universal stress protein (USP, Rv1636), which binds to cAMP with a higher affinity than ATP. Orthologs of these proteins are conserved in all mycobacteria and act as “sink” for cAMP, facilitating the availability of this second messenger for signaling when required. Here, we have used the cAMP-bound crystal structure of USP from *Mycobacterium smegmatis*, a closely related homolog of *Mtb*, to conduct a structure-guided hunt for potential binders of Rv1636, primarily employing molecular docking approach. A library of 1.9 million compounds was subjected to virtual screening to obtain an initial set of ~2,000 hits. An integrative strategy that uses the available experimental data and consensus indications from other computational analyses has been employed to prioritize 22 potential binders of Rv1636 for experimental validations. Binding affinities of a few compounds among the 22 prioritized compounds were tested through microscale thermophoresis assays, and two compounds of natural origin showed promising binding affinities with Rv1636. We believe that this study provides an important initial guidance to medicinal chemists and biochemists to synthesize and test an enriched set of compounds that have the potential to inhibit *Mtb* USP (Rv1636), thereby aiding the development of novel antitubercular lead candidates.

Keywords: virtual screening, molecular docking, universal stress protein, Rv1636, anti-tubercular compounds, computational drug discovery, MM-GBSA, experimental insights

INTRODUCTION

Tuberculosis (TB), a contagious and airborne disease caused by the pathogen *Mycobacterium tuberculosis* (*Mtb*), was one of the top 10 causes of deaths worldwide in the year 2019 as per World Health Organization (WHO) global TB report 2020 (WHO, 2020). It is also a major cause of deaths in HIV patients and deaths due to antimicrobial resistance. The WHO has identified a gap of over USD 1.2 billion per year for TB research in its global TB report 2019 (WHO, 2019). The economic distress due to the ongoing coronavirus disease 2019 (COVID-19) pandemic is further threatening to stall or reverse the progress that has been achieved (WHO, 2020). Therefore, the reduction in TB disease burden calls for the scientific community's attention to contribute toward finding rational solutions for improving the current scenario. While isoniazid, rifampicin, pyrazinamide, and ethambutol are effective against drug-susceptible TB (DST-TB), multidrug resistant-TB (MDR-TB) infections do not respond to at least isoniazid and rifampicin. Extensively drug-resistant TB (XDR-TB) is a more serious problem that is resistant not only to the two key first-line drugs (isoniazid and rifampicin) but also to fluoroquinolones and second-line aminoglycosides leaving only limited options of treatment with reserved third-line drugs that possess higher toxicities. Totally drug-resistant TB (TDR-TB) correspond to the most severe forms of the infection, where all the first and second line of drugs fail to produce any response (Bahuguna and Rawat, 2020). As of August 2020, there were 22 drugs in different stages of clinical trials, including 13 new compounds: BTZ-043, delpazolid, GSK-3036656, macozinone, OPC-167832, Q203, SQ109, SPR720, sutezolid, TBAJ-876, TBA-7371, TBI-166, and TBI-223. Six approved antimicrobial drugs, namely, clofazimine, levofloxacin, linezolid, moxifloxacin, rifampicin (high dose), and rifapentine, are also undergoing trials for repurposing against TB. Host-directed therapies such as auronofin, CC-11050 (AMG 634), and everolimus are also being evaluated (WHO, 2020). Understanding the mechanism of action (MOA) of these drugs is important to formulate novel drug regimens well-tolerated by patients with comorbidities, improve cost effectiveness, and reduce therapy time. The MOA of some of the anti-TB drugs currently in the clinical pipeline has been reviewed elsewhere (Shetye et al., 2020). The introduction of promising novel anti-TB drugs like bedaquiline and delamanid in the last decade has brought new rays of hope (Koul et al., 2007; Lakshmanan and Xavier, 2013; Xavier and Lakshmanan, 2014). Unfortunately, the emergence of resistance to these drugs has also been reported (Bloemberg et al., 2015; Ghodousi et al., 2019; Nieto Ramirez et al., 2020). This calls for a constant effort to devise strategies for combating the emerging global problem of drug resistance.

An effective way to tackle drug resistance can be by targeting novel proteins that are involved in critical biological pathways in the organism and have not been targeted in the past, such as the cAMP signaling pathways. The presence of cAMP in both slow- and fast-growing mycobacteria was first noticed in the 1970s (Lowrie et al., 1975, 1979; Padh and Venkitasubramanian, 1976). These studies also showed that a large portion of cAMP (~80% for *Mycobacterium microti*) was secreted in the culture medium

(Lowrie et al., 1975). Lowrie et al. first showed the involvement of this molecule in the pathogenicity of mycobacteria. They observed a correlation between the increase in cAMP levels and the absence of phagolysosomal fusion within macrophages. This increase in cAMP was not seen upon infection with latex beads or heat-killed mycobacteria (Lowrie et al., 1975, 1979). The genome sequences of mycobacteria further endorse the importance of cAMP in their survival and virulence. Compared to other bacteria, these organisms encode a wide array of adenyl cyclases: 16 in *M. tuberculosis* and 31 in *M. marinum*—in stark contrast to the one adenyl cyclase of *Escherichia coli* (Cole et al., 1998; Stinear et al., 2008). *M. tuberculosis* also encodes 11 cAMP binding proteins, further emphasizing the significance of the second messenger in the organism (Shenoy and Visweswariah, 2006). Studies with *Mycobacterium smegmatis* showed that synthesis of cAMP was not an exclusive characteristic of slow-growing, pathogenic mycobacteria. cAMP levels in *M. smegmatis* were found to be highest during its exponential phase along with a considerable amount of secretion (Dass et al., 2008). During the infection of host alveolar macrophages, cAMP levels inside host cells increase by several folds, possibly due to secretion of cAMP as a “toxin” by the bacteria. The different fates of pathogenic vs. non-pathogenic mycobacteria within host macrophages can also be explained by changes in host's cAMP levels—non-pathogenic *M. smegmatis* causing a sustained elevation of cAMP, whereas pathogenic *M. avium* causing a transient elevation (Yadav et al., 2004). Singh et al. also observed a similar “cAMP burst” in macrophages when infected with pathogenic *M. tuberculosis* H37Rv, in contrast to a constantly elevated cAMP level when infected with non-pathogenic *M. tuberculosis* H37Ra (Singh et al., 2012). Kalamidas et al. showed that cAMP interrupts phagosomal actin assembly and, thus, prevents fusion of lysosome with phagosome and its acidification (Kalamidas et al., 2006).

Previously, we reported that a significant fraction of intracellular cAMP is bound to a mycobacterial universal stress protein (USP), Rv1636, that is abundantly expressed in both slow-growing as well as fast-growing mycobacteria (Banerjee et al., 2015). Rv1636 could possibly act as a “sink” for cAMP and release these second messengers on demand to facilitate signaling processes when required. Thorough biochemical and thermodynamic characterization of Rv1636 was subsequently performed, and the crystal structure of *M. smegmatis* USP (MSMEG_3811, a close homolog of *Mtb* USP, Rv1636) bound to cAMP was determined (Banerjee et al., 2015). Presuming that cAMP is extremely crucial for the survival and virulence of *Mtb*, targeting Rv1636 with an inhibitor could perturb overall cAMP signaling in the pathogen leading to reduced virulence.

The available chemical space to search for a potential compound that might bind to a target of interest is huge and requires high throughout compound screenings. Computational screening pipelines serve as useful tools to rationally narrow down the chemical search space in a comparatively shorter time. Furthermore, careful design of virtual chemical libraries prior to screening also generally reduces the risk of failures of drug discovery programs triggered due to toxicity (Walters et al., 1998; Mohs and Greig, 2017). In the current study, we have used the crystal structure of cAMP-bound MSMEG_3811 (PDB code:

5AHW) (Banerjee et al., 2015) as a guide to derive important knowledge about critical protein–ligand interactions. A workflow primarily driven by *in silico* approach integrated with available experimental data helped us prioritize 22 compounds that have the potential to bind to *Mtb* USP (Rv1636). These compounds were identified by computationally screening large libraries of chemical compounds (~1.9 million), including synthetic and natural compounds. Two natural compounds identified from the virtual screening have shown promising results in *in vitro* experiments. Additionally, a library of approved drugs was screened virtually to identify potential drugs that can be repurposed against Rv1636. Therefore, this study provides many potential starting points for design, synthesis, and testing of a new class of antitubercular compounds that might bind Rv1636.

MATERIALS AND METHODS

Our search for potential binders of *Mtb* USP involved a rigorous virtual screening workflow (Figure 1) comprising of four major steps, which are elaborated below.

In silico Analyses

Target Structure Selection

It is known that protein binding site residues can show structural deviations in their ligand-bound state (holo) compared to the ligand-free (apo) state. Such structural deviations can alter the binding site's shape and volume, modulating the protein–ligand recognition pattern (Fradera et al., 2002; Cozzini et al., 2008; Clark et al., 2019). Earlier studies have shown that preformed protein binding sites in holo conformation are more likely to best distinguish between binders and non-binders in virtual screenings implemented through molecular docking protocols (McGovern and Shoichet, 2003; Rueda et al., 2010).

The structure of any inhibitor/native ligand (cAMP)-bound *Mtb* USP (Rv1636; holo conformation) is currently not available in the Protein Databank (PDB) (Berman et al., 2000). The only experimentally determined structure of Rv1636 in the PDB is an apo crystal structure (PDB code: 1TQ8) (Rajashankar et al., 2004). However, a crystal structure of *M. smegmatis* USP (MSMEG_3811; PDB code: 5AHW, 2.15 Å), which is a close homolog of Rv1636 (sequence identity: 70%; Figure 2) is available. The crystal structure of MSMEG_3811 is bound to the native ligand, cAMP. Comparative analysis of the cAMP binding site residues reveal that the binding sites of MSMEG_3811 and Rv1636 are highly conserved (Figure 2). Interestingly, overlay of the cAMP-bound MSMEG_3811 structure on to the apo Rv1636 structure revealed that a few binding sites residues show considerable backbone and side-chain deviations due to a shift of a stretch of residues (residues 117–146 in 5AHW) toward the cAMP binding pocket in the holo conformation as compared to the apo state (Supplementary Figure 1). Furthermore, when compared to the crystal structure 5AHW, residues equivalent to positions 44–64 are missing in the electron density map of all the chains of the crystal structure 1TQ8. One of these residues (Met61 in 5AHW, which is equivalent to Val60 in 1TQ8) lines the cAMP binding site and can thus influence the interaction profile of docked ligands. The electron density map of one of the chains

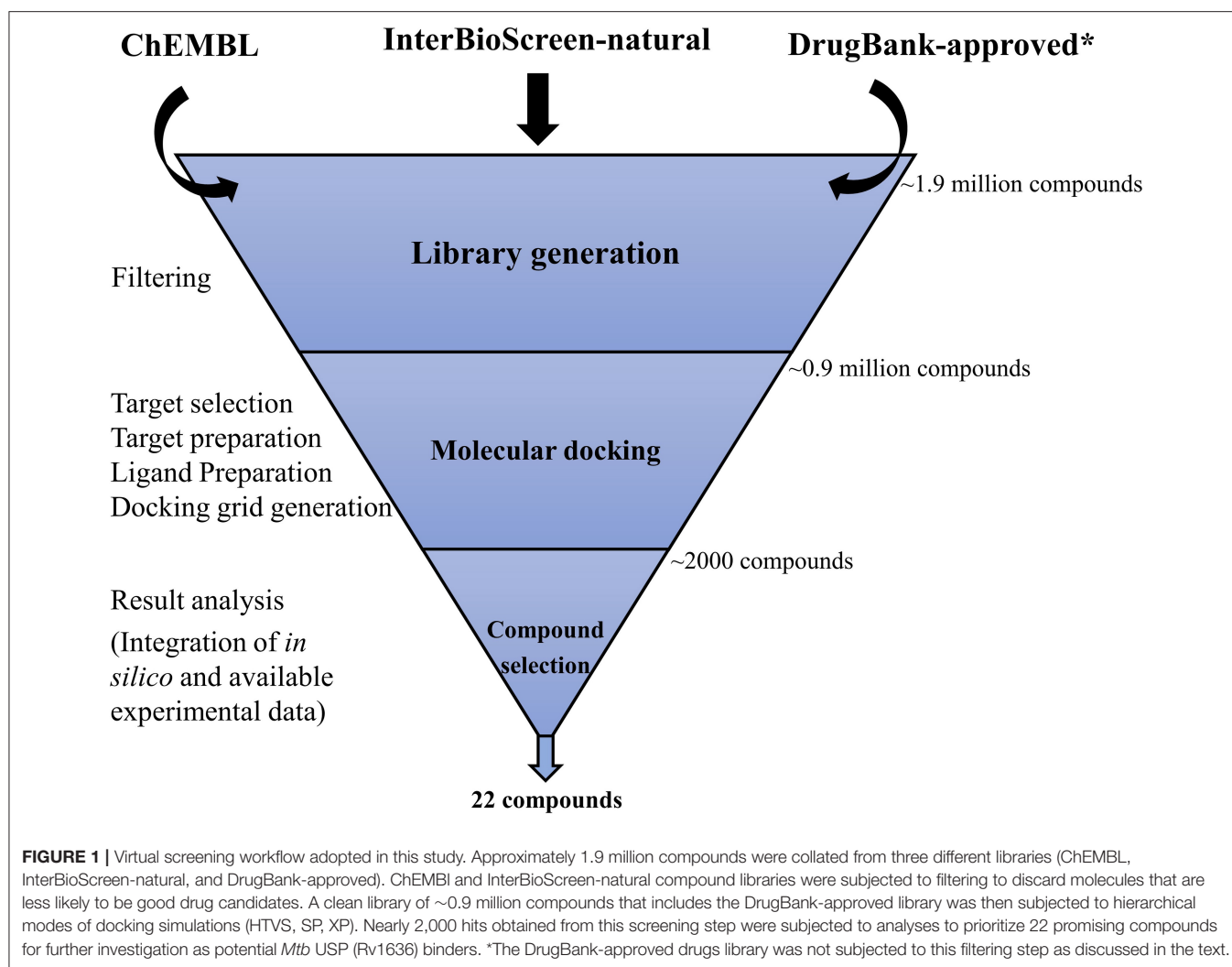
(chain C) of 5AHW has no missing residues. Therefore, the crystal structure of holo MSMEG_3811 with a preformed pocket that hosts cAMP (with no missing residues in the binding site of chain C) is more suited than the apo Rv1636 structure to screen for potential binders that can target the cAMP-binding pocket of *Mtb* USP. Therefore, here, we have used chain C of 5AHW for the docking study. Importantly, our earlier studies indicated that cAMP exhibits comparable binding affinities with Rv1636 and MSMEG_3811 (Banerjee et al., 2015). Thus, a compound that binds to the cAMP binding site of MSMEG_3811 is likely to bind to Rv1636.

Target Structure Preparation

The reliability of the predictions from docking studies is largely dependent on the accuracy of the atomic coordinates of the input structures. For a structure determined by X-ray crystallography, a good fit of the atomic model to observed electron density ensures the reliability of the position of the atoms. Earlier studies revealed instances of overenthusiastic interpretation of ligand density (Deller and Rupp, 2015). Therefore, the quality of the ligand and binding site residues of the input structure (PDB code: 5AHW) was inspected using the EDIA (electron density score for individual atoms) tool (Meyder et al., 2017). The structure was also visually inspected against its electron density map. Supplementary Table 1 shows that the quality of the binding site residues and bound cAMP in the chain C of 5AHW is satisfactory.

The binding site of cAMP in MSMEG_3811 is away from the interface of the protomers. Thus, only one chain of the homo-multimeric protein was chosen for docking experiments (Supplementary Figure 2). In the chain C of 5AHW, Val113, a residue lining the cAMP binding pocket has been modeled with dual conformations; each conformer has an occupancy of 0.5, indicating that both these conformers have equal influences on modulating ligand interactions and, thus, can differentially influence the outcomes of virtual screening (Supplementary Figure 2). Hence, during target preparation, both the conformers of Val113 were considered by fixing the coordinates of each conformer of the residue one at a time in two separate protein models, hereafter referred to as conformer I and conformer II. To minimize the chances of missing potential hits favored by only one of the two conformers, we docked the ligand library against both the available conformers (I and II). Ligands that fit well to the binding sites of both the conformers could also be identified and prioritized for testing, as these ligands are likely to have higher chances of binding to the protein at the specified site.

Protein preparation wizard available in the Schrödinger software package was used for the target preparation (Sastry et al., 2013). Hydrogen atoms were added to the structure. Water molecules and other unwanted crystallization aids were deleted from the binding site. The protonation states of the bound ligand, cAMP, were generated using Epik (Shelley et al., 2007; Greenwood et al., 2010) at pH 7.0 ± 0.5, and the protein was prepared at pH 7.4. Hydrogen bonding network in the structure was subjected to optimization followed by a restrained minimization so that heavy atoms converge to RMSD 0.3 Å,



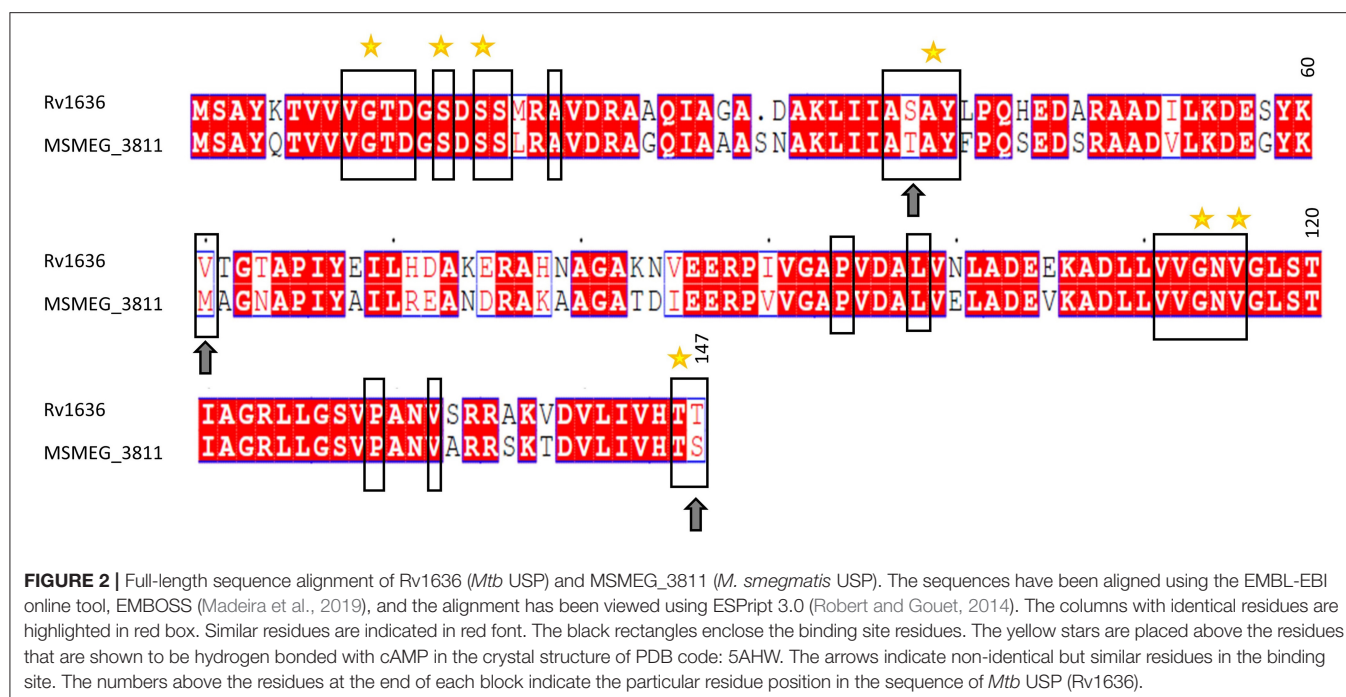
and the hydrogen atoms were fully optimized. This was done to ensure that strains in the structure are relieved alongside full relaxation of the hydrogen bonding network.

Ligand Library Generation and Preparation

A library of 1.9 million compounds was generated by collating compounds from three databases: ChEMBL, version 24.1 (Mendez et al., 2018); InterBioScreen-natural compounds (downloaded in September 2018; <https://www.ibscreen.com/natural-compounds>); and DrugBank, version 5.1.1 (approved molecules) (Wishart et al., 2006). The compounds from the ChEMBL and InterBioScreen library were subjected to cleaning by using the (i) structural and (ii) molecular property filters offered by Canvas (Duan et al., 2010; Sastry et al., 2010). The structural filters aided in removing compounds following the Rapid Elimination of Swill (REOS) (Walters et al., 1998) and Pan-Assay Interference Compounds (PAINS) (Baell and Holloway, 2010) concepts to enrich the library with molecules that are less likely to be toxic and promiscuous. The molecular property filters eliminated compounds (with molecular weight

>500 Da, hydrogen bond acceptor and donor count more than 10 and 5, respectively, and AlogP > 5) that are less likely to be successful oral drugs (Lipinski et al., 1997; Lipinski, 2000). From the DrugBank database, the subset of approved small-molecule drugs was included in our library. These molecules were not subjected to pre-filtering, as the known information on the safety and usages of these drugs could be exploited in prioritizing compounds for testing as discussed later. Finally, a clean library comprising of 0.9 million compounds was prepared by desalting and generating tautomers and stereoisomers at pH 7 ± 1 using the LigPrep module of Schrödinger package.

Additionally, we prepared a library of 14 compounds that demonstrated or were predicted to bind to Rv1636 through experimental or computational approaches, respectively. Two out of the 14 compounds include cAMP and ATP, where cAMP is known to bind to Rv1636 with a 10-fold higher affinity than ATP (Banerjee et al., 2015). cAMP and ATP served as the control compounds for our docking studies. The remaining 12 compounds include 10 polyphenolic compounds and 2 approved drugs (amikacin and kanamycin). We refer to the library of these



12 compounds as the secondary library. The 10 polyphenolic compounds were suggested to be potential binders of *Mtb* USP (Rv1636) by Aanandhi et al. (2014) based on their docking studies, where the compounds were docked at a site different from the cAMP binding site. We were interested in checking the possibility of binding of these compounds at the cAMP binding site. In a study by Sharma et al. (2016), it has been observed that Rv1636 is overexpressed in amikacin- and kanamycin-resistant *Mtb* isolates. They further performed docking studies to show that both the mentioned drugs have the potential to bind to the conserved USP domain of Rv1636. Docking of the control and secondary library of compounds was performed to ensure the validity of our protocol in reproducing the pose of the bound native ligand, cAMP, understand whether the results from our docking studies correlate with previously reported experimental binding affinities of cAMP and ATP toward MSMEG_3811, and compare the predicted binding affinities of the compounds in our primary library with those of the control and secondary library compounds.

Molecular Docking

Molecular docking of all the prepared chemical compounds was performed using Glide implemented in the virtual screening workflow (VSW) of Schrödinger software package (Friesner et al., 2004, 2006; Halgren et al., 2004). The grid box for docking the compounds was generated for both conformers I and II. Default settings in the Glide Receptor Grid Generation module were used for generating the two grid boxes enclosing the cAMP binding site in conformers I and II, which involve specifying the centroid of the bound cAMP as center of the grid box and choosing a box size that accommodates ligands similar to the size of the bound ligand. The van der Waals radii scaling factor for non-polar

atoms of the protein was kept at 1.0 with a partial charge cut-off of 0.25. The percentage of output compounds from each stage of the hierarchical VSW protocol was specified in such a way so that not more than 1,000 top-scoring compounds were reported in the hit list after the final stage of screening the ChEMBL library. Similarly, the initial number of virtual hits obtained from the screening of the InterBioScreen-natural compound and DrugBank-approved drugs libraries were restricted to 50 and 20 top scoring compounds, respectively.

The hierarchical docking modes in VSW include the following stages: (i) high throughput virtual screening (HTVS), (ii) standard precision (SP), and (iii) extra precision (XP). The first stage performs HTVS, which is the fastest of the three stages and trades sampling exhaustiveness for higher speed. The ligands that are retained are passed on to the second stage, which performs SP docking. The Glide SP docking performs more exhaustive sampling than HTVS stage. Both HTVS and SP docking use the same scoring function (SP GlideScore) to rank order the ligand poses. This score is a “softer” function that aims to minimize false negatives during the virtual screening of a large database of compounds. The ligands that survive after the SP docking stage are then passed on to the third stage, XP docking, for a more rigorous sampling. The XP docking uses a “harder” scoring function that penalizes poses that violate expected physical chemistry principles, such as large desolvation of polar and charged groups. The third stage in the VSW reduces the false positives that SP docking lets through. Thus, the three stages of screening lead to rational funneling of a large library of compounds to a small set of candidate ligands ranked on their predicted ability to bind to the specified conformation of the protein of interest at a given site (Friesner et al., 2004, 2006; Halgren et al., 2004).

Compound Selection

From an initial hit list of ~2,000 compounds (~1,000 for each conformer), 22 compounds were selected for experimental testing (18 compounds from the ChEMBL library, 2 each from the InterBioScreen-natural and DrugBank-approved drug libraries). MMGBSA (implemented in Prime v3 of Schrödinger software package) calculations were performed on all the initial hits (~2,000 compounds; ~1,000 initial hits from each of the two conformers) to estimate the relative binding affinities of these ligands in the implicit solvent model against the respective conformer (I and II) of the protein. The VSGB solvent model was used, which employs the variable-dielectric generalized Born model, incorporating a residue-dependent effect, where the solvent is water. While the Glide dock scores are based on empirical scoring functions that distinguish binders from non-binders, Prime-MMGBSA is a physics-based method that computes relative binding free energies (dG_{bind}) of bound and unbound molecules as per Equation 1, where E_{complex} is the minimized energy of the protein–ligand complex, and the E_{ligand} and E_{receptor} are the individual minimized energies in uncomplexed form. The absolute values calculated are not necessarily in agreement with experimental binding affinities. However, it has been shown earlier that ranking of ligands based on MMGBSA dG_{bind} scores agree reasonably with experimental binding energies and outperform empirical docking scores, especially in case of congeneric series of ligands (Lyne et al., 2006). These scores could serve as one of the guiding parameters for prioritizing analogous compounds while testing in an experimental setting. The docked poses of the ligands obtained from Glide-XP docking (final stage of VSW) served as the starting ligand structures for the Prime-MMGBSA calculations. The prepared protein structure for each of the two conformers used for docking the ligand library was taken as the input protein structure for the Prime-MMGBSA calculations. While the ligands' docked poses were subjected to relaxation, the protein atoms were kept rigid during the Prime-MMGBSA calculations.

$$dG_{\text{bind}} = E_{\text{complex}} - (E_{\text{ligand}} + E_{\text{receptor}}) \quad (1)$$

From the initial hit list of ~2,000 compounds obtained from the ChEMBL library, 100 top-scoring compounds based on docking scores were prioritized for further analysis. The poses and interaction profiles of each compound were visually scrutinized to ensure that the docked compounds fit well into the desired binding pocket and most of the important binding site residues (such as Ala40, Gly10, Ser14, Ser16, Gly114, Val116, Thr146) are engaged in hydrogen-bond interactions with the docked compounds. The mentioned residues are hydrogen bonded with bound cAMP in the crystal structure (PDB code: 5AHW; **Supplementary Figure 3**). While the residues Ala40, Gly10, Gly114, and Val116 establish hydrogen bonds with cAMP through backbone carbonyl oxygen or amide nitrogen or both, the remaining residues are engaged in side-chain-mediated hydrogen bonding with cAMP. It has been shown earlier that mutation of Gly10 and Gly114 (which corresponds to Gly113 in Rv1636) to Thr and Ala, respectively, significantly compromise the binding of cAMP and ATP to MSMEG_3811

(Banerjee et al., 2015). Therefore, engagement of these residues in interactions with other compounds might inhibit cAMP binding to the protein.

In addition, available information on the bioactivity of the shortlisted compounds was fetched from ChEMBL and or PubChem (Kim et al., 2018). The compounds that are already known to be effective against tuberculosis infection were assigned a higher priority for testing and designated as “biased set” compounds. The remaining compounds (“Blind set”) were chosen based on chemical diversity (as indicated by “Tanimoto coefficient”) to ensure that representative compounds from each cluster of chemical compounds are tested in an experimental setup. The Canvas module available with Schrödinger software package was used to calculate 2D Tanimoto coefficient (Syuib et al., 2013) and subsequently for the chemical diversity analysis.

From the virtual screening of InterBioScreen-natural compound library, 50 top scoring hits were subjected to MMGBSA calculations, and interaction profile analysis was performed in a similar way as mentioned for the ChEMBL library. Two compounds were selected for testing from this library. Two out of the 20 approved drugs as obtained from screening the DrugBank library were selected based on the consensus docking results against conformers I and II followed by interaction profile analysis coupled with analysis of the data pertaining to known primary targets of the compounds as available in DrugBank. Besides, curcumin from the secondary library was selected for testing.

Since docking scores predicted approximate binding affinities between the protein and ligands and Prime-MMGBSA dG_{bind} scores are approximate relative binding energies between the bound and unbound state of the molecules, more negative scores indicate the possibility of stronger binding.

OPLS3e force field (Roos et al., 2019) was used throughout the entire computational study.

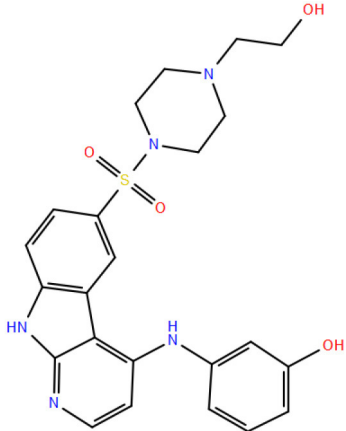
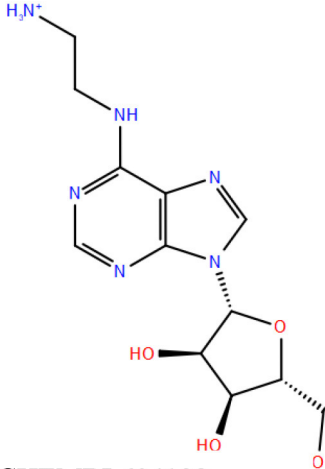
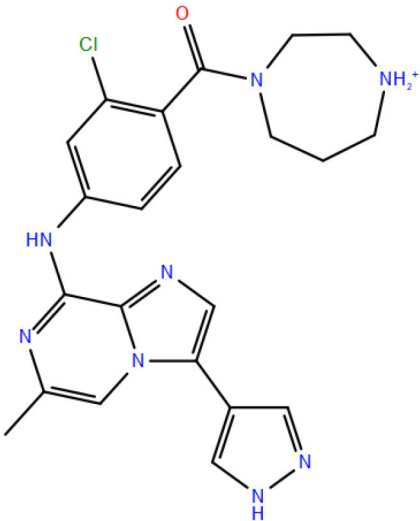
The non-covalent interactions between the protein and the docked compounds were visualized in Maestro GUI available with Schrödinger suite of programs. The geometric criteria used for the detection of these interactions are presented in **Supplementary Table 2**. The sketches of the chemical compounds provided in **Table 1** are made using the 2D sketcher implemented in Maestro GUI (Schrödinger, LLC, New York). The figures of protein–ligand complexes were generated using Maestro GUI and PyMOL (Schrödinger, LLC).

In vitro Analysis

Purification of Rv1636 and Microscale Thermophoresis

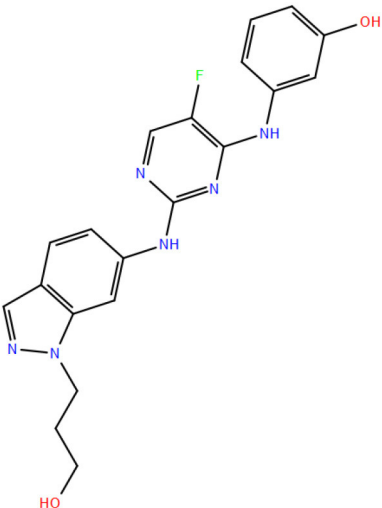
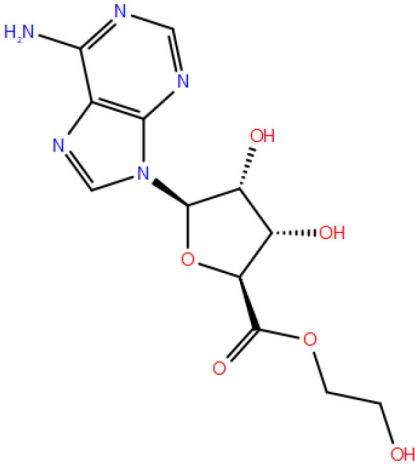
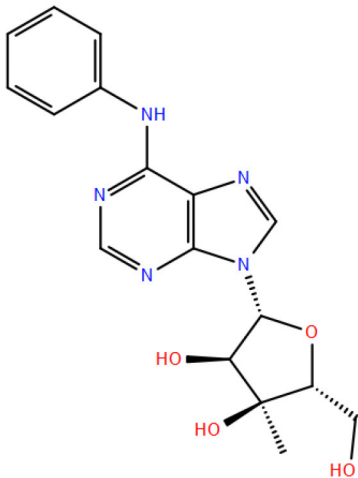
His-tagged Rv1636 was purified from *E. coli* SP850 *cyc*[−] strain in a buffer containing 50 mM Tris–Cl (pH 7.5), 100 mM NaCl, 5 mM 2-ME, and 10% glycerol as described earlier (Banerjee et al., 2015). Microscale thermophoresis (MST) was performed on a Nanotemper Technologies Monolith[®] NT.115 instrument (Munich, Germany). The protein was labeled with NT-495-NHS fluorescent dye in a buffer containing 10 mM HEPES (pH 7.5), 100 mM NaCl, 10% glycerol, and 0.05% Tween20. Labeled His-Rv1636 (100 nM) was added to varying concentrations of the ligand in buffer containing 50 mM Tris–Cl (pH 7.5), 100 mM

TABLE 1 | Results of molecular docking and Prime-MMGBSA calculation of 22 shortlisted candidates.

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG _{bind} score (kcal/mol)	Interacting residues*
1	 CHEMBL3133832	-11.8	-77.1	V9, G10 , T11, D12, S17, A20, A38, T39, A40 , Y41, F42, K60, M61, A62, P95, L99, V113, G114 , N115, V116 , G117, L118, G123, G127, S128, V129 , P130, T146
2	 CHEMBL604198	-11.8	-61.0	V9, G10 , T11, D12, S14, S16, S17, A20, A38, T39, A40 , Y41, F42, E57, M61, A62, A94, P95, L99, V112, V113, G114 , N115, V116 , G117, L118, S128, V129, P130, V133, T146
3	 CHEMBL1836814	-11.6	-68.9	V9, G10, T11, D12, S17, A20, A38, T39, A40 , Y41, F42, E57, K60, M61, A62, I67, V91, G93, A94, P95, A98, L99, V113, G114 , N115, V116, V129, P130

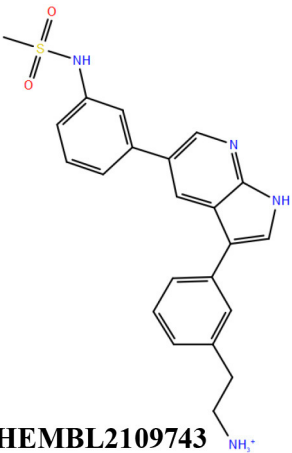
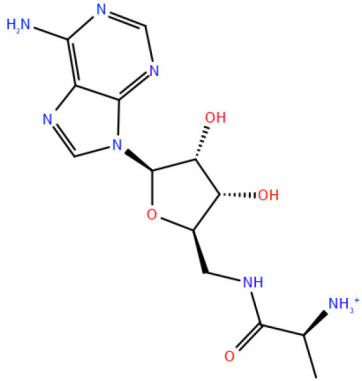
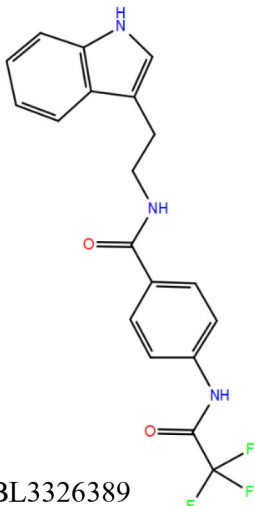
(Continued)

TABLE 1 | Continued

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG_{bind} score (kcal/mol)	Interacting residues*
4	 <p>CHEMBL3918786</p>	-11.1	-70.5	V9, G10 , T11, D12 , S14, S16, S17, A20, V21, A38, T39, A40 , Y41, F42, E57, K60, M61, A62 , G63, P95, L99, V113, G114, N115, V116, V129, P130, T146
5	 <p>CHEMBL2006195</p>	-11.1	-70.3	V9, G10 , T11, D12, S14, S16, S17, A20, A38, T39, A40 , Y41, M61, P95, L99, V112, V113, G114 , N115, V116 , G117, L118, S128, V129 , P130, T146
6	 <p>CHEMBL470932</p>	-11.0	-67.4	V9, G10 , T11, D12, S14, S17, A20, V21, A38, T39, A40 , Y41, F42, E57, K60, M61, A62, A94, P95, L99, V112, V113, G114 , N115, V116 , G117, L118, S128, V129, P130, T146

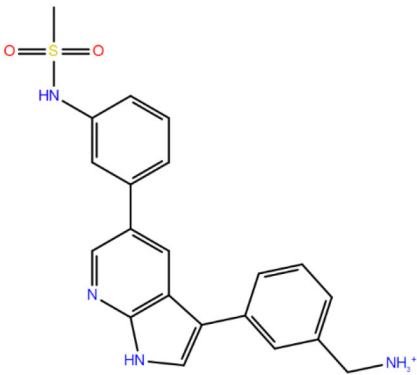
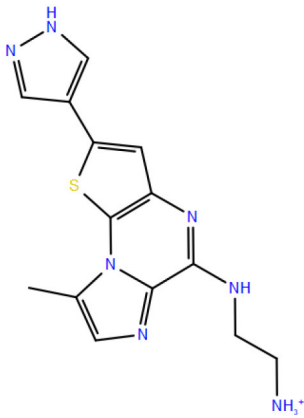
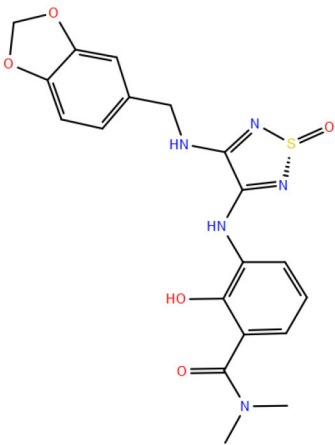
(Continued)

TABLE 1 | Continued

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG_{bind} score (kcal/mol)	Interacting residues*
7	 <p>CHEMBL2109743</p>	-10.9	-61.6	G10, T11, D12, S16, S17, A38, T39, A40, Y41, F42, E57 , G58, K60, M61, A62, I67, A94, P95, L99, V113, G114, N115, V116 , G117, S128, V129, P130, T146
8	 <p>CHEMBL1712355</p>	-10.8	-59.4	V9, G10 , T11, D12 , S14, S16 , S17, A20, A38, T39, A40 , Y41, M61, A62, P95, L99, V112, V113, G114 , N115, V116, V129, P130, V133, T146
9	 <p>CHEMBL3326389</p>	-10.7	-47.6	G10, T11, D12, G13, S14, S16, S17, A20, A38, T39, A40 , Y41, F42, E57, K60, M61, A62, I67, G93, A94, P95, L99, V113, G114, N115, V116 , G117, L118, S128, V129, P130

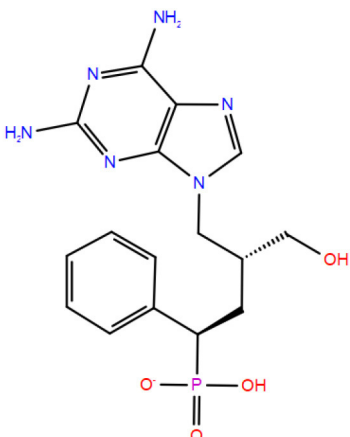
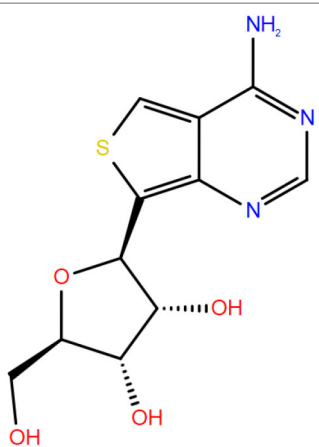
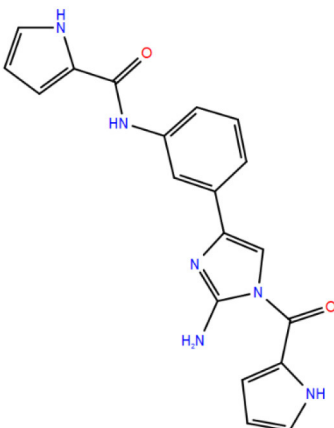
(Continued)

TABLE 1 | Continued

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG _{bind} score (kcal/mol)	Interacting residues*
10	 <p>CHEMBL536150</p>	-10.7	-61.6	G10, T11, D12, S14, S17, A38, T39, A40, Y41, F42, E57 , K60, M61, A62, A67, G93, A94, P95, L99, V113, G114, N115, V116 , G117, L128, V129, P130, T146
11	 <p>CHEMBL250106</p>	-10.7	-52.5	G10, T11, D12, S17, A38, T39, A40 , Y41, F42, E57 , K60, M61, A67, G93, A94, P95, L99, V113, G114, V116 , G117, L128, V129, P130
12	 <p>CHEMBL253104</p>	-10.5	-69.6	G10, T11, D12 , S14, S16, S17, A20, A38, T39, A40 , M61, A62, V91, P95, A98, L99, V113, G114 , N115, V116 , G117, L118, G123, L126, G127, S128, V129, P130, N132, T146

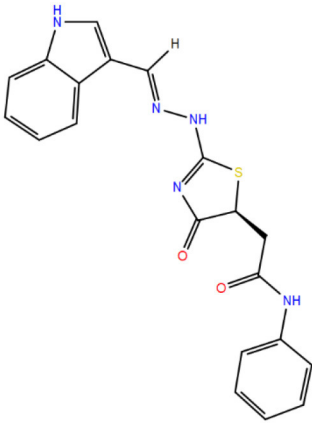
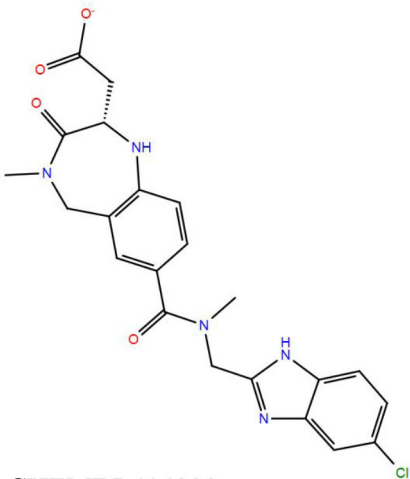
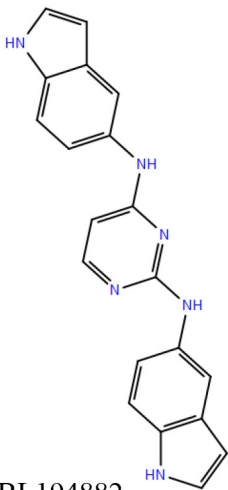
(Continued)

TABLE 1 | Continued

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG_{bind} score (kcal/mol)	Interacting residues*
13	 <p>CHEMBL2028231</p>	−10.5	−50.8	V9, G10 , T11, D12, S14, S16 , S17, A20, V21, A38, T39, A40 , Y41, M61, A62, V91, P95, L99, V112, V113, G114 , N115, V116, G117, S128, V129, P130, V133, T146, S147
14	 <p>CHEMBL1727847</p>	−10.4	−63.7	V9, G10 , T11, D12, S14, S16, S17, A20, V21, A38, T39, A40 , Y41, M61, P95, L99, V112, V113, G114 , N115, V116 , G117, L118, S128, V129, P130, T146
15	 <p>CHEMBL3113262</p>	−10.3	−62.0	V9, G10, T11, D12, S17, A20, A38, T39, A40 , Y41, F42, M61, P95, L99, V113, G114, N115, V116 , G117, L118, G123, G127, S128, V129, P130, T146

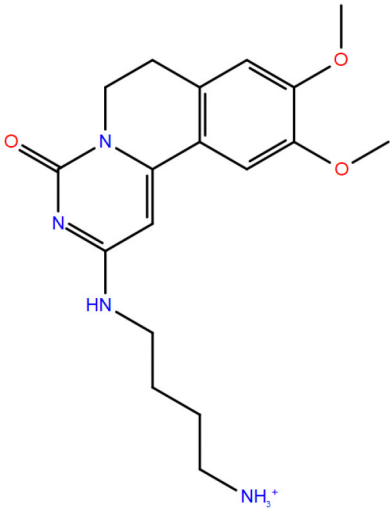
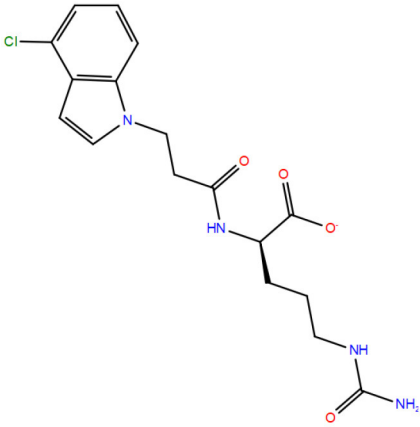
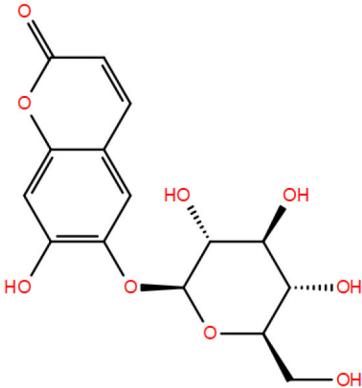
(Continued)

TABLE 1 | Continued

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG_{bind} score (kcal/mol)	Interacting residues*
16	 <p>CHEMBL3195891</p>	-10.2	-54.3	G10, T11, D12, G13, S14, S17, A38, T39, A40 , Y41, F42, E57, K60, M61, A62, I67, V91, P95, A98, L99, V113, G114, N115, V116 , G117, L118, S128, V129, P130
17	 <p>CHEMBL414232</p>	-10.1	-45.0	G10, T11, D12, G13, S14, S16 , S17, A20, A38, T39, A40 , Y41, F42, E57, K60, M61, A62, A67, P95, L99, V112, V113, G114, N115, V116 , G117, S128, V129, P130, V133, T146 , S147
18	 <p>CHEMBL194882</p>	-10.1	-45.0	G10, T11, D12, S17, A38, T39, A40 , Y41, F42, E57, K60, M61, A62, V91, P95, L99, V113, G114, N115, V116 , G117, S128, V129, P130

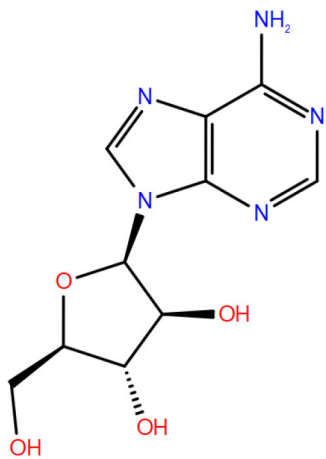
(Continued)

TABLE 1 | Continued

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG_{bind} score (kcal/mol)	Interacting residues*
19	 <p>STOCK1N42384</p>	-11.1	-58.2	V9, G10, T11, D12, S14, S16, S17, A38, T39, A40 , Y41, F42, E57 , G58, K60, M61, A62, I67, A94, P95, L99, V113, G114, N115, V116, G117, L118, V129, P130, V133, T146
20	 <p>STOCK1N74667</p>	-8.4	-48.1	V9, G10, T11, D12, S14, S16 , S17, A20, A38, T39, A40, S41, S61, V91, P95, L99, V112, V113, G114 , N115, V116 , G117, L118, G123, G127, S128, V129, P130, V133, T146 , S147
21	 <p>Esculin (DB13155)</p>	-11.1	-60.4	V9, G10 , T11, D12 , S14, S16, S17, A20, A38, T39, A40 , S61, A62, P95, L99, V112, V113, G114 , N115, V116 , G117, L118, S128, V129 , P130, T146

(Continued)

TABLE 1 | Continued

Sl. No.	Compound	Docking score (kcal/mol)	Prime-MMGBSA dG_{bind} score (kcal/mol)	Interacting residues*
22	 Vidarabine (DB00194)	−8.8	−52.6	V9, G10 , T11, D12, S14, S16, S17, A20, A38, T39, A40 , S41, M61, P95, L99, V113, G114, N115, V116 , G117, L118, S128, V129, P130, V133, T146

*This column holds the information on all binding site residues (within 5 Å) based on the docked pose of the ligand. The residue names in bold are involved in hydrogen bonding. Other residues provide favorable contacts to the ligand. Further details on other types of interaction could be found in **Supplementary Table 3**.

The alphanumeric code indicated below each compound's structure correspond to the original identification number of the compound in the respective databases. The Compound identification numbers represented in bold are the biased set molecules.

NaCl, 10% glycerol, and 0.05% Tween20. Samples were incubated at room temperature for 10 min, loaded into capillaries, and placed in the MST block. Thermophoresis was measured at an ambient room temperature of 25°C and performed using 60% excitation power for the nanoblu filter and medium MST IR-laser power. Fluorescent migration used to determine K_d was measured from 1.5 to 2.5 s and then normalized to initial fluorescence (−1.0 to 0 s). The data from three independent replicates were analyzed using MO Affinity Analysis software v2.3 and fit to the standard K_d fit model, which describes a molecular interaction with a 1:1 stoichiometry according to the law of mass action.

RESULTS

Control Library

This library consists of the two known binders of Rv1636 and MSMEG_3811, viz., cAMP and ATP. The redocking experiment yielded a reproducible binding pose for cAMP as observed in 5AHW. The two poses (experimental and predicted) perfectly superimpose on each other (**Supplementary Figure 3**), thus validating our docking protocol. The docked pose of cAMP has a docking score of −10.5 kcal/mol. The docking score of ATP against MSMEG_3811 is −1.7 kcal/mol. This result corroborates the earlier experimental observations (Banerjee et al., 2015). In the current study, too, the binding affinity of cAMP to Rv1636 has been verified through MST assay (**Table 2** and **Supplementary Figure 4**).

Primary Library

Compounds from three different sources (ChEMBL, InterBioScreen-natural, and DrugBank-approved) have been

collected in this library, as mentioned earlier. Docking results of selected compounds from these libraries are presented below.

ChEMBL Library

The initial set of ~2,000 compounds obtained from screening the ChEMBL library yielded compounds whose docking score range from −11.8 to −8.2 kcal/mol. The top 100 compounds (range of docking scores, −11.8 to −10.1 kcal/mol) were subjected to in-depth analysis (**Supplementary Table 3**). The Prime-MMGBSA dG_{bind} scores of these 100 compounds range from −82.2 to −29.8 kcal/mol. As mentioned earlier, the absolute values calculated here might not agree with experimental binding energies (for details on relevance of Prime-MMGBSA dG_{bind} scores, refer to Materials and Methods). The docking and Prime-MMGBSA dG_{bind} scores calculated in our study indicate favorable binding of the top 100 compounds. Some of these compounds are theoretically better binders than cAMP (as indicated by the scores). Analyzing the interaction profiles of top 100 compounds revealed that most of these docked compounds are engaged in hydrogen bonding with multiple critical binding site residues (like Gly10, Ala40, Ser16, Gly114, etc.). Some of the compounds are also involved in other types of electrostatic interactions (such as salt bridges, aromatic CH- π interactions, π - π stacking, and halogen bonds) with Thr11, Asp12, Phe42, Asp57, etc. Information on the known anti-tubercular property could be obtained from database search for 3 out of the 100 top compounds. These three compounds comprise the “biased” subset of molecules that were shortlisted for experimental investigations. From the remaining 97 compounds, 18 chemically diverse compounds were prioritized for testing that formed the “blind” subset (**Table 1**). The analyses of the docked poses

TABLE 2 | Binding affinity (K_d) of experimentally tested compounds as determined by MST assays.

Sl. No.	Name of the compound	Library	Docking score (kcal/mol)	Prime-MMGBSA dG_{bind} score (kcal/mol)	K_d (μM)
1	cAMP	Control (positive)	−10.5	−60.5	2.68 ± 0.07
2	STOCK1N42384	Primary (InterBioScreen-natural compound)	−11.1	−58.2	998 ± 82
3	STOCK1N74667	Primary (InterBioScreen-natural compound)	−8.4	−48.1	1717 ± 731
4	Curcumin	Secondary (literature search)	−6.6	−58.8	17.37 ± 0.8

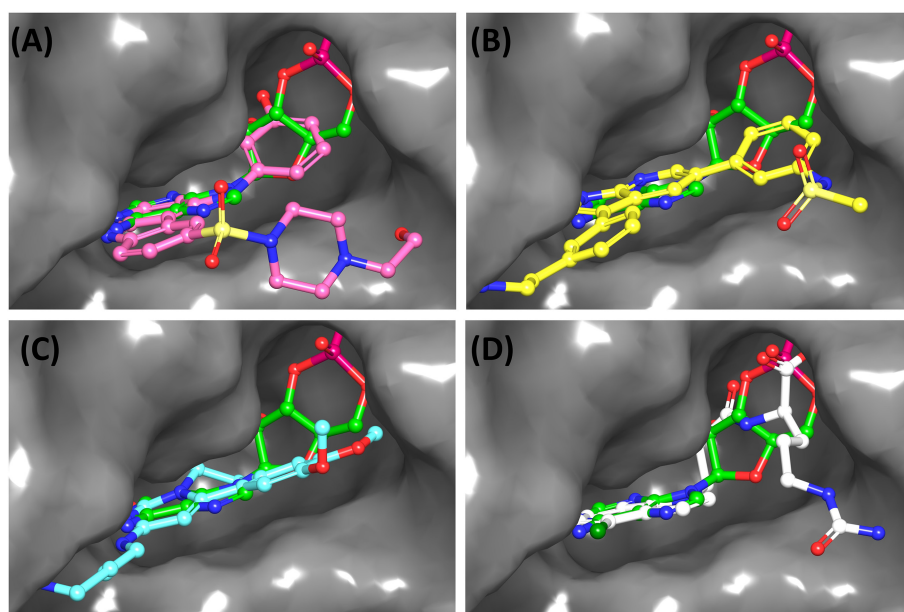


FIGURE 3 | Overlay of docked pose of selected compounds on to the bound pose of cAMP (green stick) as reported in 5AHW. The protein binding site is depicted as gray surface, and the ligands are shown in ball and stick representation: **(A)** ChEMBL3133832 (pink carbon), **(B)** ChEMBL2109743 (yellow carbon), **(C)** STOCK1N-42384 (cyan carbon), **(D)** STOCK1N-74667 (white carbon). Nitrogen, oxygen, chlorine, and sulfur atoms are shown in blue, red, dark green, and yellow, respectively. Hydrogen atoms were not displayed during image generation to maintain visual clarity.

of two selected *in silico* hits from the ChEMBL library are furnished below.

- (a) ChEMBL3133832 (IUPAC name: 3-[[6-[4-(2-hydroxyethyl)piperazin-1-yl]sulfonyl-9H-pyrido[2,3-b]indol-4-yl]amino]phenol): This compound is one of the best hits from the ChEMBL library with the best docking score (−11.8 kcal/mol) among the ~0.9 million compounds that were subjected to docking. It also has the best MMGBSA score (−77.1 kcal/mol) among all the compounds that have been selected for testing (Table 1). The compound is well-accommodated in the cAMP binding cavity of MSMEG_3811 and is predicted to be hydrogen-bonded with key residues, such as Gly10, Ala40, Gly114, Val116, and Val129 (Figure 3 and Supplementary Figure 5).
- (b) ChEMBL2109743 (IUPAC name: N-[3-[3-[3-(2-aminoethyl)phenyl]-1H-pyrrolo[2,3-b]pyridin-5-yl]phenyl]methanesulfonamide): This compound (also referred as GSK581005A) is from the biased subset. Phenotypic screening at GlaxoSmithKline (GSK) led to

the identification of ChEMBL2109743/GSK581005A to be effective against *Mtb* H37Rv [minimum inhibitory concentration (MIC) < 10 μM] with low human cell-line toxicity (Ballell et al., 2013). Owing to the known antitubercular property of ChEMBL2109743, we have selected it for testing. The docked pose of the compound in cAMP binding site of MSMEG_3811 indicates that it is well-accommodated in the binding pocket (docking score = −10.9 kcal/mol; MMGBSA dG_{bind} score = −61.6 kcal/mol) and is also predicted to be hydrogen bonded with critical binding site residues (Figure 3, Supplementary Figure 6, Table 1, and Supplementary Table 3).

InterBioScreen-Natural Library

The 50 initial hits reported from the hierarchical docking exercises have docking scores that range from −11.6 to −7.3 kcal/mol. The relative binding energies of all these 50 compounds were calculated using the Prime-MMGBSA approach. Visual inspection of the poses of all the compounds

revealed that only a few of the compounds from this library are engaged in hydrogen bonding with the important binding site residues. Only one compound, STOCK1N-42384 (IUPAC name: 2-((4-aminobutyl)amino)-9,10-dimethoxy-6,7-dihydro-4H-pyrimido[6,1-a]isoquinolin-4-one), has a docking score better than cAMP. Given the importance of natural compounds and their analogs in medicinal chemistry (Mushtaq et al., 2018; Atanasov et al., 2021), two compounds, STOCK1N-42384 and STOCK1N-74667 (IUPAC name: (R)-2-(3-(4-chloro-1H-indol-1-yl)propanamido)-5-ureidopentanoic acid), were prioritized for laboratory testing (Table 1). STOCK1N-42384 and STOCK1N-74667 demonstrated a K_d of ~ 1 and ~ 1.7 mM, respectively, against Rv1636 (Table 2 and Supplementary Figure 4). The analyses of the docked poses of the two selected compounds from the InterBioScreen-natural library are presented below.

- (a) STOCK1N-42384: This compound engages one of the key binding site residues, Ala40, in hydrogen bonding (Figure 3 and Supplementary Figure 7). Additionally, it is also hydrogen bonded with Glu57. Several binding site residues offer favorable contacts to this compound in its predicted pose, thus contributing to a docking score (-11.1 kcal/mol) comparable to some of the top-scoring virtual hits from the ChEMBL library (Table 1, Figure 3, and Supplementary Figure 7).
- (b) STOCK1N-74667: This compound has a docking score of -8.4 kcal/mol and is predicted to be engaged in hydrogen bonding with Gly114, one of the critical residues for cAMP binding as demonstrated in our previous mutagenesis studies and discussed earlier. It is also predicted to be hydrogen bonded to Ser16, Val116, and Thr146. Furthermore, several residues housed by the cAMP binding site offer favorable contacts to the docked pose of STOCK1N-74667 in MSMEG_3811 (Table 1, Figure 3, and Supplementary Figure 8).

DrugBank Library

A list of 20 initial reported hits were obtained upon screening the library of approved drugs against both the protein conformers. The docked poses of all the hits were analyzed, and two compounds were prioritized for experimental testing: esculin and vidarabine (Table 1). While esculin (<https://go.drugbank.com/drugs/DB13155>) is a glycosyl compound used as a vasoprotective agent, vidarabine is a known antiviral drug with established safety records (<https://www.drugbank.ca/drugs/DB00194>). The primary target of esculin is a human protein (androgen receptor). In addition, esculin is a coumarin derivative, and promiscuity of this chemical class of compounds is well-known (Stefanachi et al., 2018). Therefore, anticipating esculin could interfere with the functions of undesired human proteins, vidarabine was assigned a higher priority for further investigations. While docking studies predicted favorable interactions of vidarabine and esculin with critical binding site residues (Table 1), *in vitro* binding assays performed on both these compounds against Rv1636 did not show encouraging results (data not shown).

Secondary Library

The 10 polyphenolic natural compounds from this library (that includes curcumin) have docking scores between -7.3 and -4.8 kcal/mol, which indicate that the binding affinities of these polyphenolic compounds are likely to be weaker than the native ligand, cAMP (Supplementary Table 4). MST assays revealed a K_d of $17 \mu\text{M}$ for curcumin against Rv1636 (Table 2 and Supplementary Figures 4, 9). The two other approved drugs from the secondary library, amikacin and kanamycin, although predicted to be favorably accommodated (-9.5 and -8.9 kcal/mol, respectively) in Rv1636, have docking scores poorer than most of the compounds that we have selected for experimental validations. The majority of the compounds in the secondary library failed to establish hydrogen bonds with the critical binding site residues (Supplementary Table 4).

DISCUSSION

This study has shortlisted potential binders of *Mtb* USP (Rv1636) using a rigorous computational protocol primarily employing molecular docking simulations. Although docking scores are important parameters to find “needles” (potential binders) from the “haystack” (of non-binders in a large chemical universe), it is well-known that the scoring functions can have their own limitations (Huang et al., 2010). Therefore, to derive meaningful insights from computational studies, we have integrated other physics-based methods like the Prime-MMGBSA calculations and the available experimental knowledge in our workflow to ensure that the high confidence virtual hits are taken forward for experimental validations. A molecule with more negative docking and Prime-MMGBSA dG_{bind} scores and also predicted to be hydrogen bonded with the critical binding site residues (as shown previously through mutagenesis experiments) are expected to be better candidates. Furthermore, when such a candidate is also reported as hits against both conformers I and II, the confidence associated with that compound is higher as explained earlier (Supplementary Table 3). It is to be noted that the ChEMBL and InterBioScreen-Natural compound libraries of molecules have been filtered through REOS, PAINS, and Lipinski’s rule of five filters as a prescreening step. Therefore, all the shortlisted candidates from these libraries are drug-like molecules and, in general, can be assumed to be safe. The targeted pocket in Rv1636 is suited for binding nucleotide scaffolds. Therefore, any molecule with a nucleotide containing moiety or its analog targeting this pocket has a chance to cross-talk with host nucleotide-binding proteins, such as the protein kinases (Taylor et al., 2012). Nevertheless, ChEMBL2109743, while known to inhibit human serum and glucocorticoid-regulated kinase 1 (SGK-1) (James et al., 2009), has been observed to have low cellular toxicity in the GSK tuberculosis screening.

Corroboration Between Our *in silico* Studies and Experimental Findings

cAMP and ATP (Control Library)

As mentioned earlier, cAMP binds to the conserved USP domain of Rv1636 by engaging some key binding site residues in hydrogen bonding confirmed through structure-guided

mutagenesis studies. Furthermore, cAMP has been found to bind to this protein with almost 10 times higher affinity than ATP (Banerjee et al., 2015). Similar observations have also been noted in our computational studies, as discussed earlier (**Supplementary Figure 3** and **Supplementary Table 3**).

Biased Subset (ChEMBL Library)

ChEMBL2109743/GSK581005A is one of the interesting hits that we identified by screening the ChEMBL library of compounds. In a study by Mugumbate et al., it has been suggested through chemogenomic studies that *Mtb* dihydrofolate reductase (DHFR) could be a target of this compound (Mugumbate et al., 2015). Our virtual screening results suggest that GSK581005A is a potential binder of Rv1636 and is predicted to be hydrogen-bonded with critical binding site residues (**Figure 3**, **Supplementary Figure 6**, **Table 1**, and **Supplementary Table 3**). It is worthy to note that chemogenomic tools used in the mentioned study were trained on the ChEMBL database of known target–ligand pairs. Therefore, all predicted targets are biased toward well-studied proteins that are already included to the ChEMBL database (Mugumbate et al., 2015). Hence, it is unlikely that such methods would predict Rv1636 as a target of any query compound. Given that GSK581005A is already privileged with respect to its cell permeability, there is merit in future *in vitro* testing of this compound against Rv1636 (Manjunatha and Smith, 2015).

Two other compounds (ChEMBL3195891 and ChEMBL17272847) have been reported to be tested against an *Mtb* target. Both these compounds were reported to be active in a counter-screening for inhibitors of the fructose-bisphosphate aldolase (FBA) (<https://pubchem.ncbi.nlm.nih.gov/compound/135470622#section=Biological-Test-Results>). However, a detailed report that can aid in making the informed decisions could not be found.

STOCK1N-42384 and STOCK1N-74667 (InterBioScreen-Natural Library)

Both these natural compounds shortlisted from our virtual screening bound Rv1636 in the MST assays (**Table 2** and **Supplementary Figure 4**). Further investigations with analogs of these compounds that establish hydrogen bonds with the critical cAMP binding residues (such as Gly 10 and Gly114 for STOCK1N-42384; Gly10 and Ala40 for STOCK1N-74667) could improve the binding affinities of these natural compounds (**Table 1**, **Figure 3**, and **Supplementary Figures 5, 6**).

Curcumin (Secondary Library)

A previously published report (Aanandhi et al., 2014) indicated curcumin to be a potential binder of Rv1636 at a site different from the cAMP binding site. Analysis of the docking score and hydrogen-bonding profile of curcumin at cAMP binding site in our study hints that curcumin is unlikely to be a strong binder (**Supplementary Table 4** and **Supplementary Figure 9**). Contrary to our expectation, curcumin demonstrated a high binding affinity toward Rv1636 ($K_d \sim 17 \mu\text{M}$) (**Table 2** and **Supplementary Figure 4**). Therefore, curcumin may interact with Rv1636 through a site different from the cAMP binding site. Notably, curcumin and other

polyphenolic compounds are highly promiscuous and are flagged as pan-assay interference compounds (PAINS) that often act by non-drug-like mechanisms. Attempts to optimize PAINS as drug candidates in the past have proven futile, and we do not encourage medicinal chemists to consider curcumin as a good starting point for designing an inhibitor against Rv1636 or any other drug targets (Baell, 2010; Baell and Holloway, 2010; Baell and Walters, 2014).

A close inspection of the cAMP binding site in 5AHW reveals that the site can be grossly divided into five subpockets (P1, P2, P3, P4, and P5) (**Supplementary Figure 10**). The P1 comprises a mixture of hydrophobic and polar residues (Gly10, Thr11, Asp12, Ser14, Val113, Gly114, Asn115, and Val116). The P2 (Ala38, Thr39, Ala40, and Leu99), P3 (Met61 and Pro95), and P4 (Ala20, Val129, Pro130) are predominantly hydrophobic. Interestingly, the backbones of some of the hydrophobic residues in P1 and P2 are directed toward the ligand-binding cavity, thus facilitating the formation of hydrogen bonds with cAMP. The residues in P3 and P4 offer favorable hydrophobic contacts to the bound cAMP. The P5 is formed by three polar residues, *viz.*, Ser16, Ser17, and Thr146. Ser16 and Thr146 are involved in side-chain-mediated hydrogen-bonding interactions with bound cAMP in 5AHW (**Supplementary Figure 10**). Analysis of the docked poses of the top 100 compounds shows that most of the compounds in our library do not establish hydrogen bonds with the P5 residues (**Supplementary Table 3**). Introduction of functional groups at appropriate sites on the ligand that help form hydrogen bonds with P5 residues may contribute to improving the binding affinities of the compounds against Rv1636. In the future, such chemical modifications of the shortlisted compounds that show promising activity in experimental testing can be explored. Furthermore, molecular dynamics studies on promising *in vitro* hits would provide important insights into compound optimization.

The *in silico*-driven approach employed in this study has helped in shortlisting 22 virtual hits that can potentially bind to *Mtb* USP (Rv1636). This is the first report of screening a large library of publicly available compounds (~1.9 million) to identify potential binders of Rv1636. An overall analysis of the results on the shortlisted candidates suggests that these compounds are likely to be better candidate binders of cAMP binding site in Rv1636 than earlier reported polyphenolic compounds. *In vitro* testing of a few shortlisted compounds has demonstrated promising candidates from the natural compound library. Corroboration between our computational and experimental studies emphasizes the strength of a carefully designed protocol used in selecting an enriched set of potential compounds from vast chemical libraries. Relevant details of the shortlisted compounds that might help medicinal chemists, biochemists, and other researchers make an informed decision for selecting, testing, and designing experiment protocols have been provided (**Tables 1, 2** and **Supplementary Tables 3, 5**). To conclude, the findings reported in this study can serve as important starting points in the drug discovery pipeline of antitubercular leads targeted against Rv1636. It can be expected that successful inhibition of this protein combined with other established anti-tubercular therapeutic approaches could open

new avenues for effective disease management and tackling the emerging problem of drug resistance. Finally, the integrative *in silico* pipeline presented in this study is a generalized one and, in principle, can be used for any target-centric ligand screening ventures.

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/s.

AUTHOR CONTRIBUTIONS

SC designed and performed the computational studies. MC and AB performed experiments. SC prepared the manuscript with inputs from all the authors. SC, NS, and SV finalized the manuscript.

FUNDING

This research was supported by Mathematical Biology program and FIST programs sponsored by the Department of Science and Technology (DST), the DBT-IISc Partnership Program Phase-II (BT/PR27952/INF/22/212/2018/21.01.2019) and a DBT grant to SV (BT/PR15216/COE/34/02/2017). Support from UGC, India—Centre for Advanced Studies and Ministry of Human Resource Development, India is gratefully acknowledged. The authors acknowledge the support from DBT-Bioinformatics and Computational Biology Centre. SC acknowledges DST-INSPIRE for her research fellowship and MC was supported by a Senior Fellowship from the Council for Scientific and Industrial Research.

ACKNOWLEDGMENTS

The authors are thankful to Dr. Arka Banerjee and Dr. Sachchidanand for useful discussions. Prof. R. Sowdhamini was acknowledged for providing access to the Schrödinger suite of programs at NCBS, Bengaluru. The authors thank Dr. Pritesh Bhat and his team at Schrödinger, Bengaluru for the technical assistance received in facilitating local access to the software package at the IISc campus. NS and SV are J. C. Bose National Fellows and SV was a Margadarshi Fellow supported by the DBT Wellcome Trust India Alliance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.599221/full#supplementary-material>

Supplementary Figure 1 | Structural comparison of apo Rv1636 (PDB code: 1TQ8) and cAMP bound MSMEG_3811 (PDB code: 5AHW). **(A)** Superimposition of chain C of 5AHW (cyan cartoon) on to chain A of 1TQ8 (violet cartoon). The overall RMSD between the aligned residues in these two chains as calculated using DALI (Holm, 2020) is 3.1Å. The residues within 5Å of bound cAMP (depicted in dot representation with green carbon atoms) are shown in magenta in both

5AHW and 1TQ8. The parts of the protein chains that superimpose well are shown as translucent cartoons whereas the parts showing deviations are shown as opaque cartoons. The regions that show structural deviations of the secondary structures are encircled with black dashes. The region encircled with orange dashes could be seen only in 5AHW. The protein residues in the equivalent region are missing the electron density map of 1TQ8. **(B)** For better visualization, only the parts that show deviations are shown; the rest of the parts in the two proteins that superimposes well were not displayed during image generation. **(C)** Superimposition of all the binding site residues surrounding cAMP (represented as sphere with green carbon atoms). **(D)** The pair of residues which show structural deviations are shown (Y41:Y53; P95:P107; L99:L111; N115:N127; V129:V141; P130:P142; V133:V145; T146:T158; S147:T159). The co-ordinates of the residue equivalent to M61 (of 5AHW) are missing in the electron density map of 1TQ8. In **(C,D)** protein residues are shown as thin sticks; 1TQ8: violet, 5AHW: magenta. The residue identifiers are also color-coded as per the color of the corresponding residues.

Supplementary Figure 2 | Analysis of crystal structure of cAMP bound MSMEG_3811 (PDB code: 5AHW). **(A)** The six chains of MSMEG_3811 are shown in surface representation (chain A: green, chain B: blue, chain C: cyan, chain D: yellow, chain E: wheatish, chain F: white). The binding site of cAMP (represented as dots; green carbon atoms) in chain C is shown in magenta (also indicated by a black arrow). As can be seen in the figure, the magenta region only spans within the chain C and is away from the interface of the protomers. **(B)** cAMP in the binding site (shown as magenta translucent surface) of chain C of 5AHW. The residue V113 in the binding site is shown as gray stick. The atoms CA, CB, CG1, and CG2 of V113 have dual occupancies and thus two different orientations of the side chain could be seen. **(C)** The side-chain orientation of conformer I of V113 with respect to bound cAMP in chain C. **(D)** The side-chain orientation of conformer II of V113 with respect to bound cAMP in chain C. cAMP is shown in stick representation in panel B, C, and D with green carbon atoms. Nitrogen, oxygen, and phosphorous atoms are shown in blue, red and orange, respectively.

Supplementary Figure 3 | cAMP in MSMEG_3811 binding pocket (PDB code: 5AHW). **(A)** Superimposition of experimentally determined bound pose of cAMP (green) on to the re-docked pose of cAMP (yellow) in the binding pocket (gray surface) of the protein. **(B)** cAMP (green ball and stick model with gray transparent surface) bound to the protein binding site. The residues (within 5Å of the ligand) in the binding pocket are shown as thin sticks and color-coded based on their physicochemical properties (cyan: polar; green: hydrophobic; red: charged and negative; white: glycine). **(C)** 2D-interaction diagram of cAMP with residues in the binding pocket. Hydrogen bonds are shown in pink arrows. The residue identifiers are depicted as leaves, where the base of the leaves indicate the residue backbone, and the tip of the leaves indicate the direction in which the side-chains of the residues are pointed. Nitrogen, oxygen, sulfur, and phosphorus atoms are shown in blue, red, yellow and orange, respectively. Hydrogen atoms were not displayed during image generation to maintain visual clarity.

Supplementary Figure 4 | Micro-scale thermophoresis to study interaction of fluorescent tagged-His-Rv1636 (100 nM) with varying concentration of: **(A)** cAMP; **(B)** STOCK1N-42834 and STOCK1N-74667, **(C)** Curcumin.

Supplementary Figure 5 | 2D interaction map of docked pose of ChEMBL3133832 with the binding site residues of MSMEG_3811. For details related to color code, please refer to the legend to **Supplementary Figure 3**.

Supplementary Figure 6 | 2D interaction map of docked pose of ChEMBL2109743 with the binding site residues of MSMEG_3811. For details related to color code, please refer to the legend to **Supplementary Figure 3**.

Supplementary Figure 7 | 2D interaction map of docked pose of STOCK1N-42384 with the binding site residues of MSMEG_3811. For details related to color code, please refer to the legend to **Supplementary Figure 3**.

Supplementary Figure 8 | 2D interaction map of docked pose of STOCK1N-74667 with the binding site residues of MSMEG_3811. For details related to color code, please refer to the legend to **Supplementary Figure 3**.

Supplementary Figure 9 | Analysis of docked pose of Curcumin with the binding site residues of MSMEG_3811. **(A)** Overlay of docked pose of curcumin (violet stick) onto bound pose of cAMP (green stick). Nitrogen, and oxygen atoms are shown in blue and red, respectively. Hydrogen atoms were not displayed during

image generation to maintain visual clarity. **(B)** 2D interaction map of docked pose of curcumin with the binding site residues of MSMEG_3811. For details related to color code, please refer to the legend to **Supplementary Figure 3**.

Supplementary Figure 10 | Sub-pockets in cAMP binding site of MSMEG_3811 (PDB code: 5AHW). The five sub-pockets are marked P1, P2, P3, P4, and P5. The group of residues in each sub-pocket is encircled. For details related to color code, please refer to the legend to **Supplementary Figure 3**.

Supplementary Table 1 | EDIA report for cAMP and the binding site residues in chain C of 5AHW.

REFERENCES

- Aanandhi, M. V., Bhattacharjee, D., George, P. S. G., and Ray, A. (2014). Natural polyphenols down-regulate universal stress protein in *Mycobacterium tuberculosis*: an *in-silico* approach. *J. Adv. Pharm. Technol. Res.* 5, 171–178. doi: 10.4103/2231-4040.143036
- Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., Orhan, I. E., Banach, M., Röllinger, J. M., et al. (2021). Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* 20, 200–216. doi: 10.1038/s41573-020-00114-z
- Baell, J., and Walters, M. A. (2014). Chemistry: chemical con artists foil drug discovery. *Nature* 513, 481–483. doi: 10.1038/513481a
- Baell, J. B. (2010). Observations on screening-based research and some concerning trends in the literature. *Fut. Med. Chem.* 2, 1529–1546. doi: 10.4155/fmc.10.237
- Baell, J. B., and Holloway, G. A. (2010). New substructure filters for removal of Pan Assay Interference Compounds (PAIS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740. doi: 10.1021/jm901137j
- Bahuguna, A., and Rawat, D. S. (2020). An overview of new antitubercular drugs, drug candidates, and their targets. *Med. Res. Rev.* 40, 263–292. doi: 10.1002/med.21602
- Ballell, L., Bates, R. H., Young, R. J., Alvarez-Gomez, D., Alvarez-Ruiz, E., Barroso, V., et al. (2013). Fueling open-source drug discovery: 177 small-molecule leads against tuberculosis. *ChemMedChem* 8, 313–321. doi: 10.1002/cmde.201200428
- Banerjee, A., Adolph, R. S., Gopalakrishnapai, J., Kleinboelting, S., Emmerich, C., Steegborn, C., et al. (2015). A universal stress protein (USP) in mycobacteria binds cAMP. *J. Biol. Chem.* 290, 12731–12743. doi: 10.1074/jbc.M115.644856
- 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
- Bloemberg, G. V., Keller, P. M., Stucki, D., Trauner, A., Borrell, S., Latshang, T., et al. (2015). Acquired resistance to bedaquiline and delamanid in therapy for tuberculosis. *N. Engl. J. Med.* 373, 1986–1988. doi: 10.1056/NEJMc1505196
- Clark, J. J., Benson, M. L., Smith, R. D., and Carlson, H. A. (2019). Inherent versus induced protein flexibility: comparisons within and between apo and holo structures. *PLoS Comput. Biol.* 15:e1006705. doi: 10.1371/journal.pcbi.1006705
- Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544. doi: 10.1038/31159
- Cozzini, P., Kellogg, G. E., Spyraakis, F., Abraham, D. J., Costantino, G., Emerson, A., et al. (2008). Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* 51, 6237–6255. doi: 10.1021/jm800562d
- Dass, B. K. M., Sharma, R., Shenoy, A. R., Mattoo, R., and Visweswariah, S. S. (2008). Cyclic AMP in mycobacteria: characterization and functional role of the Rv1647 ortholog in *Mycobacterium smegmatis*. *J. Bacteriol.* 190, 3824–3834. doi: 10.1128/JB.00138-08
- Deller, M. C., and Rupp, B. (2015). Models of protein-ligand crystal structures: trust, but verify. *J. Comput. Aided. Mol. Des.* 29, 817–836. doi: 10.1007/s10822-015-9833-8
- Duan, J., Dixon, S. L., Lowrie, J. F., and Sherman, W. (2010). Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* 29, 157–170. doi: 10.1016/j.jmgm.2010.05.008
- Fradera, X., de la Cruz, X., Silva, C. H. T. P., Gelpi, J. L., Luque, F. J., and Orozco, M. (2002). Ligand-induced changes in the binding sites of proteins. *Bioinformatics* 18, 939–948. doi: 10.1093/bioinformatics/18.7.939
- 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
- Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *J. Med. Chem.* 49, 6177–6196. doi: 10.1021/jm051256o
- Ghodousi, A., Rizvi, A. H., Baloch, A. Q., Ghafoor, A., Khanzada, F. M., Qadir, M., et al. (2019). Acquisition of cross-resistance to bedaquiline and clofazimine following treatment for tuberculosis in Pakistan. *Antimicrob. Agents Chemother.* 63:e00915–19. doi: 10.1128/AAC.00915-19
- Greenwood, J. R., Calkins, D., Sullivan, A. P., and Shelley, J. C. (2010). Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput. Aided. Mol. Des.* 24, 591–604. doi: 10.1007/s10822-010-9349-1
- Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* 47, 1750–1759. doi: 10.1021/jm030644s
- Holm, L. (2020). DALI and the persistence of protein shape. *Protein Sci.* 29, 128–140. doi: 10.1002/pro.3749
- Huang, S.-Y., Grinter, S. Z., and Zou, X. (2010). Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* 12, 12899–12908. doi: 10.1039/c0cp00151a
- James, S. F., Marlys, H., Sharada, M., Scott, K. T., David, G. W., Kazuya, K., et al. (2009). *1H-PYRROLO[2,3-B]PYRIDINES*. Current Assignee GlaxoSmithKline LLC. U.S. Patent No. 2009/0233955 A1. <https://patents.google.com/patent/US20090233955A1> (accessed September 17, 2009).
- Kalamidas, S. A., Kuehnle, M. P., Peyron, P., Rybin, V., Rauch, S., Kotoulas, O. B., et al. (2006). cAMP synthesis and degradation by phagosomes regulate actin assembly and fusion events: consequences for mycobacteria. *J. Cell Sci.* 119, 3686–3694. doi: 10.1242/jcs.03091
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2018). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109. doi: 10.1093/nar/gky1033
- Koul, A., Dendouga, N., Vergauwen, K., Molenberghs, B., Vranckx, L., Willebrords, R., et al. (2007). Diarylquinolines target subunit c of mycobacterial ATP synthase. *Nat. Chem. Biol.* 3, 323–324. doi: 10.1038/nchembio884
- Lakshmanan, M., and Xavier, A. S. (2013). Bedaquiline - the first ATP synthase inhibitor against multi drug resistant tuberculosis. *J. Young Pharm.* 5, 112–115. doi: 10.1016/j.jyp.2013.12.002
- Lipinski, C. A. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249. doi: 10.1016/S1056-8719(00)00107-6
- Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25. doi: 10.1016/S0169-409X(96)00423-1
- Lowrie, D. B., Aber, V. R., and Jaccett, P. S. (1979). Phagosome-lysosome fusion and cyclic adenosine 3': 5'-monophosphate in macrophages infected with *Mycobacterium microti*, *Mycobacterium bovis* BCG or *Mycobacterium lepraemurium*. *Microbiology* 110, 431–441. doi: 10.1099/00221287-110-2-431

- Lowrie, D. B., Jackett, P. S., and Ratcliffe, N. A. (1975). *Mycobacterium microti* may protect itself from intracellular destruction by releasing cyclic AMP into phagosomes. *Nature* 254, 600–602. doi: 10.1038/254600a0
- Lyne, P. D., Lamb, M. L., and Saeh, J. C. (2006). Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem.* 49, 4805–4808. doi: 10.1021/jm060522a
- Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268
- Manjunatha, U. H., and Smith, P. W. (2015). Perspective: challenges and opportunities in TB drug discovery from phenotypic screening. *Bioorg. Med. Chem.* 23, 5087–5097. doi: 10.1016/j.bmc.2014.12.031
- McGovern, S. L., and Shoichet, B. K. (2003). Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* 46, 2895–2907. doi: 10.1021/jm0300330
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2018). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 47, D930–D940. doi: 10.1093/nar/gky1075
- Meyder, A., Nittinger, E., Lange, G., Klein, R., and Rarey, M. (2017). Estimating electron density support for individual atoms and molecular fragments in x-ray structures. *J. Chem. Inf. Model.* 57, 2437–2447. doi: 10.1021/acs.jcim.7b00391
- Mohs, R. C., and Greig, N. H. (2017). Drug discovery and development: role of basic biological research. *Alzheimer's Dement.* 3, 651–657. doi: 10.1016/j.trci.2017.10.005
- Mugumbate, G., Abrahams, K. A., Cox, J. A. G., Papadatos, G., van Westen, G., Lelièvre, J., et al. (2015). Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and *in vitro* validation. *PLoS ONE* 10:e0121492. doi: 10.1371/journal.pone.0121492
- Mushtaq, S., Abbasi, B. H., Uzair, B., and Abbasi, R. (2018). Natural products as reservoirs of novel therapeutic agents. *EXCLI J.* 17, 420–451. doi: 10.17179/excli2018-1174
- Nieto Ramirez, L. M., Quintero Vargas, K., and Diaz, G. (2020). Whole genome sequencing for the analysis of drug resistant strains of *Mycobacterium tuberculosis*: a systematic review for bedaquiline and delamanid. *Antibiotics* 9:133. doi: 10.3390/antibiotics9030133
- Padh, H., and Venkatasubramanian, T. A. (1976). Adenosine 3', 5'-monophosphate in *Mycobacterium phlei* and *Mycobacterium tuberculosis* H37Ra. *Microbios* 16, 183–189.
- Rajashankar, K. R., Kniewel, R., Solorzano, V., Lima, C. D., and Burley, S. K. (2004). Crystal structure of protein Rv1636 from *Mycobacterium tuberculosis* H37Rv. *New York SGX Res. Cent. Struct. Genomics*. doi: 10.2210/pdb1tq8/pdb
- Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 42, W320–W324. doi: 10.1093/nar/gku316
- 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
- Rueda, M., Bottegoni, G., and Abagyan, R. (2010). Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* 50, 186–193. doi: 10.1021/ci9003943
- Sastry, M., Lowrie, J. F., Dixon, S. L., and Sherman, W. (2010). Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* 50, 771–784. doi: 10.1021/ci100062n
- Sastry, M. G., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided. Mol. Des.* 27, 221–234. doi: 10.1007/s10822-013-9644-8
- Sharma, D., Lata, M., Singh, R., Deo, N., Venkatesan, K., and Bisht, D. (2016). Cytosolic proteome profiling of aminoglycosides resistant *Mycobacterium tuberculosis* clinical isolates using MALDI-TOF/MS. *Front. Microbiol.* 7:1816. doi: 10.3389/fmicb.2016.01816
- Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007). Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided. Mol. Des.* 21, 681–691. doi: 10.1007/s10822-007-9133-z
- Shenoy, A. R., and Visweswariah, S. S. (2006). New messages from old messengers: cAMP and mycobacteria. *Trends Microbiol.* 14, 543–550. doi: 10.1016/j.tim.2006.10.005
- Shetye, G. S., Franzblau, S. G., and Cho, S. (2020). New tuberculosis drug targets, their inhibitors, and potential therapeutic impact. *Transl. Res.* 220, 68–97. doi: 10.1016/j.trsl.2020.03.007
- Singh, V., Jamwal, S., Jain, R., Verma, P., Gokhale, R., and Rao, K. V. S. (2012). *Mycobacterium tuberculosis*-driven targeted recalibration of macrophage lipid homeostasis promotes the foamy phenotype. *Cell Host Microbe* 12, 669–681. doi: 10.1016/j.chom.2012.09.012
- Stefanachi, A., Leonetti, F., Pisani, L., Catto, M., and Carotti, A. (2018). Coumarin: a natural, privileged and versatile scaffold for bioactive compounds. *Molecules* 23:250. doi: 10.3390/molecules23020250
- Stinear, T. P., Seemann, T., Harrison, P. F., Jenkin, G. A., Davies, J. K., Johnson, P. D. R., et al. (2008). Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Res.* 18, 729–741. doi: 10.1101/gr.075069.107
- Syui, M., Arif, S. M., and Malim, N. (2013). "Comparison of similarity coefficients for chemical database retrieval," in *Proceedings of the 2013 1st International Conference on Artificial Intelligence, Modelling and Simulation AIMS '13* (Kota Kinabalu: IEEE Computer Society), 129–133.
- Taylor, S. S., Keshwani, M. M., Steichen, J. M., and Kornev, A. P. (2012). Evolution of the eukaryotic protein kinases as dynamic molecular switches. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 367, 2517–2528. doi: 10.1098/rstb.2012.0054
- Walters, W. P., Stahl, M. T., and Murcko, M. A. (1998). Virtual screening—an overview. *Drug Discov. Today* 3, 160–178. doi: 10.1016/S1359-6446(97)01163-X
- WHO (2019). *Global Tuberculosis Report 2019*. Available online at: <https://www.who.int/tb/global-report-2019>
- WHO (2020). *Global Tuberculosis Report 2020*. Available online at: <https://apps.who.int/iris/bitstream/handle/10665/336069/9789240013131-eng.pdf>
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672. doi: 10.1093/nar/gkj067
- Xavier, A. S., and Lakshmanan, M. (2014). Delamanid: a new armor in combating drug-resistant tuberculosis. *J. Pharmacol. Pharmacother.* 5, 222–224. doi: 10.4103/0976-500X.136121
- Yadav, M., Roach, S. K., and Schorey, J. S. (2004). Increased mitogen-activated protein kinase activity and TNF- α production associated with *Mycobacterium smegmatis*-but not *Mycobacterium avium*-infected macrophages requires prolonged stimulation of the calmodulin/calmodulin kinase and cyclic AMP/protein kin. *J. Immunol.* 172, 5588–5597. doi: 10.4049/jimmunol.172.9.5588

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

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



Structure-Guided Computational Approaches to Unravel Druggable Proteomic Landscape of *Mycobacterium leprae*

Sundeep Chaitanya Vedithi^{1*}, Sony Malhotra^{2†}, Marta Acebrón-García-de-Eulate¹, Modestas Matusevicius¹, Pedro Henrique Monteiro Torres³ and Tom L. Blundell^{1*}

¹ Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, ² Rutherford Appleton Laboratory, Science and Technology Facilities Council, Oxon, United Kingdom, ³ Laboratório de Modelagem e Dinâmica Molecular, Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:

Alexandre G. De Brevern,
INSERM U1134 Biologie Intégrée du
Globule Rouge, France

Reviewed by:

Stéphane Téletchéa,
Université de Nantes, France
Jeremy Esque,
Institut Biotechnologique de Toulouse
(INSA), France

*Correspondence:

Sundeep Chaitanya Vedithi
scv26@cam.ac.uk
Tom L. Blundell
tlb20@cam.ac.uk

[†] 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: 02 February 2021

Accepted: 12 April 2021

Published: 07 May 2021

Citation:

Vedithi SC, Malhotra S,
Acebrón-García-de-Eulate M,
Matusevicius M, Torres PHM and
Blundell TL (2021) Structure-Guided
Computational Approaches
to Unravel Druggable Proteomic
Landscape of *Mycobacterium leprae*.
Front. Mol. Biosci. 8:663301.
doi: 10.3389/fmolb.2021.663301

Leprosy, caused by *Mycobacterium leprae* (*M. leprae*), is treated with a multidrug regimen comprising Dapsone, Rifampicin, and Clofazimine. These drugs exhibit bacteriostatic, bactericidal and anti-inflammatory properties, respectively, and control the dissemination of infection in the host. However, the current treatment is not cost-effective, does not favor patient compliance due to its long duration (12 months) and does not protect against the incumbent nerve damage, which is a severe leprosy complication. The chronic infectious peripheral neuropathy associated with the disease is primarily due to the bacterial components infiltrating the Schwann cells that protect neuronal axons, thereby inducing a demyelinating phenotype. There is a need to discover novel/repurposed drugs that can act as short duration and effective alternatives to the existing treatment regimens, preventing nerve damage and consequent disability associated with the disease. *Mycobacterium leprae* is an obligate pathogen resulting in experimental intractability to cultivate the bacillus *in vitro* and limiting drug discovery efforts to repositioning screens in mouse footpad models. The dearth of knowledge related to structural proteomics of *M. leprae*, coupled with emerging antimicrobial resistance to all the three drugs in the multidrug therapy, poses a need for concerted novel drug discovery efforts. A comprehensive understanding of the proteomic landscape of *M. leprae* is indispensable to unravel druggable targets that are essential for bacterial survival and predilection of human neuronal Schwann cells. Of the 1,614 protein-coding genes in the genome of *M. leprae*, only 17 protein structures are available in the Protein Data Bank. In this review, we discussed efforts made to model the proteome of *M. leprae* using a suite of software for protein modeling that has been developed in the Blundell laboratory. Precise template selection by employing sequence-structure homology recognition software, multi-template modeling of the monomeric models and accurate quality assessment are the hallmarks of the modeling process. Tools that map interfaces and enable building of homo-oligomers are discussed in the context of interface stability. Other software is described to determine the druggable proteome by using information related to the chokepoint analysis of the metabolic

pathways, gene essentiality, homology to human proteins, functional sites, druggable pockets and fragment hotspot maps.

