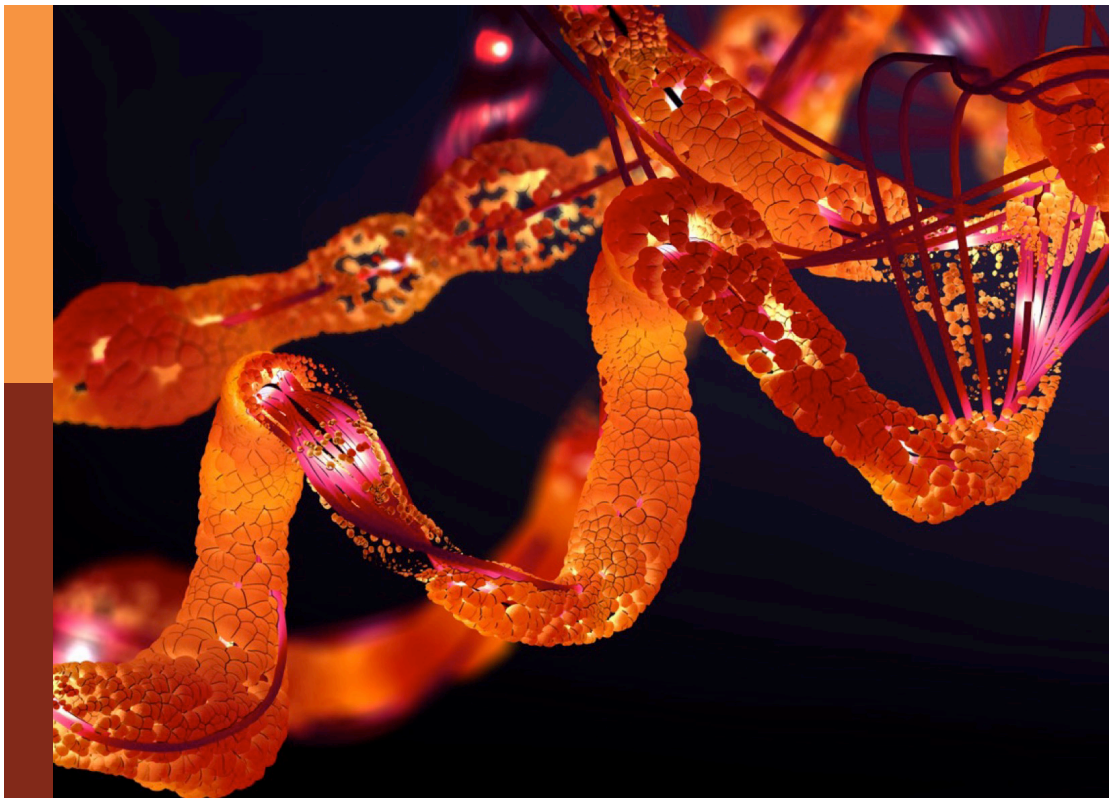
frontiers | Research Topics

Mechanisms, thermodynamics and kinetics of ligand binding revealed from molecular simulations and machine learning

Edited by
Yinglong Miao, Weiliang Zhu, Chia-en A. Chang and J. Andrew McCammon

Published in
Frontiers in Molecular Biosciences



ISSN 1664-8714
ISBN 978-2-83251-512-9
DOI 10.3389/978-2-83251-512-9

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

Mechanisms, thermodynamics and kinetics of ligand binding revealed from molecular simulations and machine learning

Topic editors

Yinglong Miao — University of Kansas, United States

Weiliang Zhu — Shanghai Institute of Materia Medica, Chinese Academy of Sciences (CAS), China

Chia-en A. Chang — University of California, Riverside, United States

J. Andrew McCammon — University of California, San Diego, United States

Citation

Miao, Y., Zhu, W., Chang, C.-e. A., McCammon, J. A., eds. (2023). *Mechanisms, thermodynamics and kinetics of ligand binding revealed from molecular simulations and machine learning*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-83251-512-9

Table of contents

frontiers | Frontiers in Molecular Biosciences

TYPE Editorial
PUBLISHED 17 January 2023
DOI 10.3389/fmolb.2023.1139471

OPEN ACCESS

EDITED AND REVIEWED BY
Ray Luo,
University of California, Irvine,
United States

*CORRESPONDENCE
Yinglong Miao,
miao@ku.edu
Chia-En A. Chang,
chiaenc@ucr.edu
Weiliang Zhu,
wlzhu@simm.ac.cn
J. Andrew McCammon,
jmccammon@ucsd.edu

SPECIALTY SECTION
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 07 January 2023
ACCEPTED 11 January 2023
PUBLISHED 17 January 2023

CITATION
Miao Y, Chang C-EA, Zhu W and McCammon JA (2023), Editorial: Mechanisms, thermodynamics and kinetics of ligand binding revealed from molecular simulations and machine learning.
Front. Mol. Biosci. 10:1139471.
doi: 10.3389/fmolb.2023.1139471

Editorial: Mechanisms, thermodynamics and kinetics of ligand binding revealed from molecular simulations and machine learning

Yinglong Miao[1]*, Chia-En A. Chang[2]*, Weiliang Zhu[3]* and J. Andrew McCammon[4]*

[1]Center for Computational Biology and Department of Molecular Biosciences, University of Kansas, Lawrence, KS, United States, [2]Department of Chemistry, University of California, Riverside, Riverside, CA, United States, [3]Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China, [4]Departments of Chemistry & Biochemistry and Pharmacology, University of California San Diego, San Diego, CA, United States

KEYWORDS

ligand binding, thermodynamics, kinetics, molecular docking, molecular dynamics, brownian dynamics, machine learning, mechanisms

Editorial on the Research Topic
Mechanisms, thermodynamics and kinetics of ligand binding revealed from molecular simulations and machine learning

Ligand binding plays an essential role in cellular signaling. Detailed understanding of the mechanisms, structures, thermodynamics and kinetics of ligand binding is central to drug discovery in the pharmaceutical industry and academia (Baron and McCammon, 2013; Peng et al., 2019). Despite this critical importance, such tasks remain challenging in computational chemistry and biophysics. Molecular docking has proven useful in rapid virtual screening of small molecules for drug discovery, although it is often difficult to fully incorporate receptor flexibility into the docking calculations. Recent developments in computing hardware and simulation algorithms have enabled molecular dynamics (MD) simulations to capture dynamic ligand binding and dissociation processes. These simulations can then be analyzed to compute both thermodynamic free energies and kinetic rates of ligand binding (Pang and Zhou, 2017; Tang et al., 2017; Nunes-Alves et al., 2020; Wang et al., 2022). In addition, Brownian dynamics simulations have been very efficient in generating a large number of ligand binding trajectories and estimating the binding kinetic rates (Huber and McCammon, 2019; Muñiz-Chicharro et al., 2022). Finally, emerging machine learning techniques have greatly enhanced molecular simulations and facilitated analysis of the simulation trajectories (Glielmo et al., 2021).

This Research Topic is focused on studies of the pathways, mechanisms, free energies and kinetics of ligand binding to target receptors. We encouraged both method development and application papers. Potential techniques used to address these problems include molecular docking, MD, Brownian dynamics, and machine learning approaches. Systems of interest broadly involve ligand binding to any type of receptors, including proteins, nucleic acids, materials, and so on.

Carloni et al. have reviewed recent major advancements in molecular simulation methodologies for predicting dissociation rate (k_{off}), a parameter of fundamental importance in drug design. They further discuss the impact of the potential energy

function models on the accuracy of the prediction, and provide a perspective from high-performance computing and machine learning for highly efficient and accurate prediction of the constants. Roussey and Dickson have uncovered important factors of host-guest unbinding through detailed analysis of a large dataset of simulation trajectories. They have found that differences in ion densities as well as guest-ion interactions strongly correlate with differences in the probabilities of reactive paths, and play a significant role in the guest unbinding.

Joshi et al. describe the extension of their clever method using curvilinear coordinate-based sampling to study the thermodynamics of rapamycin associating with the FKBP12 enzyme, the first step in the action of this antiproliferative agent. The method uses a multiple-walker umbrella sampling simulation approach to characterizing the protein–protein interaction energetics along the curvilinear paths, and yields binding free energies and mechanistic details of rapamycin binding with wild-type FKBP12 and modifications of these molecules.

Shinobu et al. have optimized practical protocols for a 2D replica-exchange MD (REMD) method that combines generalized replica exchange with solute tempering and replica-exchange umbrella sampling (gREST/REUS). As demonstrated on ligand binding to three protein kinase systems, the method ensures good random walks in the 2D replica spaces, which are important for enhanced sampling of kinase-inhibitor binding.

Chai et al. have carried out multi-microsecond length MD simulations of STK17B in three different states. They observed the conformational dynamics of its P-loop that could flip into the ADP-binding site upon the inhibitor binding to interact with inhibitors and the protein C-lobe, leading to strengthened communications between the C- and N-lobes. Their simulation results could be useful for designing highly selective inhibitors.

Zhang et al. have carried out MD simulations and Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) calculations and revealed stronger binding of rivaroxaban in the Y99C mutant of coagulation factor X than in the Y99A mutant. Their simulations have also shown that ligand binding may not only be a dynamic process but also a dynamic state involving multiple binding poses, which could be important for drug design. Cai et al. have performed MD simulations and absolute binding free energy calculations for exploring the drug resistance mechanism of epidermal growth factor receptor (EGFR), a target protein of many non-small cell lung cancer (NSCLC) drugs. They found the binding affinity of ATP to L858R/T790M mutant is higher than that to the L858R mutant, due to the significant changes of the protein conformation and the van der Waals interactions. Their findings could be valuable for designing new drugs for NSCLC.

Girame et al., Garcia-Borràs and Feixas have applied MD simulations to investigate changes in protonation states of in-pathway residues during protein-ligand binding processes. The authors found that binding of benzamidine to trypsin was infrequent when His57 was positively charged, where His57 was part of the catalytic triad and located more than 10 Å away from the gorge of the substrate binding pocket. Their findings illustrate the importance in properly accounting for protonation states of distal residues when using MD simulations to study ligand binding pathways.