Keywords: *Mycobacterium leprae*, amino acid substitution, chokepoint reactions, drug binding sites, homology (comparative) modeling, protein interface

INTRODUCTION

Mycobacterium leprae causes leprosy in about 200,000 people each year globally. Leprosy is a dermato-neurological infectious disease with varied clinical manifestations, often resulting in peripheral sensorimotor/demyelinating neuropathy leading to permanent nerve damage and disability. The World Health Organization currently recommends a combinatorial therapy [multidrug therapy (MDT)] with Dapsone, Rifampicin (Rifampin) and Clofazimine to treat leprosy (Manghani and Arif, 2006). MDT has proven effective in reducing the prevalence and controlling the incidence from about 5 million new cases in the 1990s to ~200,000 new cases from the year 2005 (after India declared the elimination of leprosy). However, with the emergence of single and multidrug-resistant strains of *M. leprae*, novel therapies are essential to curb ongoing transmission of the disease. Also, the current therapy duration with MDT is 1 year leading to reduced treatment compliance and increased defaulter rates globally (Cambau et al., 2018).

Mycobacterium leprae is phylogenetically the closest bacterial species to *Mycobacterium tuberculosis* (*M. tuberculosis*). However, the *M. leprae* has a reduced genome of 3.2 Mbp, compared to 4.4 Mbp in *M. tuberculosis*, and survive with only 1,614 protein coding genes which are largely annotated based on the features of their homologues in *M. tuberculosis* and other mycobacterial species (Cole et al., 2001). Dapsone interacts with bacterial dihydropteroate synthase, an enzyme essential for folic acid biosynthesis in bacteria. It is absent in humans (Cambau et al., 2006). Rifampin interacts with RNA polymerase, an enzyme critical for DNA dependent RNA synthesis (transcription) in *M. leprae*. These drugs are either bacteriostatic or bactericidal. However, they do not interfere with predilection of *M. leprae* for human nerve cells, which is a significant cause for demyelinating neuropathy in leprosy (Lockwood and Saunderson, 2012). Newer antibacterial agents that can effectively treat the disease within a short duration of time and prevent nerve damage are essential to reduce morbidity associated with the disease (Rao and Jain, 2013). Currently known drugs for leprosy, their drug target proteins and references related to their mechanisms of action are listed in **Table 1**.

Knowledge of the structural components of the proteome of *M. leprae* is critical for identifying drug target proteins and deciphering their essential roles in the survival of the pathogen. Key enzymes that catalyze chokepoint reactions can act as potential drug targets for antimycobacterial discovery. However, information related to 3D structures of these proteins is scarce for *M. leprae*, necessitating a more focussed structural genomics effort to unravel the druggable proteomic landscape of this bacillus long known to humankind.

Software tools that predict stability and affinity changes in drug-target proteins due to substitution mutations are discussed

in the context of antimicrobial resistance. While the emphasis is on deciphering the druggable proteome, we provide a brief overview of the structure-guided virtual screening tools and methods that facilitate the chemical expansion of fragment-like small molecules to lead-like or drug-like compounds in the active or allosteric sites of the target protein.

Proteome Modeling in *Mycobacterium leprae* and Its Relevance to Structure-Guided Drug Discovery

Of the 1,614 annotated genes that are expressed in *M. leprae*, the structures of only 17 proteins are available (see **Table 2**) to date in the publicly available databases [Protein Data Bank (PDB) (Berman et al., 2000)], as opposed to around 1,277 entries for *Mycobacterium tuberculosis*. Solving the crystal/cryoEM structures of all the potential drug targets in *M. leprae* requires costly and labor intensive effort. Given the high sequence identity of many of the *M. leprae* proteins with their homologous counterparts in *M. tuberculosis* with solved structures in the PDB, employing computational tools to perform comparative modeling of proteins in *M. leprae* can be a robust alternative for acquiring a preliminary understanding of the functional sites and small molecule interactions.

Different groups have made several attempts to model the proteins of *M. leprae*. **Table 3** lists two web-resources where such information is available. Computational protein structure modeling can reduce the sequence-structure gaps and enable genome-scale modeling of infectious pathogens (Bienert et al., 2017). Although the paucity of structural proteomics information for *M. leprae* in the publicly available databases is a challenge, the software developed in the Blundell laboratory will be useful in performing proteome scale modeling pipeline (*Vivace*) for proteomes of Mycobacterial pathogens and other bacterial species (Skwark et al., 2019). *Vivace* optimizes template selection, enables sequence-structure alignments, and constructs optimal quality models in both apo- and ligand-bound states. To facilitate multi-template modeling, protein structures from the entire PDB are initially organized in a structural profile database named TOCCATA (Ochoa-Montano et al., 2015). Protein structures within each profile are classified based on domain annotations in CATH (Sillitoe et al., 2019) and SCOP (Andreeva et al., 2020) databases, pre-aligned and functionally annotated with information derived from UniProt (The UniProt Consortium, 2021) and PDB. The query protein sequence is aligned with representative structures from each cluster using a sequence-structure alignment tool named FUGUE (Shi et al., 2001). FUGUE recognizes distant homologues by sequence-structure comparison using environment-specific substitution tables and

structure-dependent gap penalties. Alignments generated by FUGUE are fed into Modeler 9.24 (Webb and Sali, 2016) for model building. The ligands and other small molecules are modeled into corresponding protein structure models at sites recognized from the ligand-bound templates. Multiple models are generated based on the number of cluster hits, ranging from 3 to ~1,000 models per query sequence in the *M. leprae* proteome. These models are of different states (ligand-bound and apomeric) and of varying quality based on the templates used in each profile.

Once modeled, each of the protein structure models undergoes a rigorous quality assessment by employing methods such as NDOPE, GA341 (Shen and Sali, 2006), GOAP (Zhou and Skolnick, 2011), SOAP (Webb and Sali, 2016), Molprobit (Chen et al., 2010) and secondary structure

agreement score (Eramian et al., 2006). Models with extensive chain clashes, poorly resolved loops and improperly fitted ligands are ranked low in the consensus quality scoring process described in CHOPIN—a web resource for structural and functional proteome of *Mycobacterium tuberculosis* (Ochoa-Montañón et al., 2015).

Vivace is being used to model the proteome of *M. leprae*. Sequence and structural features at the genome-scale are being analyzed to identify essential enzymes that drive chokepoint metabolic reactions. Models in apomeric, ligand-bound and oligomeric (discussed in the later sections) states are being generated and analyzed for surface topology, cavities (Binkowski et al., 2003) and fragment hotspots (sites for potential small molecule binding) (Radoux et al., 2016). The schematic workflow shown in **Figure 1** illustrates the modeling

TABLE 1 | Drugs and their corresponding target proteins in *M. leprae*.

Drug	Target proteins/Ribosomal subunits	Gene (gene name)	References
Dapsone	Dihydropteroate synthase (DHPS)	<i>folP1</i> (ML0224)	Williams et al., 2000
Rifampin	β-subunit of the DNA-dependent RNA polymerase	<i>rpoB</i> (ML1891)	Lin et al., 2017
Clofazimine	Unknown	-	Lechartier and Cole, 2015
Fluoroquinolones	DNA gyrase subunit A	<i>gyrA</i> (ML0006)	Blower et al., 2016
	DNA gyrase subunit B	<i>gyrB</i> (ML0005)	Yamaguchi et al., 2016
Macrolides	50S subunit (23S rRNA in particular)	-	Ji et al., 1996
Minocycline	30S ribosomal subunit, blocking the binding of aminoacyl-tRNA to the 16S rRNA	-	Ji et al., 1996
Thioamides	Enoyl-ACP-reductase	<i>inhA</i> (ML1806)	Wang et al., 2007
Bedaquiline	Proton pump of ATP synthase	<i>atpE</i> (ML1140)	Guo et al., 2021
Epiroprim	Dihydrofolate reductase	<i>folA</i> (ML1518)	Dhople, 2002

TABLE 2 | List of protein structures available for *M. leprae* in Protein Data Bank

Gene Id	PDB Id	Description	References
ML2441	4EO9	Crystal structure of a phosphoglycerate mutase gpm1 from <i>Mycobacterium leprae</i>	Baugh et al., 2015
ML0210	4ECP	X-ray crystal structure of Inorganic Pyrophosphate PPA from <i>Mycobacterium leprae</i>	Unpublished
ML0560	4J07	Crystal structure of a PROBABLE RIBOFLAVIN SYNTHASE, BETA CHAIN RIBH (6,7-dimethyl-8-ribityllumazine synthase, DMRL synthase, Lumazine synthase) from <i>Mycobacterium leprae</i>	Unpublished
ML1382	5IE8	The pyrazinoic acid binding domain of Ribosomal Protein S1 from <i>Mycobacterium tuberculosis</i> *	Huang B. et al., 2016
ML0482	1BVS	RUVA Complexed to a Holliday Junction	Roe et al., 1998
ML2684	3AFP	Crystal structure of the single-stranded DNA binding protein from <i>Mycobacterium leprae</i> (Form II)	Kaushal et al., 2010
ML2640	2CKD	Crystal structure of ML2640 from <i>Mycobacterium leprae</i>	Graña et al., 2007
ML0380	1LEP	Three-Dimensional Structure of the Immunodominant Heat-Shock Protein Chaperonin-10 of <i>Mycobacterium Leprae</i>	Mande et al., 1996
ML1962	3I4O	Crystal Structure of Translation Initiation Factor 1 from <i>Mycobacterium tuberculosis</i> *	Hatzopoulos and Mueller-Dieckmann, 2010
ML2428A	5O61	The complete structure of the <i>Mycobacterium smegmatis</i> 70S ribosome*	Hentschel et al., 2017
ML1485	4WKW	Crystal Structure of a Conserved Hypothetical Protein from <i>Mycobacterium leprae</i> Determined by Iodide SAD Phasing	Unpublished
ML2174	3R2N	Crystal structure of cytidine deaminase from <i>Mycobacterium leprae</i>	Baugh et al., 2015
ML1806	2NTV	<i>Mycobacterium leprae</i> InhA bound with PTH-NAD adduct	Wang et al., 2007
ML2684	3AFQ	Crystal structure of the single-stranded DNA binding protein from <i>Mycobacterium leprae</i> (Form II)	Kaushal et al., 2010
ML2069	4EX4	The Structure of GlcB from <i>Mycobacterium leprae</i>	Unpublished
ML2640	2UYO	Crystal structure of ML2640c from <i>Mycobacterium leprae</i> in an hexagonal crystal form	Graña et al., 2007
ML2640	2UYQ	Crystal structure of ML2640c from <i>Mycobacterium leprae</i> in complex with S-adenosylmethionine	Graña et al., 2007

*The solved region of the protein structure is 100% in sequence identity with *M. leprae*.

TABLE 3 | Web resources with models of *M. leprae* proteins (modelled using single templates).

Web resource	Description	Availability	References
ModBase	A database of annotated comparative protein structure models and associated resources	https://modbase.compbio.ucsf.edu/modbase/cgi/index.cgi	Pieper et al., 2011
SwissModel Repository	A database of annotated 3D protein structure models generated by the SWISS-MODEL homology-modeling pipeline	https://swissmodel.expasy.org/repository	Bienert et al., 2017

procedures adopted by our group to model proteomes of mycobacterial pathogens.

Approaches to Predict Homo/Hetero-Oligomeric Complexes

Protein-protein interactions (homo/hetero) govern a majority of the cellular processes. Structure determination of these complexes is crucial for understanding their functions. Usually, the experimental techniques used to unravel interacting protein partners are time consuming, challenging and expensive. There have been significant advances in the development of computational methods and tools to identify interacting pairs and predict the structures of protein-protein complexes (Das and Chakrabarti, 2021).

The computational tools for predicting protein-protein interactions developed over the years can be classified into the knowledge-based or *de novo* prediction methods. If the structures of the interacting partners are known, the interactions can be predicted using template-based, or template free and/or restraint-based docking. Template-based docking can provide the multi-component modeled complex but requires the presence of multi-component template structures (Ogmen et al., 2005; Mukherjee and Zhang, 2011). If the homologous multi-component template is unavailable, protein-protein docking approaches can be used to sample the conformational space and predict the docked complexes which are further scored using different schemes to discriminate native-like conformations from a pool of docked solutions. These different approaches for computational modeling of protein interactions were recently reviewed by Soni and Madhusudhan (2017).

Recently, tools have been developed which can make use of the wealth of sequence information available for protein sequences to predict/model interactions accurately. Machine learning approaches including deep learning have played a significant role in training models which can predict the interactions using the features derived from protein sequences alone (Huang Y.-A. et al., 2016; Du et al., 2017; Sun et al., 2017; Chen et al., 2019). The inspection of co-evolving sites in two protein partners can provide strong signals to elucidate interacting partners (Yu et al., 2016). A recent method CoFex (Hu and Chan, 2017) used co-evolutionary features to predict protein interactions. Ensemble based approaches which use multiple machine learning methods to vote for predictions have been reported to achieve high sensitivity and accuracy (Zhang et al., 2019; Li et al., 2020). Deep learning has also been employed to train a convolutional neural network (CNN) to predict the protein interacting pairs with high accuracy (Wang et al., 2019; Torrisi et al., 2020).

However, *in-silico* approaches can often give false positive or negative results as well, hence one also needs validation strategies to assess the quality of predicted interactions. Efforts in the community such as CASP (Critical Assessment of Structure Prediction) and CAPRI (Critical Assessment of Prediction of Interfaces) competitions, aim to assess the field and the state-of-the-art methods and their ability to “correctly” model protein structures and their interactions, respectively. They define and use multiple scores for assessing the quality of protein structure and interfaces in the modeled complexes. CASP14 is the present ongoing competition, where deep learning approach-AlphaFold2 has outperformed and were able to accurately predict the protein structures (AlQuraishi, 2019).

To illustrate the modeling process adopted by *Vivace*, **Figure 2** depicts the models of three potential drug targets in *M. leprae*, the *menB* [1,4-dihydroxy-2-naphthoyl-CoA synthase (ML2263)], *menD* [2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase (ML2270)] and *coaA* [Pantothenate kinase (ML1954)].

The gene product of *menB* converts o-succinylbenzoyl-CoA (OSB-CoA) to 1,4-dihydroxy-2-naphthoyl-CoA (DHNA-CoA) and its homologue in *M. tuberculosis* (Rv0548c) is reported as essential (DeJesus et al., 2017). We built the model using the structure of its orthologous protein in *M. tuberculosis* (PDB Id: 4QII) as the template with sequence identity of 93% and sequence coverage of 100% (**Figure 2A**). The gene product of *menD* catalyzes the thiamine diphosphate-dependent decarboxylation of 2-oxoglutarate. Its homologue in *M. tuberculosis* (Rv0555) is noted to be essential for bacterial survival. We modeled *menD* of *M. leprae* using the structure of the *M. tuberculosis* orthologue (PDB Id: 5ESD) as the template with the sequence identity of 86% and sequence coverage of 99% (**Figure 2B**). Finally, we modeled *coaA* which synthesizes coA from (R)—Pantothenate. CoaA has been recognized as a drug target in tuberculosis (Chiarelli et al., 2018). We modeled *coaA* using its orthologue in *M. tuberculosis* (PDB Id: 2GET) as the template with sequence identity of 93% and sequence coverage of 98% (**Figure 2C**).

Structural Implications of Substitution Mutations

Development of drug resistance is recognized as one of the major hurdles to disease management and control. For *M. leprae*, it is even more challenging as it relies on mouse footpad models (Vedithi et al., 2018). Antimicrobial resistance was noted in Dapson, Rifampicin and Ofloxacin (a second-line drug). Treating and managing the disease is a big hurdle due to emerging drug resistance for these three drugs and lack of alternative effective treatments.

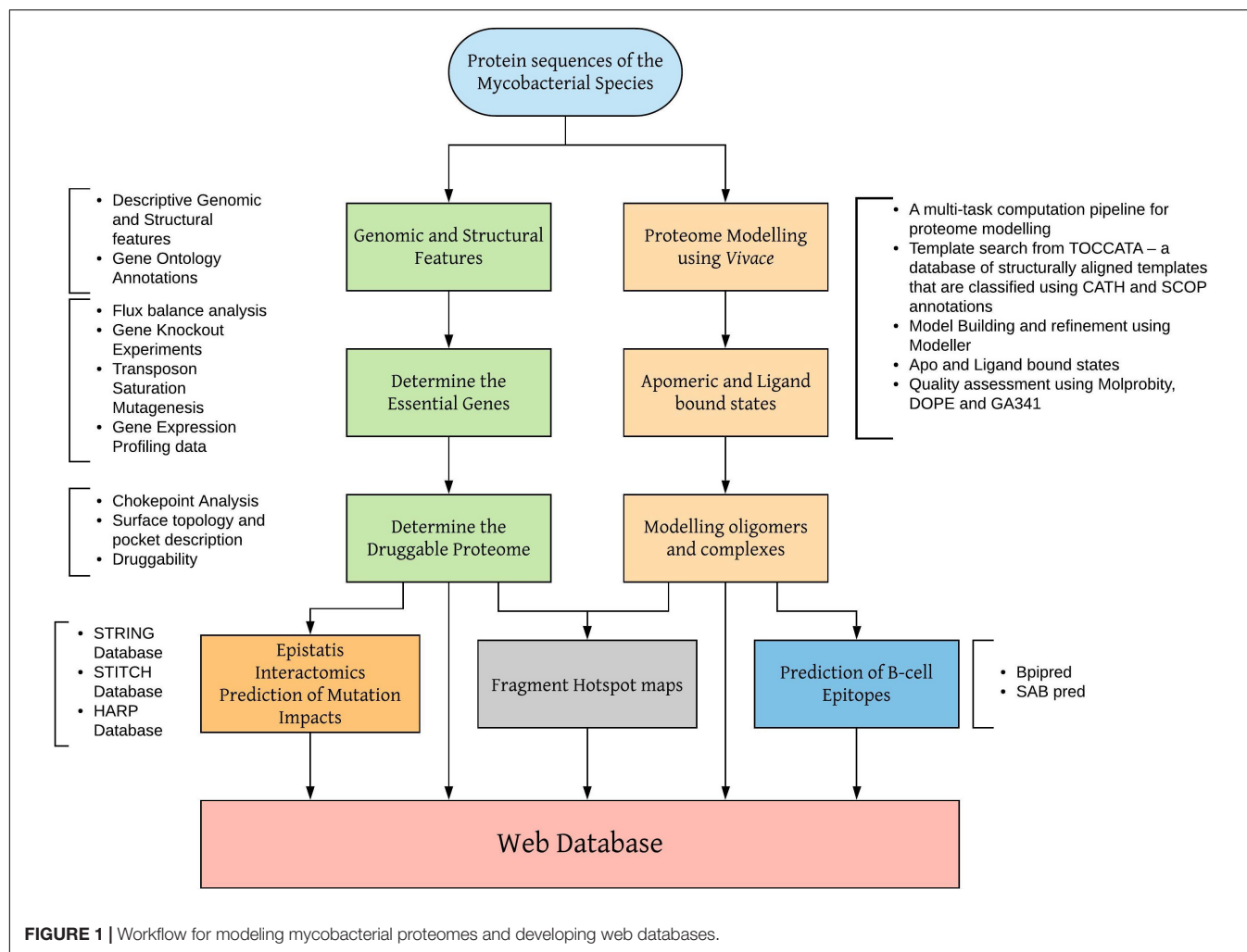


FIGURE 1 | Workflow for modeling mycobacterial proteomes and developing web databases.

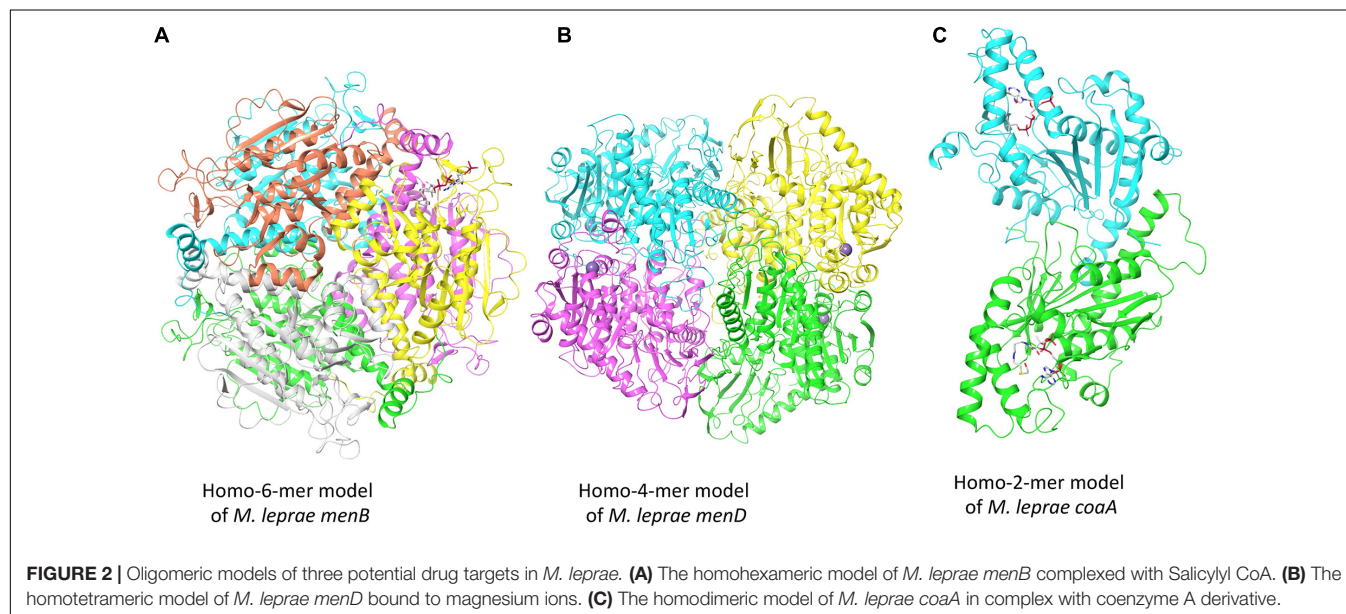


FIGURE 2 | Oligomeric models of three potential drug targets in *M. leprae*. **(A)** The homohexameric model of *M. leprae menB* complexed with Salicylyl CoA. **(B)** The homotetrameric model of *M. leprae menD* bound to magnesium ions. **(C)** The homodimeric model of *M. leprae coaA* in complex with coenzyme A derivative.

The drug-resistance mutations when mapped on to the three-dimensional structure of the drug target can provide crucial insights into effects on protein structure and function. There are several available software/web servers that can predict the impacts of mutation on protein stability and interactions with other proteins, ligands and nucleic acids. We have provided a list of some of the commonly used software tools for investigating the effects of mutations on protein structure and function (Table 4).

Our own group have developed over the past decade the mCSM suite of computer programmes that use ML/AI approaches to predict the impacts of amino acid mutations not only on protomer stability (Pires et al., 2014a) but also on protein-protein, protein nucleic acid and protein-ligand interactions (Pires et al., 2016; Pires and Ascher, 2017). Recently, there have been further developments in the field where machine learning (ML)-based methods are gaining

popularity. Many more recent ML methods also use features derived from protein structure and/or sequence to predict the effect of mutations (Hopf et al., 2017). A recent review, summarizes the performance of different ML methods and emphasizes the need for selecting reliable training dataset and informative features (Fang, 2020). Deep learning is an advanced training which is composed of multiple layers to learn different features from the input data and is proven to learn from the high-dimensional data. Recently, a method called DeepCLIP (Grønning et al., 2020) has been proposed which can predict protein binding to RNA using only sequence data. Another recently developed deep learning framework-MuPIPR (Zhou et al., 2020) (Mutation Effects in Protein-protein Interaction Prediction Using Contextualized Representations), can predict the effects of mutation on protein-protein interactions in terms of changes in buried surface area and binding affinity.

TABLE 4 | Some of the commonly used tools for predicting the effect of mutations on protein structure and function.

Software	Description	Availability	References
SIFT	Amino acid substitution effect on protein function	https://sift.bii.a-star.edu.sg/	Ng and Henikoff, 2003
PolyPhen-2	Amino acid substitution effect on protein structure and function using protein sequence	http://genetics.bwh.harvard.edu/pph2/	Adzhubei et al., 2010
SNPs3D	Amino acid substitution effect on protein structure and function using SVM based model	http://snps3d.org/	Yue et al., 2006
MutPred2	Machine learning approach to quantify pathogenicity of mutation	http://mutpred.mutdb.org/index.html	Pejaver et al., 2020
PROVEAN	Impact of mutation on protein function by using multiple sequence alignment	http://provean.jcvi.org/index.php	Choi and Chan, 2015
mCSM	Effect of mutation on protein structure and interactions using graph-based signatures	http://biosig.unimelb.edu.au/mcsm/	Pires et al., 2014a
SDM2	Effect of mutation on protein structure and interactions using environment-specific amino-acid substitution frequencies	http://marid.bioc.cam.ac.uk/sdm2	Pandurangan et al., 2017, 2
DUET	Consensus prediction of mCSM and SDM2 for protein stability	http://biosig.unimelb.edu.au/duet/	Pires et al., 2014b
PoPMuSiC-2	Effects of mutation on protein stability using statistical potentials	http://dezyme.com/en/Services	Dehouck et al., 2009
FoldX	Change in free energy using force fields-based method	http://foldxsuite.crg.eu/	Schymkowitz et al., 2005
Hunter	Predicting protein stability upon mutation using side chain interactions	http://bioinfo41.weizmann.ac.il/hunter/	Potapov et al., 2010
MAESTRO	Measures changes in free energy upon mutation using machine learning	https://pbwww.che.sbg.ac.at/?page_id=416	Laimer et al., 2015
I-Mutant3.0	SVM based prediction of protein stability change upon mutation using either sequence and/or structure	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi	Capriotti et al., 2008
MUPro	SVM and neural network-based prediction of changes in protein stability	http://mupro.proteomics.ics.uci.edu/	Cheng et al., 2006
iStable	Change in free energy using SVM based predictor	http://predictor.nchu.edu.tw/istable/	Chen et al., 2013
MutaBind	Change in free energy using force fields, statistical potentials and side-chain optimisation methods	https://www.ncbi.nlm.nih.gov/research/mutabind/index.fcgi/	Li et al., 2016
BeAtMuSiC	Impact of mutations on protein-protein interactions using statistical potentials	http://babylone.ulb.ac.be/beatmusic/	Dehouck et al., 2013
SNAP2	Predict functional impacts of mutations using neural network-based model	https://roslab.org/services/snap2web/	Hecht et al., 2015
Envision	Supervised, stochastic gradient boosting algorithm to quantify the effect of mutation	https://envision.gs.washington.edu/shiny/envision_new/	Gray et al., 2018
EVmutation	Unsupervised statistical method to predict effect of mutations using residue dependencies between positions	https://marks.hms.harvard.edu/evmutation/	

In-silico Saturation Mutagenesis

Using the tools described above, computational efforts exploiting recent growth in the availability of computing power can be immensely helpful to perform saturation mutagenesis, which can be used as a surveillance tool for drug resistance in leprosy. These mutational scanning exercises can provide crucial insights into the structure-function relationships by exploring all possible substitutions at a given site. This can provide a glimpse into the functional consequences of mutations in antimicrobial-resistance phenotypes. The extensive quantitative data from computational saturation mutagenesis experiments can guide experimental approaches and prove helpful for validation and/or engineering purposes. Recently published HARP (a database of Hansen's Disease Antimicrobial Resistance Profiles) database (Vedithi et al., 2020) is a comprehensive repository of *in-silico* mutagenesis experiments for three important drug targets for *M. leprae* namely dihydropteroate synthase, RNA polymerase and DNA gyrase. A consensus impact for all the possible mutations on protein stability and function of these drug targets is provided in this database.

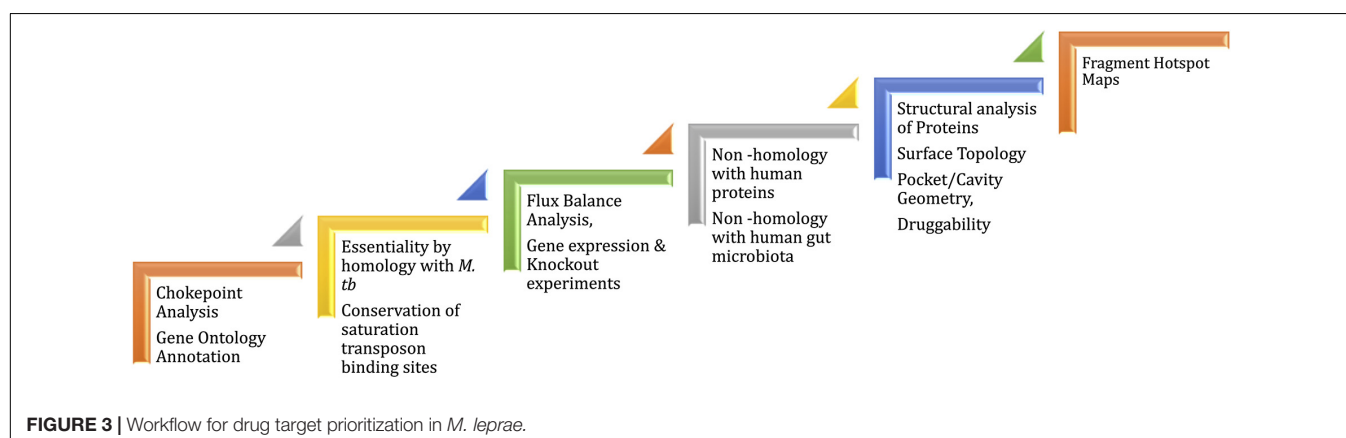
Druggability

Mycobacterium leprae genome is reduced to 3,268,203 bp preserving only 1,614 ORFs (Cole et al., 2001; Liu et al., 2004) of the *Mycobacterial* genus. The genome reduction is due to evolutionary adaptation of this intracellular obligate bacillus to Schwann and macrophages cells. Gene essentiality in *M. leprae* is deciphered based on essentiality of homologous genes, mainly in *M. tuberculosis* that are determined by experiments (Sasseti et al., 2003; DeJesus et al., 2017). Because of the evolutionary loss of non-essential genes by pseudogenization, only 65% of the existing total of *M. leprae* genes have been demonstrated to be essentials (Borah et al., 2020). In order to identify potential drug targets, a chokepoint reaction analysis helps to find proteins that are either consumers of unique substrates, or are unique producers of metabolites. It is predicted that the inhibition of chokepoint proteins produces an interruption of essential cell functions (Yeh et al., 2004). Determining the druggability of protein targets is important to avoid intractable targets. A druggable protein has the ability to bind with high affinity to a drug. In leprosy, the

dihydropteroate synthase (DHPS), RNA polymerase (RNAP) and DNA gyrase (GYR) are known druggable proteins as they are the targets of Dapsone, Rifampicin and Ofloxacin, respectively. Nevertheless, protein druggability properties can be predicted by different bioinformatics tools based on the 3D structure /model of the protein. For example, the α -1,2-mannosyltransferase and mannosyltransferase proteins related to lipoarabinomannan pathway were identified as a possible drug targets using CASTp (Computer Atlas of Surface Topography of proteins) (Gupta et al., 2020). CASTp determines the volume and the area of each cavity and pocket. Furthermore, for each pocket the solvent accessible surface and the molecular surface are calculated. Small-molecule virtual screening is another *in-silico* strategy used to ascertain druggability of the protein target. This approach provides an understanding of the physicochemical features of the binding sites and potential ligands that bind at these sites. In *Mycobacterium tuberculosis*, 2,809 proteins are identified as druggable using this *in-silico* approach (Anand and Chandra, 2014). Mammalian cell entry proteins of the class *mce1A-E* have been reported in *M. leprae* to facilitate bacterial entry into human nasal epithelial cells (Fadlitha et al., 2019). *Mce1A* has a significant role in the cell predilection and a comprehensive understanding of the structure and druggability of this protein can provide insights into host pathogen interactions and transmission in leprosy (Sato et al., 2007). In the case of ML2177c, this gene encodes for uridine phosphorylase and shows significantly high expression during leprosy infection. This is predicted as druggable using fragment-hotspot-map analysis (Malhotra et al., 2017). The fragment hotspots contain a juxtaposition of charge and lipophilicity that are essential for effective ligand binding through both enthalpic and entropic contributions. The hotspot map software uses different molecular probes (toluene, aniline and phenol) to calculate affinity maps that provide a visual guide of the pocket (Radoux et al., 2016). **Figure 3** illustrates the recommended pathway to target prioritization in mycobacterial drug discovery.

Structure-Guided Virtual Screening

Structure-guided virtual screening is a cost-effective computational tool for preliminary screening of proteins that are potential drug targets with chemical libraries ranging



from small core fragments to large macrocyclic compounds in size and scaling from a few hundred molecules to billions (used in ultra-large-scale virtual screening campaigns). Since physical synthesis of drug molecules is not required, millions of virtual chemical entities can be swiftly docked into the active site of the protein structure/model and appropriate chemical scaffolds that fit with high scores and form relevant stabilizing interactions can be shortlisted for experimental validations. Virtual screening can be applied to novel drug discovery and also in drug repositioning experiments (screening with existing approved drugs to identify new target-protein interactions). A repurposing screen of LipU, a lipolytic protein that is conserved across mycobacterial species and noted to be essential for survival of *M. leprae*, revealed high docking scores for anti-viral drugs and anti-hypertensive (Kaur et al., 2019). Molecular docking software, such as Glide (Friesner et al., 2006), CCDC-GOLD (Jones et al., 1997), Autodock (Goodsell and Olson, 1990), Ledock (Wang et al., 2016), FlexX (Kramer et al., 1999), and SwissDock (Grosdidier et al., 2011) are used in virtual screening campaigns. Each algorithm has a unique scoring function to assess the fitness, number of stable interatomic interactions, and changes in energy landscape.

DISCUSSION AND CONCLUSION

Here, we have reviewed the tools and the advances made in protein structure prediction, modeling of genomes and impacts of amino acid replacements on protein structure and function. We have discussed these areas particularly focusing on the mycobacterial genomes, more specifically *M. leprae*. Given the paucity of information related to structural proteomic studies in leprosy, we discussed a multi-task protein modeling pipeline that enables proteome-scale template-based modeling of individual proteins encoded by various annotated genes in *M. leprae*. Homology-based structural and functional annotation of these protein models (Ochoa-Montañón et al., 2015; Skwark et al., 2019) using appropriate computational tools for modeling and druggability assessment can expedite characterization of the structural proteome of *M. leprae* and accelerate structure-guided novel drug discovery to combat nerve damage associated with leprosy.

Using the latest advancements in the field of protein structure bioinformatics, we describe our attempts to perform proteome scale modeling of mycobacterial genomes using in-house databases and pipelines. The modeled protein monomers or (homo/hetero) oligomers are subjected to multiple state-of-the-art validation scores. These models can be very helpful

and provide useful insights to understand protein structure and function, identify drug targets and unravel their functional roles in the pathogen. The structures of selected drug targets can also help in experimental design and prioritizing the protein targets for validation.

The emergence of drug resistance to the multidrug therapy is a major challenge in treating mycobacterial infections especially leprosy where structural features of drug-target interactions are poorly understood. To complement the computational findings, our group has employed cryoEM methods to understand the impact of mutations on the structure of catalase peroxidase in *M. tuberculosis* (Munir et al., 2019, 2021). Protein sequences and structures can be used to model the impacts of drug resistance mutation on protein structure and function. We have described various software available for predicting the impacts of mutations using protein sequence or structure or both. *In-silico* saturation mutagenesis experiments can guide the experimental design and help in saving the time and labor required to conduct laboratory experiments on animal models. Structure-based drug design (Blundell et al., 2002; Blundell and Patel, 2004) is a way forward and is a promising approach to design new drugs and treatments.

AUTHOR CONTRIBUTIONS

SV, SM, and MADGE conducted the review and written the manuscript. MM, PT, and TB reviewed the manuscript and provided necessary additions to the text. All authors contributed to the article and approved the submitted version.

FUNDING

SV, MADGE, and MM were supported by American Leprosy Missions Grant, United States of America (Grant No: G88726), SM was supported by the MRC DBT Grant, United Kingdom and India (RG78439), and TB was supported by the Wellcome Trust Programme Grant, United Kingdom (200814/Z/16/Z).

ACKNOWLEDGMENTS

Authors would like to thank the rest of the computational biology team at the Department of Biochemistry, University of Cambridge, United Kingdom, for their overarching support and guidance in the review process.

REFERENCES

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248
- AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics* 35, 4862–4865. doi: 10.1093/bioinformatics/btz422
- Anand, P., and Chandra, N. (2014). Characterizing the pocketome of *Mycobacterium tuberculosis* and application in rationalizing polypharmacological target selection. *Sci. Rep.* 4, 6356. doi: 10.1038/srep06356
- Andreeva, A., Kulesha, E., Gough, J., and Murzin, A. G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 48, D376–D382. doi: 10.1093/nar/gkz1064
- Baugh, L., Phan, I., Begley, D. W., Clifton, M. C., Armour, B., Dranow, D. M., et al. (2015). Increasing the structural coverage of tuberculosis drug targets. *Tuberculosis (Edinb)* 95, 142–148. doi: 10.1016/j.tube.2014.12.003
- 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.

- Bienert, S., Waterhouse, A., de Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., et al. (2017). The SWISS-MODEL repository—new features and functionality. *Nucleic Acids Res.* 45, D313–D319. doi: 10.1093/nar/gkw1132
- Binkowski, T. A., Naghibzadeh, S., and Liang, J. (2003). CASTp: computed Atlas of surface topography of proteins. *Nucleic Acids Res.* 31, 3352–3355. doi: 10.1093/nar/gkg512
- Blower, T. R., Williamson, B. H., Kerns, R. J., and Berger, J. M. (2016). Crystal structure and stability of gyrase–fluoroquinolone cleaved complexes from *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* 113, 1706–1713. doi: 10.1073/pnas.1525047113
- Blundell, T. L., Jhoti, H., and Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* 1, 45–54. doi: 10.1038/nrd706
- Blundell, T. L., and Patel, S. (2004). High-throughput X-ray crystallography for drug discovery. *Curr. Opin. Pharmacol.* 4, 490–496. doi: 10.1016/j.coph.2004.04.007
- Borah, K., Kearney, J.-L., Banerjee, R., Vats, P., Wu, H., Dahale, S., et al. (2020). GSMN-ML- a genome scale metabolic network reconstruction of the obligate human pathogen *Mycobacterium leprae*. *PLoS Negl. Trop. Dis.* 14:e0007871. doi: 10.1371/journal.pntd.0007871
- Cambau, E., Carthage, L., Chauffour, A., Ji, B., and Jarlier, V. (2006). Dihydropteroate synthase mutations in the FolP1 gene predict dapson resistance in relapsed cases of leprosy. *Clin. Infect. Dis.* 42, 238–241. doi: 10.1086/498506
- Cambau, E., Saunderson, P., Matsuoka, M., Cole, S., Kai, M., Suffys, P., et al. (2018). Antimicrobial resistance in leprosy: results of the first prospective open survey conducted by a WHO surveillance network for the period 2009–15. *Clin. Microbiol. Infect.* 24, 1305–1310.
- Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* 9:S6. doi: 10.1186/1471-2105-9-S2-S6
- Chen, C.-W., Lin, J., and Chu, Y.-W. (2013). iStable: off-the-shelf predictor integration for predicting protein stability changes. *BMC Bioinformatics* 14:S5. doi: 10.1186/1471-2105-14-S2-S5
- Chen, K.-H., Wang, T.-F., and Hu, Y.-J. (2019). Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. *BMC Bioinformatics* 20:308. doi: 10.1186/s12859-019-2907-1
- Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., et al. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66, 12–21. doi: 10.1107/S0907444909042073
- Cheng, J., Randall, A., and Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62, 1125–1132. doi: 10.1002/prot.20810
- Chiarelli, L. R., Mori, G., Orena, B. S., Esposito, M., Lane, T., de Jesus Lopes Ribeiro, A. L., et al. (2018). A multitarget approach to drug discovery inhibiting *Mycobacterium tuberculosis* PyrG and PanK. *Sci. Rep.* 8:3187. doi: 10.1038/s41598-018-21614-4
- Choi, Y., and Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31, 2745–2747. doi: 10.1093/bioinformatics/btv195
- Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., et al. (2001). Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011. doi: 10.1038/35059006
- Das, S., and Chakrabarti, S. (2021). Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Sci. Rep.* 11:1761. doi: 10.1038/s41598-020-80900-2
- Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., and Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* 25, 2537–2543. doi: 10.1093/bioinformatics/btp445
- Dehouck, Y., Kwasigroch, J. M., Rooman, M., and Gilis, D. (2013). BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Res.* 41, W333–W339. doi: 10.1093/nar/gkt450
- DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., et al. (2017). Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *mBio* 8:e02133–16. doi: 10.1128/mBio.02133-16
- Dhople, A. M. (2002). In vivo activity of epiroprim, a dihydrofolate reductase inhibitor, singly and in combination with dapson, against *Mycobacterium leprae*. *Int. J. Antimicrob. Agents* 19, 71–74. doi: 10.1016/S0924-8579(01)00470-8
- Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y., and Zhang, Y. (2017). DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J. Chem. Inf. Model.* 57, 1499–1510. doi: 10.1021/acs.jcim.7b00028
- Eramian, D., Shen, M., Devos, D., Melo, F., Sali, A., and Marti-Renom, M. A. (2006). A composite score for predicting errors in protein structure models. *Protein Sci.* 15, 1653–1666. doi: 10.1110/ps.062095806
- Fadlitha, V. B., Yamamoto, F., Idris, I., Dahlan, H., Sato, N., Aftitah, V. B., et al. (2019). The unique tropism of *Mycobacterium leprae* to the nasal epithelial cells can be explained by the mammalian cell entry protein 1A. *PLoS Negl. Trop. Dis.* 13:e0006704. doi: 10.1371/journal.pntd.0006704
- Fang, J. (2020). A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief Bioinform* 21, 1285–1292. doi: 10.1093/bib/bbz071
- Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *J. Med. Chem.* 49, 6177–6196. doi: 10.1021/jm0512560
- Goodsell, D. S., and Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins* 8, 195–202. doi: 10.1002/prot.340080302
- Graña, M., Haouz, A., Buschiazio, A., Miras, I., Wehenkel, A., Bondet, V., et al. (2007). The crystal structure of M. leprae ML2640c defines a large family of putative S-adenosylmethionine-dependent methyltransferases in mycobacteria. *Protein Sci.* 16, 1896–1904. doi: 10.1110/ps.072982707
- Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst.* 6, 116–124.e3. doi: 10.1016/j.cels.2017.11.003
- Grønning, A. G. B., Doktor, T. K., Larsen, S. J., Petersen, U. S. S., Holm, L. L., Bruun, G. H., et al. (2020). DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. *Nucleic Acids Res.* 48, 7099–7118. doi: 10.1093/nar/gkaa530
- Grosdidier, A., Zoete, V., and Michielin, O. (2011). SwissDock, a protein–small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* 39, W270–W277. doi: 10.1093/nar/gkr366
- Guo, H., Courbon, G. M., Bueler, S. A., Mai, J., Liu, J., and Rubinstein, J. L. (2021). Structure of mycobacterial ATP synthase bound to the tuberculosis drug bedaquiline. *Nature* 589, 143–147. doi: 10.1038/s41586-020-3004-3
- Gupta, E., Gupta, S. R. R., and Niraj, R. R. K. (2020). Identification of drug and vaccine target in *Mycobacterium leprae*: a reverse vaccinology approach. *Int. J. Pept. Res. Ther.* 26, 1313–1326. doi: 10.1007/s10989-019-09936-x
- Hatzopoulos, G. N., and Mueller-Dieckmann, J. (2010). Structure of translation initiation factor 1 from *Mycobacterium tuberculosis* and inferred binding to the 30S ribosomal subunit. *FEBS Lett.* 584, 1011–1015. doi: 10.1016/j.febslet.2010.01.051
- Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics* 16:S1. doi: 10.1186/1471-2164-16-S8-S1
- Hentschel, J., Burnside, C., Mignot, I., Leibundgut, M., Boehringer, D., and Ban, N. (2017). The complete structure of the *Mycobacterium smegmatis* 70S ribosome. *Cell Rep.* 20, 149–160. doi: 10.1016/j.celrep.2017.06.029
- Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., et al. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35, 128–135. doi: 10.1038/nbt.3769
- Hu, L., and Chan, K. C. C. (2017). Extracting coevolutionary features from protein sequences for predicting protein–protein interactions. *IEEE ACM Trans. Comput. Biol. Bioinform.* 14, 155–166. doi: 10.1109/TCBB.2016.2520923
- Huang, B., Fu, J., Guo, C., Wu, X., Lin, D., and Liao, X. (2016). (1)H, (15)N, (13)C resonance assignments for pyrazinoic acid binding domain of ribosomal protein S1 from *Mycobacterium tuberculosis*. *Biomol. NMR Assign.* 10, 321–324. doi: 10.1007/s12104-016-9692-9
- Huang, Y.-A., You, Z.-H., Chen, X., Chan, K., and Luo, X. (2016). Sequence-based prediction of protein–protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinformatics* 17:184. doi: 10.1186/s12859-016-1035-4

- Ji, B., Jamet, P., Perani, E. G., Sow, S., Lienhardt, C., Petinon, C., et al. (1996). Bactericidal activity of single dose of clarithromycin plus minocycline, with or without ofloxacin, against *Mycobacterium leprae* in patients. *Antimicrob. Agents Chemother.* 40, 2137–2141. doi: 10.1128/aac.40.9.2137
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748. doi: 10.1006/jmbi.1996.0897
- Kaur, G., Pandey, B., Kumar, A., Garewal, N., Grover, A., and Kaur, J. (2019). Drug targeted virtual screening and molecular dynamics of LipU protein of *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *J. Biomol. Struct. Dyn.* 37, 1254–1269. doi: 10.1080/07391102.2018.1454852
- Kaushal, P. S., Singh, P., Sharma, A., Muniyappa, K., and Vijayan, M. (2010). X-ray and molecular-dynamics studies on *Mycobacterium leprae* single-stranded DNA-binding protein and comparison with other eubacterial SSB structures. *Acta Crystallogr. D Biol. Crystallogr.* 66, 1048–1058. doi: 10.1107/S0907444910032208
- Kramer, B., Rarey, M., and Lengauer, T. (1999). Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. *Proteins* 37, 228–241. doi: 10.1002/(sici)1097-0134(19991101)37:2<228::aid-prot8>3.0.co;2-8
- Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. (2015). MAESTRO – multi agent stability prediction upon point mutations. *BMC Bioinformatics* 16:116. doi: 10.1186/s12859-015-0548-6
- Lechartier, B., and Cole, S. T. (2015). Mode of Action of Clofazimine and Combination Therapy with Benzothiazinones against *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* 59, 4457–4463. doi: 10.1128/AAC.00395-15
- Li, M., Simonetti, F. L., Goncarencu, A., and Panchenko, A. R. (2016). MutaBind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic Acids Res.* 44, W494–W501. doi: 10.1093/nar/gkw374
- Li, Y., Golding, G. B., and Ilie, L. (2020). DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* btaa750. doi: 10.1093/bioinformatics/btaa750 [Epub ahead of print].
- Lin, W., Mandal, S., Degen, D., Liu, Y., Ebright, Y. W., Li, S., et al. (2017). Structural basis of *Mycobacterium tuberculosis* transcription and transcription inhibition. *Mol. Cell* 66, 169–179.e8. doi: 10.1016/j.molcel.2017.03.001
- Liu, Y., Harrison, P. M., Kunin, V., and Gerstein, M. (2004). Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.* 5:R64. doi: 10.1186/gb-2004-5-9-r64
- Lockwood, D. N., and Saunderson, P. R. (2012). Nerve damage in leprosy: a continuing challenge to scientists, clinicians and service providers. *Int. Health* 4, 77–85. doi: 10.1016/j.inhe.2011.09.006
- Malhotra, S., Vedithi, S. C., and Blundell, T. L. (2017). Decoding the similarities and differences among mycobacterial species. *PLoS Negl. Trop. Dis.* 11:e0005883. doi: 10.1371/journal.pntd.0005883
- Mande, S. C., Mehra, V., Bloom, B. R., and Hol, W. G. (1996). Structure of the heat shock protein chaperonin-10 of *Mycobacterium leprae*. *Science* 271, 203–207. doi: 10.1126/science.271.5246.203
- Mangani, P. R., and Arif, M. A. (2006). Multidrug therapy in leprosy. *J. Indian Med. Assoc.* 104, 686–688.
- Mukherjee, S., and Zhang, Y. (2011). Protein–protein complex structure predictions by multimeric threading and template recombination. *Structure* 19, 955–966. doi: 10.1016/j.str.2011.04.006
- Munir, A., Kumar, N., Ramalingam, S. B., Tamilzhalagan, S., Shanmugam, S. K., Palaniappan, A. N., et al. (2019). Identification and characterization of genetic determinants of isoniazid and rifampicin resistance in *Mycobacterium tuberculosis* in Southern India. *Sci. Rep.* 9:10283. doi: 10.1038/s41598-019-46756-x
- Munir, A., Wilson, M. T., Hardwick, S. W., Chirgadze, D. Y., Worrall, J. A. R., Blundell, T. L., et al. (2021). Using cryo-EM to understand antimycobacterial resistance in the catalase-peroxidase (KatG) from *Mycobacterium tuberculosis*. *Structure* doi: 10.1016/j.str.2020.12.008 [Epub ahead of print].
- Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509
- Ochoa-Montaña, B., Mohan, N., and Blundell, T. L. (2015). CHOPIN: a web resource for the structural and functional proteome of *Mycobacterium tuberculosis*. *Database (Oxford)* 2015:bav026. doi: 10.1093/database/bav026
- Ogmen, U., Keskin, O., Aytuna, A. S., Nussinov, R., and Gursoy, A. (2005). PRISM: protein interactions by structural matching. *Nucleic Acids Res.* 33, W331–W336. doi: 10.1093/nar/gki585
- Pandurangan, A. P., Ochoa-Montaña, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.* 45, W229–W235. doi: 10.1093/nar/gkx439
- Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H.-J., et al. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* 11, 1–13. doi: 10.1038/s41467-020-19669-x
- Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., et al. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* 39, D465–D474. doi: 10.1093/nar/gkq1091
- Pires, D. E. V., and Ascher, D. B. (2017). mCSM-NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Res.* 45, W241–W246. doi: 10.1093/nar/gkx236
- Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014a). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* 42, W314–W319. doi: 10.1093/nar/gku411
- Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014b). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342. doi: 10.1093/bioinformatics/btt691
- Pires, D. E. V., Blundell, T. L., and Ascher, D. B. (2016). mCSM-lig: quantifying the effects of mutations on protein–small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.* 6:29575. doi: 10.1038/srep29575
- Potapov, V., Cohen, M., Inbar, Y., and Schreiber, G. (2010). Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. *BMC Bioinformatics* 11:374. doi: 10.1186/1471-2105-11-374
- Radoux, C. J., Olsson, T. S. G., Pitt, W. R., Groom, C. R., and Blundell, T. L. (2016). Identifying interactions that determine fragment binding at protein hotspots. *J. Med. Chem.* 59, 4314–4325. doi: 10.1021/acs.jmedchem.5b01980
- Rao, P. N., and Jain, S. (2013). Newer Management Options in Leprosy. *Indian J. Dermatol.* 58, 6–11. doi: 10.4103/0019-5154.105274
- Roe, S. M., Barlow, T., Brown, T., Oram, M., Keeley, A., Tsaneva, I. R., et al. (1998). Crystal structure of an octameric RuvA-holliday junction complex. *Mol. Cell* 2, 361–372. doi: 10.1016/s1097-2765(00)80280-4
- Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* 48, 77–84. doi: 10.1046/j.1365-2958.2003.03425.x
- Sato, N., Fujimura, T., Masuzawa, M., Yogi, Y., Matsuoka, M., Kanoh, M., et al. (2007). Recombinant *Mycobacterium leprae* protein associated with entry into mammalian cells of respiratory and skin components. *J. Dermatol. Sci.* 46, 101–110. doi: 10.1016/j.jdermsci.2007.01.006
- Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388. doi: 10.1093/nar/gki387
- Shen, M., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15, 2507–2524. doi: 10.1110/ps.062416606
- Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310, 243–257. doi: 10.1006/jmbi.2001.4762
- Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., et al. (2019). CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.* 47, D280–D284. doi: 10.1093/nar/gky1097
- Skwark, M. J., Torres, P. H. M., Copoiu, L., Bannerman, B., Floto, R. A., and Blundell, T. L. (2019). Mabellini: a genome-wide database for understanding the structural proteome and evaluating prospective antimicrobial targets of the emerging pathogen *Mycobacterium abscessus*. *Database* 2019, baz113. doi: 10.1093/database/baz113
- Soni, N., and Madhusudhan, M. S. (2017). Computational modeling of protein assemblies. *Curr. Opin. Struct. Biol.* 44, 179–189. doi: 10.1016/j.sbi.2017.04.006
- Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* 18:277. doi: 10.1186/s12859-017-1700-2

- The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi: 10.1093/nar/gkaa1100
- Torrisi, M., Pollastri, G., and Le, Q. (2020). Deep learning methods in protein structure prediction. *Comput. Struct. Biotechnol. J.* 18, 1301–1310. doi: 10.1016/j.csbj.2019.12.011
- Vedithi, S. C., Malhotra, S., Das, M., Daniel, S., Kishore, N., George, A., et al. (2018). Structural implications of mutations conferring rifampin resistance in *Mycobacterium leprae*. *Sci. Rep.* 8:5016. doi: 10.1038/s41598-018-23423-1
- Vedithi, S. C., Malhotra, S., Skwark, M. J., Munir, A., Acebrón-García-De-Eulate, M., Waman, V. P., et al. (2020). HARP: a database of structural impacts of systematic missense mutations in drug targets of *Mycobacterium leprae*. *Comput. Struct. Biotechnol. J.* 18, 3692–3704. doi: 10.1016/j.csbj.2020.11.013
- Wang, F., Langley, R., Gulten, G., Dover, L. G., Besra, G. S., Jacobs, W. R., et al. (2007). Mechanism of thioamide drug action against tuberculosis and leprosy. *J. Exp. Med.* 204, 73–78. doi: 10.1084/jem.20062100
- Wang, L., Wang, H.-F., Liu, S.-R., Yan, X., and Song, K.-J. (2019). Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-019-46369-4
- Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., et al. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* 18, 12964–12975. doi: 10.1039/C6CP01555G
- Webb, B., and Sali, A. (2016). Comparative Protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics* 54, 5.6.1–5.6.37. doi: 10.1002/cpbi.3
- Williams, D. L., Spring, L., Harris, E., Roche, P., and Gillis, T. P. (2000). Dihydropteroate synthase of *Mycobacterium leprae* and dapsone resistance. *Antimicrob. Agents Chemother.* 44, 1530–1537. doi: 10.1128/aac.44.6.1530-1537.2000
- Yamaguchi, T., Yokoyama, K., Nakajima, C., and Suzuki, Y. (2016). DC-159a shows inhibitory activity against DNA gyrases of *Mycobacterium leprae*. *PLoS Negl. Trop. Dis.* 10:e0005013. doi: 10.1371/journal.pntd.0005013
- Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D., and Altman, R. B. (2004). Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.* 14, 917–924. doi: 10.1101/gr.2050304
- Yu, J., Vavrusa, M., Andreani, J., Rey, J., Tufféry, P., and Guerois, R. (2016). InterEvDock: a docking server to predict the structure of protein–protein interactions using evolutionary information. *Nucleic Acids Res.* 44, W542–W549. doi: 10.1093/nar/gkw340
- Yue, P., Melamud, E., and Moulton, J. (2006). SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 7:166. doi: 10.1186/1471-2105-7-166
- Zhang, L., Yu, G., Xia, D., and Wang, J. (2019). Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 324, 10–19. doi: 10.1016/j.neucom.2018.02.097
- Zhou, G., Chen, M., Ju, C. J. T., Wang, Z., Jiang, J.-Y., and Wang, W. (2020). Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR Genom. Bioinform.* 2. doi: 10.1093/nargab/lqaa015
- Zhou, H., and Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* 101, 2043–2052. doi: 10.1016/j.bpj.2011.09.012

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

Copyright © 2021 Vedithi, Malhotra, Acebrón-García-de-Eulate, Matusevicius, Torres and Blundell. 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:

Agnel Praveen Joseph,
Science and Technology Facilities
Council, United Kingdom

Reviewed by:

Sayan Chatterjee,
University School of Biotechnology,
Guru Gobind Singh Indraprastha
University, India
Eliza Wyszko,
Institute of Bioorganic Chemistry
(PAS), Poland

*Correspondence:

Radhakrishnan Surendrakumar
surendrakumar@nmc.ac.in

Specialty section:

This article was submitted to
Molecular Diagnostics
and Therapeutics,
a section of the journal
Frontiers in Molecular Biosciences

Received: 04 December 2020

Accepted: 26 March 2021

Published: 11 May 2021

Citation:

Gobinath P, Packialakshmi P,
Vijayakumar K, Abdellattif MH,
Shahbaaz M, Idhayadhulla A and
Surendrakumar R (2021) Synthesis
and Cytotoxic Activity of Novel Indole
Derivatives and Their *in silico*
Screening on Spike Glycoprotein
of SARS-CoV-2.
Front. Mol. Biosci. 8:637989.
doi: 10.3389/fmolb.2021.637989

Synthesis and Cytotoxic Activity of Novel Indole Derivatives and Their *in silico* Screening on Spike Glycoprotein of SARS-CoV-2

Perumal Gobinath¹, Ponnusamy Packialakshmi¹, Kaliappillai Vijayakumar²,
Magda H. Abdellattif³, Mohd Shahbaaz^{4,5}, Akbar Idhayadhulla¹ and
Radhakrishnan Surendrakumar^{1*}

¹ PG & Research, Department of Chemistry, Nehru Memorial College (Affiliated Bharathidasan University), Puthanampatti, India, ² Department of Chemistry, M. Kumarasamy College of Engineering, Karur, India, ³ Department of Chemistry, College of Science, Deanship of Scientific Research, Taif University, Taif, Saudi Arabia, ⁴ South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa, ⁵ Laboratory of Computational Modeling of Drugs, South Ural State University, Chelyabinsk, Russia

This work investigated the interaction of indole with SARS-CoV-2. Indole is widely used as a medical material owing to its astounding biological activities. Indole and its derivatives belong to a significant category of heterocyclic compounds that have been used as a crucial component for several syntheses of medicine. A straightforward one-pot three-component synthesis of indole, coupled with Mannich base derivatives 1a–1j, was synthesized without a catalyst. The products were confirmed by IR, ¹H-NMR, ¹³C-NMR, mass spectra, and elemental analysis. The indole derivatives were tested for cytotoxic activity, using three cancer cell lines and normal cell lines of Human embryonic kidney cell (HEK293), liver cell (LO2), and lung cell (MRC5) by MTT assay using doxorubicin as the standard drug. The result of cytotoxicity indole compound 1c (HepG2, LC₅₀–0.9 μm, MCF–7, LC₅₀–0.55 μm, HeLa, LC₅₀–0.50 μm) was found to have high activity compared with other compounds used for the same purpose. The synthesized derivatives have revealed their safety by exhibiting significantly less cytotoxicity against the normal cell line (HEK-293), (LO2), and (MRC5) with IC₅₀ > 100 μg/ml. Besides, we report an *in silico* study with spike glycoprotein (SARS-CoV-2-S). The selective molecules of compound 1c exhibited the highest docking score

–2.808 (kcal/mol) compared to other compounds. This research work was successful in synthesizing a few compounds with potential as anticancer agents. Furthermore, we have tried to emphasize the anticipated role of indole scaffolds in designing and discovering the much-awaited anti-SARS CoV-2 therapy by exploring the research articles depicting indole moieties as targeting SARS CoV-2 coronavirus.

Keywords: indole, Mannich base, cytotoxic activity, COVID-19, spike protein

INTRODUCTION

Coronavirus has proved to be the most deadly of the 21st-century epidemics by being responsible for emergent communicable disorders. It first manifested its presence through the onset of dangerous pneumonia, started by the (SARS-CoV) infestation in 2003 (Rahman et al., 2020). In December 2019, many lung fever patients infected by a novel coronavirus were announced in Wuhan, China (Chan et al., 2020; Li et al., 2020; Zhu et al., 2020). The SARS-CoV-2 has been the cause of greater than 1.27 million deaths as of November 11, 2020 (Lu et al., 2020; Wu et al., 2020; Zhou et al., 2020). The acronym for coronavirus, namely, SARS-CoV-2, was assigned by the World Health Organization (WHO) on February 11, 2020 (Gorbalenya et al., 2020). SARS-CoV-2 has become a global health crisis involving around 212 countries (World Health Organization, 2020). Several drug mixtures are still being used.

However, the remedial outcome has been meager with secondary response (Cao et al., 2020). Adenosine triphosphate (ATP) analog was used as an antiviral drug to counter the effects of COVID-19, but more statistics are required to demonstrate its efficiency (Cohen, 2020; Holshue et al., 2020; Wang et al., 2020). On August 11, 2020, Russia became the first nation to approve a vaccine (sputnik V) to protect against infection by COVID-19 (Talha, 2020). The inherent RNA of coronaviruses and its structure information is described and discussed by other researchers (Hussain et al., 2005; Chen et al., 2020). In the biorhythms of coronaviruses, some functional and non-functional proteins are involved (Ramajayam et al., 2011; Ren et al., 2013). The emergence of drug-resistance for antiviral activity and defective antiviral drugs stimulates a great demand to develop a less toxic and more potent antiviral agent. In this regard, researchers have recently focused on naturally available indoles and their derivatives.

The inclusion of indole is the most significant structural modification in drug development, and it is labeled as one of the “privileged scaffolds” (Evans et al., 1988; deSa Alves et al., 2009; Welsch et al., 2010). The enlargement of a new technique for the pattern of C-N and C-C bonds that evade the pore functional group is tremendously significant in current organic chemistry (Ricci, 2008). Amino methylation is a crucial method for direct carbon-carbon and carbon-nitrogen bond-forming reactions (Hwang and Uang, 2002). Usually, amino methylation is done by the Mannich reaction using aldehyde as a methylene group source (Mannich and Krosche, 1912). Indole is perhaps the most ubiquitous motif in nature (Humphrey and Kuethe, 2006). Many natural and synthetic indole derivatives

have been in great demand in medical and pharmaceutical applications since they can bind with high affinity to many receptors (Sundberg, 1970, 1996; Lounasmaa and Tolvanen, 2000; Horton et al., 2003; Gu and Hamann, 2005; Somei and Yamada, 2005; Shiri, 2012). Previously reported natural products of indole derivatives are shown in **Figure 1** (Chen et al., 2019), and the biological activities of indole derivatives are offered in **Figure 2** (Kumari and Singh, 2019). Indole regulates numerous aspects of microorganism physiology, including reproductive structure formation, body stability, resistance to medication, biofilm formation, and virulence (Chadha and Silakari, 2017). Based on the above properties, we prepared new indole derivatives 1a–1j via the Mannich reaction. As the indole compounds have been rigorously involved in ailments including viral infections and cancer, there exists a profound scope of exploring these multiple nuclei to curb coronaviruses (Zhang et al., 2015). Here we demonstrated that the indole moiety potentially blocked the infectivity of SARS CoV-2 by targeting glycoproteins. They also potentially block the enzymatic activity of SARS CoV-2 and replication of coronavirus (Hattori et al., 2021). Therefore, through this, indole derivatives developed against SARS-CoV-2 epidemics using *in vitro* and *in silico* approaches may be of immense value at this hour of global emergency and in the future.

EXPERIMENTAL

General

All the chemicals were purchased from Merck. The melting point was determined using an open capillary tube, and it is uncorrected. The IR spectra were recorded in KBr on a Shimadzu 8201pc (4000–400 cm^{-1}). ^1H and ^{13}C -NMR spectra were recorded on Bruker Avance II NMR spectrometer 300 MHz with DMSO- d_6 as solvent using tetramethylsilane (TMS) as an internal standard. Mass spectra were recorded using Clarus SQ8 (Perkin Elmer), and the elemental analysis (C, H, and N) was performed on a Varian EL III instrument.

General Procedure for the Synthesis of Compounds 1a–1j

We compounded Furan-2-ylmethylenhydrazine (0.01 mol), indole (0.01 mol), and substituted aldehydes (0.01 mol) in ethanol solution to give a yellow solution with light brown color precipitate. The residue was recrystallized with ethanol. The obtained compound was purified by thin-layer chromatography (TLC). Hexane was used as eluting and solvent in TLC. All the synthesized compounds were separated by column chromatography.

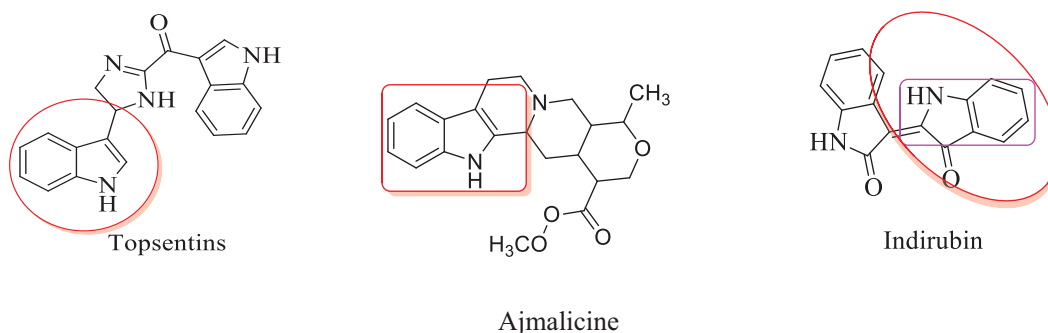


FIGURE 1 | Natural products of indole derivatives.

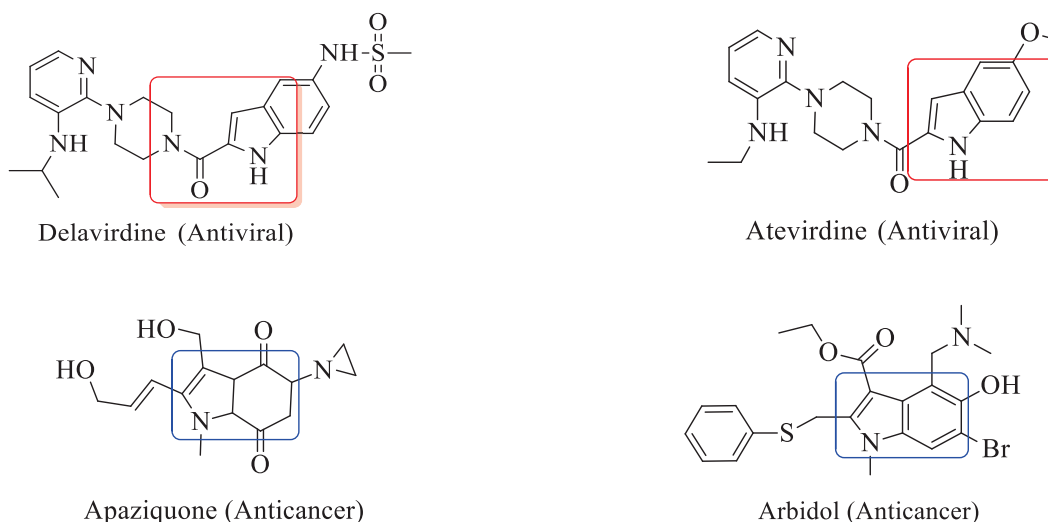


FIGURE 2 | Biological activities of indole derivatives.

(*E*)-1-((2furan-2ylmethylene)hydrazinyl)phenyl)methyl) 1H-indole (1a)

Light yellowish brown solid: mp 250°C; IR (KBr) (cm^{-1}) 3440 (NH str), 3080 (CH-str Ar-ring), 1623 (C = N), 1092 (N-CH-N). ^1H NMR (DMSO- d_6), δ (ppm) J (Hz): 8.41 (s, 1H, CH = N), 7.99–7.79 (d, 2H, indole), 7.82–6.70 (m, 3H, furan), 7.40–6.95 (m, 9H, Ar), 7.08 (s, 1H, NH), 7.08 (s, 1H, CH); ^{13}C NMR (DMSO- d_6) δ (ppm): 150.46, 145.73, 118.96, 113.43 (4C, Furyl ring), 143.52, 128.57, 127.78, 126.95, 126.90, 125.25 (6C, Ph ring), 136.25, 126.62, 125.23, 121.69, 120.09, 118.23, 111.57, 110.76 (8C, indole ring), 135.23 (1C, C = N), 40.59 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 315.14 (M^{++} , 20); Elemental analysis: *Anal.* $\text{C}_{20}\text{H}_{17}\text{N}_3\text{O}$: C, 76.20; H, 5.45; N, 13.35; Found C, 76.25; H, 5.55; N, 13.28.

1-((2furan-2ylmethylene)hydrazinyl)3-nitrophenyl)methyl) 1H-indole (1b)

Light brown solid: mp 258°C; IR (KBr) (cm^{-1}) 3435 (NH str), 3058 (CH-str Ar-ring), 1590 (NO_2^-), 1623

(C = N), 1094 (N-CH-N). ^1H NMR (DMSO- d_6), δ (ppm) J (Hz): 8.33–6.75 (m, 9H, Ar), 8.10 (s, 1H, CH = N), 7.95–7.82 (d, 2H, indole), 7.69–6.74 (m, 3H, furan), 7.09 (s, 1H, NH), 6.73 (s, 1H, CH); ^{13}C NMR (DMSO- d_6) δ (ppm): 150.62, 145.64, 118.96, 113.50 (4C, Furyl ring), 147.75, 135.85, 127.84, 126.69, 114.20, 114.12 (6C, Ph ring), 136.53, 126.62, 125.25, 121.62, 120.16, 118.25, 111.48, 110.70 (8C, indole ring), 135.27 (1C, C = N), 40.54 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 360.12 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{20}\text{H}_{16}\text{N}_4\text{O}_3$: C, 66.67; H, 4.49; N, 15.56; Found C, 66.62; H, 4.55; N, 15.48.

1-((2furan-2ylmethylene)hydrazinyl)1H-indole-1-yl)methyl)phenol (1c)

Brown solid: mp 272°C; IR (KBr) (cm^{-1}) 3585 (OH), 3449 (NH str), 3086 (CH-str Ar-ring), 1629 (C = N), 1091 (N-CH-N). ^1H NMR (DMSO- d_6), δ (ppm) J (Hz): 10.24 (s, 1H, OH), 8.52 (s, 1H, CH = N), 7.95–7.52 (d, 2H, indole), 7.68–6.97 (m, 3H, furan), 7.42–6.72 (m, 9H, Ar), 7.06 (s, 1H, NH), 6.74 (s, 1H, CH); ^{13}C NMR (DMSO- d_6) δ (ppm): 176.36

(1C, Ph-OH), 150.25, 145.24, 118.42, 113.47 (4C, Furyl ring), 142.18, 129.48, 129.50, 114.25, 114.22 (5C, Ph ring), 136.04, 127.89, 125.24, 121.67, 120.15, 118.29, 111.52, 110.76 (8C, indole ring), 135.29 (1C, C = N), 40.51 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 331.13 (M^+ , 20); Elemental analysis: *Anal.* $C_{20}H_{17}N_3O_2$: C, 72.50; H, 5.18; N, 12.69; Found C, 72.45; H, 5.24; N, 12.65.

1-((4-chlorophenyl)2-furan-2ylmethylene)hydrazinyl methyl)-1H-indole (1d)

Light brown solid: mp 260°C; IR (KBr) (cm^{-1}) 3442 (NH str), 3082 (CH-str Ar-ring), 1626 (C = N), 1094 (N-CH-N), 818 (C-Cl). 1H NMR (DMSO- d_6), δ (ppm) J (Hz): 8.11 (s, 1H, CH = N), 7.98–7.81 (d, 2H, indole), 7.84–6.74 (m, 3H, furan), 7.32–6.70 (m, 9H, Ar), 7.06 (s, 1H,

TABLE 1 | Physicochemical data of synthesized compounds (1a-1j).

Compounds	R	Product	Solvent	Yield (%)
1a			EtOH	89
1b			EtOH	86
1c			EtOH	82
1d			EtOH	95
1e			EtOH	88
1f			EtOH	85
1g			EtOH	80
1h			EtOH	78
1i			EtOH	67
1j	HCHO		EtOH	75

NH), 6.74 (s, 1H, CH); ^{13}C NMR (DMSO- d_6) δ (ppm): 150.31, 145.35, 118.85, 113.40 (4C, Furyl ring), 136.10, 131.67, 130.50, 129.8, 129.59, 129.57 (6C, Ph ring), 136.25, 127.78, 125.26, 121.59, 120.10, 118.20, 111.59, 110.73 (8C, indole ring), 135.21 (1C, C = N), 40.50 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 349.10 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{20}\text{H}_{16}\text{ClN}_3\text{O}$: C, 68.68; H, 4.62; N, 12.03; Found C, 68.65; H, 4.63; N, 12.05.

1-((2-furan-2-ylmethylene)hydrazinyl)4-methoxyphenyl methyl)1H-indole (1e)

Light yellowish brown solid; mp 275°C; IR (KBr) (cm^{-1}) 3444 (NH str), 3056 (CH-str Ar-ring), 2854 (OCH_3), 1618 (C = N), 1089 (N-CH-N). ^1H NMR (DMSO- d_6) δ (ppm) J (Hz): 8.22 (s, 1H, CH = N), 7.96–7.86 (d, 2H, indole), 7.62–6.80 (m, 3H, furan), 7.47–6.69 (m, 9H, Ar), 7.09 (s, 1H, NH), 6.75 (s, 1H, CH). 3.86 (s, 3H, OCH_3); ^{13}C NMR (DMSO- d_6) δ (ppm): 158.62 (1C, Ph, OCH_3), 150.24, 145.62, 118.94, 113.48 (4C, Furyl ring), 136.51, 126.61, 125.27, 121.60, 120.18, 118.28, 111.50, 110.72 (8C, indole ring), 135.83, 127.82, 126.67, 114.18, 114.09 (5C, Ph ring), 135.29 (1C, C = N), 55.74 (1C, OCH_3), 40.54 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 345.15 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{21}\text{H}_{19}\text{N}_3\text{O}_2$: C, 73.04; H, 5.55; N, 12.16; Found C, 70.59; H, 5.60; N, 12.30.

2-((2-furan-2-ylmethylene)hydrazinyl)(1H-indol-1-yl)methyl)phenol (1f)

Light brown solid; mp 265°C; IR (KBr) (cm^{-1}) 3447 (NH str), 3089 (CH-str Ar-ring), 2915 (C- OCH_3), 1628 (C = N), 1090 (N-CH-N). ^1H NMR (DMSO- d_6) δ (ppm) J (Hz): 9.82 (s, 1H, OH), 8.25 (s, 1H, CH = N), 8.22–6.83 (m, 9H, Ar), 7.97–7.81 (d, 2H, indole), 7.71–6.49 (m, 3H, furan), 7.19 (s, 1H, NH), 6.71 (s, 1H, CH); ^{13}C NMR (DMSO- d_6) δ (ppm): 148.90, 144.45, 119.08, 113.06 (4C, Furyl ring), 139.89 (1C, Ph, OH), 136.56, 126.79, 128.41, 121.07, 120.79, 119.82, 109.76, 100.24 (8C, indole ring), 131.01, 126.02, 120.42, 118.92, 116.05 (5C, Ph) 134.19 (1C, C = N), 40.49 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 331.37 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{20}\text{H}_{17}\text{N}_3\text{O}_2$: C, 72.50; H, 5.72; N, 12.69; Found C, 72.40; H, 5.51; N, 12.48.

4-((2-furan-2-ylmethylene)hydrazinyl)1H-indol-1-yl)methyl)-2-Methoxyphenol (1g)

Brown solid; mp 289°C; IR (KBr) (cm^{-1}) 3442 (NH str), 3082 (CH-str Ar-ring), 2910 (C- OCH_3), 1626 (C = N), 1094 (N-CH-N). ^1H NMR (DMSO- d_6) δ (ppm) J (Hz): 10.48 (s, 1H, OH), 8.93 (d, 1H, J = 6.1 Hz, Furyl ring), 8.16 (s, 1H, methylene), 7.92–7.80 (d, 2H, J = 7.6 Hz, indole), 7.87 (d, 1H, J = 6.7 Hz, Furyl ring), 7.20, 7.18, 6.84, 6.75 (m, 4H, indole ring), 7.09 (s, 1H, NH), 6.64 (t, 1H, Furyl ring), 6.65–6.74 (m, 2H, Ar-CH), 6.71 (s, 1H, CH), 3.85 (s, 3H, OCH_3) 3.50 (s, 1H, CH); ^{13}C NMR (DMSO- d_6) δ (ppm): 150.21, 145.38, 118.82, 113.41 (4C, Furyl ring), 147.52, 146.82, 142.30, 126.37, 120.18, 115.39, 56.09 (7C, Ph ring), 136.15, 127.79, 125.23, 121.49,

120.00, 118.28, 111.52, 110.71 (8C, indole ring), 135.22 (1C, C = N), 40.51 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 349.10 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{21}\text{H}_{19}\text{N}_3\text{O}_3$: C, 69.75; H, 5.31; N, 11.66; Found C, 69.69; H, 5.20; N, 11.53.

4-((2-furan-2-ylmethylene)hydrazinyl)(1H-indol-1-yl)methyl)-N,N-dimethylaniline (1h)

Light brownish yellow solid; mp 270°C; IR (KBr) (cm^{-1}) 3440 (NH str), 3085 (CH-str Ar-ring), 2946 (NH_2), 1620 (C = N), 1096 (N-CH-N). ^1H NMR (DMSO- d_6) δ (ppm) J (Hz): 8.29 (s, 1H, CH = N), 7.93–7.83 (d, J = 7.0, 2H, indole), 7.72–6.54 (m, 3H, furan), 7.46–6.68 (m, 9H, Ar), 7.13 (s, 1H, NH), 6.76 (s, 1H, CH), 3.80–3.12 (s, 3H, N- CH_3); ^{13}C NMR (DMSO- d_6) δ (ppm): 149.17, 128.13, 127.87, 127.80, 112.75, 112.72 (6C, Ph) 149.10, 144.43, 118.93, 112.67 (4C, Furyl ring), 136.57, 128.92, 128.51, 121.75, 120.77, 119.78, 109.57, 100.84, (8C, indole ring), 134.61 (1C, C = N), 41.29 (2C, N- CH_3) 40.58 (1C, CH); EI-MS (Relative intensity %): m/z 331.37 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{22}\text{H}_{22}\text{N}_4\text{O}$: C, 73.42; H, 6.20; N, 15.63; Found C, 73.22; H, 6.09; N, 15.43.

2-(furan-2-ylmethylene)hydrazinyl)-3,7-dimethylocta-2,6-dien-1-yl)-1H-indole (1i)

Light brown solid; mp 280°C; IR (KBr) (cm^{-1}) 3446 (NH str), 3080 (CH-str Ar-ring), 1620 (C = N), 1094 (N-CH-N). ^1H NMR (DMSO- d_6) δ (ppm) J (Hz): 8.30 (d, 1H, J = 6.2 Hz, furyl ring), 7.58–7.53 (d, 2H, J = 7.0, indole ring), 7.40, 7.31, 7.06, 6.58 (m, 4H, indole ring), 7.12 (s, 1H, NH), 6.90, 6.48 (m, 2H, Furyl ring), 6.68 (s, 1H, CH), 5.47, 5.30 (d, 2H, J = 7.7 Hz, citral), 2.00 (d, 2H, J = 6.3, citral), 2.04 (d, 2H, J = 6.0, citral), 1.82–1.71 (s, 9H, citral); ^{13}C NMR (DMSO- d_6) δ (ppm): 149.98, 144.38, 118.97, 112.68 (4C, Furyl ring), 136.52, 128.91, 127.83, 121.68, 120.75, 119.81, 109.65, 100.95 (8C, indole ring), 135.52, 132.02, 123.52, 118.76, 39.45, 26.41, 24.61, 18.63, 16.10 (9C, citral) 134.62 (1C, C = N), 40.60 (1C, N-CH-N); EI-MS (Relative intensity %): m/z 361.22 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{23}\text{H}_{27}\text{N}_3\text{O}$: C, 76.43; H, 7.54; N, 11.63; Found C, 76.66; H, 7.43; N, 11.42.

1-((2-furan-2-ylmethylene)hydrazinyl)methyl)-1H-indol (1j)

Brown solid; mp 285°C; IR (KBr) (cm^{-1}) 3442 (NH str), 3082 (CH-str Ar-ring), 1626 (C = N), 1092 (N-CH-N). ^1H NMR (DMSO- d_6) δ (ppm) J (Hz): 8.24 (s, 1H, CH = N), 7.78–6.59 (d, J = 6.2 Hz, 3H, Furan), 7.62–7.59 (m, 2H, indole), 7.47–6.45 (m, 9H, Ar), 7.16 (s, 1H, NH), 5.54 (d, J = 6.2 Hz, 2H, CH_2); ^{13}C NMR (DMSO- d_6) δ (ppm): 149.17, 145.02, 118.91, 112.62 (4C, Furyl ring), 136.49, 128.98, 127.89, 121.76, 120.72, 119.86, 109.67, 100.91 (8C, indole ring), 134.62 (1C, C = N), 40.59 (1C, CH_2); EI-MS (Relative intensity %): m/z 239.27 (M^+ , 20); Elemental analysis: *Anal.* $\text{C}_{14}\text{H}_{13}\text{N}_3\text{O}$: C, 70.29; H, 5.49; N, 17.57; Found C, 70.35; H, 5.78; N, 17.88.