Xue et al. have performed MD simulations on glucocorticoid receptor (GR) complexed with cofactor TIF2 and five different agonists. They have uncovered a communication mechanism between the ligand-binding and cofactor-binding pockets, and identified a pair of important residues (D590 and T739) in the allosteric communication pathway, which could be useful for GR-targeted drug discovery. Shen et al. have examined 130+ RORγ complex structures with different agonists and inverse agonists, identified specific changes in the contact interaction for distinguishing active and inactive conformations, and observed essential modes for separating allosteric binding vs canonical binding and active vs inactive structures. Their simulations and analyses have also revealed some essential contacts to the constitutive activity of RORγ.

Huang et al. have built the most likely 3D structures of alpha/beta hydrolase domain-containing 5 (ABHD5) and the ABHD5-ligand complexes by combining various computational and experimental methods. Their simulations have also identified three residues and some hydrophobic interactions important for protein structure, function and the interactions with ligands and membrane.

Xiao et al. have introduced the Protein Allosteric Sites Server (PASSer2.0), which uses a geometry-based algorithm and automated machine learning to predict allosteric sites. The authors tested a total of 204 proteins from the Allosteric Database (ASD) and ASBench database. The server performed well under multiple indicators. It will provide a valuable tool to facilitate allosteric drug discovery. McKay et al. have developed an essential dynamics ensemble docking (EDED) approach to identify the most relevant receptor conformations for virtual screening. They have demonstrated the approach on docking of small-molecule antagonists of the PAC1 class B GPCR. With four representative receptor models selected from simulations and screening of three million ZINC compounds and 23 experimentally validated ligands of PAC1, they show that EDED can effectively reduce the number of false positives and improve the accuracy of docking.

The paper "Big Data Analytics for Improved Prediction of Ligand Binding and Conformational Selection" by Gupta et al. continues the work by these authors to enhance our understanding of the binding of small molecules to proteins through the conformational selection mechanism. The authors make use of modern machine learning approaches and provide valuable tools for identifying proteins that utilize these mechanisms.

In summary, remarkable advances have been made in both method development and applications in computational predictions of ligand binding free energies and kinetics (especially the dissociation rate). Advanced MD simulations have revealed mechanisms of ligand recognition and associated protein conformational changes, which often involves allosteric modulation. Novel approaches have been developed to select important receptor conformations for molecular docking and improve the docking accuracy. A new server (PASSer2.0) has been developed for predicting allosteric sites in proteins based on machine learning. It will greatly facilitate allosteric drug discovery. These advances are expected to expand our capabilities in simulations of ligand binding and drug discovery.

Author contributions

YM, C-EC, WZ and JM wrote the manuscript.

Funding

YM is supported by the National Institutes of Health (R01GM132572) and National Science Foundation (2121063). C-EC is supported by the National Institutes of Health (R01GM109045) and National Science Foundation (MCB1932984). JM is supported by the National Institutes of Health grant GM31749.

WZ is supported by the National Science Foundation of China (82273851) and National Key R&D Program (2022YFA1004304).

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

References

Baron, R., and Mccammon, J. A. (2013). Molecular recognition and ligand association. *Annu. Rev. Phys. Chem.* 64, 151–175. doi:10.1146/annurev-physchem-040412-110047

Glielmo, A., Husic, B. E., Rodriguez, A., Clementi, C., Noe, F., and Laio, A. (2021). Unsupervised learning methods for molecular simulation Data. *Chem. Rev.* 121, 9722–9758. doi:10.1021/acs.chemrev.0c01195

Huber, G. A., and Mccammon, J. A. (2019). Brownian dynamics simulations of biological molecules. *Trends Chem.* 1, 727–738. doi:10.1016/j.trechm.2019.07.008

Muñiz-Chicharro, A., Votapka, L. W., Amaro, R. E., and Wade, R. C. (2022). Brownian dynamics simulations of biomolecular diffusional association processes. *WIREs Comput. Mol. Sci. n/a*, e1649.

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, X., and Zhou, H. X. (2017). Rate constants and mechanisms of protein-ligand binding. *Annu. Rev. Biophys.* 46, 105–130. doi:10.1146/annurev-biophys-070816-033639

Peng, C., Wang, J., Yu, Y., Wang, G., Chen, Z., Xu, Z., et al. (2019). Improving the accuracy of predicting protein-ligand binding-free energy with semiempirical quantum chemistry charge. *Future Med. Chem.* 11, 303–321. doi:10.4155/fmc-2018-0207

Tang, Z., Roberts, C. C., and Chang, C. A. (2017). Understanding ligand-receptor non-covalent binding kinetics using molecular modeling. *Front. Biosci. (Landmark Ed.* 22, 960–981. doi:10.2741/4527

Wang, J., Bhattarai, A., Do, H. N., and Miao, Y. (2022). Challenges and frontiers of computational modelling of biomolecular recognition. *QRB Discov.* 3, e13. doi:10.1017/qrd.2022.11

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fmolb.2022.858316

Local Ion Densities can Influence Transition Paths of Molecular Binding

Nicole M. Roussey[1] *and Alex Dickson*[1,2]*

[1]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, United States, [2]Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, United States

Improper reaction coordinates can pose significant problems for path-based binding free energy calculations. Particularly, omission of long timescale motions can lead to overestimation of the energetic barriers between the bound and unbound states. Many methods exist to construct the optimal reaction coordinate using a pre-defined basis set of features. Although simulations are typically conducted in explicit solvent, the solvent atoms are often excluded by these feature sets—resulting in little being known about their role in reaction coordinates, and ultimately, their role in determining (un)binding rates and free energies. In this work, analysis is done on an extensive set of host-guest unbinding trajectories, working to characterize differences between high and low probability unbinding trajectories with a focus on solvent-based features, including host-ion interactions, guest-ion interactions and location-dependent ion densities. We find that differences in ion densities as well as guest-ion interactions strongly correlate with differences in the probabilities of reactive paths that are used to determine free energies of (un)binding and play a significant role in the unbinding process.