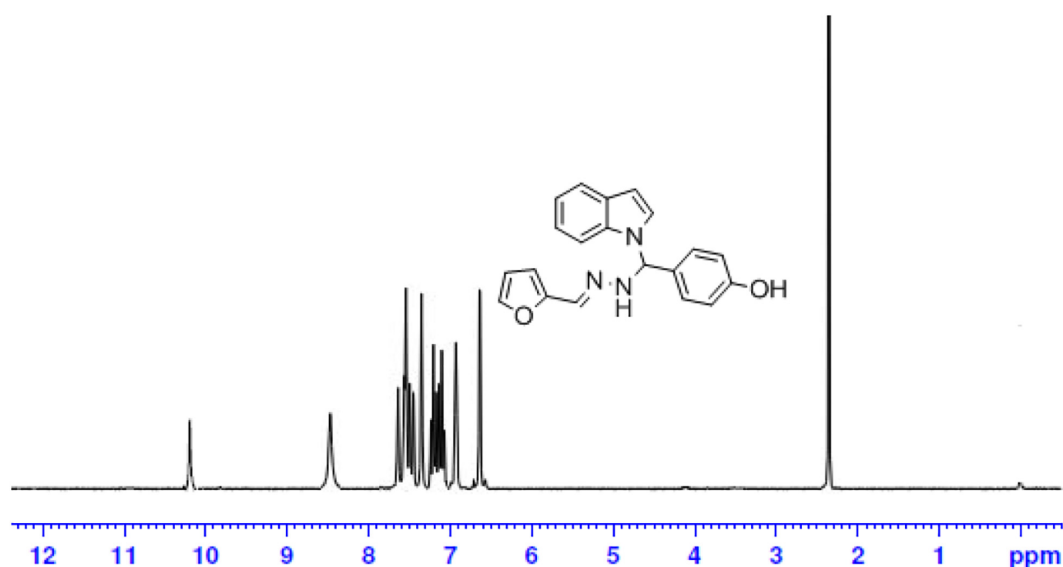


FIGURE 3 | ¹H-NMR spectrum of compound-1c.

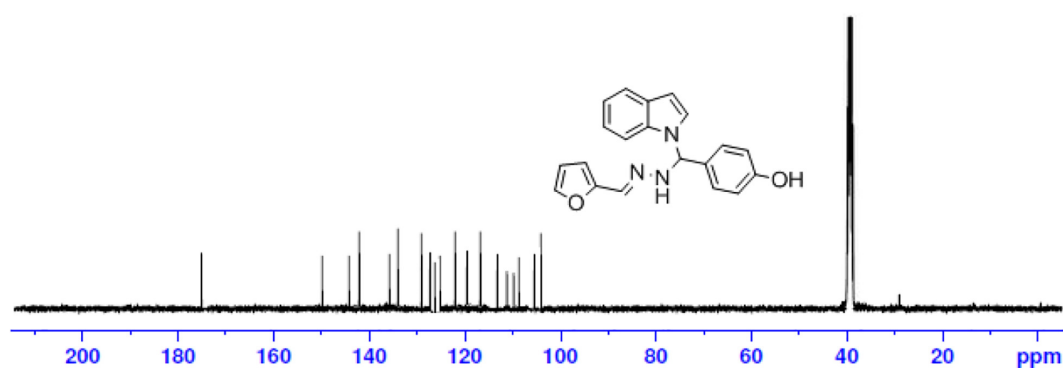
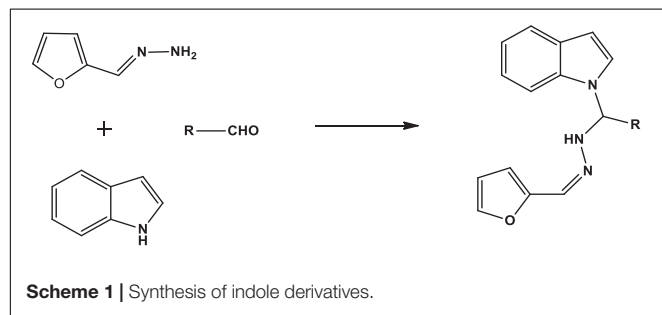


FIGURE 4 | ¹³C-NMR spectrum of compound-1c.



Scheme 1 | Synthesis of indole derivatives.

Biological Screening

Cytotoxic Activity

The cytotoxicity experiment was performed according to the United States NCI protocol, previously reported method. A detailed experimental procedure was given in **Supplementary Material** (Premnath et al., 2015).

Molecular Docking

Molecular docking was performed to confirm the molecular interaction with Covid-19 spike core protein to ensure the secondary biological mechanism based on the molecular pose on the binding moiety. The molecular structure of the selected ligand was drawn using Chem. Draw. Before it being considered for molecular interaction, it was 2D optimized by the energy minimization process. The 3D molecular protein crystal structure of spike glycoprotein of SARS-CoV-2 PDB ID 6WPT protein was downloaded. The protein structure was prepared using Schrodinger 12.4 software to remove water molecules and optimize the structure to become suitable to execute flexible docking. In protein preparation, hydrogen atoms were added to increase the hydrophilicity, and already existed co-crystal molecules, and missed loops were optimized (Boyd and Paull, 1995). The ligand preparation module optimized the ligand 3D structure of selected molecules to remove unwanted atomic orientation by molecular and

TABLE 2 | Cytotoxic activity of synthesized compounds (1a–1j).

Cpds	HepG2			MCF-7			HeLa		
	GI ₅₀ (μm)	TGI (μm)	LC ₅₀ (μm)	GI ₅₀ (μm)	TGI (μm)	LC ₅₀ (μm)	GI ₅₀ (μm)	TGI (μm)	LC ₅₀ (μm)
1a	43.2	78.3	> 100	–	–	> 100	–	–	100
1b	42.2	23.1	56.8	46.1	85.1	> 100	51.0	89.2	> 100
1c	36.3	65.3	0.9	–	–	0.55	41.3	87.2	0.50
1d	01.0	0.25	54.0	0.89	09.3	65.0	08.9	06.8	0.50
1e	15.9	38.2	44.8	24.4	59.3	66.3	34.2	72.1	> 100
1f	29.1	46.8	57.5	18.6	41.8	42.3	21.6	54.7	77.4
1g	48.0	61.3	59.3	34.0	67.4	30.5	37.9	51.9	58.9
1h	21.3	41.0	72.1	12.9	45.3	10.3	52.9	81.3	> 100
1i	19.3	28.3	> 100	45.6	56.0	59.3	49.0	49.3	55.0
1j	40.3	45.3	66.8	56.0	49.3	78.6	2.3	58.3	48.3
Doxorubicin (standard)	0.01	0.13	0.58	0.02	0.21	0.74	0.05	0.41	0.88

TABLE 3 | *In vitro* cytotoxicity of indole derivatives (1a–1j) on normal cells^a.

Compounds	MRC5	HEK-293	LO2
	IC ₅₀ (μm)	IC ₅₀ (μm)	IC ₅₀ (μm)
1a	76.36	70.06	67.48
1b	67.21	62.12	72.17
1c	86.66	81.14	87.10
1d	56.25	79.14	66.24
1e	72.76	57.09	56.01
1f	51.24	66.17	68.24
1g	62.61	58.24	70.54
1h	66.32	67.01	58.22
1i	79.41	77.44	66.70
1j	75.14	52.71	70.12

^aEach compound was tested in triplicate. All error bars represent mean ± SD from three independent experiments.

quantum mechanics. Molecular docking, with flexible SP followed by XP, was executed. The grid-based technique, evaluation, and minimization of grid approximation procedure were followed by Premnath et al. (2016) and Muthiah et al. (2020). The confirmation of the best interactive molecule with 6WPT protein was concluded based on the G score and number of hydrogen bonds and bonding efficiency and binding energy.

RESULTS AND DISCUSSION

Chemistry

The one-pot Mannich reactions of substituted benzaldehyde, indole, and Furan-2-lymethylenhydrazine were done by reflux for 2 h using ethanol, a solvent, without any catalyst. The obtained solid 1-((2furan-2ylmethylene)hydrazinyl)phenyl)methyl)1H-indole (1a) was washed with cooled water and recrystallized using ethanol. It was purified by TLC. Hexane was used as an eluting solvent in TLC. All the synthesized compounds were separated by column chromatography. A similar procedure was carried out to synthesize the other nine compounds (1b–1j) Physicochemical

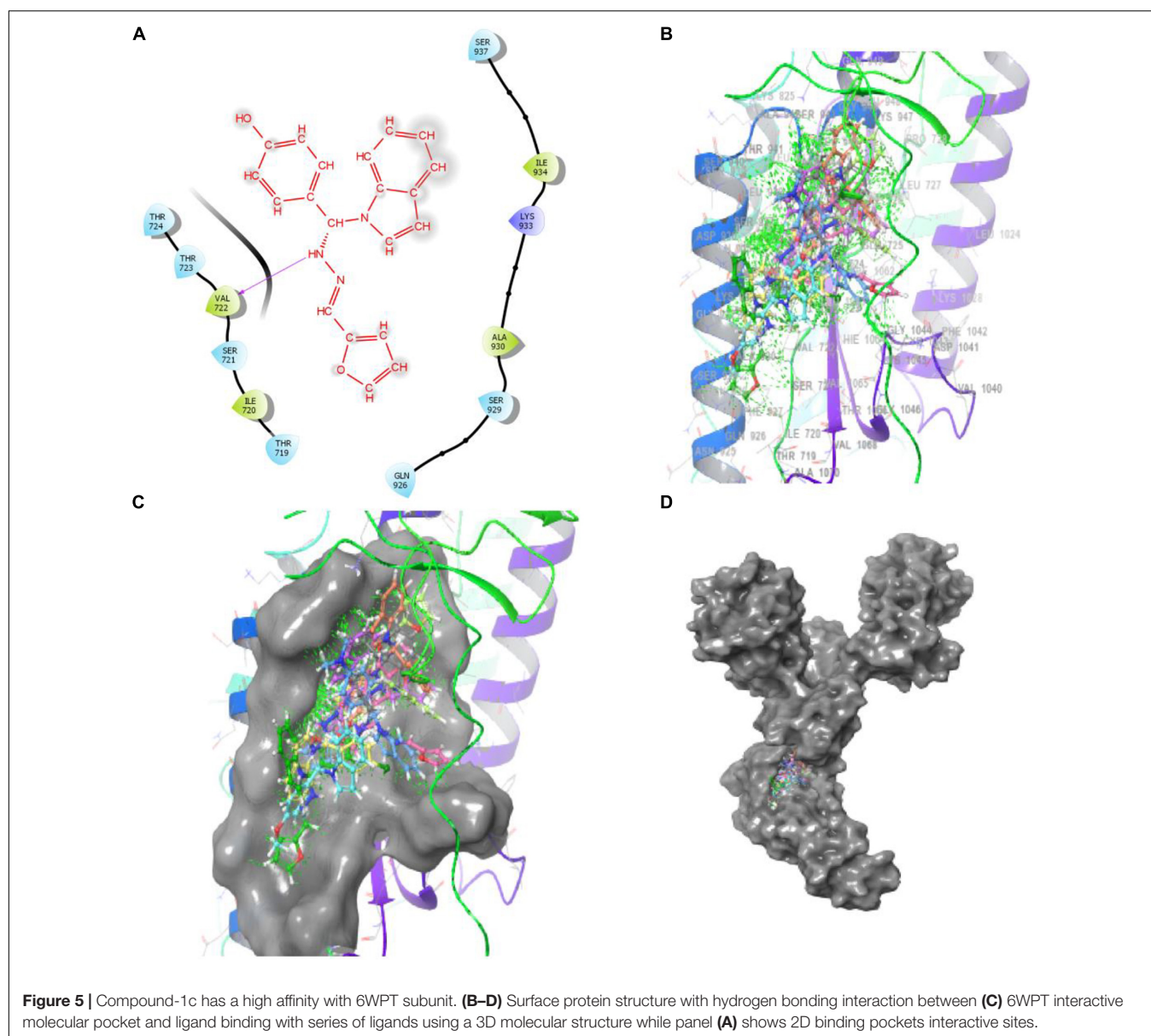
data of synthesized compounds (1a–1j) are given in Table 1. The Figure 3 indicates ¹H-NMR spectra of compound 1c, and Figure 4 displays ¹³C-NMR spectra of compound 1c. Scheme 1 represents the synthesis of compounds 1a–1j.

All the newly synthesized indole derivatives were characterized by FT-IR, which showed various functional groups. The ¹H-NMR spectra of compounds (1a–1j) indicate frequency observed at 7.16–7.07 and 6.79–5.54, corresponding to the NH-CH and CH-Ph protons. The ¹³C-NMR spectra exhibit the peak at 144.42–118.76 and 40.60–40.53, corresponding to the NH-CH and CH-Ph carbon, respectively.

Cytotoxic Activity

The newly prepared compounds 1a–1j are examined for their cytotoxic activity according to the United States NCI protocol, which was a previously reported method (Chadha and Silakari, 2017). The 50% growth inhibition (GI₅₀), tumor growth inhibition (TGI), and lethal concentration (LC₅₀) values were determined. The compounds 1c were a significant activity against (HepG2, LC50-0.9 μm, MCF-7, LC50-0.55 μm, HeLa, LC50-0.50 μm). Doxorubicin was used as a standard drug. None of the tested derivatives had shown significant activity toward the cancer cell lines. The compounds were also evaluated for their possible cytotoxicity in human embryonic kidney cells (HEK-293), lung cells (MRC-5), and liver cells (LO2) by employing MTT assay. The assay results suggested that these compounds did not significantly affect normal kidney cells' growth (As most of the compound's IC₅₀ values are > 100). Hence, these compounds revealed their safety for the normal cells, and the compound 1c can be taken as lead compounds for further development of more potent agents for HepG2 (Liver), MCF-7 (Breast), HeLa (Cervical) cancer cell lines. The results of cytotoxic screening of compounds (1a–1j) are shown in Table 2, and *in vitro* cytotoxicity of indole derivatives (1a–1j) on normal cells are shown in Table 3.

None of the tested derivatives had shown significant activity toward the cancer cell lines. The compounds were also evaluated for the possible cytotoxicity in human embryonic kidney cells (HEK-293), lung cells (MRC5), and liver cells (LO2) by

**TABLE 4** | Docking results of synthesized compounds.

Entry Name	Glide g score	Glide e model	Glide energy	XP H-Bond	Bonded Amino acid	Bond length Å
3 (1c)	−2.808	−37.395	−29.608	−1.33	VAL 722	2.18
7 (1g)	−2.715	−38.457	−27.458	−0.7	THR 724 ALA 944	1.91 2.08
5 (1e)	−2.174	−28.608	−24.17	−0.641	VAL 722	2.37
2 (1b)	−1.912	−35.298	−30.366	−0.7	LYS 947 ALA 944	2.51 1.80
1 (1a)	−1.636	−31.767	−26.491	−0.7	ALA 944	2.03
6 (1f)	−1.537	−32.323	−25.423	−0.7	THR 724	1.98
4 (1d)	−1.213	−35.839	−28.794	−0.37	SER 937	2.10
8 (1h)	−1.18	−30.149	−29.338	−1.29	THR 724	1.94
9 (1i)	−0.499	−31.625	−26.095	−1.56	THR 724	1.99

employing MTT assay. The assay results suggested that these compounds did not significantly affect the growth of normal cells (as most of the compounds $IC_{50} > 100$). Hence this compounds

revealed their safety for the normal cells and the compound 1c can be taken as lead compound for further development of more potential agent for HepG2 (Liver), MCF-7 (Breast), HeLa

(Cervical) cancer cell lines, and *in vitro* cytotoxicity of indole derivatives (1a–1j) on normal cells are shown in **Table 3**.

Molecular Docking

PAT binds 6WPT with strong affinity via computer docking studies

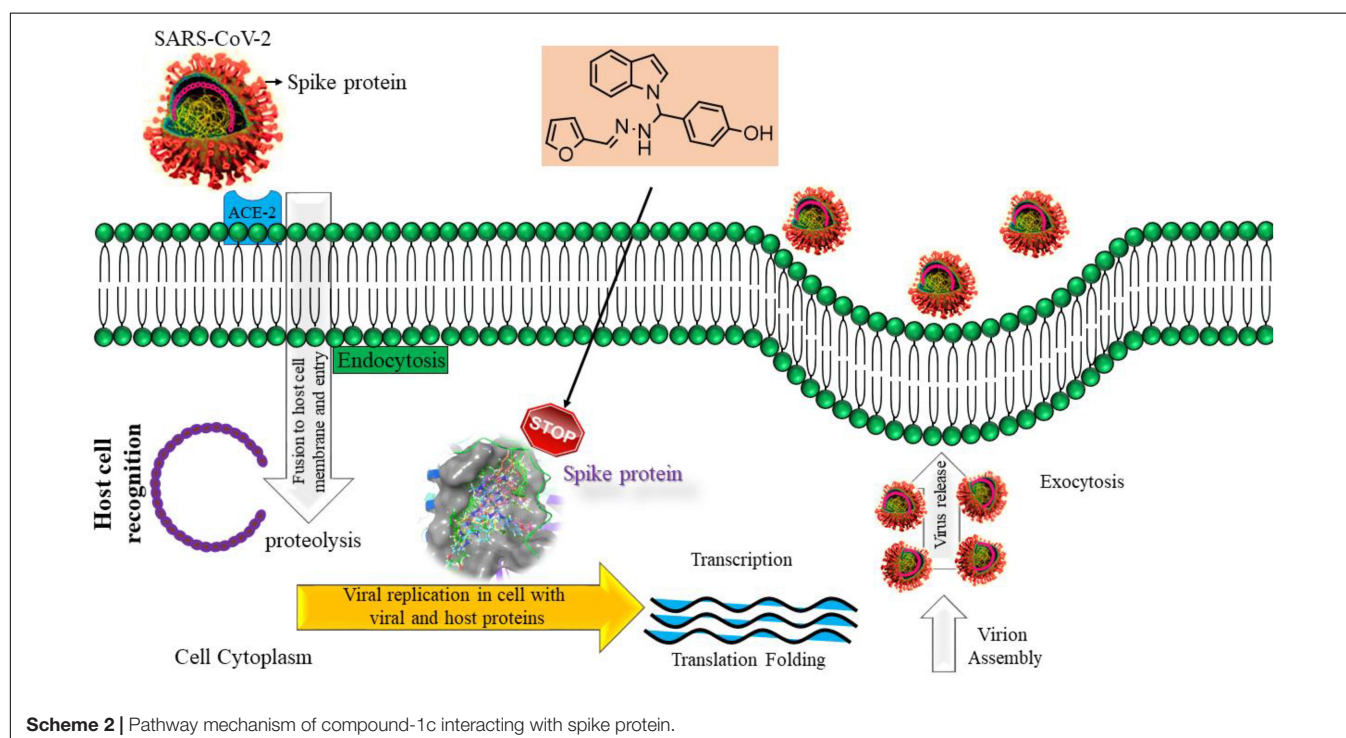
Bimolecular interaction studies were used to characterize the interaction between selected drug-like molecule and protein biomolecular binding sites. The protein interaction study was executed to forecast the interactive visualization modes and binding of small molecule and their respective protein receptors. An investigation of the interactive molecular complex of ligand series disclosed very informative and important connections between the drug like molecular series and the (6WPT) protein receptor. The two-dimensional and three-dimensional protein molecular structural images were perfectly visualized using Schrodinger integrative python software to analyze the molecular interaction between the selective ligand series and protein macromolecule (6WPT) (**Figure 5**). The overall optimized G score for binding ligands is predicted as -2.808 kcal/mol. This is taken to indicate an expected favorable reaction. Ligand 3 (1c) were perfectly interacted and formed close molecular interactions with amino acid residues on the predicted selective binding sites of VAL367, LEU368, PHE342, GLY339, GLY112, ARG55, LEU47, ASN343, ASP115, TYR32 of receptor protein (6WPT) (XXX) during different biochemical communications of hydrogen bonding, and hydrophobic interaction (Pinto et al., 2020). The binding score of (1c) to 6WPT was moderately burly with a predictable affinity of -2.808 kcal/mol. The docking analysis characterization of novel synthesized molecules are shown in **Table 4**. Further,

ligand series bonded with 6WPT through interactions with hydrogen bonding interaction and Pi-Pi interaction, Pi-Pi stacking interactive protein amino acids side chains of valine (Val), threonine (Thr), serine (Ser), alanine (Ala), and lysine (Lys) were analyzed and predicted important interactive bioactive binding site molecules. The molecular interaction analysis with selected small molecules of (1c) with 6WPT protein was much stronger than other series of selected ligands with a predictable affinity of -2.808 kcal/mol (**Figure 5**). The pathway mechanisms of spike protein interactions with highly active compound 1c are shown in **Scheme 2**. All the 2D structures of synthesized compounds are given in **Supplementary Materials**.

Structure Activity Relationship

A structure-activity relationship analysis (SAR) was performed to find the link between the chemical structure of a dynamic molecule and its cytotoxic activity. SAR analysis makes it possible to identify the chemical group/atom that plays a critical function in modulating the cytotoxic activity of compounds within the specific system. Using the cytotoxic activity results of the indole Mannich base derivatives, preliminary SARs could be evaluated. The data of the selected indole Mannich base derivatives (1a–1j) showed that compound 1c is the most effective (HepG2, LC_{50} -0.9 μ m, MCF-7, LC_{50} -0.55 μ m, HeLa, LC_{50} -0.50 μ m) control doxorubicin.

Due to the presence of an indole ring fused to a hydroxyl benzaldehyde, it was found that the compound acquires a high cytotoxic activity against cancer cell lines. This was due to the presence of electron releasing hydroxyl group on phenyl ring



attached with an indole skeleton. The rest of the compounds demonstrate feeble cytotoxic activity against all the tested cancer cell lines.

Moreover, from the docking results, it can be assumed that the docking score for indole derivatives (1a–1i) have an acceptable range except 1j compound along with essential interaction which can stabilize the compound in the active site of a protein. Compound 1j has no active site because of an absence of electron withdrawing /electron releasing group on it. From the results, the compound 1c has exhibited the highest docking score of -2.808 (Kcal/mol) compared to other compounds.

CONCLUSION

We have reported a facile, high-yielding, one-pot procedure for the synthesis of (1a–1j) via Mannich reaction using various kinds of protected aldehydes which was successfully employed and gave very high yields. Moreover, there were no requirements for dry solvents or protective gas atmospheres. All the newly synthesized compounds (1a–1j) were screened for *in vivo* cytotoxicity activities against Hep-G2 (Liver), HeLa (Cervical), and MCF-7 (Breast) cancer cell lines and normal cell lines in Human embryonic kidney cell (HEK293), liver cell (LO2), and lung cell (MRC5). Among the indole derivatives, compound 1c (HepG2, LC_{50} -0.9 μ m), (MCF-7, LC_{50} -0.55 μ m), and (HeLa, LC_{50} -0.50 μ m) was that the most active compound against the Doxorubicin standard. All other compounds were less active against the standard. The synthesized derivatives revealed a high safety level by exhibiting very low cytotoxicity against the normal cell line (HEK-293), (LO2), and (MRC5). Furthermore, we report *in silico* molecular docking studies against SARA-CoV-2 spike proteins and the biological characterization of the results reveal that compound 1c (-2.808 Kcal/mol) has the best multiple biological activities and can be used as a model for future derivatives based on the 1c molecular structure. It may identify the route to develop the best drug against Covid-19.

REFERENCES

- Boyd, M. R., and Paull, K. D. (1995). Some practical considerations and applications of the national cancer institute in vitro anticancer drug discovery screen. *Drug Dev. Res.* 34, 91–109. doi: 10.1002/ddr.430340203
- Cao, B., Wang, Y., Wen, D., Liu, W., Wang, J., Fan, G., et al. (2020). A trial of Lopinavir–Ritonavir in Adults hospitalized with severe Covid-19. *N. Engl. J. Med.* 382, 1787–1799.
- Chadha, N., and Silakari, O. (2017). Indoles as therapeutics of interest in medicinal chemistry: bird's eye view. *Eur. J. Med. Chem.* 134, 159–184.
- Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 395, 514–523. doi: 10.1016/S0140-6736(20)30154-9
- Chen, X. B., Xiong, S. L., Xie, Z. X., Wang, Y. C., and Liu, W. (2019). Three-component one-pot synthesis of highly functionalized bis-indole derivatives. *ACS Omega* 4, 11832–11837. doi: 10.1021/acsomega.9b01159
- Chen, Y., Liu, Q., and Guo, D. (2020). Emerging coronaviruses: genome structure, replication, and pathogenesis. *J. Med. Virol.* 92, 418–423. doi: 10.1002/jmv.25681

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

PG: Organic compounds preparation; PP: Preparation of synthetic compound and chemical data analysis; KV: Manuscript editing; MA: Validation; MS: Molecular docking analysis; AI: All kinds of spectral analysis; RS: Investigation and writing original draft preparation through the contributions of all authors. All the authors contributed to the article and approved the submitted version.

FUNDING

MA thankfully acknowledges the Taif University researcher supporting project Number TURSP/91, Taif University, Taif, Saudi Arabia.

ACKNOWLEDGMENTS

We wish to thank Nehru Memorial College (Affiliated Bharathidasan University), Puthanampatti, India, for taking GC-MS spectra and providing necessary facilities. We wish to thank Tamil Nadu Government for providing DST-FIST fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.637989/full#supplementary-material>

- Cohen, J. (2020). *Science* 368, 14–16. doi: 10.1126/science.abb0659
- deSa Alves, F. R., Barreiro, E. J., and Fraga, C. A. (2009). From nature to drug discovery: the indole scaffold as a 'privileged structure'. *Mini Rev. Med. Chem.* 9, 782–793. doi: 10.2174/138955709788452649
- Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., and Whitter, W. L. (1988). Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* 31, 2235–2246. doi: 10.1021/jm00120a002
- Gorbalenya, A. E., Baker, S. C., Baric, R. S., Groot, R. J., Drosten, C., Gulyaeva, A. A., et al. (2020). Severe acute respiratory syndrome-related coronavirus: the species and its viruses – a statement of the coronavirus study group. *bioRxiv [preprint]* doi: 10.1101/2020.02.07.937862
- Gu, W., and Hamann, M. T. (2005). Indole alkaloid marine natural products: an established source of cancer drug leads with considerable promise for the control of parasitic, neurological and other diseases. *Life Sci.* 78, 442–453. doi: 10.1016/j.lfs.2005.09.007
- Hattori, S., Higashi-Kuwata, N., Hayashi, H., Allu, S. R., Raghavaiah, J., Bulut, H., et al. (2021). A small molecule compound with an indole moiety inhibits the main protease of SARS-CoV-2 and blocks virus replication. *Nat. Commun.* 12:668.

- Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., et al. (2020). First case of 2019 novel coronavirus in the United States. *N. Engl. J. Med.* 382, 929–936. doi: 10.1056/NEJMoa2001191
- Horton, D. A., Bourne, G., and Smythe, T. (2003). The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev.* 103, 893–930. doi: 10.1021/cr020033s
- Humphrey, G. R., and Kueth, T. J. (2006). Practical methodologies for the synthesis of indoles. *Chem. Rev.* 106, 2875–2911. doi: 10.1021/cr0505270
- Hussain, S., Pan, J., Chen, Y., Yang, Y., Xu, J., Peng, Y., et al. (2005). Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus. *J. Virol.* 79, 5288–5295. doi: 10.1128/JVI.79.9.5288-5295.2005
- Hwang, D. R., and Uang, B. J. (2002). A modified Mannich-type reaction catalyzed by VO(acac)(2). *Org. Lett.* 4, 463–466. doi: 10.1021/ol017229j
- Kumari, A., and Singh, R. K. (2019). Medicinal chemistry of indole derivatives: current to future therapeutic perspectives. *Bioorganic Chem.* 89:103021. doi: 10.1016/j.bioorg.2019.103021
- Li, Q. M., Xuhua, G., Peng, W., Xiaoye, W., Lei, Z., Yeqing, T., et al. (2020). Early transmission dynamics in wuhan, China, of novel coronavirus-infected Pneumonia. *N. Engl. J. Med.* 382, 1199–1207. doi: 10.1056/NEJMoa2001316
- Lounasmaa, M., and Tolvanen, A. (2000). Simple indole alkaloids and those with a non-rearranged monoterpenoid unit. *Nat. Prod. Rep.* 17, 175–191. doi: 10.1039/A809402K
- Lu, R., Zhao, X., Li, J., and Niu, P. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395, 565–574. doi: 10.1016/S0140-6736(20)30251-8
- Mannich, C., and Krosche, W. (1912). Uebere in Kondensations produktaus Formaldehyd, Ammoniakund Antipyrin. *Arch. Pharm.* 250:647. doi: 10.1002/ardp.19122500151
- Muthiah, I., Karthikeyan, R., Dhanaraj, P., and Vallinayagam, S. (2020). In silico structure prediction, molecular docking and dynamic simulation studies on G Protein-Coupled Receptor 116: a novel insight into breast cancer therapy. *J. Biomol. Structure Dynamics [Online ahead of print]* doi: 10.1080/07391102.2020.1783365
- Pinto, D., Park, Y. J., Beltramello, M., Walls, A. C., Tortorici, M. A., Bianchi, S., et al. (2020). Structural and functional analysis of a potent sarbecovirus neutralizing antibody. *BioRxiv [preprint]* doi: 10.1101/2020.04.07.023903
- Premnath, D., Enoch, I. V. M. V., Selvalumar, P. M., Indiraleka, M., and Vennila, J. J. (2016). Design, synthesis, spectral analysis, in vitro anticancer evaluation and molecular docking studies of some fluorescent 4-Amino-2, 3-Dimethyl-1-Phenyl-3-Pyrazolin-5-One, ampyrone derivatives. *Interdiscip. Sci.: Computational Life Sci.* 9, 130–139.
- Premnath, D., Selvakumar, P. M., Ravichandiran, P., Tamil Selvan, G., Indiraleka, M., and Jannet Vennila, J. (2015). Synthesis and spectroscopic characterization of fluorescent 4-amino anti pyrine analogues: molecular docking and invitro cytotoxicity studies. *Spectrochimica Acta Part A: Mol. Biomol. Spectroscopy* 153, 118–123.
- Rahman, N., Basharat, Z., Yousuf, M., Castaldo, G., Rastrelli, L., and Khan, H. (2020). Virtual screening of natural products against type II transmembrane serine protease (TMPRSS2), the priming agent of coronavirus 2 (SARS-CoV-2). *Molecules* 25:2271. doi: 10.3390/molecules2510227
- Ramajayam, R., Tan, K. P., and Liang, P. H. (2011). Recent development of 3C and 3CL protease inhibitors for anti-coronavirus and anti-picornavirus drug discovery. *Biochem. Soc. Trans.* 39, 1371–1375. doi: 10.1042/BST0391371
- Ren, Z., Yan, L., and Zhang, N. (2013). The newly emerged SARS-Like coronavirus HCoV-EMC also has an “Achilles’ heel”: current effective inhibitor targeting a 3C-like protease. *Protein Cell* 4, 248–250. doi: 10.1007/s13238-013-2841-3
- Ricci, A. (2008). *Amino Group Chemistry, From Synthesis to the Life Sciences*. Weinheim: Wiley-VCH.
- Shiri, M. (2012). Indoles in multicomponent processes (MCPs). *Chem. Rev.* 112, 3508–3549. doi: 10.1021/cr2003954
- Somei, M., and Yamada, F. (2005). Simple indole alkaloids and those with a non-rearranged monoterpenoid unit. *Nat. Prod. Rep.* 22, 73–103. doi: 10.1039/b316241a
- Sundberg, R. J. (1970). *The Chemistry of Indoles*. New York, NY: Academic Press.
- Sundberg, R. J. (1996). Indoles, academic press, San Diego. *J. Med. Chem.* 39:5286. doi: 10.1021/jm960712t
- Talha, K. B. (2020). *The Russian vaccine for COVID19*, News, sep-4, 2020, 8.
- Wang, M., Cao, R., and Zhang, L. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* 30, 269–271. doi: 10.1038/s41422-020-0282-0
- Welsch, M. E., Snyder, S. A., and Stockwell, B. R. (2010). Privileged scaffolds for library Design and drug discovery. *Curr. Opin. Chem. Biol.* 14, 347–361. doi: 10.1016/j.cbpa.2010.02.018
- World Health Organization [WHO] (2020). *WHO Director-General’s opening remarks at the media briefing on COVID-19-11 March 2020*. Geneva: World Health Organization.
- Wu, F., Zhao, S., and Yu, B. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi: 10.1038/s41586-020-2008-3
- Zhang, M.-Z., Chen, Q., and Yang, G.-F. (2015). A review on recent developments of indole-containing antiviral agents. *Eur. J. Med. Chem.* 89, 421–441. doi: 10.1016/j.ejmech.2014.10.065
- Zhou, P., Yang, X. L., and Wang, X. G. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273.
- Zhu, N., Zhang, D., Wenling, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi: 10.1056/NEJMoa2001017

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

Copyright © 2021 Gobinath, Packialakshmi, Vijayakumar, Abdellattif, Shahbaaz, Idhayadhulla and Surendrakumar. 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.



Cryo-EM Map-Based Model Validation Using the False Discovery Rate Approach

Mateusz Olek^{1,2} and Agnel Praveen Joseph^{3*}

¹Department of Chemistry, University of York, York, United Kingdom, ²Electron BioImaging Center, Rutherford Appleton Laboratory, Didcot, United Kingdom, ³Scientific Computing Department, Science and Technology Facilities Council, Research Complex at Harwell, Didcot, United Kingdom

OPEN ACCESS

Edited by:

Giulia Palermo,
University of California, Riverside,
United States

Reviewed by:

Pavel Afonine,
Lawrence Berkeley National
Laboratory, United States
Carlos Oscar Sanchez Sorzano

*Correspondence:

Agnel Praveen Joseph
agnel-praveen.joseph@stfc.ac.uk

Specialty section:

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

Received: 12 January 2021

Accepted: 26 April 2021

Published: 18 May 2021

Citation:

Olek M and Joseph AP (2021) Cryo-EM Map-Based Model Validation Using the False Discovery Rate Approach.
Front. Mol. Biosci. 8:652530.
doi: 10.3389/fmolb.2021.652530

Significant technological developments and increasing scientific interest in cryogenic electron microscopy (cryo-EM) has resulted in a rapid increase in the amount of data generated by these experiments and the derived atomic models. Robust measures for the validation of 3D reconstructions and atomic models are essential for appropriate interpretation of the data. The resolution of data and availability of software tools that work across a range of resolutions often limit the quality of derived models. Hence, the final atomic model is often incomplete or contains regions where atomic positions are less reliable or incorrectly built. Extensive manual pruning and local adjustments or rebuilding are usually required to address these issues. The presented research introduces a software tool for the validation of the backbone trace of atomic models built in the cryo-EM density maps. In this study, we use the false discovery rate analysis, which can be used to segregate molecular signals from the background. Each atomic position in the model can be associated with an FDR backbone validation score, which can be used to identify potential mistraced residues. We demonstrate that the proposed validation score is complementary to existing validation metrics and is useful especially in cases where the model is built in the maps having varying local resolution. We also discuss the application of the score for automated pruning of atomic models built *ab-initio* during the iterative model building process in Buccaneer. We have implemented this score in the CCP-EM software suite.

Keywords: cryo-EM, model validation, FDR map, CCP-EM, automated model building

INTRODUCTION

Improvements in cryo-EM data collection and processing techniques in recent years have enabled structure determination at near-atomic resolutions (Subramaniam, 2019). For structure interpretation, a number of tools for *ab-initio* model building have been developed and used in recent years (Hoh et al., 2020; Terwilliger et al., 2020; Pfab et al., 2021; Lawson et al., 2021). Despite the resolution revolution, the majority of maps (92%) deposited in the EM Data Bank (<https://www.ebi.ac.uk/pdbe/emdb/>) are at resolutions worse than 3 Å, and the average resolution of maps this year is around 5 Å (https://www.ebi.ac.uk/pdbe/emdb/statistics_sp_res.html/). Moreover, some local areas in the cryo-EM map can be poorly resolved. These issues may result in some parts of the derived atomic model being incorrectly built or traced into background noise. Model validation tools that are based on the analysis of stereochemical properties of the atomic model, such as MolProbity

(Williams et al., 2018), CaBLAM (Prisant et al., 2020), or Ramachandran plots, detect potential issues with the geometry of the model. The users can inspect the possible incorrect regions of the model and attempt to fix these in interactive visualization tools like Coot (Emsley et al., 2010).

Another set of validation tools evaluate the agreement of the atomic model with the cryo-EM map. Some of these scores can estimate the agreement of each residue against the area of the map covered by the residue. The agreement is either quantified as Manders' overlap coefficient in SMOC (Joseph et al., 2016), real space cross-correlation coefficient in PHENIX local CCC (Afonine et al., 2018), or a score of atomic resolvability in MapQ (Pintilie 2020). One of the observations from the recent model challenge is that the absolute values of some of these metrics are sensitive to the map resolution (Lawson et al., 2021). One reason is the underlying sensitivity of the metric toward differences in the shape of map distributions at different resolutions. Another reason is the fact that the synthetic map calculation from the model may not be optimal to represent experimental data at different resolutions.

The recently introduced FSC-Q score allows us to assess the local agreement of a model with the cryo-EM density map, and is normalized to account for local resolution variation (Ramírez-Aportela et al., 2021). The map-model local Fourier shell correlation (FSC) is normalized with respect to the local FSC obtained from the halfmaps. The FSC-Q score is calculated as the difference between these two and the values fluctuates around 0. A threshold of ± 0.5 is recommended to detect poorly fitted atoms. Although the FSC-Q calculation is not directly affected by the B-factor values used for map sharpening, the mask applied can have an effect in the local FSC calculation.

MapQ scores atoms in the residues by comparing the distance-dependent map value fall-off against a Gaussian-like reference derived from a map of apoferritin resolved at 1.54 Å and an associated well-fitted atomic model. The Q-score is calculated as a correlation between the map values and the reference Gaussian. Values close to 1 indicate that the atom is well resolved (Pintilie, 2020).

Metrics such as the atom-inclusion score (Lagerstedt et al., 2013), implemented as part of EMDB validation analysis, identify atoms in the model that are outside a selected map contour. The score is hence very sensitive to the choice of map contour, which is often subjective. Also, in cases where the local resolution varies across the map, a single contour may not be optimal to cover the entire molecular volume without including the background noise.

The resolution of cryo-EM maps may vary as a result of molecular flexibility, partial occupancy, non-uniform particle orientation, and other factors associated with the reconstruction process. Often, the resolution is better in the core and it gradually worsens toward the edges or other flexible parts of the molecular assembly. The statistical analysis used in the false discovery rate (FDR) approach allows associating confidence in distinguishing molecular signals from the background and detecting weak features in the map based on the statistical significance estimate. The FDR calculation (Beckers et al., 2019) generates confidence maps with values at each voxel reflecting the fraction of voxels expected to contain molecular

signals at this threshold (the voxel value). The 1% FDR threshold (confidence map threshold of 0.99) was demonstrated to reliably discriminate voxels associated with the molecular volume from the background noise over a wide resolution range, including maps at near-atomic resolutions to 6.8 Å and the subtomogram averages in the resolution range 7–90 Å.

In this study, we present a tool for validating the backbone trace of an atomic model by estimating the confidence that the backbone atoms are in the molecular volume rather than the background. Each residue in the model are assigned scores based on the confidence map calculated using the FDR approach. We demonstrate the utility of the approach to detect mistraced residues, using datasets from the EMDB model challenges 2015/16 and 2019, and compare it against other metrics used in the field for estimating local fit to maps.

This procedure is also useful for pruning mistraced regions of the model generated by *ab-initio* modeling tools like Buccaneer (Hoh et al., 2020). Especially at areas of the map with resolution worse than 3.5 Å, it is not uncommon that the chain may be mistraced into the background. Also, Buccaneer often traces a few polypeptide fragments in the background areas with noisy features or artifacts from map reconstruction and postprocessing. These fragments are not connected with the main chains of the model and usually are only of a few residues long. Currently, there is no automated tool to locate and remove mistraced residues. Where possible, the pruned models can then be extended with one of the automated model building tools or rebuilt in an interactive tool like Coot.

Initial results indicate that our approach is effective in detecting mistraced regions of the model and for automated pruning of models as part of Buccaneer. The FDR backbone validation score assesses whether the backbone coordinates are within the molecular volume and is complementary to existing validation tools that either assess the model quality or evaluate agreement with the map. The described tool is implemented and available as a part of the CCP-EM (Burnley et al., 2017) software suite.

METHODS

The FDR backbone validation method assesses positions of the atoms in the input model based on the confidence map derived using the FDR approach (Beckers et al., 2019). For the confidence map calculation, a processed/sharpened but unmasked map is preferred. Masked maps that exclude the majority of solvent background are not useful for confidence map calculations. The procedure estimates the background noise distribution from four density cubes placed outside of the particle volume in the x, y, and z central axes by default. Each voxel of the map is then compared against the background estimate to detect significant deviations and a *p*-value is associated to quantify significance. To account for the number of voxels and their dependencies, the *p*-values are further adjusted using false discovery rate (Benjamini and Yekutieli, 2001). Each voxel is assigned with an FDR-adjusted significance score between 0 and 1, 0 refers to noise only and 1 to a clear molecular signal. A score of 0.99 indicates that a maximum

of 1% of voxels (1% FDR) is expected to be background noise, beyond this threshold.

We use the following steps to calculate the FDR backbone score:

- (1) The minimal input for the FDR-validation is the pdb or cif/mmCIF format file of the atomic model and the confidence map calculated using the FDR control approach. The confidence map can be calculated using the “confidence map” implementation in the CCP-EM software suite or using a standalone installation from the source (<https://git.embl.de/mbeckers/FDRthresholding>).
- (2) The input model coordinates are extracted and mapped onto the confidence map grid by associating the voxel(s) around the atomic coordinate (within 1 Å).
- (3) Each atom of the model is then associated with the corresponding map value from the confidence map. In the default mode, the FDR backbone score of each residue is calculated as an average of map values at the coordinates of the C-alpha, C, and N atoms. We use this approach primarily to detect mistraced residues based on the positions of backbone coordinates in the map. We exclude the backbone carbonyl oxygen as they are often associated with weak map information at resolutions worse than 3 Å. This approach can be used to detect misplacement of the side chains as well, although missing map data at the ends of acidic and highly flexible side chains can lead to false detections. For nucleic acids, the average score is calculated based on the C1', C2', C3', C4', C5', O3', O4', O5', and P atoms positions. For ligands and waters, all of the atoms are taken into consideration. The users can also choose an optional validation mode based solely on the Ca positions of the residues and C1' for nucleic acids. This mode is useful with models with only Ca atoms, usually built in low-resolution cryo-EM maps.
- (4) Additionally, this tool offers an option to prune the atomic model, which can be used to automatically remove the residues with a score lower than 0.9 as well as the preceding and following residues. A model pruned this way can be used in the next stages of the iterative model building procedures, where the missing segments can be extended or rebuilt. This is useful when dealing with the *ab-initio* models from the automated model building tools. In some cases, particularly when building in areas of the map with a local resolution worse than 3.5 Å, parts of the chains can be traced into the background.
- (5) As an output, we provide a CSV format file containing the list of residues with the associated FDR backbone scores. The models after pruning will have the low scoring residues removed. They are saved in the selected folder with the original model names with a suffix “pruned” added depending on the mode used. We also provide an attribute file that can be used to associate the FDR score for each residue in the atomic model in UCSF Chimera (Pettersen et al., 2004). The model can then be colored using the FDR score attribute to identify areas with low scores.

The FDR backbone validation tool is written in Python 3. To handle the I/O model files in pdb and cif/mmCIF format the GEMMI package (Wojdyr, 2017) is used. The map files are processed with the mrcfile python package (Palmer, 2016). The tool also requires NumPy (tested with v1.16.2 (NumPy v1.16 Manual' 2019)). The GUI implementation with CCP-EM software suite was done using PyQt ('PyQt 4.9.4 Reference Guide' 2011).

In this study, we compare the FDR backbone validation approach against other metrics for estimating local fit to maps. For a fair comparison, the other metrics were also calculated only on the backbone atoms of the models. The map deposited in EMDB as a “primary” map was used for the analysis, the FSC-Q score calculation also requires the half-maps.

The Q-score values for the backbone atoms were calculated using the MapQ plugin (v1.6) for UCSF Chimera, at the resolution reported for the deposited primary map.

The FSC-Q score was calculated using the tool (validate fsc-q) integrated in the Scipion v3.0.7-Eugenius. The FSC-Q value for the backbone is calculated as an average for the C, N, and Ca atoms.

The SMOC and SCCC scores were calculated using the score_smoc.py script available from TEMPy1 in CCP-EM v1.5, for the minimal backbone of the models (C, N, and Ca atoms). The script was run with the “-distance” mode option which uses distance from the atoms for identifying voxel-covered. SMOC estimates the Manders' overlap coefficient while SCCC calculates the cross correlation coefficient.

The PHENIX map/model CCC (v1.18.2) scores were calculated on the models obtained from 10 cycles of atomic B-factor refinement with REFMAC5 (Murshudov et al., 2011) (using the keyword option “refi bonly”). This was done to ensure that the atomic B-factors are refined as PHENIX uses the atomic B-factors as part of the map calculation from the atomic model.

The box size of the input map was trimmed wherever possible to improve the speed of the computations. The FDR-validation requires a sharpened but unmasked map, with the background features present in order to estimate the noise distribution.

RESULTS

To demonstrate the application of our approach, we used the following examples, the majority of which are models submitted to the EMDB model challenges for target maps resolved at a range of resolutions. In each case, we compare the FDR backbone scores against other metrics that estimate local fit to map. Using a set of residues detected as “mistraced” by the FDR backbone score, we assess agreement with other scores and also highlight cases where there is a disagreement. We use the reference model from the model challenge to compare the backbone conformation and fit to map. Please note that the reference model does not always have the best fit to map for all residues in the model, and often several of the models submitted to the challenge have a better fit (Lawson et al., 2021). In some cases, there are obvious backbone misfits in the reference model, as discussed below. In such cases, we also compare the model of interest against other models reported with

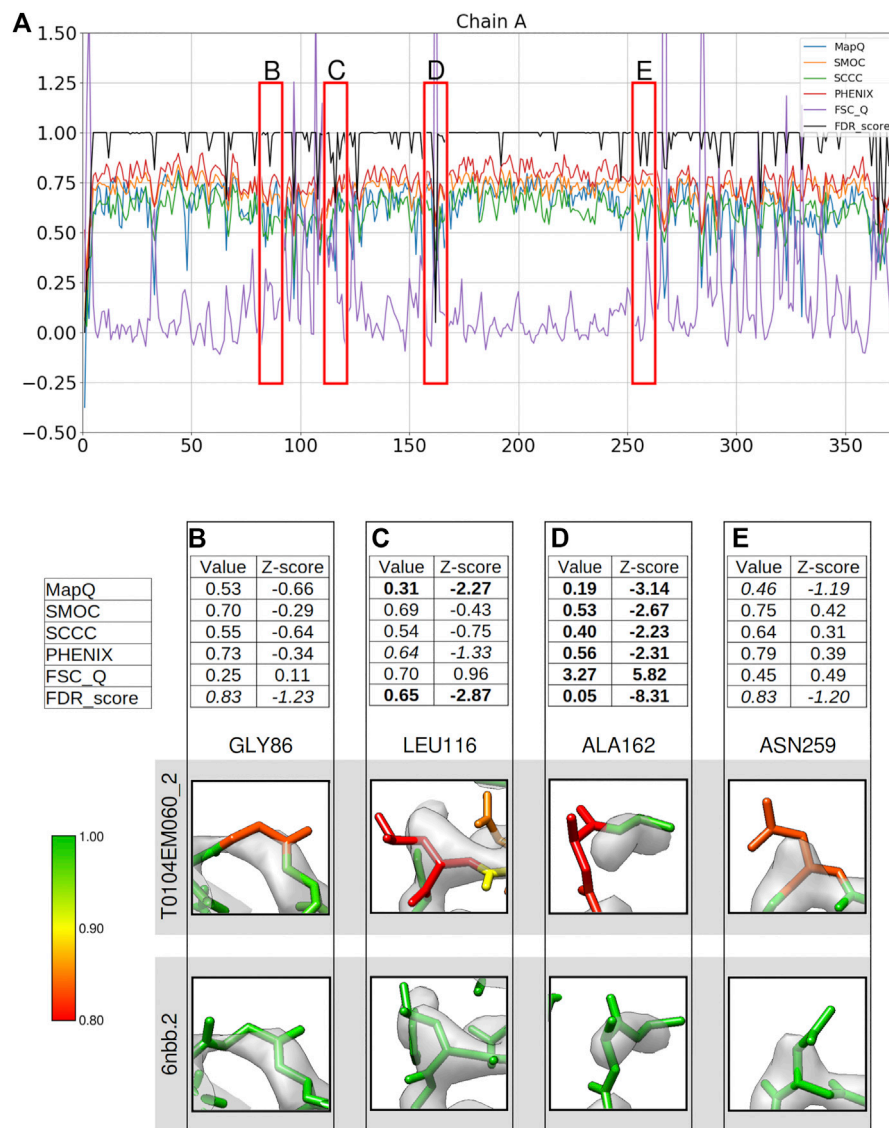


FIGURE 1 | Comparison of local assessment metrics for the atomic model of alcohol dehydrogenase (color bar shows the correspondence with the FDR scores assigned, residues in red have FDR scores around 0.8 or worse, yellow around 0.9, and green around 1.0). The metrics are calculated only for the backbone atoms. **(A)** Per-residue plot of MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone scores for the chain A of the atomic model T0104EM060_2 from the EMD model challenge, the red boxes highlight the residues selected for detailed analysis in the panels below. For Gly 86 **(B)**, Leu 116 **(C)**, Ala 162 **(D)**, and Asn 259 **(E)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-0406, gray) displayed at the recommended contour level and rendered in UCSF Chimera; and the residue fit in map as modeled in the reference (PDB ID: 6nbb.2).

a higher rank higher in the model challenge based on a number of validation metrics (<https://model-compare.emdataresource.org/>).

Alcohol Dehydrogenase (2.9 Å, Target T0104)

We computed per-residue backbone scores based on different metrics for the chain A of model T0104EM060_2 submitted to the Model Challenge 2019 for the target alcohol dehydrogenase map (EMD-0406) resolved at 2.9 Å resolution (Herzik et al. 2019; **Figure 1A**). We checked residues either associated with lower

confidence scores (0.95 or lower) or where the scores disagree in detecting a mistrace, and compared against the reference model used in the model challenge (PDB ID: 6nbb). The reference structure has 10 models representing local conformational variability. We chose the second model (6nbb.2) for our analysis as it has a relatively better fit with the map when inspected in UCSF Chimera, for cases we discuss in **Figures 1, 2**, especially at the N-terminus (**Figure 2B**). The metrics used for comparison includes Q-score, SMOC, SCCC, PHENIX local CCC, and FSC-Q, calculated only for the backbone atoms (see Methods). The residues highlighted in red boxes in **Figure 1A**

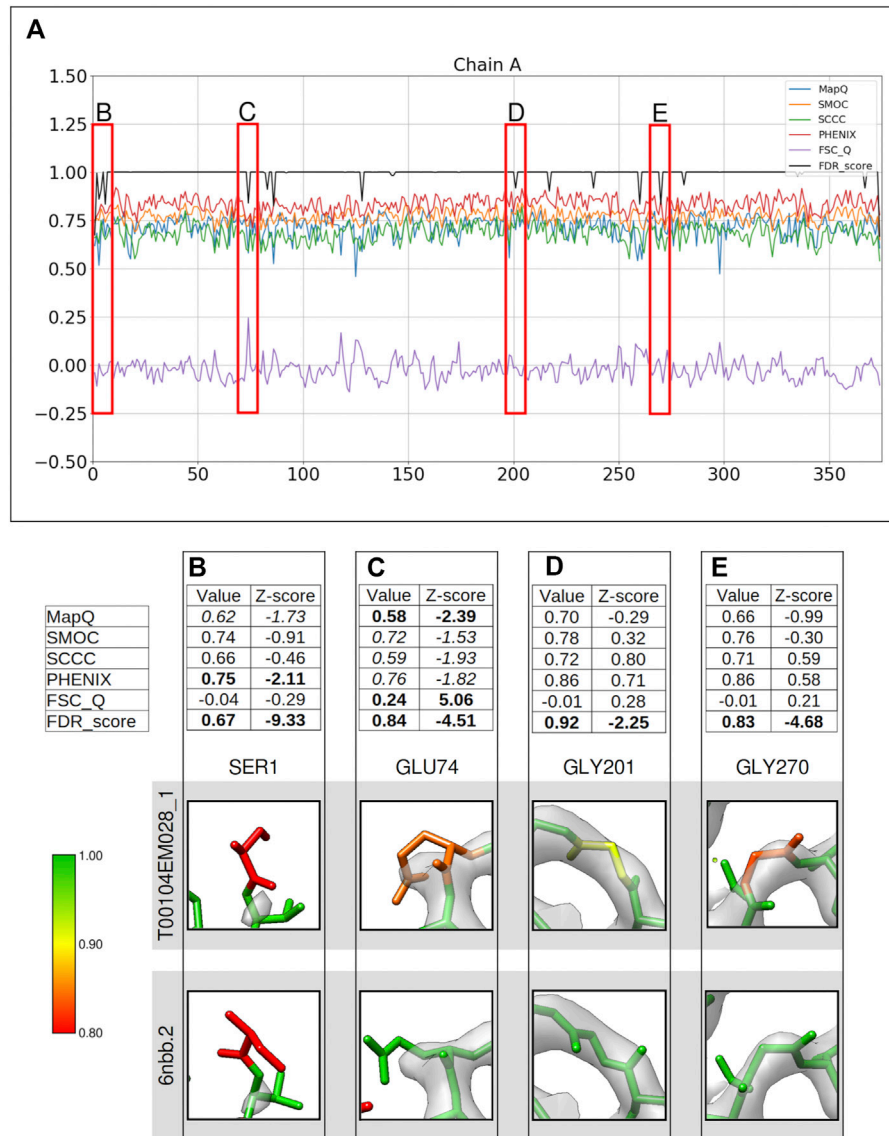


FIGURE 2 | Comparison of metrics for the atomic model of alcohol dehydrogenase (color bar shows the correspondence with the FDR scores assigned, residues in red have FDR scores around 0.8 or worse, yellow around 0.9, and green around 1.0). The metrics are calculated only for the backbone atoms. **(A)** Per-residue plot of the scores from MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the chain A of the atomic model T0104EM028_1 from the EMDB model; the red boxes highlight the residues selected for detailed analysis in the panels below. For Ser1 **(B)**, Glu 74 **(C)**, Gly 201 **(D)**, and Gly 270 **(E)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-0406, grey) displayed at the recommended contour level and rendered in UCSF Chimera; and the residue fit in map as modeled in the reference (PDB ID: 6nbb.2).

indicate some of the regions where the scores differ. We provide a detailed analysis of these residues. **Figures 1B–E** show a table with the values of each metric and corresponding Z-scores, along with a snapshot of the residue colored by the FDR backbone score. Also, the corresponding view of the residue from the reference model 6nbb.2 ((Herzik et al., 2019), model 2) is provided. The models are overlaid with the deposited cryoEM density map EMD-0406 (Herzik et al., 2019) and rendered at the author-recommended contour level 0.02 (0.6 σ).

Compared to a few other models submitted to the model challenge, the model T0104EM060_2 is ranked lower by the

validation metrics used in the challenge (https://model-compare.emdataresource.org/2019/cgi-bin/em_multimer_results.cgi?target_map=T0104emd_0406). Plot of per-residue backbone scores for the chain A (**Figure 1A**), also shows that many residues in this model are associated with lower scores (drops in the plot). We investigated a few residues including cases where the metrics disagree. Gly86 is highlighted as a potential mistrace by the FDR backbone score of 0.83 (**Figure 1B**). The residue in the reference model has a high FDR score (0.991) and the backbone shows better fit to map with a different conformation involving a shift and differences in backbone dihedrals. Z-scores computed for

different metrics reflect that none of the other scores identify this mistrace with any significance (absolute value of Z-scores < 1). Note that the Z-scores for FDR backbone assessment are less reliable, especially because the majority of the residues often have a score of 1.0 and the distribution is not close to normal. We recommend using the absolute values of this score to detect potential mistraces.

Figure 1C shows Leu116 associated with a low FDR backbone score of 0.65. It can be seen that the backbone C α and C are out of the map at the recommended contour level. In comparison, the reference model shows a better fit of backbone atoms. The mistrace is also detected by the MapQ score (0.31, Z-score -2.27), while PHENIX-CCC (0.63, Z-score -1.33) and FSC-Q (0.70, Z-score 0.96) have lower scores but associated with relatively low significance (Z-scores of -1.33 and 0.96, respectively).

Another residue Ala162 is located at a relatively disordered or low-resolution area of the map (**Figure 1D**), and the backbone is partly out of the map contour when compared to the reference model. All the scores identify the mistrace with significance (absolute Z-scores > 2.0), and the residue is associated with a low FDR backbone score of 0.05. Even though the map at the recommended contour level does not fully support the backbone, the reference model shows a better fit and has an FDR backbone score of 0.978. This reflects that the FDR backbone score detects voxels covering molecular volume even in the low resolution areas of the map.

Figure 1E shows Asn259 associated with an FDR backbone score of 0.83 with part of the backbone outside the contoured map. The reference model shows a better fitted backbone conformation (**Figure 1E**). MapQ also points to the potential mistrace in the submitted model with a Q-score of 0.46 although with a less significant Z-score of -1.19 . The other metrics fail to identify this issue with the backbone fit. Hence, in comparison to other metrics tested in this study, the FDR backbone score detects cases of mistrace where one or more backbone atoms are displaced into background noise.

Figure 2 presents a similar analysis of the model T0104EM028_1 submitted to the same target map. Ser1 at the N-terminus of chain A is associated with an FDR backbone score of 0.67, clearly indicating a potential mistrace. Ser1 is associated with a disordered area of the map with no prominent map information at the recommended contour (**Figure 2B**). PHENIX_CC (0.75, Z-score -2.11) and MapQ (0.62, Z-score -1.73) scores also suggest poor agreement with data. The map trace is more obvious at a lower contour level (**Supplementary Figure S1**), and the terminal N atom is outside the map even at this level. Hence, there is less confidence associated with the backbone atom positions and this is also highlighted by PHENIX_CC and MapQ scores.

Figure 2C highlights Glu74 with both backbone and side chain atoms out of the recommended contour. The residue, as modeled in the reference, shows better fit with backbone atoms (and most of the side chain) inside the recommended contour. The lack of map information for the end of side chain is a common trait observed in cryo-EM maps for negatively charged side chains. FSC-Q and MapQ indicate a backbone mistrace with

Z-score values less than -2.0 (>2.0 in case of the FSC-Q score). The other metrics also highlight this, although with a relatively lower significance (Z-score < -1.5).

Gly201 is associated with an FDR validation score of 0.92. Other scores do not seem to indicate mistrace with any significance (all Z-scores were between -1 and 1) (**Figure 2D**). This residue has a different backbone conformation in the reference model and is associated with an FDR backbone score of 1.0. The backbone has a better fit in the reference with all atoms except carbonyl oxygen inside the recommended contour. Another case where only the FDR-validation score detects a mistrace of backbone is Gly270, where the reference model shows a better fit with the map with a slight shift in atom positions (**Figure 2E**). The backbone residue shifted outside of the map density is presented in **Figure 2E**. These cases highlight that the FDR backbone score can work in complementarity to the scores that quantify agreement with the map.

The structure of alcohol dehydrogenase has zinc ions bound but the ions are not modeled in all of the structures submitted to the model challenge. **Supplementary Figure S2A** presents a comparison of two models submitted (Model Challenge IDs: T0104EM010_1 on the left panel and T0104EM028_1 on the center) where the zinc atoms are modeled, along with the reference model (PDB ID: 6nbb) on the right. In the model T0104EM010_1, the zinc atom is highlighted as a potential misfit based on our approach, and no obvious map data can be seen at this position. It can be seen that the ligands in both the model T0104EM028_1 and the reference structure are placed in a position justified by map density and supported by the higher FDR backbone scores. It is worth mentioning that many of the automated model building software do not support ligand fitting, and therefore this is often done interactively. The presented validation technique can be useful for validating the modeled ligands in cryo-EM maps.

T20s Proteasome (2.8 Å)

Another set of models used for the evaluation of the FDR backbone validation approach were chosen from those submitted to the model challenge for the target map of the T20s proteasome (EMD-6287), resolved at 2.8 Å resolution. **Figure 3A** presents the comparison of scores from difference metrics obtained for the chain L of the model T0002EM133_1. Again, the areas where the scores disagree were inspected closely.

Several residues in this model are associated with lower score values as evaluated by different metrics (**Figures 3B–E**). In this case, the Z-scores are less meaningful as the distribution of scores is likely to deviate significantly from normal because of the presence of several low-scoring residues (outliers). Therefore, we considered a less stringent absolute Z-score cutoff of 1.0 to associate significance to the scores. Again, as the FDR scores do not follow a normal distribution (often many residues have a score of 1.0 and a few scoring lower), the Z-scores are less useful. We recommend using the absolute FDR scores to detect potential mistraces.

Val14 is associated with a low FDR score of 0.83, also supported by a lower MapQ score of 0.32 (Z-score -1.73)

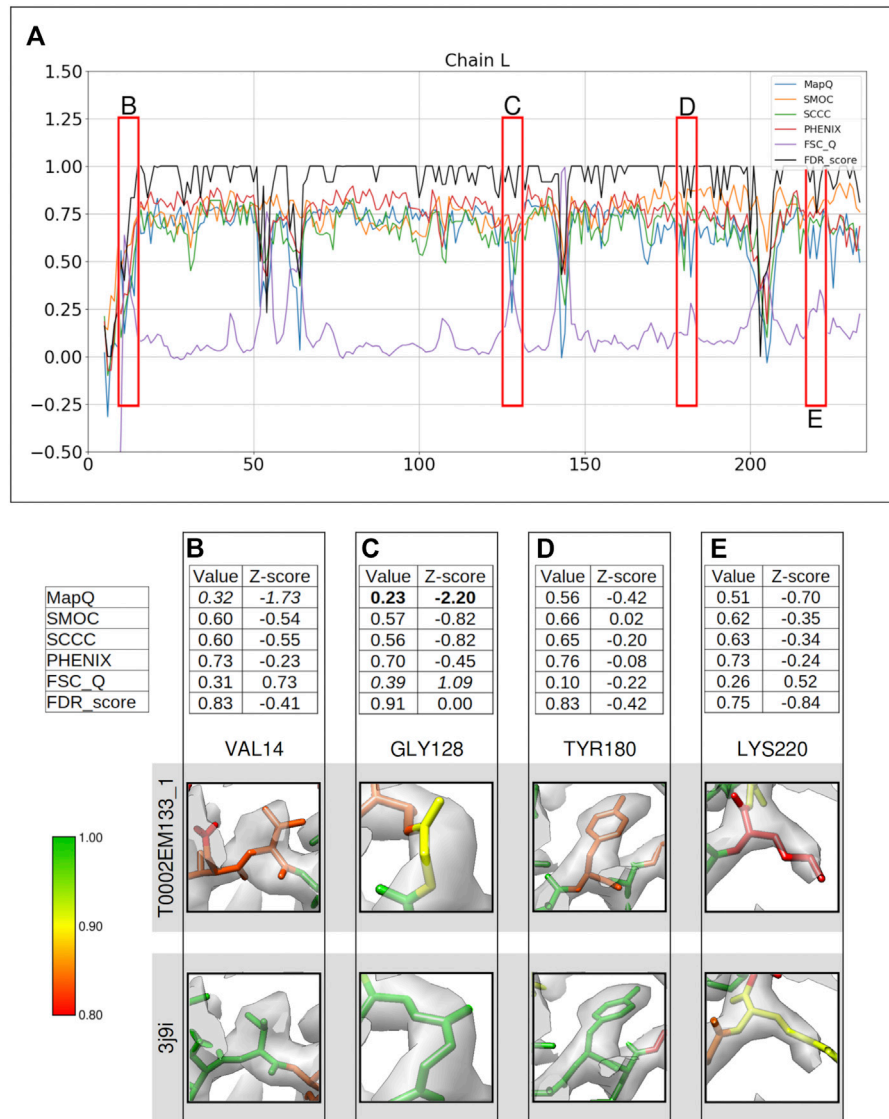


FIGURE 3 | Comparison of metrics for the atomic model of T20s proteasome, calculated only for the backbone atoms. **(A)** Comparison of the per-residue scores from MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the chain L of the atomic model T0002EM133_1 from the EMDB model; the red boxes highlight the residues selected for detailed analysis in the panels below. For Val 14 **(B)**, Gly 128 **(C)**, Tyr 180 **(D)**, and Lys 220 **(E)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-6287, grey) displayed at the recommended contour level 0.025 (3.3σ) and rendered in UCSF Chimera; and the residue fit in map as modeled in the reference (PDB ID: 3j9i).

(Figure 3B). The backbone N atom is out of the map contoured at the recommended level. The reference model (PDB ID: 3j9i) shows a better fit and has an FDR score of 1.0. Hence, the FDR backbone score and MapQ detect the mistrace with a greater significance compared to other metrics.

Figure 3C shows Gly128 which has been scored lower by MapQ (0.23, Z-score: -2.20), and FSC-Q has a score of 0.39, although with a relatively less significant Z-score of 1.09. The backbone of this residue is associated with an FDR backbone score of 0.91. Visual inspection of the backbone shows that the Ca and carbonyl C atoms are partly out of the contoured map. The

reference model shows a better fit of this residue with a higher FDR score of 0.99.

Tyr180 is assigned a low FDR backbone score of 0.83 **(Figure 3D)**, and the other scores do not highlight a backbone misplacement with all absolute Z-score values less than 0.5. The backbone N atom of the modeled residue is partly outside the contoured map. The reference model (chain B) shows a better backbone and a side chain fit and has an FDR backbone score of 0.99. Hence, the FDR score detects backbone misplacements compared to other metrics used in this study and is thus effective in identifying mistraced residue backbone.

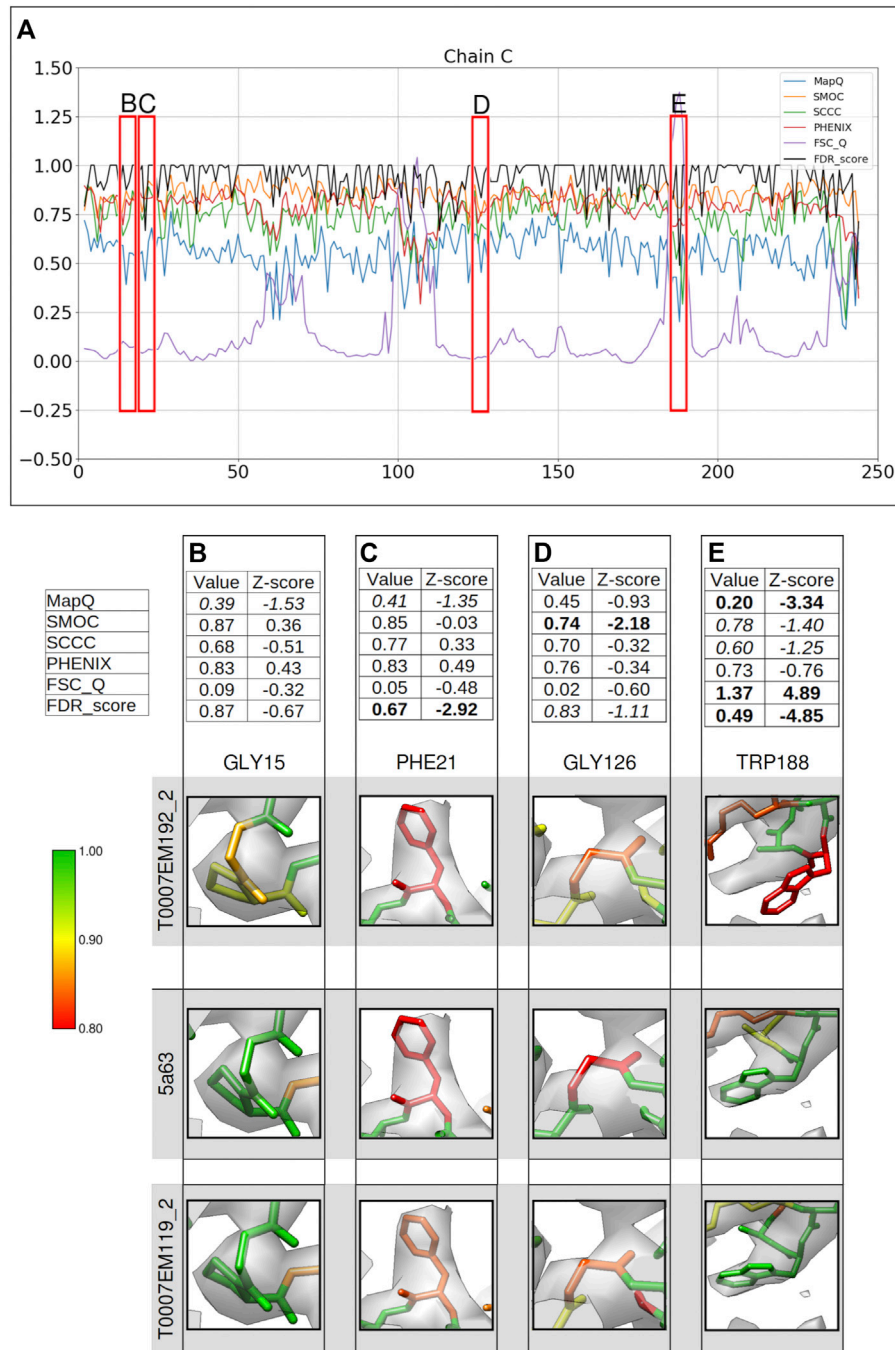


FIGURE 4 | Comparison of metrics for the atomic model of γ -secretase, calculated only for the backbone atoms. **(A)** Comparison of per-residue scores from MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the chain C of the atomic model T0007EM192_2 from the EMDB model challenge; the red boxes highlight the residues selected for detailed analysis in the panels below. For Gly 15 **(B)**, Phe 21 **(C)**, Gly 126 **(D)**, and Trp 188 **(E)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-3061, grey) displayed at the recommended contour level and rendered in UCSF Chimera (first row); the residue fit in map as modeled in the reference (PDB ID: 5a63, center), and the fit of model T0007EM119_2 in the map (last row).

This is another case where the scores disagree is Lys220 **(Figure 3E)**, which is associated with a low FDR backbone score of 0.75, while the other scores do not highlight a mistrace with significance. The MapQ score is relatively lower with a value of 0.51 (Z-score: -0.70). A closer inspection and

comparison with reference suggests that the residue has a better placement in the reference with a shift of backbone atoms accompanied by better positioning of side chain. The residue backbone in the reference was assigned an FDR score of 0.92 and the residues on either side of Lys220 also score low. This

highlights the possibility of further improvement of backbone atom placement in this segment of the reference.

γ -Secretase (3.4 Å, Target T0007)

The FDR backbone validation tool was used to assess another model (Model Challenge ID: T0007EM192_2) submitted to the EMDB Model Challenge 2015/2016 (Lawson and Chiu, 2018) for the gamma secretase map EMD-3061, solved at 3.4 Å resolution (Bai et al., 2015). **Figure 4** shows the comparison of different scores associated with residues in the chain C of the model. **Figures 4B–E** provide a closer look into some of the areas of the model where the scores disagree. As discussed below, the reference model (PDB ID: 5a63) does not show a better fit for most of these cases. Hence, we also compared the backbone fit against T0007EM119_2, which is another model submitted to the model challenge for this target and ranked higher than the reference by multiple metrics used in the challenge.

Gly15 is associated with an FDR backbone score of 0.87 (**Figure 4B**) and MapQ also associates a low score of 0.39 with the backbone (Z-score: -1.53). Other scores do not highlight a mistrace of backbone atoms for this residue. Visual inspection shows that the N and C α atoms are outside the map at the recommended contour. In the reference model (PDB ID: 5a63), the backbone shows a slight shift of the backbone toward the map volume. In the model T0007EM119_2, which scored higher than the reference in the model challenge, the atoms are shifted well into the map and Gly15 has an FDR score of 1.0. Hence, the slight backbone misplacement is highlighted by FDR and MapQ scores in this case.

Phe21 is also associated with a low FDR validation score of 0.67 and MapQ associates a relatively lower Q-score of 0.41 (Z-score: 1.35) (**Figure 4C**). The reference model shows similar backbone atom positions but associated with a lower FDR backbone score (0.58). Upon closer inspection of the model at a higher contour level, we find that the carbonyl C atom is out of the map. In the model T0007EM119_2, the residue shows a slightly better fit with the backbone shifted into the map, and has an FDR backbone score of 0.83. Multiple metrics (the FDR score and MapQ) point to a potential backbone misfit and further investigation is required in this case to establish this and check for improvement upon refitting.

Figure 4D shows Gly126 with the C α atom outside the map at the recommended contour. The modeled residue is detected as potential mistrace with an FDR backbone validation score of 0.83 and a lower SMOC score of 0.74 (Z-score -2.18). Other scores do not highlight this with a significant Z-score. The backbone of Gly126 in the reference model (PDB ID: 5a63) is also partly outside the contoured map and associated with a lower FDR score (0.75). In the model T0007EM119_2, a similar scenario was found where the C α atom is partly outside the map contour. Both the FDR score and SMOC identify a misfit in this case reflecting a potential for improvement of the backbone fit. In the absence of a good reference fit, further investigation and refitting is required to confirm the backbone misplacement.

Another residue associated with a low FDR backbone score of 0.49 is Trp188 (**Figure 4E**). The backbone mistrace is evident in this case when compared to the reference structure (PDB ID:

5a63), where Trp188 is better fitted in the map (**Figure 4E**) and has an FDR backbone score of 1.0. FSC-Q and MapQ scores also detect the backbone misplacement with significant Z-scores. The model T0007EM119_2 also shows a well-fitted backbone with an FDR score of 0.995. Hence, in this case, the FDR score works in complementarity with the metrics that calculate CCC or similar (SMOC).

Supplementary Figure S2B highlights another segment of the model (T0007EM192_2) with a polysaccharide, where the atomic positions in the terminal monosaccharide units have relatively lower FDR scores. The FDR validation score is calculated as an average of scores of the atoms in each unit. These terminal units of the carbohydrate are expected to be more flexible and the range of values of the score reflects this as well, suggesting higher uncertainty of the positions at the edges for being associated with molecular signals. The terminal monosaccharide unit in the reference model (PDB ID: 5a63) is also associated with lower FDR validation scores.

RNA Polymerase Complex From SARS-CoV-2 (2.5 Å)

We also applied our approach to assess the atomic model (PDB ID: 7bv2) deposited with the recently published structure of RNA polymerase complex (EMD-30210, 2.5 Å) from SARS-CoV-2 virus (Yin et al., 2020). A few residues in the model have lower confidence scores assigned (**Figure 5A**).

Figure 5 shows a comparison of the validation metrics for residues in the chain B of the model. The FSC-Q score was not calculated for this case as the half maps were not available from EMDB. **Figures 5B–D** provide insights into selected regions of the model, fitted in the map contoured at the recommended level 0.058 (4.3 σ). At this contour the map data corresponding to most of the backbone of low-scoring residues at the N-terminus is disconnected, possibly indicating relatively lower local resolutions. We also assessed the backbone atom placement at a lower contour level (0.035) (**Figures 5B–D**, second row). To check whether the disconnected map data is due to local over-sharpening (often resulting from a global sharpening factor applied to the map), we calculated a locally sharpened map using LocScale (Jakobi et al., 2017) implemented in the CCP-EM software suite (**Figures 5B–D**, last row).

Figure 5B shows Val83 highlighted as a potential mistrace by the FDR score with a value of 0.67 and MapQ (0.45, Z-score -2.30). The poor quality of fit is also indicated by other metrics including SMOC (0.69, Z-score -1.27), SCCC (0.51, Z-score -1.80), and PHENIX (0.58, Z-score -1.81). It can be seen that this residue backbone is not fully supported by the map even at a lower contour level (**Figure 5B**, second row) and the backbone peptide N atom is partly out of the map. As expected, the locally sharpened map from LocScale is less disordered with the peptide N atom at the edge of the map contour. The peptide N has a low FDR score of 0.0 compared to C α and carbonyl C which have scores of 1.0. In this case, the backbone is likely to be misplaced as highlighted by multiple scores.

A similar case involving Leu98 is presented in **Figure 5C**. Leucine 98 scores low with all other metrics (Z-scores lower than

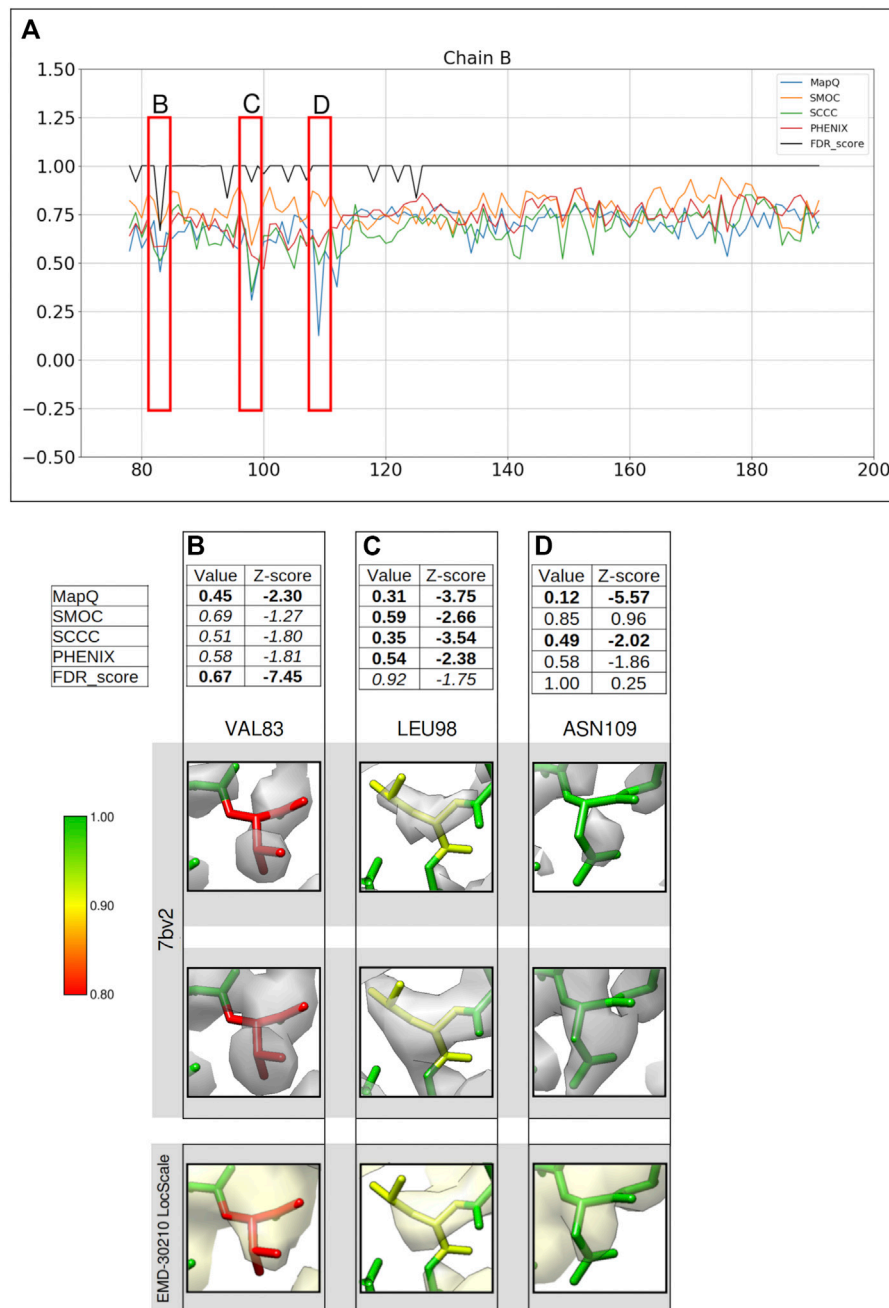


FIGURE 5 | Comparison of backbone validation metrics for the atomic model of the RNA polymerase complex (PDB ID: 7bv2). **(A)** Comparison of the per-residue scores from MapQ, SMOC, SCCC, PHENIX, and FDR backbone score for the chain B of the atomic model, the red boxes highlights the residues selected for detailed analysis in the panels below. For Val 83 **(B)**, Leu 98 **(C)**, and Asn 109 **(D)**, the panel shows a table with values of scores obtained with each metric and corresponding Z-scores; the residue fit in the target map (EMD-30210, grey) displayed at the recommended contour level and rendered in UCSF Chimera; the residues fit in the map rendered at a lower contour level.

–2.0) and this residue has an FDR score of 0.92. The position of carbonyl C atom is not fully supported by the map even at a lower contour and this atom has an FDR score of 0.75. In this case, multiple metrics highlight a potential backbone misfit and require further investigation to explore the possibility of improving the fit. The locally sharpened map also shows disconnected map trace

at the selected contour, with the carbonyl C atom placed outside the contour. In the absence of a good reference fit for this residue, Asn109 on the other hand, is highlighted as a misfit by MapQ (0.12, Z-score –5.57), SCCC (0.49, Z-score –2.02), and PHENIX (0.58, Z-score –1.86) (**Figure 5D**). However, the FDR validation score assigns a value of 1.00 (Z-score 0.13) for this residue

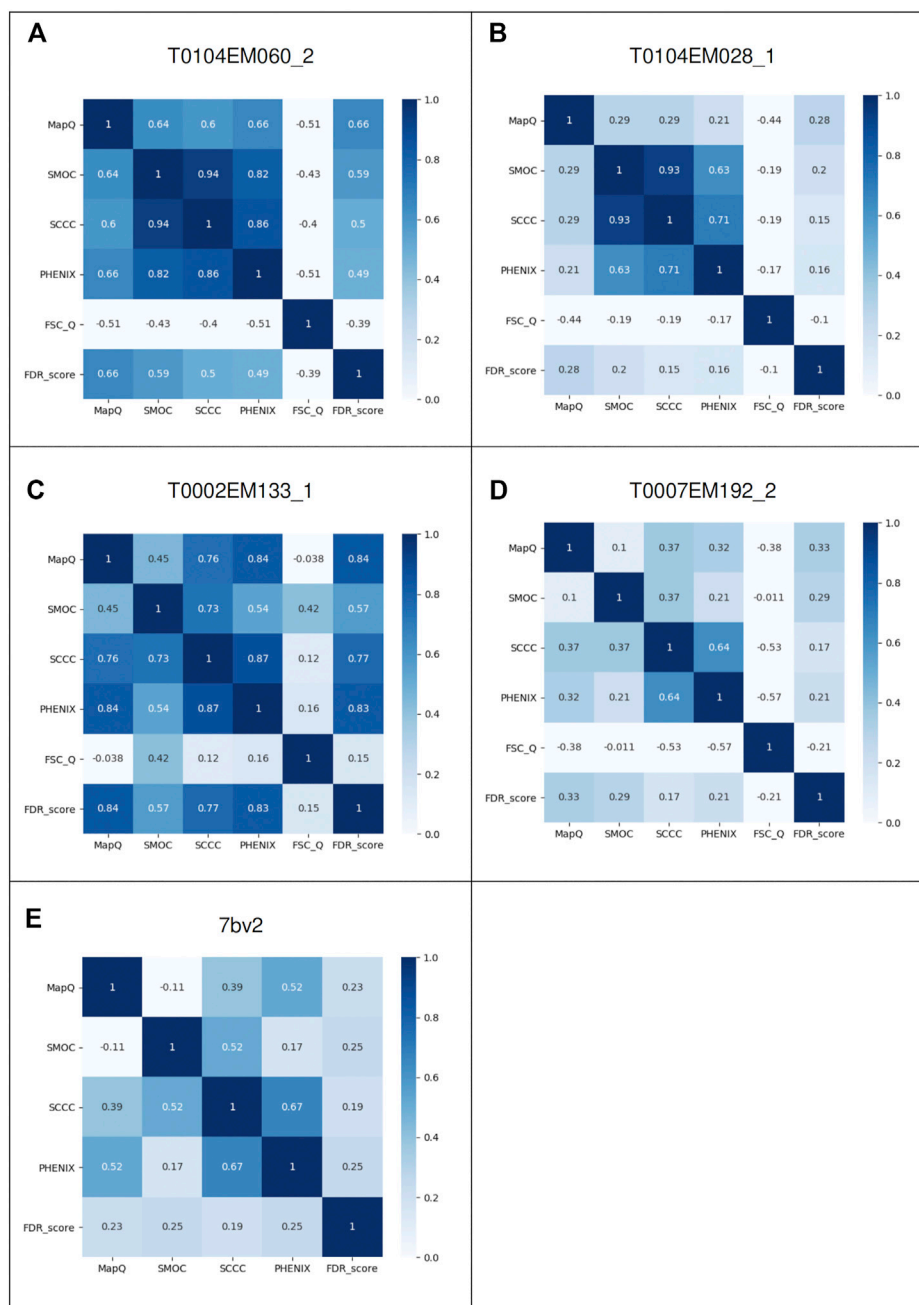


FIGURE 6 | Pairwise correlations of different metrics: MapQ, SMOC, SCCC, PHENIX, FSC-Q, and FDR backbone score for the atomic models: **(A)** Chain A of alcohol dehydrogenase T0104EM060_2, **(B)** chain A of T0104EM028_1, **(C)** chain L of T20s proteasome T0002EM133_1, **(D)** chain C of γ -secretase T0007EM192_2, and **(E)** chain B of the RNA polymerase complex model 7bv2.

backbone, suggesting no mistrace of the backbone. A closer inspection of this position in the model shows that the lower contour level covers most of the backbone atoms of the residue, except carbonyl C atom where the map is still disconnected. The locally sharpened map is smoother with no disorder and shows a better coverage of backbone atoms. The residue is located at a low resolution area of the map and

it is likely that the backbone is within the molecular volume but the atoms are misfitted, as highlighted by the other metrics.

Supplementary Figure S2C shows an area of the modeled RNA where the terminal nucleotide has a lower FDR score. The map density is also disordered at this position likely due to the higher flexibility of this part of the RNA.