Keywords: free energy, binding affinity, molecular dynamics, weighted ensemble, ligand unbinding, mechanisms, SAMPL system

OPEN ACCESS

Edited by:
Yinglong Miao,
University of Kansas, United States

Reviewed by:
Jing Huang,
Westlake University, China
Nanjie Deng,
Pace University, United States
Hidemi Nagao,
Kanazawa University, Japan

***Correspondence:**
Alex Dickson
alexrd@msu.edu

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *19 January 2022*
Accepted: *01 April 2022*
Published: *26 April 2022*

Citation:
Roussey NM and Dickson A (2022) Local Ion Densities can Influence Transition Paths of Molecular Binding. Front. Mol. Biosci. 9:858316. doi: 10.3389/fmolb.2022.858316

1 INTRODUCTION

Atomistic simulations are a broadly used method to better understand the microscopic interactions that govern ligand binding and unbinding and to calculate critical values such as transition rates and free energies. Both rates and free energies can in principle be computed with straightforward molecular simulations, starting in either the bound or unbound state. However, the cost required to simulate binding transition paths is typically prohibitive due to high energetic barriers separating the bound and unbound states. To overcome these barriers, a variety of enhanced sampling techniques can be employed, which commonly require the use of a predefined reaction coordinate: a single collective variable that describes the progress of the (un)binding reaction.

The use of proper reaction coordinates can lead to improvements in the convergence of free energies for enhanced sampling methods Tiwary and Berne (2016) and is necessary for accurate path-based free energy calculations in biological systems Zhang and Voth (2011). Many methods have been developed to seek out optimal reaction coordinates including but not limited to VAMPnets Mardt et al. (2018), DiffNets Ward et al. (2021), Deep-TICA Bonati et al. (2021), SGOOP Tiwary and Berne (2016), and AMINO Ravindra et al. (2020). All of the above methods construct a reaction coordinate from a set of candidate features that are either predefined or require user intuition of the (un)binding process.

Significant effort has been dedicated to understanding the role of water in the ligand (un)binding process, including binding pocket solvation effects and bulk and single molecule effects Chau (2004);

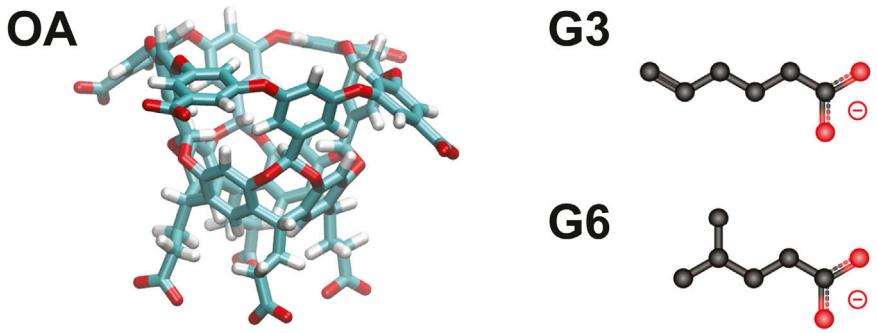


FIGURE 1 | The OA-G3 and OA-G6 systems. The OA host molecule (left). The G3 (top right) and G6 (bottom right) guest molecules.

Tiwary et al. (2015); Maurer and Oostenbrink (2019); Rizzi et al. (2021). Water molecule density has been included in reaction coordinates through the utilization of Deep-LDA Bonati et al. (2020). This method successfully found a complex reorganization of the water structure in unbinding for use as a reaction coordinate and has been able to produce accurate binding free energies Rizzi et al. (2021). The role of ions along molecular binding pathways is much less understood. Ion distributions surrounding molecules such as double stranded DNA Kolesnikov et al. (2021) and RNA Auffinger et al. (2004) have been studied and it has been found that ion affinity for molecules such as cyclodextrins and DNA is dependent on the force field used Erdos et al. (2021) as well as the water model employed Kolesnikov et al. (2021). A difference in unbinding rates has been found between implicit and explicit ions in simulation, with implicit ion representations overestimating unbinding rates across a broad range of ion concentrations Erbas et al. (2018). However, it appears that little is known about the effects of changes in ion densities along ligand (un)binding pathways.