TABLE 1 | Outlier detection by different metrics for ten of the models submitted to the 2019 Model Challenge for the target alcohol dehydrogenase map (EMD-0406). For each model, the table number of residues of chain A associated with Z-scores lower than -2.0 . For the FDR backbone score, the table also shows the number of residues with an FDR backbone score less than 0.9

ModelID	CC	Method	FDR_score <0.9	FDR_score Z-score < -2	MapQ Z-score < -2	SMOC Z-score < -2	SCCC Z-score < -2	PHENIX Z-score < -2	FSC-Q Z-score > 2	FSC-Q Z-score < -2
T0104EM035_1	0.32	<i>Ab-initio</i>	3	6	15	9	8	8	12	7
T0104EM027_1	0.32	<i>Ab-initio</i>	8	7	15	11	9	10	10	6
T0104EM010_1	0.32	<i>Ab-initio</i>	7	9	14	8	13	6	11	7
T0104EM041_1	0.32	<i>Ab-initio</i>	10	10	10	9	13	9	13	5
T0104EM090_1	0.31	<i>Ab-initio</i>	21	19	12	13	12	12	12	0
T0104EM028_1	0.31	<i>Ab-initio</i>	9	15	13	12	10	10	14	2
T0104EM025_1	0.31	Optimized	8	9	12	17	14	11	14	5
T0104EM082_1	0.31	<i>Ab-initio</i>	9	9	15	18	18	12	12	2
T0104EM060_2	0.28	<i>Ab-initio</i>	39	66	17	14	14	16	8	0
T0104EM054_1	0.27	<i>Ab-initio</i>	83	26	17	16	18	17	18	3

Correlation Between Different Metrics

In the cases discussed above, we show a number of cases where different metrics disagree in the detection of backbone mistrace and cases where the FDR backbone scores can be complementary. To check how different metrics rank models based on local backbone fit to maps, we scored ten of the models submitted to the 2019 Model Challenge for the target alcohol dehydrogenase map (EMD-0406), and nine of them were built using *ab-initio* model building approaches. For each model, the number of residues of chain A associated with a Z-score lower than -2.0 were counted (Table 1). The table also shows the number of residues with the FDR backbone score less than 0.9. The models in the table are sorted by the global CCC scores derived from the assessment results of the model challenge (https://model-compare.emdataresource.org/2019/cgi-bin/em_multimer_results.cgi?target_map=T0104emd_0406). No two metrics completely agree in the ranks assigned to the models based on the number of potential backbone misfits. However, there is a general agreement on best scoring the models and those with the lowest ranks. Note that the Z-scores are less meaningful in cases where the distribution of the score is far from normal. This is expected to affect the ranks, especially in the case of FSC-Q and MapQ where the outliers have significantly lower scores than the rest of the distribution.

To further investigate the pairwise agreement between metrics, we computed pairwise correlations between scores for the case studies discussed above. Figures 6A–E present the correlation matrices highlighting pairwise correlations between metrics for each case (corresponding to Figures 1–5). In a general scenario where an atomic model fits well overall in a map but includes a few mistraced residues, the majority of the residues have FDR scores of 1.0 and we expect lower scores for mistraced residues. Hence, the FDR score being less variable relative to other scores, the pairwise correlations involving the FDR score are expected to be low. Indeed, we observe this for most of the models except for T0104EM060_2 and T0002EM133_1 where many of the residues are associated with low backbone scores (Figure 4). In these two cases, the FDR score shows better correlation with MapQ with pairwise correlation coefficients of 0.66 and 0.84,

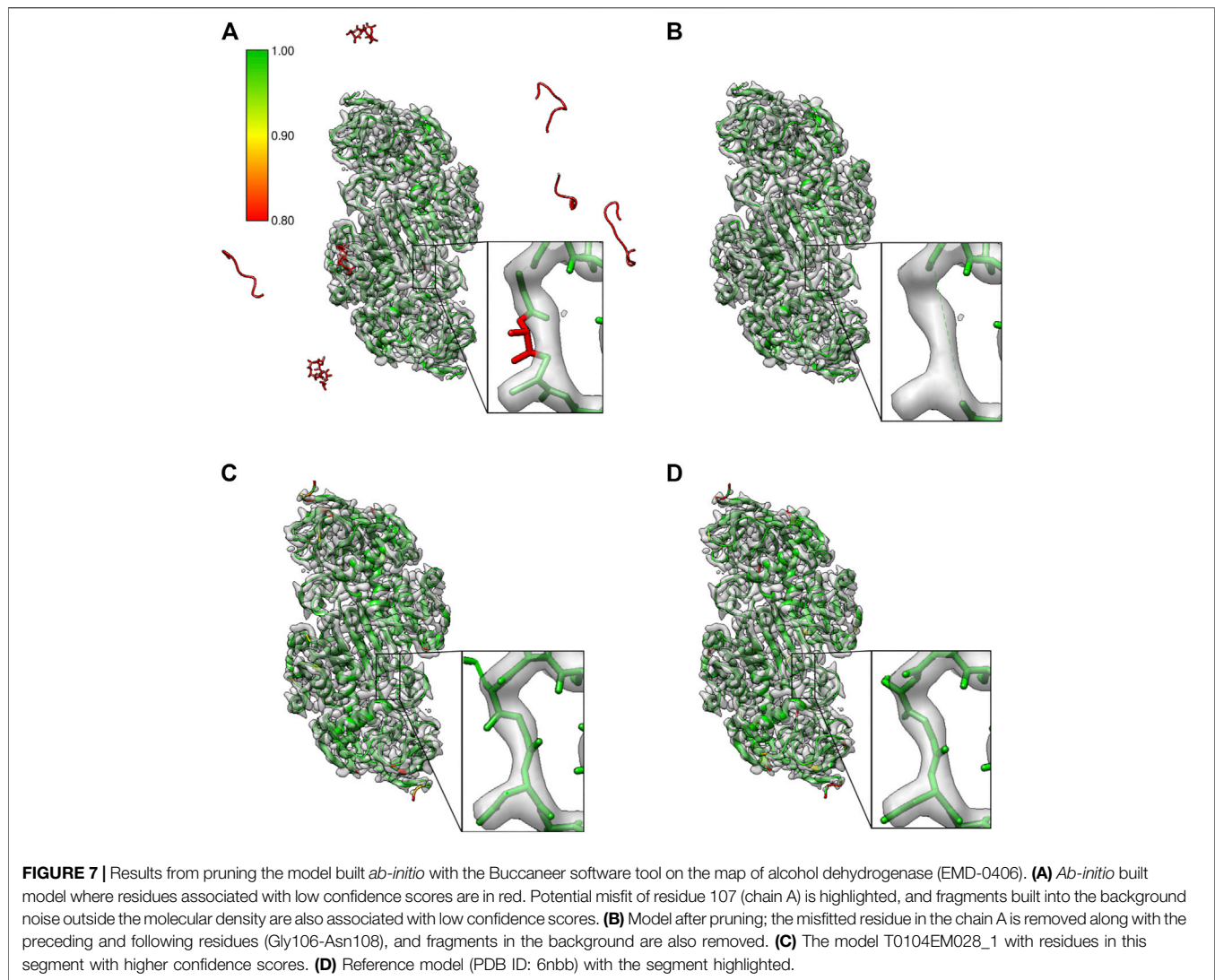
respectively. MapQ scores also correlate with SCCC and PHENIX_CC scores for these two cases.

Overall, SCCC and PHENIX CCC show a good correlation in most cases with pairwise correlations in the range 0.64 to 0.87, which is expected as both scores involve calculation of cross correlation coefficient. SCCC and SMOC scores are largely correlated as well with pairwise correlations spread between 0.37 and 0.94. These two scores use similar underlying procedures for synthetic map generation from model and identification of voxels covered by atoms. FSC-Q does not correlate with any of the other scores as the score reflects the model-map (and map-model) differences, unlike the other scores.

Pruning *Ab-Initio* Built Models

The proposed approach was used to prune models generated by Buccaneer (Cowtan, 2006) which is an *ab-initio* model building tool that works by an iterative process involving finding backbone seed positions, growing them to fragments, connecting and pruning fragments to chains and pruning the resulting chains. Often the final model from Buccaneer needs to be pruned interactively in Coot to remove any fragments and fix any obvious mistraces. Identifying parts of the model that are fitted into low confidence regions of the map enables automated pruning of the models.

We tested this using the *ab-initio* model built using the Buccaneer software for the 2.9 Å reconstruction of alcohol dehydrogenase (EMD-0406). Figure 7A shows the model built from four Buccaneer cycles. The confidence map-based approach identifies fragments built into the background noise outside the molecular density (highlighted in red). The zoomed area provides a closer look at the loop where one of the residues is mistraced and backbone atoms are out of the contoured map. Figure 7B shows the same model after pruning based on our approach. All the fragments and mistraced residues were removed. Residues on either side of the low scoring residue are also removed while pruning. This helps to rebuild this whole region in the next round of the automated model building. Figures 7C,D shows the confidence scores for the same segment from the model T0104EM028_1 and the reference model (6nbb.2), respectively. The residues of these models have higher



confidence scores and the residues are fitted better in the contoured map.

DISCUSSION

The majority of cryo-EM reconstructions in EMDB are determined at resolutions worse than 3 Å and often the local resolution varies significantly in maps that are otherwise resolved at higher resolutions on an average. Hence, the chances of errors in the model are higher and validation tools that can detect errors and areas with high uncertainty, are necessary. In this study we tested an approach that evaluates the backbone trace of atomic models based on the local molecular signal (compared to background noise) in the map. The confidence scores calculated per voxel from the original map using the FDR control approach (Beckers et al., 2019) are mapped to individual backbone atoms in the model.

For the purpose of testing the approach, we used examples covering a range of reported resolutions from 2.5 to 3.4 Å. The residue backbones that have an FDR score less than 0.9 are included in **Supplementary Table S1**. Most of these models are built using model building and refinement tools commonly used in the field, as part of the EMDB model challenges. These challenges act as platforms to assess models derived using a wide range of modeling techniques and compare metrics which can be used to evaluate these atomic models. It also provides a repository of models built from a range of map targets and a reference model to compare against, which can be extremely useful for development and testing of new validation software.

We show that the FDR backbone score is complementary to existing model evaluation tools. The proposed score evaluates only the atomic positions and not the model agreement with the map. Hence, it is not useful for detecting any misfits within the molecular contour. Also, the current implementation of the score does not identify side chain rotamer misfits. However, as seen in

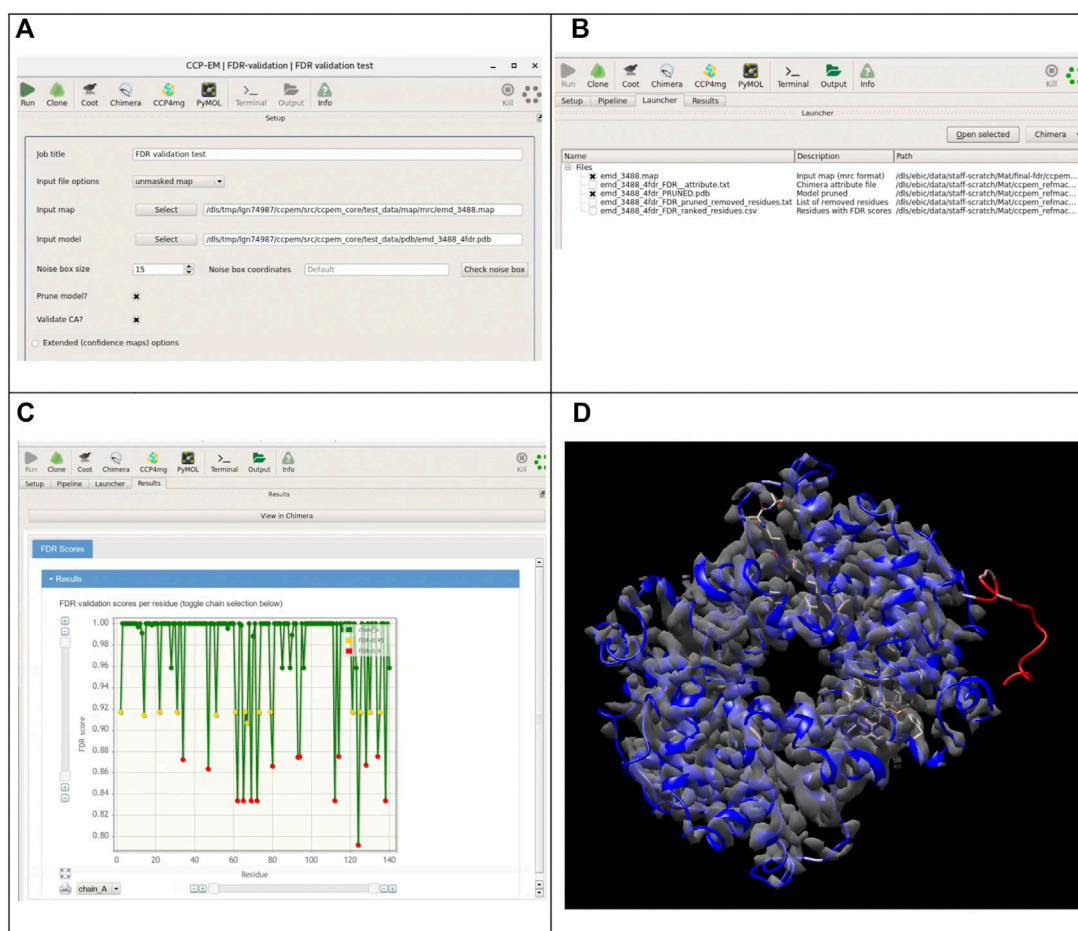


FIGURE 8 | Integration of the presented tool with the CCP-EM software suite. **(A)** Input interface for the FDR-validation task. The original map (preferably unmasked) or a precalculated confidence map and a model are required as inputs. If the original map is provided as input, a confidence map will be calculated internally. Users have access to the advanced setup options for confidence map job. **(B)** Overview of the launcher tab listing the output files. **(C)** Results tab with the FDR backbone scores plots for each chain. **(D)** Resulting models colored according to calculated confidence score, overlaid with the confidence map in UCSF Chimera (accessible from the results tab of the CCP-EM task).

many of the cases discussed in results, often backbone mistraces are associated with side chain misfits as well.

On the other hand, as demonstrated in results, some residues where one or more atoms are fitted into the background noise may still have fit-to-map scores within tolerable limits. This misplacement of backbone atoms is evident when compared to the reference, where a better backbone fit can be found. In such cases, the FDR backbone score works in a complementary manner.

Potential backbone mistraces involving a number of glycine residues were detected by the FDR backbone score (see Results), and not by other metrics. One explanation could be that glycine is often seen in flexible loops associated with low-resolution areas of the map, and some of these scores are sensitive to map resolution. In general, limiting the score calculations to backbone atoms, might also affect some of the scores like CCC, where a sufficiently large distribution of values is expected for meaningful estimation of mean and standard deviation and hence a reliable score calculation.

We also show that the approach detects weak molecular signals that are at low resolution areas of the map and not otherwise obvious. We recommend that residues associated with FDR scores less than 0.95 usually require attention and residues with scores less than 0.9 usually reflect clear cases of backbone mistrace.

We also demonstrate that the approach is useful in detecting residue mistraces in a model. Hence, the tool is useful as part of iterative model building pipelines or to evaluate the final model. Automated pruning of models based on this approach can be a useful step in the iterative model building and refinement process. Models after pruning can be also a starting point for extending or iterative building with the Buccaneer model building tool. As presented in the results, the approach is useful to validate ligands, carbohydrates, and nucleic acids as well.

The implementation of this score as a tool in the CCP-EM software suite makes it easily accessible for the cryo-EM community. The described software tool is available from the CCP-EM suite as “FDR validation task” confidence map

calculation can be run as part of this task, where the user has to provide the original map (preferably unmasked) and the model to validate. As an option, the user can adjust the size of the noise box used for calculation of background statistics. In some cases, especially if the specimen is significantly elongated in one direction, users should also check the preview of the noise boxes to make sure that the noise box does not contain any part of the molecular volume. The extended options for the confidence maps section allows to set the advanced parameters for the FDR maps calculation (**Figure 8A**).

If the user has already generated the FDR map, it can be used directly as an input (**Figure 8B**). Instead of a confidence map, any custom map can be used as well and residues will be assigned scores based on values in the map. Validation based on the scores of backbone atoms is run by default, users can additionally choose to validate only the CA positions. Optionally, the model can be pruned further to remove residues associated with low confidence scores. A model file with atomic b-factors replaced by the confidence scores and a CSV file containing the confidence scores for each residue are generated as outputs. If the option to prune the model was chosen, a pruned model is provided as the additional output, along with a text file containing the list of all removed residues. **Figure 8C** shows the launcher tab with a list of output files generated from the job. On the results tab, a link is included to open the resulting models directly in UCSF Chimera with the model colored based on the confidence scores. **Figure 8D** shows the results open directly in UCSF Chimera with the resulting model colored according to the confidence score.

For a confidence map of size $192 \times 192 \times 192$ voxels, the assignment of FDR scores to residues takes about 0.22 s on a PC with specification: Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz x 8, 8 GB RAM. The latest CCP-EM nightly release available from <https://www.ccpem.ac.uk/download.php> includes the implementation of “FDR validation task” The source code of the tool for evaluating atomic models based on confidence maps is available from (<https://github.com/m-olek/FDR-validation>).

CONCLUSION

In this study, we present a tool for validating atomic models derived from cryoEM maps. It works based on the calculation of the confidence maps, which estimates molecular signal to noise at every voxel, and also detects weak signals from the low resolution areas of the map. This helps to assess atomic positions based on the

local information in the map and identify mistraced residues in the model. This approach is complementary to other validation tools that quantify agreement with the map, as it evaluates atom positions based on the local map information. We believe that, with the integration with the CCP-EM software suite, the presented tool will be a useful addition to the existing validation tools.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

MO developed the tool, performed analysis on different examples, and drafted the manuscript. AJ and MO conceived the idea. AJ helped with analysis, implementation of the tool in CCP-EM, and manuscript writing.

FUNDING

This work was supported by PhD studentship from the University of York and Diamond (eBIC) and Wellcome Trust research grant WT (208398/Z/17/Z).

ACKNOWLEDGMENTS

We thank Prof. Kevin Cowtan and Prof. Peijun Zhang for useful discussions throughout the course of this study. We thank Maxmillian Beckers, Arjen Jakobi, and Prof. Carsten Sachse for help with implementation of FDR approach in CCP-EM and useful comments on this work. We also thank Sony Malhotra for comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.652530/full#supplementary-material>

REFERENCES

- Afonine, P. V., Klaholz, B. P., Moriarty, N. W., Poon, B. K., Sobolev, O. V., Terwilliger, T. C., et al. (2018). New Tools for the Analysis and Validation of Cryo-EM Maps and Atomic Models. *Acta Cryst. Sect D Struct. Biol.* 74 (9), 814–840. doi:10.1107/s2059798318009324
- Bai, X., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., et al. (2015). An Atomic Structure of Human γ -Secretase. *Nature* 525 (7568), 212–17. doi:10.1038/nature14892
- Beckers, M., Jakobi, A. J., and Sachse, C. (2019). Thresholding of Cryo-EM Density Maps by False Discovery Rate Control. *Int. Union Crystallogr. J.* 6 (1), 18–33. doi:10.1107/s2052252518014434
- Benjamini, Y., and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* 29 (4), 1165–1188. doi:10.1214/aos/1013699998
- Burnley, T., Palmer, C. M., and Winn, M. (2017). Recent Developments in the CCP-EM software Suite. *Acta Cryst. Sect D Struct. Biol.* 73 (6), 469–477. doi:10.1107/s2059798317007859
- Cowtan, K. (2006). The Buccaneer Software for Automated Model Building. *Acta Crystallogr. D* 62. *Acta Crystallogr. Section D, Biol. Crystallogr.* 62 (October), 1002–100211. doi:10.1107/s0907444906022116
- Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010). Features and Development of Coot. *Acta Crystallogr. D Biol. Cryst.* 66 (4), 486–501. doi:10.1107/s0907444910007493

- Herzik, M. A., Wu, M., and Lander, G. C. (2019). High-Resolution Structure Determination of Sub-100 KDa Complexes Using Conventional Cryo-EM. *Nat. Commun.* 10 (1), 1032. doi:10.1038/s41467-019-08991-8
- Hoh, S. W., Burnley, T., and Cowtan, K. (2020). Current Approaches for Automated Model Building into Cryo-EM Maps Using Buccaneer with CCP-EM. *Acta Cryst. Sect D Struct. Biol.* 76 (6), 531–541. doi:10.1107/s2059798320005513
- Jakobi, A. J., Wilmanns, M., and Sachse, C. (2017). Model-Based Local Density Sharpening of Cryo-EM Maps. *ELife* 6 (October), e27131. doi:10.7554/elife.27131
- Joseph, A. P., Malhotra, S., Burnley, T., Wood, C., Clare, D. K., Winn, M., et al. (2016). Refinement of Atomic Models in High Resolution EM Reconstructions Using Flex-EM and Local Assessment. *Methods* 100 (May), 42–49. doi:10.1016/j.ymeth.2016.03.007
- Lagerstedt, I., Moore, W. J., Patwardhan, A., Sanz-García, E., Best, C., Swedlow, J. R., et al. (2013). Web-Based Visualisation and Analysis of 3D Electron-Microscopy Data from EMDB and PDB. *J. Struct. Biol.* 184 (2), 173–181. doi:10.1016/j.jsb.2013.09.021
- Lawson, C. L., and Chiu, W. (2018). Comparing Cryo-EM Structures. *J. Struct. Biol.* 204 (3), 523–526. doi:10.1016/j.jsb.2018.10.004
- Lawson, C. L., Kryshtafovych, A., Adams, P. D., Afonine, P. V., Baker, M. L., Barad, B. A., et al. (2021). Cryo-EM Model Validation Recommendations Based on Outcomes of the 2019 EMDDataResource Challenge. *Nat. Methods* 18 (2), 156–164. doi:10.1038/s41592-020-01051-w
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Long, F., Vagin, A. A., et al. (2011). REFMAC5 for the Refinement of Macromolecular Crystal Structures. *Acta Crystallogr. Section D: Biol. Crystallogr.* 67 (Pt 4), 355–367. doi:10.1107/s0907444911001314
- Palmer, C. (2016). Mrcfile: MRC File I/O Library. Available at: <https://github.com/ccpem/mrcfile>.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera?A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25 (13), 1605–1612. doi:10.1002/jcc.20084
- Pfaff, J., Phan, N. M., and Si, D. (2021). DeepTracer for Fast De Novo Cryo-EM Protein Structure Modeling and Special Studies on CoV-Related Complexes. *Proc. Natl. Acad. Sci.* 118 (2). doi:10.1073/pnas.2017525118
- Pintilie, G. (2020). Measurement of Atom Resolvability in CryoEM Maps with Q-Scores. *Microsc. Microanal.* 26 (S2), 2316. doi:10.1017/s1431927620021170
- Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S., and Richardson, D. C. (2020). New Tools in MolProbity Validation: CaBLAM for CryoEM Backbone, UnDowser to Rethink "waters," and NGL Viewer to Recapture Online 3D Graphics. *Protein Sci.* 29 (1), 315–329. doi:10.1002/pro.3786
- Ramírez-Aportela, E., Erney, D. M., and Fonseca, Y. C. (2011). 'FSC-Q: A CryoEM Map-To-Atomic Model Quality Validation Based on the Local Fourier Shell Correlation'. *Nat. Commun.* 12 (1), 42. 10.1038/s41467-020-20295-w
- Subramaniam, S. (2019). The Cryo-EM Revolution: Fueling the Next Phase. *Int. Union Crystallogr. J.* 6 (1), 1–2. doi:10.1107/s2052252519000277
- Terwilliger, T. C., Adams, P. D., Afonine, P. V., and Sobolev, O. V. (2020). Cryo-EM Map Interpretation and Protein Model-building Using Iterative Map Segmentation. *Protein Sci.* 29 (1), 87–99. doi:10.1002/pro.3740
- Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci.* 27 (1), 293–315. doi:10.1002/pro.3330
- Wojdyr, M. (2017). Project-Gemmi/Gemmi. C++. GEMMI. Available at: <https://github.com/project-gemmi/gemmi> (Accessed April 9, 2021)
- Yin, W., Mao, C., Luan, X., Shen, D.-D., Shen, Q., Su, H., et al. (2020). Structural Basis for Inhibition of the RNA-dependent RNA Polymerase from SARS-CoV-2 by Remdesivir. *Science* 368 (6498), 1499–1504. doi:10.1126/science.abc1560
- 'NumPy Reference NumPy v1.16 Manual' (2019). Available at: <https://docs.scipy.org/doc/numpy-1.16.0/reference/> (Accessed April 9, 2021).
- 'PyQt 4.9.4 Reference Guide' (2011). 'PyQt 4.9.4 Reference Guide'. Available at: <https://doc.bccnsoft.com/docs/PyQt4/>

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

Copyright © 2021 Olek and Joseph. 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.



Structural Design and Analysis of the RHOA-ARHGEF1 Binding Mode: Challenges and Applications for Protein-Protein Interface Prediction

Ennys Gheyouche¹, Matthias Bagueneau¹, Gervaise Loirand², Bernard Offmann¹ and Stéphane Téletchéa^{1*}

¹ UFIP, Université de Nantes, UMR CNRS 6286, Nantes, France, ² Université de Nantes, CHU Nantes, CNRS, Inserm, L'institut Du Thorax, Nantes, France

OPEN ACCESS

Edited by:

Joseph Rebehmed,
Lebanese American University,
Lebanon

Reviewed by:

Sophie Sacquin-Mora,
UPR9080 Laboratoire de Biochimie
Théorique (LBT), France
Guillaume Postic,
INSERM U973 Molécules
Thérapeutiques in Silico (MTI), France

*Correspondence:

Stéphane Téletchéa
stephane.teletchea@univ-nantes.fr

Specialty section:

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

Received: 18 December 2020

Accepted: 13 April 2021

Published: 24 May 2021

Citation:

Gheyouche E, Bagueneau M,
Loirand G, Offmann B and
Téletchéa S (2021) Structural Design
and Analysis of the RHOA-ARHGEF1
Binding Mode: Challenges and
Applications for Protein-Protein
Interface Prediction.
Front. Mol. Biosci. 8:643728.
doi: 10.3389/fmolb.2021.643728

The interaction between two proteins may involve local movements, such as small side-chains re-positioning or more global allosteric movements, such as domain rearrangement. We studied how one can build a precise and detailed protein-protein interface using existing protein-protein docking methods, and how it can be possible to enhance the initial structures using molecular dynamics simulations and data-driven human inspection. We present how this strategy was applied to the modeling of RHOA-ARHGEF1 interaction using similar complexes of RHOA bound to other members of the Rho guanine nucleotide exchange factor family for comparative assessment. In parallel, a more crude approach based on structural superimposition and molecular replacement was also assessed. Both models were then successfully refined using molecular dynamics simulations leading to protein structures where the major data from scientific literature could be recovered. We expect that the detailed strategy used in this work will prove useful for other protein-protein interface design. The RHOA-ARHGEF1 interface modeled here will be extremely useful for the design of inhibitors targeting this protein-protein interaction (PPI).

Keywords: PPI, protein-protein docking, molecular dynamics simulation, ARHGEF1, RHOA

1. INTRODUCTION

Precise interactions between proteins allow a tight control on many functions and pathways, eventually leading to gene expression or silencing, to protein release or degradation, and even to cell death. To date, there are between 130,000 and up to 650,000 protein-protein interactions (PPIs) described (Ottmann, 2015), but only a fraction of PPIs is validated experimentally, ranging from 14,000 (Rolland et al., 2014) to 125,000 PPIs (<http://interactome3d.irbbarcelona.org/>, Mosca et al., 2013). This structural gap comes from the difficulties of obtaining experimentally full-length interacting proteins and then to resolving their structures using crystallography, NMR, or electron microscopy (EM). As a result, in most cases for a specific PPI, one has to combine existing incomplete experimental structures with *in silico* approaches. This virtual step is even critical for drug discovery, as exemplified with the successful targeting of 50 PPIs by small molecules (Skwarczynska and Ottmann, 2015). When no protein-protein structure is available, one has thus to perform protein-protein docking predictions.

The binding mode prediction of two proteins is very challenging since: (i) minor to major local structural rearrangements may be triggered upon protein recognition, (ii) one protein may recognize multiple proteins, and (iii) cofactors/nucleic acids may be involved to enhance or stabilize the interaction. The analysis of the various existing protein-protein interfaces available in the Protein Data Bank indicates also that (i) the interface area varies greatly between protein families, (ii) the composition in amino acids in this interface may be biased, and (iii) the binding lifetime is transient. Due to the complexity of modeling these diverse PPIs, there are many methods developed, and a global evaluation called CAPRI is performed periodically. We use the most robust and successful methods validated in this competition for protein-protein docking evaluation of the optimal binding mode of our example (Lensink et al., 2007, 2018, 2019).

To illustrate the process of modeling a protein-protein interface, we selected a protein complex where some unbound and bound experimental structures are available, and a complex between other members of the family is available. The first protein partner in our study is RHOA (gene *RHOA*), the second protein partner is Rho Guanine nucleotide Exchange Factor 1 (gene *ARHGEF1*). RHOA is a member of the RAS superfamily of small GTPases recognized as a master regulator of the actin cytoskeleton, thus driving multiple cellular processes, such as cell contraction, migration, proliferation, and gene transcription. While basal and controlled RHOA activity is required for homeostatic functions in physiological conditions, its uncontrolled overactivation plays a causative role in the pathogenesis of several diseases, such as cancer, neurodegenerative, or cardiovascular diseases (Guilluy, 2010; Cherfils and Zeghouf, 2011; Vetter, 2014; Loirand, 2015; Prieto-Dominguez et al., 2019; Arrazola Sastre et al., 2020). RHOA is a molecular switch that couples cell surface receptors to intracellular effector pathways by cycling between a cytosolic inactive state bound to guanosine 5'-diphosphate (GDP), and an active GTP-bound state that translocates to the membrane. The activation of RHOA is mediated by Rho nucleotide guanine exchange factors (GEFs) that promote the exchange of GDP for GTP, which are themselves turned on by the activation of upstream membrane receptors (Cherfils and Zeghouf, 2013). ARHGEF1 is the Rho GEF responsible for the activation of RHOA by angiotensin II through type 1 angiotensin II receptor in vascular smooth muscle cells (Loirand and Pacaud, 2010; Luigia et al., 2015). This signaling pathway participates in the physiological control of the vascular tone and blood pressure, and is causally involved in the pathophysiology of hypertension (Guilluy, 2010).

Small GTPases structure consists of a six-stranded β -sheet (β strands B1 to B6) linked by helices and loops (Ihara et al., 1998). In RHOA, the β -sheet is made up of the anti-parallel association of B1 and B2 and the parallel association of B3, B1, B4, B5, and B6, and there are five α helices (A1, A3, A3', A4, and A5) and three 3_{10} helices (H1–H3). RHOA possess two hinge regions, a

loop called switch I (29–42) and an helix called switch II (62–68), which are described to be more flexible than the core β -sheet (Dvorsky and Ahmadian, 2004), as shown in **Figure 1** for various bound and unbound experimental structures. RHOGEF proteins catalyze the exchange of GDP with 5'-triphosphate (GTP) on RHOA (Felline et al., 2019). Two domains on these proteins, Pleckstrin Homology (PH) and Dbl Homology (DH), are involved in the nucleotide exchange mechanism, the RHOA-bound DH domain being more rigid than the PH domain (**Figure 1C**). The DH domain consists of six α -helices arranged in an oblong shape, which interact with switch I and switch II regions of RHOA. The PH domain contains seven antiparallel β -strands forming a roll architecture, connected to helix $\alpha 6$ of the DH domain. The mechanism of nucleotide exchange involves large displacements of the PH domain relative to the DH-RHOA interaction (Felline et al., 2019).

In this article, we show how to combine virtual approaches with experimental data to predict reliably the formerly non-existing structure of a PPI. By using a rough structural superimposition or more advanced protein-protein docking methods, we build, analyze, and refine the RHOA-ARHGEF1 model and discuss the strengths and weaknesses of this approach. We, then, derive general recommendations to reproduce our approach to model other PPIs.

2. MATERIALS AND METHODS

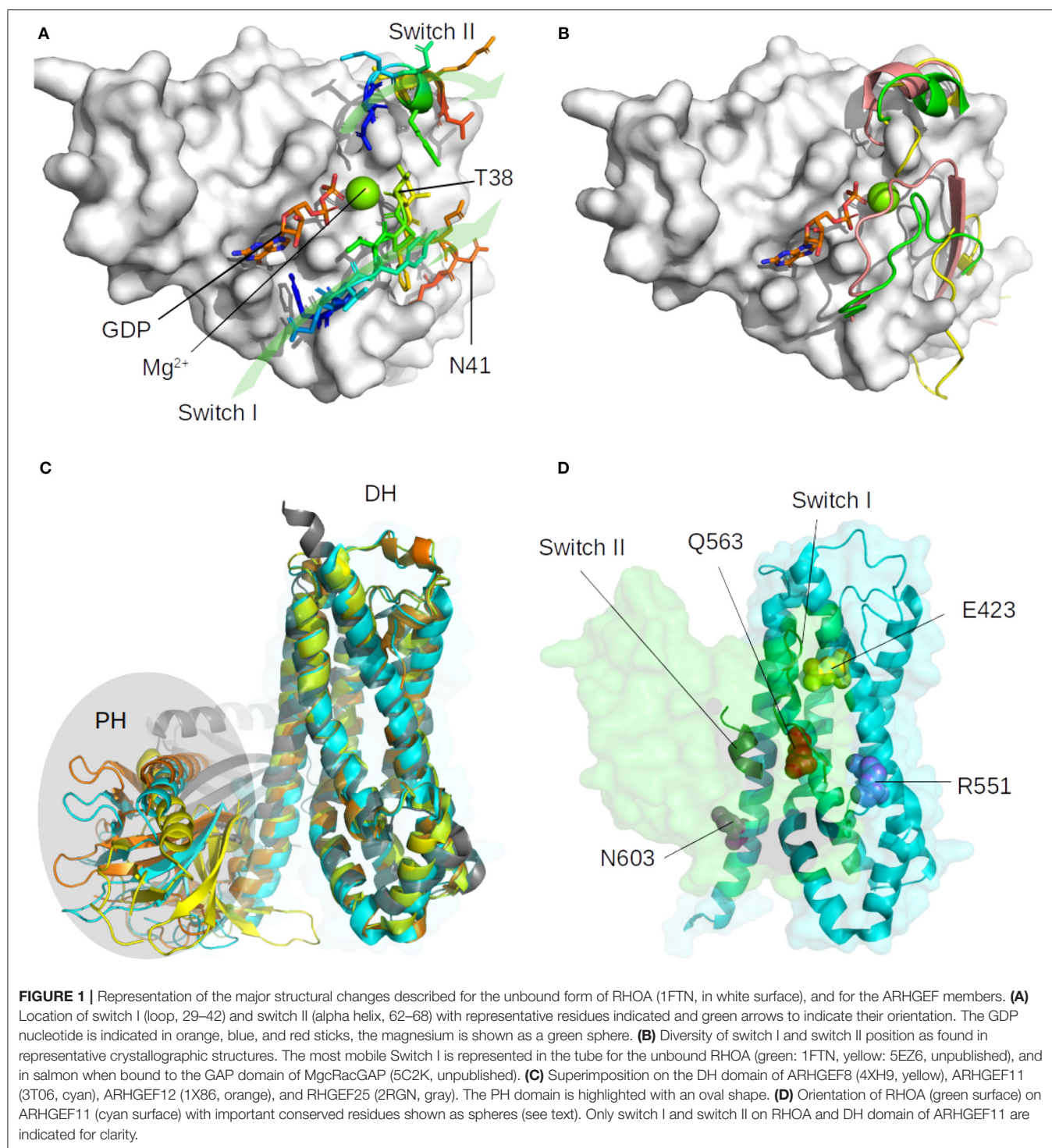
2.1. Sequence Analysis

A BLAST sequence search on the Non-Redundant (NR) database with RHOA or ARHGEF1 sequence as query was performed in June, 2018. The resulting sequences were aligned using clustal Omega (Madeira et al., 2019), amino acids conservation was estimated using Jalview (Waterhouse et al., 2009) and in-house scripts in Biopython (Cock et al., 2009). The phylogenetic analysis of human conserved sequences was performed in Genious version 2019.0.4 (<http://www.geneious.com/>).

2.2. Structure and Interface Analysis

We used all structures of RHOA complexed with all ARHGEFs available in the Protein Databank (Berman et al., 2003) in January 2018 and their respective unbound form when available. Only one representative chain by crystallographic structure was taken as reference (**Supplementary Table 1**). Experimental structures were analyzed using PDBePISA version 1.54 (Krissinel and Henrick, 2007; Krissinel, 2010). Two methods were selected to analyze protein-protein interfaces in order to obtain useful insights of the important residues involved in the interaction. The first one was 2P2I Inspector, version 2.0 (http://2p2idb.cnrs-mrs.fr/2p2i_inspector.html) (Basse et al., 2016) which computes a series of 51 chemical and physical descriptors from three-dimensional (3D) structures. The second one, PPCheck (<http://caps.ncbs.res.in/ppcheck/>) is a webserver for quantifying the strength of a protein-protein interface (Sukhwil and Sowdhamini, 2013). It can also be used to predict hotspots, perform computational alanine scanning, and to differentiate possible native-like conformations from the non-native ones

Abbreviations: RMSD, root mean square deviation; SAS, solvent accessible surface.



given a set of decoy ensembles as obtained through the protein-protein docking as it computes the strength of non-bonded interactions between any two proteins/chains present in the complex. Robetta Server was used to perform virtual alanine scanning of the interface (Kortemme et al., 2004). Models from docking or molecular dynamics simulations were visually

assessed and analyzed in The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. Root Mean Square Deviation (RMSD) was computed using the PyMOL rms_cur command. As RMSD is a global measure, we use two specific measures for rigid body docking and molecular dynamics simulations interface analysis as described in Takemura and

Kitao (2019): (i) Ligand-RMSD (L-RMSD) where one protein is Fixed (F) and the second protein is Mobile (M) to compute the L-RMSD. First the Fixed protein is superimposed on the same structure in the crystallographic reference, then the RMSD is computed on the Mobile (M) protein alone. (ii) Interface-RMSD (i-RMSD): in this case, only the amino acids known to be involved in the interface between both proteins are evaluated. Again, a first step consists of superimposing the one protein (RHOA or the PH domain of ARHGEFs) on the reference crystallographic structure to remove translation and rotation degrees of freedom potentially coming from the docking methods process. The angle between helices is computed using a plugin by Thomas Holder, which computes the angle between two vectors created from the coordinates of the C α atoms of each helix.

2.3. Superimposition Model of RHOA-ARHGEF1

A preliminary structure of RHOA bound to ARHGEF1 was derived from the co-crystal of RHOA-ARHGEF11 (PDB id:3T06) (Bielnicki et al., 2011). The complexes were created by superimposition of ARHGEF1 (3ODO) (Chen et al., 2011) on ARHGEF11 in PyMOL. This superimposition was submitted to MolProbity (Williams et al., 2018) for analysis and steric clashes were removed using Chiron (Ramachandran et al., 2011). Chiron performs rapid energy minimization of protein molecules using discrete molecular dynamics with an all-atom representation for each residue in the protein, this process allows to remove most of the steric clashes.

2.4. Protein-Protein Docking

The binding mode prediction was done in default mode for all methods using their respective webserver: (i) ATTRACT ff2g (<http://chemosimserver.unice.fr/attract/>) (Chéron et al., 2017), (ii) ClusPro version 2 (<https://cluspro.bu.edu/home.php>) (Kozakov et al., 2017), (iii) Haddock (van Zundert et al., 2016), (iv) PyDockWeb, Oct 2017 (<https://life.bsc.es/servlet/pydock/home/>) (Jiménez-García et al., 2013), (v) ZDOCK version 3.0.2f (Pierce et al., 2014).

2.5. Molecular Dynamics Simulation

All-atom simulations of unbound and bound proteins were performed using GROMACS 2016.3 (Abraham et al., 2015), the starting structures are detailed in **Supplementary Table 1**. Each system was prepared with the AMBER forcefield FF99SB-ILDN (Lindorff-Larsen et al., 2010) in explicit solvent (TIP3P) (Jorgensen et al., 1983) with a specific attention to protonation states as reported in PROPKA. A NVT followed by the anisotropic pressure coupling (NPT ensemble) protocol was applied until equilibration was reached and the full molecular dynamics simulation was computed for 500 ns up to 1 μ s. All simulations were run on the CCIPL cluster facility at the University of Nantes using GPUs. The force field parameters for GDP and GTP were gathered from the AMBER parameters database (<http://research.bmh.manchester.ac.uk/bryce/amber>) and converted to GROMACS format files using acpype (da Silva and Vranken, 2012). The resulting trajectories were visualized in the VMD version 1.9.1 (Humphrey et al., 1996), and GROMACS tools were used for various measurements.

TABLE 1 | Description of the protein-protein docking methods evaluated.

ATTRACT	<i>Ab-initio</i> protocol using a coarse-grained forcefield (ff2g) and manage to predict an estimation of the binding energy between the two proteins.	Chéron et al., 2017
CLUSPRO	FFT based program PIPER (pairwise potential based on the decoy at the reference approach) for the docking and using RMSD based clustering for filtering models then scoring using four different Scoring functions	Kozakov et al., 2017
HADDOCK	High Ambiguity-Driven DOCKing allows the use of external data, either experimental or from bioinformatics analysis to drive the modeling process	van Zundert et al., 2016
PyDockWeb	Rigid-body docking using FTDock, Gabb et al. (1997) and Jiménez-García et al. (2013) then an energy scoring based on empirical potential composed by of electrostatic and desolvation terms.	Cheng et al., 2007
ZDOCK	Fast Fourier Transform based protein docking program, version 3.0.2f (IFACE Statistical Potential, Shape Complementarity, and Electrostatics)	Pierce et al., 2014

PyContact was used to analyze protein contacts type, strength, and lifetime throughout the simulations (Scheurer et al., 2018). These contacts were plot with the R package MDplots (Margreitter and Oostenbrink, 2017).

3. RESULTS

In order to determine which docking method was the best for our specific needs, we have evaluated their performance (i) on recovering existing crystallographic structures of RHOA bound to a GEF, a process called re-docking, and (ii) on assembling the bound RHOA from a given crystallographic with a GEF from another crystallographic structure, this process being known as cross-docking.

3.1. Protein-Protein Docking Strategies

3.1.1. Z-Dock Is the Best Method for Building Our Complex

As docking strategies are based on different methods, it is difficult to determine *a priori* which method will produce the most reasonable starting complex for further studies. We assessed the top five performing docking software from CAPRI assessment, briefly introduced in **Table 1**, to produce RHOA/ARHGEF1 complexes: ATTRACT, ClusPro, HADDOCK, PyDockWeb, and ZDOCK.

3.1.2. Assessment of Webserver Performance in Re-docking Experiments

We evaluated the performance of each software according to its ability to recover the existing structure of bound RHOA and ARHGEFs. This method called re-docking allows to discriminate the accuracy of the algorithm studied on our system. The web servers define the first input protein given as the receptor, so

it stays fixed (F) and considers the second protein provided as the mobile protein (M). We performed our analysis with the small GTPase or the GEF as (F) or (M). The results are presented in **Table 2**. We computed the L-RMSD of the predicted complex by superimposing the Fixed protein with the same protein in the crystallographic structure and by computing the RMSD on the Mobile protein to assess the performance of each method. When comparing individually the predicted pose against the reference, ATTRACT and PyDockWeb perform equally with lower L-RMSD values for ATTRACT on its best predictions. ZDOCK and ClusPro present more diverse results and larger deviations. Although it should not be important, we observed that the input order of the fixed and mobile protein was affecting the prediction. This is of limited importance for ATTRACT and PyDockWeb, these methods are, therefore, less sensitive to the size of the mobile protein (RHOA being 200 AA long and GEF DH/PH domains being 600 AA long). We observe that for 4XH9, there is a dramatic decrease in the quality of the prediction although the calculations were performed three times. It is also possible to determine which models ranks the best among all poses and not only the first one: ZDOCK and ATTRACT proposes 10 binding modes, ClusPro offers multiple weights for their scoring functions for clusters of poses, and PyDockWeb presents the 100 best binding poses.

TABLE 2 | Analysis of docking software performance for complex formation using small GTPase (RHOA) or GEF (ARHGEF 8/11/12/25) as a mobile (M) or fixed (F) structure.

GEF(F). RHOA(M)/ GEF(M). RHOA(F)	ZDOCK	ATTRACT	ClusPro	PyDockWeb
3T06	1.60/1.92	0.70/0.97	3.21/7.10	1.66/1.34
4XH9	1.99/3.06	20.25/14.09**	4.27/1.80	1.52/1.53
1X86	2.61/3.23	1.11/1.71	3.32/5.81	1.44/0.92
2RGN	2.14/3.00	1.46/2.24	3.93/3.01	0.92/0.80
Computing time	24 h	24 h	24 h	1 week

The L-RMSD values are in (Å). The best prediction analyzed is the first of the best 10 poses or the best cluster of poses classified by each method. The best predictions are in bold, the wrong predictions are in italics. HADDOCK could not be evaluated at this stage since we chose to not use data guidance. Computing time is indicative of one docking experiment.

**Calculation were performed independently in triplicate to exclude any temporary issue.

All the poses identified as close to the crystallographic structure with a global RMSD under 2 Å were present in the top 10 solutions of the best cluster for each method. Altogether, this lead us to consider the PyDockWeb at the end of the re-docking study, although its computation time is significantly higher than the other methods (**Table 2**). As ZDOCK and PyDockWeb provided more reliably poses close to the original crystallographic structures, these two methods were kept for the following analysis.

3.1.3. Assessment of the Best Performing Methods in Cross-Docking Experiments

We verified the dependence on the initial structure for the protein-protein binding mode prediction. We applied a cross-docking experiment where each bound RHOA in one crystal structure is evaluated against another GEF partner. We used the free RHOA structure as a sensitivity control for our cross-docking measurements. There is no dependence for the docking result linked to the pdb input for RHOA or GEF. The results are available in **Table 3**: as one would expect, it is more complicated to build hybrid complexes than re-docking complexes. Out of the 12 combinations of partners for the cross-docking, ZDOCK is better for six predictions, PyDockWeb for five, and their prediction is good and close in one case. For two cases, pydockweb finds a very different orientation than the crystal structure : RHOA: 1X86/GEF: 2RGN (30.27 Å), RHOA: 4XH9/GEF: 1X86 (34.25 Å), where ZDOCK finds a close conformation (3.19/3.31). For both methods, the docking of the unbound RHOA produces very unrealistic protein-protein interfaces, indicating the sensitivity of the method toward switch I and II adaptations of RHOA for GEF binding although the RMSD between bound RHOA (3T06) and unbound RHOA (1FTN) is small (0.7 Å).

Considering the results of the re/cross-docking, we selected ZDOCK as the best method to predict the more favorable binding mode between RHOA and ARHGEF1 (**Figure 2**).

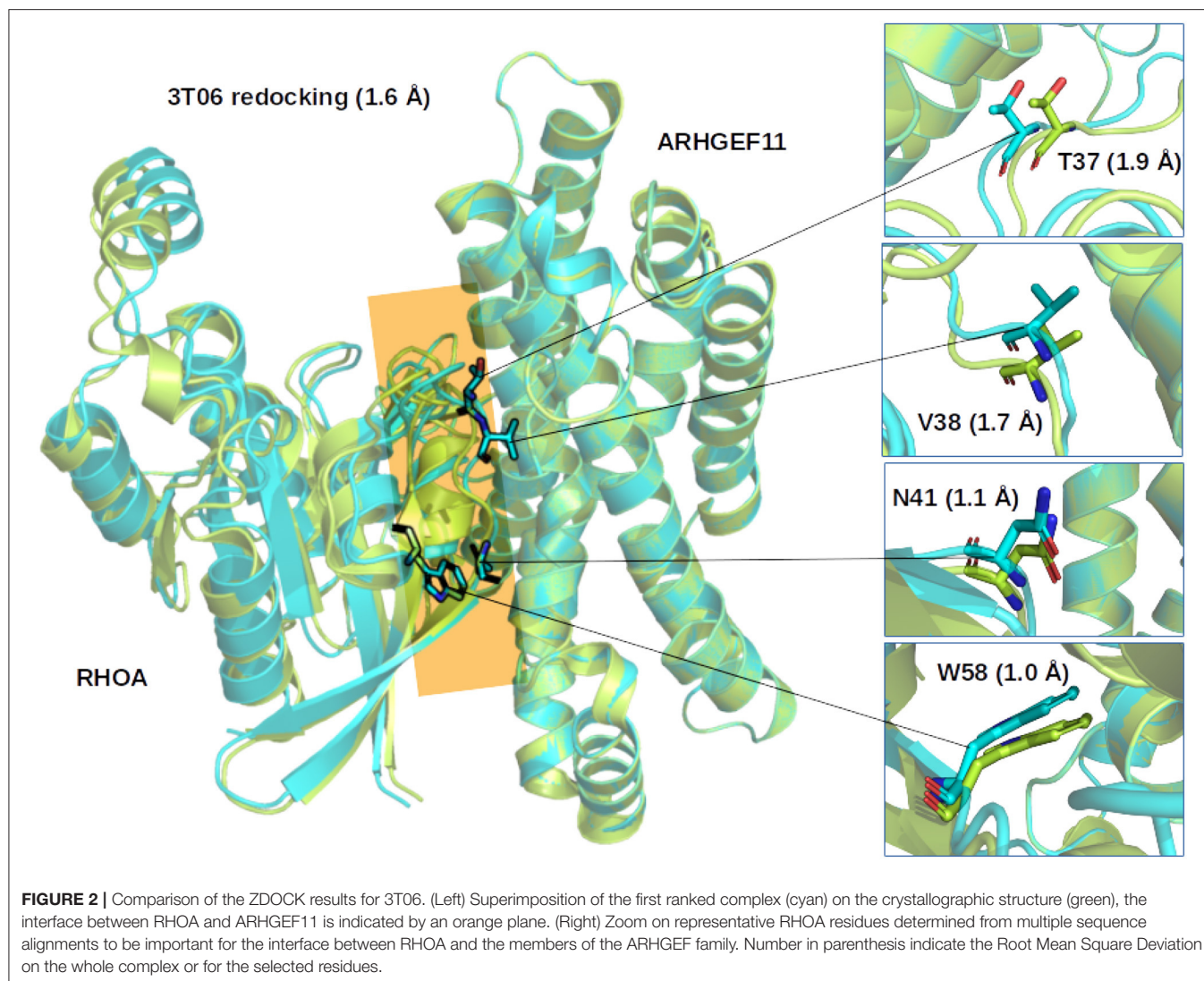
3.1.4. Evaluation of the Best Models Based on Known Interactions to Select the Best RHOA Candidate Structure

Since no experimental structure of the RHOA-ARHGEF1 interface is available, we used the crystallographic structures

TABLE 3 | Analysis of the crossdocking performance for the ZDOCK and PyDOCKweb. RHOA was considered as the ligand (M), and each GEF was fixed (F).

ZDOCK/pydockweb		GEF			
		GEF11 (3T06)	GEF8 (4XH9)	GEF12 (1X86)	GEF25 (2RGN)
Bound RHOA	3T06	<u>1.60/1.66</u>	2.15/2.10	3.22/ 1.86	3.51/ 2.46
	4XH9	3.35/5.58	<u>1.99/1.52</u>	3.31/34.25	2.59/3.56
	1X86	3.15/ 2.46	2.66/ 2.02	<u>2.61/1.44</u>	3.19/30.27
	2RGN	3.01/5.38	2.39/3.38	2.53/2.16	<u>2.14/0.92</u>
Free RHOA	1FTN	14.42/29.35	17.02/16.39	31.08/38.60	14.44/38.13

The RMSD value (in Å) for the mobile part is displayed for the best prediction in the first 10 poses. The results for re-docking experiments, underlined, are identical to **Table 2**. The best predictions are in bold.



of other GEF paralogs bound to RHOA to determine which amino acids are shared in all complexes. The amino acids mapping to RHOA and ARHGEF1 was done using multiple sequence alignments comparison (data not shown). Four residues are conserved in GEFs interfaces, namely, E423, Q563, R551, and N603 in ARHGEF1. As can be seen in **Figure 3**, these shared residues for all GEF interfaces can be split in zones or individual amino acids contacts. By analogy to existing complexes, ARHGEF1 E423 has to be present close to Y34/T37/V38 (a region called switch I in RHOA), R551 has to be close to V43/D45/E54 of RHOA, and N603 has to be close to D67/R68/L69 (a region called switch II in RHOA). Only one amino acid in RHOA, N41 seems to bind exclusively to Q563. It is well-established that RHOA is very rigid due to the strong structural requirements imposed by the GTP recognition and hydrolysis mechanism (Dvorsky and Ahmadian, 2004). Only two regions called switch I and switch II are more flexible with or without binding partners, as seen in **Figure 1**. Our study allows a more detailed understanding of

the interaction between amino acids pairs important for the RHOA-ARHGEF binding.

As done for ARHGEF1, we also analyzed the interface residues of the other GEFs present in each crystallographic structures in the complex with RHOA and assessed their amino acids conservation using multiple sequence alignments. Four residues of ARHGEF1 are present at the interface, and 10 for RHOA. As can be seen in **Table 4**, we have listed all amino acids present in interaction between both proteins, i.e., at least one amino acid of RHOA is in contact with one amino acid or more of a given ARHGEF. On average, the number of contact pairs recovered after docking represents at least half of the residues known to be present on both sides of the interface. This indicates that our docking strategy allows to build a reasonable starting structure for the RHOA-ARHGEF1 complex. As sequence conservation on the DH+PH domains modeled here is the most important among ARHGEF11, ARHGEF12, and ARHGEF1 (**Supplementary Table 2**), and provided the docking validation steps showed that ARHGEF11 docking

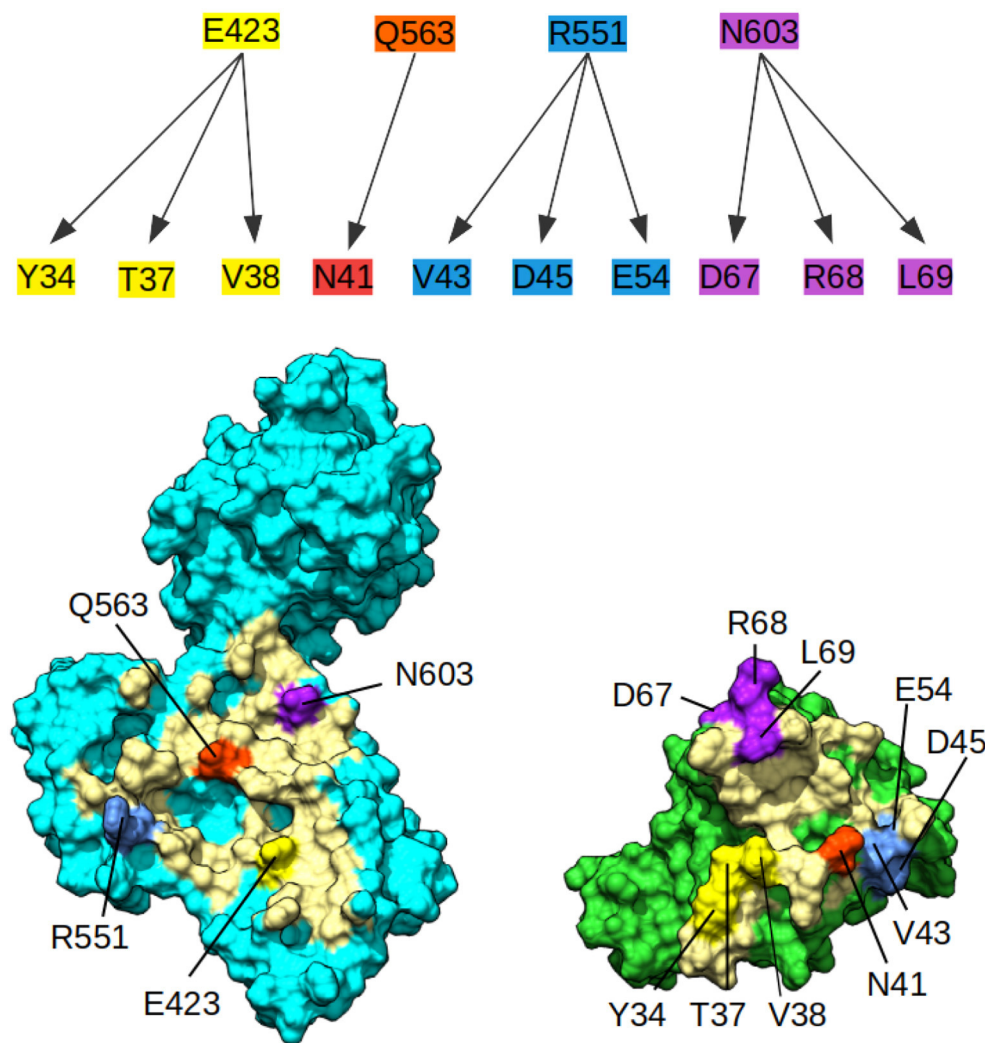


FIGURE 3 | Conserved contacts in the interface between RHOA and all its GEFs. (Top) Diagram of conserved contacts by amino acids, amino acids E423, Q563, R551, and N603 in ARHGEF1, and residues linked by arrows pertaining to RHOA. (Bottom) Split view of ARHGEF1 (cyan) RHOA (green) with matching residues between proteins highlighted in yellow, white, red, blue, and purple, the rest of the interface is indicated in pale yellow.

allowed to recover more contacts than with ARHGEF12 docking, we chose the RHOA structure found in 3T06 to dock it using ZDOCK with the unbound structure of ARHGEF1 (3ODO). The resulting complex will be referenced thereafter as complexD (Docking).

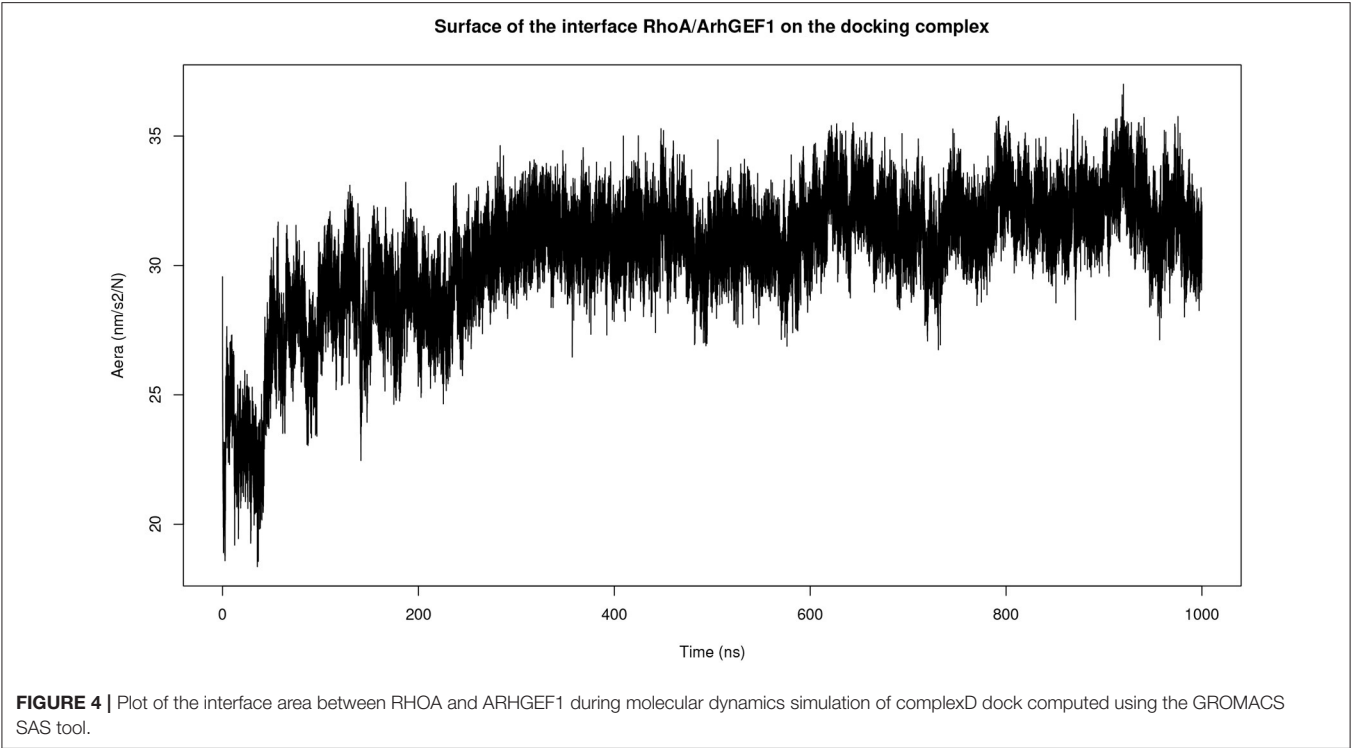
3.2. Template-Based Complex Modeling

Since there are some experimental structures of RHOA bound to ARHGEF1 homologs, we also predicted the bound structure of both proteins with a simpler approach, based on structural superimposition in PyMOL. We first used a rhoA-bound crystal structure and superimposed the free ARHGEF1 on all the homologous GEFs. By doing so with rigid models, we could not take into account the concerted induced-fit required for finely tuning the interaction. We used a webserver from Dokholyan Team named Chiron which allows to relax

the most important steric clashes (<http://redshift.med.unc.edu/chiron/>) (Ramachandran et al., 2011). In order to solve all the bumps, many rounds were necessary. A preliminary structure of RHOA bound to ARHGEF1 was derived from the RHOA-ARHGEF11 crystal structure (3T06). The complexes were created by superimposition of ARHGEF1 (3ODO) on ARHGEF11 in PyMOL. This superimposition was submitted to MolProbity for analysis, and steric clashes were removed using Chiron. From there, we used classical descriptors to evaluate the resulting complex (delta SAS, RMSD, ...) with a special look into the interface size. This interface was analyzed with PDBePISA. The best binding was found when using ARHGEF1 from 3ODO and RHOA from 3T06: the interface is 2,949 Å² corresponding to around 5% of the total surface of ARHGEF1 and to around 11% of RHOA total surface area. In the complexD, this interface comprises amino acids 3 to 181 from RHOA and 392 to 761

TABLE 4 | Numbers of amino acids determined to be in interaction between both proteins at the interface, numbers from RHOA (x/10) or its complexed ARHGEF (y/4), as found in the different binding mode generated by ZDOCK during the crossdocking experiments, while the ARHGEF partner was kept fixed.

	ARHGEF12 (1X86)	ARHGEF25 (2RGN)	ARHGEF11 (3T06)	ARHGEF8 (4XH9)
RHOA (1X86)	5/10-3/4	5/10-2/4	2/10-1/4	3/10-1/4
RHOA (2RGN)	3/10-3/4	6/10-3/4	8/10-4/4	5/10-2/4
RHOA (3T06)	5/10-1/4	7/10-3/4	6/10-3/4	6/10-2/4
RHOA (4XH9)	3/10-3/4	4/10-2/4	1/10-1/4	7/10-3/4
RHOA (1FTN)	0	0	0	0



from ARHGEF1, for a total interface size of 2,830 Å². Those values are smaller than the values found for the homologs complexes where this interface area is on average 3,371 Å². Before exploring further the complexD and complexT, we verified our models with PPcheck, which decomposes the interaction energy in three terms: (i) hydrogen bonding (E_{hyd}), (ii) inter-chain van der Waals interactions (E_{vw}), and (iii) inter-chain electrostatic interactions (E_{ele}). The total stabilizing energy is then divided by the total number of interface residues to obtain the energy per residue. No significant deviation requiring further refinements with the Chiron webserver was present in both models.

3.3. Molecular Dynamics Simulation Refinement for ComplexT and ComplexD

Both complexes were modeled using molecular dynamics simulations to determine if the interface could be refined during this procedure. The initial surface area in complexD, 2,830 Å² at the beginning of the simulation, stabilizes to

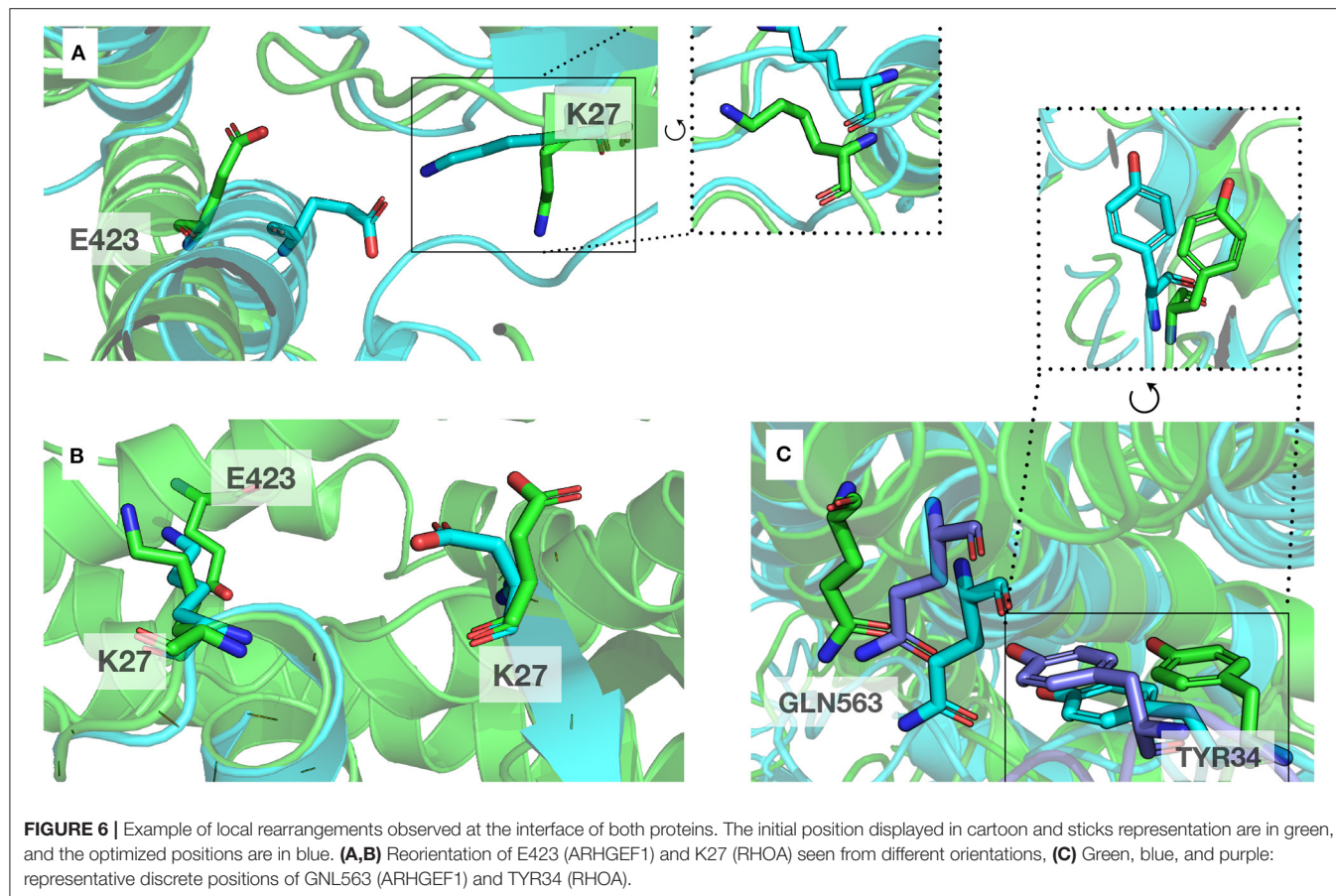
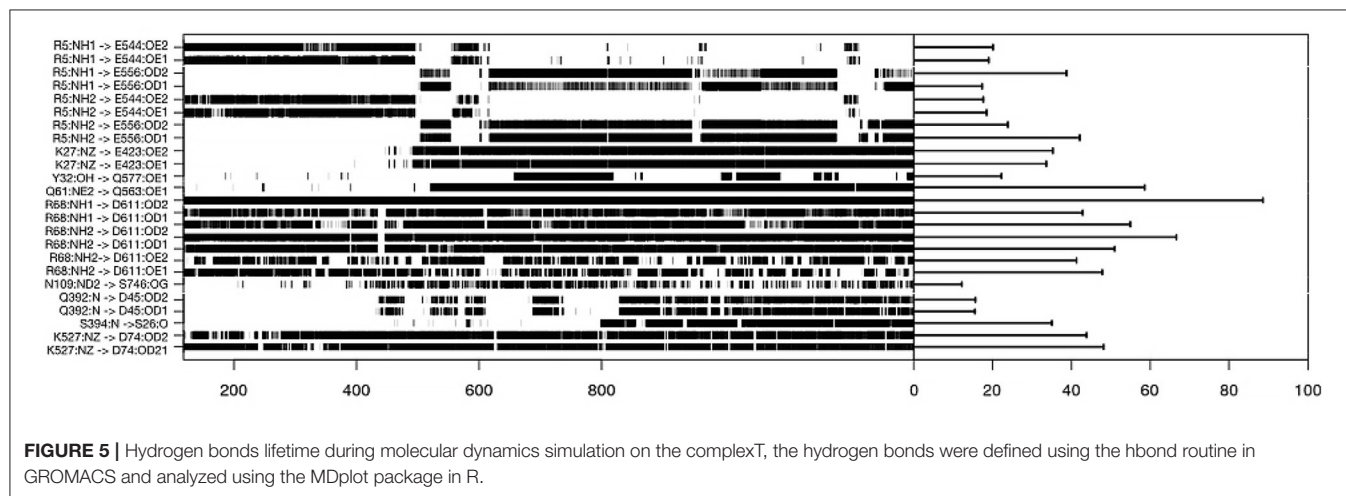
3,056 Å² between 200 and 1,000 ns of the simulation (**Figure 4**). The molecular dynamics simulation of RHOA-ARHGEF1 complexT also remained stable for most of the time with a rapid initial increase in the interface area followed by a plateau after 250 ns. The average interface size in this plateau is 3,150 Å² (data not shown). Starting with two different models, we observe an augmentation in protein surface contact driven by local adjustments. When considering individually each protein at the RMSD level, there is a higher deviation for RHOA than ARHGEF1, implying that RHOA undergoes most of the conformational changes, as we will see in details below.

3.4. Interface Contacts Evolution Over Time

To understand the evolution of interface complex during the simulations, we analyzed the hydrogen bonds between the two partners. The result is shown in **Figure 5** where we only plotted hydrogen bonds with a lifetime in the simulation

over 15%. With this plot, we can identify which amino acids, side chains are seeing important rotations. For instance, an important hydrogen bond is conserved between RHOA-ARG5 and ARHGEF1-GLU544 or ARHGEF1-ASP556. The most stable contact is RHOA-ARG68—ARHGEF1-ASN603/ASP611, since it is observed for 900 ns, or 90% of the simulation time. For others contacts, some were present from the beginning

of the simulation, others appeared and disappeared. Since it may take time to stabilize contacts, we observe that an important interaction appears between RHOA-GLN61 and ARHGEF1-GLN563 at 500 ns. Interestingly, some of these hydrogen bonds are members of the very conserved list of amino acids listed above, for instance, for RHOA-ARG68 and ARHGEF1-ASN603.



4. DISCUSSION

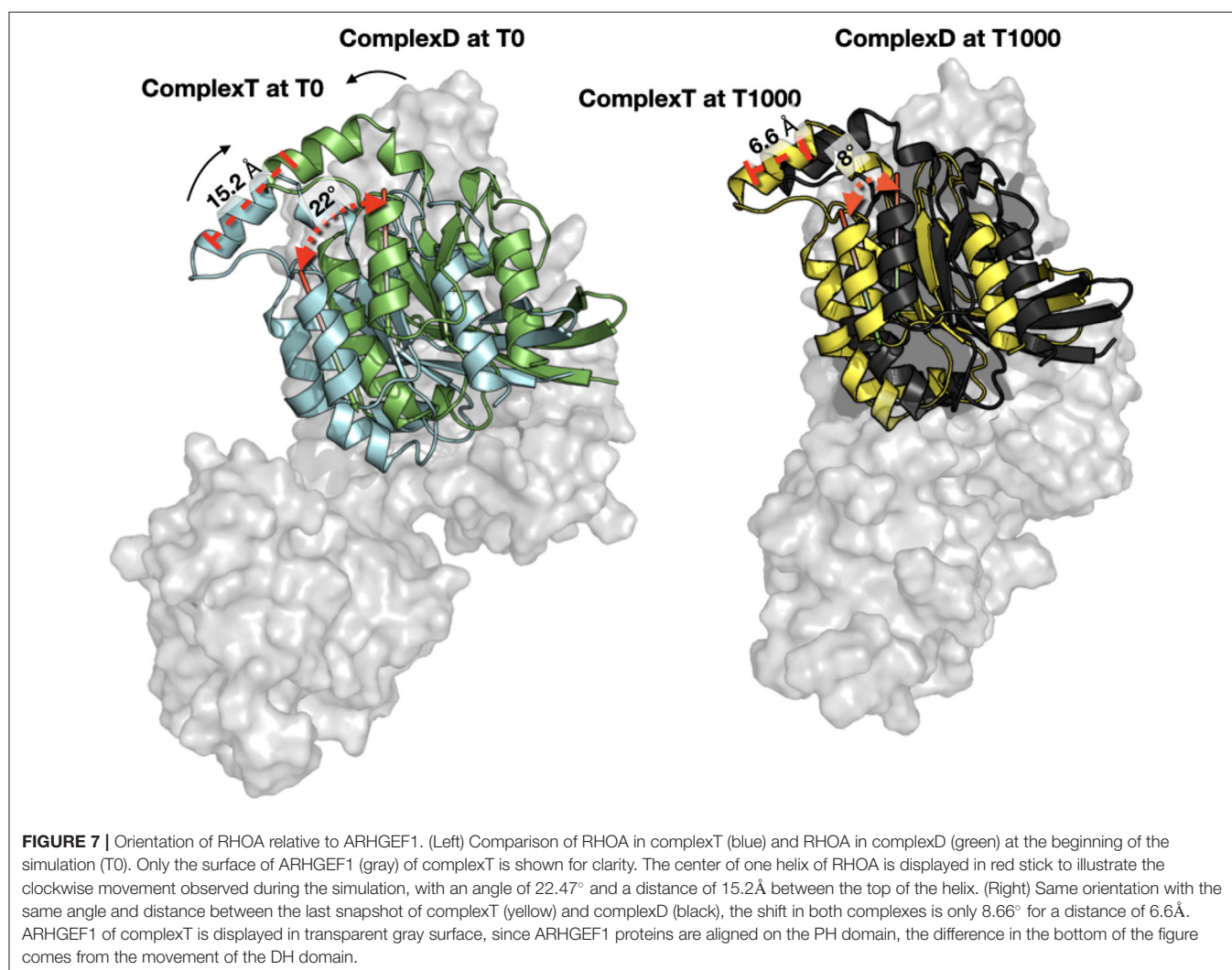
Since no experimental structure of RHOA-ARHGEF1 was available from X-ray studies, NMR, or EM studies, we had to model it. Using the protein sequences of Rho and GEF families, and the existing protein structures of bound homologs RHOA-ARHGEF8 (Petit et al., 2018), RHOA-ARHGEF11 (Bielnicki et al., 2011), RHOA-ARHGEF12 (Kristelly et al., 2004), and RHOA-ARHGEF25 (Lutz et al., 2007), we analyzed to determine which amino acids were shared at the interface of the complex (Figure 3). This sequence and structure-based information was important to assess the validity of our models.

4.1. Initial Models of RHOA-ARHGEF1 Complex

The prediction of protein-protein interface using docking methods is still an important field of research (Smith and Sternberg, 2002; Lensink et al., 2007) but the predictive power of these methods greatly varies depending on the protein families (Bendell et al., 2014; Wang et al., 2017). As no GEF

or RHOA experimental structure was used as target to assess the methods in recent CASP experiments, we benchmarked how these methods could perform on our specific case, using re-docking and cross-docking experiments. We selected ZDOCK after a careful quantification and inspection of the re-docking/cross-docking experiments since its results were the most robust across most predictions and in agreement with our sequence+structure derived data. The best model (**complexD**) selected from ZDOCK contains 5 out of 10 shared amino acids in RHOA and 3 out of 4 shared amino acids from ARHGEF1, with an interface surface area of 2,830 Å².

As the docking experiments are time-consuming and contain also uncertainties, we did also a more crude approach using PyMOL. We analyzed the existing structures of RHOA bound to other members of the ARHGEF family to find the best starting template for our structural comparison. A superimposition of ARHGEF1 (3ODO) on RHOA-ARHGEF11 (3T06) was then performed in PyMOL and further refined using Chiron (Ramachandran et al., 2011). This modeled interface



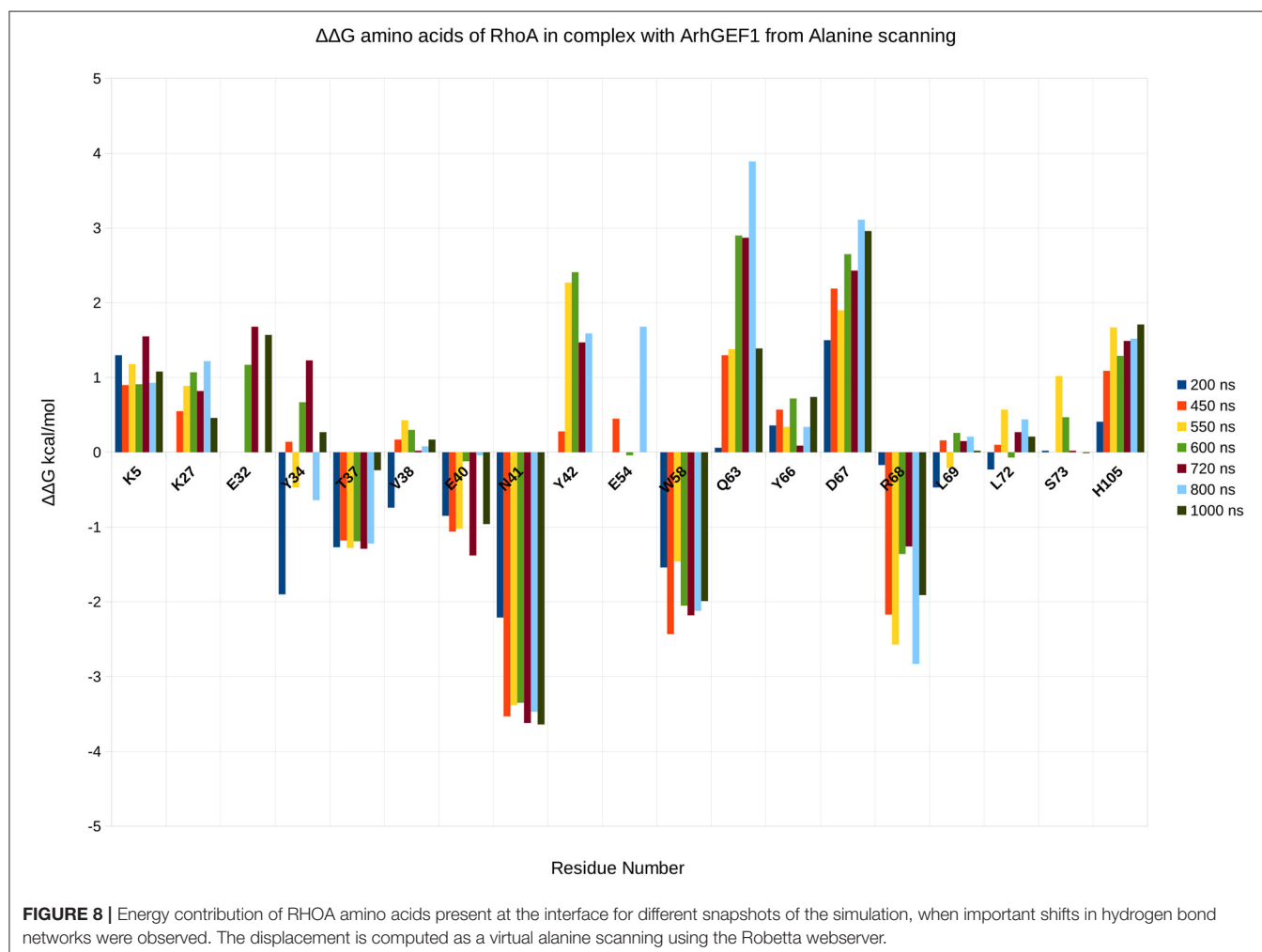
(complexT) allowed to correctly find the position of 2 out of 10 amino acids in RHOA and 2 out of 4 amino acids from ARHGEF1.

Both methods allowed to define comparable starting complexes of the RHOA-ARHGEF1 interface from rigid templates. Major steric clashes were carefully examined using Chiron (Ramachandran et al., 2011) and visual inspection, but no further amino acids adjustment was required. The RMSD difference between both models is 0.4 Å for ARHGEF1 and 9.5 Å for RHOA. This larger difference in RHOA position comes from an alternate orientation of the protein relative to ARHGEF1 with a clockwise rotation of 22° between RHOA in complexT and RHOA in complexD (Figure 7). This alternative positioning of RHOA in comparison to other members of the GEF family is also present in crystallographic structures. Both complexT and complexD seemed therefore reasonable starting complexes, with a comparable building time of 1 day for both protocols: instant for PyMOL superimposition plus 1 day for removal of clashes in Chiron and 1 day for ZDOCK prediction.

4.2. Molecular Dynamics Interface Refinement

A classical method to enhance protein models is molecular dynamics simulations (MD) (Mirjalili et al., 2014). We performed MD on complexT and complexD for 1 μs each. During this simulation of the complexT, the interface area in the complex increased (3,480 Å²) in comparison to the initial complex (2,949 Å²). We identified amino acids conserved in all RHOA-ARHGEF complexes by combining structural sequence analysis (Figure 3A). These contacts are stable throughout the simulation (R5, R68/E544, D556, N603, and D611). Interestingly new contacts are observed K27-E423, R68-D611 led by local rearrangements of amino acids, in particular K27, Y34 (RHOA), and E423 (ARHGEF1) (Figure 6).

Both complexT and complexD lead to a similar RHOA-ARHGEF1 interface at the end of the simulation. At the beginning of the simulation ($t = 0$ ns), ComplexD and complexT only have a global RMSD difference of 1.4 Å between them, but at the end ($t = 1,000$ ns), the RMSD rises to 3.7 Å. Since the RMSD is a global measure of the movements, i.e., both proteins moved, it



is important to understand how proteins evolved independently during the simulations.

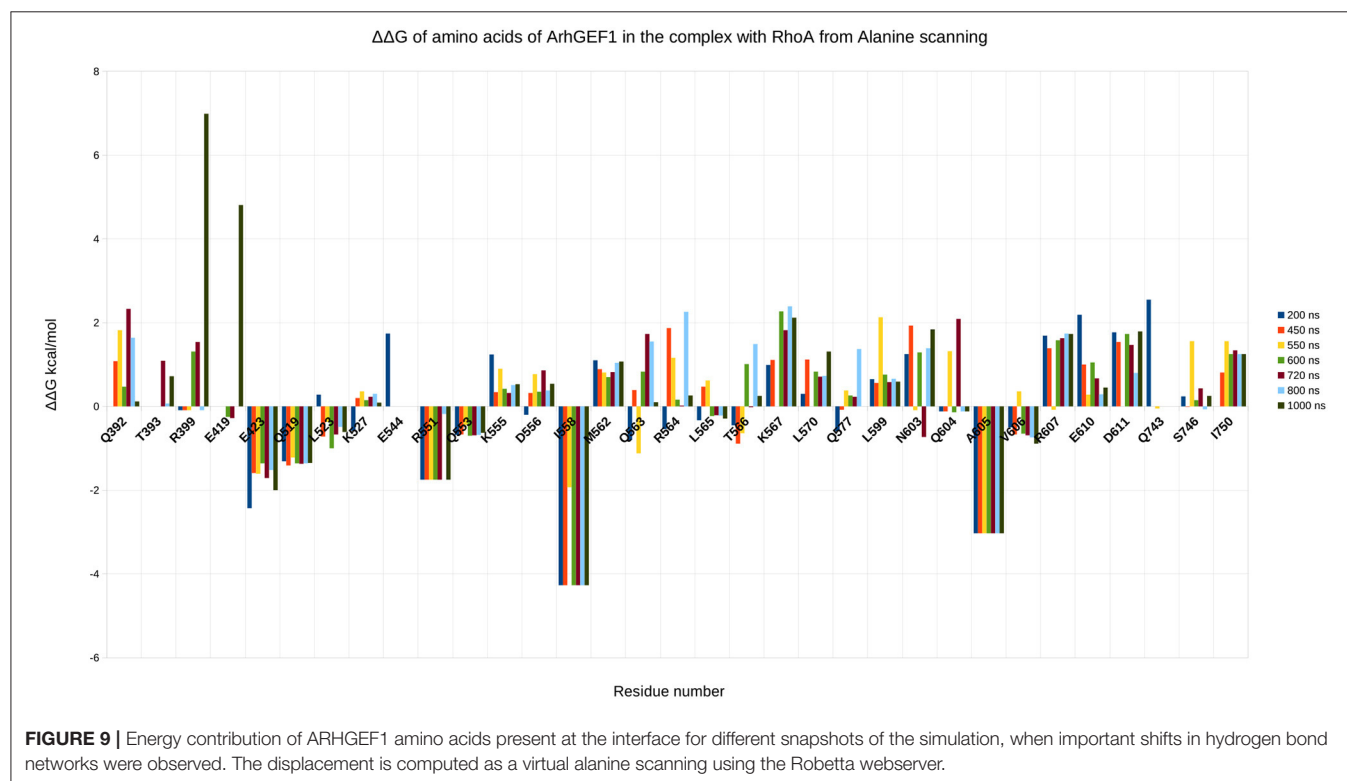
When each trajectory is taken individually, we observe that ComplexD moved more (5.1 Å) from its initial structure than complexT (2.9 Å). This apparent difference comes mostly from a larger movement of the PH domain in the complexD simulation since the RMSD between the initial structure and the end of the simulation concerning only the ARHGEF1 protein is 18 Å. This apparently large difference in ARHGEF1 position for complexD is the consequence of two movements: (i) the local rearrangements of the DH domain (the core RHOA binding domain of GEFs), which is similar in both complexD and complexT (6 Å), and (ii) a larger movement of the PH domain which may be involved in the nucleotide exchange. When aligning the DH domain in both trajectories, we, therefore, see a more important rotation of RHOA relative to ARHGEF1 for complexD (6 Å) than for complexT (3.5 Å) as illustrated in **Figure 7**. The main interface enhancements thus appear locally at the ARHGEF1-RHOA interface, mostly on the DH domain of ARHGEF1, and more globally with a clockwise (+22°) rotation in the complexT, and a slighter anti-clockwise (−3°) rotation in complexD.

Both complexes are refined after molecular dynamics simulations, many important amino acids saw an increase in contact frequency and the position of RHOA relative to ARHGEF1, either inherited from the superposition onto ARHGEF1 or from the docking studies with ZDOCK, led to a strong convergence of the interaction. This study confirms

the interest of using molecular dynamics simulation to increase model quality.