Recent studies have demontrated that adaptations of the weighted ensemble method Huber and Kim (1996); Dickson and Brooks, (2014); Donyapour et al. (2019) can efficiently generate ligand binding and unbinding pathways that can then be used to determine rates and binding free energies Dixon et al. (2018); Lotz and Dickson (2018b); Hall et al. (2020). Specifically, an extensive analysis was conducted on a series of host-guest systems containing small, organic guest molecules ("G3" and "G6") interacting with "octa-acid" hosts ("OA") (**Figure 1**), which were originally part of the SAMPL6 (Statistical Assessment of the Modeling of Proteins and Ligands) SAMPLing challenge Rizzi et al. (2018, 2020). The REVO variant of the weighted ensemble method allowed for efficient generation of large numbers of binding and unbinding events without employing biasing forces that could perturb the (un)binding mechanism. This is notable as mean first passage times of unbinding ranged up to hundreds of seconds for these systems. It accomplishes this by running an ensemble of trajectories and periodically "resampling" this ensemble to shift computational emphasis toward unique trajectories that are moving towards a target state, and adjusting the probabilities of the trajectories accordingly. As a result,

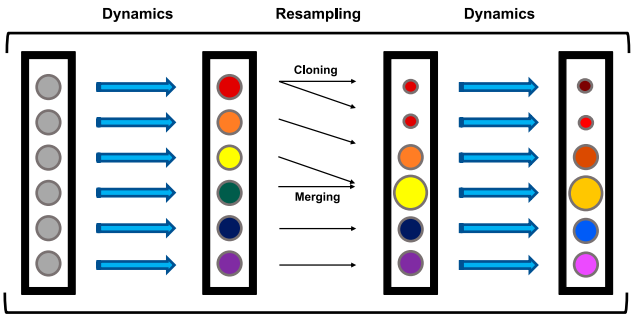


FIGURE 2 | General WE Framework. Every circle represents a trajectory in the ensemble. Colors represent conformations and circle size represents probability, with all trajectories beginning with the same conformation and probability. Trajectories are run for a predetermined number of steps (dynamics), followed by a resampling step containing merging and cloning procedures. This cycle repeats until the end of the simulation.

each unbinding pathway has an associated statistical weight (ranging from 10^{-12} to 10^{-6}) that governs how strongly it contributes to the calculation of observables, including the unbinding rate constant, k_{off}.

During these resampling steps, only the geometric relationship between the host and guest molecules was used; the positions of the water molecules and ions were neglected. Here, a time- and probability-dependent analysis of solvent based features including water and ions is presented for unbinding trajectories from the OA-G3 and OA-G6 SAMPL systems. We explore the significant differences in guest-ion interactions between high- and low-probability unbinding events, also referred to as "exit points", as well as differences in spatial arrangements of ions during unbinding. In these simulations, we have found that the generation of the most probable reactive paths requires fluctuations toward low ion densities within certain regions of the simulation box, particularly in the space immediately above the binding pocket. Differences in these ion densities along transition paths are associated with up to 10^6-fold differences in unbinding probabilities, which motivates the future inclusion of ion densities in (un)binding progress variables.

2 MATERIALS AND METHODS

2.1 Weighted Ensemble Sampling

The simulations analyzed here were previously generated Dixon et al. (2018); Hall et al. (2020) with a variant of the weighted ensemble (WE) Huber and Kim (1996) method called "REVO" Donyapour et al. (2019) utilizing the Wepy Lotz and Dickson (2020) software package. A generalized framework for WE is as follows (**Figure 2**). WE uses an ensemble of trajectories that are evolved forward in time in a parallel fashion. Each trajectory carries with it a statistical weight (w) that governs the extent to which it contributes to ensemble averages. Generally, WE simulations include two main steps: 1) An MD simulation step that moves trajectories forward in time by a predetermined time interval, and 2) a resampling step that include cloning and merging operations. Resampling is designed to both use cloning to increase the number of trajectories that have a desirable value for a feature of interest, and to decrease redundancy by merging trajectories that are similar based on the feature of interest. Together, this process aims to diversify the trajectories within the ensemble with the goal of increasing the probability of sampling rare or long-timescale events of interest for a given system. When cloning a trajectory, two new independent trajectories with the same conformation are created with half the probability, or weight (w) of the original. Merging two trajectories A and B leads to the creation of trajectory C with weight $w_c = w_a + w_b$. Trajectory C inherits either conformation A or B with a probability proportional to w_a or w_b, respectively.

A central feature of a WE simulation is the resampling function (also referred to as a "resampler") that determines which trajectories are selected for cloning and which are selected for merging. The resampler takes in an initial set of trajectories and returns a new set, which is the outcome of a series of merging and cloning steps following the rules described above. These new trajectories thus have conformations that are a subset of the initial conformation set and the sum of trajectory weights is unchanged (typically equal to 1).

In order to determine transition rates, these WE simulations were run in a nonequilibrium ensemble, where trajectories are created in the bound state and terminated in the unbound state. The unbound state was defined using a boundary condition (BC) that is satisfied when the minimum host-guest distance is greater than 1.0 nm, following previous work Lotz and Dickson (2018a). When the BC is reached, the trajectory contributes to the reactive flux calculation according to its weight at the time of crossing, which we refer to as its "exit point probability". The exit point probability can be anything between the minimum and maximum values set when the simulation was run. An exit point or unbinding event being considered "high-weight" or "low-weight"is relative, with this being dependent on the weights of all exit points within the dataset. The weights of trajectories vary because they are changed during the resampling steps that are done between rounds of dynamics in the weighted ensemble algorithm.

2.2 Resampling of Ensembles by Variation Optimization

Resampling of Ensembles by Variation Optimization (REVO) Donyapour et al. (2019) is a resampling algorithm for use with Wepy that works by maximizing a function called the **trajectory variation** (V). V is a scaled sum of the all to all pairwise distances between trajectories in the ensemble (**Eq. (1)**), where d_{ij} is the distance between trajectory i and trajectory j and V_i is the variation for trajectory i.

$$V = \sum_i V_i = \sum_i \sum_j \left(\frac{d_{ij}}{d_\star}\right)^\alpha \phi_i \phi_j \tag{1}$$

The measurement of distance between two trajectoies can be arbitrarily defined in the REVO method. In this case it was defined as the root mean squared deviation of the ligand after aligning the host molecules. As the host molecules have four-fold symmetry, four separate distances were calculated after aligning the hosts in the four symmetrically-equivalent postitions, upon which the smallest such distance was used for d_{ij}. ϕ_i is a non-negative function referred to as a "novelty" that signifies the importance of individual trajectories. In this work is was solely a function of walker weight. $d_\star$, the "characteristic distance" is the average distance after one cycle of dynamics, and is only used to make the varation function unitless. The α parameter balances the value of the distance and novelty terms and was set equal to 4. Other methodological details pertinent to data generation are available in Ref. Dixon et al. (2018) and Ref. Hall et al. (2020). The overall goal of REVO is to optimize the value of V by cloning trajectories with a high value of V_i and merging trajectories with a low value of V_i. See Ref. Donyapour et al. (2019) for more details of the REVO method.

2.3 Dataset Information

The weighted ensemble data used for this analysis comes from papers published in 2020 Hall et al. (2020) (the primary OA-G6 data set) and 2018 Dixon et al. (2018) (OA-G3 data set and a secondary OA-G6 data set). Briefly, the primary OA-G6 data set contains 10 simulations with 48 trajectories each and 1,500 cycles per trajectory that begin in the initial OA-G6-0 pose provided in the SAMPL6 SAMPLing challenge Rizzi et al. (2020). The 2018 data sets contain five simulations each with 48 trajectories and 2000 cycles per trajectory, each beginning at one of the five initial poses for the corresponding system. Reactive paths begin in the bound state and end in the unbound state when a BC is hit. The BC is defined as a 1.0 nm minimum distance between the host and guest molecules.

3 RESULTS

We find that each reactive path can be split into two phases: 1) initial departure from the bound state, and 2) full separation of the host and guest. There are often many cycles between the guest physically leaving the binding pocket of the host and the BC being hit. It was determined that in all of the reactive paths generated, a

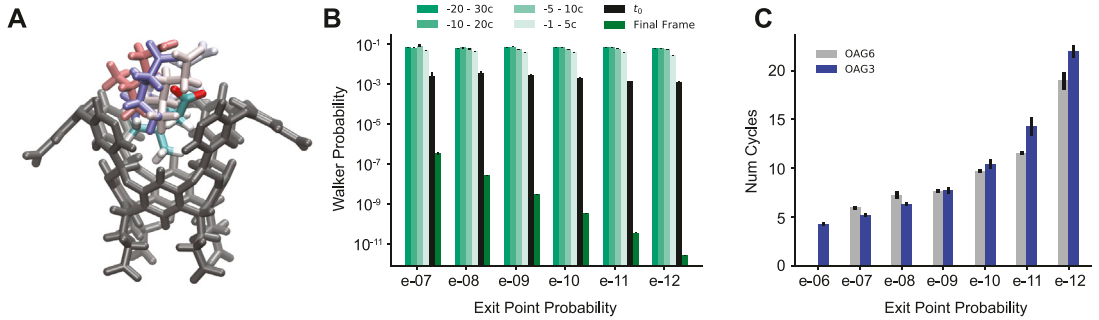


FIGURE 3 | Analysis of t_0 poses. **(A)** The OA host molecule with the G6 ligand in the starting pose (multi-color) and example t_0 poses (pastels). Some atoms from the host have been removed in A for clarity. **(B)** Average probabilities from -30 cycles to the final unbinding event organized by unbinding probability for the 2020 OA-G6 data set. **(C)** The average number of cycles between t_0 and the unbinding event for OA-G3 (blue) and the 2020 OA-G6 data set (gray) organized by unbinding probability.

TABLE 1 | The number of observed unbinding events grouped by exit point probability. The OA-G6 row corresponds to the OA-G6 2020 dataset.

	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-11}	10^{-12}
OA-G6	0	7	17	112	359	652	2220
OA-G3	10	18	88	195	483	1,116	4103

center-of-mass (COM) to COM distance of 0.7 nm indicated an irreversible transition between these two parts (**Supplementary Figure S1**). This can be seen as a physical "commitment to unbinding" point after which rebinding does not occur, where the guest has just been released from the partially solvated binding pocket (**Figure 3A**). The cycle corresponding to this point is found for all reactive paths and used for analysis; we refer to this point as t_0.

When the BC is hit for the reactive paths, the unbinding probabilities varied between 10^{-12} and 10^{-6} for OA-G3 and between 10^{-12} and 10^{-7} for OA-G6. The low probability exit points are highly abundant for both OA-G6 and OA-G3, whereas the high probability exit points occur with a very low frequency for both systems. Overall, the number of exit points *increases* as the probability of the exit points *decreases* (**Table 1**).