4.3. Validation of the Binding Mode

To validate the binding mode from the simulations, we used the interactions as a starting point and analyzed specific interaction of this complex found in the literature. We found back couples of interactions, which have a lifetime of over 15%, are conserved in all RHOA/GEFs binding modes, with some interactions specific to the RHOA-ARHGEF1 interface. We observed that the complex tends to go toward a more stable conformation when the PH domain moves to enclose RHOA, with an increase in the number of interactions, with a mean SAS going over 3,100 Å², the mean surface area for all complexes of RHOA/GEFs available so far. We could identify some specific contacts for RHOA-ARHGEF1 from the complexT, namely D59-K567, Q63-T566, which were not described previously (Hoffman and Cerione, 2002; Derewenda et al., 2004). In the complexD, the specific contacts E97-S746 has already been described by Gasmi-seabrook (Gasmi-Seabrook et al., 2010) as an essential contact in the nucleotide exchange for PDZ-ARHGEF1 and RHOA. This contact is not observed in the complexT simulation. Starting from two complexes built with different strategies, we were able to have a perfect compatibility between experimental predictions and our *in silico* methods. The identified additional contacts, specific for the RHOA-ARHGEF1, will require experimental exploration since there were differences from the two simulations, potentially coming for the overall dynamics of the interface. To guide experimental validations, we



studied *via* virtual alanine scanning if some amino acids could be qualified as hotspots (Kortemme et al., 2004; Jiang et al., 2017) of the interface (**Figures 8, 9**). Only three amino acids seem to contribute strongly to the binding of both proteins: (1) N41 for RHOA, already identified by multiple sequence alignments and structure comparison, (2,3) I558 and A605 in ARHGEF1. These residues seem to be robustly involved in the interface during all the simulations, either starting with the complexD or the complexT.

4.4. Selection of the Best Model

The knowledge acquired with our strategies helped us to understand the most relevant elements for the binding of the two partners altogether with insights for selecting/computing relatively good refined models. In the initial models after minimization for complexT, there were already 5 over the 10 conserved (E423, Q563, and N603) contacts and for the complexD 4 over 10 (mostly with E423). After simulation for complexT, we can see 8 over 10 conserved contacts and for complexD, there are 6 over 10 contacts during a short time frame where the interface SAS is the highest. Some contacts are seen only thanks to the simulations and one question rises, what is more important to select for qualifying the best model? Its higher number of contacts or the presence

of conserved/important contacts? In our case, both models show conserved/important contacts and new contacts specific to each model. The amino acids detected as hotspots are not conclusive since they are present in both trajectories. During the simulations, even if starting from somewhat different structures, the interaction between ARHGEF1 and RHOA converges. Our model building strategy clearly indicates that a molecular dynamics simulation, starting from rationally designed PPI, improves the initial models.

4.5. Comparison of the MD Model With Information-Driven HADDOCK Docking

HADDOCK is very efficient for protein-protein docking when experimental/bioinformatics constraints can be added for driving the docking. As we had determined the important residues for the binding interface, we used them in HADDOCK webserver first for redocking experiments on the 3T06 crystal structure as shown for other methods in **Table 2**, and obtained a RMSD of 0.75 Å with 90% of the structure in the first same cluster, better than all other protein-protein docking methods. We then did the prediction of the interaction model using ARHGEF1 from 3ODO and RHOA from 3T06. This prediction gave us nine clusters, and after careful analysis only two seemed to have the expected binding mode: cluster 1 and cluster 5. When

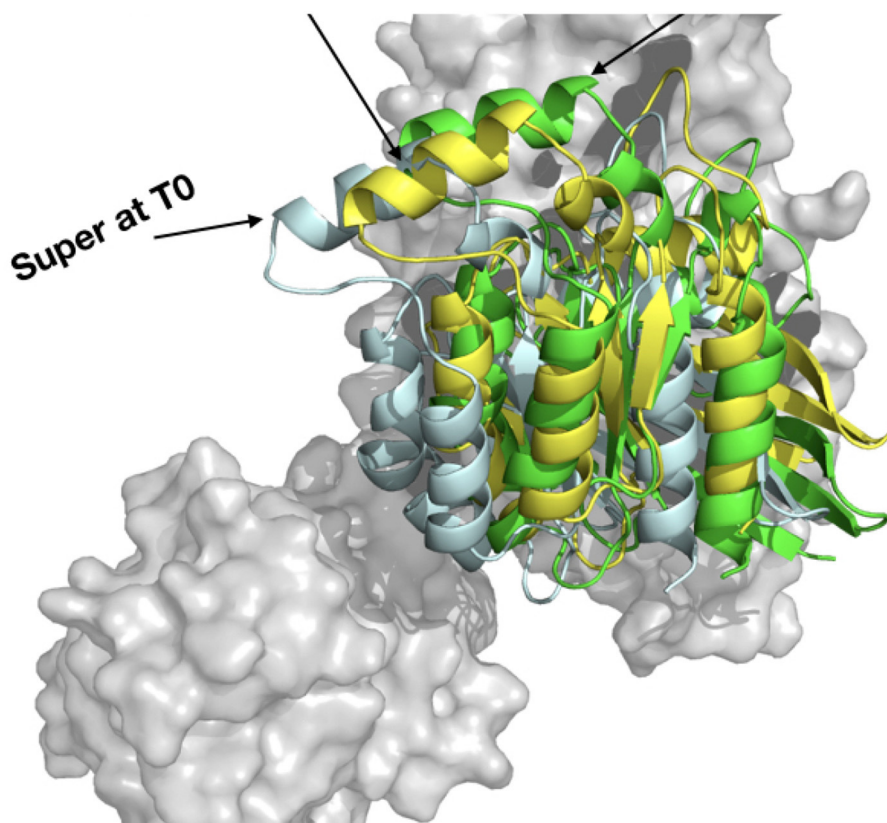


FIGURE 10 | Comparison of the HADDOCK webserver cluster 1 model built from 3T06 RHOA and 3ODO ARHGEF1 (yellow), complexT (light cyan), and complexD (green).

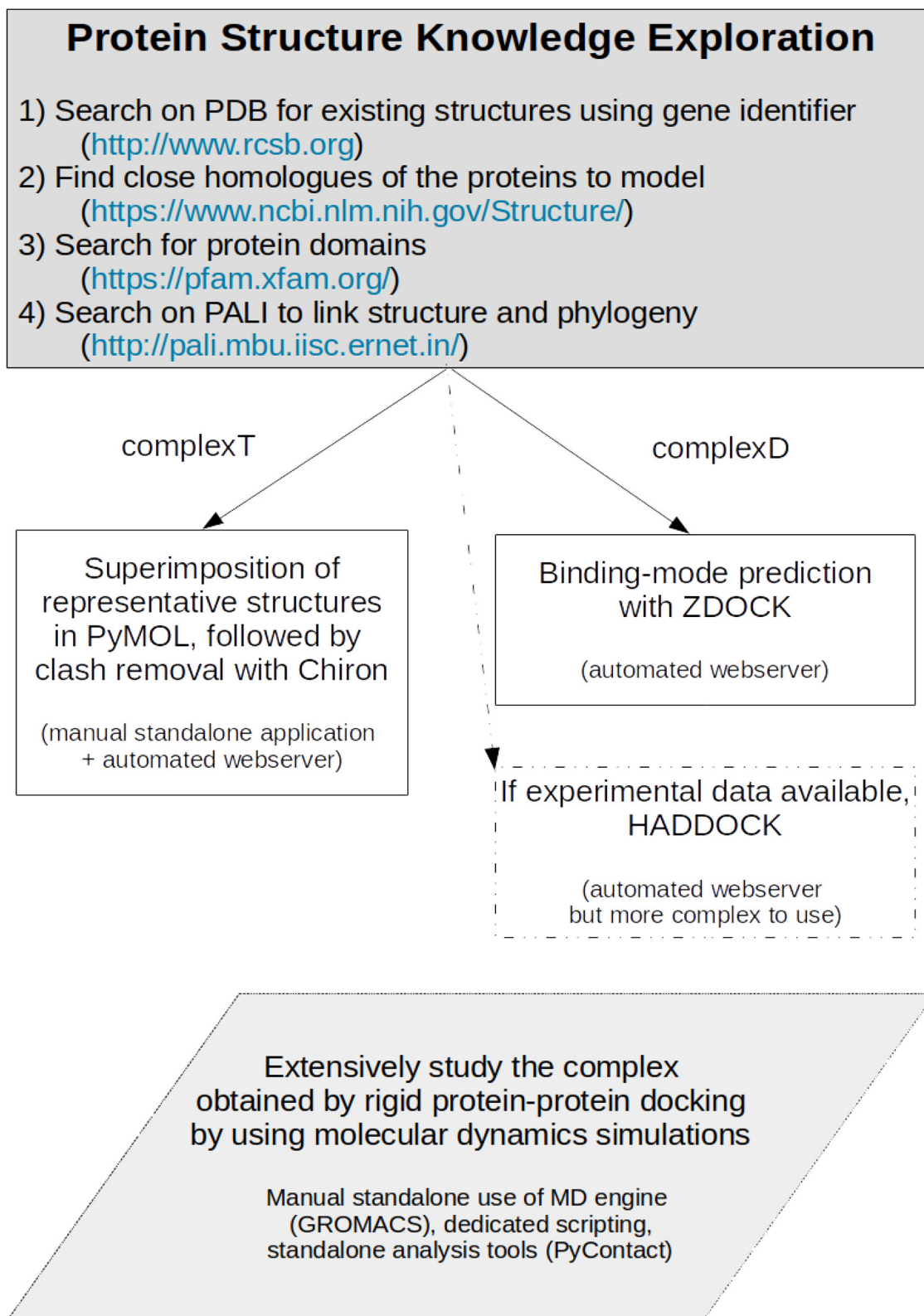


FIGURE 11 | Protocol for producing a protein-protein complex where close homologs of the proteins of interest can be identified.

cluster 1 is compared to ZDOCK's derived complexD (without information driven construction), this cluster 1 has a RMSD of 0.62 Å, also very close to complexT with a RMSD value of 0.65 Å (**Figure 10**). Qualitatively, the cluster1 model displays 6 over 10 of the contacts given as input. If experimental data are available, for instance, coming from mutagenesis experiments, HADDOCK allows their incorporation to guide the binding mode. In this situation, HADDOCK is certainly the best strategy to build a PPI, provided these data can be transformed in sufficient constraints as input. However, the resulting model provided by HADDOCK in 1 day compared to other docking methods is very interesting. The interface area for cluster 1 is 1,478 Å², slightly better than complexD model before refinement (1,420 Å²), but far from the refined interface obtained after molecular dynamics simulations (>3,000 Å²). Even if it is possible to build a reliable model by integrating various data in HADDOCK, a long molecular dynamics simulation, with a simulation time above 250 ns is still required to enhance the quality of a PPI (Feig and Mirjalili, 2016).

5. CONCLUSION

Most biological processes involve transient protein-protein interactions, in particular for cellular signaling. The RHOA-ARHGEF1 interaction is responsible for the activation of RHOA downstream of type 1 angiotensin II receptor signaling in vascular smooth muscle cells, thereby controlling vascular tone and blood pressure (Loirand, 2015).

Our study aims at exemplifying how one can model a protein-protein interaction when sufficient experimental structures are present, but only experimental data for close homologs are available. We set up two different strategies summarized in **Figure 11**. One has first to identify the close homologs. If the members of the family have a standardized name, they should rapidly be identified directly in the Protein Data Bank (Berman et al., 2003). If not, a search on the National Center for Biotechnology Information (NCBI) structure service, in the Protein Families Database (Mistry et al., 2021), or in the PALI database (Balaji et al., 2001) should help in finding close homologs. If not, the protocols described in our work should be considered with caution. When the close homologs are identified, it is possible to apply the protocols previously described. The first one, based on structural superimposition of partners, allows a rapid building of the complex, but provides required local adjustments to avoid steric clashes. We expect it to be useful for a preliminary study of how the proteins interact. The second strategy based on most advanced methods combining the search of the best binding mode *via* the assessment of the results of protein-protein docking, followed by the refinement of the best docked model using molecular dynamics simulations. This model showed not only increased shape complementarity and

increased contacts but also provides insights into the dynamics of the detailed amino acids interactions between the partners. This more advanced strategy is probably only accessible to experts and should only be required for atomic-level analysis and mechanistic studies. In our study, both strategies gave close initial models, but we do not expect the results on RHOA-ARHGEF1 to be amenable for general purpose. We, therefore, recommend to use a protein-protein rigid-body docking study (complexD) for producing the initial interaction mode. In our study, ZDOCK was better if precision, robustness, and time are taken altogether into consideration. When possible, we recommend to perform long molecular dynamics simulations to enhance the network of interaction between both proteins and to get a better overview of the lifetime of each interaction. More generally, we expect these strategies will be successfully applied to a variety of targets where a partial structural coverage of both partners is known, provided the complex to model has characteristics comparable with the two proteins described in this article.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ST and GL obtained the funding. EG did the experiments and analysis. BO and GL supervised the work. EG and ST wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was part of the PIRAMID project funded by the French Regional Council of Pays-de-la-Loire and the TROPIC project, supported by ANR grant Programme d'Investissements d'Avenir (ANR-16-IDEX-0007), the French Regional Council of Pays-de-la-Loire, and Nantes Métropole.

ACKNOWLEDGMENTS

The authors thank the Center de Calcul Intensif des Pays de la Loire (CCIPL) for computational resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.643728/full#supplementary-material>

REFERENCES

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi: 10.1016/j.softx.2015.06.001
- Arrazola Sastre, A., Luque Montoro, M., Gálvez-Martín, P., Lacerda, H. M., Lucia, A., Llaverro, F., et al. (2020). Small GTPases of the Ras and Rho families

- switch on/off signaling pathways in neurodegenerative diseases. *Int. J. Mol. Sci.* 21:6312. doi: 10.3390/ijms21176312
- Balaji, S., Sujatha, S., Kumar, S. S. C., and Srinivasan, N. (2001). PALI—a database of Phylogeny and ALIGNment of homologous protein structures. *Nucleic Acids Res.* 29, 61–65. doi: 10.1093/nar/29.1.61
- Basse, M. J., Betzi, S., Morelli, X., and Roche, P. (2016). 2P2idb v2: update of a structural database dedicated to orthosteric modulation of protein–protein interactions. *Database* 2016:baw007. doi: 10.1093/database/baw007
- Bendell, C. J., Liu, S., Aumentado-Armstrong, T., Istrate, B., Cernek, P. T., Khan, S., et al. (2014). Transient protein–protein interface prediction: datasets, features, algorithms, and the RAD-T predictor. *BMC Bioinformatics* 15:82. doi: 10.1186/1471-2105-15-82
- Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* 10, 980–980. doi: 10.1038/nsb1203-980
- Bielnicki, J. A., Shkumatov, A. V., Derewenda, U., Somlyo, A. V., Svergun, D. I., and Derewenda, Z. S. (2011). Insights into the molecular activation mechanism of the RhoA-specific guanine nucleotide exchange factor, PDZRhoGEF. *J. Biol. Chem.* 286, 35163–35175. doi: 10.1074/jbc.M111.270918
- Chen, Z., Guo, L., Sprang, S. R., and Sternweis, P. C. (2011). Modulation of a GEF switch: autoinhibition of the intrinsic guanine nucleotide exchange activity of p115-RhoGEF. *Prot. Sci.* 20, 107–117. doi: 10.1002/pro.542
- Cheng, T. M. K., Blundell, T. L., and Fernandez-Recio, J. (2007). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins* 68, 503–515. doi: 10.1002/prot.21419
- Cherfils, J., and Zeghouf, M. (2011). Chronicles of the GTPase switch. *Nat. Chem. Biol.* 7, 493–495. doi: 10.1038/nchembio.608
- Cherfils, J., and Zeghouf, M. (2013). Regulation of small GTPases by GEFs, GAPs, and GDIs. *Physiol. Rev.* 93, 269–309. doi: 10.1152/physrev.00003.2012
- Chéron, J. B., Zacharias, M., Antonczak, S., and Fiorucci, S. (2017). Update of the ATTRACT force field for the prediction of protein–protein binding affinity. *J. Comput. Chem.* 38, 1887–1890. doi: 10.1002/jcc.24836
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163
- da Silva, A. W. S., and Vranken, W. F. (2012). Acypype-antechamber python parser interface. *BMC Res. Notes* 5:367. doi: 10.1186/1756-0500-5-367
- Derewenda, U., Oleksy, A., Stevenson, A. S., Karczyska, J., Dauter, Z., Somlyo, A. P., et al. (2004). The crystal structure of RhoA in complex with the DH/PH fragment of PDZRhoGEF, an activator of the Ca²⁺ sensitization pathway in smooth muscle. *Structure* 12, 1955–1965. doi: 10.1016/j.str.2004.09.003
- Dvorsky, R., and Ahmadian, M. R. (2004). Always look on the bright side of Rho: structural implications for a conserved intermolecular interface. *EMBO Rep.* 5, 1130–1136. doi: 10.1038/sj.embor.7400293
- Feig, M., and Mirjalili, V. (2016). Protein structure refinement via molecular-dynamics simulations: what works and what does not? *Proteins* 84, 282–292. doi: 10.1002/prot.24871
- Felline, A., Belmonte, L., Raimondi, F., Bellucci, L., and Fanelli, F. (2019). Interconnecting flexibility, structural communication, and function in RhoGEF oncoproteins. *J. Chem. Inform. Model.* 59, 4300–4313. doi: 10.1021/acs.jcim.9b00271
- Gabb, H. A., Jackson, R. M., and Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* 272, 106–120. doi: 10.1006/jmbi.1997.1203
- Gasmi-Seabrook, G. M. C., Marshall, C. B., Cheung, M., Kim, B., Wang, F., Jang, Y. J., et al. (2010). Real-time NMR study of guanine nucleotide exchange and activation of RhoA by PDZ-RhoGEF. *J. Biol. Chem.* 285, 5137–5145. doi: 10.1074/jbc.M109.064691
- Guilluy, C. (2010). The Rho exchange factor Arhgef1 mediates the effects of angiotensin II on vascular tone and blood pressure. *Nat. Med.* 16:9. doi: 10.1038/nm.2079
- Hoffman, G. R., and Cerione, R. A. (2002). Signaling to the Rho GTPases: networking with the DH domain. *FEBS Lett.* 513, 85–91. doi: 10.1016/S0014-5793(01)03310-5
- 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
- Ihara, K., Muraguchi, S., Kato, M., Shimizu, T., Shirakawa, M., Kuroda, S., et al. (1998). Crystal structure of human RhoA in a dominantly active form complexed with a GTP analogue. *J. Biol. Chem.* 273, 9656–9666. doi: 10.1074/jbc.273.16.9656
- Jiang, J., Wang, N., Chen, P., Zheng, C., and Wang, B. (2017). Prediction of protein hotspots from whole protein sequences by a random projection ensemble system. *Int. J. Mol. Sci.* 18:1543. doi: 10.3390/ijms18071543
- Jiménez-García, B., Pons, C., and Fernández-Recio, J. (2013). pyDockWEB: a web server for rigid-body protein–protein docking using electrostatics and desolvation scoring. *Bioinformatics* 29, 1698–1699. doi: 10.1093/bioinformatics/btt262
- 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
- Kortemme, T., Kim, D. E., and Baker, D. (2004). Computational alanine scanning of protein–protein interfaces. *Sci.* 2004:pl2. doi: 10.1126/stke.2192004pl2
- Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., et al. (2017). The ClusPro web server for protein–protein docking. *Nat. Protoc.* 12, 255–278. doi: 10.1038/nprot.2016.169
- Krissinel, E. (2010). Crystal contacts as nature's docking solutions. *J. Comput. Chem.* 31, 133–143. doi: 10.1002/jcc.21303
- Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* 372, 774–797. doi: 10.1016/j.jmb.2007.05.022
- Kristelly, R., Gao, G., and Tesmer, J. J. G. (2004). Structural determinants of RhoA binding and nucleotide exchange in leukemia-associated Rho guanine-nucleotide exchange factor. *J. Biol. Chem.* 279, 47352–47362. doi: 10.1074/jbc.M406056200
- Lensink, M. F., Brysbaert, G., Nadzirin, N., Velankar, S., Chaleil, R. A. G., Gerguri, T., et al. (2019). Blind prediction of homo- and hetero-protein complexes: the CASP13-CAPRI experiment. *Proteins* 87, 1200–1221. doi: 10.1002/prot.25838
- Lensink, M. F., Méndez, R., and Wodak, S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins* 69, 704–718. doi: 10.1002/prot.21804
- Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C., and Wodak, S. J. (2018). The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins* 86, 257–273. doi: 10.1002/prot.25419
- 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* 78, 1950–1958. doi: 10.1002/prot.22711
- Loirand, G. (2015). Rho kinases in health and disease: from basic science to translational research. *Pharmacol. Rev.* 67, 1074–1095. doi: 10.1124/pr.115.010595
- Loirand, G., and Pacaud, P. (2010). The role of Rho protein signaling in hypertension. *Nat. Rev. Cardiol.* 7, 637–647. doi: 10.1038/nrcardio.2010.136
- Luigia, C. M., Jérémy, B., Nabila, D., Gilliane, C., Anne, B., Michel, A., et al. (2015). Angiotensin II activates the RhoA exchange factor Arhgef1 in humans. *Hypertension* 65, 1273–1278. doi: 10.1161/HYPERTENSIONAHA.114.05065
- Lutz, S., Shankaranarayanan, A., Coco, C., Ridilla, M., Nance, M. R., Vettel, C., et al. (2007). Structure of goq-p63RhoGEF-RhoA complex reveals a pathway for the activation of RhoA by GPCRs. *Science* 318, 1923–1927. doi: 10.1126/science.1147554
- Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268
- Margreiter, C., and Oostenbrink, C. (2017). MDplot: visualise molecular dynamics. *R J.* 9, 164–186. doi: 10.32614/RJ-2017-007
- Mirjalili, V., Noyes, K., and Feig, M. (2014). Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins* 82, 196–207. doi: 10.1002/prot.24336
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913
- Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nat. Methods* 10, 47–53. doi: 10.1038/nmeth.2289
- Ottmann, C. (2015). “New compound classes: protein–protein interactions,” in *New Approaches to Drug Discovery*, Vol. 232, eds U. Nielsch, U. Fuhrmann, and S. Jaroch (Cham: Springer International Publishing), 125–138. doi: 10.1007/164_2015_30

- Petit, A. P., Garcia-Petit, C., Bueren-Calabuig, J. A., Vuillard, L. M., Ferry, G., and Boutin, J. A. (2018). A structural study of the complex between neuroepithelial cell transforming gene 1 (Net1) and RhoA reveals a potential anticancer drug hot spot. *J. Biol. Chem.* 293, 9064–9077. doi: 10.1074/jbc.RA117.001123
- Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T., and Weng, Z. (2014). ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* 30, 1771–1773. doi: 10.1093/bioinformatics/btu097
- Prieto-Dominguez, N., Parnell, C., and Teng, Y. (2019). Drugging the small GTPase pathways in cancer treatment: promises and challenges. *Cells* 8:255. doi: 10.3390/cells8030255
- Ramachandran, S., Kota, P., Ding, F., and Dokholyan, N. V. (2011). Automated minimization of steric clashes in protein structures. *Proteins* 79, 261–270. doi: 10.1002/prot.22879
- Rolland, T., Tasan, M., Charlotiaux, B., Pevzner, S., Zhong, Q., Sahni, N., et al. (2014). A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226. doi: 10.1016/j.cell.2014.10.050
- Scheurer, M., Rodenkirch, P., Siggel, M., Bernardi, R. C., Schulten, K., Tajkhorshid, E., et al. (2018). PyContact: rapid, customizable, and visual analysis of noncovalent interactions in MD simulations. *Biophys. J.* 114, 577–583. doi: 10.1016/j.bpj.2017.12.003
- Skwarczynska, M., and Ottmann, C. (2015). Protein–protein interactions as drug targets. *Future Med. Chem.* 7, 2195–2219. doi: 10.4155/fmc.15.138
- Smith, G. R., and Sternberg, M. J. (2002). Prediction of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.* 12, 28–35. doi: 10.1016/S0959-440X(02)00285-3
- Sukhwal, A., and Sowdhamini, R. (2013). Oligomerisation status and evolutionary conservation of interfaces of protein structural domain superfamilies. *Mol. Biosyst.* 9, 1652–1661. doi: 10.1039/c3mb25484d
- Takemura, K., and Kitao, A. (2019). More efficient screening of protein-protein complex model structures for reducing the number of candidates. *Biophys. Physicobiol.* 16, 295–303. doi: 10.2142/biophysico.16.0_295
- van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastiris, P. L., Karaca, E., et al. (2016). The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* 428, 720–725. doi: 10.1016/j.jmb.2015.09.014
- Vetter, I. R. (2014). “The structure of the G domain of the Ras superfamily,” in *Ras Superfamily Small G Proteins: Biology and Mechanisms*, Vol. 1, ed A. Wittinghofer (Vienna: Springer Vienna), 25–50. doi: 10.1007/978-3-7091-1806-1_2
- Wang, W., Yang, Y., Yin, J., and Gong, X. (2017). Different protein-protein interface patterns predicted by different machine learning methods. *Sci. Rep.* 7:16023. doi: 10.1038/s41598-017-16397-z
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033
- Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: more and better reference data for improved all-atom structure validation. *Prot. Sci.* 27, 293–315. doi: 10.1002/pro.3330

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

Copyright © 2021 Gheyouche, Bagueneau, Loirand, Offmann and Téletchéa. 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-Site Models of *Streptococcus pyogenes* Cas9 in DNA Cleavage State

Honghai Tang¹, Hui Yuan¹, Wenhao Du¹, Gan Li¹, Dongmei Xue¹ and Qiang Huang^{1,2*}

¹ State Key Laboratory of Genetic Engineering, Shanghai Engineering Research Center of Industrial Microorganisms, Ministry of Education Engineering Research Centre of Gene Technology, School of Life Sciences, Fudan University, Shanghai, China,

² Multiscale Research Institute of Complex Systems, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:

Agnel Praveen Joseph,
Science and Technology Facilities
Council, United Kingdom

Reviewed by:

Kai Tittmann,
University of Göttingen, Germany
Riccardo Miggiano,
University of Eastern Piedmont, Italy

*Correspondence:

Qiang Huang
huangqiang@fudan.edu.cn

Specialty section:

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

Received: 14 January 2021

Accepted: 22 March 2021

Published: 21 April 2021

Citation:

Tang H, Yuan H, Du W, Li G,
Xue D and Huang Q (2021)
Active-Site Models of *Streptococcus*
pyogenes Cas9 in DNA Cleavage
State. *Front. Mol. Biosci.* 8:653262.
doi: 10.3389/fmolb.2021.653262

CRISPR-Cas9 is a powerful tool for target genome editing in living cells. Significant advances have been made to understand how this system cleaves target DNA. However, due to difficulty in determining active CRISPR-Cas9 structure in DNA cleavage state by X-ray and cryo-EM, it remains uncertain how the HNH and RuvC nuclease domains in CRISPR-Cas9 split the DNA phosphodiester bonds with metal ions and water molecules. Therefore, based on one- and two-metal-ion mechanisms, homology modeling and molecular dynamics simulation (MD) are suitable tools for building an atomic model of Cas9 in the DNA cleavage state. Here, by modeling and MD, we presented an atomic model of SpCas9-sgRNA-DNA complex with the cleavage state. This model shows that the HNH and RuvC conformations resemble their DNA cleavage state where the active-sites in the complex coordinate with DNA, Mg²⁺ ions and water. Among them, residues D10, E762, H983 and D986 locate at the first shell of the RuvC active-site and interact with the ions directly, residues H982 or/and H985 are general (Lewis) bases, and the coordinated water is located at the positions for nucleophilic attack of the scissile phosphate. Meanwhile, this catalytic model led us to engineer new SpCas9 variant (SpCas9-H982A + H983D) with reduced off-target effects. Thus, our study provides new mechanistic insights into the CRISPR-Cas9 system in the DNA cleavage state, and offers useful guidance for engineering new CRISPR-Cas9 editing systems with improved specificity.

Keywords: Gene editing, CRISPR-Cas9, DNA cleavage mechanism, Molecular dynamics, off-target effects

INTRODUCTION

The RNA-guided CRISPR-Cas9 nuclease from *Streptococcus pyogenes* (SpCas9) has been widely used as a powerful and versatile tool for genome engineering (Hsu et al., 2014; Wang et al., 2016; Chen and Doudna, 2017). Guided by a pre-designed sgRNA with a 20-base long sequence for targeting DNA, SpCas9 could cleave corresponding complementary sequences in the genome through RNA: DNA hybridization (Jinek et al., 2012; Cong et al., 2013). Many studies have demonstrated that this cleavage process includes: first, the Recognition lobe (REC) domains of SpCas9 interact with the common sgRNA scaffold to form an Ribonucleoprotein (RNP) complex for recognizing the N is any one of bases (Adenine, Thymine, Cytosine, or Guanine), G is Guanine (NGG) motif just downstream the 20-mer targeting sequence, and then mediates the formation of the RNA: DNA heteroduplex; next, the HNH and RuvC endonuclease domains catalyze the hydrolysis of two

DNA phosphodiester bonds in the complementary and non-complementary strands, respectively (Anders et al., 2014; Jinek et al., 2014; Nishimasu et al., 2014; Jiang et al., 2015, 2016). Although significant advances have been made in the past years, a complete understanding of the DNA hydrolysis mechanisms of the HNH and RuvC active sites, especially those of the RuvC domain is still lacking.

Structural and biochemical studies have suggested that the catalytic residues in the HNH active sites are D839, H840, and N863, which may employ the one-metal-ion mechanism to hydrolyze the complementary strand of the target DNA (Jinek et al., 2012; Nishimasu et al., 2014), in agreement with recent research (Zhu et al., 2019). Also, some studies reported that D10, E762, H983, and D986 are the catalytic residues of the RuvC domain, and then suggested that they utilize the two-metal-ion hydrolysis mechanism to split the non-complementary (non-target) strand (De Vivo et al., 2008; Ho et al., 2010; Nishimasu et al., 2014; Palermo et al., 2015; Chen and Doudna, 2017). However, unlike those of the HNH residues, the precise roles of these RuvC residues in the DNA cleavage remain debated. For example, for the HNH domain, it has been firmly established that H840 acts as the general (Lewis) base to activate the water molecule for nucleophilic attack of the scissile phosphate (Nishimasu et al., 2014). For the RuvC domain, in contrast, several views considered that H983 is the general base for the DNA hydrolysis (Zuo and Liu, 2016; Chen and Doudna, 2017), and others proposed that this residue plays a role in coordinating the metal ion (Fernandes et al., 2016). To clarify such mechanistic problems, direct structural information about the wild-type HNH and RuvC active-sites in the DNA cleavage state is critical. So it becomes very important to obtain the cleavage-activating structure of SpCas9 in complex with sgRNA and the target DNA.

Unfortunately, similar to structural studies of many enzymes, it is currently difficult to use the wild-type, active SpCas9 protein to form a stable complex with a full-length target DNA; alternatively, activity-dead mutants were often used for the structural determination (Anders et al., 2014; Jinek et al., 2014; Nishimasu et al., 2014; Jiang et al., 2016; Huai et al., 2017). Probably due to such a technical limitation, available x-ray crystal structures of the mutants were often resolved in the cleavage-inactive states, in which the HNH active-site was found far from the scissile bond of the complementary strand (>13 Å) (Nishimasu et al., 2014; Jiang et al., 2016). Also, in most of the resolved structures, certain numbers of nucleotides from the 20-mer non-target sequence are missing (Anders et al., 2014; Jiang et al., 2016). These made it difficult to build atomic models in the DNA cleavage state. To address this problem, we previously used cryo-EM to capture the active conformations of SpCas9, and obtained a structure of the SpCas9–sgRNA–DNA ternary complex in which the HNH active-site is nearest to the split bond of the complementary strand among all the available structures (Huai et al., 2017). However, this structure did not enable us to accurately build atomic models for the wild-type active-sites with metal ions and water molecules. Even though the current studies have obtained two cryo-EM structures (6O0Y and 6VPC) around 3.3 Å, the atomic models of their active sites in the cleavage (catalytic) state cannot be completely and accurately built due to

their HNH and RuvC domains are not in the cleavage (catalytic) state at the same time: among them, the conformation of the HNH active sites in 6O0Y resembles catalytic state (Zhu et al., 2019), whereas the conformation of the RuvC active sites in 6VPC is in the catalytic state (Zhu et al., 2019; Lapinaite et al., 2020). Thus, how the HNH and RuvC active sites in a sharing structure organize with Mg^{2+} ions and water molecules to hydrolyze the DNA phosphodiester bonds remains open.

Here, we further refined the cryo-EM structure of the SpCas9–sgRNA–DNA ternary complex to be in the DNA cleavage state, and thereby rebuild the atomic model for the active complex by molecular modeling according to one- and two-metal-ion mechanisms. For this, we will present the model of the wild-type HNH and RuvC active-sites in complex with DNA, Mg^{2+} ions, and water molecules, and obtain a refined cryo-EM structure of SpCas9 in DNA cleavage state. Meanwhile, our mechanistic models indicate that the residues D10, E762, H983, and D986 are in the first coordination shell of the RuvC active-site with the ions, suggesting that H983 coordinates with the Mg^{2+} ion, and that the general base for the RuvC catalysis is likely H982 or/and H985.

RESULTS AND DISCUSSION

Atomic Models of Nuclease Active-Sites

SpCas9 can cleave DNA by RNA guiding and has developed a powerful genome-editing tool that is widely used in many fields (Hsu et al., 2014; Wang et al., 2016; Chen and Doudna, 2017). To understand the conformational changes of SpCas9 in the atomic level and its interaction with nucleic acids, researchers used x-ray or cryo-electron microscopy to resolve SpCas9 structures with various forms, including SpCas9 monomer (4CMP) (Jinek et al., 2014), SpCas9–sgRNA binary complex (4ZT0) (Jiang et al., 2015), and SpCas9–sgRNA–DNA ternary complexes (4O08, 4UN3, 5F9R, 5Y36, 6O0Y, and 6VPC) (Anders et al., 2014; Nishimasu et al., 2014; Jiang et al., 2016; Huai et al., 2017; Zhu et al., 2019; Lapinaite et al., 2020; **Supplementary Figure 1**). However, the aforementioned atom models are not all-atom structures of SpCas9 in the DNA shearing-active state and are thereby limitations of our understanding of the precise roles of the critical sites in the SpCas9 catalytic centers, especially the key sites in the RuvC active center. To overcome the difficulties of resolving an active structure with a high resolution for elucidating the precise roles of the key sites in the RuvC catalytic center, based on the currently available structures of SpCas9, through homology modeling, and molecular dynamics simulation, we intended to rebuild a model of SpCas9–sgRNA–DNA ternary complex in the DNA cleavage state from theories.

To screen out suitable atom models as the structural templates for constructing SpCas9 in shearing active state, we aligned and analyzed the protein structures of the SpCas9–sgRNA–DNA ternary complexes (Anders et al., 2014; Nishimasu et al., 2014; Jiang et al., 2016; Huai et al., 2017; Zhu et al., 2019; Lapinaite et al., 2020), and observed that the structural orientations of SpCas9 were highly consistent below the oblique axis, while those above the oblique axis were slightly different

(**Supplementary Figure 2**). This indicated that any single atomic model among SpCas9 structures can also be selected as a template to build the structure of REC1 and PI domains, but the construction of REC2, REC3, HNH, and RuvC domains needs to be analyzed and screened further. Considering the structural characteristics of these three atomic models (5Y36, 6O0Y, and 6VPC) (Huai et al., 2017; Zhu et al., 2019; Lapinaite et al., 2020) in the existing structures, they are the only active structures of SpCas9 so far. Among them, 5Y36 has magnesium ions in the catalytic centers and the atomic structure of the full-length DNA, while 6O0Y and 6VPC displayed the catalytic states of the HNH and RuvC domains in the DNA cleavage, respectively. Therefore, we presumed that SpCas9–sgRNA–DNA ternary complex in the DNA cleavage state might be rebuilt from theories, through selecting these three atomic models as references and using complementary strategies among their structures. First, we aligned these three structures and saved the coordinates of their relative positions; Next, we extracted atomic coordinates of nucleic chains and magnesium ions from 5Y36 (Huai et al., 2017); Subsequently, we transferred the atomic coordinates of REC2 and RuvC domains in 6VPC and the atomic coordinates of REC1, REC3, PI, and HNH domains in 6O0Y to the aforementioned nucleic chains (Zhu et al., 2019; Lapinaite et al., 2020). Finally, these structural fragments were assembled into a complete-initial structure of the SpCas9–sgRNA–DNA ternary complex (**Figure 1A**). To obtain the structure of ternary complex in the active state, the aforementioned ternary complex was accommodated into two cryo-EM densities maps (end-21308.map and end-0584.map, respectively) (Zhu et al., 2019; Lapinaite et al., 2020), which was processed by Situs-3.1 and Rosetta programs (Ashworth and Baker, 2009; Wriggers, 2010). This result showed that secondary structural elements in the ternary complex matched well into the EM density maps (**Supplementary Figure 3**). And in this active structure of the ternary complex, we observed that the locations of the key sites in the HNH and RuvC catalytic centers were consistent with those of 6O0Y and 6VPC (**Supplementary Figure 4**; Zhu et al., 2019; Lapinaite et al., 2020), respectively. In addition, in the two-metal center, the N8 atom of H983 points toward the Mg^{2+} ion A, thus H983 may have a potential ability to stabilize the Mg^{2+} ion A (**Supplementary Figure 5**), rather than a general. Thus these data suggested that we successfully refined the structure of SpCas9 in the active state by the structural-complementary strategy.

In our atom model, we observed that the key sites of HNH and RuvC active centers, including residues D839, N863, E762, H983, and D986, embraced Mg^{2+} ions, consistent with the previous report (Nishimasu et al., 2014). Therefore, the result suggested that these residues are critical sites of the active centers in SpCas9, which is supported by an alanine scan. As shown in **Figure 1B**, alanine (Ala) substitution of D839, N863, E762, H983, or D986 may convert the wild-type SpCas9 into a nickase. Compared with the wild-type (lanes 3 and 8 in **Figure 1B**), the cleavage activities of the mutants D839A and N863A are sharply reduced and disappeared, respectively (lanes 4 and 5 in **Figure 1B**). Likewise, the cleavage activities of the mutants E762A, H983A, and D986A are significantly weakened (lanes 9, 10, and 11 in **Figure 1B**). Thus, these residues are

essential for binding the metal ion and catalyzing the target DNA. However, we did not know whether this model is a catalytic state in DNA cleavage. To obtain the atomic model of SpCas9 in the DNA catalytic cleavage state, based on the one- and two-metal principles (Steitz and Steitz, 1993; Yang, 2008), we fixed magnesium ions and the coordination residues around them to resemble an octahedral symmetry, and performed molecular dynamics (MD) simulation for this. With coordination fixation, we set up coordination distances (near 2.0 Å) between magnesium ions and the corresponding coordination residues in the catalytic centers. After MD, in the HNH catalytic center, we observed that an Mg^{2+} ion is surrounded by D839 and N863, water molecules, and the phosphate group (**Figure 1C**). The Mg^{2+} ion coordinates with the O8 atoms of D839 and N863, the O atoms of water molecules and the phosphate group; and the coordination geometry between the Mg^{2+} ion and the oxygen atoms is a near octahedral symmetry with the ligand-to-metal ion distance between 1.8 and 2.0 Å (**Figure 1C**); these features are almost consistent with those of one-metal-ion model (Yang, 2008). Meanwhile, the N8 atom of H840 in the second structural shell points toward the corresponding phosphate group (**Figure 1C**), and could act as the general base to activate the water molecule for the nucleophilic attack on the phosphodiester bond. Similarly, in the RuvC catalytic center, two Mg^{2+} ions are enclosed by D10, E762, D986, and H983, water molecules, and the phosphate group (**Figure 1D**). In line with the two-metal-ion active-site model (Steitz and Steitz, 1993; Nowotny and Yang, 2006; Yang, 2008), these two Mg^{2+} ions are about 3.2 Å apart, and jointly coordinate with the O8 atoms of D10, E762, and D986, the N8 atom of H983, and the O atoms of water molecules and the phosphate group (**Figure 1D**). The distance between the ligands and corresponding Mg^{2+} ion is about 2.0 Å, which forms two near octahedrons (**Figure 1D**). At the same time, in this active state, the Root mean square deviation (RMSD) of the ternary complex was minor and stable, approximately 3 Å (**Supplementary Figure 6**). Thus, we succeed in refining the atomic model of the SpCas9 in cleavage (catalytic) state and revealing the potential catalysis roles for the key sites of the HNH and RuvC catalytic centers in hydrolyzing DNA.

Catalytic Role of RuvC Residue H983

Previous studies have shown that the active sites of the RNase H [Protein data bank (PDB) ID: 1ZBL] consist of four conserved residues (D71, E109, D132, D192), and coordinate with two metal ions (Nowotny and Yang, 2006). Meanwhile, earlier studies also revealed that the function of the RNase H is similar to those of the RuvCs (PDB IDs: 4LD0, 1HJR, and 4UN3), and mainly cleaves nucleic acid chains by the two-metal-ion mechanism (Ariyoshi et al., 1994; Nowotny et al., 2005, 2007; Chen et al., 2013; Gorecka et al., 2013; Nishimasu et al., 2014). To investigate the catalytic role of H983 in the RuvC active center, we performed structure superpositions on the above four PDBs. In their catalytic domains, aside from the secondary structures ($\alpha\beta\beta$), the orientations of these key catalytic residues are semblable, for example, site-1, site-2, site-3, and site-4 (D192, D138, H143, and H983) (**Supplementary Figure 7**). Therefore, the function of

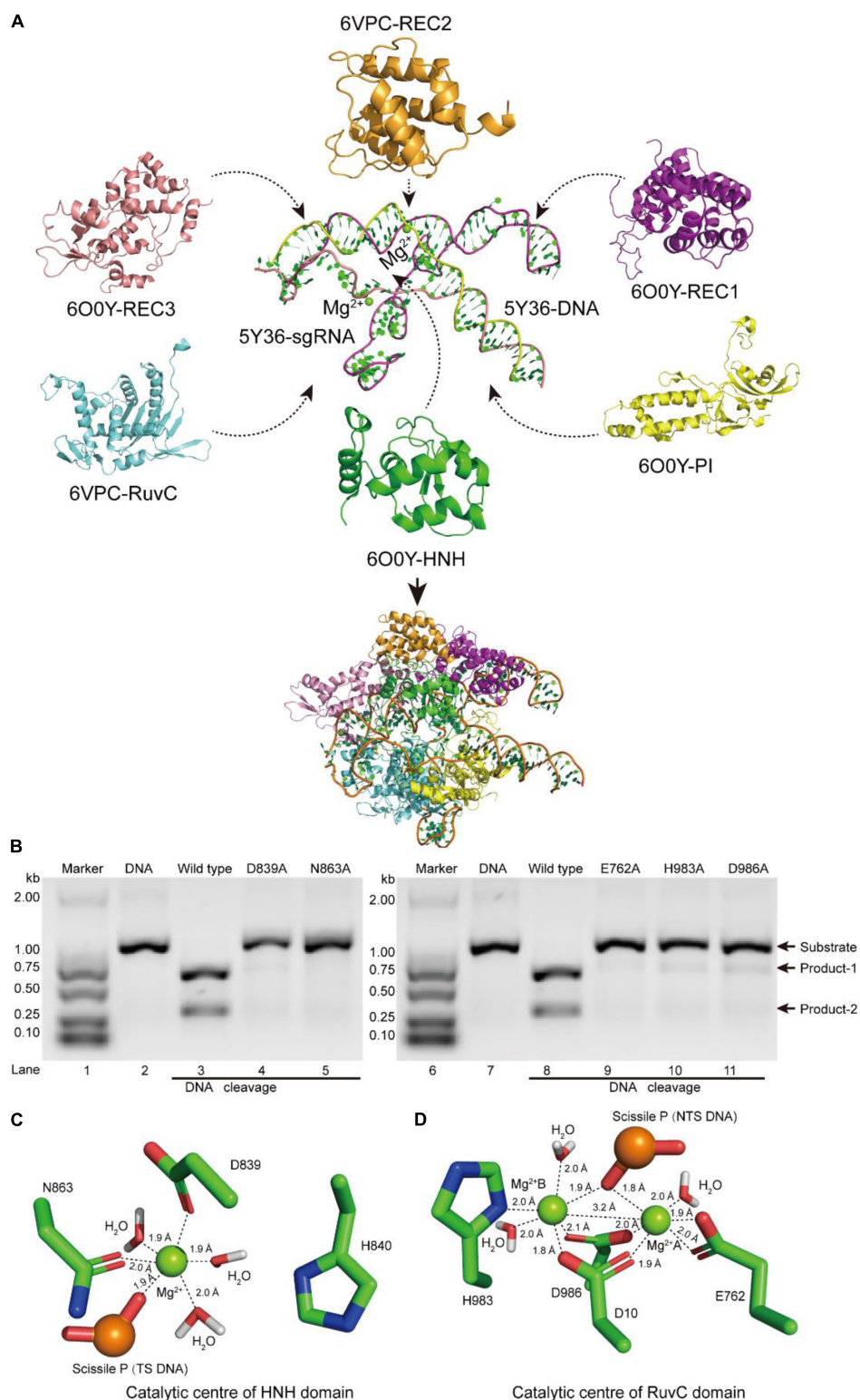


FIGURE 1 | Atomic model of the ternary complex in the DNA cleavage state. **(A)** The structural assembly of SpCas9-sgRNA-DNA ternary complex. Domains HNH, REC1, REC3, and PI come from atomic model 6O0Y, and domains REC2 and RuvC derive from atomic model 6VPC. **(B)** *In vitro* cleavage activities of SpCas9 and its mutants. These mutation sites are located in the HNH and the RuvC domains, respectively. The target DNA molecules are cleaved into two products (product-1 and -2). **(C,D)** The Mg²⁺ ion coordination of the catalytic centers (HNH and RuvC, respectively). The distances between ligands and corresponding Mg²⁺ ions are about 2.0 Å.

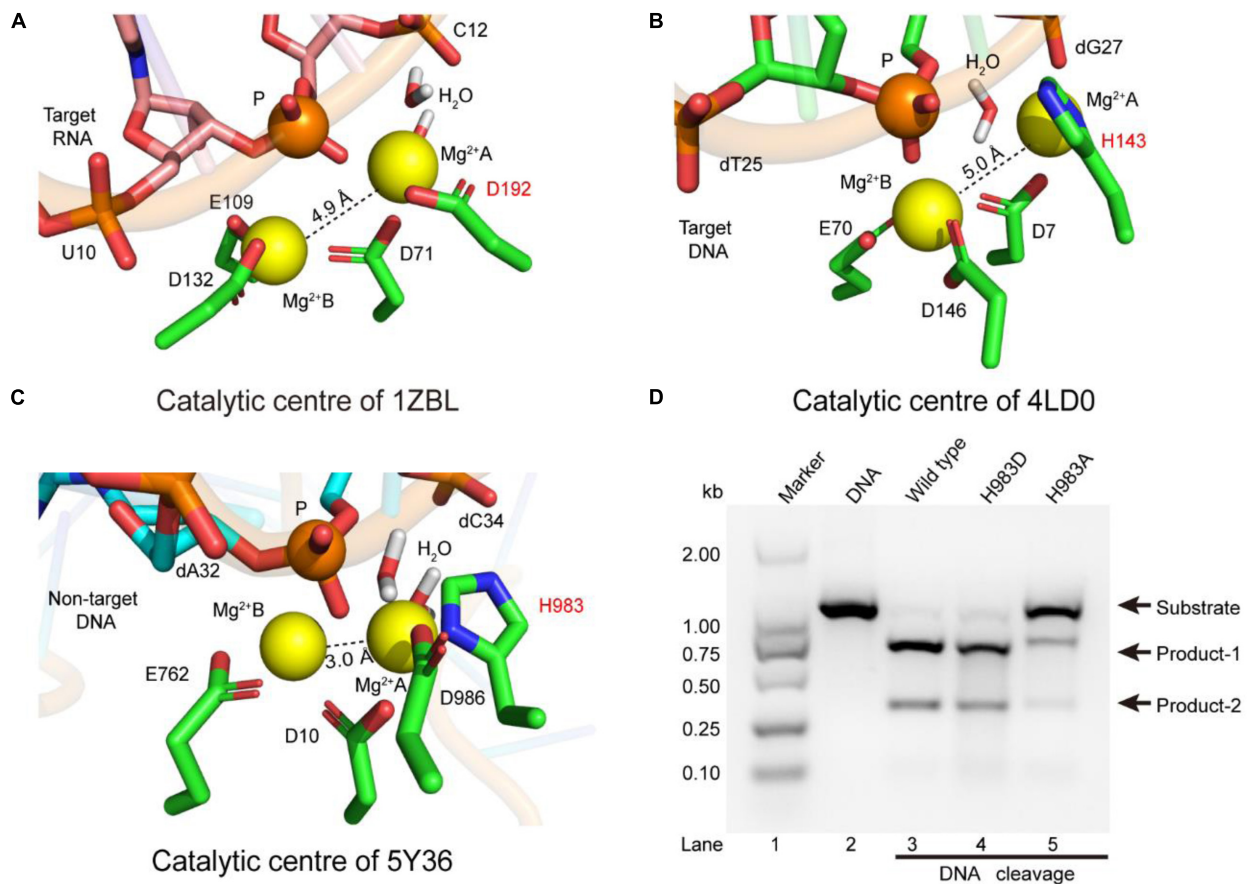


FIGURE 2 | The coordination configurations of the catalytic centers. **(A–C)** The stereoviews of the two-metal catalytic centers in the atomic models (PDB IDs: 1ZBL, 4LD0, and 5Y36, respectively). The Mg²⁺ ions are coordinated by the ligands (amino acids, phosphate groups, and water molecules). **(D)** *In vitro* cleavage activities of SpCas9 and its mutants. The target DNA molecules are cleaved into two products (product-1 and -2).

the residue H983 in the SpCas9 may be to coordinate with the Mg²⁺ ion A.

To identify the catalytic role of H983 in the RuvC active site, we analyzed whether the conformations of residues H983 (5Y36), H143 (4LD0), and D192 (1ZBL) are consistent in the Mg²⁺ coordination by MD simulation after inserting Mg²⁺ into active centers (**Figure 2**). In the 1ZBL or the 4LD0, the distance between these two Mg²⁺ ions is about 5.0 Å; the scissile phosphorus is nearly located between the Mg²⁺ ions A and B; the D192 or H143 and at least one water molecule coordinate with the Mg²⁺ ion A (**Figures 2A,B**). Similarly, in the 5Y36, these two Mg²⁺ ions are spaced 3.0 Å apart; the Mg²⁺ ions A and B are bisected by the scissile phosphorus; the H983 and two water molecules coordinate with the Mg²⁺ ion A (**Figure 2C**). Thus, these results suggest that H983 of SpCas9 may be involved in immobilizing (anchoring) the Mg²⁺ ion A to maintain the reaction conformation of the catalytic center, consistent with the previous suggestion (Fernandes et al., 2016).

Since structure studies have shown that the action of the H983 in the SpCas9 is equivalent to that of the D192 in the RNase H, we presumed that the His mutation into Asp (H983D) at site 983 may retain the native cleavage activity; in contrast, the

cleavage activity with the His mutation into Ala (H983A) will be weakened or may disappear. As expected, the cleavage activity of the mutant H983D is similar to that of the wild-type SpCas9 (lanes 3 and 4 in **Figure 2D**), in agreement with the cleavage activity of the mutant H983N (data not shown); and that of the mutant H983A was sharply reduced (lane 5 in **Figure 2D**). Thus, these results demonstrate that H983 can stabilize the Mg²⁺ ion A during SpCas9 cleaving the DNA, rather than the general base as the previous deductions (Chen and Doudna, 2017; Palermo, 2019). Moreover, the distance analysis between the Mg²⁺ ions A and B in the SpCas9 and its mutants also supports this view (**Supplementary Figure 8**), because the substitution of H983 with Ala (A) or Asp (D) affects the stability of the Mg²⁺ ion A.

Lewis (General) Bases of RuvC Active-Site

The above results indicate that the RuvC residue H983 coordinates with the Mg²⁺ ion, and is not a general base (Chen and Doudna, 2017; Palermo, 2019). Therefore, it might be other residues to serve as general bases during the RuvC domain hydrolyzing the No-target strand (NTS). To confirm

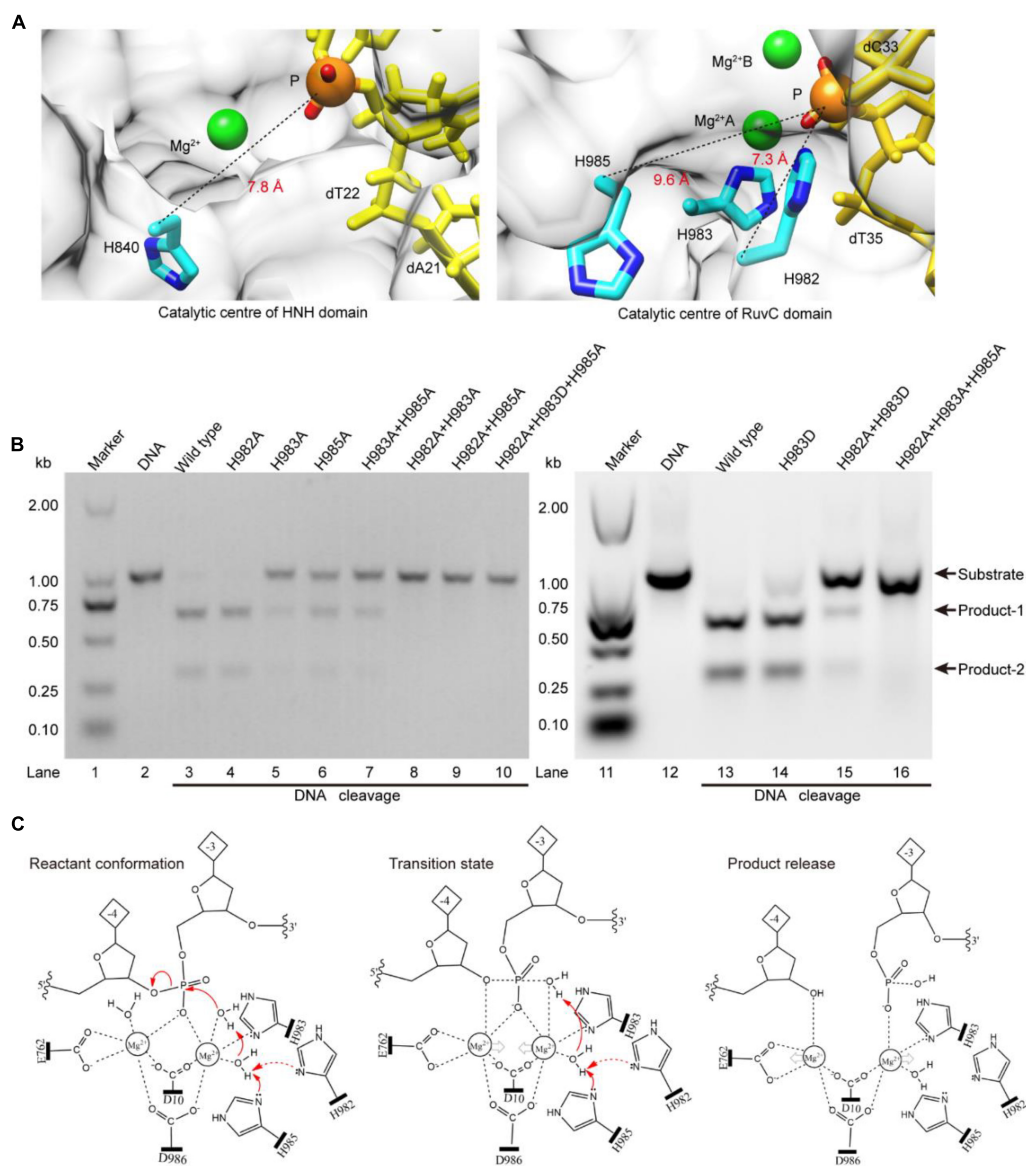


FIGURE 3 | The Lewis bases in the SpCas9 catalytic centers. **(A)** The Lewis bases in the HNH and the RuvC. H840 is the Lewis base in the one-metal domain; H982 and H985 are the potential Lewis bases in the two-metal domain. **(B)** *In vitro* cleavage activities of the SpCas9 mutants. The target DNA molecules are cleaved into two products (product-1 and -2). **(C)** The two-metal-ion model of the RuvC domain. **Reactant conformation:** the cooperative motion of two metal ions. **Transition state:** the high-energy intermediate. **Product release:** the cleavage and breakage of the target DNA strands.

this, referring to the distance (7.8 Å) between the C_α atom of the general base H840 and the corresponding scissile phosphorus, we analyzed residues around the scissile phosphorus in the RuvC catalytic center. As shown in **Figure 3A**, we observed that the C_α atoms of the residues H982 and H985 and the corresponding scissile phosphorus are 7.3 and 9.6 Å apart, respectively; meanwhile, N_δ atoms of their imidazole groups point toward the scissile phosphorus. Therefore, these two amino acids might be general bases. To identify the roles of these residues, through site-directed mutagenesis and cleavage activity detection, we first confirmed whether they

are critical during the NTS hydrolysis. In our results, the mutant H982A is still able to cleave the target DNA and almost maintains the same cleavage activity as the wild-type SpCas9 (lanes 3 and 4 in **Figure 3B**). In contrast, the cleavage activity of the mutant H985A is significantly decreased (lane 6 in **Figure 3B**). Therefore, H985 is a key residue in the RuvC catalytic center and appears to play a more critical role than H982.

To further clarify the precise roles of the H982 and H985, we continued to examine the cleavage activities of several mutants, including H983A + H985A, H982A + H983A,

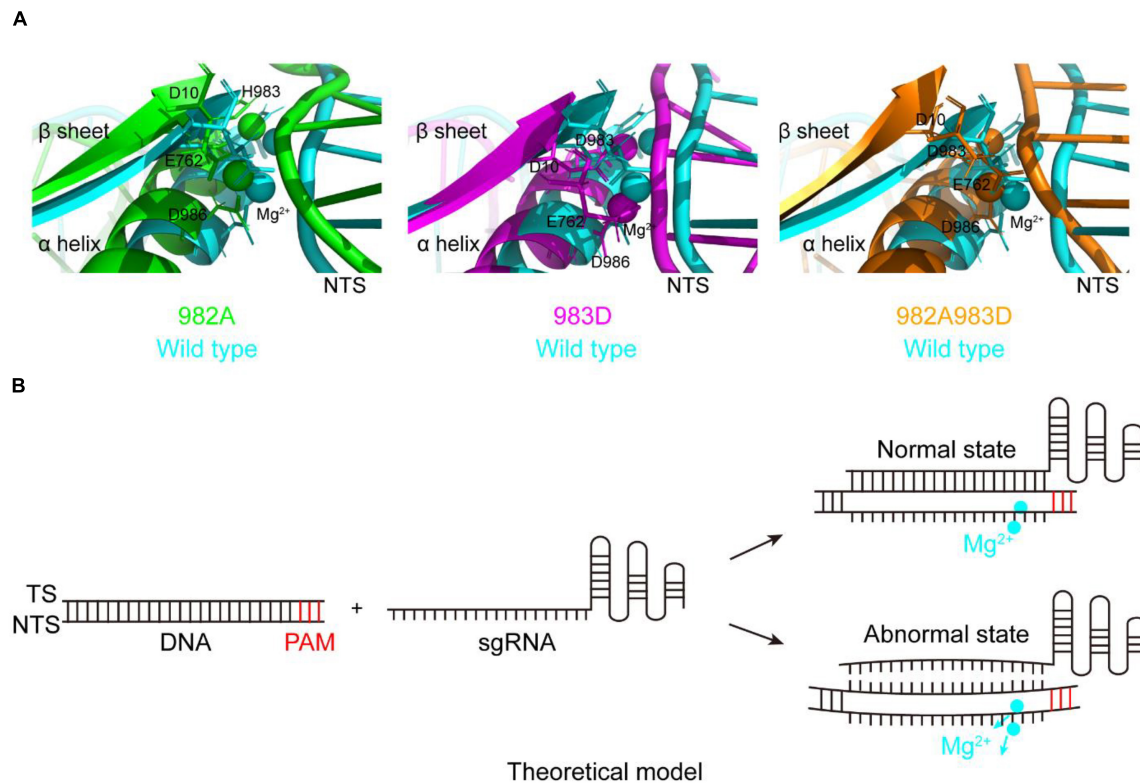


FIGURE 4 | The movement of the non-target strands (NTS) in the PAM distal ends. **(A)** In the RuvC active centers of SpCas9 mutants, the NTSes of the PAM distal ends are shifted due to the conformational change of the RuvC catalytic centers (especially magnesium ions). **(B)** The model of reducing off-targets of SpCas9. In these mutants (H982A, H983D, and H982A + H983D), when sgRNAs are not mated with the NTS of the substrate DNA (in abnormal state), the DNA may be not cleaved by SpCas9 mutants.

H982A + H985A, H982A + H983D, H982A + H983A + H985A, and H982A + H983D + H985A (**Figure 3B**). Among them, the cleavage activity of the mutant H983A + H985A is still partly retained (lane 7 in **Figure 3B**). Therefore, H982 could be a general base to activate the water molecule for the nucleophilic attack of the phosphodiester bond. However, the cleavage activity of the mutant H982A + H983A is removed (lane 8 in **Figure 3B**). This seems to indicate that the H985 is not a Lewis base. But, it is because the distance between the H985 and the scissile phosphorus is too far in the mutant H982A + H983A that the H985-activated water molecule may not complete the nucleophilic attack of the scissile phosphorus (**Supplementary Figure 9**). This is also supported by site-directed mutagenesis because the lost activity of the mutant H982A + H983A is reversed by the mutant H982A + H983D (lane 15 in **Figure 3B**). Thus, H985 could be also a general base like H982.

Meanwhile, in the RuvC catalytic center, to rule out the possibility of independence on a general basis during the NTS hydrolysis, we evaluated the cleavage activities of the mutants H982A + H985A and H982A + H983D + H985A. These two mutants could not cleave the target DNA (lanes 9 and 10 in **Figure 3B**), similar to the mutant H982A + H983A + H985A

(lane 16 in **Figure 3B**). Therefore, in the RuvC domain of SpCas9, the cleavage of the NTS is dependent on the general bases. Moreover, the chemical rescue by imidazole also supports this point (**Supplementary Figure 10**), because the lost activity of SpCas9 mutant H982A + H983D + H985A can be recovered by adding imidazole. Thus, the hydrolyzing of the NTS needs the residues of acting as the general bases besides the residues of stabilizing the Mg²⁺ ions. In the earlier studies, although the understanding of the RuvC domain has made significant advances, the residue H983 has always been regarded as a pseudo general base (Nishimasu et al., 2014; Jiang et al., 2016; Zuo and Liu, 2016; Palermo, 2019). However, the general bases could be residues of H982 and H985 from our studies. Taken together, the above results indicate that the H983 acts as a stabilizer of the Mg²⁺ ion, and reveal that both H982 and H985 serve as general bases to activate water molecules for the nucleophilic attack of the phosphodiester bond.

Based on the above results and the two-metal-ion mechanism (Steitz and Steitz, 1993; Yang, 2008), we propose a revised model for the RuvC catalytic center to hydrolyze the NTS (**Figure 3C**). First, two Mg²⁺ ions are jointly coordinated by the phosphate group of the NTS (between bases -3 and -4), water molecules,

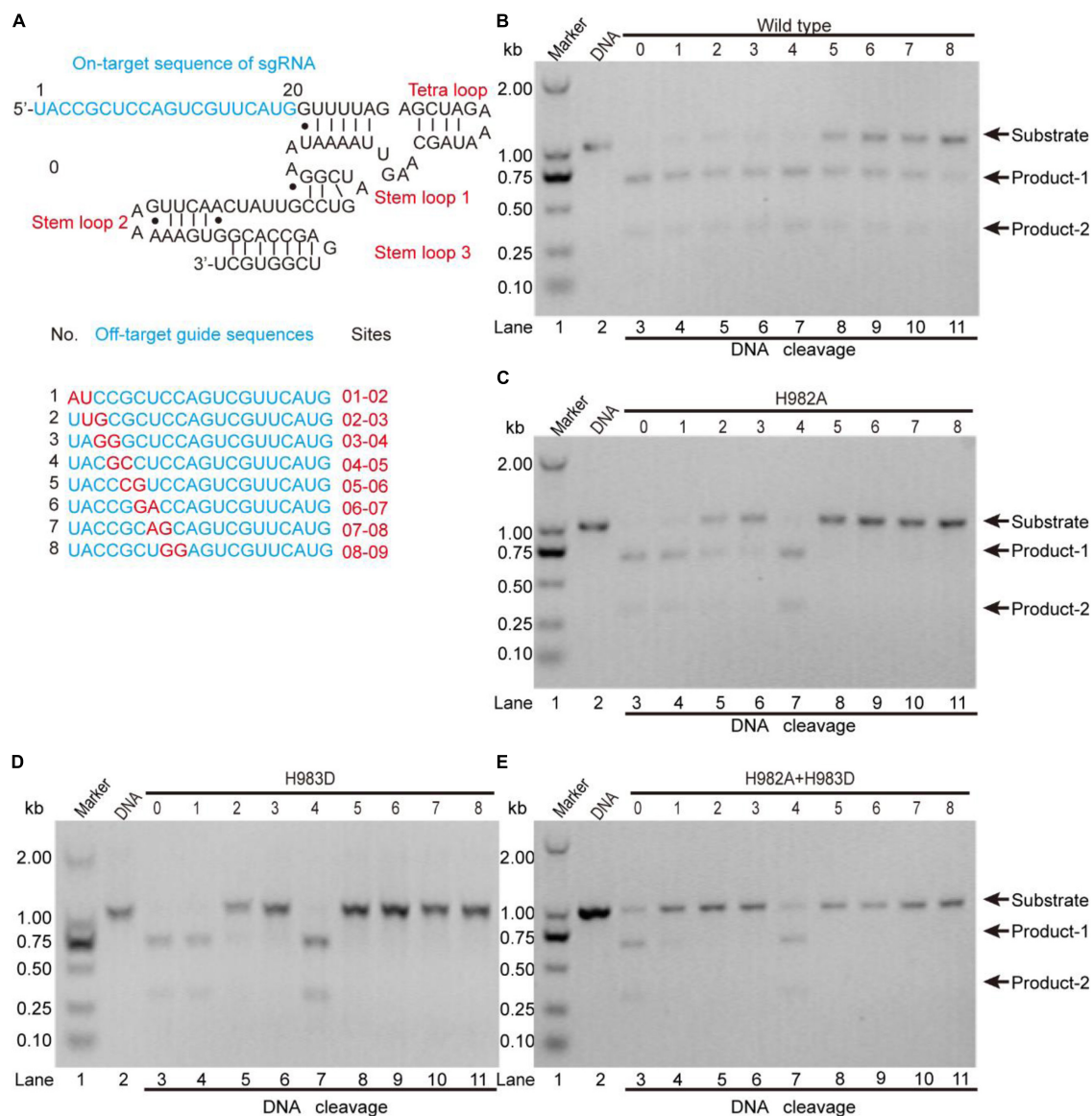


FIGURE 5 | The detection of the *in vitro* off-target effects. **(A)** The sequences of the sgRNA and its mutants. The on-target sequence of the sgRNA shows in No. 0; the off-target sequences show in Nos. 1, 2, 3, 4, 5, 6, 7, and 8. **(B–E)** The detection of the off-target effects for the wild-type SpCas9 and its mutants. The target DNA molecules are cleaved into two products (product-1 and -2).

and four residues (D10, E762, H983, and D986). Among them, the Mg^{2+} ions A and B are about 4 Å apart, and the distances between Mg^{2+} ions and the corresponding ligands are about 2 Å. This coordination architecture facilitates the RuvC catalytic center and the phosphate backbone to get close to each other and then forms the reaction conformation for the nucleophilic attack on the scissile phosphodiester (reactant conformation in **Figure 3C**). Then, the imidazole groups of H982 and/or H985 capture the proton of the coordinated water and generate the nucleophile; meanwhile, two Mg^{2+} ions move toward each other to facilitate the activated water to attack the phosphodiester bond (pro-Rp oxygen). Subsequently, the nucleophile and the

leaving group moieties are transiently bonded to the phosphorus for forming a transition state (transition state in **Figure 3C**). Finally, the nucleophile attacks the phosphorus (P5') and the leaving groups depart simultaneously (product release in **Figure 3C**).

SpCas9 Variants With Reduced Off-Target Effects

During studying for the functions of the residues H982, H983, and H985, we have obtained three mutants with the DNA cleavage activity: H982A, H983D, and H982A + H983D.

Interestingly, in these mutants, the MD simulations showed that the conformation changes of the RuvC catalytic centers remarkably cause a shift of the NTS that interacts with the Target strand (TS) of the PAM-distal end (**Figure 4A**), comparison of that of wild type; Also, stability shift assay showed that the stability of these mutants become high, especially H982A + H983D (**Supplementary Figure 12**). For this, we proposed a model: in the active center where the conformation of the key sites was changed, when sgRNA combined to DNA, if sgRNA completely matched with NTS in DNA, the DNA may be cleaved by SpCas9 variants; on the contrary, the DNA may not be sheared by SpCas9 variants (**Figure 4B**). Therefore, we speculated that these mutants may hinder complete hybridization between the sgRNA and the DNA to improve their specificity when there are no highly matched Watson–Crick base pairs between the sgRNA and the TS.

To clarify this, we detected the off-target effects of the mentioned mutants by agarose gel electrophoresis (**Figure 5**). As seen, for the wild-type, all sgRNAs can guide the DNA cleavage (**Figure 5B**), indicating that the off-target effects of the wild-type SpCas9 are very severe. For the mutants H982A and H983D, only sgRNAs 1, 2, 3, and 4 can guide them to cleavage the target DNA (**Figures 5C,D**), demonstrating that the mutants H982A and H983D could reduce the off-target effects. For the mutant H982A + H983D, only sgRNAs 1 and 4 are capable of guiding the DNA cleavage (**Figure 5E**), illustrating that this mutant greatly decreases the off-target effects. In addition, to evaluate the kinetics of these mutants, we also performed the time-course cleavage reaction, and qualitatively observed that their reaction rates had the following trend: SpCas9 > H982A > H983D \approx H982A + H983D (**Supplementary Figure 13**). In general, in the three mutants, the mutant H982A + H983D may have more potential application value. Despite it being necessary to conduct *in vivo* studies for these mutants, it is not the focus of the thesis; thus, *in vivo* studies related to these mutants will be exhibited in future work.

In the previous studies, the strategies to decrease the SpCas9 off-target effects mainly include the following several aspects: (1) using a pair of catalytic-inactive SpCas9 nucleases which fused to a *FokI* nuclease domain (Tsai et al., 2014); (2) truncating the guide sequence of the sgRNA at 5' end (Fu et al., 2014); (3) decreasing the number of the active SpCas9 in the cell (Hsu et al., 2013); (4) using a pair of SpCas9 nickase mutants to produce double nicks for DNA (Ran et al., 2013); and (5) neutralizing positively charged residues within the NTS groove (Slaymaker et al., 2016). Nevertheless, few studies focused on the relationship between the specificity of SpCas9 and the residue types of its catalytic center. In our results, we found that the SpCas9 mutants may possess high catalytic specificity by changing the catalytic residues of the RuvC catalytic center.

CONCLUSION

In summary, we refined the cryo-EM structure of the SpCas9–sgRNA–DNA complex in the DNA cleavage state. This active

structure presents an atomic model for the HNH and RuvC active-sites in complex with DNA, Mg^{2+} ions, and water molecules, and thereby identifying the catalytic roles of the residues D10, E762, H982, H983, H985, and D986 in the RuvC active sites. Moreover, our catalytic model led to a new engineered SpCas9 variant (SpCas9-H982A + H983D) with reduced off-target effects. Hence, this study not only provides new mechanistic insights into the DNA cleavage by CRISPR-Cas9 but also offers useful guidance for engineering new CRISPR-Cas9 systems with improved specificity.

MATERIALS AND METHODS

Expression and Purification of Cas9 Proteins

The sequences of encoding SpCas9 and its mutants (primers in **Supplementary Table 1**, mutants in **Supplementary Table 2**) were inserted between the restriction sites (*NdeI* and *XhoI*) of the plasmid pET-21, and their N-terminals were fused with a His \times 6 tag and Tobacco etch virus (TEV) protease cleavage site (Huai et al., 2017). The integrative plasmid was transformed into Rosetta (DE3) competent cells (TANGEN), which were cultivated in Terrific Broth (Sangon Biotch, Shanghai, China) containing antibiotics (100 μ g/ml Amp and 30 μ g/ml Cm) at 37°C on a 200 rpm shaker. When the absorption value (OD600) of the cell concentration in the bacteria solution was approximately 0.8–1.0 (Huai et al., 2017), a 5 mM-IPTG was added to the bacteria solution to induce protein expression at 16°C on a 160 rpm shaker for approximately 20 h. Next, harvest cells were collected by centrifugation at 4°C and 5,000 rpm for 10 min and were lysed by sonication (power output 5, pulse-on 3 s, pulse-off 3 s, for a total of 10 min) to release the protein in the lysis buffer (20 mM HEPES, 500 mM KCl, pH = 7.5, 0.2 μ M filtered and degassed) (Huai et al., 2017). Then, the protein was bound to Ni-NTA agarose beads (Qiagen, Shanghai, China) in the ice bath on a 150 rpm shaker for at least 1.5 h, was washed by the wash buffer (20 mM HEPES, 500 mM KCl, 1% sucrose, pH = 7.5, 0.2 μ M filtered and degassed) at 2 ml/min, and eluted by the eluent buffer (wash buffer with 20, 30, 50, 100, and 250 mM imidazole, respectively; pH = 7.5, 0.2 μ M filtered and degassed) at 2 ml/min (Anders et al., 2014; Huai et al., 2017). Subsequently, the protein was incubated with TEV protease overnight at 4°C to remove the 6 \times His tag. Finally, the protein was further purified by SP Sepharose HiTrap column (elution with a linear gradient of 0.1–1 M KCl) and Superdex 200 16/60 column from GE Life Sciences (elution with 20 mM HEPES, 500 mM KCl, pH = 7.5) (Anders et al., 2014; Jiang et al., 2016; Huai et al., 2017), and was concentrated by 100 kDa MWCO centrifugal filter (Merck Millipore) to store in the storage buffer (20 mM HEPES, 150 mM KCl, 1 mM DTT, glycerol 50%, pH = 7.5) (Huai et al., 2017). All proteins were detected by SDS-PAGE (SDS-PAGE image of part proteins are shown in **Supplementary Figure 11**).

Note: The base sequences (**Supplementary Table 3**) of SpCas9 were mutated using the Muta-directTM site-Directed Mutagenesis

Kit (Beijing SBS Genetech Co, Ltd., Beijing, China¹, cat.no.SDM-15) to obtain every mutant (**Supplementary Table 2**) in the present text. Reaction system (50 μ l) includes 10 \times reaction buffer (5 μ l), sample plasmid (10 ng/ μ l, 2 μ l), primer F/R (10 pmol/ μ l, 1 μ l), dNTP mixture (each 2.5 mM, 2 μ l), ddH₂O (38 μ l), Muta-directTM enzyme (1 μ l). A reaction is ended under the conditions (95°C 30 s, 55°C 30 s, 72°C 10 min, 15–18 cycles), MutazymeTM (10 U/ μ l, 1 μ l) was added to the reaction system to digest methylated plasmids; Then, mutant plasmids in the reaction system were transformed into competent cells.

Preparation of Nucleic Acids

A 98-nt sgRNA was *in vitro* transcribed and purified using the MEGAShortscript T7 Transcription Kit and the MEGAclean Transcription Clean-Up Kit from Thermo Fisher Scientific (China) Co., Ltd., Pudong New Area, Shanghai, China. A 920-bp target DNA contained PAM (TGG) was amplified by the Ultra HiFidelity PCR Kit (TIANGEN, Sichuan, China) at the following conditions (94°C 30 s, 55°C 30 s, 72°C 1.5 min, 30 cycles), and was purified using the AxyPrepTM DNA Gel Extraction Kit (Axygen Biotechnology, Taizhou, China). The purified sgRNA and DNA were resuspended using the desired solution and volume, and stored at –80°C.

Activity Detection of Cas9 Proteins

Methods of activity can be detected in three ways: (1) *In vitro* activity assays: the mole numbers (200 nM, 1 μ l) of Cas9 proteins (**Supplementary Table 2**) are equal to that of sgRNA, and both cas9 protein and sgRNA were mixed and incubated in the buffer (20 mM HEPES, 100 mM KCl, 1 mM DTT, 0.5 mM EDTA, 2 mM MgCl₂, 5% glycerol, pH = 7.5) at 37°C for 30 min; the buffer was added with 100–150 ng target DNA, and incubated another 1.5 h at 37°C; the cleavage products were detected by gel electrophoresis on 1% agarose gel stained with 1 \times GeneGreen nucleic acid dye (TIANGEN). (2) Chemical rescue of His982 and His985 by imidazole: this process was similar to (1), with the only difference being the addition of a step: “the solution was mixed with 2 μ l imidazole (500 mM) and incubated overnight at 37°C” before the detection of products. (3) *In vitro* detection of the off-target effects: the process was similar to (1), with the difference being sgRNA (continuous mutation with two bases as a unit in 20-nt sequence for target DNA) and the incubation time (4 h) after adding target DNA.

Building and Refinement of Atomic Models

The atomic model of the ternary complex (SpCas9–sgRNA–DNA) was built according to the cryo-EM density maps (end-21308.map and end-0584.map) (Zhu et al., 2019; Lapinaite et al., 2020), and the previous conformation of SpCas9 (PDB IDs: 5Y36, 6O0Y, and 6VPC) (Huai et al., 2017; Zhu et al., 2019; Lapinaite et al., 2020). First, the initial model

of the all-atom ternary complex was constructed through the homology modeling method MODELLER using the aforementioned PDBs as templates (Fiser and Sali, 2003; Huai et al., 2017; Zhu et al., 2019; Lapinaite et al., 2020). Next, the initial position of the HNH domain was modified based on reference structures (PDB ID 6O0Y) and its cryo-EM density (end-0584.map) (Zhu et al., 2019); at the same time, the RuvC domain was amended using 6VPC and its cryo-EM density (end-21308.map) as references. Then, the position of the Mg²⁺ ions in the active sites was confirmed by NAMD with a 12–6–4-type multisite model to obtain the experimental coordination patterns (Li and Merz, 2014; Liao et al., 2017). Finally, the initial all-atom model of the ternary complex was automatically refined by Situs-3.1, Rosetta macromolecular modeling suite, and MD simulations (Wriggers, 2010; Leaver-Fay et al., 2011; Mirjalili et al., 2014).

Molecular Dynamics Simulations

The MD simulations with explicit water models were conducted using NAMD (Phillips et al., 2005). The CHARMM36 force field (Hart et al., 2012) and the TIP3P water (Jorgensen et al., 1983) model were employed to model the simulation system of the ternary complex with the water solvent. To build a simulation system, some simulation parameters were adjusted: the all-atom structure of the ternary complex was solvated in the center of a cubic water box with a minimum distance of 12 Å from the complex surface to the edge of the box; the Na⁺ and Cl[–] ions were used to mimic an ionic concentration of 15 M in the system, including the certain number of additional Na⁺ ions that neutralize the net negative charge of the complex. To conduct the simulations, periodic boundary conditions were used; the van der Waals interactions were treated with a cut-off distance of 10 Å using a smooth switching function from 8 Å; the electrostatic interactions were calculated with particle mesh Ewald (PME) method using a local interaction distance of 10 Å; the SHAKE algorithm was employed to constrain bonds involving hydrogen atom, and thereby a time step of 2.0 fs was used. Ultimately, the simulations were performed in the isobaric-isothermal (NPT) ensemble, at a constant pressure of 1 bar and a constant temperature of 298 K controlled by Langevin dynamics.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

QH supervised the project. HT, HY, WD, GL, and DX expressed and purified proteins and conducted the DNA cleavage assay. HT and QH carried out the building, refinement of atomic models,

¹<http://www.sbsbio.com>

and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the grants from National Major Scientific and Technological Special Project for “Significant New Drugs Development” (2018ZX09J18112), the National Natural Science Foundation of China (31671386, 31971377, and 91430112), the Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), and ZJLab.

REFERENCES

- Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* 513, 569–573. doi: 10.1038/nature13579
- Ariyoshi, M., Vassilyev, D. G., Iwasaki, H., Nakamura, H., Shinagawa, H., and Morikawa, K. (1994). Atomic-structure of the ruvc resolvase – a holliday junction-specific endonuclease from *escherichia-coli*. *Cell* 78, 1063–1072. doi: 10.1016/0092-8674(94)90280-1
- Ashworth, J., and Baker, D. (2009). Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.* 37, e73. doi: 10.1093/nar/gkp242
- Chen, J. S., and Doudna, J. A. (2017). The chemistry of Cas9 and its CRISPR colleagues. *Nat. Rev. Chem.* 1:0078. doi: 10.1038/s41570-017-0078
- Chen, L., Shi, K., Yin, Z., and Aihara, H. (2013). Structural asymmetry in the thermus thermophilus RuvC dimer suggests a basis for sequential strand cleavages during holliday junction resolution. *Nucleic Acids Res.* 41, 648–656. doi: 10.1093/nar/gks1015
- Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823. doi: 10.1126/science.1231143
- De Vivo, M., Dal Peraro, M., and Klein, M. L. (2008). Phosphodiester cleavage in ribonuclease H occurs via an associative two-metal-aided catalytic mechanism. *J. Am. Chem. Soc.* 130, 10955–10962. doi: 10.1021/ja8005786
- Fernandes, H., Pastor, M., and Bochtler, M. (2016). Type II and type V CRISPR effector nucleases from a structural biologist's perspective. *Postepy Biochem.* 62, 315–326.
- Fiser, A., and Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* 374, 461–491. doi: 10.1016/S0076-6879(03)74020-8
- Fu, Y., Reyon, D., and Joung, J. K. (2014). Targeted genome editing in human cells using CRISPR/Cas nucleases and truncated guide RNAs. *Methods Enzymol.* 546, 21–45. doi: 10.1016/B978-0-12-801185-0.00002-7
- Gorecka, K. M., Komorowska, W., and Nowotny, M. (2013). Crystal structure of RuvC resolvase in complex with holliday junction substrate. *Nucleic Acids Res.* 41, 9945–9955. doi: 10.1093/nar/gkt769
- Hart, K., Foloppe, N., Baker, C. M., Denning, E. J., Nilsson, L., and Mackerell, A. D. Jr. (2012). Optimization of the CHARMM additive force field for DNA: improved treatment of the BI/BII conformational equilibrium. *J. Chem. Theory Comput.* 8, 348–362. doi: 10.1021/ct200723y
- Ho, M. H., De Vivo, M., Dal Peraro, M., and Klein, M. L. (2010). Understanding the effect of magnesium ion concentration on the catalytic activity of ribonuclease H through computation: does a third metal binding site modulate endonuclease catalysis? *J. Am. Chem. Soc.* 132, 13702–13712. doi: 10.1021/ja102933y
- Hsu, P. D., Lander, E. S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157, 1262–1278. doi: 10.1016/j.cell.2014.05.010
- Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* 31, 827–832. doi: 10.1038/nbt.2647
- Huai, C., Li, G., Yao, R., Zhang, Y., Cao, M., Kong, L., et al. (2017). Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nat. Commun.* 8:1375. doi: 10.1038/s41467-017-01496-2
- Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., et al. (2016). Structures of a CRISPR-Cas9 r-loop complex primed for dna cleavage. *Science* 351, 867–871. doi: 10.1126/science.aad8282
- Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J. A. (2015). Structural biology. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* 348, 1477–1481. doi: 10.1126/science.aab1452
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided dna endonuclease in adaptive bacterial immunity. *Science* 337, 816–821. doi: 10.1126/science.1225829
- Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., et al. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* 343, 1215–1226. doi: 10.1126/science.1247997
- 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
- Lapinaite, A., Knott, G. J., Palumbo, C. M., Lin-Shiao, E., Richter, M. F., Zhao, K. T., et al. (2020). DNA capture by a CRISPR-Cas9-guided adenine base editor. *Science* 369, 566–571. doi: 10.1126/science.abb1390
- Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., et al. (2011). Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 487, 545–574. doi: 10.1016/B978-0-12-381270-4.00019-6
- Li, P., and Merz, K. J. (2014). Taking into account the ion-induced dipole interaction in the nonbonded model of ions. *J. Chem. Theory Comput.* 10, 289–297. doi: 10.1021/ct400751u
- Liao, Q., Pabis, A., Strodel, B., and Kamerlin, S. (2017). Extending the nonbonded cationic dummy model to account for ion-induced dipole interactions. *J. Phys. Chem. Lett.* 8, 5408–5414. doi: 10.1021/acs.jpclett.7b02358
- Mirjalili, V., Noyes, K., and Feig, M. (2014). Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins* 82(Suppl. 2), 196–207. doi: 10.1002/prot.24336
- Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., et al. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156, 935–949. doi: 10.1016/j.cell.2014.02.001
- Nowotny, M., Gaidamakov, S. A., Crouch, R. J., and Yang, W. (2005). Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell* 121, 1005–1016. doi: 10.1016/j.cell.2005.04.024
- Nowotny, M., Gaidamakov, S. A., Ghirlando, R., Cerritelli, S. M., Crouch, R. J., and Yang, W. (2007). Structure of human RNase H1 complexed with an RNA/DNA hybrid: insight into HIV reverse transcription. *Mol. Cell* 28, 264–276. doi: 10.1016/j.molcel.2007.08.015
- Nowotny, M., and Yang, W. (2006). Stepwise analyses of metal ions in RNase H catalysis from substrate destabilization to product release. *EMBO J.* 25, 1924–1933. doi: 10.1038/sj.emboj.7601076
- Palermo, G. (2019). Structure and dynamics of the CRISPR-Cas9 catalytic complex. *J. Chem. Inform. Model.* 59, 2394–2406. doi: 10.1021/acs.jcim.8b00988

ACKNOWLEDGMENTS

We thank Drs. Jianping Liu and Peng Nan for their assistance in the experimental study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.653262/full#supplementary-material>

- Palermo, G., Cavalli, A., Klein, M. L., Alfonso-Prieto, M., Dal Peraro, M., and De Vivo, M. (2015). Catalytic metal ions and enzymatic processing of DNA and RNA. *Acc. Chem. Res.* 48, 220–228. doi: 10.1021/ar500314j
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802. doi: 10.1002/jcc.20289
- Ran, F. A., Hsu, P. D., Lin, C. Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., et al. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154, 1380–1389. doi: 10.1016/j.cell.2013.08.021
- Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. *Science* 351, 84–88. doi: 10.1126/science.aad5227
- Steitz, T. A., and Steitz, J. A. (1993). A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl. Acad. Sci. U.S.A.* 90, 6498–6502. doi: 10.1073/pnas.90.14.6498
- Tsai, S. Q., Wyvekens, N., Khayter, C., Foden, J. A., Thapar, V., Reyon, D., et al. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* 32, 569–576. doi: 10.1038/nbt.2908
- Wang, H., La Russa, M., and Qi, L. S. (2016). CRISPR/Cas9 in genome editing and beyond. *Annu. Rev. Biochem.* 85, 227–264. doi: 10.1146/annurev-biochem-060815-014607
- Wriggers, W. (2010). Using Situs for the integration of multi-resolution structures. *Biophys. Rev.* 2, 21–27. doi: 10.1007/s12551-009-0026-3
- Yang, W. (2008). An equivalent metal ion in one- and two-metal-ion catalysis. *Nat. Struct. Mol. Biol.* 15, 1228–1231. doi: 10.1038/nsmb.1502
- Zhu, X., Clarke, R., Puppala, A. K., Chittori, S., Merk, A., Merrill, B. J., et al. (2019). Cryo-Em structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nat. Struct. Mol. Biol.* 26, 679–685. doi: 10.1038/s41594-019-0258-2
- Zuo, Z., and Liu, J. (2016). Cas9-catalyzed DNA cleavage generates staggered ends: evidence from molecular dynamics simulations. *Sci. Rep.* 5:37584. doi: 10.1038/srep37584

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

Copyright © 2021 Tang, Yuan, Du, Li, Xue and Huang. 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.