At and before the t_0 point, the probabilities of the reactive paths are roughly the same, with a value of 10^{-3} with only the probabilities following t_0 varying based on exit point probability (**Figure 3B**). The number of cycles between t_0 and the unbinding event also correlates to the exit point probability, with high probability exit points having ~5 cycles between the two points, and ~20 cycles for low probability exit points (**Figure 3C**). There is a steady increase in the average number of cycles between t_0 and the unbinding event as the probability of the trajectories decreases.

These differences prompt the question: are there differences in physical features associated with this large variation in exit point probability? To answer this question, a set of physical features was chosen and the values of those features were calculated for every cycle of every reactive path generated. The features in question include the number of waters in the binding site of the host, the number of ions around the upper negative charges of the host molecule, the number of ions around the guest, and the number of waters around the guest molecule (**Figure 4A**). To calculate these features, a continuous logistic function was used: $f(r) = 1 - \frac{1}{1+(e^{-S(r-r_0)})}$, where r is the minimum atomic distance between the two entities. We use two different sets of values for the interaction radius (r_0) and steepness (S) parameters: $r_0 = 3$ Å, $S = 17$ or $r_0 = 5$ Å and $S = 12$) (**Figure 4A**). The sum of $f(r)$ across all ions (or waters) is a continuous count of the number of molecules of that species surrounding the host (or guest) for that cycle.

It was found that some features were consistent or had only a slight variation across all exit point probabilities, such as the number of binding site waters and the number of waters surrounding the guest molecule (**Supplementary Figure S2**). However, some features were found to show trends that differentiated the high- and low-probability exit points. These features included the total number of positive ions surrounding the upper negative charges of the host (**Figure 4B,C**) and the number of positive ions surrounding the guest molecule (**Figure 4D,E**). In both OA-G6 and OA-G3 there is a general trend of the number of ions surrounding the upper negative charges of the host increasing as the exit point probability *increases*, although this is not observed for $1e-7$ exit points in the OA-G6 dataset. There is also a clear trend of increasing guest-Na^+ interaction as the exit point probability *decreases* including before, at, and after the t_0 point. Similar trends were observed for features on the 3 Å scale (**Supplementary Figure S3**).

As we find that the interaction between the guest and Na^+ ions correlates with the probability of the unbinding trajectories, we now examine Na^+ ion densities in the region of space directly above the host. Specifically, we examine a cylindrical region of space beginning immediately above the host and ending at the top of the box (**Figure 5A**). We find that this region is critical to determine the outcome of dissociation trajectories that have reached t_0. The autocorrelation of ion density in this region ($C(\tau)$) is surprisingly long-lived; it follows a single exponential decay with a timescale of 77 ns (**Figure 5B**).

Figure 5C shows the average number of ions in the cylinder for cycles $[t_0 - 3, t_0 + 3]$ for each reactive trajectory. An average of

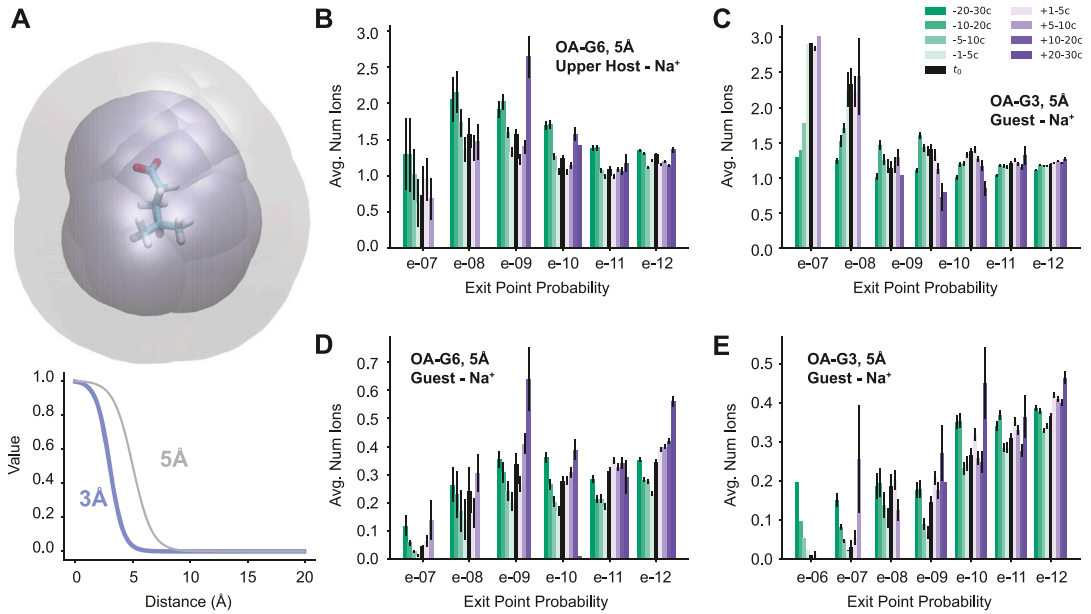


FIGURE 4 | Feature Analysis. **(A)** A visualization of the region of space considered for the guest-ion features using the G6 ligand. The maximum distance for the 5 Å scale is in gray and the 3 Å scale is in blue (top). The two logistic functions used to determine the molecule counts (bottom). **(B–E)** Molecule counts for Na^+ ions with results organized by both time and exit point probability. The legend in C applies to all four plots. The average total ion count (5 Å scale) around the upper negative charges of the host for **(B)** OA-G6 and **(C)** OA-G3. The average total ion count (5 Å scale) around the guest for **(D)** OA-G6 and **(E)** OA-G3. OA-G6 results correspond to the 2020 data set.