Conformational Characterization of the Co-Activator Binding Site Revealed the Mechanism to Achieve the Bioactive State of FXR

Anita Kumari^{1,2}, Lovika Mittal¹, Mitul Srivastava¹, Dharam Pal Pathak^{2,3} and Shailendra Asthana^{1*}

¹Translational Health Science and Technology Institute (THSTI), Faridabad, India, ²Department of Pharmaceutical Chemistry, Delhi Pharmaceutical Sciences and Research University (DPSRU), New Delhi, India, ³Delhi Institute of Pharmaceutical Sciences and Research (DIPSAR), New Delhi, India

OPEN ACCESS

Edited by:

Ramanathan Sowdhamini,
National Centre for Biological
Sciences, India

Reviewed by:

Supriyo Bhattacharya,
City of Hope National Medical Center,
United States
Amit Kumar,
University of Cagliari, Italy

*Correspondence:

Shailendra Asthana
sasthana@thsti.res.in

Specialty section:

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

Received: 25 January 2021

Accepted: 14 July 2021

Published: 31 August 2021

Citation:

Kumari A, Mittal L, Srivastava M,
Pathak DP and Asthana S (2021)
Conformational Characterization of the
Co-Activator Binding Site Revealed the
Mechanism to Achieve the Bioactive
State of FXR.
Front. Mol. Biosci. 8:658312.
doi: 10.3389/fmolb.2021.658312

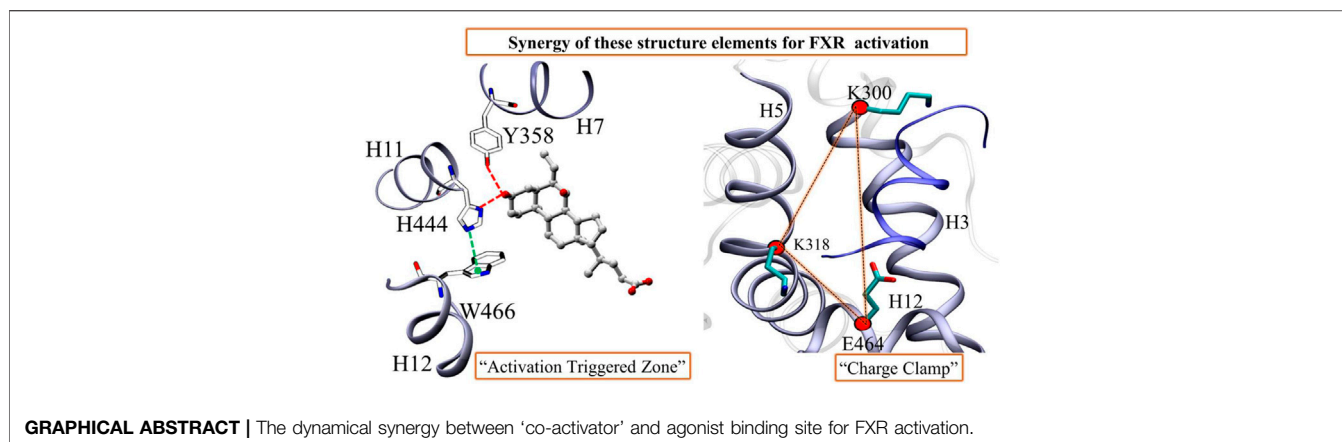
FXR bioactive states are responsible for the regulation of metabolic pathways, which are modulated by agonists and co-activators. The synergy between agonist binding and 'co-activator' recruitment is highly conformationally driven. The characterization of conformational dynamics is essential for mechanistic and therapeutic understanding. To shed light on the conformational ensembles, dynamics, and structural determinants that govern the activation process of FXR, molecular dynamic (MD) simulation is employed. Atomic insights into the ligand binding domain (LBD) of FXR revealed significant differences in inter/intra molecular bonding patterns, leading to structural anomalies in different systems of FXR. The sole presence of an agonist or 'co-activator' fails to achieve the essential bioactive conformation of FXR. However, the presence of both establishes the bioactive conformation of FXR as they modulate the internal wiring of key residues that coordinate allosteric structural transitions and their activity. We provide a precise description of critical residue positioning during conformational changes that elucidate the synergy between its binding partners to achieve an FXR activation state. Our study offers insights into the associated modulation occurring in FXR at bound and unbound forms. Thereafter, we also identified hot-spots that are critical to arrest the activation mechanism of FXR that would be helpful for the rational design of its agonists.

Keywords: farnesoid X receptor, agonist, molecular dynamics simulation, binding free energy calculations, principal component analysis

INTRODUCTION

Upon bile acid (BA) binding, FXR regulates a network of genes in synthesis, uptake, and secretion along with intestinal absorption, thus regulating the level of BAs in the cells. An abnormal BA metabolism is associated with liver injury, metabolic disorders, cardiovascular and cardiovascular and digestive system diseases (Li and Chiang, 2014; Chiang, 2017). FXR is a nuclear receptor that

Abbreviations: FXR, Farnesoid X receptor; MD, molecular dynamics; AF1, activation function 1; DBD, DNA-binding domain; LBD, ligand-binding domain; 6-ECDCA, 6-ethylchenodeoxycholic acid; MM-GBSA, molecular mechanics-generalized born surface area; GAFF, general amber force field; MM, molecular mechanics; PCA, principal component analysis; FEL, Free energy landscape; CAS, computational alanine scanning.



belongs to the NR superfamily and is predominantly found in the liver, intestine, and kidney (Makishima et al., 1999; Parks et al., 1999; Wang et al., 1999; Aranda and Pascual, 2001). FXR is essential in regulating the network of genes involved in maintaining BA and lipid homeostasis (Sinal et al., 2000; Kumari et al., 2020) and, therefore, has a considerable pharmacological relevance (Zhang and Edwards, 2008; Hollman et al., 2012; Arab et al., 2017). Significant work has been carried out to discover many synthetic molecules viz. steroidal and non-steroidal agonists for the FXR. Accordingly, the *first-in-class* FXR agonist 6 α -ethyl-CDCA (2, 6-ECDCA, INT-747, obeticholic acid, OCA) has gained approval for primary biliary cirrhosis (PBC) and is undergoing development for several other liver-related disorders such as NASH and NAFLD (Mudaliar et al., 2013; Neuschwander-Tetri et al., 2015). It is reported that the chemical manipulation on CDCA (chenodeoxycholic acid) scaffold helps to improve potency, efficacy, and metabolic stability of bile acid ligands (Di Leva et al., 2013; Sepe et al., 2015; Festa et al., 2017). Among them, the introduction of an ethyl group at C6 in CDCA makes the 6-EDCA ('OCA') (Figures 1A,B) approximately 100-fold more potent than CDCA (Pellicciari et al., 2002). 'OCA' is the "*first in class*" selective agonist for FXR having anti cholestatic and hepatoprotective properties (Abenavoli et al., 2018; Connolly et al., 2018). In addition to this, hepatic inflammation and intestinal inflammation can be inhibited by 'OCA' induced FXR activation. However, these effects could be problematic in a patient population with an elevated risk for cardiovascular diseases (Hirschfield et al., 2015; Neuschwander-Tetri et al., 2015; Bowlus, 2016; Pencek et al., 2016). Recently, it was reported that 'OCA' failed to achieve a first therapy against NASH (AuthorAnonymous, 2020), as it was reported that the complete FXR activation inhibits metabolic cholesterol breakdown and limits bile acid production, resulting in increased cholesterol levels in 'OCA' clinical studies (Neuschwander-Tetri et al., 2015). Therefore, it seems that complete and/or pronounced agonism possibly not favorable. Hence, it is essential to discern the binding mechanism, dynamics and determinants of FXR at molecular level.

Similarly to other NRs, the FXR protein exhibits a modular structure and contains few autonomous functional domains. It includes an N-terminal region with a ligand-independent activation function (AF1), a highly conserved zinc-finger DNA-binding domain (DBD) that is connected to the LBD by a flexible hinge region (Massafra et al., 2018). Additionally, the LBD contains two well-conserved regions. A signature motif and the AF2 motif are located at the C-terminal end of the LBD, responsible for the ligand-dependent transactivation function. In recent years, a considerable number of crystallographic structures of the LBD of several NRs have appeared in the literature, which suggested that upon agonist binding to FXR, it results in a large conformational rearrangement of FXR, causing the dissociation of co-repressors and the recruitment of 'co-activator' which promote the transcriptional initiation (Downes et al., 2003; Costantino et al., 2005; Merk et al., 2019). The crystal comparison of apo- and agonist-bound structures help to identify the key residues and structural determinants for FXR agonism. The static picture from the X-ray structures indicates that significant conformational changes were observed to establish a connection between the apo form and the active state of FXR (bounded with agonist and 'co-activator'). The co-crystal structure of FXR with 'OCA' (PDB-ID: 1OSV) has revealed that helix H12 adopts the https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/agonist agonistic conformation and stabilizes the 'co-activator' peptide binding (Mi et al., 2003). The binding of 'OCA' recruits the helix H12 against the helices 3, 4, and 10, corresponding to the "active state" of FXR, where the helix H12 stabilizes the binding of the 'co-activator' (Figure 1C). It seems that the 'OCA' has a higher affinity between BAs due to the placement of the 6 α -ethyl group into a hydrophobic cavity between the side chains of I359, F363 and Y366 (Pellicciari et al., 2002). This analysis indicates that the binding of the 'co-activator' significantly contributes to the stabilization of the FXR+OCA complex and thereby affects the conformation. It has been observed that the recruitment of that agonist and 'co-activator' binding are necessary to produce these significant conformational changes and induces a loss or gain of interaction networks stabilized through hydrogen bonding and vdW interactions in FXR. Also, the architecture of 'co-activator' site and its dynamical synergy with agonist site is not explored in details. Therefore, we are exploring the

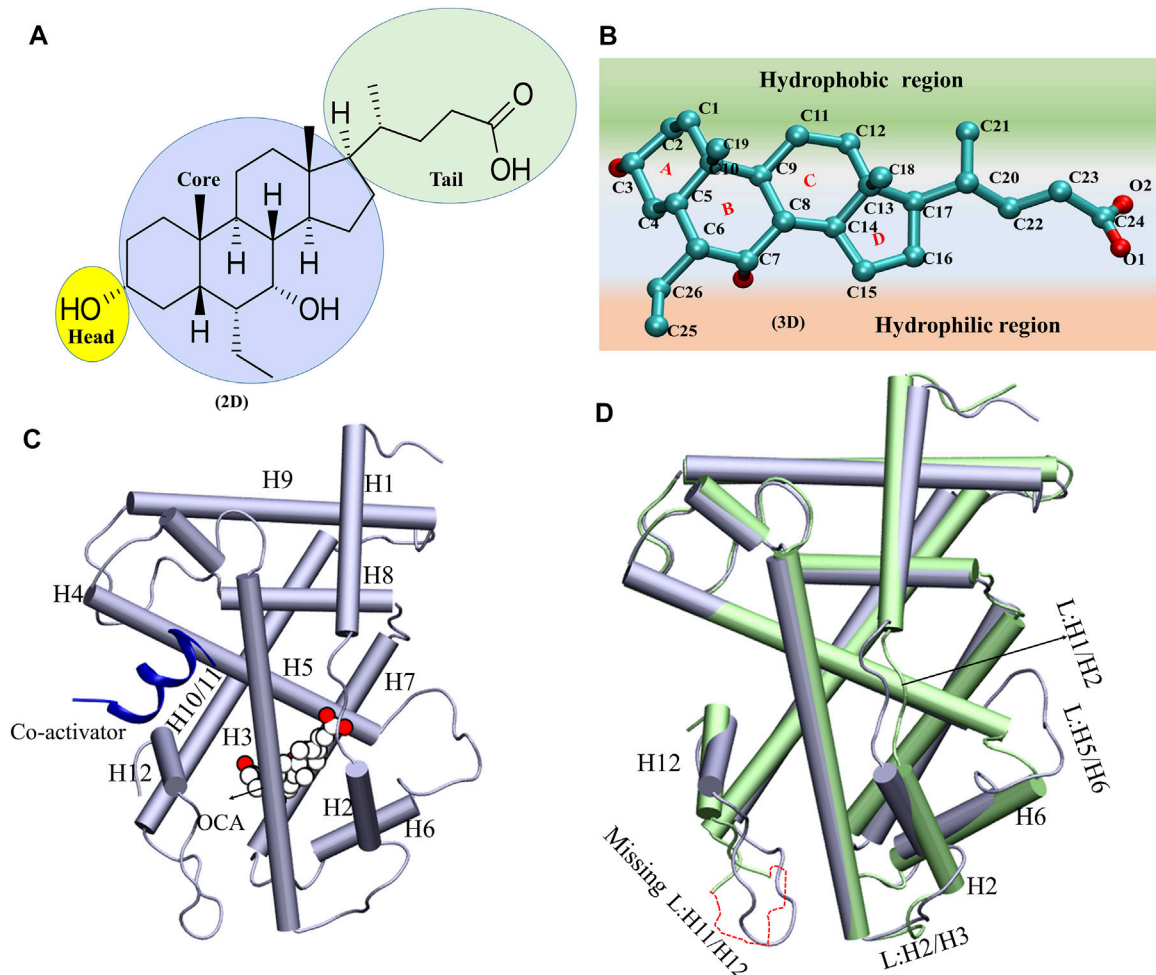


FIGURE 1 | The schematic presentation of the OCA and overlay of the system A and Systems D. **(A,B)** The 2D and 3D representations of 'OCA'. The different regions of 'OCA' were highlighted in different colors. **(C)** The structure of ligand binding domain of FXR. The binding of 'OCA' (VDW: white) and 'co-activator' (cartoon: blue) are highlighted. **(D)** The superimposition of crystal structures of System C (PDB: ID 5Q0K; in lime) and System D (PDB: ID 1 OSV; ice blue) and induce changes in the structure are mentioned.

dynamical changes of FXR with 'co-activator' in the presence and absence of agonists (i.e., 'OCA') through molecular dynamics simulations. The precise description of the positioning of critical residues during conformational changes will help to elucidate the synergy with its binding partners and how FXR is able to achieve its activation state using MD simulations, MM-GBSA free energy calculations, essential dynamics, and thermodynamic analysis.

MATERIALS AND METHOD

Structure Retrieval

The X-ray structures of the FXR complexes with 'OCA' only with a 'co-activator' and without any binder (APO) (**Supplementary Table S1**) were retrieved from Protein Data Bank (Bank, 2021). The crystal structure of human FXR (PDB-ID: 5Q0K) bound with a 'co-activator', and rat FXR (PDB-ID: 1OSV) which is bound with 'OCA' and 'co-activator' both is used for comparative analysis.

Since the binding of the coactivator SRC1 (KDHQLRLYLLDKD) in human FXR is similar the binding of 'co-activator' GRIP-1 (ENALLRYLLDKD) in rat FXR and both 'co-activators' share the high homology between them (Wang et al., 1998; Soisson et al., 2008). These peptides shared the conserved LXXLL motif in the sequence. There are also crystal structures available for human FXR with GRIP-1 (Kudlinzki et al., 2019; Merk et al., 2019). Therefore, we have considered the 'co-activator' of rat FXR with human FXR for the study to maintain uniformity. Thus, the FXR without 'OCA' and 'co-activator' is System A and the FXR with 'co-activator' is System C. The FXR with 'OCA' is System B and FXR with 'OCA' and 'co-activator' is System D. The 'OCA' without protein is System E. The details of all FXR systems are given in **Supplementary Table S1**. The LBD of FXR consists of 230 amino acids in structure (total length). The residues involved in the interaction are conserved from the comparative analysis of the binding pocket in both human and rat FXR, which were well studied earlier (Downes et al., 2003; Mi et al., 2003; Kemper, 2011).

Protein Structure Preparation

Here, we have explored four systems, APO-protein of FXR (System A), APO+agonist (System B), APO + 'co-activator' (System C), and APO+Agonist + 'co-activator' (System D), to identify the transition dynamics between the different conformational states of FXR with its binding partners. Before MD simulation, each targeted protein structure was prepared using the Protein Preparation Wizard encoded in the Schrodinger 3.5 suite (Sastry et al., 2013; Anang et al., 2018; Sarkar et al., 2021). The crystal waters were also removed, and hydrogens were added. The breaks in the crystal structures were interpolated by using the Prime (Sastry et al., 2013) module of the Schrodinger Suite. The capping was done to the uncapped -N and -C termini of the FXR protein. The hydrogen bond optimization was performed using PROPKA (Rostkowski et al., 2011) at pH7, and the restrained minimizations were also done for the systems using OPLS3 (Optimized Potentials for Liquid Simulations) force field (Harder et al., 2016). To study the sequence similarity between the human FXR and rat FXR crystal structures, the multiple sequence alignment (MSA) was performed by using the PRIME module of Maestro (Proteins, 2004; Mittal et al., 2020).

MD Simulations

All the systems defined above were subjected to MD simulations. The details of the simulated systems are listed in **Supplementary Table S2**. In total, we have generated 6.5 μ s long MD simulations including the triplicates for each system of FXR. The general Amber force field (GAFF) and Amber ff14SB force field were used for ligand, 'co-activator', and protein. The antechamber was used to automatically calculate charges and atom types for the ligand ('OCA') using GAFF (Wang et al., 2004). The different protein systems for FXR were prepared for simulations using the LEaP program implemented in the Amber package (Pradhan et al., 2018). All the energy minimization and MD simulations are carried out by using the *sander* and *pmemd* modules of AMBER16, respectively (Case et al., 2005). In LEaP, the AMBER ff14SB (Maier et al., 2015) force field was assigned to the protein. Counter ions were added to neutralize the system and the protein system was solvated using a TIP3P water model in an orthorhombic box with a span 10 Å from the periphery of the protein. Each system was neutralized by adding counterion ions. Periodic boundary conditions and particle mesh Ewald methods were employed to treat long-range electrostatic interactions (Darden et al., 1993). Hydrogen bonds were constrained by applying the SHAKE algorithm (Ryckaert et al., 1977). The integration time step for all MD simulations was set at 2 fs. The nonbonded cutoff was 8 Å. The solvated models were first minimized with the module *sander* in constant volume by 2,000 cycles of steepest descent minimization followed by 1,000 cycles of conjugate gradient minimization. The systems were then equilibrated for 500ps at 300 K and 1 atm pressure. For MD simulations, isobaric (NPT) conditions were maintained with the target pressure of 1 bar utilizing the Berendsen barostat. The temperature was regulated using a Langevin thermostat. MD was eventually run for 500 ns, and atomic coordinates were saved every 5ps as snapshots. In addition, the study of MD simulation trajectories was carried out. The simulations have been performed

using the GPU version of AMBER16. The last 150ns stable trajectories for all four systems were used for the analysis. To evaluate the stability and dynamics of the FXR systems, triplicate all-atom MD simulations were performed using AMBER 16.

MD Trajectory Analysis

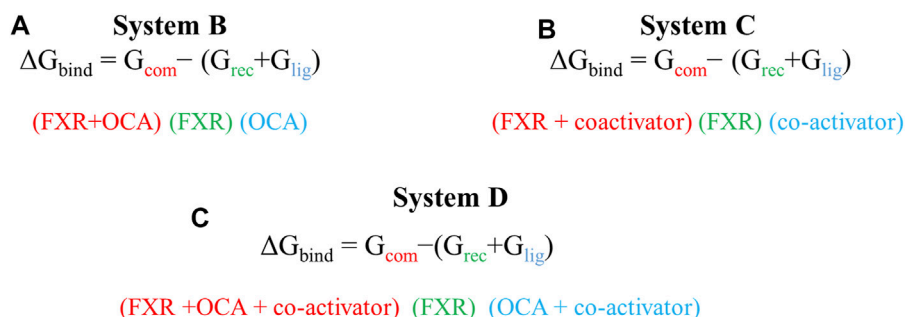
The root mean square deviations (RMSD) of backbone atoms, root mean square fluctuations of Ca atoms (RMSF), salt bridges, solvent-accessible solvent area (SASA), and radius of gyration (Rg) were calculated for whole trajectories by the Tcl scripts implemented in (visual molecular dynamics) VMD (Humphrey et al., 1996) to assess the overall molecular systems stability and fluctuation in the systems. To explore the systems in terms of compactness, the Rg was calculated. SASA was computed for different systems of FXR bound and unbound with 'OCA' and 'co-activator'. Hydrogen bond (HB) analysis was done using CPPTRAJ of AMBER to search the bonds with in the two selection criteria that is an acceptor-donor distance of 3.5 Å, and acceptor ... H-donor ... Angle cutoff is 120°. We have calculated the HB for the stable of MD trajectories of the simulation. The CPPTRAJ of Amber16 was used for secondary structure analysis (DSSP), principal component analysis (PCA) analysis, and dynamic cross-correlation matrix (DCCM) plot. The DCCM map and the DSSP plots were generated by using the Ca atoms of all FXR systems throughout the MD simulation (Roe and Cheatham, 2013; Manjula et al., 2019). Following that, we have used the plugin for Pymol (Molecular Graphics System, Version 2.0 Schrödinger, LLC), xPyder, which is an interface that provides the 3D depiction of cross-correlations between residues in dynamics (Pasi et al., 2012). The graphs were plotted using XMGRACE (Turner, 2005). To calculate cation- π interaction between the residues W466 and H444, the angle between the atoms of CD1@W466-CE1@H444-CH2@W466 were calculated in all systems of FXR (Khandelja and Kaznessis, 2007).

Cavity Volume Calculations

As the pocket analysis is useful for the study of structural dynamics of the proteins, therefore we have performed the pocket volume analysis of all FXR systems with the help of the POVME2 (Pocket Volume Measurer) algorithm (Durrant et al., 2014; Srivastava et al., 2018). All water molecules and counterions were stripped from the trajectory. Thereafter, the trajectory was aligned, and the frames were extracted from VMD for all the systems, which is used as initial input for this method. Next, we defined the inclusion and exclusion regions where the inclusion region entirely encompassed all the binding-pocket conformations of the trajectory while the exclusion region is an area that does not associate with the pocket. In our systems, we chose Ca atoms of residues M325 and F365 that lie at the center of a cavity and protrude inwards to it to define the inclusion sphere. The volume of a whole pocket was calculated by simply summing the individual volumes associated with each grid point in the inclusion spheres.

Free Energy Calculations

The free energies for FXR systems were calculated by using the MM-GBSA method in AMBER tools and AMBER16 (Chipot and



SCHEME 1 | The scheme is used for the calculation of binding energy of 'OCA' in presence and absence of 'co-activator' in both Systems B and D. The total binding energy of the 'co-activator' and 'OCA' for System D.

Pohorille, 2007; Suri et al., 2014; Mittal et al., 2019). For this, the frames were extracted from the most stable state from the MD trajectories of all FXR systems. The binding free energy (ΔG_{bind}) on each system is evaluated by using the following equation:

$$\Delta G_{\text{bind}} = G_{\text{com}} - (G_{\text{rec}} + G_{\text{lig}}) \quad (1)$$

where G_{com} , G_{rec} , and G_{lig} are the absolute free energies of a complex, receptor, and ligand, respectively, arranged over the equilibrium trajectory. The calculations are performed as per **Scheme 1**. The free energy, G , for each species can be calculated by using MM-GBSA and MM-PBSA approaches and can be calculated as follows:

$$G = E_{\text{gas}} + G_{\text{sol}} - TS \quad (2)$$

$$E_{\text{gas}} = E_{\text{int}} + E_{\text{ele}} + E_{\text{vdw}} \quad (3)$$

$$G_{\text{polar, PB(GB)}} = E_{\text{ele}} + G_{\text{sol-polar, PB(GB)}} \quad (4)$$

$$G_{\text{non-polar, PB(GB)}} = E_{\text{vdw}} + G_{\text{sol-np, PB(GB)}} \quad (5)$$

$$G_{\text{sol}} = G_{\text{PB(GB)}} + G_{\text{sol-np}} \quad (6)$$

$$G_{\text{sol-np}} = \gamma \text{SAS} \quad (7)$$

Where G is described as a Gibbs free energy, E_{gas} is the gas phase energy which is the sum of internal energy (E_{int}), electrostatic interaction (E_{ele}), and the van der Waals interaction (E_{vdw}). G_{sol} is the solvation free energy is the sum of polar [$G_{\text{PB(GB)}}$] and nonpolar contributions ($G_{\text{sol-np}}$). It is computed using the parameters defined in the Amber ff14SB force field. $G_{\text{sol-polar, PB(GB)}}$ is the contribution of polar solvents determined by solving the equations Poisson-Boltzmann (PB) and Generalized-Boltzmann (GB) (Genheden and Ryde, 2015). The overall polar contributions were determined as a summation of the contribution from electrostatics (E_{ele}) and polar solvation [$G_{\text{sol-polar, PB(GB)}}$]. The sum of the obtained total nonpolar interaction contributions by E_{vdw} and $G_{\text{sol-np, PB(GB)}}$. $G_{\text{sol-np}}$ is the non-polar solvent contribution measured using $0.0072 \text{ kcal/mol } \text{\AA}^{-2}$ (value of constant γ) and using a water probe radius of 1.4 \AA to determine the solvent-accessible surface area (SASA) (Sitkoff et al., 1994). The dielectric constants were set to 1 and 80, respectively, for solute and solvents. Free energy decomposition in terms of contributions from structural subunits of both binding partners provides insight into the origin of binding on an atomic level.

Principal Component Analysis

In this work, the PCA, also known as essential dynamics (ED) analysis, is used to study the broad concerted motions in FXR-LBD in their bound and unbound state (Kumari et al., 2021; Mittal et al., 2021; Singh et al., 2021). The analysis was carried out to identify the large-scale average motion of an FXR in all systems by the CPPTRAJ module of AmberTools. The frames were taken from the MD simulation trajectories after the evolution of the systems. To obtain the proper trajectory matrix in PCA, the overall translation or rotation motion was removed by fitting the coordinate data to the average structure. Only the backbone atoms were included during the PCA study. The elements of the positional covariance matrix C are defined by the following equation:

$$C_{ij} = \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle \quad (i, j = 1, 2, 3 \dots \dots, 3N) \quad (8)$$

where x_i and x_j are the Cartesian coordinates of the i^{th} and j^{th} Ca atom, N is the number of Ca atoms considered, and $\langle x_i \rangle$ and $\langle x_j \rangle$ represent the time average over all the configurations obtained in the MD simulation (van Aalten et al., 1995; Ivetic and McCammon, 2009). The $\langle \rangle$ sign indicates the ensemble average of the atomic position in the Cartesian space. Major protein motion that contributes to the overall motion was visualized using the Normal Mode wizard plugin in VMD.

Free Energy Landscape

The protein global minimum energy can be derived from Free Energy Landscape (FEL). The FEL represents a mapping of all possible conformations which a molecule can adopt during a simulation, together with their corresponding energy typically reported as the Gibbs free energy. The calculation was carried out using the first two principal components (PC1 and PC2) obtained from individual trajectories. The first two PCs of the respective systems served as reaction coordinates to generate two-dimensional FEL plots for all FXR systems. This was implemented using the g_{sham} module of Gromacs (Van Der Spoel et al., 2005).

$$G\alpha = -kT \ln P(q\alpha) / P_{\text{max}}(q) \quad (9)$$

where k is the Boltzman constant, T is the temperature of simulation, $P(q\alpha)$ estimates the probability density function

obtained from a histogram of the MD data, and $P_{\max}(q)$ is the probability of the most populated state.

Computational Alanine Scanning

We have carried out CAS for the highlighted residue-wise energy decomposition results to confirm the *hot-spot* amino acids in FXR. The calculations were run on the stable MD trajectory by using the MM-GBSA approach. The amino acid of interest is replaced with alanine, and absolute binding free energy is recalculated. Finally, the difference in the binding free energies of the wild type and mutant, $\Delta\Delta G_{\text{bind}}$, was computed as follows:

$$\Delta\Delta G_{\text{bind}} = \Delta G_{\text{bind}}[\text{Wild Type}] - \Delta G_{\text{bind}}[\text{Mutant}] \quad (10)$$

Negative values of $\Delta\Delta G_{\text{bind}}$ indicate the favorable contributions of residues in wild type while positive values indicate the unfavorable contributions. The mutant models of all the *hot-spot* residue were generated by using the maestro module.

RESULTS

As of now, several FXR-LBD crystal structures have been resolved in complex with a range of distinct ligands, which reveals that FXR possesses a highly flexible binding pocket wherein binder dependent conformational changes play an indispensable role to achieve the activation state of FXR (Merk et al., 2019). The reported crystal structures contained only LBD along with agonists/partial-agonists/antagonists and/or co-activators/co-repressors. Since the co-crystal structure represents only a single snapshot of a dynamic binding equilibrium of agonist, 'co-activator', or both, it was not sufficient to gain mechanistic understanding.

It remained unclear whether agonist, 'co-activator', or both altered the internal wiring of FXR and modulated the structure and function. Further understanding of FXR regulation requires a more in-depth knowledge of the interactions between FXR and its binding partner. It is always interesting to explore the structural determinants responsible to convert active protein to inactive or *vice versa*. The conformational changes that occur during the binding or unbinding processes of different binders induce the essential conformational changes which are required for the transitions of the protein from their different biological states. In total, 13 MD simulations were performed as in triplicates for each system of FXR and we report the outcomes from the consensus of the three simulations (triplicates).

Exploration of Conformational Changes in the Presence and Absence of 'OCA'

The structure of 'OCA' comprises one 5-membered ring and three six-membered rings fused in **Figure 1A**. The 2D steroidal rings are named as core region, the OH group as head, and the carboxylic group as a tail region (**Figure 1A**). The 'OCA' displays a convex hydrophobic and a concave hydrophilic face as shown in **Figure 1B**. **Figure 1C** clearly shows that ring A of the 'OCA' faces the C-terminal of helix H12, and this orientation is opposite in other receptors like progesterone, estrogen, testosterone, and

glucocorticoids as their ring D faces the helix H12 (Brzozowski et al., 1997; Shiao et al., 1998; Williams and Sigler, 1998; Sack et al., 2001; Bledsoe et al., 2002).

The superimposition of System C and D has shown an average RMSD of backbone atoms of 1.65 Å, indicating the deviations in the backbone atoms of both systems. We have found that helices (H2, H6, and H12) and loops between [H1 and H2 (L: H1/H2), H2 and H3 (L: H2/H3), H5 and H6 (L: H5/H6), and H11 and H12 (L: H11/H12)] have shown deviations in System D relative to the System C. The loop L: H11/H12 is not crystallized in System C, due to low electron density (**Figure 1D**). It has also been reported that the loop L: H11/H12 is very essential for the stability of the helix H12 position, and the presence of agonists and 'co-activators' makes this loop stable in the whole NR family (Costantino et al., 2005). In System D, the helices H2 and H6 were shortened compared to System C, which resulted in significant variations in the loops (L: H2/H3 and L: H5/H6) in the respective systems (**Figure 1D**). Since these conformational changes in helices and loops reported an impact on binding of 'OCA' and 'co-activator' therefore, the MD simulations were implemented to discern their mechanism of action (**Supplementary Table S1**).

The Dynamical Exploration of FXR in Four Different Systems

The time-dependent RMSD of backbone atoms of each simulated system was used to analyze the stability of the systems. As the simulation progressed, each of the systems evolved for a short period of time, from 5 to 40ns, and after that the systems converged, however, the plateau are achieved after 200ns which is consistent till the end of the simulation (**Figure 2A**). The resulting average values of RMSD are remarkably similar, as expected for each system of three runs, where the backbone atoms of System C appear the most stable (**Supplementary Figure S1**). The RMSD plots suggest that the binding of 'OCA' alone (System B) causes the deviations in the backbone atom from its initial state while the 'co-activator' tries to stabilize the FXR protein both as alone and with 'OCA' (System C and D). The Rg plot showed that System A became more compact during the simulation as compared to other systems (**Figure 2B** and **Supplementary Table S3**). Since System A was formed by removing the 'co-activator', its initial Rg was like other systems until ~100ns but afterward became relatively more compact. This hints that the presence of any binder ('co-activator' and/or ligand) causes conformational changes that decrease the compactness of the protein.

Furthermore, the RMSF calculation (of Ca atoms) was performed to identify the regions with high fluctuations, and their average values are summarized in **Figure 3A** and **Supplementary Table S3**. The overall RMSF profile reflects that all systems have minimal Ca fluctuations with average values ranging from 0.97 Å to 1.18 Å (**Supplementary Table S3**). Upon comparing all FXR systems, fluctuations have been observed in helix H2, and loops L: H5/H6 and L: H9/H10 exhibited higher flexibility in Systems B (**Figures 3B–E**). In System D, the loops L: H1/H2 and L: H2/H3 showed higher fluctuation than the other systems (**Figures 3B,C**). It seems that 'OCA' alone induced fluctuations in helix H2 and in loop areas L: H5/H6, L: H9/H10, whereas the presence of 'OCA' with 'co-

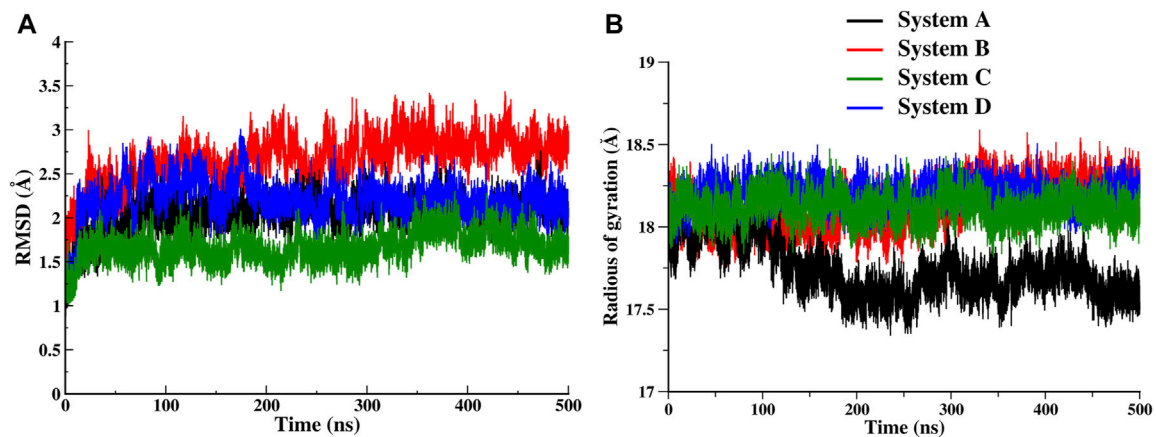


FIGURE 2 | Time series evolution of FXR systems. **(A)** Backbone RMSD during simulation for System A (black), System B (red), System C (green) and System D (blue). **(B)** The Rg plot for all FXR systems.

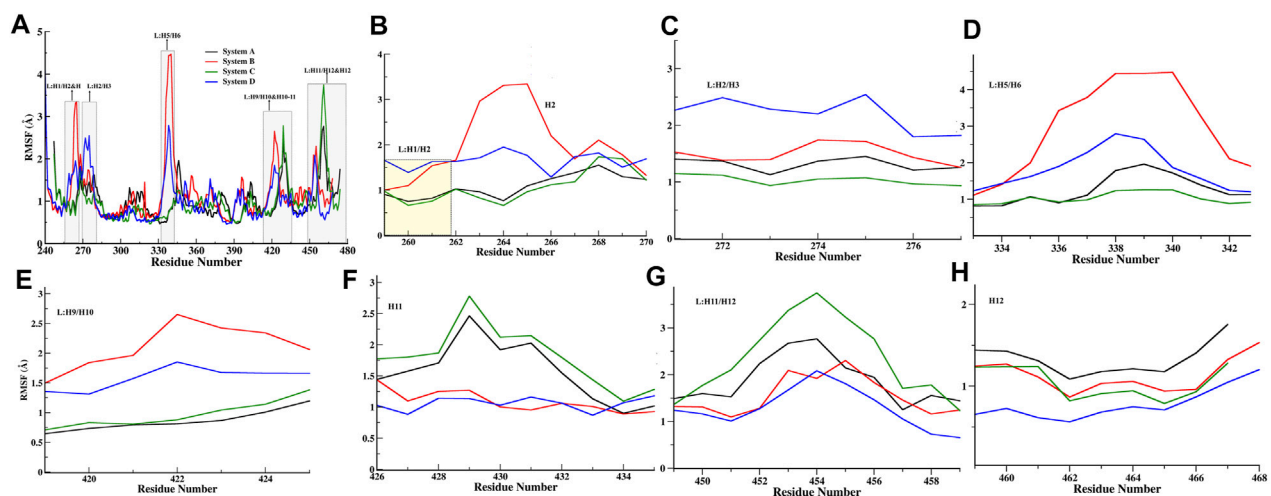


FIGURE 3 | Quantitative analysis of fluctuation from the MD simulation. **(A)** RMSF of $C\alpha$ atoms of all FXR Systems. The values were presented during the 500ns MD simulations time (ns) scale. The different regions of fluctuations observed in the RMSF plot are mentioned. The region-wise RMSF plots were shown in **(B)** for loop L: H1/ H2 and helix H2 region, **(C)** loop between helices 2 and 3 regions (L: H2/ H3), **(D)** Loop between helices H5 and H6 region (L: H5/ H6), **(E)** loop between H9 and H10 (L: H9/ H10), **(F)** helix H11, **(G)** loop between helix 11 and 12 (L: H11/ H12) region, and **(H)** helix H12 regions.

activator' tends to decrease these variations, but the presence of both raises the fluctuation in loops L: H1/H2 and L: H2/H3 (Figures 3B–E). However, the helix H11 and loop L: H11/H12 showed higher fluctuation in System C but these regions experienced the lowest fluctuations in System D (Figures 3F–G). It is also seen that helix H12 has the least fluctuation in System D and highest in System A (Figure 3H). It reflects that binding of 'OCA' significantly minimizes the fluctuations in helices H11, H12, and loop L: H11/H12 in the presence of a 'co-activator'. We also found that the high fluctuation in helix H12 is mainly due to the absence of 'OCA' and 'co-activator' in System A than other systems (Figure 3H).

Thus, the overall analysis demonstrated that the dynamicity of the FXR highly depends upon the binding of 'OCA' and 'co-

activator' that causes the conformational changes in the FXR. This suggests that the agonist is required to induce a conformational shift in helix H12 so that the 'co-activator' can be correctly positioned in the FXR.

Secondary Structural Changes During the Simulation

To assess secondary structural stability, the secondary structure transitions in each of the FXR systems were analysed during the MD simulation (Supplementary Figure S2). We observe that secondary structure content was retained in all the systems, except the residues located in the helices H2, H6 and loops L: H1/H2, L: H2/H3, L: H5/H6, and L: H11/H12 region of FXR as

shown in **Supplementary Figures S3–S5**. The residues of the loop L: H2/H3 and helix H2 form the stable coil during the simulation in Systems A and C (**Supplementary Figure S3**). The residues of the loop L: H1/H2 change from the coil to bend secondary structure in System A and B and form the stable secondary structure in Systems C and D. In System B, the residues of the loop L: H2/H3 interchange turn to bend throughout the simulation. Whereas, in System D, these residues form the bend and turn up to 200ns and eventually form the stable structure until the simulation's end. Similarly, the helices H5, H6 form stable structures in all the systems of FXR (**Supplementary Figure S4**). The residues of loops L: H5/H6 changes from turn or loop to coil and then bends throughout simulation in System B as compared to other systems. The loop L: H11/H12 shows the characteristic fluctuation in the different systems of FXR (**Supplementary Figure S5**). These residue forms a pi helix and bends in System A and B, respectively. However, in System C, the loop residues change from loop to pi helix then bend and eventually gain loop form towards the end. In System D, the residue forms the bend and then eventually regains its turn or loop form during the end of the simulation as depicted in **Supplementary Figure S5**. It is noticeable that this change in regions was not highlighted earlier for 'OCA' (Costantino et al., 2005). The secondary structure analysis indicated the presence of 'OCA' caused the significant conformational changes in the loops forming the binding cavity of FXR. The 'co-activator' binding stabilized these loops and systems showed the minimum secondary structure changes in these regions.

Conformational Flexibility in LBD of FXR

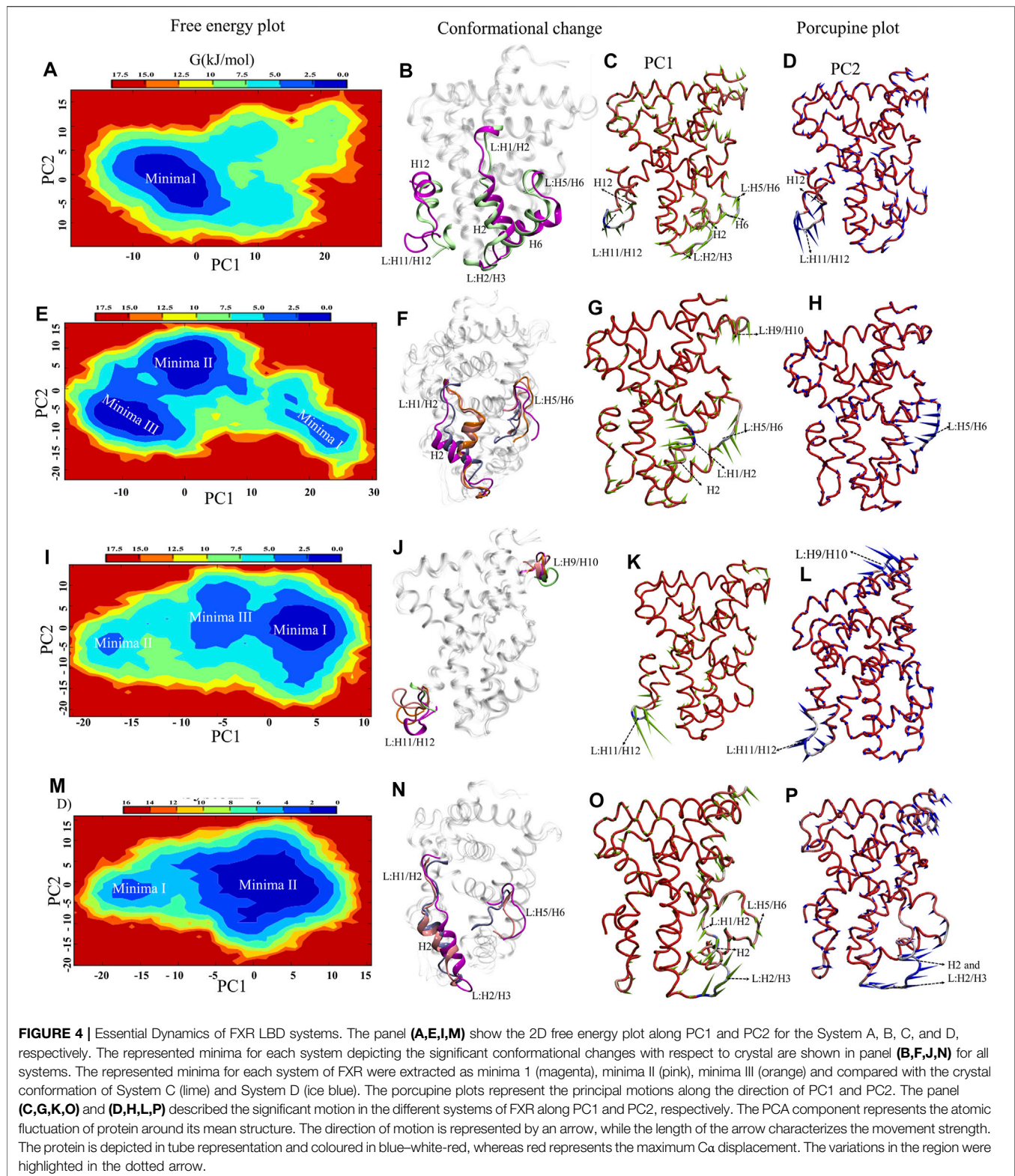
The conformations accessed by LBD are very flexible with different binding partners. To explore it we have performed PCA analysis. The PCA reflects the collective motions of a protein during simulation (Maisuradze et al., 2009). We have shown the cumulative contribution with respect to the PC components for each system (**Supplementary Figure S6A**). It can be observed that overall contributory motion in System B is more than 72% of total fluctuation due to the first 10 PCs, while the top 10 PCs contribute 65, 65, and 64% of total motion respectively in the other Systems A, C, and D (**Supplementary Figure S6B**). This observation suggests that System B shows the highest fluctuation among the other systems as well as RMSD and RMSF plots. On the other hand, System D has shown the least fluctuation, which signifies that both 'OCA' and 'co-activator' binding stabilize the overall FXR systems. In addition, the fractional contribution plot of the top 10 PCs, the first two PCs, PC1 and PC2 appear to capture the notable variations between the systems (**Supplementary Figure S6B**).

Understanding the structural dynamics of FXR complexes is important therefore, we constructed FEL along with the first two PCs as reaction coordinates in the 2D plot that reflect specific properties of the systems and measure conformational variability (**Figure 4**). The size and shape of the minimal energy area (basin: in blue) indicate the stability of a system. Smaller and more centralized basins suggest that the corresponding complex is more stable. Porcupine plots utilizing PC1 and PC2 are

constructed in **Figure 4** to indicate the locations with high atomic fluctuations and their directionality in the FXR simulated systems. System A reflects one deep basin (Minima1) (**Figure 4A**). These basins correspond to the conformational changes in helices H6, H12, and loops L: H2/H3, L: H5/H6, L: H11/H12 (**Figure 4B**) and based on the direction and magnitude of the porcupine vector, the highest fluctuation is seen in loop L: H11/H12 and anticorrelated movement to loops L: H2/H3, L: H5/H6, and helix H6 along PC1 (**Figure 4C**). PC2 captured the highest fluctuation in helix H12 and loop L: H11/H12 (**Figure 4D**). This shows that in absence of any binder, these regions constituting the binding pocket of agonist are flexible and move inward, resulting in reduced gyration and binding pocket volume as in concordance with the above-discussed sections. In System B, we observed three low-energy basins (Minima I, Minima II, and Minima III) along PC1, but the deepest is the Minima II and III (**Figure 4E**).

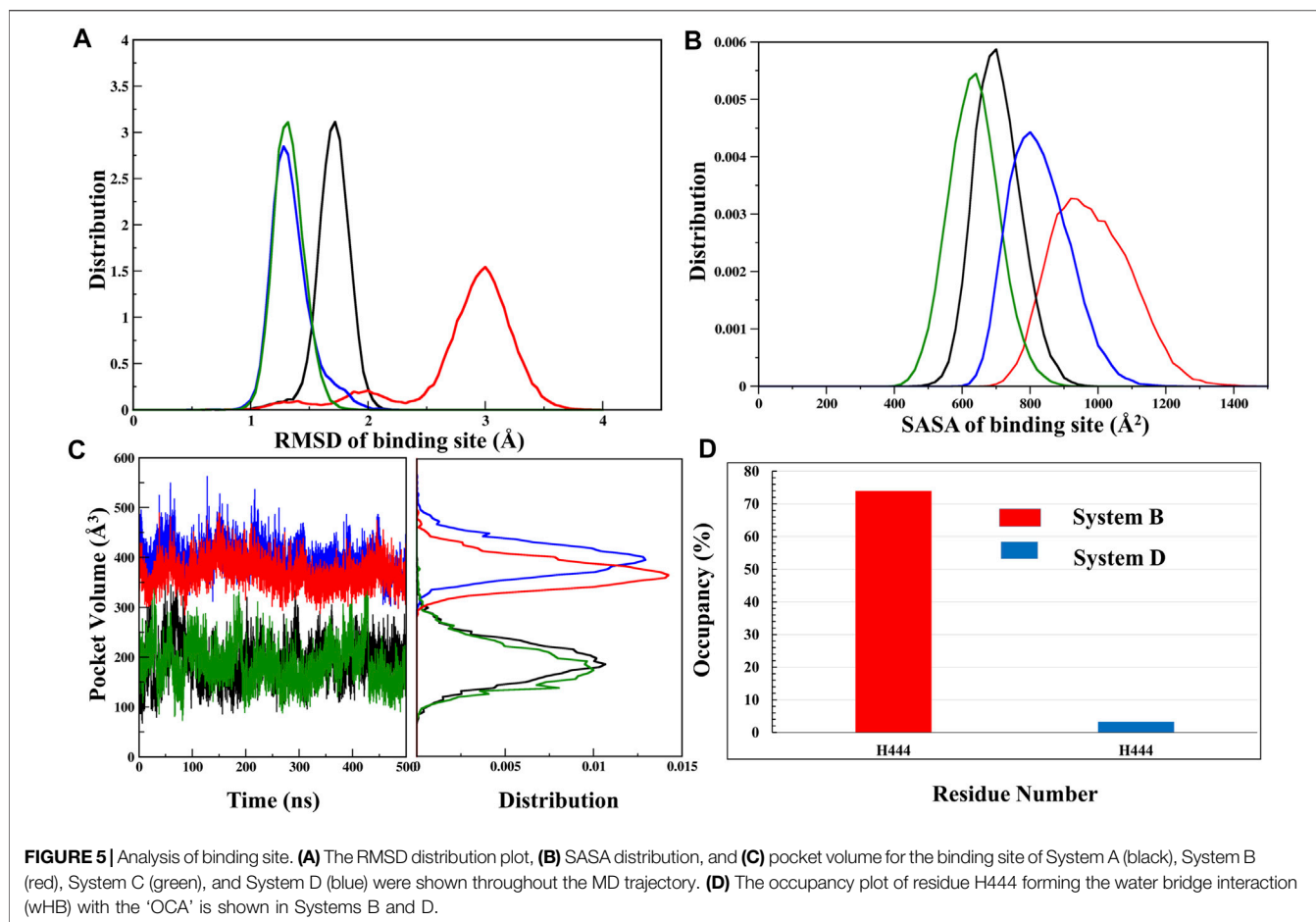
These basins correspond to the conformational changes in loops L: H1/H2, L: H2/H3, L: H5/H6, and helix H2 (**Figure 4F**). In porcupine plots, the helix H2 and loops L: H1/H2 and L: H5/H6 show the anticorrelated movement with each other and capture the highest fluctuation along PC1 (**Figure 4G**). The slight outward movement in loop L: H9/H10 was also observed along PC1. In PC2 the higher fluctuations were captured in loop L: H5/H6 (**Figure 4H**). The superimposition of crystal structure and the representative structures from each minimum were compared in terms of deviation from each other through RMSD analysis (**Supplementary Table S4**). This signifies that the binding of 'OCA' caused the subtle changes in these regions and significant stabilization is seen in the loop L: H11/H12 and helix H12. In System C, the FEL plot revealed three low-energy basins (Minima I, Minima II, and Minima III) along PC1, but the deepest basin was Minima I (**Figure 4I**). The superimposition of the structures reflects the conformational changes in loops L: H9/H10 and L: H11/H12 (**Figure 4J**). Despite the presence of a 'co-activator' in the porcupine plot of System C, it shows the highest fluctuation in the loop L: H11/H12 as compared to the other systems along PC1 (**Figure 4K**). The PC2 shows the fluctuation in loop L: H9/H10 higher (**Figure 4L**). This indicates that the binding of the 'co-activator' and 'OCA' alone causes the internal fluctuation in the FXR which is far away from the binding region of 'OCA' and 'co-activator'. In System D, we observed the two basins (Minima I and Minima II), out of which the minima II is a deep basin along PC2 (**Figure 4M**). Upon comparing conformation changes, we found the significant changes in helix H2 and loops L: H1/H2, L: H2/H3, and L: H5/H6 in System D (**Figure 4N**). The porcupine plot for PC1 shows inward movements in the helix H2, loops L: H2/H3, and L: H2/H3 shows the anti-correlated motion with the loop L: H5/H6 (**Figure 4O**). PC2 captured the highest fluctuation in the helix H2 and loop L: H2/H3 (**Figure 4P**).

In System D the represented structure does not deviate much from the crystal structure as compared to System B (**Supplementary Table S4**). In general, the binding of the agonist to FXR the helix H12 adopts the conformation and



stabilizes ‘co-activator’ peptide binding (Mi et al., 2003). However, the ‘co-activator’ can bind with the FXR in the absence of ‘OCA’ with the weaker binding affinity and causes more fluctuation in loop L: H11/H12 which can be seen in System

C compared to other systems. The ‘OCA’ and ‘co-activator’ binding alone as in Systems B and C bring out the subtle changes in the helix H2 and loop regions of the LBD of FXR, while binding of both to the FXR stabilizes the system.



Binding Site Analysis of 'OCA' With/Without 'Co-Activator'

FXR's LBD comprises a hydrophobic pocket leading to lipophilic molecules such as BAs. As per the previously described results, it is noticed that the binding 'OCA' and 'co-activator' causes the conformational changes in the LBD of FXR. The binding site of FXR is known to have considerable flexibility to accommodate the various chemotypes (Massafra et al., 2018). To discern this dynamic of the binding site of 'OCA', we have explored the binding site RMSD, SASA, and pocket volume throughout the MD simulation. The backbone RMSD distribution plot for the binding site of 'OCA' alone in System B has a wider distribution with multiple peaks with most of the population at nearly 3.0 Å as compared to other simulated systems (Figure 5A, Supplementary Table S5). This confirms the flexible nature of the FXR pocket. The SASA distribution plot of the binding site of System B is found to be more solvent-exposed than the other three systems (Figure 5B, Supplementary Table S5), which signify that the binding of 'co-activator' make the pocket more stable. Further, it is also seen that the presence of 'OCA' and 'co-activator' increased the pocket volume which significantly reduced in System A indicating that the agonist increases the pocket volume of LBD of FXR (Figure 5C, Supplementary Table S5).

Upon multiple sequence alignment of rat and human FXR sequences, the similarity and identity are 96 and 92%, respectively, (we have shown similarity here) however, binding site residues are 100% conserved in rats and humans within 4.0 Å of from the center of 'OCA' (Supplementary Figure S7A) (Downes et al., 2003; Mi et al., 2003). The helices and loops which are involved in the binding are highlighted in Supplementary Figure S7B. The 'OCA' binding is mediated by the 25 residues mainly involving the hydrophobic interaction, among which only 5 residues, R328, S329, Y358, Y366, and H444, are involved in the establishment of the HBs with 'OCA' (Supplementary Figure S7C and Table 1).

Upon superimposition of the binding pocket of Systems C and D, there were significant conformational changes were observed in residues M262, M287, M325, F326, R328, S329, F333, Y358, Y366, M447, and W451 in System D that are responsible to accommodate 'OCA' in the binding pocket of FXR (Supplementary Figure S7D). The changes in configurations of the HBs forming residues in both Systems C and D are shown in Supplementary Figure S7E. We have divided the 2D structure of 'OCA' (Figure 1A) to explore the residue-wise contribution, marked as three main regions, the head includes only one OH group, core (steroidal rings), and tail (carboxyl group). In the crystal (System D), the head moiety was surrounded by residues

TABLE 1 | Interaction analysis of the FXR complex system with 'OCA' within the 4.0 Å area of the pocket.

Helices involved	Types of interaction	Residue Number
L:H1/H2	Hydrophobic	M ²⁶²
H3	Hydrophobic	L ²⁸⁴ , M ²⁸⁷ , A ²⁸⁸
	Polar	H ²⁹¹
H5	H-bond	S ³²⁹ , F ³²⁸
	Hydrophobic	M ³²⁵ , F ³²⁶ , I ²⁸³ , I ³³² , F ³³³
H6	Hydrophobic	L ³⁴⁵ , I ³⁴⁹
L:H6/H7	Hydrophobic	I ³⁵⁴
H7	H-bond	Y ³⁵⁸ , Y ³⁶⁶
	Hydrophobic	I ³⁵⁹ , M ³⁶² , F ³⁶³
H11	H-bond	H ⁴⁴⁴
	Hydrophobic	M ⁴⁴⁷
L:H11/H12	Hydrophobic	W ⁴⁵¹
H12	Hydrophobic	W ⁴⁶⁶

Y358, H444, M447; core region is near to residues I283, L284, V322, M325, F326, S329, F333, I345, I349, I354, I359, M362, F363, Y366, W451, and Y466; while the tail region is lined by residues M262, M287, A288, H291, R328 and I332 (**Supplementary Figure S7D**). The RMSF plot for 'OCA' in Systems B and D showed significant fluctuation in its different functional groups. Although both follow the same pattern, 'OCA' experienced more atomic fluctuation (atom 1–23) in System B than System D (**Figure 6A**).

Furthermore, the comparison of the binding site of the representative structures from the MD simulations sheds light on important residue displacements which are crucial for the binding phenomenon (**Figures 6B,C**). In System D, we observed

that 'OCA' retained the HB interaction with the residues S329 (91.92%), Y358 (76%), Y366 (88.75%), and H444 (88.16%) and lost the interaction with residue R328 with respect to the crystal structure (**Figure 6A, Table 2**). However, in the absence of the 'co-activator', the 'OCA' lost its interaction with the residue H444 (17.34%), Y358, and R328 in System B. It gained water-mediated interaction (wHB) with the residue H444 with the occupancy of 74% as compared to System D (**Figures 5D, 6C and Table 2**). However, in System B the 'OCA' form the HB with the residues S329 (97.75%), and Y366 (92.29%) during the simulation (**Figure 6C and Table 2**). During MD, we found that the new residues P263, Q264, and T267 surround the tail region of the 'OCA' in System B, whereas in System D, the residue I294 is found in the vicinity of 'OCA' (**Supplementary Figures S8A, S8B**), which is not yet reported in previous FXR based studies. This is due to the significant fluctuation in the helix H2, loop L: H1/H2 of System B than the System D. The interaction between 'OCA' and residues M262 and T267 possibly transient, however, seem important for their movements between stable states. We have also observed the time-line conformational changes in the interacting residues M262, T267, Y358, and H444 in both the systems (**Figures 6D,E**). In the case of System D, we observe the least changes in the conformation of the residues Y358, H444, and in 'OCA' as compared to System B, therefore form the stable interaction with it (**Figure 6D**). Both the residues are cryptic in nature as their interactions were missing in the initial state but came into light at intermediate state and eventually got stabilized (**Figure 6D**). In System B the conformational changes in the 'OCA' and the residue Y358 is more from the initial state which causes the loss of interaction between the residue Y358 and the 'OCA' and

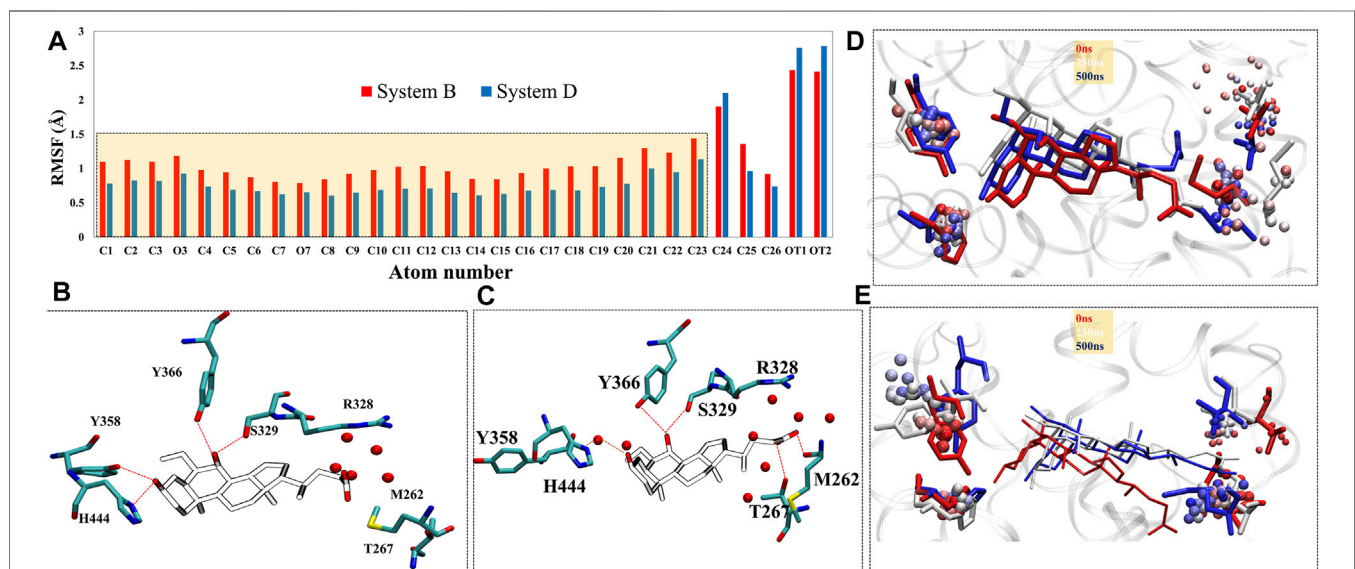


FIGURE 6 | The change in the residual positioning of the LBP (ligand binding pocket) of FXR. **(A)** The ligand RMSF plot for System B (red) and System D (blue) highlighting the difference in their values. The atomic fluctuations in the head and core regions of 'OCA' are higher in System B than System D. **(B,C)** the panels represent HB interaction between the FXR residues and 'OCA' for Systems D and B, respectively. The red dotted line is the distance between the 'OCA' and the residue of the FXR protein. The panels **(D,E)** represent the conformational sampling for these residues in Systems D and B during MD simulation shown in time step coloring method. The representatives of initial state (red: 0 ns), intermediate state (white: 250 ns) and final state (blue: 500 ns) are shown in licorice representation while the beads representation shows the sampling of these residues throughout the trajectory by the stride of 1,000 frames at equal interval.

TABLE 2 | HB Occupancy (cut-off > 50%) of the interacting residues with 'OCA' in both systems B and D.

System B				
Donor	Donor atom	Acceptor	Acceptor atom	Occupancy
Y366	OH	'OCA'	O7	97.75%
'OCA'	O7	S329	OG	92.29%
'OCA'	O3	H444	ND1	17.34%
System D				
Donor	Donor atom	Acceptor	Acceptor atom	Occupancy
'OCA'	O7	S329	OG	91.92%
Y366	OH	'OCA'	O7	88.75%
'OCA'	O3	H444	ND1	88.16%
'OCA'	O3	Y358	OH	76%

gain the transient interaction with the residues M262 and T267 (Figure 6E). The changes in the conformation of 'OCA' give a place for water mediate wHB interaction with the residue H444 and stabilize it in the pocket of System B. HB trajectories depicting time-dependent bond distance variations are illustrated in Systems B and D (Supplementary Figures S8C, S8D). As we observed the ligand and pocket flexibility for FXR, we speculated next about changes in the torsion angle distribution of the 'OCA' (tail region) in Systems B, D, and E, (Supplementary Figures S9A, S9B, see the details in supplementary results Section 3.1). Although the 'OCA' tail region is free to move in System D but unable to form the bond with residues M262 and T267, due to stable core region (1–23) interaction (Figure 6A). This indicates that in the presence of protein and 'co-activator', the 'OCA' behave differently and these differences in angle play a certain role in the conformational diversity of ligand 'OCA' (Supplementary Figures S9A, S9B). This mainly provides insights into the conformational strain undergone to maintain the protein-bound conformation.

Role of Cation- π Interactions Between the Residues H444 and W466 (Activation Trigger Zone)

The stabilization of FXR in active conformation is based upon the interaction established between an aromatic triad tyrosine-histidine-tryptophan (Y358/H444/W466) i.e. called the ("activation trigger") and the ring A of the 'OCA' (Figure 7A) (Gioiello et al., 2014; Massafra et al., 2018). It involves the HB interaction between the residues Y358 (H7) and H444 (H11) with the 'OCA' and the Cation- π interaction between the NE@H444 atom with the center of the indole ring of residue W466 (H12) shown in Figure 7A. It is also known that active conformation of the LBD requires the stability of loop L: H11/H12 than helix H12 which is achieved by the physical constraint in residue H444 (Costantino et al., 2005). The HBs formed between the 3-OH group of 'OCA's and the residue H444 and Y358. These interactions restrict the mobility of residue H444 and stabilize the trigger zone. The loss of their interaction would

remove the necessary support for helix H12 in its active position (Mi et al., 2003). As we have discussed above, the interaction between the residues Y358 and H444 with 'OCA' is more stable in System D than B (Figure 7B). We also found the least fluctuation in the loop L:H11/H12 and helix H12 in System D than the other systems. Secondly, cation- π interactions between the indole ring of the residue W466 and NE2 atom on perpendicularly oriented residue H444 have been known to stabilize the helix H12 (Mi et al., 2003). This is the T-shaped conformation where the two planes are perpendicular, and the angle fluctuates between 45° and 145° (Khandelia and Kaznessis, 2007). To calculate cation- π interaction, throughout the dynamics, we calculate the angle between the atoms of CD1@W466-CE1@H444-CH2@W466. We also computed the Ca distance between the H444:W466 and H444:Y358 residues (Figure 7B). We noticed that the angle in all three Systems B, C, and D fluctuated within a range of 45°–75° during the simulation. In System A, the distance between these atoms increases during the simulation due to which the angle decreases and fails to maintain the required criteria for the angle formation (Figure 7B). This signifies that the angle between the residue H444 and W466 is stable in the presence of both 'OCA' and 'co-activator' in comparison to the APO form of FXR.

We observe that the distribution plot for Ca distance between the residues Y358 and H444 showed the distance is higher in Systems B and D than in Systems C and A (Figure 7C). However, the distance between the residues H444 and W466 is substantially more observed in System A as compared to other systems (Figure 7D). This indicates the binding of 'OCA' and 'co-activator' causing the significant conformational changes in these residues as binding decreased the distance between the residues H444 and W466 and stabilized the cation- π interaction in Systems B, C, and D than System A (Figure 7D). The overall analysis suggested that the binding of both 'co-activator' and 'OCA' to the FXR is necessary for the increased binding affinity. The HB distance pattern between the residues Y358 and H444 with 'OCA' as described above and observed in the crystal structure is maintained during the simulation in System D only and not achieved in Systems B or C.

'Co-Activator' Binding Site Analysis is Essential to Achieve Activation State of FXR

The FXR's LBD acts as a molecular switch after ligand binding, undergoing the conformational changes that result in the recruitment of the 'co-activator' protein by forming the "charge clamp" and a hydrophobic groove that interact with the LXXLL motifs of 'co-activators' (Weikum et al., 2018; Merk et al., 2019). It is reported in the agonistic conformation of FXR, the 'co-activator' (LxxLL motif) is bound by "charge clamp" with residues K300 (H3) and E464 (H12) (Figure 8A) (Merk et al., 2019). Therefore, we explored the conformational residual changes which are responsible for stabilizing helix H12 throughout the dynamics. The 'co-activator' binding surface on FXR comprises the helices H3, H4, H5, and H12. This

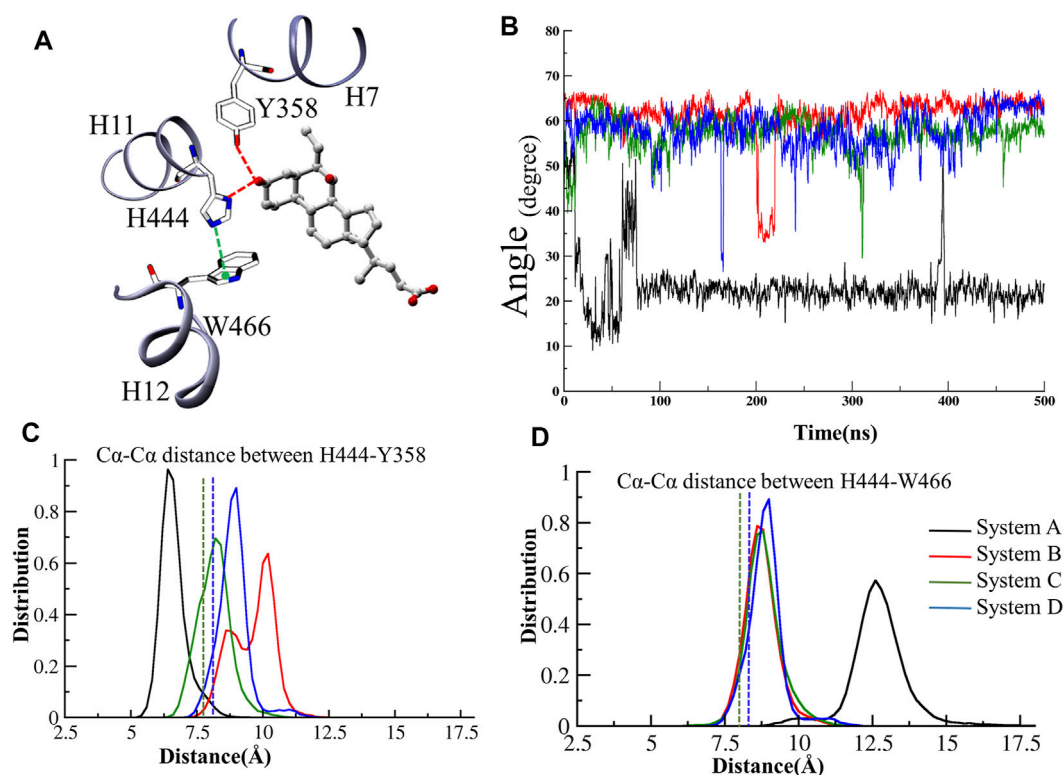


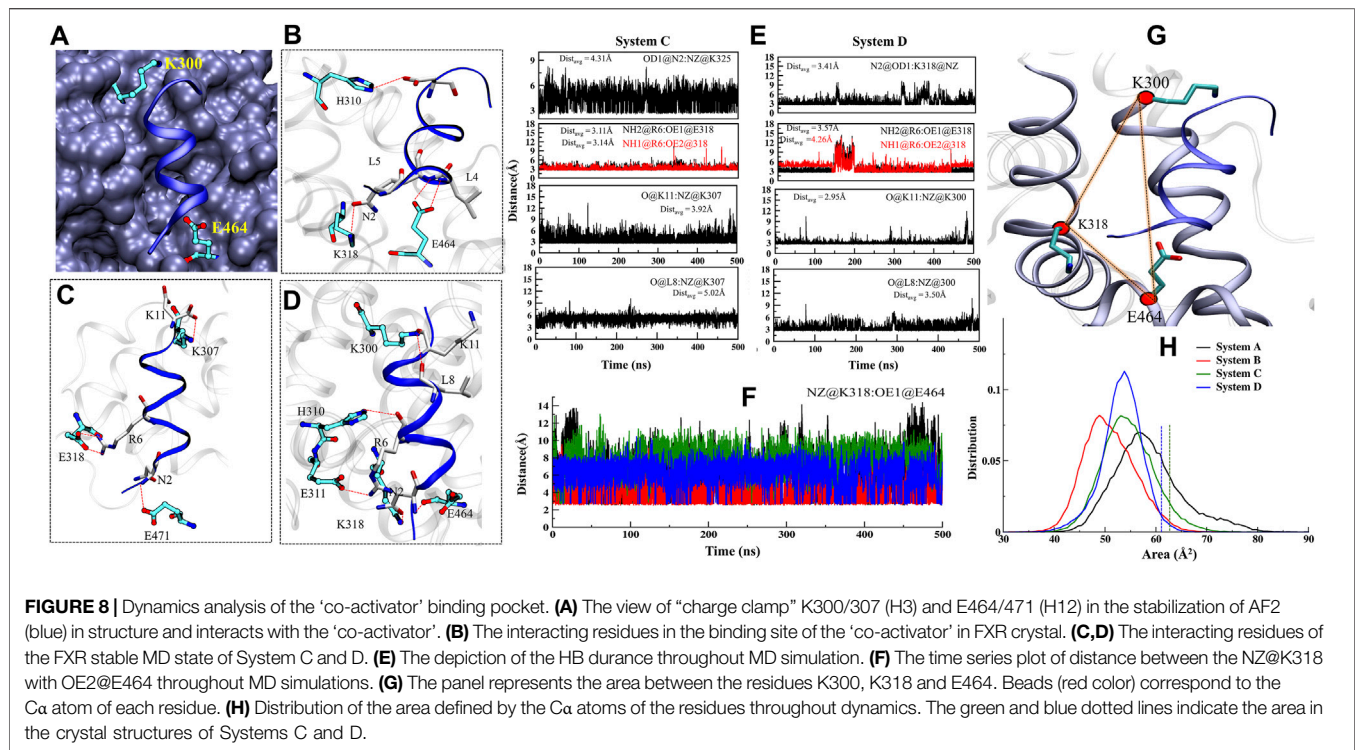
FIGURE 7 | The activation trigger zone analysis. **(A)** The view of the “activation trigger” zone in the FXR structure involves the interaction between a tyrosine-histidine-tryptophan (Y358/H444/W466) triad and the ring A of the steroid bile acid backbone. The HB between the Y358 and H444 with the ‘OCA’ is shown in red dotted line. The cation- π interaction between the NE@H444 (H10-11) atom with the center of indole ring of residue W466 (H12) is shown in green dotted line. **(B)** The time series of angle between the atoms of CD1@W466-CE1@H444-CH2@W466 is depicted. The distribution plot for the distance between C α atoms of **(C)** residues of H444 and W466 **(D)** residues Y358 and H444, throughout MD simulations. The green and blue dotted lines represent the C α distances in the crystal structures of System C and D, respectively.

analysis have shown that, with the binding of agonist in FXR, the ‘co-activator’ typically forms the four HB with the residues K300, H310, E311, and E464 of the FXR (Merk et al., 2019). In System D/C, the ‘co-activator’ forms the HBs with residues K300/307, H310/317, K318/325, E464/471 and non-bonded contacts with the FXR residues Q293/300, V296/303, E297/304, F305/312, Q313/320, I314/321, L317/324, P460/467, L461/468 and E464/471 (Supplementary Figure S10A and Table 3). We have found that in System D, HB interaction between the ‘co-activator’ residues N2, L4, L5, and D10 with the FXR residues K318, E464, and H310, respectively (Figure 8B). We observe that the residue E311 is not the vicinity of the ‘co-activator’ binding site in 4.5 Å in System D (Supplementary Figure S10A and Table 3).

Further to see the residue interaction with the ‘co-activator’, we have analyzed the stable state for Systems C and D. The residues of ‘co-activator’ N2, R6, D12 gain the interaction with FXR in terms of HBs with K325, E318, and K300 residues (Figure 8C) and non-bonded interaction with the I472, T306, I321, I469, T292, V296, V299, L302, V303, Q300, L324 and I321 in System C (Supplementary Figure S9B). In the case of System D, the residues of the ‘co-activator’ N2, R6, K11, and L8 gain the HB interaction with FXR residues K318, E464, H310, E311, and K300 (Figure 8D) and non-bonded with the residues I465, L317,

Q313, I314, T299, R301, Q306, and V292 in System D (Supplementary Figure S10C). The loss of interaction was also observed between the residues L4, L5, D12 with the E464 and K300 (Figures 8C,D).

Further, we analyzed the HB distance during dynamics (Figure 8E). We noticed that the interaction between the ‘co-activator’: FXR atom O@K11: NZ@K300 is most stable throughout dynamics in System D among the other interactions (Figure 8E). While this interaction is unstable in System C. This signifies that possibly the ‘OCA’ helps to establish the interaction with the “charge clamp” residue more stable. The interaction between the O@L8: NZ@K300 in System D is more stable than System C. While the interaction between OD1@D10:NE2@H310 becomes unstable during dynamics in both Systems C and D (Supplementary Figure S9D). In System C, the interaction between the NH2@R6:OE1@E311 and NH1@R6: OE2@E311 is more stable as compared to System D. This could be the region of retaining the ‘co-activator’ in the FXR without the presence of ‘OCA’. In System D, the interaction with OD1@N2:NZ@318 is comparably stable than System C. However, the interaction O@N2:OE2@E464, N@L4: OE1@E464, and N@L5: OE2@E464 is not stable throughout the dynamics in both the



systems (Supplementary Figure S9C). This is interesting to note that the "charge clamp" residues are playing an important role in the recruitment of 'co-activator'. Therefore, we have calculated the area between the Ca atom of residues K300-E464-K318 in all systems of FXR (Figure 8G). We take CA@E464 as an anchor residue linking the K300 and K318 residues. Throughout MD simulation, we measured the region for all FXR systems to see shifts in the "charge clamp" forming residue presence and absence of 'OCA' and 'co-activator'. During the simulation timescale, the conformational changes result in a dramatic expansion in the area of the clamp in System A with respect to other systems (Figure 8H). This increase in area is due to the high flexibility of the Ca atom due to the absence of the 'co-activator'. We found, however, that the region for System D is least extended; this could be due to the presence in the FXR of both 'OCA' and 'co-activator' binding. System C also has less distribution area than Systems A and B, which supports that, in the absence of a ligand, the FXR can retain the 'co-activator'. While this area

estimation approach is not a precise procedure, it still provides tentative details on the selection of "charge clamp" residues to explore the co-activator/co-repressor binding site with or without binding of 'OCA'. Laying down an assumption, we could propose that the agonist binding to the FXR is always required for the strong binding of the 'co-activator' to the FXR. From MD analysis, we found that the agonists alone or 'co-activator' can bind and retain in their binding site as we have achieved confident data about their binding. However, they are unable to achieve their active state.

Per-Residue Wise Free Energy Contributions to Identify the Critical Residues in FXR Binding Free Energy of 'OCA' and 'Co-Activator' in FXR

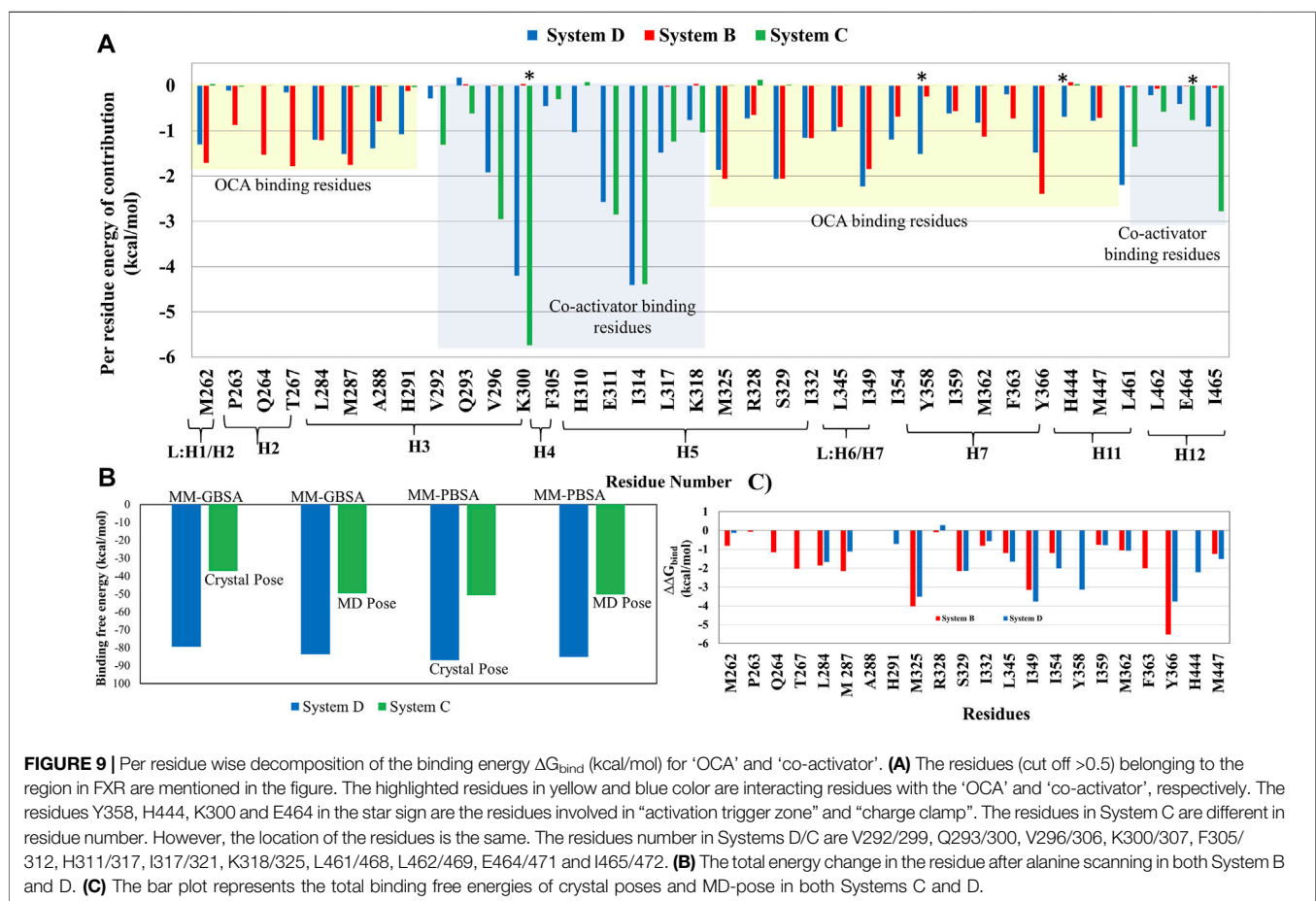
The total binding free energies calculated as per Scheme 1 are listed in Table 4. The total ΔG_{bind} of 'OCA' (System B) in the absence of a 'co-activator' is -30.45 kcal/mol, by using the MM-PBSA method (Table 4). The total ΔG_{bind} of the 'co-activator' in the absence of 'OCA' (System C) -50.14 kcal/mol by using the MM-PBSA method. However, the total ΔG_{bind} of 'OCA' and 'co-activator' in FXR (System D) is -86.83 kcal/mol, and their MM-GBSA values are listed in Table 4. The per residue energy of each contributing residue is given in Figure 9A. Here, we noticed that the total ΔG_{bind} of 'OCA' is increased in the presence of a 'co-activator'. This reflects that binding of 'OCA' is more energetically favored upon binding of 'co-activator'. The higher contribution to the ΔG_{bind} in the presence and absence of 'co-activator' in Systems B and D is due to the difference in the $\Delta G_{\text{solvent}}$ GB

TABLE 3 | Interaction analysis of FXR with peptide within 4.5 Å.

Helices involved	Types of interaction	Residue number
H3	Hydrophobic	V ²⁹⁶
	Polar	Q ²⁹³ , E ²⁹⁷ , K ³⁰⁰
H4	Hydrophobic	F ³⁰⁵
H5	H-bond	K ³¹⁸
	Hydrophobic	L ³¹⁷ , I ³¹⁴
	Polar	Q ³¹³ , H ³¹⁰
H12	Hydrophobic	P ⁴⁶⁰ , L ⁴⁶¹
	H-bond	E ⁴⁶⁴

TABLE 4 | The contribution of the binding free energy for 'co-activator' in Systems B, C and D. In the bracket, the standard deviation and the standard error of mean values are specified. The standard error of mean values (i.e., the standard deviation divided by the square root of the number of snapshots) to depict the precision of the MM-GBSA and MM-PBSA methods to estimate the binding free energies. The bold values indicates the total binding free energy.

Contribution	System B ('OCA' only)	System C (only 'co-activator')	System D ('OCA'+ 'co-activator')
ΔE_{int}	0	0	0
ΔE_{vdw}	-53.99 (3.22 \pm 0.05)	-53.54 (4.89 \pm 0.08)	-108.91 (6.45 \pm 0.11)
ΔE_{ele}	-182.24 (11.93 \pm 0.22)	-209.92 (35.81 \pm 0.65)	-301.84 (38.30 \pm 0.60)
ΔE_{GB}	201.77 (11.66 \pm 0.22)	221.75 (33.52 \pm 0.61)	339.98 (35.39 \pm 0.64)
ΔE_{surf}	-7.23 (0.34 \pm 0.006)	-8.93 (0.57 \pm 0.01)	-16.06 (0.83 \pm 0.01)
ΔG_{gas}	-236.24 (12.75 \pm 0.23)	-263.47 (36.18 \pm 0.66)	-410.75 (38.84 \pm 0.70)
$\Delta G_{solv\ GB}$	194.54 (11.53 \pm 0.21)	212.82 (33.33 \pm 0.60)	323.92 (35.39 \pm 0.64)
ΔG_{GB}	-41.67 (3.45 \pm 0.06)	-50.65 (6.22 \pm 0.11)	-86.83 (7.86 \pm 0.14)
$\Delta G_{solv\ PB}$	205.78 (13.85 \pm 0.25)	-53.54 (4.89 \pm 0.08)	325.60 (36.24 \pm 0.66)
ΔE_{PB}	210.71 (13.83 \pm 0.25)	220.06 (33.35 \pm 0.60)	336.56 (36.39 \pm 0.66)
$\Delta E_{n-polar}$	-4.92 (0.10 \pm 0.001)	6.73 (0.26 \pm 0.004)	-10.96 (0.37 \pm 0.006)
ΔG_{PB}	-30.45 (6.63 \pm 0.12)	-50.14 (7.95 \pm 0.14)	-85.15 (10.90 \pm 0.19)



and $\Delta G_{solv\ PB}$. Besides, the ΔG_{bind} differences in System C and D appeared due to the electrostatic interactions (ΔE_{ele}) in the gas-phase and polarization contributions (ΔG_{pol}), indicating that the two energetic components had remarkable effects on the binding free energy between 'co-activator' and FXR. The calculated total ΔG_{bind} for the crystal poses of Systems C and D is -49.54 kcal/mol and -83.65 kcal/mol, respectively by using the MM-PBSA method (Figure 9B). In terms of dynamics, it is observed that the

binding free energy of System D is even improved to crystal pose, indicating the possibility of better structural fit is achieved during the simulation.

Per Residue Wise Energy Contribution in FXR-'OCA' Interactions

Analysis of the residue wise free energy decomposition was also carried out to analyze the individual energetic

contributions of each residue involved in the stabilization of protein-ligand complexes. To understand the interactions at the atomic level, binding free energy contributions were determined for each residue for Systems B and D (Figure 9A, Supplementary Table S6). The residues having a contribution of (-0.5 kcal/mol) or above were considered *hot-spot* amino acids and were positioned to contribute most to the stability of the complex. As per the cutoff of the residues M262, P263, Q264, T267, L284, M287, A288, M325, R328, S329, I332, L345, I349, I354, I359, M362, F363, Y366, and M447 have shown high energy contribution in System B (Figure 9A). In the case of System D, the residues M262, L284, M287, A288, H291, R328, S329, M325, I332, L345, I349, I354, Y358, I359, M362, Y366, H444, and M447 has shown the highest contribution in the presence of 'co-activator'. Most of the residues are common in both the systems except the M262, T267, L284, and F363 in System B and H291, H444, and Y358 in System D. The residues M262, Q264, T267, and S329 make remarkably high free energy contributions hence, making a considerably enormous contribution to the overall binding free energy of System B and in System D the residues M325, S329, and I349 contribute more to the overall binding energy. To determine the detailed contribution of each important residue, the binding energy was decomposed into electrostatic, VdW, solvation (polar and nonpolar), and total contribution (Supplementary Table S6). The thermodynamic profiling suggests that the electrostatic and VdW are the major contributors to the 'OCA' net binding. The residue R328, which forms a HB with 'OCA' in the crystal structure, reveals an unfavorable contribution towards the total binding free energy, as this interaction was not sustained in both the systems. In System B the 'OCA' also gained interaction with residues M262 and T267 which can see their higher contribution in System B than D. Thus, the thermodynamic profiling suggests that the contribution of the residue plays a major role in the binding of the 'OCA' to FXR. The analysis revealed that 'OCA' is stable and gains substantially favorable interactions with the pocket residues. However, we further perform the CAS to elucidate their impact on the binding energy of the systems.

Key Residue Contributions in FXR and 'Co-Activator' Interactions

We have calculated residue-wise decomposition to identify critical residues involved in protein-'co-activator' interaction in Systems C and D and take the same cut-off, shown to be above -0.5 kcal/mol (Figure 9A, Supplementary Table S7). We observed the "charge clamp" residue K300/307 and the highest contribution in total binding energy in the presence (System D) and absence of 'OCA' (System C) (Figure 9A). Besides this, we found the residues V292/299, V296/306, E311/318, I314/321, L317/324, K318/321, L461/468, I465/472 contributed higher to the binding energy (-1.0 kcal/mol). This indicates that the hydrophobic residues facilitate the repacking of the helix H12 as the non-polar residues V296/306, I314/321, L461/468, I465/472 contribute the above -2.0 kcal/mol to total binding energy in

FXR (Figure 9A). These values indicated the possibility that the 'co-activator' can bind to FXR in the absence of 'OCA'.

Cross-Validation of Residue Wise Contribution in the Stability of 'OCA' via Computational Alanine Scanning

To accomplish the contribution of the identified residues to the total free energy, we performed computational alanine scanning. The obtained results indicate that the mutation in residues has significantly dropped the binding energy by more than -1.0 kcal/mol (cutoff) in both the complexes (Figure 9C). In residues M287, S329, M325, I349, and I354 typical in the presence and absence of a 'co-activator', a substantial decrease in binding energy (-2.0 kcal/mol) was observed. The residues T267 and F363 had a significant reduction in binding energy while they were interacting in presence of 'OCA' alone. The residues Y358 and H444 form the direct interaction with the 'OCA' in the presence of a 'co-activator' and have a significant drop in the (>-2.0 kcal/mol) binding energy. Therefore, one can infer the key *hot-spot* residues T267, M325, I349, Y358, S329, F363, Y366, and H444 important for the 'OCA' recognition mechanism in FXR. The presence of a 'co-activator' establishes a stable interaction of 'OCA' with the FXR, which is responsible for the activation mechanism of FXR.

DISCUSSION

To unveil the binding event of 'OCA' and 'co-activator' at its functional level, we have evaluated the four systems of FXR and the possible mechanism for activation at the molecular level by using the triplicates of MD simulations.

The conformational change of FXR-LBD in response to different molecular binding, such as agonist and partial agonist is significant for recruitment of 'co-activator' protein and release of co-repressor. Several studies have been proposed to analyze the interactions of FXR with agonist, antagonist, or with and without co-through molecular modeling (Costantino et al., 2005; Meyer et al., 2005; Zhang et al., 2006; Di Leva et al., 2013). However, here, we tried to speculate how the active state of FXR has been obtained at the structural conformational level and how this internal motion helps to modulate the specific region, which provides the specific platform for activation of FXR in synergy between the binding sites of 'co-activator' and agonist.

Perturbed Mobility of Loop L: H11/H12 is Essential for the Activation of LBD

During the simulation, the systems with the binding partner as a 'co-activator' alone and with 'OCA' remained remarkably stable. The compactness in the system without 'co-activator' and 'OCA' signifies that the binding of both disturbs the internal dynamic behavior of FXR (Figure 2B). Compared to Systems A and C, the binding of both 'OCA' alone and with a 'co-activator' had significantly decreased the RMS fluctuation in loop L: H11/

H12 and helix H12 (**Figures 3G,H**). This indicates that the binding of ‘OCA’ is essential for the conformational changes at helix H12. The DCCM map revealed that the helix H3 shows the correlated motion with loop L: H11/H12 and helix H12 in presence of ‘OCA’ (System B) and both ‘OCA’ and ‘co-activator’ (System D) (**Supplementary Figure S11**, see the details in supplementary results Section 3.2). The enhanced correlation in the agonistic conformation comes from the increased stability of the loop L: H11/H12 and the helix H12 whose stability is critical to maintain the agonistic conformation. This is in concordance with RMSF results as well where the fluctuations in helix H12 and loop L: H11/H12 get reduced in the presence of ‘OCA’ and ‘co-activator’ both. The essential dynamics also reveal that the presence of ‘OCA’ and ‘co-activator’ maintain the stability in loop L: H11/H12, which is not found in the presence of either ‘co-activator’ or agonist alone (**Figures 4J–L**). Since the loop L: H11/H12 controls the flexibility of helix H12 and a critical determinant for its orientation (Costantino et al., 2005; Merk et al., 2019). Therefore, the study of this region is essential to understand the mechanism of different types of ligand binding in FXR.

Flexibility Allows Reaching the Activate State Conformation by Modulating *via* ‘OCA’ at Agonist and ‘Co-Activator’ Binding Sites

We observed that ‘OCA’ and ‘co-activator’ are substantially stable according to the different analyses. The binding of both either only ‘OCA’ alone or with a ‘co-activator’ caused the higher fluctuation in helix H2 and loops between L: H1/H2, L: H2/H3, L: H5/H6, and L: H9/H10 regions which signify the flexible nature of LBD of FXR. However, the ‘co-activator’ binding stabilizes the helix H2 and loop L: H2/H3 with ‘OCA’. This is in concordance with secondary structure analysis results where the conformational flexibility of these regions is higher in the presence of ‘OCA’ (**Supplementary Figures S3–S5**). Furthermore, we found that in the presence of ‘OCA’ there is an anticorrelated movement in the helix H2 and loops L: H1/H2, L: H5/H6 region of FXR, which is absent in the presence of ‘co-activator’ alone. These changes in loop conformation account for the increase in SASA and RMSD values of the binding site in the presence of ‘OCA’ alone which stabilized in the presence of a ‘co-activator’. The volume of the LBD pocket was shown to vary significantly during the simulation and revealed the flexible nature of the pocket of FXR. Combined with RMSF analysis of the ‘OCA’, one possible explanation is that the core region of the ‘OCA’ is stable in the presence of a ‘co-activator’, and just increased the fluctuation at the tail region of ‘OCA’. The binding of ‘OCA’ induced the significant expansion of the LBD, which is why for its increased pocket volume.

Changes in Hydrogen-Bond Network Upon ‘OCA’ and ‘Co-Activator’ Binding

The HBs between a protein and ligands provides directionality and specificity of interaction, an important aspect for molecular recognition. The considerable changes were observed in the binding pocket of FXR in the presence of ‘OCA’ and ‘co-

activator’. In our study, the residues S329 and Y366 are noticed to be common in the presence and absence of a ‘co-activator’ (occupancy >50%) **Table 2**. It is surprising that, in the absence of a ‘co-activator’, the Y358 and H444 residues lose their contact completely with ‘OCA’ during the simulation and stay in the pocket. This is attributable to the establishment of the stable association of wHB with H444 residues (occupancy >70%) (**Figure 5D**). In the crystal structure of ‘OCA’ the HB is formed with R328, which is lost during the simulation, and its tail region also forms the transient interaction with the M262 and T267 residues, which are absent in the presence of a ‘co-activator’. The ‘co-activator’ interacting residues with FXR is also more stable in System D than System C (**Figure 8E**). Free energy per residue decomposition and alanine scanning confirm the contribution of the key residues maximum in System D. However, the comparable energy contribution from Systems B and C indicates that achieving the critical orientation is important for the individual residue.

The Role of “Activation Trigger Zone” and “Charge Clamp” in Stability of Helix H12

We noted in the presence of ‘OCA’ and ‘co-activator’ the persistent formation of the angle between the residues Y358, H444, and W466; while in apo these residues are unable to form this angle during dynamics. The α distance between them is also found to be stable in presence of the ‘OCA’ and ‘co-activator’ compared to APO. Costantino et al. reported that the HB interaction between the ‘OCA’ and residues Y358 and H444 is not sufficient to stabilize the helix H12 since it is already in an active conformation in the absence of ‘OCA’, and there is stable interaction between the residues H444 and W466 (Costantino et al., 2005), and the same was achieved in our studies. Interestingly, we found that in the presence of ‘OCA’, the wHB played an important role in restricting the movement of residue H444 in the LBP of FXR and stabilizing the helix H12 conformation. The residue of “charge clamp” (K300) formed the stable interaction with the residues L8 and K11 of the ‘co-activator’ in System D than C, which confirms that the binding of ‘OCA’ enhances the association of the ‘co-activator’ with FXR (**Figure 8E**). It has been also reported that in the absence of an agonist, the ‘co-activator’ is inaccessible to FXR due to the formation of a salt-bridge between the residues K318 and E464 (Costantino et al., 2005). But in our result, we confirm that there is no stable salt bridge formation between the residues K318 and E464 during the simulation (**Figure 8F**). Henceforth, the FXR can be bound with the ‘co-activator’ in the absence of ‘OCA’ but the stability of loop L: H11/H12 is necessary for stabilizing helix H12, which is not consistent.

Key Feature Determining the Binding of ‘OCA’

The quantitative characterization of binding free energies of specific residues in protein–ligand binding is critical as these residues are capable of modulating the internal wiring from function to non-function state and vice-versa. Here, we have elucidated the key interaction captured by the ‘OCA’ in the presence and absence of a ‘co-activator’. Binding free energy

calculations suggest the 'OCA' affinity is highest with 'co-activator' binding to FXR. Based on the results of per residue binding free energy decomposition, we can observe that the number of residues stabilizing the complex as well as the energetic weight of each interaction contributes to the main differences in the total binding free energy. 'OCA' is well stabilized in the presence of a 'co-activator', as a result of forming similar interactions with comparable per residue-free energy contributions to the total binding free energy.

The residues L284, M325, S329, I345, I349, I354, I359, M362, F363, Y366, and W466 occupied the core region of 'OCA' have shown the higher contribution towards the binding energy alone and with the presence of a 'co-activator'. The residues P263, Q264, T267 near the tail region of 'OCA' which gain interaction during dynamics have a higher contribution in 'OCA' at System B. This signifies the 'OCA' stability in the pocket of FXR in absence of a 'co-activator'. As reported earlier, the 6 α -ethyl group (head region) in 'OCA' binds into the hydrophobic cavity that exists between the side chains of residues I359, F363, and Y366 increases the affinity of 'OCA' (Pellicciari et al., 2002). We also found that these residues show a substantial contribution towards the total binding energy in the presence of 'OCA' and confirmed the important role of residues in the molecular recognition of 'OCA' in the FXR pocket. In the presence of a 'co-activator', the residues Y358 and H444 contribution is higher along with the other residues. However, we did not find any contribution from the residue F363 in presence of a 'co-activator'. This means that 'OCA' governs the stability in the FXR pocket in the presence and absence of 'co-activator' differently and only 'co-activator' binding is required for agonist discovery.

Key Feature Determining the Binding of 'Co-Activator'

The thermodynamic profiling suggests that the electrostatic and VdW are the major contributors in the net binding of the 'co-activator' in the presence and absence of 'OCA'. The residues of the "charge clamp" formation play an important role in the co-activator-interacting surface, exist in many NRs, and can stabilize their active conformations (Nolte et al., 1998; Wang et al., 2017). In our study, we also noticed the stable interaction with the "charge clamp" residue K300/307 and the highest contribution in net binding energy (Figure 9A) in Systems D and C. However, the lowest contribution came from the residue E464. Furthermore, Merk et al. had studied that the unliganded form of FXR was also able to recruit the 'co-activator' (Merk et al., 2019). This result also suggested that the contribution of residue is substantial in absence of 'OCA' and the 'co-activator' can bind with the FXR.

Binding Hot Spot for 'OCA'

Based on the consistent information of interaction analysis and CAS, a significant drop in the binding energy more than -2.0 kcal/mol were noticed in the residues T267, M287, S329, M325, I349, I354, Y358, F363, and H444 in the presence and absence of 'co-activator' (Figure 9C). Overall, these data indicate that to achieve a specific active state of FXR certain residue orientation must be targeted to activate the FXR agonism. Overall, conformational, and residual synergy has been observed between agonist and 'co-activator' binding sites. The

RMSD, SASA plots, and distance variation between residues H444-Y358 and residues H444-W466 reflect the binder-dependent dynamical adjustment at the architectures of binding sites that correlated well with thermodynamic outcomes. This indicates that both sites and their residual position must be considered to improve and discover modulators.

CONCLUSION

The complete activation of FXR by OCA blocks the BAs synthesis and hinders metabolic cholesterol degradation. As a result, studying FXR conformational changes in the presence and absence of 'OCA' and 'co-activator' seems essential to explore. Here, we have leveraged our understanding in molecular association between binding sites of 'co-activator' and agonist using detailed dynamics analysis of four comparable systems.

MD simulations divulged profound shifts of the different helices in the FXR systems. Our work mainly explores the binding mechanism of 'OCA' in FXR. Further, correlation analysis reveals that a global network of the correlated motions exists in the FXR, whose components include all regions identified so far to be critical for the binding of 'OCA'. The increase in ΔG_{bind} energy suggested that the presence of a 'co-activator' increased the binding affinity of 'OCA' with FXR. The ΔE_{ele} energy is more favorable in presence of both 'OCA' and 'co-activator' alone. However, the ΔE_{ele} is most favorable in binding 'OCA' with a 'co-activator'. The CAS analysis further confirms the individual contribution to the total binding energy. Our results pointed to residues M262, T267, M287, M325, S329, I349, Y358, Y366, and H444, in an FXR, found to be more crucial for binding of 'OCA'. The agonist and 'co-activator' binding with FXR, an activation state in which the loop L: H11/H12 and helix H12 are completely stabilized and the interactions remain intact to keep the architecture of their binding pockets. The lack of 'OCA' in the binding pocket of FXR makes loop L: H11/H12 extremely unstable. However, the 'co-activator' binds to FXR. It implies that to keep this loop stable, the 'OCA' binding is necessary. In the absence of a 'co-activator', the 'OCA' loses its significant interaction with the residues Y358 and H444 which is necessary for the "activation trigger" in FXR. Thereby improving our understanding of 'OCA' and 'co-activator' binding sites in FXR provide a promising basis for future agonist discovery. Overall, the conformational characterization and dynamical synergy between the binding sites and residues of the 'co-activator' and agonist could be explore further for better mechanistic understanding.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SA conceived and designed the study. AK performed and analyzed MD and ED simulations. LM and MS help with the

experiment and analysis. DP provided input in understanding the biology of the FXR systems. AK and SA wrote the manuscript. All authors have read and approved the manuscript.

ACKNOWLEDGMENTS

We thank the Translational Health Science and Technology Institute for funding this work. The authors acknowledge the

reviewers for their valuable suggestions and comments that have improved the content and presentation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.658312/full#supplementary-material>

REFERENCES

- Aalten, D. M. F. v., Findlay, J. B. C., Amadei, A., and Berendsen, H. J. C. (1995). Essential Dynamics of the Cellular Retinol-Binding Protein Evidence for Ligand-Induced Conformational Changes. *Protein Eng. Des. Sel* 8, 1129–1135. doi:10.1093/protein/8.11.1129
- Abenavoli, L., Falalyeyeva, T., Boccuto, L., Tsyryuk, O., and Kobylak, N. (2018). Obeticholic Acid: A New Era in the Treatment of Nonalcoholic Fatty Liver Disease. *Pharmaceuticals* 11, 104. doi:10.3390/ph11040104
- Anang, S., Kaushik, N., Hingane, S., Kumari, A., Gupta, J., Asthana, S., et al. (2018). Potent Inhibition of Hepatitis E Virus Release by a Cyclic Peptide Inhibitor of the Interaction between Viral Open Reading Frame 3 Protein and Host Tumor Susceptibility Gene 101. *J. Virol.* 92. doi:10.1128/JVI.00684-18
- Arab, J. P., Karpen, S. J., Dawson, P. A., Arrese, M., and Trauner, M. (2017). Bile Acids and Nonalcoholic Fatty Liver Disease: Molecular Insights and Therapeutic Perspectives. *Hepatology* 65, 350–362. doi:10.1002/hep.28709
- Aranda, A., and Pascual, A. (2001). Nuclear Hormone Receptors and Gene Expression. *Physiol. Rev.* 81, 1269–1304. doi:10.1152/physrev.2001.81.3.1269
- AuthorAnonymous (2020). Intercept's NASH Hopes Dashed. *Nat. Biotechnol.* 38, 911. doi:10.1038/s41587-020-0638-5
- Bank, R. P. D. (2021). RCSB PDB: Homepage. Available at: <https://www.rcsb.org/>. (Accessed June 8, 2021).
- Bledsoe, R. K., Montana, V. G., Stanley, T. B., Delves, C. J., Apolito, C. J., McKee, D. D., et al. (2002). Crystal Structure of the Glucocorticoid Receptor Ligand Binding Domain Reveals a Novel Mode of Receptor Dimerization and Coactivator Recognition. *Cell* 110, 93–105. doi:10.1016/s0092-8674(02)00817-6
- Bowlus, C. (2016). Obeticholic Acid for the Treatment of Primary Biliary Cholangitis in Adult Patients: Clinical Utility and Patient Selection. *Hmer* 8, 89–95. doi:10.2147/HMER.S91709
- Brzozowski, A. M., Pike, A. C. W., Dauter, Z., Hubbard, R. E., Bonn, T., Engström, O., et al. (1997). Molecular Basis of Agonism and Antagonism in the Oestrogen Receptor. *Nature* 389, 753–758. doi:10.1038/39645
- Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., et al. (2005). The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* 26, 1668–1688. doi:10.1002/jcc.20290
- Chiang, J. Y. L. (2017). Bile Acid Metabolism and Signaling in Liver Disease and Therapy. *Liver Res.* 1, 3–9. doi:10.1016/j.livres.2017.05.001
- Chipot, C., and Pohorille, A. (2007). *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Springer Science & Business Media. Available at: https://books.google.com/books/about/Free_Energy_Calculations.html?hl=&id=EbOSq7IRPuYC.
- Connolly, J. J., Ooka, K., and Lim, J. K. (2018). Future Pharmacotherapy for Non-alcoholic Steatohepatitis (NASH): Review of Phase 2 and 3 Trials. *J. Clin. Transl Hepatol.* 6, 1–12. doi:10.14218/JCTH.2017.00056
- Costantino, G., Entrena-Guadix, A., Macchiarulo, A., Gioiello, A., and Pellicciari, R. (2005). Molecular Dynamics Simulation of the Ligand Binding Domain of Farnesoid X Receptor. Insights into helix-12 Stability and Coactivator Peptide Stabilization in Response to Agonist Binding. *J. Med. Chem.* 48, 3251–3259. doi:10.1021/jm049182o
- 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 Leva, F. S., Festa, C., D'Amore, C., De Marino, S., Renga, B., D'Auria, M. V., et al. (2013). Binding Mechanism of the Farnesoid X Receptor marine Antagonist Suvanine Reveals a Strategy to Forestall Drug Modulation on Nuclear Receptors. Design, Synthesis, and Biological Evaluation of Novel Ligands. *J. Med. Chem.* 56, 4701–4717. doi:10.1021/jm400419e
- Downes, M., Verdecia, M. A., Roecker, A. J., Hughes, R., Hogenesch, J. B., Kast-Woelbern, H. R., et al. (2003). A Chemical, Genetic, and Structural Analysis of the Nuclear Bile Acid Receptor FXR. *Mol. Cell* 11, 1079–1092. doi:10.1016/s1097-2765(03)00104-7
- Durrant, J. D., Votapka, L., Sørensen, J., and Amaro, R. E. (2014). POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theor. Comput.* 10, 5047–5056. doi:10.1021/ct500381c
- Festa, C., De Marino, S., Carino, A., Sepe, V., Marchianò, S., Cipriani, S., et al. (2017). Targeting Bile Acid Receptors: Discovery of a Potent and Selective Farnesoid X Receptor Agonist as a New Lead in the Pharmacological Approach to Liver Diseases. *Front. Pharmacol.* 8, 162. doi:10.3389/fphar.2017.00162
- Genheden, S., and Ryde, U. (2015). The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* 10, 449–461. doi:10.1517/17460441.2015.1032936
- Gioiello, A., Cerra, B., Mostarda, S., Guercini, C., Pellicciari, R., and Macchiarulo, A. (2014). Bile Acid Derivatives as Ligands of the Farnesoid X Receptor: Molecular Determinants for Bile Acid Binding and Receptor Modulation. *Cmc* 14, 2159–2174. doi:10.2174/156802661466614112100208
- Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., et al. (2016). OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theor. Comput.* 12, 281–296. doi:10.1021/acs.jctc.5b00864
- Hirschfield, G. M., Mason, A., Luketic, V., Lindor, K., Gordon, S. C., Mayo, M., et al. (2015). Efficacy of Obeticholic Acid in Patients with Primary Biliary Cirrhosis and Inadequate Response to Ursodeoxycholic Acid. *Gastroenterology* 148, 751–761. e8. doi:10.1053/j.gastro.2014.12.005
- Hollman, D. A. A., Milona, A., van Erpecum, K. J., and van Mil, S. W. C. (2012). Anti-inflammatory and Metabolic Actions of FXR: Insights into Molecular Mechanisms. *Biochim. Biophys. Acta (Bba) - Mol. Cell Biol. Lipids* 1821, 1443–1452. doi:10.1016/j.bbalip.2012.07.004
- Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14, 33–38. doi:10.1016/0263-7855(96)00018-5
- Ivetac, A., and McCammon, J. A. (2009). Elucidating the Inhibition Mechanism of HIV-1 Non-nucleoside Reverse Transcriptase Inhibitors through Multicopy Molecular Dynamics Simulations. *J. Mol. Biol.* 388, 644–658. doi:10.1016/j.jmb.2009.03.037
- Kemper, J. K. (2011). Regulation of FXR Transcriptional Activity in Health and Disease: Emerging Roles of FXR Cofactors and post-translational Modifications. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1812, 842–850. doi:10.1016/j.bbdis.2010.11.011
- Khandel, H., and Kaznessis, Y. N. (2007). Cation- π Interactions Stabilize the Structure of the Antimicrobial Peptide Indolicidin Near Membranes: Molecular Dynamics Simulations. *J. Phys. Chem. B* 111, 242–250. doi:10.1021/jp064776j
- Kudlinzki, D., Merk, D., Linhard, V. L., Saxena, K., Schubert-Zsilavecz, M., and Schwalbe, H. (2019). Crystal Structure of Farnesoid X Receptor (FXR) with Bound NCoA-2 Peptide and CDCA. doi:10.2210/pdb6h11/pdb
- Kumari, A., Mittal, L., Srivastava, M., and Asthana, S. (2021). Binding Mode Characterization of 13b in the Monomeric and Dimeric States of SARS-CoV-2

- Main Protease Using Molecular Dynamics Simulations. *J. Biomol. Struct. Dyn.* 1, 1–19. doi:10.1080/07391102.2021.1927844
- Kumari, A., Pal Pathak, D., and Asthana, S. (2020). Bile Acids Mediated Potential Functional Interaction between FXR and FATP5 in the Regulation of Lipid Metabolism. *Int. J. Biol. Sci.* 16, 2308–2322. doi:10.7150/ijbs.44774
- Li, T., and Chiang, J. Y. L. (2014). Bile Acid Signaling in Metabolic Disease and Drug Therapy. *Pharmacol. Rev.* 66, 948–983. doi:10.1124/pr.113.008201
- 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
- Maisuradze, G. G., Liwo, A., and Scheraga, H. A. (2009). Principal Component Analysis for Protein Folding Dynamics. *J. Mol. Biol.* 385, 312–329. doi:10.1016/j.jmb.2008.10.018
- Makishima, M., Okamoto, A. Y., Repa, J. J., Tu, H., Learned, R. M., Luk, A., et al. (1999). Identification of a Nuclear Receptor for Bile Acids. *Science* 284, 1362–1365. doi:10.1126/science.284.5418.1362
- Manjula, S., Sivanandam, M., and Kumaradhas, P. (2019). Probing the “Fingers” Domain Binding Pocket of Hepatitis C Virus NS5B RdRp and D559G Resistance Mutation via Molecular Docking, Molecular Dynamics Simulation and Binding Free Energy Calculations. *J. Biomol. Struct. Dyn.* 37, 2440–2456. doi:10.1080/07391102.2018.1491419
- Massafa, V., Pellicciari, R., Gioiello, A., and van Mil, S. W. C. (2018). Progress and Challenges of Selective Farnesoid X Receptor Modulation. *Pharmacol. Ther.* 191, 162–177. doi:10.1016/j.pharmthera.2018.06.009
- Merk, D., Sreeramulu, S., Kudlinzki, D., Saxena, K., Linhard, V., Gande, S. L., et al. (2019). Molecular Tuning of Farnesoid X Receptor Partial Agonism. *Nat. Commun.* 10, 2915. doi:10.1038/s41467-019-10853-2
- Meyer, U., Costantino, G., Macchiarulo, A., and Pellicciari, R. (2005). Is Antagonism of α -Z-Guggulsterone at the Farnesoid X Receptor Mediated by a Noncanonical Binding Site? A Molecular Modeling Study. *J. Med. Chem.* 48, 6948–6955. doi:10.1021/jm0505056
- Mi, L.-Z., Devarakonda, S., Harp, J. M., Han, Q., Pellicciari, R., Willson, T. M., et al. (2003). Structural Basis for Bile Acid Binding and Activation of the Nuclear Receptor FXR. *Mol. Cell* 11, 1093–1100. doi:10.1016/s1097-2765(03)00112-6
- Mittal, L., Kumari, A., Suri, C., Bhattacharya, S., and Asthana, S. (2020). Insights into Structural Dynamics of Allosteric Binding Sites in HCV RNA-dependent RNA Polymerase. *J. Biomol. Struct. Dyn.* 38, 1–14. doi:10.1080/07391102.2019.1614480
- Mittal, L., Srivastava, M., and Asthana, S. (2019). Conformational Characterization of Linker Revealed the Mechanism of Cavity Formation by 227G in BVDV RDRP. *J. Phys. Chem. B* 123, 6150–6160. doi:10.1021/acs.jpcc.9b01859
- Mittal, L., Srivastava, M., Kumari, A., Tonk, R. K., Awasthi, A., and Asthana, S. (2021). Interplay Among Structural Stability, Plasticity, and Energetics Determined by Conformational Attuning of Flexible Loops in PD-1. *J. Chem. Inf. Model.* 61, 358–384. doi:10.1021/acs.jcim.0c01080
- Mudaliar, S., Henry, R. R., Sanyal, A. J., Morrow, L., Marschall, H. U., Kipnes, M., et al. (2013). Efficacy and Safety of the Farnesoid X Receptor Agonist Obeticholic Acid in Patients with Type 2 Diabetes and Nonalcoholic Fatty Liver Disease. *Gastroenterology* 145, 574–582. e1. doi:10.1053/j.gastro.2013.05.042
- Neuschwander-Tetri, B. A., Loomba, R., Sanyal, A. J., Lavine, J. E., Van Natta, M. L., Abdelmalek, M. F., et al. (2015). Farnesoid X Nuclear Receptor Ligand Obeticholic Acid for Non-cirrhotic, Non-alcoholic Steatohepatitis (FLINT): a Multicentre, Randomised, Placebo-Controlled Trial. *The Lancet* 385, 956–965. doi:10.1016/S0140-6736(14)61933-4
- Nolte, R. T., Wisely, G. B., Westin, S., Cobb, J. E., Lambert, M. H., Kurokawa, R., et al. (1998). Ligand Binding and Co-activator Assembly of the Peroxisome Proliferator-Activated Receptor- γ . *Nature* 395, 137–143. doi:10.1038/25931
- Parks, D. J., Blanchard, S. G., Bledsoe, R. K., Chandra, G., Consler, T. G., Kliewer, S. A., et al. (1999). Bile Acids: Natural Ligands for an Orphan Nuclear Receptor. *Science* 284, 1365–1368. doi:10.1126/science.284.5418.1365
- Pasi, M., Tiberti, M., Arrigoni, A., and Papaleo, E. (2012). xPyder: a PyMOL Plugin to Analyze Coupled Residues and Their Networks in Protein Structures. *J. Chem. Inf. Model.* 52, 1865–1874. doi:10.1021/ci300213c
- Pellicciari, R., Fiorucci, S., Camaioni, E., Clerici, C., Costantino, G., Maloney, P. R., et al. (2002). 6 α -Ethyl-Chenodeoxycholic Acid (6-ECDCA), a Potent and Selective FXR Agonist Endowed with Anticholestatic Activity. *J. Med. Chem.* 45, 3569–3572. doi:10.1021/jm025529g
- Pencek, R., Marmon, T., Roth, J. D., Liberman, A., Hooshmand-Rad, R., and Young, M. A. (2016). Effects of Obeticholic Acid on Lipoprotein Metabolism in Healthy Volunteers. *Diabetes Obes. Metab.* 18, 936–940. doi:10.1111/dom.12681
- Pradhan, M., Suri, C., Choudhary, S., Naik, P. K., and Lopus, M. (2018). Elucidation of the Anticancer Potential and Tubulin Isotype-specific Interactions of β -sitosterol. *J. Biomol. Struct. Dyn.* 36, 195–208. doi:10.1080/07391102.2016.1271749
- Proteins (2004). *Proteins: Structure, Function, and Bioinformatics*, 55. doi:10.1002/prot.v55:2
- Roe, D. R., and Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theor. Comput.* 9, 3084–3095. doi:10.1021/ct400341p
- Rostkowski, M., Olsson, M. H., Søndergaard, C. R., and Jensen, J. H. (2011). Graphical Analysis of pH-dependent Properties of Proteins Predicted Using PROPKA. *BMC Struct. Biol.* 11, 6. doi:10.1186/1472-6807-11-6
- Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 23, 327–341. doi:10.1016/0021-9991(77)90098-5
- Sack, J. S., Kish, K. F., Wang, C., Attar, R. M., Kiefer, S. E., An, Y., et al. (2001). Crystallographic Structures of the Ligand-Binding Domains of the Androgen Receptor and its T877A Mutant Complexed with the Natural Agonist Dihydrotestosterone. *Proc. Natl. Acad. Sci.* 98, 4904–4909. doi:10.1073/pnas.081565498
- Sarkar, R., Sharma, K. B., Kumari, A., Asthana, S., and Kalia, M. (2021). Japanese Encephalitis Virus Capsid Protein Interacts with Non-lipidated MAP1LC3 on Replication Membranes and Lipid Droplets. *J. Gen. Virol.* 102. doi:10.1099/jgv.0.001508
- Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., Sherman, W., and Sherman, W. (2013). Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided Mol. Des.* 27, 221–234. doi:10.1007/s10822-013-9644-8
- Sepe, V., Distrutti, E., Fiorucci, S., and Zampella, A. (2015). Farnesoid X Receptor Modulators (2011–2014): a Patent Review. *Expert Opin. Ther. Patents* 25, 885–896. doi:10.1517/13543776.2015.1045413
- Shiau, A. K., Barstad, D., Loria, P. M., Cheng, L., Kushner, P. J., Agard, D. A., et al. (1998). The Structural Basis of Estrogen Receptor/coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell* 95, 927–937. doi:10.1016/s0092-8674(00)81717-1
- Sinal, C. J., Tohkin, M., Miyata, M., Ward, J. M., Lambert, G., and Gonzalez, F. J. (2000). Targeted Disruption of the Nuclear Receptor FXR/BAR Impairs Bile Acid and Lipid Homeostasis. *Cell* 102, 731–744. doi:10.1016/s0092-8674(00)00062-3
- Singh, M., Srivastava, M., Wakode, S. R., and Asthana, S. (2021). Elucidation of Structural Determinants Delineates the Residues Playing Key Roles in Differential Dynamics and Selective Inhibition of Sirt1-3. *J. Chem. Inf. Model.* 61, 1105–1124. doi:10.1021/acs.jcim.0c01193
- Sitkoff, D., Lockhart, D. J., Sharp, K. A., and Honig, B. (1994). Calculation of Electrostatic Effects at the Amino Terminus of an Alpha helix. *Biophysical J.* 67, 2251–2260. doi:10.1016/S0006-3495(94)80709-X
- Soisson, S. M., Parthasarathy, G., Adams, A. D., Sahoo, S., Sitlani, A., Sparrow, C., et al. (2008). Identification of a Potent Synthetic FXR Agonist with an Unexpected Mode of Binding and Activation. *Proc. Natl. Acad. Sci.* 105, 5337–5342. doi:10.1073/pnas.0710981105
- Srivastava, M., Suri, C., Singh, M., Mathur, R., and Asthana, S. (2018). Molecular Dynamics Simulation Reveals the Possible Druggable Hot-Spots of USP7. *Oncotarget* 9, 34289–34305. doi:10.18632/oncotarget.26136
- Suri, C., Hendrickson, T. W., Joshi, H. C., and Naik, P. K. (2014). Molecular Insight into γ - γ Tubulin Lateral Interactions within the γ -tubulin Ring Complex (γ -TuRC). *J. Comput. Aided Mol. Des.* 28, 961–972. doi:10.1007/s10822-014-9779-2
- Turner, P. (2005). *XMGRACE, Version 5.1.19*. Center for Coastal and Land-Margin Research. Beaverton, OR: Oregon Graduate Institute of Science and Technology.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* 26, 1701–1718. doi:10.1002/jcc.20291
- Wang, H., Chen, J., Hollister, K., Sowers, L. C., and Forman, B. M. (1999). Endogenous Bile Acids Are Ligands for the Nuclear Receptor FXR/BAR. *Mol. Cell* 3, 543–553. doi:10.1016/s1097-2765(00)80348-2
- Wang, J.-C., Stafford, J. M., and Granner, D. K. (1998). SRC-1 and GRIP1 Coactivate Transcription with Hepatocyte Nuclear Factor 4. *J. Biol. Chem.* 273, 30847–30850. doi:10.1074/jbc.273.47.30847

- 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
- Wang, S.-R., Xu, T., Deng, K., Wong, C.-W., Liu, J., and Fang, W.-S. (2017). Discovery of Farnesoid X Receptor Antagonists Based on a Library of Oleanolic Acid 3-O-Esters through Diverse Substituent Design and Molecular Docking Methods. *Molecules* 22, 690. doi:10.3390/molecules22050690
- Weikum, E. R., Liu, X., and Ortlund, E. A. (2018). The Nuclear Receptor Superfamily: A Structural Perspective. *Protein Sci.* 27, 1876–1892. doi:10.1002/pro.3496
- Williams, S. P., and Sigler, P. B. (1998). Atomic Structure of Progesterone Complexed with its Receptor. *Nature* 393, 392–396. doi:10.1038/30775
- Zhang, T., Dong, X.-C., and Chen, M.-B. (2006). Recognition of LXXLL by Ligand Binding Domain of the Farnesoid X Receptor in Molecular Dynamics Simulation. *J. Chem. Inf. Model.* 46, 2623–2630. doi:10.1021/ci060112v
- Zhang, Y., and Edwards, P. A. (2008). FXR Signaling in Metabolic Disease. *FEBS Lett.* 582, 10–18. doi:10.1016/j.febslet.2007.11.015

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

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

Copyright © 2021 Kumari, Mittal, Srivastava, Pathak and Asthana. 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.



RIGI, TLR7, and TLR3 Genes Were Predicted to Have Immune Response Against Avian Influenza in Indigenous Ducks

Aruna Pal^{1*}, Abantika Pal² and Pradyumna Baviskar³

¹West Bengal University of Animal and Fishery Sciences, Kolkata, India, ²Indian Institute of Technology Kharagpur, Kharagpur, India, ³St. Jude Children's Research Hospital, Memphis, TN, United States

OPEN ACCESS

Edited by:

Agnel Praveen Joseph,
Science and Technology Facilities
Council, United Kingdom

Reviewed by:

Tarun Bhattacharya,
Directorate of Poultry Research (DPR),
ICAR, India
Vikas Vohra,
National Dairy Research Institute
(ICAR), India

*Correspondence:

Aruna Pal
arunachatterjee@gmail.com

Specialty section:

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

Received: 25 November 2020

Accepted: 29 September 2021

Published: 14 December 2021

Citation:

Pal A, Pal A and Baviskar P (2021) RIGI,
TLR7, and TLR3 Genes Were
Predicted to Have Immune Response
Against Avian Influenza in
Indigenous Ducks.
Front. Mol. Biosci. 8:633283.
doi: 10.3389/fmolb.2021.633283

Avian influenza is a disease with every possibility to evolve as a human-to-human pandemic arising out of frequent mutations and genetic reassortment or recombination of avian influenza (AI) virus. The greatest concern is that till date, no satisfactory medicine or vaccines are available, leading to massive culling of poultry birds, causing huge economic loss and ban on export of chicken products, which emphasizes the need to develop an alternative strategy for control of AI. In the current study, we attempt to explore the molecular mechanism of innate immune potential of ducks against avian influenza. In the present study, we have characterized immune response molecules such as duck TLR3, TLR7, and RIGI that are predicted to have potent antiviral activities against the identified strain of avian influenza through *in silico* studies (molecular docking) followed by experimental validation with differential mRNA expression analysis. Future exploitation may include immunomodulation with the recombinant protein, and transgenic or gene-edited chicken resistant to bird flu.

Keywords: *Anas platyrhynchos*, avian influenza, RIGI, TLR3, TLR7

INTRODUCTION

Ducks are reported to be relatively resistant to common poultry diseases, including viral disease, compared to chicken (Pal et al., 2017), and are commonly asymptomatic to avian influenza virus infection (Kim et al., 2009; Fleming-Canepa et al., 2019; CDC, Centre for Disease control and Prevention, 2021). There is clear lack of further systematic characterization of the indigenous ducks at the molecular level. In an effort to understand, we have studied this as a first step. Hence, there is an urgent need to explore the innate immune response genes, particularly against viral infection.

Avian influenza is caused by single-stranded RNA virus, which is negatively stranded, and belongs to Orthomyxoviridae family (WHO, 2019). It is commonly known as bird flu since birds are the main host. Based on the antigenic differences, two surface proteins, namely, hemagglutinin and neuraminidase of avian influenza virus have been mostly subtyped and the nomenclature is provided accordingly. Till date, 18 subtypes of HA (H1–H18) and 11NA (N1–N11) have been detected (Tong et al., 2013; Ying et al., 2014). H5, H7, and H9 were observed to be the most pathogenic subtypes of bird. Most of the H5 and H7 subtypes were regarded as highly pathogenic avian influenza (HPAI) virus, owing to the higher incidence and mortality of birds. The greatest concern is the lack of definite treatment or vaccination due to frequent mutation and reassortment of viral strain, regarded as antigenic shift and antigenic drift (Webster and Govorkova, 2014; WHO,

2020). Due to massive culling of birds in the affected area and the ban on the export of poultry products, the WHO has regarded avian influenza as one of the most economically effected zoonotic diseases (WHO, 2018b). It has been reported that in West Bengal during the 2008 outbreak, there was loss of 500.42 crores, 6 percent of the poultry population was culled, and five lakh families were affected. In Manipur at the 2007 outbreak, the loss due to the disease has been found to be 14 per cent of the total value of livestock outputs in the entire state. More than three lakh birds were culled, and 24 tonnes of poultry feed was destroyed post-flu (Otte et al., 2008a). Vietnam estimated the direct losses of 109 village and backyard producers with flock sizes smaller than 50 birds at US\$69 (VND1084000) per farm (Otte et al., 2008a). An average loss of US\$22 was estimated per household from the loss of birds in Egypt (Otte et al., 2008a).

The basic mechanism of host immunity against viral infection is generally different from that of other infectious agents such as bacteria and protozoa. Viruses utilize the host immune mechanism for their infection and further survival, thus allowing it to act as hijackers. Accordingly, viruses employ the host cellular machinery for living normal cells through the process of invasion; multiplication within the host might in turn kill, damage, or change the cells, and make the individual sick (Maarouf et al., 2018). The molecular mechanism of replication of avian influenza virus involves certain proteins. The HA protein present from the surface of the AI virus aids in recognition and binding to sialic acid on the surface of host cells, thereby aiding in the entry of the virus in the host cell (Matrosovich et al., 2009a). Following the binding, virus particles are endocytosed, leading to endosome maturation, and pH is lowered, resulting in the conformational change in HA, thereby causing fusion of the endosome and virion membranes. The viral M2 protein (matrix 2) acts as an ion channel for further lowering of the pH of the viral particle. This leads to the dissociation of the M1 protein (matrix 1) virion “shell” in such a way that the eight vRNPs [NP (nucleoprotein)-coated and polymerase complex (PB1, PA, and PB2)-bound viral RNAs] are released into the cytosol (Basler and Aguilar, 2008a). In the next step, the viral RNPs transport into the cell nucleus wherein accessory cellular components necessary for influenza viral replication and transcription are present. Following the process of genome replication, transcription, and protein synthesis, NEP (nuclear export protein) and M1 act to traffic newly synthesized vRNPs out of the nucleus, into the cytoplasm, and to the plasma membrane, leading to the assembly of progeny virions. At this stage, several viral proteins contribute to budding, including M1 and M2. In the next step, NA aids in the removal of sialic acid from glycoproteins in both the viral and cell membranes, resulting in the prevention of the interaction between HA and host cell receptors, and release of new infectious virus particles (Medina and García-Sastre, 2011). NS1 (nonstructural protein 1) plays a role within the infected cell in order to counteract innate host-cell defense systems, for example, interferon (IFN), which may otherwise limit efficient virus replication (Hale et al., 2008).

As the virus gets an entry in the body, an immune response is triggered, followed by local inflammatory signaling. Innate immune reaction is initially activated by conserved pathogen-

associated molecular pattern (PAMP), pattern recognition receptors (PRRs), retinoic acid-inducible gene (RIG)-I like receptors, MDA5, LGP 2, and toll-like receptor (TLRs) such as TLR3 and TLR7 (Kannaki et al., 2010). Viral nucleic acid binds to these receptors expressed on macrophages, microglia, dendritic cells, and astrocytes; releases type-I interferon (IFN-I); and helps in the production of interferon-stimulated genes (ISGs) (Ali et al., 2019). Interferon-I upregulates antiviral proteins, and accordingly, peripheral immune cells are stimulated and alter endothelial tight junction (Otte et al., 2008b). It has been observed that the absence of IFN-I signaling leads to the prevention of microglial differentiation and decrease of peripheral myeloid cell patrolling (Otte et al., 2008b).

TLR7 is a member of the Toll-like receptor family, which recognizes single-stranded RNA in endosomes, which is a common feature of viral genomes (Mangani and McGavern, 2018). TLR7 can recognize GU-rich single-stranded RNA. TLR7 was reported to have influences on viral infection in poultry and has been regarded as a vital component of antiviral immunity, particularly in ducks (Mangani and McGavern, 2018). RIG-I (retinoic acid-inducible gene I) or RIG-I-like receptor dsRNA helicase enzyme is part of the RIG-I-like receptor family, which also includes MDA5 and LGP2. These have been reported to function as a pattern recognition receptor that is a sensor for viruses such as influenza A, others such as Sendai virus, and Flavivirus (Matrosovich et al., 2009b). RIG-I typically recognizes short 5' triphosphate uncapped double-stranded or single-stranded RNA (Basler and Aguilar, 2008b). RIG-I and MDA5 are the viral receptors, acting through a common adapter MAVS and trigger an antiviral response through type-I interferon response. RIG1 is an important gene conferring antiviral immunity for ducks, particularly avian influenza (Matrosovich et al., 2009b).

TLR3 is another member of the toll-like receptor (TLR) family. Infectious agents express PAMP (pathogen-associated molecular patterns), which is readily recognized by TLR3, which in turn secretes cytokines responsible for effective immunity. It recognizes dsRNA associated with a viral infection, and induces the activation of IRF3, unlike all other toll-like receptors which activate NF- κ B (Kannaki et al., 2010). IRF3 ultimately induces the production of type I interferons, which is ultimately responsible for host defense against viruses (Kell and Gale, 2015). In our lab, earlier we had studied immunogenetics against bacterial disease with identified immune response molecule such as CD14 gene in goat (Vercammen et al., 2008a; Schlee et al., 2009), cattle (Stetson and Medzhitov, 2006), and buffalo (Pal and Chatterjee, 2009; Pal et al., 2013). We reported for the first time the role of mitochondrial cytochrome B gene for immunity in sheep (Pal et al., 2011) and immune-response genes in Haringhata Black chicken (Pal et al., 2014).

Indigenous duck population in the Indian subcontinent was observed to have better immunity against viral infections, and so far, no systematic studies were undertaken. Certain reports revealed that ducks were mostly asymptomatic and were better resistant to avian influenza infection. Thus, the present study was conducted with the aim of molecular characterization of immune

response genes (TLR3, TLR7, and RIGI) of duck, providing the initial proteomics study and prediction of the binding site with multiple strains of avian influenza virus through *in silico* studies (molecular docking), establishment of disease-resistant genes of ducks through quantitative PCR, and experimental validation of the identified genes through differential mRNA expression profiling of the identified gene with respect to healthy and challenged embryonated eggs as *in vitro* studies.

MATERIALS AND METHODS

Animals, Sample Collection, and RNA Isolation

Duck samples were collected from different agro-climatic regions of West Bengal, India, from farmer's herd. The chicken breeds such as Haringhata Black and Aseel were maintained in the university farm (West Bengal University of Animal and Fishery Sciences). Samples from other poultry species such as guineafowl and goose were also collected from the university farm. Samples from turkey and quail were collected from State Poultry farm, Animal Resource Development Dept, Tollygunge, Govt. of West Bengal, India. The birds were vaccinated against routine diseases such as Ranikhet disease and fowl pox. Six male birds (aged 4–5 months) were considered under each group for this study and are maintained under uniform managerial conditions.

All experiments were conducted in accordance with relevant guidelines and regulations of the Institutional Animal Ethics Committee, and all experimental protocols were approved by the Institutional Biosafety Committee, West Bengal University of Animal and Fishery Sciences, Kolkata.

The total RNA was isolated from the ileocecal junction of duck, Haringhata Black chicken, Aseel, and other poultry species such as guineafowl and goose, using RiboPure Kit (Invitrogen), following the manufacturer's instructions and was further used for cDNA synthesis (Schlee et al., 2009; Pal et al., 2011).

Materials

Taq DNA polymerase, 10X buffer, and dNTP were purchased from Invitrogen, and SYBR Green qPCR Master Mix (2X) was obtained from Thermo Fisher Scientific Inc. (PA, United States). L-Glutamine (Glutamax 100x) was purchased from Invitrogen corp., (Carlsbad, CA, United States). Penicillin-G and streptomycin were obtained from Amresco (Solon, OH, United States). Filters (Millex GV. 0.22 µm) were purchased from Millipore Pvt. Ltd., (Billerica, MA, United States). All other reagents were of analytical grade.

Synthesis, Confirmation of cDNA, and PCR Amplification of TLR3, RIGI, and TLR7 Genes

The 20 µl reaction mixture contained 5 µg of total RNA, 0.5 µg of oligo dT primer (16–18 mer), 40 U of ribonuclease inhibitor, 10 M of dNTP mix, 10 mM of DTT, and 5 U of MuMLV reverse transcriptase in the reverse transcriptase buffer. The reaction

mixture was gently mixed and incubated at 37°C for 1 h. The reaction was stopped by heating the mixture at 70°C for 10 min and chilled on ice. The integrity of the cDNA was checked by PCR. To amplify the full-length open reading frame (ORF) of the gene sequence, a specific primer pair was designed based on the mRNA sequences of *Gallus gallus* by DNASTAR software. The primers have been listed in **Table 1**. 25 µl of the reaction mixture contained 80–100 ng cDNA, 3.0 µl 10X PCR assay buffer, 0.5 µl of 10 mM dNTP, 1 U Taq DNA polymerase, 60 ng of each primer, and 2 mM MgCl₂. PCRs were carried out in a thermocycler (PTC-200, MJ Research, United States) with the following cycling conditions: initial denaturation at 94°C for 3 min, denaturation at 94°C for 30 sec, and varying annealing temperature (as mentioned in **Table 1**) for 35 sec, and extension at 72°C for 3 min was carried out for 35 cycles followed by final extension at 72°C for 10 min.

cDNA Cloning and Sequencing

PCR amplicons verified by 1% agarose gel electrophoresis were purified from gel using Gel Extraction Kit (Qiagen GmbH, Hilden, Germany) and ligated into a pGEM-T easy cloning vector (Promega, Madison, WI, United States) following the manufacturer's instructions. The 10 µl of the ligated product was directly added to 200 µl competent cells, heat shock was given at 42°C for 45 s in a water bath, and cells were then immediately transferred on chilled ice for 5 min, and SOC was added. The bacterial culture was pelleted and plated on the LB agar plate containing ampicillin (100 mg/ml) added to the agar plate @ 1:1000, IPTG (200 mg/ml) and X-Gal (20 mg/ml) for blue-white screening. Plasmid isolation from overnight-grown culture was done by the small-scale alkaline lysis method. Recombinant plasmids were characterized by PCR using gene-specific primers and restriction enzyme digestion based on the reported nucleotide sequence for cattle. The enzyme EcoRI (MBI Fermentas, United States) is used for fragment release. Gene fragment insert in the recombinant plasmid was sequenced by an automated sequencer (ABI prism) using the dideoxy chain termination method with T7 and SP6 primers (Chromous Biotech, Bangalore).

Sequence Analysis

The nucleotide sequence so obtained was analyzed for protein translation, sequence alignments, and contig comparisons by DNASTAR version 4.0, Inc., United States. The novel sequence was submitted to the NCBI GenBank, and the accession number was obtained, which is available in a public domain now.

Study of Predicted TLR3, TLR7, and RIG1 Peptides Using Bioinformatic Tools

The predicted peptide sequence of TLR3, TLR7, and RIG1 of indigenous duck was derived by Edit sequence (Lasergene Software, DNASTAR) and then aligned with the peptide of other chicken breed and avian species using Megalign sequence Programme of Lasergene Software (DNASTAR). Prediction of the signal peptide of the CD14 gene was conducted using the software (Signal P 3.0 Server-prediction

TABLE 1 | List of primers used for amplification of TLR3, RIG1, and TLR7 genes in indigenous duck.

Gene	Primer	Product length	Annealing temp
Primers used for amplification of TLR3 for duck			
Duck TLR3.1	FP: TGGAAAACATGTCAAATCAG RP: TCACGGAGGTTCTTCAG	450	49.9
Duck TLR3.2	FP: TCCGTGAGCTTGTGTTGT RP: AGATGTTTGAGCCTGGAC	460	50.2
Duck TLR3.3	FP: GATAAATTCGCTCACTGG RP: TCTAAGGCTTGAACGA	436	48.8
Duck TLR3.4	FP: TCAGCAATAACAACATAGCAAACA RP: GGGTCGCATTAAGCCAACT	456	51.8
Duck TLR3.5	FP: ATATACCTGGATTGCAGTCTCAGT RP: CTGGGCTGGCCACTTCAAG	650	52.7
Primers used for amplification of RIG1 for duck			
Duck RIG1.1	FP: CTGCAGTGCTACCGCCGCTACATC RP: TATCCGACCGACAGAGACATTCAA	460	59.6
Duck RIG1.2	FP: AAAGATGTTGACAGTGAAATG RP: TCCTTGAACAGAGTATCCTT	402	50.8
Duck RIG1.3	FP: CAGGACGAAAGGCGAAAGTT RP: TGTATGTCAAGGTAGGAGCAGAGA	448	53.8
Duck RIG1.4	FP: ATCCCTTTGCAGCCATTATCC RP: CGCGCCCCATCAAACAC	585	55.2
Duck RIG1.5	FP: TAACTACATAAAGCCAGGTG RP: TACTTTAGGTTTTATTTCTTTC	448	50.4
Duck RIG1.6	FP: CCAGAAGGAAAGAAATAAAACC RP: TGGTGGGTACAAGTTGGACAT	416	52.3
Primers used for amplification of TLR7 of duck			
Duck TLR7.1	FP: TCAAGCATATTCATGAAGACTTT RP: TGGGCCCCAACCTGACAG	513	58.4
Duck TLR7.2	FP: TTGAGAATGGCAGTTTTG RP: AGCCTTTGAATGTATCTTA	500	48.8
Duck TLR7.3	FP: ACATTCAAAGGCTTTTTATTCCT RP: TATTGCATTACCTGACAAAGTTGAG	754	52.4
Duck TLR7.4	FP: GATGCCTCAACTTGTGAGTAATG RP: TTTTCGGGGAAGCTAGATTCTT	751	53.5
Duck TLR7.5	FP: CTAGCTTCCCGAAATGTCAT RP: TTCTGCACAGCCTTTTCCTCAG	736	54.8
Duck TLR7.6	FP: AGCGCCTTCTAGATGAAAA RP: TTTTAGTTTATGAGATTTTATTAT	400	48.8
List of primers used for QPCR study			
β -Actin	FP: 5'-GAGAAATTGTGCGTGACATCA-3' RP: 5'-CCTGAACCTCTCATTGCCA-3'	152	60
TLR2	FP: 5'-CATTACCATGAGGCAGGGATAG-3' RP: 5'-GGTGCAGATCAAGGACACTAGGA-3'	157	60
TLR4	FP: 5'-TTCAGAACGGACTCTTGAGTGG-3' RP: 5'-CAACCGAATAGTGGTGACGTTG-3'	131	60
TLR7	5'-TTGCTGCTGTTGTCTTGAGTGAG-3' 5'-AACACAGTGCATTTGACGTCCT-3'	182	60
Bu-1	5'-GGCTGTTGTGCTCCTCACTCATCT-3' 5'-CACCACCGACATTGTTATTCCAT-3'	106	60

*The final and complete sequence is obtained by joining the fragments of the amplified products of the gene consecutively.

1TLR2, Toll-like receptor 2; TLR4, Toll-like receptor 4; TLR7, Toll-like receptor 7; Bu-1, chicken B-cell marker chB6.

results, Technical University of Denmark). Estimation of leucine percentage was conducted manually from the predicted peptide sequence. Di-sulfide bonds were predicted using suitable software (<http://bioinformatics.bc.edu/clotelab/DiANNA/>) and by homology search with other species.

The protein sequence-level analysis study was carried out with specific software (<http://www.expasy.org/tools/blast/>) for the determination of leucine-rich repeats (LRRs), leucine zipper, N-linked glycosylation sites, detection of leucine-rich nuclear export signals (NESs), and detection of the position of the GPI

anchor. The detection of leucine-rich nuclear export signals (NESs) was carried out with NetNES 1.1 Server, Technical University of Denmark. The analysis of O-linked glycosylation sites was carried out using NetOGlyc 3.1 server (<http://www.expasy.org/>), whereas the N-linked glycosylation site was detected by NetNGlyc 1.0 software (<http://www.expasy.org/>). The detection of leucine-zipper was conducted through Expasy software, Technical University of Denmark (Pal et al., 2020). Regions for alpha-helix and beta-sheet were predicted using NetSurfP-Protein Surface Accessibility and Secondary

Structure Predictions, Technical University of Denmark (Schlee et al., 2009; Pal et al., 2018a). Domain linker prediction was done according to the software developed (Pal et al., 2019a). The LPS-binding site (Glick, 1977) and LPS-signaling sites (Petersen et al., 2009) were predicted based on homology studies with other polypeptide species.

Three-Dimensional Structure Prediction and Model Quality Assessment

The templates which possessed the highest sequence identity with our target template were identified by using PSI-BLAST (<http://blast.ncbi.nlm.nih.gov/Blast>). The homology modeling was used to build a 3D structure based on homologous template structures using PHYRE2 server (Ebina et al., 2009). The 3D structures were visualized by PyMOL (<http://www.pymol.org/>), which is an open-source molecular visualization tool. Subsequently, the mutant model was generated using the PyMOL tool. Swiss PDB Viewer was employed for controlling energy minimization. The structural evaluation along with a stereochemical quality assessment of predicted model was carried out by using the SAVES (Structural Analysis and Verification Server), which is an integrated server (<http://nihserver.mbi.ucla.edu/SAVES/>). The ProSA (Protein Structure Analysis) webserver (<https://prosa.services.came.sbg.ac.at/prosa>) was used for refinement and validation of the protein structure (Cunningham et al., 2000). The ProSA was used for checking model structural quality with potential errors, and the program shows a plot of its residue energies and Z-scores which determine the overall quality of the model. The solvent accessibility surface area of the IR genes was generated by using NetSurfP server (<http://www.cbs.dtu.dk/services/NetSurfP/>) (Muroi et al., 2002). It calculates relative surface accessibility, Z-fit score, the probability for Alpha-Helix, probability for beta-strand and coil score, etc. TM-align software was used for the alignment of 3D structure of IR protein for different species and RMSD estimation to assess the structural differentiation (Kelley et al., 2015). The I-mutant analysis was conducted for mutations detected to assess the thermodynamic stability (Wiederstein and Sippl, 2007). PROVEAN analysis was conducted to assess the deleterious nature of the mutant amino acid (Pal et al., 2018b).

Molecular Docking

Molecular docking is a bioinformatic tool used for *in silico* analysis for the prediction of the binding mode of a ligand with a protein 3D structure. PatchDock is an algorithm for molecular docking based on the shape complementarity principle (Zhang and Skolnick, 2005). The PatchDock algorithm was used to predict ligand-protein docking for surface antigen for avian influenza (H antigen and NA antigen) with the molecules for innate immunity against viral infections such as TLR3, TLR7, and RIG1. FireDock was employed for further confirmation (Capriotti et al., 2005). The amino acid sequence for the surface antigen (hemagglutinin and neuraminidase) from different strains of avian influenza was retrieved from gene bank. Hemagglutinin segment 4 sequence was collected from the Indian subcontinent as H5N1 (Acc no.

KR021385, Protein id. AKD00332), H4N6 (Acc no. JX310059, Protein id. AF082958), H6N2 (Acc no. KU598235, Protein id. AMH93683), and H9N2 (Acc no. 218091, Protein id. AAG53040). Neuraminidase segment 6 was collected from the Indian subcontinent as H5N1 (Acc no. KT867346, Protein id. ALK80150), H4N6 (Acc no. JX310060, Protein id. AF082959), and H6N2 (Acc no. KU598237, Protein id. AMH93685) to be employed as the ligand. The receptor molecules employed were TLR3 (Gene bank accession number KX865107, NCBI, and derived protein as ASW23003), RIGI (Gene bank accession number KX865107, protein ASW23002 from NCBI), and TLR7 (Gene bank Accession no. MK986726, NCBI) for duck sequenced and characterized in our lab.

Assessment of Antigenic Variability Among Different Strains of Avian Influenza

MAFFT software (Choi and Chan, 2015) was employed for the detection of amino acid variability and construction of phylogenetic tree for different strains of avian influenza detected in duck in the Indian subcontinent.

The amino acid sequence for the surface antigen (hemagglutinin and neuraminidase) from different strains of avian influenza was retrieved from gene bank. Hemagglutinin segment 4 sequence was collected from the Indian subcontinent as H5N1 (Acc no. KR021385, Protein id. AKD00332), H4N6 (Acc no. JX310059, Protein id. AF082958), H6N2 (Acc no. KU598235, Protein id. AMH93683), and H9N2 (Acc no. 218091, Protein id. AAG53040).

Neuraminidase segment 6 was collected from the Indian subcontinent as H5N1 (Acc no. KT867346, Protein id. ALK80150), H4N6 (Acc no. JX310060, Protein id. AF082959), and H6N2 (Acc no. KU598237, Protein id. AMH93685.1).

Protein-Protein Interaction Network Depiction

In order to understand the network of TLR3, TLR7, and RIG1 peptides, we performed analysis submitting FASTA sequences to STRING 9.1 (Schneidman-Duhovny et al., 2005). Confidence scoring was used for functional analysis. Interactions with score <0.3 are considered as low confidence, scores ranging from 0.3 to 0.7 are classified as medium confidence, and scores >0.7 yield high confidence. The functional partners were depicted.

KEGG analysis also depicts the functional association of TLR3, TLR7, and RIG1 peptides with other related proteins (KEGG: Kyoto Encyclopedia of Genes and Genomes-GenomeNet, <https://www.genome.jp/kegg/>).

Real-Time PCR

An equal amount of RNA (quantified by Qubit fluorometer, Invitrogen), wherever applicable, were used for cDNA preparation (Superscript III cDNA synthesis kit; Invitrogen). All qRT-PCR reactions were conducted on the ABI 7500 fast system. Each reaction consisted of 2 µl cDNA template, 5 µl of 2X SYBR Green PCR Master Mix, 0.25 µl each of forward and reverse primers (10 pmol/µl), and nuclease-free water for a final volume

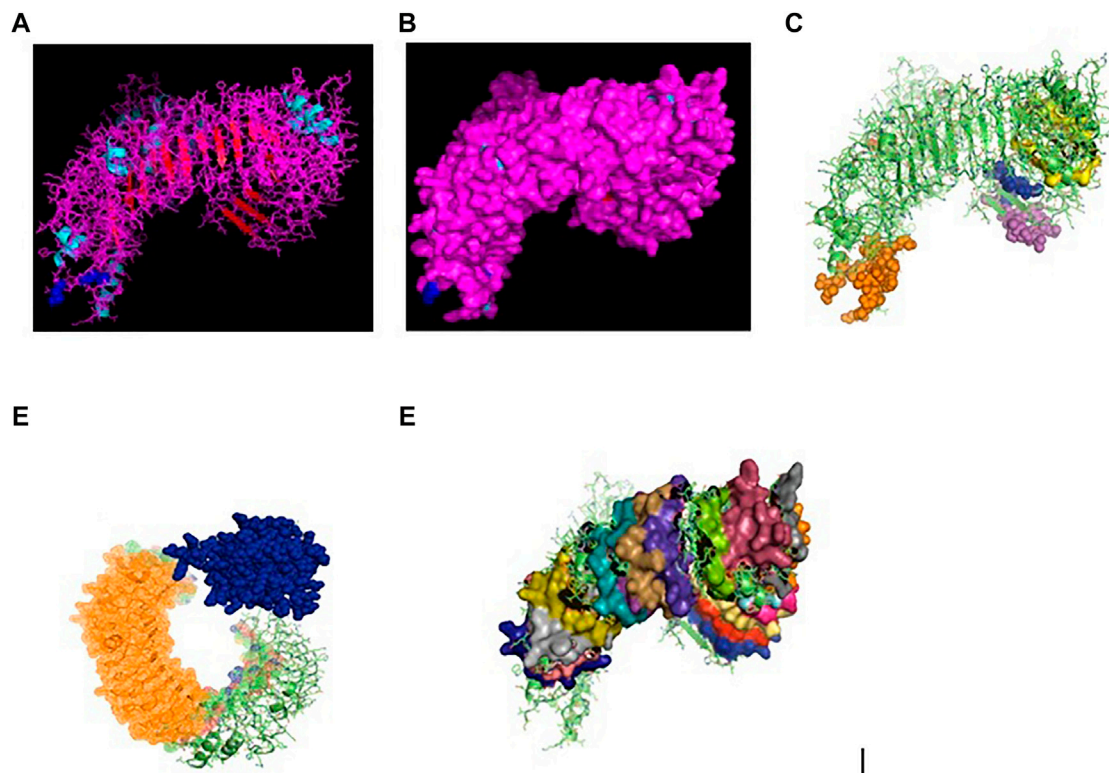


FIGURE 1 | (A) TLR3 molecule of duck (secondary structure with disulphide bond) as blue sphere. (B) TLR3 molecule of duck (secondary structure with disulphide bond) surface view. (C) 3D structure of TLR3 of duck: yellow surface: leucine zipper (151–172), red sphere: GPI anchor (879), blue sphere: leucine-rich nuclear export signal, magenta sphere: LRRNT (37–51), orange sphere: LRRCT (664–687). (D) 3D structure of TLR3 of indigenous duck, blue sphere: TIR (748–890), orange mesh: leucine-rich receptor-like protein kinase. (E) 3D structure of TLR3 of duck with leucine-rich repeat.

of 10 μ l. Each sample was run in duplicate. Analysis of real-time PCR (qRT-PCR) was performed by the delta-delta-Ct ($\Delta\Delta$ Ct) method. The primers used for QPCR analysis have been listed as per **Table 1**.

Studies were also conducted for differential mRNA expression profiling of TLR3, TLR7, and RIGI as an *in vitro* study in embryonic fibroblast cell of both chicken (Haringhata Black breed) and duck (Bengal duck) after the challenge study with the H5N1 strain of avian influenza virus in comparison to control in BSL3 lab. Samples were collected after 72 h of infection and subjected to RNA isolation, and the same steps were followed as explained earlier.

Comparison of TLR3, TLR7, and RIGI Structures of Indigenous Ducks With Respect to Chicken

Nucleotide variation for the proteins was detected from their nucleotide sequencing, and amino acid variations were estimated (DNASTAR). The 3D structure of the derived protein was estimated for both indigenous ducks and chicken by Pymol software. The PDB structure of the respective proteins was derived from *PHYRE* software (Mashiach et al., 2008). We also employed *Modeller* software for protein structural

modeling (Katoh and Standley, 2013) for better confirmation. Alignment of the structure of TLR3, TLR7, and RIGI duck with chicken was conducted by TM Align software (Szkłarczyk et al., 2015).

RESULTS

Molecular Characterization of TLR3 Gene

Toll-like receptors are a group of pattern recognition receptors effective against a wide range of pathogens. TLR3 gene of indigenous ducks has been characterized with 2688 bp nucleotide (Gene bank accession number KX865107, NCBI) and derived protein as ASW23003.1. The 3D protein structure (**Figure 1A**) with surface view (**Figure 1B**) has been depicted, with helix light blue, sheet red, loop pink, and blue spheres as disulfide bonds.

Posttranslational modification sites for TLR3 of duck have been depicted in **Figures 1C–E**. **Figure 1C** reveals the 3D structure of TLR3 of duck with the sites for leucine zipper (151–172 amino acid position, yellow surface), GPI anchor (aa position 879, red sphere), leucine-rich nuclear export signal (aa position 75–83, blue sphere), LRRNT (aa position 37–51, magenta sphere), and LRRCT (aa position 664–687, orange

TABLE 2 | Amino acid variations for TLR3 gene in duck with other poultry species.

SI no.	Position	Duck	Chicken	Turkey	Goose	Domain
1	42	K	E	K	K	LRRNT
2	61	H	L	H	H	LRR1
3	68	C	V	V	C	LRR1
4	69	H	P	P	P	LRR1
5	70	A	E	E	A	LRR1
6	74	R	Q	E	K	LRR1
7	77	K	N	N	K	LRR2
8	92	Q	K	Q	Q	LRR2
9	94	E	O	E	E	LRR2
10	106	V	K	K	V	LRR3
11	119	A	V	V	T	LRR3
12	137	D	E	E	D	LRR4
13	166	L	L	L	W	LRR5
14	179	C	Y	Y	C	LLR6
15	187	K	N	K	K	LLR6
16	192	S	K	K	S	LLR6
17	200	N	N	N	K	LLR7
18	212	F	V	F	F	LRR7
19	213	H	Q	H	H	LRR7
20	285	Y	S	S	S	LRR8
21	299	N	K	N	N	LRR9
22	306	K	E	E	K	LRR9
24	310	S	I	I	I	LRR9
28	317	S	L	L	S	LRR9
29	319	Y	Y	Y	H	LRR9
30	346	Y	H	Y	Y	LRR10
31	355	N	N	H	N	LRR10
32	360	R	R	Q	R	LRR10
33	370	N	K	K	N	LRR11
34	378	S	Y	Y	S	LRR11
35	382	I	T	I	I	LRR11
36	390	T	K	K	T	LRR11
37	423	H	Q	Q	H	LRR12
38	435	S	N	N	S	LRR12
39	444	K	E	E	K	LRR12
40	468	S	I	I	S	LRR13
41	497	Q	R	R	Q	LRR14
42	521	H	H	Y	H	LRR15
43	522	K	E	K	K	LRR15
44	539	H	C	Q	H	LRR15
45	571	Q	H	H	Q	LRR
46	577	F	H	Q	F	LRR
47	581	Y	D	N	Y	LRR
48	601	T	T	N	T	LRR
49	619	E	N	D	V	LRR
50	686	V	A	A	A	LRRCT
51	766	I	T	T	I	TIR1

sphere). **Figure 1D** depicts the 3D structure of TLR3 of duck with the site for TIR (amino acid position 748–890, blue sphere) and the sites for leucine-rich receptor–like protein kinase (amino acid position 314–637, orange mesh). **Figure 1E** represents the sites for leucine-rich repeats as spheres at aa sites 53–74 (blue), 77–98 (red), 101–122 (yellow-orange), 125–145 (hot pink), 148–168 (cyan), 172–195 (orange), 198–219 (gray), 275–296 (raspberry), 299–319 (split pea), 346–367 (purple-blue), 370–393 (sand), 422–444 (violet), 447–468 (deep teal), 497–518 (olive), 521–542 (green), 553–574 (gray), 577–598 (salmon), and 601–622 (density). The other sites for posttranslational modification as observed were 16 sites for N-linked glycosylation, 8 sites for casein kinase 2 phosphorylation, 8

sites for myristoylation, and 9 sites for phosphokinase phosphorylation.

In comparison for TLR3 among the avian species, 51 amino acid variations were observed, which contribute to various important domains of TLR3, including LRR, LRRCT, and TIR domains (**Table 2**).

Molecular Characterization of RIGI of Duck

RIGI of duck has been characterized (Gene bank accession number KX865107, protein ASW23002 from NCBI). The 3D structure of RIGI of duck is depicted in **Figures 2A,B** (surface view). RIGI is an important gene conferring antiviral immunity. A series of posttranslational modification and various domains for its important function have been represented.

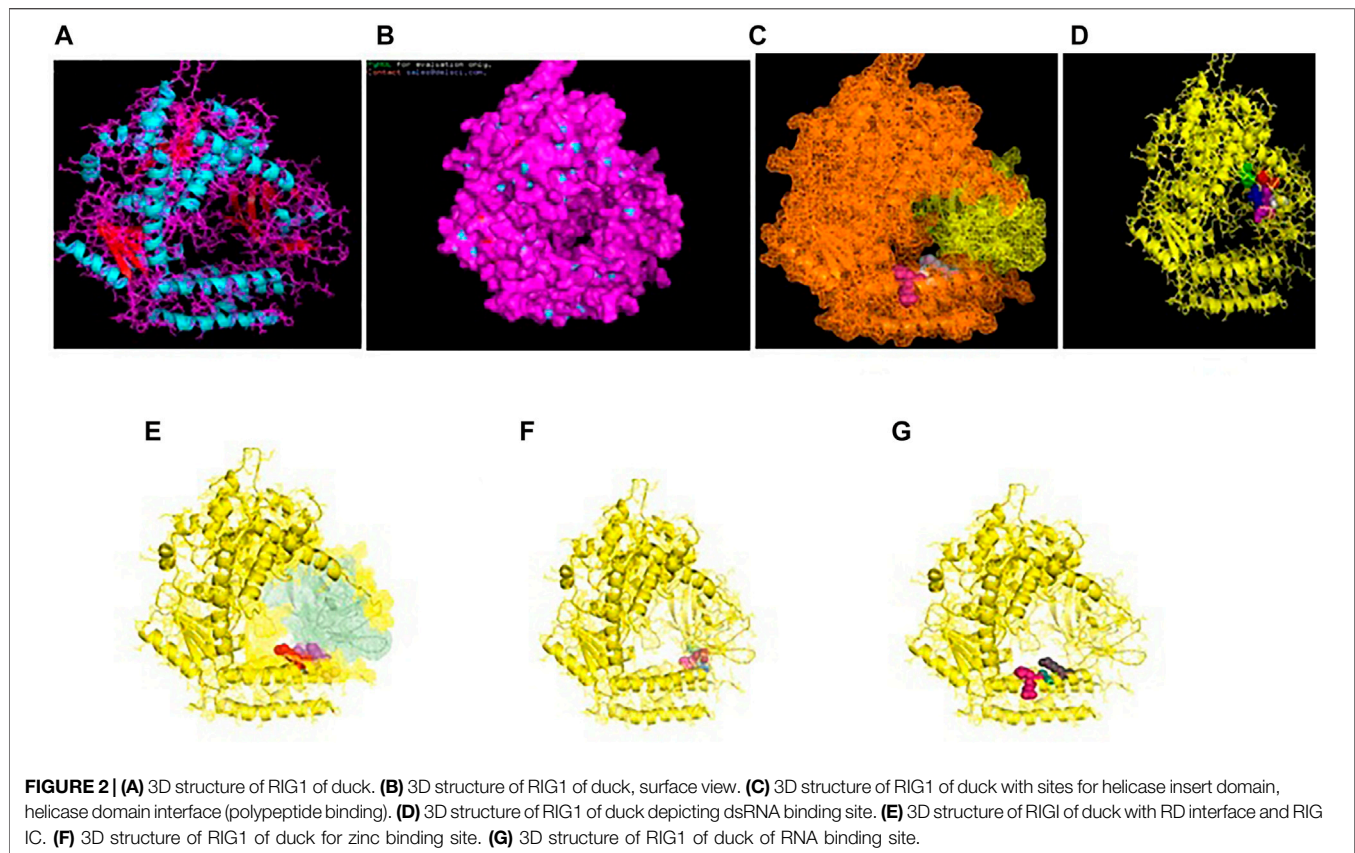
CARD_RIG1 (caspase activation and recruitment domain found in RIG1) have been depicted in amino acid positions 2–91 and 99–188. CARD2 interaction site (17–20, 23–24, 49–50, and 79–84), CARD1 interface (100, 103, 130–135, 155, 159, and 161–162), and helical insert domain interface (101, 104–105, 107–108, 110–112, 114–115, 139, 143–145, 147–148, 151, 180, 183–184, and 186 aa) have been depicted at RIG1 of duck. **Figure 3C** depicts helicase insert domain (242–800 aa) as orange mesh, helicase domain interface (polypeptide binding) as (511–512aa warm pink, 515aa white, 519aa gray). **Figure 2D** depicts the double-stranded RNA binding site (nucleotide-binding) at amino acid positions 832 (red), 855 (green), 876–877 (blue), 889–891 (magenta), and 911 (white). The sites for RD interface (polypeptide binding) and RIG-I-C (C terminal domain of retinoic acid-inducible gene, RIG-I protein, a cytoplasmic viral RNA receptor) have been depicted in **Figure 2E**. The site for RIG-I-C as amino acid position 807–921 is represented by a mesh of pale green tints. The sites for RD interface have been depicted as amino acid positions 519 (red sphere), 522–523 (magenta), 536–537 (orange), and 540 (gray).

Figure 2F depicts the sites for the zinc-binding domain of RIG1 of duck as amino acid positions 812 (firebrick), 815 (marine blue), 866 (green), and 871 (hot pink). The sites for RNA binding have been depicted as 511–512 (hot pink), 515 (cyan-deep teal), and 519 (gray) in **Figure 2G**.

Molecular Characterization of TLR7 of Duck

TLR7 gene has been characterized in duck (Gene bank Accession no. MK986726, NCBI). The 3D structure of TLR7 is depicted in **Figures 3A,B** (surface view). TLR7 is rich in leucine-rich repeat (LRR) as depicted in **Figure 3C**. The LRR sites are 104–125 (red sphere), 166–187 (green), 188–210 (blue), 243–264 (yellow), 265–285 (magenta), 288–309 (cyan), 328–400 (orange), 435–455, 459–480 (gray), 534–555 (warm pink), 558–628 (split pea), 691–712 (purple-blue), 715–762 (sand), and 764–824 (deep teal).

The other domains are GPI anchor at 1072 amino acid position (red sphere), domain linker sites such as 294–317 (green sphere) and 467–493 (split pea sphere) (**Figure 3D**). The TIR site had been identified as 929–1076 amino acid position (blue sphere) and cysteine-rich flanking region, and C-terminal as 823–874 amino acid position (hot pink) as



depicted. The site for LRRNT (leucine-rich repeat N-terminal domain) of TLR7 had been identified at amino acid position 75–107 (red surface), TPKR-C2 (tyrosine-protein kinase receptor C2 Ig-like domain) at amino acid position 823–869 (blue surface) and GPI anchor as a green sphere (Figure 3E). Figure 3F represents the transmembrane site for TLR7 of duck.

Molecular Docking of TLR3, RIGI, and TLR7 Peptides With the Antigenic Binding Sites of H and N Antigens of Avian Influenza Virus

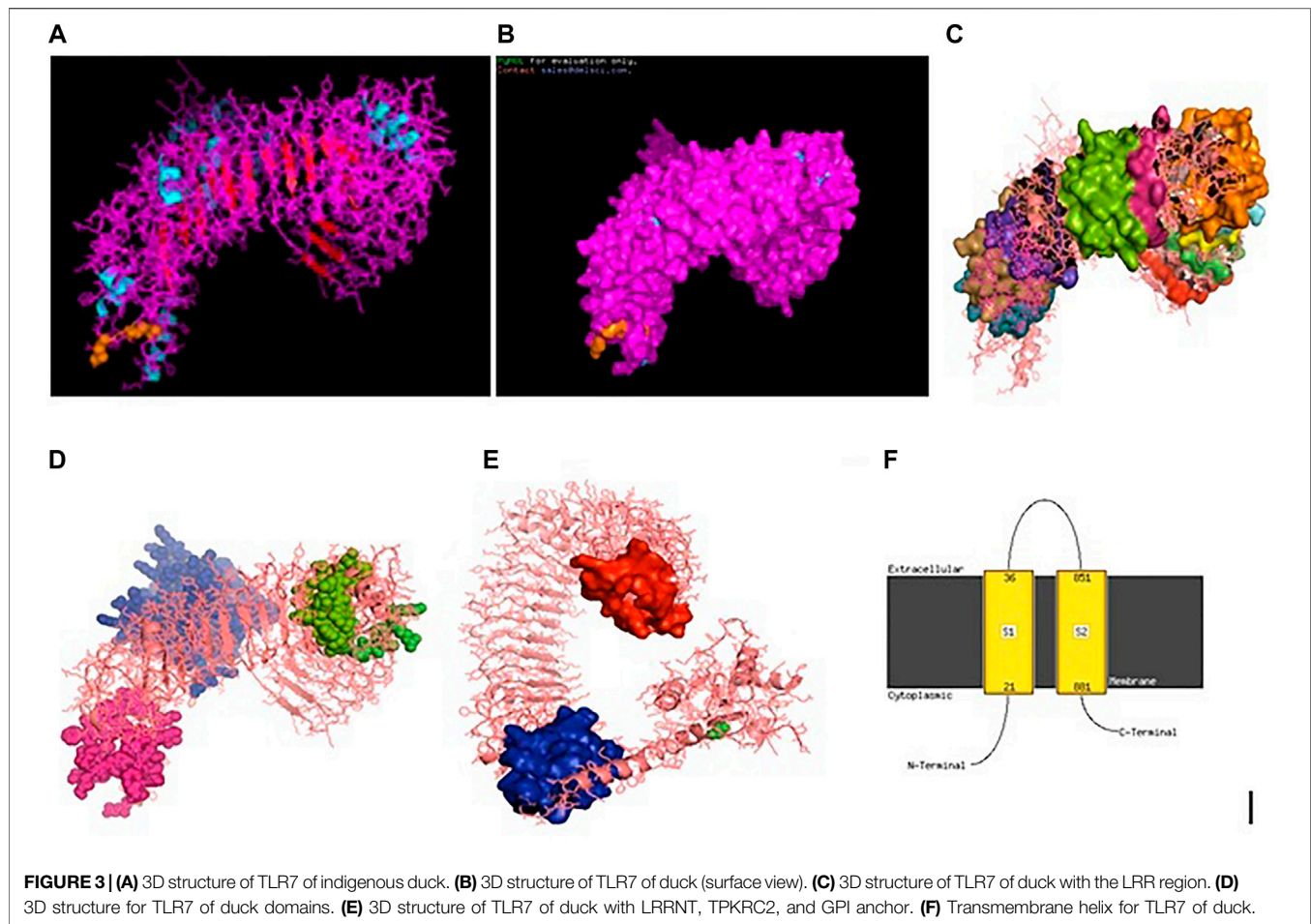
Binding for H and N antigens was observed for different strains of avian influenza with RIGI, TLR7, and TLR3 (Figure 4). PatchDock analysis has revealed a high score for hemagglutinin and neuraminidase antigen for the H5N1 strain of avian influenza virus. The PatchDock score for H antigen for RIGI, TLR7, and TLR3 was observed to be 19920, 20532, and 22880, respectively, whereas the PatchDock score for N antigen for RIGI, TLR7, and TLR3 was 21570, 20600, and 21120, respectively, as detailed in **Supplementary Figure S1**. The binding scores were observed to be sufficiently high. The highest score was obtained for H antigen with TLR3.

Ligand binding is very much important for the receptor molecule. In our current study, we had studied only the surface antigens such as hemagglutinin and neuraminidase that are involved in binding with the immune molecules.

The binding of RIGI of duck with the hemagglutinin H5N1 strain of avian influenza is being depicted with certain domains highlighted. The binding site of RIGI with H antigen of the H5N1 strain of avian influenza virus extends from 466 to 900 amino acid positions as blue spheres (Figure 5A). The red surface indicates the helicase interface domain (511–512, 515, 519 aa position of RIGI). Amino acid position 519 is a predicted site for helicase interface domain as well as a site for RD interface. The site for helicase interface domain is depicted as the yellow surface. Another important domain within the site includes zinc binding domain depicted as the green surface.

The binding of RIGI of duck with the neuraminidase H5N1 strain of avian influenza is being depicted with certain domains highlighted. The binding site of RIGI with the N antigen of the H5N1 strain of avian influenza virus extends from lysine 245 to isoleucine 914 amino acid positions as blue spheres (Figure 5B). Figure 5B depicts only the aligned region of neuraminidase and RIGI. The site for RIG-I-C (C terminal domain of retinoic acid-inducible gene ranging from 807–921 aa position by yellow stick).

An interesting observation was that in the CARD domains such as CARD_RIG1 (caspase activation and recruitment domain found in RIG1) and CARD2 interaction site, CARD1 interface was not involved in binding with both the surface protein hemagglutinin and neuraminidase of avian influenza virus. This was proved through the pdb structure of RIGI developed with Modeller software (Figures 5C,D), respectively, for H- and N-antigen.



The binding of TLR3 of duck with the neuraminidase H5N1 strain of avian influenza is being depicted with certain domains highlighted. The binding site of TLR3 with N antigen of the H5N1 strain of avian influenza virus extends from threonine 34 to isoleucine 459 amino acid positions as green spheres (**Figure 6A**). Identifiable domains within this region include LLR 1 to 12, site for leucine zipper, leucine-rich nuclear export signal, and LRRNT. The domains within the 3D structure of TLR3 have been visualized already in **Figures 1A–E**.

The binding of TLR3 of duck with the hemagglutinin H5N1 strain of avian influenza is being depicted with certain domains highlighted. The binding site of TLR3 with the H antigen of the H5N1 strain of avian influenza virus extends from asparagine 53 to arginine 895 amino acid positions as green spheres (**Figure 6B**). The important domains within this region include LRR region 1–18, site for leucine zipper, GPI anchor, leucine-rich nuclear export signal, LRRCT, site for TIR, and the sites for leucine-rich receptor-like protein kinase.

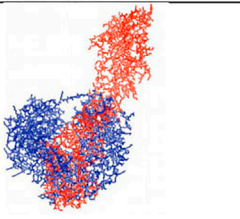
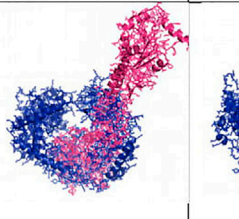
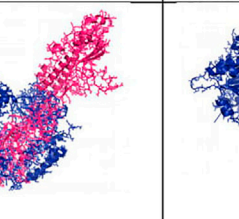
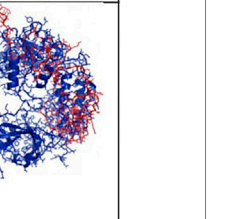
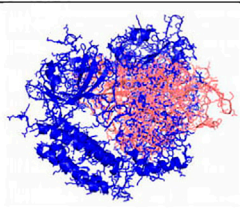
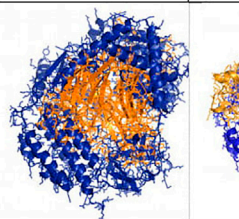
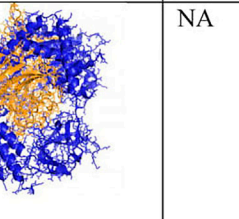
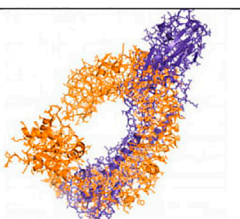
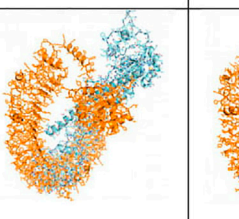
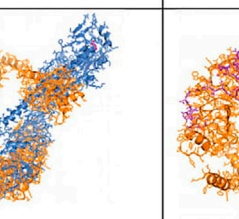
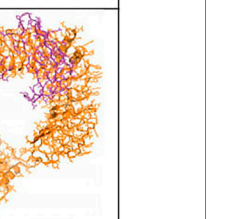
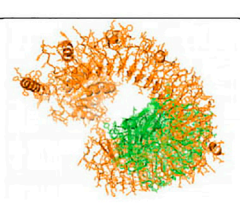
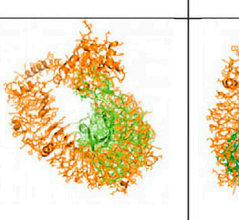
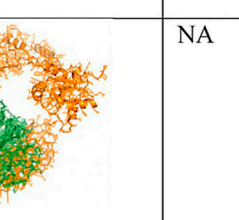
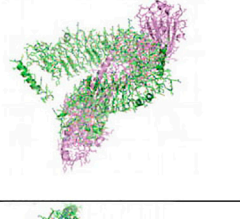
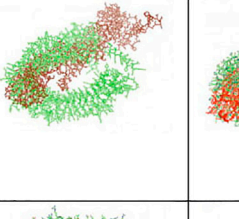
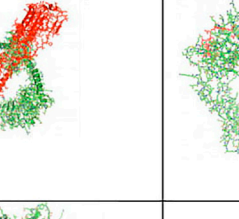
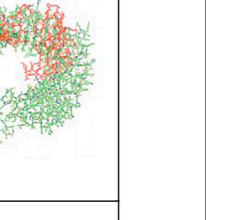
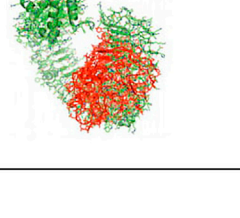
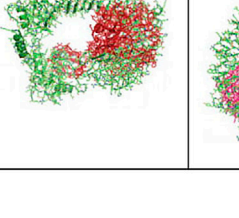
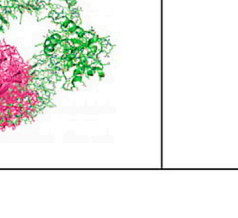
The binding of TLR7 of duck with the neuraminidase H5N1 strain of avian influenza is being depicted with certain domains highlighted. The binding site of TLR7 with the N antigen of the H5N1 strain of avian influenza virus extends from valine 87 to glutamine 645 amino acid positions as orange spheres (**Figure 7A**). Identifiable important domains within this region

include LRR region 1–11 and domain linker sites. The detail visualization of these domains is present in **Figures 3A–F** in the molecular visualization tool.