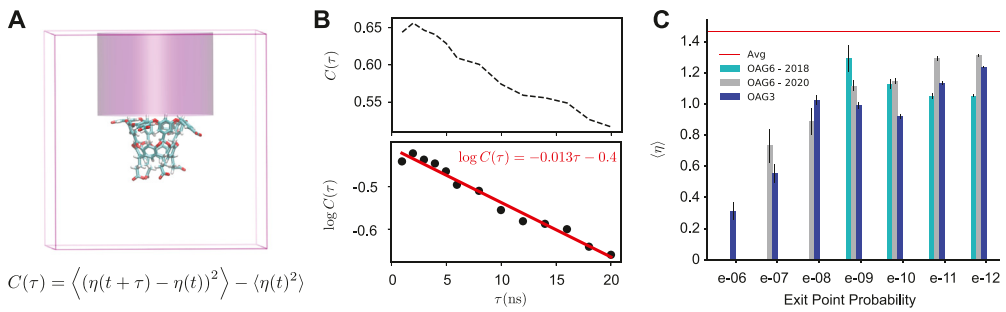


FIGURE 5 | Ion Density Analysis. **(A)** A diagram showing the simulation box and the cylindrical space above the host where the number of ions ($\eta(t)$) is determined. The equation for calculating the autocorrelation of this quantity ($C(\tau)$) is shown. **(B)** An autocorrelation plot of the cylindrical ion density ($\eta(t)$) is calculated using all reactive and non-reactive trajectory data. **(C)** The average number of ions in the cylinder space above the host for OA-G3 (dark blue), OA-G6 (2020, gray), and OA-G6 (2018, cyan). The average from 4500 random simulation cycles is shown in red.

1.47 ions was found in upper cylinder space when averaged over all available data (including reactive and non-reactive trajectories). The cylinder ion densities of *reactive trajectories* were found to be significantly lower than the bulk average regardless of exit point probability. A striking relationship was observed between this ion density and the exit point probability that was consistent across all data sets with highly weighted exit point probabilities (**Figure 5C**), where highly-weighted exit points had a significantly lower average number of ions in the cylinder. Overall, highly weighted exit points had less ions above the host, and subsequently near the guest at t_0, with this number gradually increasing as the exit point probability decreased (**Supplementary Figure S4**).

To explain these findings, we first analyzed the electrostatic forces on the guest molecule for all OA-G6 reactive trajectories at the t_0 point for one ensemble of the OA-G6 2020 data set. This was done by first removing all forces from the system other than the nonbonded (electrostatic) forces. Then the force on the ligand was determined at key points along the unbinding trajectories and average forces were determined for each exit point probability group. Results are shown for the initial bound cycles (cycles 0–6) and for the t_0-surrounding cycles used for

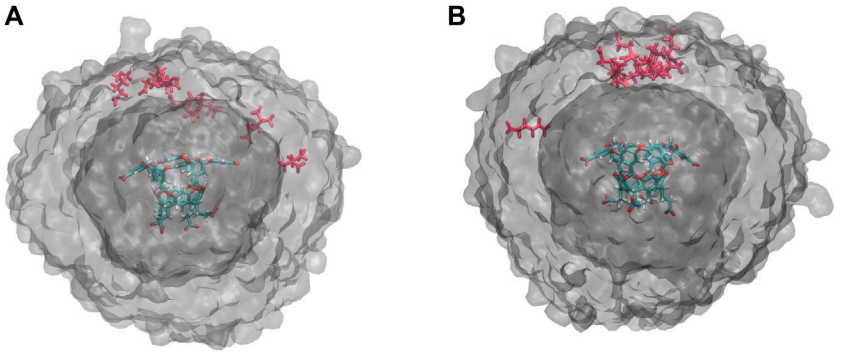


FIGURE 6 | Exit Point Analysis. **(A)** Unbinding event locations for exit points with probabilities 10^{-7} (red) and 10^{-12} (gray VolMap) for OA-G6. **(B)** Unbinding event locations for exit points with probabilities 10^{-6} (red) and 10^{-12} (gray VolMap) for OA-G3. The surfaces show a density contour (Isoval) of 0.0001 in both panels.

the cylinder-ion analysis (**Supplementary Figure S5**). We find that the net electrostatic force is pushing the guest outward from the host, and that the magnitude of this force is about 20 kJ/mol/Å higher in the initial pose (80 kJ/mol/Å) than it is at t_0 (60 kJ/mol/Å). No significant difference is found for exit points of different weights for both the overall magnitude of the electrostatic force or the z-axis contribution to the force (**Supplementary Figure S5A,B**). We found no significant difference at t_0 across all exit point probabilities despite the difference in cylinder ion-occupancy.

An alternative explanation is that differences in occupancy change the likelihood of ion interaction after the t_0 point. This is consistent with our observations in **Figure 4D,E** and would increase