The binding of TLR7 of duck with the hemagglutinin H5N1 strain of avian influenza is being depicted with certain domains highlighted. The binding site of TLR7 with the H antigen of the H5N1 strain of avian influenza virus extends from aspartic acid 293 to tyrosine 909 amino acid positions as orange spheres (**Figure 7B**). The important domains responsible within this binding site include LRR 7 to LRR14, domain linker sites, and cysteine-rich flanking region C-terminal, and TPKR-C2 (tyrosine-protein kinase receptor C2 Ig like domain).

Amino Acid Sequence Variability and Molecular Phylogeny Among Different Strains of Avian Influenza

High degree of sequence variability has been observed in **Figures 8A,B** in the hemagglutinin segment of avian influenza and **Figures 8C,D** in the neuraminidase segment of avian influenza. The H5N1 strain was observed to be clustered with the H6N2 strain of avian influenza virus. The H4N6 strain which is comparatively less virulent and hence causing LPAI was found to possess certain uniqueness

		H5N1	H4N6	H6N2	H9N2 (partial seq)
RIGI (blue-3D structure)	H- antigen				
	N- antigen				NA
TLR7 (Orange 3D structure)	H- antigen				
	N- antigen				NA
TLR3	H- antigen				
	N- antigen				

3

FIGURE 4 | Alignment/binding of identified immune response molecules with hemagglutinin and neuraminidase for different strains of avian influenza.

in amino acid sequence. Deletions of four consecutive amino acids at positions 65–68 and 140–141 were observed in hemagglutinin. Likewise, insertion mutations were also

observed at amino acid positions 19–24 and 77–79 in hemagglutinin. Cysteine residues were observed to be conserved across the strains (**Figure 8B**).

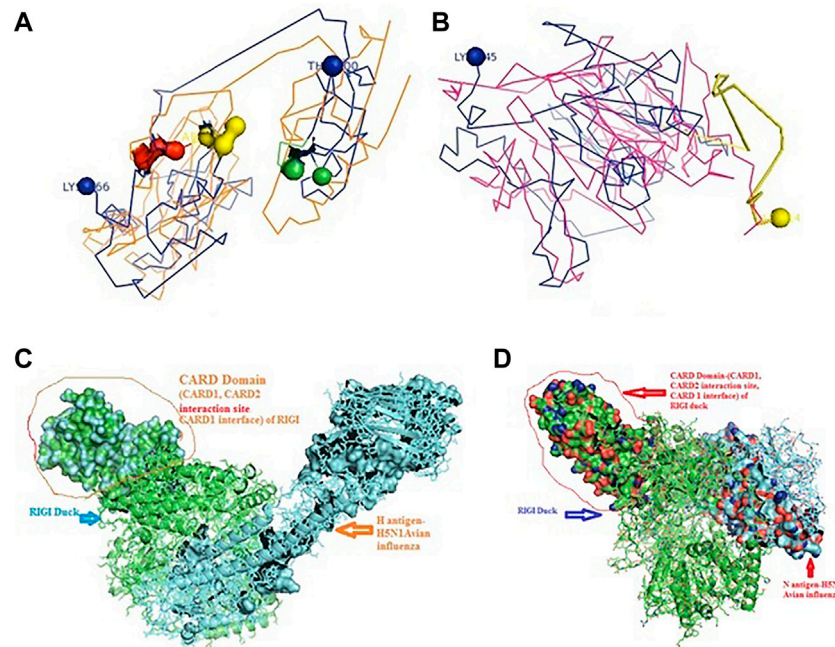


FIGURE 5 | Molecular docking of RIGI with surface antigen for avian influenza. **(A)** Molecular docking image RIG1 of duck with antigen (AI) ligand H antigen-binding site detection. **(B)** Molecular docking image RIG1 of duck with antigen (AI) NS antigen ligand (**left**) and binding site (**right**). **(C)** Duck RIGI model binding with H-antigen of avian influenza virus, with CARD domain—no binding for CARD with virus. **(D)** Duck RIGI model binding with N-antigen of avian influenza virus, with CARD domain—no binding for CARD with virus.

However, in such a case, highly pathogenic H5N1 depicts certain deletion mutations at amino acid positions 37–45, 53–63, and 80–82 of neuraminidase (**Figure 8D**).

Comparative Structural Analysis of TLR3 and TLR7 of Duck With Respect to Chicken

Ducks were reported to be genetically more resistant to chicken, particularly in terms of viral infections. Accordingly, the structural alignment of the 3D structure of TLR3 of duck with chicken has been described (**Figure 9A**). 3D structural alignment of TLR7 of duck with chicken has been visualized (**Figure 9B**). It was not possible to study the structural alignment of duck RIG-I since RIGI was not expressed in chicken.

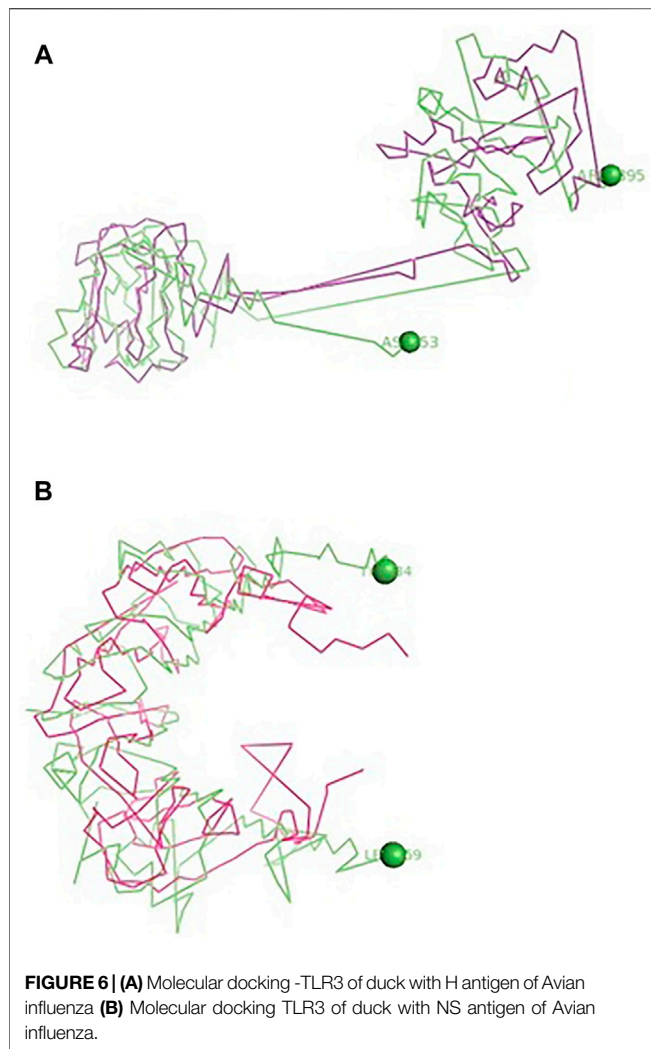
The sites for non-synonymous mutations have been depicted for TLR3 gene in duck with respect to other poultry species such as chicken, goose, and guinea fowl (**Table 2**). 51 sites for amino acid substitutions have been detected, ranging from amino acid position 42 to 766 in duck with respect to other poultry species, which actually contribute to changes in functional domains of TLR3. A comparison of TLR3 of duck with chicken actually revealed 46 sites of amino acid substitution resulting due to non-synonymous mutations, which are of much importance to our present study. Most of the substitutions caused changes in leucine-rich repeats, which is an inherent characteristic for pattern recognition receptor such as TLR2. 20 sites of amino acid substitutions that were specific for anseroides (duck and goose) were identified.

Protein–Protein Interaction Network Depiction for TLR3 and TLR7 With Respect to Other Functional Proteins

Interaction of TLR3 with other proteins has been depicted in **Figure 10A** with STRING analysis. Interaction of TLR7 with other proteins of functional interest has been depicted in **Figure 10B**. KEGG analysis depicts a mode of the defense mechanism of influenza A and the possible role of antiviral molecules in combating the infection and the role of antiviral molecules through the TLR signaling pathway.

Differential mRNA Expression Pattern of TLR7 and Other TLR Genes of Duck With Respect to Chicken and Other Poultry Species

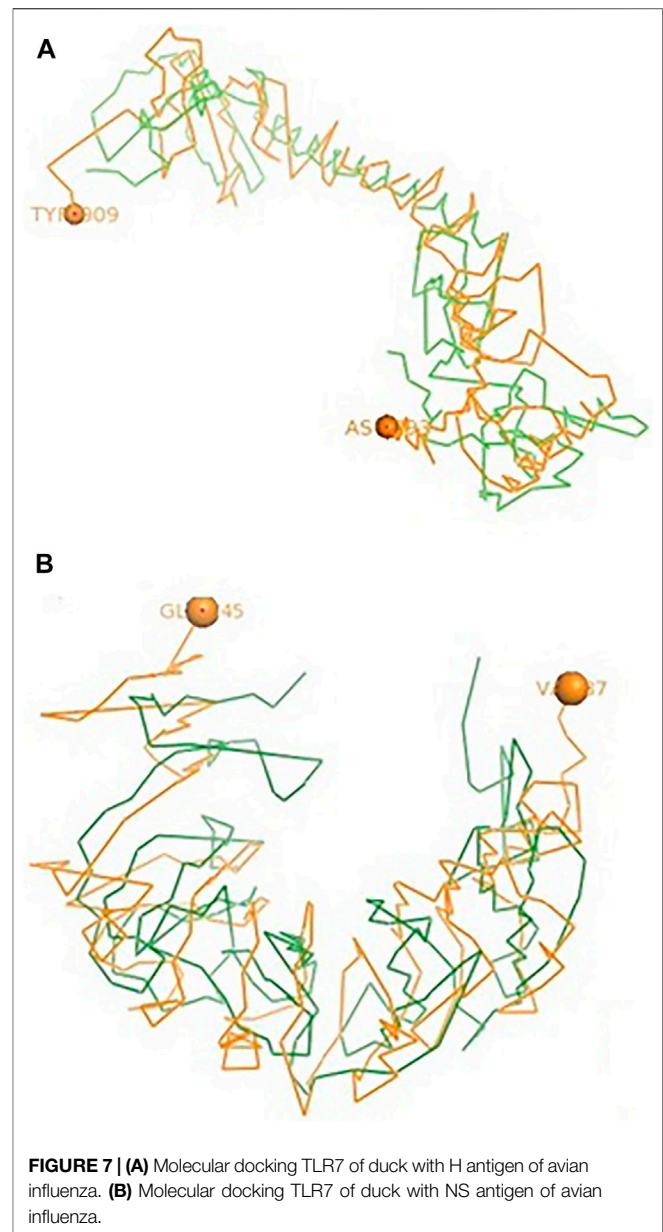
We conducted differential mRNA expression profiling of TLR2, TLR4, and TLR7. Expression profiling of TLR2 and TLR4 was observed to be better in indigenous chicken (Aseel and Haringhata Black) and guinea fowl than in anseroides (duck and guinea fowl). Both TLR2 and TLR4 are known to impart antibacterial immunity. Quantitative mRNA expression analysis clearly depicts TLR7 gene expression was definitely better in duck than in other poultry species such as goose, guinea fowl, and indigenous chicken breed (Aseel and Haringhata Black chicken) (**Figure 11**). This gives an indication that better immune response of indigenous duck may be due to the increased expression level of TLR7, which confers antiviral resistance (**Figure 11**).



Differential mRNA expression profiling for TLR7, TLR3, and RIGI genes with respect to infected versus control Bengal duck and HB chicken have been depicted in **Figures 12–14**, respectively. **Figures 12A,B** depict the TLR7 gene expression level in the infected condition with respect to healthy control in duck and chicken, respectively. Similarly, **Figures 13A,B** represent the TLR3 gene expression level in the infected condition with respect to healthy control in duck and chicken, respectively. However, **Figure 14** represents only the expression of RIGI in duck. In all the cases, the higher expression level of TLR7, TLR3, and RIGI was detected in infected cases than in healthy control birds. RIGI was reported to have significantly pronounced better expression in infected ducks than healthy control.

Phylogenetic Analysis of Indigenous Ducks With Other Poultry Species and Other Duck Population Globally

With an aim for the identification of the status of molecular evolution of duck, the indigenous duck gene sequence of West Bengal, India, was compared with other duck sequences globally.



Phylogenetic analysis was performed with respect to TLR7 (**Figure 15A**) and TLR3 (**Figure 15B**). Phylogenetic analysis revealed that ducks of West Bengal were observed to be genetically more closely related to the duck population of China (**Figure 15A**). Ducks were observed to be genetically closest to goose (**Figure 15B**). Chicken, quail, and turkey were observed to be genetically distinct from duck (**Figure 12B**).

DISCUSSION

Indigenous duck population was characterized to be very hardy, and usually asymptomatic to common avian diseases. But there is a paucity of information regarding the systemic genetic studies on duck involved in its unique immune status. It is evident that duck

A

CLUSTAL format alignment by MAFFT (v7.467)

```

AAG53040.1 -----
AKD00332.1 MEKIVLLFATISLVKSDH-----ICIGYHANNSTEQVDTIMEKNVTVTTHAQDILEKTHN
AMH93683.1 MIAFIVIAILVATGKSDK-----ICIGYHANNSTTTVDITILEKNVTVTTHSVELLENQKE
AFO82958.1 MLSIVILFLLVAE3SSQNYTGNFVICMGHHAVANGTMVKTLTDDQVEVWTAQELVESQNL

AAG53040.1 -----
AKD00332.1 GKLCDLNGVKPLILKD---CSVAGWLLGNPLCGEPTNVPEWSYIVEKANPANDLCYFGNF
AMH93683.1 ERFCRISNKAFLDLRD---CTLEGWILGNFRCGVLLADQSWSYIVERPNARNGICYPGTL
AFO82958.1 PELC----P8PLRLVDGQTCIDIINGALGSPGC-DHLNDAKWDIFIERPNAV-DTCYPTDV

AAG53040.1 -----
AKD00332.1 NDVEELKHLISRINHFEKIQIIPKDSWSDHEASLGVSAACSYQGNSSFFRMVWVLIK-KD
AMH93683.1 NEAEELKALIGSGEKVERFEMFFKNTWAGVDTDRGVSSACPLGNGPSTFYRNLLWIIKLQS
AFO82958.1 PDVQSLRSILANNGKFEPIA--EEFQWSTVKQN-GKSGACKRANVNDFFNRLNWLKVSNG

AAG53040.1 -----WGIHHPPTDTAQTNLYIRTDITTSVTTEDINRTFKPVIG
AKD00332.1 NAYFTIKKGYMNTNREDLLILWGIHHPNDEAEQTKLYQNPTTYISIGTSTLNQRLVPKIA
AMH93683.1 SEYFVIRGTFNNTGDKSILYFWGVHHPVTVDDEQKALYSGDRYVRMGTSMMNFARSPPIA
AFO82958.1 DAYFLQNLTKVNNGDYARLYIWGVHHPSTDTEQTNLYKNNPGRVTVSTKTSQTSTVVPNIG
          +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+

AAG53040.1 PRPLVNGQQGRIDYVWSVLKPGQTLRVRSGNGLIAPWYGHV-LSGGSHGRILKTDNLN3GK
AKD00332.1 TRSKINGQSGRIDFFWTILKPNDIAHPESNGNFIAPFYAYK-IVKKGDSITMRSEVEYGN
AMH93683.1 ARPAVNGQGRIDYFWSILKPGETLNVESNGNLIAPWYAYRFVNKDSKGAIFRSDLPIEN
AFO82958.1 SRPWVAGQSGRISFYWTIVEPGDLIVFNTIGNLIAPRGHYK-LNSQKKSTILNTATPIGS
          +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+

AAG53040.1 CVUQCQTEKGGNLSTLPFHNVSKYAFGTCPKYVGVKSLKLAUGLRNVPAVS---SRGLF-
AKD00332.1 CNTRCQTPLGAINSSMPFHNHPLTIGECPKYVKSNNKLVLATGLRNSPQRERRRKRGLFG
AMH93683.1 CDATCQTAEGVLRTNKTFCQNVSEPLWIGECPKYVKSNSLRLATGLRNVQIE---TRGLFG
AFO82958.1 CVSRCHTDKGSLSSTTKPFQNIARIGDCPKYVKGSLKLATGMRNIPKA---SRGLFG
          +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+ +*+

AAG53040.1 -----
AKD00332.1 AIAGFIEGGWQGMVDGWYGYHHSNEQSGYAADKESTQKAIDGVITNKVNSIIDKMNTQFE
AMH93683.1 AIAGFIEGGWTGMIDGWYGYHHSNEQSGYAADKESTQKAIDGITNKVNSIIDKMNTQFE
AFO82958.1 AIAGFIENGWQGLIDGWYGFHQNAREGTGTAADLKSTQAADQINGKLNRLIEKTNERKYH

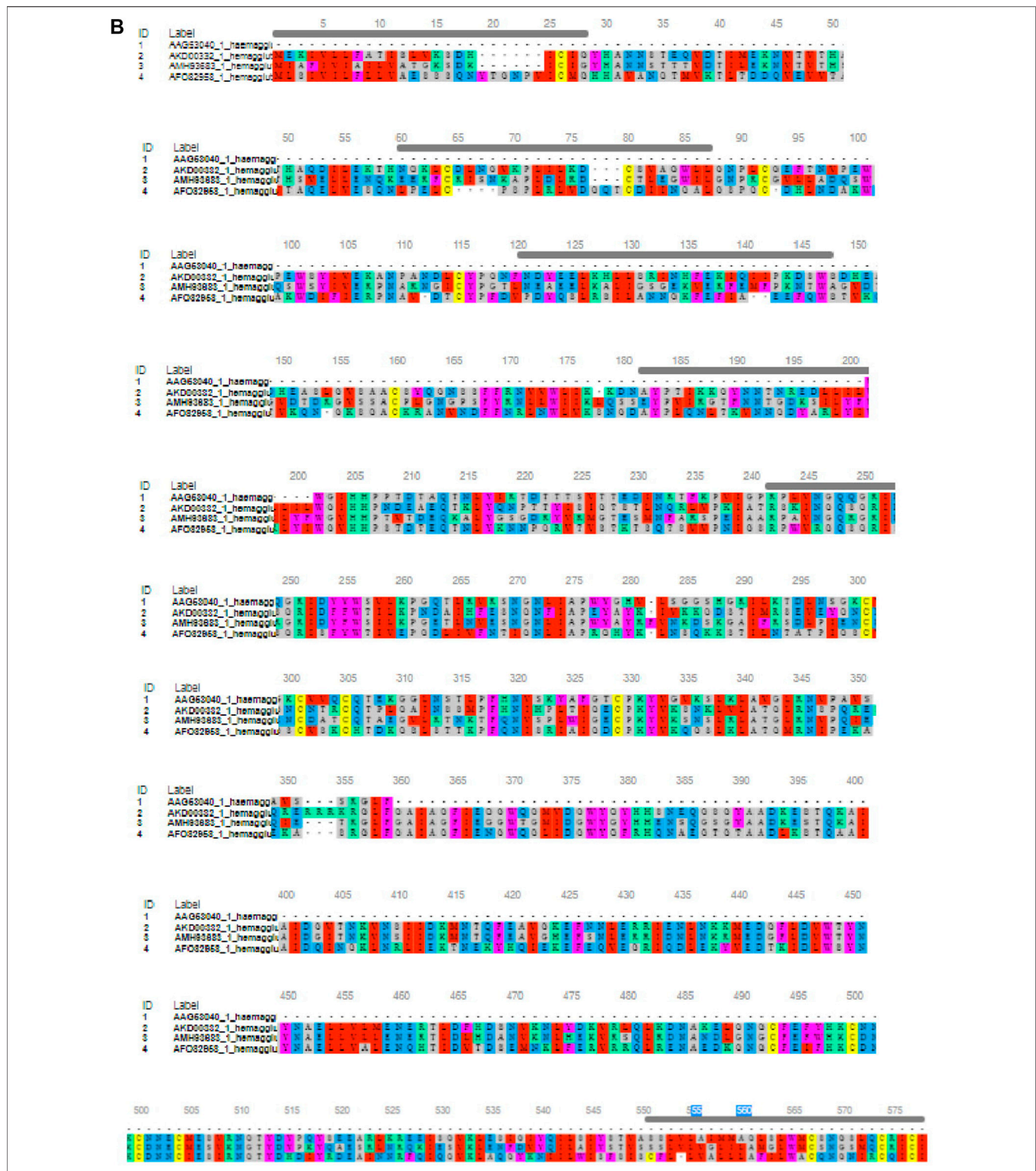
AAG53040.1 -----
AKD00332.1 AVGKEFNLERRRIENLNKRMEDGFLDWWTYNALLVLMENERTLDFHDSNVKNLYDKVRL
AMH93683.1 AVGHEFSNLERRRIDNLRMEDGFLDWWTYNALLVLENERLTLDLHDANVKNLHEKVRAS
AFO82958.1 QIEKEFEQVEGRIQDLEKYVEDTKIDLWSYNAELLVALENQHTIDWTDSEMKNKLFERVR

AAG53040.1 -----
AKD00332.1 QLKDNAKELGNGCFEFYHKCNNECMESVRNGTYDYPQYSEEARLKREEISGVKLESIGIY
AMH93683.1 QLRDNANDLNGNGCFEFWHKCDNECMESVKNNGTYDYPKYQAESRLNRQKIESVKLENFDVY
AFO82958.1 QLRENAEDKNGNGCFEIFHKCDNNCIESIRNGTYDHDIIYRDEAINNRFQIQGVKLAQGYKN

AAG53040.1 -----
AKD00332.1 QILSIYSTVASSLVLAIMMAGLSLNMCSNGSLQCRICI
AMH93683.1 QILAIYSTV88SLVLVGLILAMGLNMCSNGSMQCRICI
AFO82958.1 IILWISF818CFL-LVALLAFILMACQMGNIROQICI

```

FIGURE 8 | (A) Alignment report for Haemagglutinin antigen for different strains for Avian influenza **(B)** Alignment report for Haemagglutinin antigen for different strains for Avian influenza by MAFFT software (MSA viewer) **(C)** Alignment report for neuraminidase antigen for different strains for Avian influenza **(D)** Alignment report for neuraminidase antigen for different strains for Avian influenza by MAFFT software (MSA viewer).



possesses some unique genetic makeup which enables it to provide innate immunity against viral infection, particularly avian influenza.

In this current study, we identified three immune response molecules, which were earlier known to have immunity against

viral infection such as RIGI, TLR7, and TLR3. We characterized these proteins of duck, and attempted to identify the SNPs or variations in nucleotide among duck and chicken. Through molecular docking, the most promising IR molecules conferring innate immunity against avian influenza have been

C

CLUSTAL format alignment by MAFFT (v7.467)

```

ALK80138.1  MNPNQKIIVTIGSICMVGIIISLMLQIGNIISIWVSH-----SIQTGNQ-----
AFO82959.1  MNPNQKIICISATGMTLSVSVLLIGIANLGLNIGLHYKVGDTPDVNISSMKNKTSTTTTII
AMH93685.1  MNPNQKIITIGSVSLTIATVCFMLQIAILATTITLHFKQN---ECSIPSNQVWPCEPII
*****; *: : : : : : * : * : * :

ALK80138.1  ---HQTEPIRNTNFLTENT---VASVTLAGNSSLCPIKGWAVHSKDNSIRIGSKGDFVI
AFO82959.1  NNNTQNNFTNITNIIIVNKE---EERTFLNLTKPLCEVNSWHILSKDNAIRIGEDAHILVT
AMH93685.1  VERNITEIVYLNNTTIEKELCPKLTEYRDWSKPPQQITGFAPFSKDNSIRLSAGGDIWVT
.: : * : : : : * : : : : : : : : : :

ALK80138.1  REPFISCSHLECRITFFLTQGALLNDKHSNGTVKDRSPHRTLMSCPIGEAPSPYNSRFESV
AFO82959.1  REPYLSCDPQGCRCMFALSQGTTLRGQHANGTIHDRSPFRALVSWEMGOAPSPYNVRVECV
AMH93685.1  REPYVSCSPNKCYQFALGQGTTLDNKHSNGTIHDRIPHRTLMLNELG-VFHLGKQVCI
***:***. * * * **: * :*:***:*** *.**: :* . * . : .:

ALK80138.1  AWSASACHDGTSWLIIGISGPDNGAVAVLKYNIGIITDTIKSWRNNILRTQESECACVNGS
AFO82959.1  GWSSTSCHDGISRMSICSGPNNNASAVVWYGGRPVTEIPSWAGNILRTQESECVCCHKGI
AMH93685.1  AWSSSSCHDGRWLHVCTVGDDRNATASFIYDGVLDVSIGSWSQNILRTQESECICINGT
.**:***: : : : : : : : * * . * * . * * * : : : : : : : : : :

ALK80138.1  CFTVMIDGPSNGQASYKIFKIEKGKVVKSVELNAPNYHYEECSYPEAGEIICVCRDNWH
AFO82959.1  CPVVMIDGPANNRAATKIIFKEGKIQKIEELEGNAQHIEECSCYGAAGVIKICICRDNWK
AMH93685.1  CTVVMIDGSASGRADTRILFIKEGKIVHISPLSGSAQHIEECSCYPRYPDVRCVCRDNWK
* .*****: : : : : : : : * . * : : : : : : : : : : : : : : :

ALK80138.1  GSNRFWVSFN-QNLEYQIGYICSGVFGDNPRPNDGT--GSCG-P-VSSNGAYGVKGFSEK
AFO82959.1  GANRFVIIIDPEMMHTSKYLCRVLTDTSRPNDPTS-GNCDAPVTGGSPDPGVKGF AFL
AMH93685.1  GSNRFVIDINMADYSIDSSVCSGLVGDTPRNDSSSNCKDP-NNERGNPVGKGFADF
*:*** : : : : : : : : : : : : : : : : : : : : : : : : : :

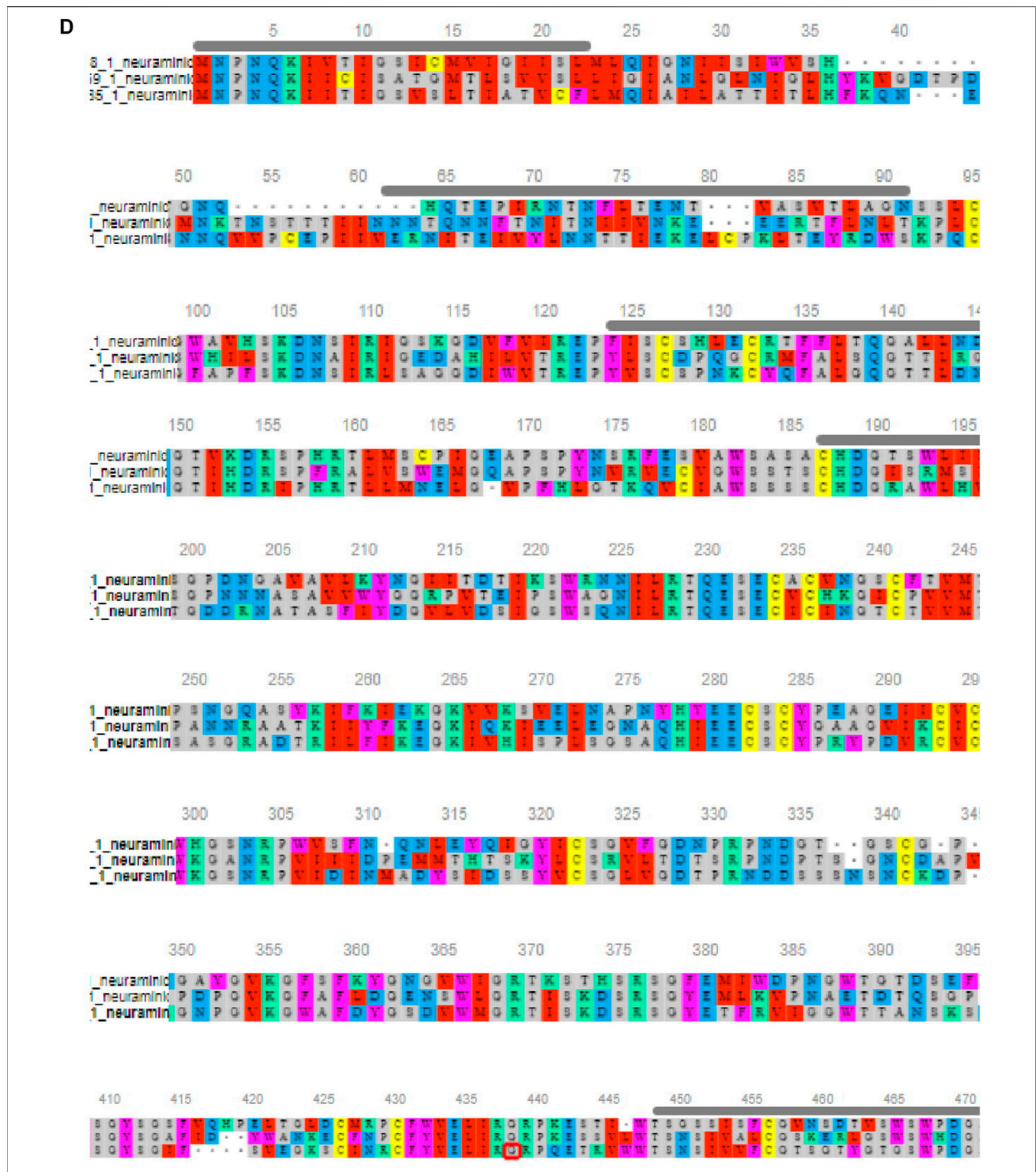
ALK80138.1  YGNGVWIGRTKSTHSRSGFEMIWDPNGWGTGTDSEF-SMKQDIVAITDWSGYSGSFVQHPE
AFO82959.1  DGNSWLGRITISKDSRSGYEMLKVPNAETDTQSGP-ISHQVIVNNQNWSGYSGAFID--Y
AMH93685.1  YGSDVWMGRITISKDSRSGYETFRVIGGWTTANSKSKQVNRQVIVDNNWWSGYSGIF----S
*.. *.**** *..***: : : : : : : : : : : : : : : : : : : : : :

ALK80138.1  LTGLDCMRPCFWVELIRGRPESTI-WTSGSSISFCGVNSDTVSWSWPDGAELPFT-I
AFO82959.1  WANKECFNPCFYVELIRGRPESSVLWTSNSIVALOGSKERLGSWSWHDGAEIIFY-K
AMH93685.1  VEGKSCINRCFYVELIRGRPEQETRVWWTNSIVVFCGTSGTYGTGSWPDGANINFMPI
. .*: . **:*:***:***: : : : : : : : : : : : : : : : : : : : :

```

identified, which were later on confirmed through the wet lab study as differential mRNA expression profiling. We attempt to explore the unique genetic constitution of duck immune response with respect to that of chicken. However, RIGI was reported to be expressed only in duck, not in chicken; hence, comparison was not available. TLR7 expression profiling was observed to be significantly better in duck than in chicken and other poultry species, indicative of better antiviral immunity in ducks. 53 non-synonymous mutations with amino acid variations were observed while comparing amino acid sequence of duck with other poultry species, including chicken, most of which are confined to the LRR domain. We had already depicted that LRR is an important domain for pathogen binding site as in this case of avian influenza. For the effective antiviral activity, binding of viral protein with the immune response molecule is the primary criterion. Leucine-rich repeats (LRRs) were observed to be

important domain involved in binding with hemagglutinin and neuraminidase surface protein in case of both TLR3 and TLR7. Similar studies have also reported LRR as the important domain against bacterial infections in case of CD14 molecule in cattle (Stetson and Medzhitov, 2006), goat (Vercammen et al., 2008a; Schlee et al., 2009), and buffalo (Pal and Chatterjee, 2009; Pal et al., 2013). Other important domains identified were LRRNT, LRRCT, site for TIR, and the sites for leucine-rich receptor-like protein kinase, including certain posttranslational modification sites. Similar reports were also identified in different species (Stetson and Medzhitov, 2006; Vercammen et al., 2008a; Pal and Chatterjee, 2009; Schlee et al., 2009; Pal et al., 2013). An important observation identified was that although the CARD domain was believed to be an important binding site for RIGI for some identified virus, it has no binding ability with avian influenza virus. Other studies have reported



the role of the CARD domain in binding with the MAVS domain as a part of antiviral immunity (Kelley and Sternberg, 2009; Šali et al., 2020).

TLR3 (CD283 or cluster of differentiation 283) is a pattern recognition receptor rich in leucine-rich repeats as revealed in duck TLR3 of the current study. Other important domains

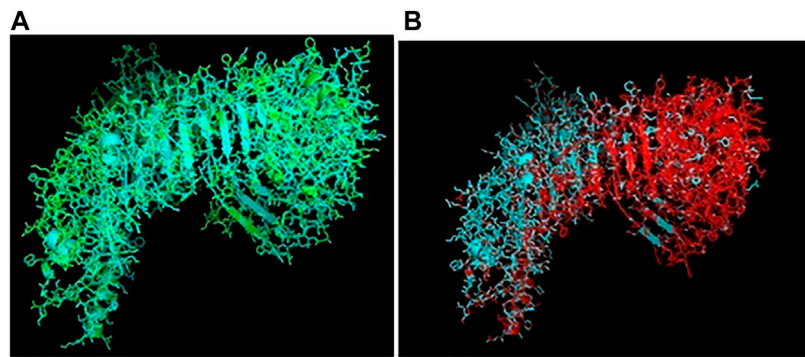


FIGURE 9 | (A) Comparison of TLR3 of duck and chicken-3D structural analysis **(B)** Comparison of TLR7 of duck and chicken-3D structural analysis.

include LRRNT, LRRCT, TIR, leucine-rich receptor-like protein kinase, leucine zipper, GPI anchor, and leucine-rich nuclear export signal. The other sites for posttranslational modification as observed were N-linked glycosylation, casein kinase, phosphorylation, myristoylation, and phosphokinase phosphorylation. Variability of amino acids in important domains was observed for duck TLR3 compared to that of chicken, goose, and turkey. It was observed that genetic similarity between duck and goose was more than that of chicken and turkey. Some amino acids which are conserved for ducks and denote for important domains such as LRR, LRRCT, and TIR have been identified. In the LRRCT domain, valine is present in a duck in contrast to alanine in chicken, turkey, and goose. The current study identified 45 sites of non-synonymous substitutions between duck and chicken, which affect important domains for TLR3 as a pattern recognition receptor. Although it is the first report of characterization of TLR3 in indigenous duck, earlier studies were conducted in Muscovy duck, when full-length cDNA of TLR3 was characterized to be of 2836 bp encoding polypeptide of 895

amino acids (Zhang and Skolnick, 2005). The characterization of the deduced amino acid sequence contained 4 main structural domains: a signal peptide, an extracellular leucine-rich repeats domain, a transmembrane domain, and a Toll/IL-1 receptor domain (Zhang and Skolnick, 2005), which is in agreement to our current study. It is to be noted that TLR3 is a PRR, with the secondary structure being visualized as helix, loop, and sheet with the sites for disulfide bond being depicted in blue spheres. It recognizes dsRNA associated with a viral infection, and induces the activation of IRF3, unlike all other Toll-like receptors which activate NF- κ B. IRF3 ultimately induces the production of type I interferons, which aid in host antiviral immunity (Morgan et al., 2019). In the current study, we observed sites for leucine zipper in duck TLR3, which is an inherent characteristic for dimerization. Earlier studies have also reported that TLR3 forms a large horseshoe shape that contacts with a neighboring horseshoe, forming a “dimer” of two horseshoes (Niu et al., 2019). As already explained that glycosylation is an important PTM (posttranslational modification site), it acts a glycoprotein. But in the proposed interface between the two horseshoe structures,

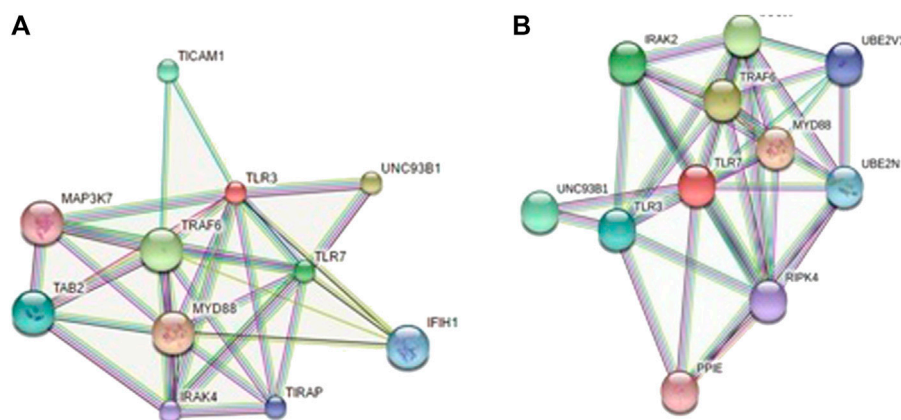


FIGURE 10 | (A) String analysis revealing molecular interaction of TLR3 **(B)** String analysis revealing molecular interaction of TLR7.

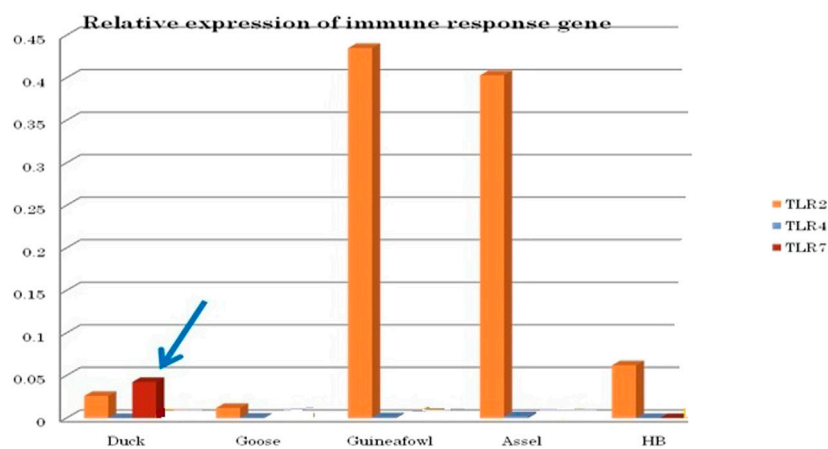


FIGURE 11 | Differential mRNA expression profile of immune response genes of healthy duck with other poultry species.

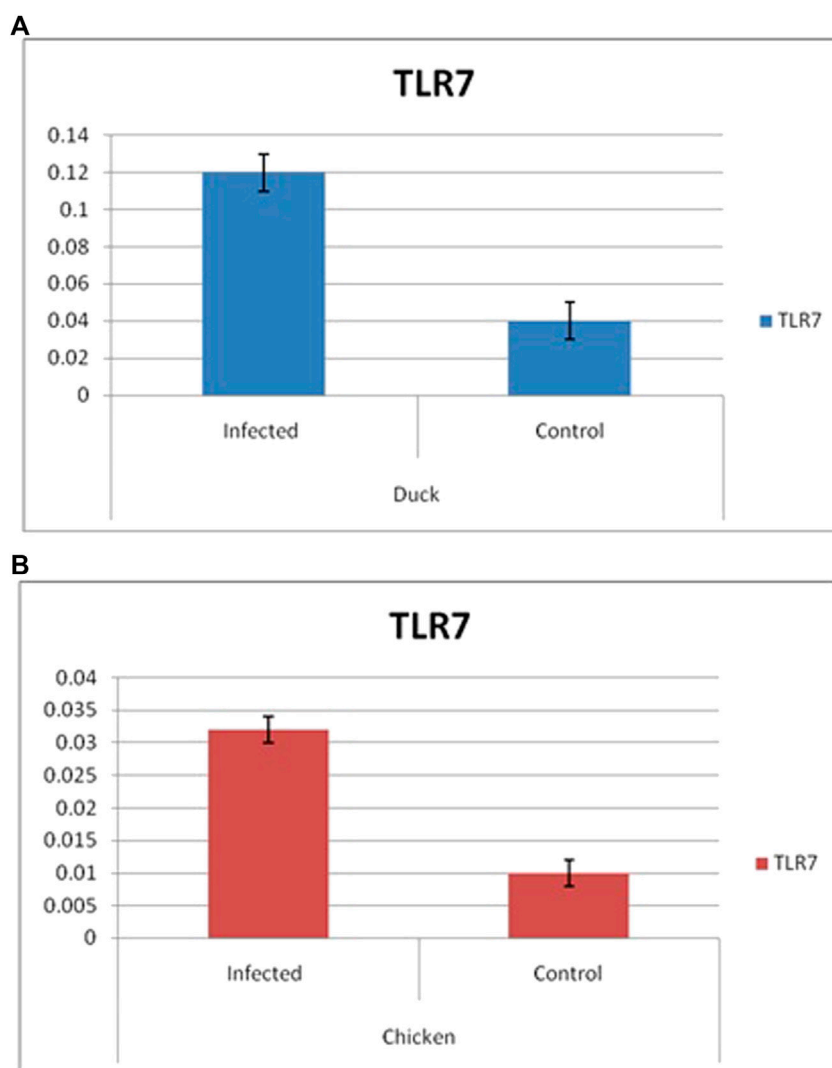


FIGURE 12 | Differential mRNA expression profiling for TLR7 gene with respect to infected versus control Bengal duck and HB Chicken.

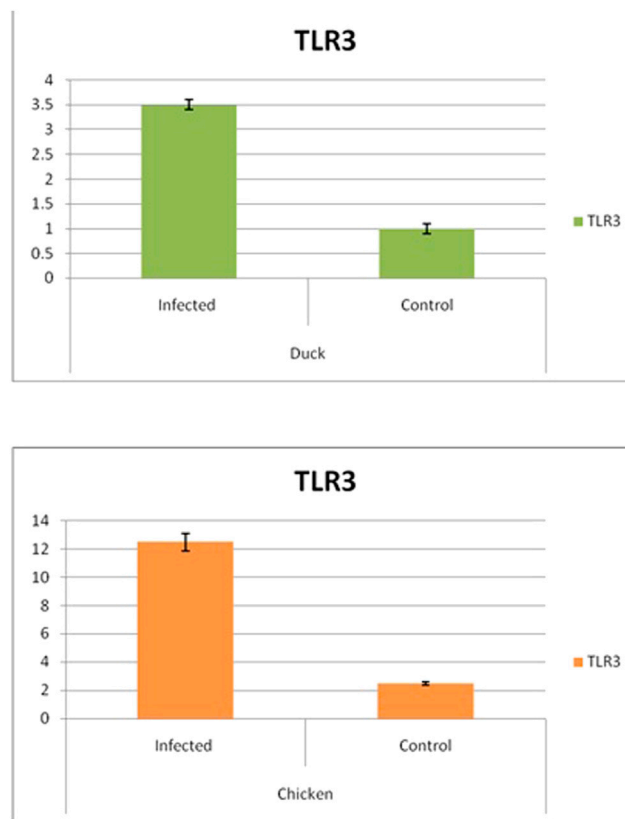


FIGURE 13 | Differential mRNA expression profiling for TLR3 with respect to infected versus control Bengal duck and HB Chicken.

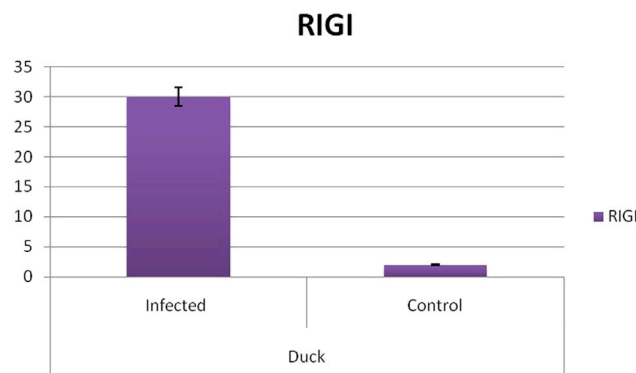


FIGURE 14 | Differential mRNA expression profiling for RIGI gene with respect to infected versus control Bengal duck and HB Chicken.

two distinct patches which are rich in positively charged amino acids and may be responsible for the binding of negatively charged viral dsRNA were observed.

RIG1 [also known as DEAD-box protein 58 (DDX58)] is an important molecule conferring antiviral immunity. Various important domains have been identified, such as CARD_RIG1 (caspase activation and recruitment domain found in RIG1), CARD2 interaction site, CARD1 interface, helicase insert domain, double-stranded RNA binding site, RIG-I-C (C

terminal domain of retinoic acid-inducible gene, RIG-I protein, and a cytoplasmic viral RNA receptor), RD interface, zinc-binding domain, and RNA binding. CARD proteins were observed to be responsible for the recognition of intracellular double-stranded RNA, a common constituent of a number of viral genomes. Unlike NLRs, these proteins, RIG-I contain twin N-terminal CARD domains and C-terminal RNA helicase domains that directly interact with and process the double-stranded viral RNA. CARD domains act through the

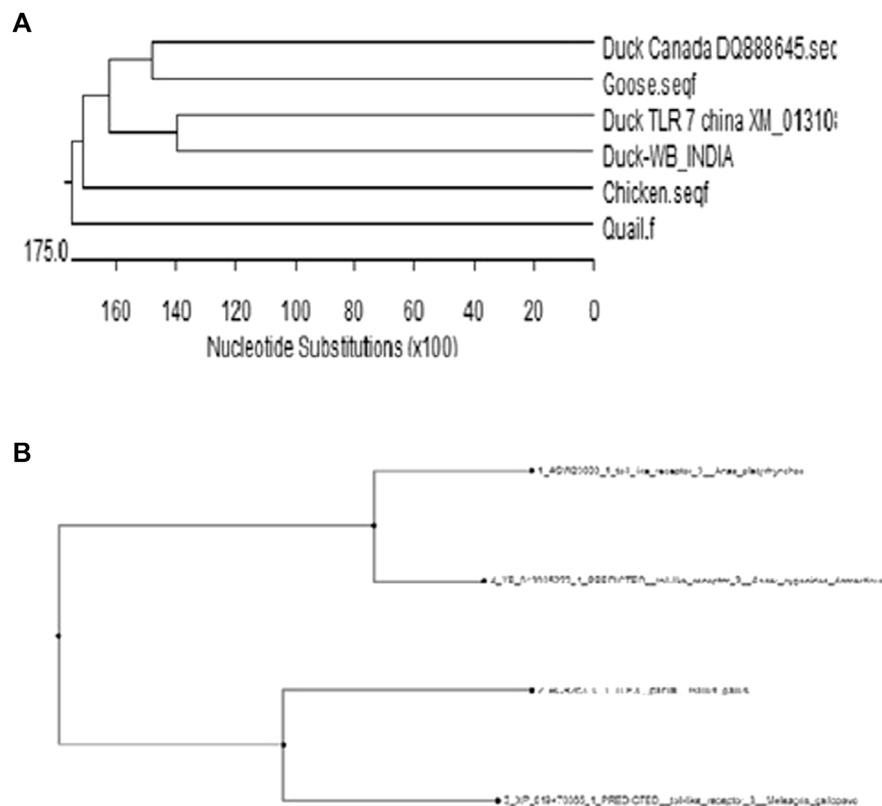


FIGURE 15 | (A) Molecular evolution of ducks reared globally in relation to other poultry species (Based on TLR7) **(B)** Molecular evolution of ducks reared globally (Based on TLR3).

interaction with the CARD motif (IPS-1/MAVS/VISA/Cardiff) which is a downstream adapter anchored in the mitochondria (Vercammen et al., 2008b; Botos et al., 2011).

Through the *in silico* alignment study, it was clearly observed that TLR3 binds with the H antigen of the avian influenza virus (H5N1). It was known that avian influenza virus (H5N1) strain is 100–200 nm spherical, enveloped, and includes 500 projecting spikes containing 80% hemagglutinin and 20% neuraminidase (Bouvier and Palese, 2008), with genome being segmented antisense ssRNA. In order to combat the infection, TLR3 binds with the hemagglutinin spikes of the influenza virus. TLR3 gene was observed to have a role to combat against Marek's disease (Said et al., 2018): seven amino acid polymorphism sites in ChTLR3 with 6 outer part sites and 1 inner part site (Barber et al., 2010). TLR3 cannot act alone. It acts while interacting with a series of molecules such as TICAM1, MAP3K7, TAB2, TRAF6, Myd88, IRAK4, IFIH1, and even TLR7. TLR3 acts through the RIG1-like receptor signaling pathway and acts through TRIF (Said et al., 2018). Along with *in silico* studies, we validated these findings with experimental challenge with avian influenza virus in embryonated fibroblast cell. The expression was observed to be more in infected egg than in healthy control for these genes in duck and chicken.

RIG1 is an important molecule which is only expressed in ducks, not in a chicken. Duck RIG1 transfected cells were

observed to recognize the RIG-I ligand, and a series of antiviral genes were expressed such as IFN- β , MX1, PKR, IFIT5, and OAS1, and consequently, HPAIV (highly pathogenic avian influenza virus) titers were reduced significantly (Haunshi and Cheng, 2014a; Ruan et al., 2015). RIG-I belongs to the IFN-stimulated gene family, and it acts through the RIG-I-like receptor signaling pathway. RIG1 detects dsRNA virus in the cytoplasm and initiates an antiviral response by producing type-I and Type-III IFN, through the activation of the downstream signaling cascade. RIG-I is an IFN-inducible viral sensor and is critical for amplifying the antiviral response (Potter et al., 2008; Barber et al., 2012). Although RIG1 expression is absent in chicken; it can produce INF α by another pathway. It has been observed that IFN- β expression upon influenza infection is mediated principally by RIG1 (Yoneyama et al., 2004). INF α expression induced by chicken is unable to protect the host from avian influenza infection as IFN- β , produced in ducks (Bouvier and Palese, 2008; Takahasi et al., 2008). This may be one of the major reasons why ducks are resistant to HPAIV, but not chicken. Since the RIG1 gene is not expressed in chicken, a comparative study was not possible. Avian influenza virus was observed to have surface glycoproteins as hemagglutinin and neuraminidase spikes on its outer surface (Loo et al., 2008). It was observed that duck RIG1 can bind with both H antigen and NA antigen of avian influenza virus. It is interesting to note how RIG1 acts on virus

and causes destruction. RIG-I acts through the RIG-I-like signaling pathway, and secretes NRLX1 and IPS1, leading to the production of IKK β , which in turn causes the secretion of NF κ B and I κ B (Koemer et al., 2007). These substances ultimately cause viral myocarditis and the destruction of the virus (Loo and Gale, 2011). A sequence of reactions occurs such as high fever, acute respiratory distress syndrome, chemoattraction of monocytes and macrophages, T-cell activation, and antibody response (Haunshi and Cheng, 2014b).

TLR7 is another important molecule responsible for antiviral immunity; it recognizes single-stranded RNA as the genetic material. It acts through the Toll-like receptor signaling pathway. However, from the current study, it was observed that TLR7 binds well with the NA antigen. The identified domains for TLR7 are mainly LRR (leucine-rich repeat), TIR, cysteine-rich flanking region, LRRNT (leucine-rich repeat N-terminal domain), and TPKR-C2 (tyrosine-protein kinase receptor C2 Ig-like domain). TLR7 releases Myd88, which in turn releases IRAK (Yamada et al., 2018). Ultimately, IRF7 is released, which causes viral myocarditis. It is interesting to note that in human and mice, TLR7 is alternatively spliced and expressed as two protein isoforms (Brisse and Ly, 2019). Another interesting observation was that chicken erythrocytes do not express TLR7 (Heil et al., 2003). While studying the TLR7 expression pattern in different avian species, an interesting observation in our current study was that TLR 7 gene expression was significantly better in duck than in other poultry species, such as indigenous chicken breeds (Aseel and Haringhata Black chicken), goose, and guineafowl. This is the first report for such a comparative study. It was reported that chicken TLR7 follow a restricted expression pattern. TLR7 expression was better in a macrophage cell line, chicken B-cell-like cell line, but the expression was observed to be lower in kidney cell line (Philbin et al., 2005). Following Marek's disease virus expression, TLR7 expression was observed to be increased in the lungs (St Paul et al., 2013).

Similarly, increased TLR7 expression was noted in IBDV (infectious bursal disease virus) (Iqbal et al., 2005). With regard to avian influenza infection, it was observed that at the early stage of low pathogenic avian influenza virus (LPAIV) infection of H11N9, in both duck and chicken, TLR7 is transiently expressed in peripheral blood mononuclear cells (PBMCs), while as infection progresses, the expression declines. Hence, it was observed that in chicken, TLR7 expression depended on the interaction between host and RNA virus (Abdul-Careem et al., 2009). Thus, differences in the expression pattern of TLR7 in chicken and duck were suggested (Abdul-Careem et al., 2009). Even in chicken, the TLR7 expression pattern was found to vary between HPAIV and LPAIV. Thus, TLR7 was observed to be an important immune response gene for avian influenza; TLR7 ligands show considerable potential for antivirals in chicken (Abdul-Careem et al., 2009). Although no direct report was available for better TLR7 expression in a duck than in a chicken, it was reported that tissue tropism and immune function of duck TLR7 are different from those of chicken TLR7, which result in a difference in susceptibility between chicken and duck, when infected by the same pathogen (Abdul-Careem et al., 2009). A high expression

pattern of duck TLR7 in respiratory and lymphoid tissue was observed to be different from that of chicken.

TLR3, TLR7, and TLR21 localize mainly in the ER in the steady-state and traffic to the endosome, where they engage with their ligands. The recognition triggers the downstream signal transduction to activate NF- κ B or IRF3/7, finally induces interferon and inflammatory cytokine production (Abdul-Careem et al., 2009). We can explore these identified and characterized genes for production of transgenic or gene-edited chicken resistant to avian influenza as a future control strategy against avian influenza through immunomodulation, devoid of side effects as in case of use of drugs (Rauf et al., 2011). It is to be noted that since the control of avian influenza virus has been difficult and challenging either through vaccination (Chen et al., 2013; Rios et al., 2017; Rx List, 2019) or treatment through antiviral drugs (FDA, 2020; Science News, 2016; Principi et al., 2019; WebMD, 2020) due to frequent mutation and genetic reassortment (regarded as antigenic shift or antigenic drift) of the single stranded RNA genome which is prone to mutations (NHS, 2014; Zhang et al., 2019; Richard et al., 2002a). An interesting observation revealed that unlike antibodies (comprising of immunoglobulins) which were highly specific, arising due to variability of Fab site and variable region (Nguyen, 2020), immune response molecules for innate immunity can bind Avian influenza virus (H, N antigen), irrespective of strains. As we analyze the binding sites, some important domains were identified, which may be involved in antiviral activity. This led to the finding that the therapeutic approach may be attempted with the recombinant product corresponding to the identified domain. Gene editing with gene insert from identified gene may lead to the evolution of disease-resistant strains/lines of chicken or duck.

Although the receptors for human influenza and avian influenza are different, mutations may overcome the barrier. As a case report in 2018, a human infection with a novel H7N4 avian influenza virus was reported in Jiangsu, China. Circulating avian H9N2 viruses were reported to be the origin of the H7N4 internal segments, unlike both the human H5N1 and H7N9 viruses that had H9N2 backbones. The major concern is that genetic reassortment and adaptive mutation of avian influenza virus give rise to human influenza virus strain H7N4 (Jiao et al., 2012; Ayora-Talavera, 2018; Karen, 2018; Li et al., 2020). The WHO has also warned about the pandemic on human flu resulting from genetic reassortment of avian influenza (Qu et al., 2020). We observed in this study that the H5N1 strain of avian influenza, a highly pathogenic strain, was genetically closer to H6N2. Recent reports revealed that H6N2 is continuously evolving in different countries such as South Africa (Quan et al., 2019), Egypt (WHO, 2018a), India (Worldometer, 2020), and North America (Zanaty et al., 2019) due to genetic reassortment. It is gradually evolving from the low pathogenic form to the high pathogenic form and observed to overcome species barrier with interspecies genetic assortment (Kumar et al., 2018; Richard et al., 2002b; Gillim-Ross et al., 2008; FAO, 2019) and every possibility to evolve as a pandemic for human. Reports are available depicting human influenza virus arising due to genetic reassortment of avian influenza in China (Jiao et al., 2012;

Ayora-Talavera, 2018; Karen, 2018; Li et al., 2020). These findings highlight the growing importance of the study in the current era, when the world is suffering from a pandemic.

Although molecular docking analysis is available for the identification of various drug molecules with avian influenza virus (Amir et al., 2011; Liu et al., 2015; Khan et al., 2017; Pal and Chakravarty, 2019b), this is the first report of molecular docking analysis with the immune response molecules responsible for antiviral immunity against avian influenza and is the basis for finding drug for a disease. A series of immune response molecules are responsible for providing antiviral immunity with their respective interaction in various pathways as we depicted through String and KEGG pathway analyses in our current study.

The future outcome for the current study is the possible utilization of the identified disease resistant genes (RIGI, TLR3, and TLR7) for the development of avian influenza-resistant chicken with the identified gene insert from duck through gene editing or a transgenic approach.

CONCLUSION

RIG1 detects the virus that is present within the cytosol of infected cells (cell intrinsic recognition), whereas TLR3 detects virus-infected cells, and TLR7 detects viral RNA that has taken up into the endosomes of sentinel cells (cell-extrinsic recognition). TLR7 may be regarded as the promising gene for antiviral immunity with pronounced expression profiling in duck in contrast to other poultry birds. Molecular docking revealed RIGI, TLR3, and TLR7 as the promising genes conferring antiviral immunity against avian influenza. Point mutations have been detected in chicken TLR3 with respect to that of duck indicative of reduced antiviral immunity in chicken in comparison to duck utilization of the identified disease-resistant genes (RIGI, TLR3, and TLR7) for the development of avian influenza-resistant chicken with the identified gene insert from duck.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and

accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Ethics Committee, West Bengal University of Animal and Fishery Sciences.

AUTHOR CONTRIBUTIONS

ARP has designed the research work, conducted the research work, analyzed data, and written the manuscript. ABP has conducted the bioinformatic analysis. PB has analyzed and revised the article.

FUNDING

The authors are thankful to the Department of Biotechnology, Ministry of Science and Technology, Govt. of India (Grant number BT/PR24310/NER/95/649/2017) and SERB, the Department of Science and Technology, Govt. of India (Grant no. EMR/2016/003554) for providing financial support.

ACKNOWLEDGMENTS

The technical and financial support by vice-chancellor, West Bengal University of Animal and Fishery Sciences, is duly acknowledged. The authors thank the director, AH & VS, Animal Resource Development Department, Govt. of West Bengal.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.633283/full#supplementary-material>

REFERENCES

- Abdul-Careem, M. F., Haq, K., Shanmuganathan, S., Read, L. R., Schat, K. A., Heidari, M., et al. (2009). Induction of Innate Host Responses in the Lungs of Chickens Following Infection with a Very Virulent Strain of Marek's Disease Virus. *Virology* 393, 250–257. doi:10.1016/j.virol.2009.08.001
- Ali, S., Mann-Nüttel, R., Schulze, A., Richter, L., Alferink, J., and Scheu, S. (2019). Sources of Type I Interferons in Infectious Immunity: Plasmacytoid Dendritic Cells Not Always in the Driver's Seat. *Front. Immunol.* 10, 778. doi:10.3389/fimmu.2019.00778
- Amir, A., Siddiqui, M. A., Kapoor, N., Arya, A., and Kumar, H. (2011). In Silico Molecular Docking of Influenza Virus (PB2) Protein to Check the Drug Efficacy. *Trends Bioinformatics* 4, 47–55. doi:10.3923/tb.2011.47.5510.3923/tb.2011.47.55

- Ayora-Talavera, G. (2018). Sialic Acid Receptors: Focus on Their Role in Influenza Infection. *Jrlcr* Vol. 10, 1–11. doi:10.2147/JRLCR.S140624
- Barber, M. R., Aldridge, J. R., Fleming-Canepa, X., Wang, Y. D., Webster, R. G., and Magor, K. E. (2012). Identification of Avian RIG-I Responsive Genes during Influenza Infection. *Mol. Immunol.* 54, 89–97. doi:10.1016/j.molimm.2012.10.038
- Barber, M. R. W., Aldridge, J. R., Webster, R. G., and Magor, K. E. (2010). Association of RIG-I with Innate Immunity of Ducks to Influenza. *Proc. Natl. Acad. Sci.* 107, 5913–5918. doi:10.1073/pnas.1001755107
- Basler, C. F., and Aguilar, P. V. (2008). Progress in Identifying Virulence Determinants of the 1918 H1N1 and the Southeast Asian H5N1 Influenza A Viruses. *Antiviral Res.* 79, 166–178. doi:10.1016/j.antiviral.2008.04.006
- Basler, C. F., and Aguilar, P. V. (2008). Progress in Identifying Virulence Determinants of the 1918 H1N1 and the Southeast Asian H5N1 Influenza A Viruses. *Antiviral Res.* 79, 166–178. doi:10.1016/j.antiviral.2008.04.006

- Botos, I., Segal, D. M., and Davies, D. R. (2011). The Structural Biology of Toll-like Receptors. *Structure* 19 (4), 447–459. doi:10.1016/j.str.2011.02.004
- Bouvier, N. M., and Palese, P. (2008). The Biology of Influenza Viruses. *Vaccine* 26 (Suppl. 4Suppl 4), D49–D53. doi:10.1016/j.vaccine.2008.07.039
- Brise, M., and Ly, H. (2019/2019). Comparative Structure and Function Analysis of the RIG-I-like Receptors: RIG-I and MDA5. *Front. Immunol.* 10. doi:10.3389/fimmu.2019.01586
- Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Res.* 33, W306–W310. Jul 1 (Web Server issue): W306–W310. doi:10.1093/nar/gki375
- CDC, Centre for Disease control and Prevention. 2021. Avian Influenza in Birds. Available at: <https://www.cdc.gov/flu/avianflu/avian-in-birds.htm>
- Chen, S., Cheng, A., and Wang, M. (2013). Innate Sensing of Viruses by Pattern Recognition Receptors in Birds. *Vet. Res.* 44, 82. doi:10.1186/1297-9716-44-82
- Choi, Y., and Chan, A. P. (2015). PROVEAN Web Server: a Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics* 31 (16), 2745–2747. doi:10.1093/bioinformatics/btv195
- Cunningham, M. D., Shapiro, R. A., Seachord, C., Ratcliffe, K., Cassiano, L., and Darveau, R. P. (2000). CD14 Employs Hydrophilic Regions to "Capture" Lipopolysaccharides. *J. Immunol.* 164, 3255–3263. doi:10.4049/jimmunol.164.6.3255
- Ebina, T., Toh, H., and Kuroda, Y. (2009). Loop-length-dependent SVM Prediction of Domain Linkers for High-Throughput Structural Proteomics. *Biopolymers* 92 (1), 1–8. doi:10.1002/bip.21105
- FAO (2019). Chinese-origin H7N9 Avian Influenza Spread in Poultry and Human Exposure. Available at: <http://www.fao.org/3/CA3206EN/ca3206en.pdf>
- FDA (2020). H5N1 Influenza Virus Vaccine, Manufactured by Sanofi Pasteur, Inc. Questions and Answers. Available at: <https://www.fda.gov/vaccines-blood-biologics/vaccines/h5n1-influenza-virus-vaccine-manufactured-sanofi-pasteur-inc-questions-and-answers>
- Fleming-Canepa, X., Aldridge, J. R., Jr, Canniff, L., Kobewka, M., Jax, E., Webster, R. G., et al. (2019). Duck Innate Immune Responses to High and Low Pathogenicity H5 Avian Influenza Viruses. *Vet. Microbiol.* 228, 101–111. Jan. doi:10.1016/j.vetmic.2018.11.018
- Gillim-Ross, L., Santos, C., Chen, Z., Aspelund, A., Yang, C.-F., Ye, D., et al. (2008). Avian Influenza H6 Viruses Productively Infect and Cause Illness in Mice and Ferrets. *J. Virol.* 82, 10854–10863. doi:10.1128/JVI.01206-08
- Glick, D. M. (1977). *Glossary of Biochemistry and Molecular Biology*. Revised edition. London, UK: Portland Press.
- Hale, B. G., Randall, R. E., Ortín, J., and Jackson, D. (2008). The Multifunctional NS1 Protein of Influenza A Viruses. *J. Gen. Virol.* 89, 2359–2376. doi:10.1099/vir.0.2008/004606-0
- Haunshi, S., and Cheng, H. H. (2014). Differential Expression of Toll-like Receptor Pathway Genes in Chicken Embryo Fibroblasts from Chickens Resistant and Susceptible to Marek's Disease. *Poult. Sci.* 93 (3), 550–555. doi:10.3382/ps.2013-03597
- Haunshi, S., and Cheng, H. H. (2014). Differential Expression of Toll-like Receptor Pathway Genes in Chicken Embryo Fibroblasts from Chickens Resistant and Susceptible to Marek's Disease. *Poult. Sci.* 93 (3), 550–555. doi:10.3382/ps.2013-03597
- Heil, F., Ahmad-Nejad, P., Hemmi, H., Hochrein, H., Ampenberger, F., Gellert, T., et al. (2003). The Toll-like Receptor 7 (TLR7)-specific Stimulus Loxoribine Uncovers a strong Relationship within the TLR7, 8 and 9 Subfamily. *Eur. J. Immunol.* 33, 2987–2997. doi:10.1002/eji.200324238
- Iqbal, M., Philbin, V. J., and Smith, A. L. (2005). Expression Patterns of Chicken Toll-like Receptor mRNA in Tissues, Immune Cell Subsets and Cell Lines. *Vet. Immunol. Immunopathology* 104, 117–127. doi:10.1016/j.vetimm.2004.11.003
- Jiao, P. R., Wei, L. M., Cheng, Y. Q., Yuan, R. Y., Han, F., Liang, J., et al. (2012). Molecular Cloning, Characterization, and Expression Analysis of the Muscovy Duck Toll-like Receptor 3 (MdTLR3) Gene. *Poult. Sci.* 91 (10), 2475–2481. doi:10.3382/ps.2012-02394
- Kannaki, T. R., Reddy, M. R., Shanmugam, M., Verma, P. C., Sharma, R. P., et al. (2010). Chicken Toll-like Receptors and Their Role in Immunity. *World's Poult. Sci. J.* 66 (04), 727–738. doi:10.1017/s0043933910000693
- Karen, S. (2018) Antigenic Drift vs Antigenic Shift. Available at: <https://www.technologynetworks.com/immunology/articles/antigenic-drift-vs-antigenic-shift-311044>
- Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30 (4), 772–780. doi:10.1093/molbev/mst010
- Kell, A. M., and Gale, M. (2015). RIG-I in RNA Virus Recognition. *Virology* 479–480, 110–121. doi:10.1016/j.virol.2015.02.017
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 Web portal for Protein Modeling, Prediction and Analysis. *Nat. Protoc.* 10, 845–858. doi:10.1038/nprot.2015.053
- Kelley, L. A., and Sternberg, M. J. E. (2009). Protein Structure Prediction on the Web: a Case Study Using the Phyre Server. *Nat. Protoc.* 4 (3), 363–371. doi:10.1038/nprot.2009.2
- Khan, J., Masood, A., Noor, A., Munir, A., and Qadir, M. I. (2017). Molecular Docking Studies on Possible Neuraminidase Inhibitors of Influenza Virus. *Ann. Antivir. Antiretrovir* 1 (1), 005–007. doi:10.17352/aaa.000002
- Kim, J.-K., Negovetich, N. J., Forrest, H. L., and Webster, R. G. (2009). Ducks: The "Trojan Horses" of H5N1 Influenza. *Influenza Other Respir. Viruses* 3 (4), 121–128. Jul. doi:10.1111/j.1750-2659.2009.00084.x
- Koerner, I., Kochs, G., Kalinke, U., Weiss, S., Staeheli, P., et al. (2007). Protective Role of Beta Interferon in Host Defense against Influenza A Virus. *J. Virol.* 81, 2025–2030. doi:10.1128/jvi.01718-06
- Kumar, M., Nagarajan, S., Murugkar, H. V., Saikia, B., Singh, B., Mishra, A., et al. (2018). Emergence of Novel Reassortant H6N2 Avian Influenza Viruses in Ducks in India. *Infect. Genet. Evol.* 61, 20–23. doi:10.1016/j.meegid.2018.03.005
- Li, X., Sun, J., Lv, X., Wang, Y., Li, Y., Li, M., et al. (2020). Novel Reassortant Avian Influenza A(H9N2) Virus Isolate in Migratory Waterfowl in Hubei Province, China. *Front. Microbiol.*, 11, , 2020 13 February2020. doi:10.3389/fmicb.2020.00220
- Liu, Z., Zhao, J., Li, W., Wang, X., Xu, J., Xie, J., et al. (2015). Molecular Docking of Potential Inhibitors for Influenza H7N9. *Comput. Math. Methods Med.* 2015, 480764, 2015 . Article ID 480764. doi:10.1155/2015/48076410.1155/2015/480764
- Loo, Y.-M., Fornek, J., Crochet, N., Bajwa, G., Perwitasari, O., Martinez-Sobrido, L., et al. (2008). Distinct RIG-I and MDA5 Signaling by RNA Viruses in Innate Immunity. *J. Virol.* 82, 335–345. doi:10.1128/jvi.01080-07
- Loo, Y.-M., and Gale, M., Jr. (2011). Immune Signaling by RIG-I-like Receptors. *Immunity* 34 (5), 680–692. doi:10.1016/j.immuni.2011.05.003
- Maarouf, M., Rai, K., Goraya, M., and Chen, J.-L. (2018). Immune Ecosystem of Virus-Infected Host Tissues. *Ijms* 19, 1379. doi:10.3390/ijms19051379
- Mangani, M., and McGavern, D. B. (2018). New Advances in CNS Immunity against Viral Infection. *Curr. Opin. Virol.* 28, 116–126. doi:10.1016/j.coviro.2017.12.003
- Mashiach, E., Schneidman-Duhovny, D., Andrusier, N., Nussinov, R., and Wolfson, H. J. (2008). FireDock: a Web Server for Fast Interaction Refinement in Molecular Docking. *Nucleic Acids Res.* 36, W229–W232. Web Server issue). doi:10.1093/nar/gkn186
- Matrosovich, M. N., Stech, J., and Klenk, H.-D. (2009). Influenza Receptors, Polymerase and Host Range. *Rev. Sci. Tech. OIE* 28, 203–217. doi:10.20506/rst.28.1.1870
- Matrosovich, M. N., Stech, J., and Klenk, H.-D. (2009). Influenza Receptors, Polymerase and Host Range. *Rev. Sci. Tech. OIE* 28, 203–217. doi:10.20506/rst.28.1.1870
- Medina, R. A., and García-Sastre, A. (2011). Influenza A Viruses: New Research Developments. *Nat. Rev. Microbiol.* 9, 590–603. doi:10.1038/nrmicro2613
- Morgan, B., Ly, H., et al. (2019). Comparative Structure and Function Analysis of the RIG-I-like Receptors: RIG-I and MDA5. *Front. Immunol.* doi:10.3389/fimmu.01586
- Muroi, M., Ohnishi, T., and Tanamoto, K.-i. (2002). Regions of the Mouse CD14 Molecule Required for Toll-like Receptor 2- and 4-mediated Activation of NF- κ B. *J. Biol. Chem.* 277 (44), 42372–42379. doi:10.1074/jbc.m205966200
- Nguyen, H. H. (2020). What Is the Role of Antigenic Shift in the Pathogenesis of Influenza. Available at: <https://www.medscape.com/answers/219557-3453/what-is-the-role-of-antigenic-shift-in-the-pathogenesis-of-influenza>
- NHS. Effectiveness of Tamiflu and Relenza Questioned. Available at: <https://www.nhs.uk/news/medication/effectiveness-of-tamiflu-and-relenza-questioned> (2014).
- Niu, B., Lu, Y., Wang, J., Hu, Y., Chen, J., Chen, Q., et al. (2019). 2D-SAR, Topomer CoMFA and Molecular Docking Studies on Avian Influenza Neuraminidase Inhibitors. *Comput. Struct. Biotechnol. J.* 17, 39–48. doi:10.1016/j.csbj.2018.11.007

- Otte, J., Hinrichs, J., Rushton, J., Roland-Holst, D., and Zilberman, D. (2008). Impacts of Avian Influenza Virus on Animal Production in Developing Countries. *CAB Rev.* 3, 080. doi:10.1079/PAVSNNR20083080
- Otte, J., Hinrichs, J., Rushton, J., Roland-Holst, D., and Zilberman, D. (2008). Impacts of Avian Influenza Virus on Animal Production in Developing Countries. *CAB Rev.* 3, 080, 2008. doi:10.1079/PAVSNNR20083080
- Pal, A., Chatterjee, P. N., and Sharma, A. (2013). A Molecular Evolution and Structural Analysis of Caprine CD14 Deduced from cDNA Clones. *Indian J. Anim. Sci.* 83 (10), 1062–1067.
- Pal, A., Sharma, A., Bhattacharya, T. K., Mitra, A., and Chatterjee, P. N. (2014). Sequence Characterization and Polymorphism Detection in Bubaline CD14 Gene. *Buffalo Bull.* 32 (2), 138–156.
- Pal, A., Pal, A., Banerjee, S., Batabyal, S., and Chatterjee, P. N. (2018). Mutation in Cytochrome B Gene Causes Debility and Adverse Effects on Health of Sheep. *Mitochondrion* 46, 393–404. doi:10.1016/j.mito.2018.10.003.46
- Pal, A., Pal, A., Banerjee, S., Batabyal, S., and Chatterjee, P. N. (2018). Mutation in Cytochrome B Gene Causes Debility and Adverse Effects on Health of Sheep. *Mitochondrion* 46, 393–404. doi:10.1016/j.mito.2018.10.003.46
- Pal, A., Pal, A., Mallick, A. I., Biswas, P., and Chatterjee, P. N. (2019). Molecular Characterization of Bu-1 and TLR2 Gene in Haringhata Black Chicken. *Genomics* 112 (1), 472–483. doi:10.1016/j.ygeno.2019.03.010
- Pal, A., Abantika, P., Baviskar, P., et al. (2017). “Molecular Characterization by Next Generation Sequencing and Study of the Genetic Basis of Antiviral Resistance of Indigenous Ducks,” in the National symposium on National Symposium on Biodynamic Animal Farming for the Management of Livestock Diversity under changing climatic scenario and 14 th Annual Convention of SOCADAB, Mannuthy, Feb 8-10(2017) (CVAS).
- Pal, A., and Chatterjee, P. N. (2009). Molecular Cloning and Characterization of CD14 Gene in Goat. *Small Ruminant Res.* 82, 84–87. doi:10.1016/j.smallrumres.2008.11.016
- Pal, A., and Chakravarty, A. K. (2019). *Genetics and Breeding for Disease Resistance*. Paperback: Academic Press. ISBN: 9780128164068 eBook ISBN: 9780128172674.
- Pal, A., Pal, A., Sharma, A., Bhattacharya, T. K., et al. (2020). Mutations in CD14 Gene Causes Mastitis in Different Breeds of buffalo as Confirmed by In Silico Studies and Experimental Validation. *BMC Genet.* Under review. doi:10.21203/rs.2.10779/v1
- Pal, A., Sharma, A., Bhattacharya, T. K., Chatterjee, P. N., and Chakravarty, A. K. (2011). Molecular Characterization and SNP Detection of CD14 Gene of Crossbred Cattle. *Mol. Biol. Int.* 2011, 1–13. Article ID 507346, 13 pages. doi:10.4061/2011/507346
- Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A Generic Method for Assignment of Reliability Scores Applied to Solvent Accessibility Predictions. *BMC Struct. Biol.* 9, 51. doi:10.1186/1472-6807-9-51
- Philbin, V. J., Iqbal, M., Boyd, Y., Goodchild, M. J., Beal, R. K., Bumstead, N., et al. (2005). Identification and Characterization of a Functional, Alternatively Spliced Toll-like Receptor 7 (TLR7) and Genomic Disruption of TLR8 in Chickens. *Immunology* 114, 507–521. doi:10.1111/j.1365-2567.2005.02125.x
- Potter, J. A., Randall, R. E., and Taylor, G. L. (2008). Crystal Structure of Human IPS-1/MAVS/VISA/Cardif Caspase Activation Recruitment Domain. *BMC Struct. Biol.* 8, 11. doi:10.1186/1472-6807-8-11
- Principi, N., Camilloni, B., Alunno, A., Polinori, I., Argentiero, A., and Esposito, S. (2019). Drugs for Influenza Treatment: Is There Significant News? *Front. Med. (Lausanne)*. doi:10.3389/fmed.2019.00109
- Qu, B., Li, X., Cardona, C. J., and Xing, Z. (2020). Reassortment and Adaptive Mutations of an Emerging Avian Influenza Virus H7N4 Subtype in China. *PLoS ONE* 15 (1), e0227597. doi:10.1371/journal.pone.0227597
- Quan, C., Wang, Q., Zhang, J., Zhao, M., Dai, Q., Huang, T., et al. (2019). Avian Influenza A Viruses Among Occupationally Exposed Populations, China, 2014–2016. *Emerg. Infect. Dis.* 25 (12), 2215–2225. doi:10.3201/eid2512.190261
- Rauf, A., Khatri, M., Murgia, M. V., Jung, K., and Saif, Y. M. (2011). Differential Modulation of Cytokine, Chemokine and Toll like Receptor Expression in Chickens Infected with Classical and Variant Infectious Bursal Disease Virus. *Vet. Res.* 42 (1), 85. doi:10.1186/1297-9716-42-85
- Richard, J., Peter, W., Woolcock, R., Scott, L., Robert, K., Webster, G., et al. (2002). Reassortment and Interspecies Transmission of North American H6N2 Influenza Viruses. *Virology* 295, 44–53. doi:10.1006/viro.2001.134
- Richard, J., Peter, W., Woolcock, R., Scott, L., Robert, K., Webster, G., et al. (2002). Reassortment and Interspecies Transmission of North American H6N2 Influenza Viruses. *Virology* 295, 44–53. doi:10.1006/viro.2001.134
- Rios, M. C., Fraga, L. T. S., Nascimento, T. V. S. B. d., Antonioli, A. R., Lyra Júnior, D. P. d., França, A., et al. (2017). Interferon and Ribavirin for Chronic Hepatitis C: Should it Be Administered in the New Treatment Era? *Acta Gastroenterol. Latinoam.* 47, 14–22.
- Ruan, W., An, J., and Wu, Y. (2015). Polymorphisms of Chicken TLR3 and 7 in Different Breeds. *PLoS ONE* 10 (3), e0119967. doi:10.1371/journal.pone.0119967
- Rx List (2019). Tamiflu. Available at: <https://www.rxlist.com/tamiflu-side-effects-drug-center.htm>
- Said, E. A., Tremblay, N., Al-Balushi, M. S., Al-Jabri, A. A., and Lamarre, D. (2018). Viruses Seen by Our Cells: The Role of Viral RNA Sensors. *J. Immunol. Res.* 2018, 1–14. doi:10.1155/2018/9480497
- Šali, A., Sánchez, R., et al. (2020). MODELLER A Program for Protein Structure Modeling. *Release* 9.24, r11614, 2020. 10.1385/1-59259-368-2:97.email: modeller-care AT salilab.org. URL <https://salilab.org/modeller>.
- Schlee, M., Roth, A., Hornung, V., Hagmann, C. A., Wimmenauer, V., Barchet, W., et al. (2009). Recognition of 5' Triphosphate by RIG-I Helicase Requires Short Blunt Double-Stranded RNA as Contained in Panhandle of Negative-Strand Virus. *Immunity* 31 (1), 25–34. Epub 2009 Jul 2. doi:10.1016/j.immuni.2009.05.008
- Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: Servers for Rigid and Symmetric Docking. *Nucleic Acids Res.* 33, W363–W367. 1Web Server issue: W363–W367. doi:10.1093/nar/gki481
- Science News (2016). New NDV-H5nx Avian Influenza Vaccine Has Potential for Mass Vaccination of Poultry. Available at: <https://www.sciencedaily.com/releases/2016/01/160107131015.htm>.
- St Paul, M., Paolucci, S., Barjesteh, N., Wood, R. D., and Sharif, S. (2013). Chicken Erythrocytes Respond to Toll-like Receptor Ligands by Up-Regulating Cytokine Transcripts. *Res. Vet. Sci.* 95, 87–91. doi:10.1016/j.rvsc.2013.01.024
- Stetson, D. B., and Medzhitov, R. (2006). Type I Interferons in Host Defense. *Immunity*, 25, 373–381. Immunity25Elsevier Inc–381. doi:10.1016/j.immuni.2006.08.007
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (20152015). STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* 43 (Database issue), D447–D452. doi:10.1093/nar/gku1003
- Takahasi, K., Yoneyama, M., Nishihori, T., Hirai, R., Kumeta, H., Narita, R., et al. (2008). Nonself RNA-Sensing Mechanism of RIG-I Helicase and Activation of Antiviral Immune Responses. *Mol. Cell* 29, 428–440. doi:10.1016/j.molcel.2007.11.028
- Tong, S., Zhu, X., Li, Y., Shi, M., Zhang, J., Bourgeois, M., et al. (2013). New World Bats Harbor Diverse Influenza A Viruses. *Plos Pathog.* 9 (10), e1003657–84. doi:10.1371/journal.ppat.1003657
- Vercammen, E., Staal, J., and Beyaert, R. (2008). Sensing of Viral Infection and Activation of Innate Immunity by Toll-like Receptor 3. *Clin. Microbiol. Rev.* 21 (1), 13–25. doi:10.1128/CMR.00022-07
- Vercammen, E., Staal, J., and Beyaert, R. (2008). Sensing of Viral Infection and Activation of Innate Immunity by Toll-like Receptor 3. *Clin. Microbiol. Rev.* 21 (1), 13–25. doi:10.1128/CMR.00022-07
- WebMD.Flu Treatment with Antiviral Drugs. Available at: [https://www.webmd.com/cold-and-flu/flu-medications#1\(2020\)](https://www.webmd.com/cold-and-flu/flu-medications#1(2020)).
- Webster, R. G., and Govorkova, E. A. (2014). Continuing Challenges in Influenza. *Ann. N.Y. Acad. Sci.* 1323 (1), 115–139. doi:10.1111/nyas.12462
- WHO. How Pandemic Influenza Emerges. Available at: [http://www.euro.who.int/en/health-topics/communicable-diseases/influenza/pandemic-influenza/how-pandemic-influenza-emerges\(2020\)](http://www.euro.who.int/en/health-topics/communicable-diseases/influenza/pandemic-influenza/how-pandemic-influenza-emerges(2020)).
- WHO (2018a). Human Infection with Avian Influenza A(H7N4) Virus – China. Available at: <https://www.who.int/csr/don/22-february-2018-ah7n4-china/en>.
- WHO (2018b). Influenza (Avian and Other Zoonotic). Available at: <https://www.who.int/biologicals/vaccines/influenza/en>.
- Wiederstein, M., and Sippl, M. J. (2007). ProSA-web: Interactive Web Service for the Recognition of Errors in Three-Dimensional Structures of Proteins. *Nucleic Acids Res.* 35, W407–W410. Issue suppl_2–W410. doi:10.1093/nar/gkm290

- Worldometer (2020). Coronavirus Death Toll and Trends. Available at: <https://www.worldometers.info/coronavirus/coronavirus-death-toll>.
- Yamada, S., Shimojima, M., Narita, R., Tsukamoto, Y., Kato, H., Saijo, M., et al. (2018). RIG-I-Like Receptor and Toll-like Receptor Signaling Pathways Cause Aberrant Production of Inflammatory Cytokines/Chemokines in a Severe Fever with Thrombocytopenia Syndrome Virus Infection Mouse Model. *J. Virol.* 92 (13), e02246–17. doi:10.1128/JVI.02246-17
- Ying, W., Yan, W., Boris, T., Yi, S., George, F. G., et al. (2014). Bat-derived Influenza-like Viruses H17N10 and H18N11. *Trends Microbiol.* 22, 183–191. doi:10.1016/j.tim.2014.01.010
- Yoneyama, M., Kikuchi, M., Natsukawa, T., Shinobu, N., Imaizumi, T., Miyagishi, M., et al. (2004). The RNA Helicase RIG-I Has an Essential Function in Double-Stranded RNA-Induced Innate Antiviral Responses. *Nat. Immunol.* 5, 730–737. doi:10.1038/ni1087
- Zanaty, A. M., Erfan, A. M., Mady, W. H., Amer, F., Nour, A. A., Rabie, N., et al. (2019). Avian Influenza Virus Surveillance in Migratory Birds in Egypt Revealed a Novel Reassortant H6N2 Subtype. *Avian Res.* 10, 41. doi:10.1186/s40657-019-0180-7
- Zhang, Y., Dong, J., Bo, H., Dong, L., Zou, S., Li, X., et al. (2019). Genetic and Biological Characteristics of Avian Influenza Virus Subtype H1N8 in Environments Related to Live Poultry Markets in China. *BMC Infect. Dis.* 19, 458. doi:10.1186/s12879-019-4079-z
- Zhang, Y., and Skolnick, J. (2005). TM-align: a Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* 33, 2302–2309. doi:10.1093/nar/gki524

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

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

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

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership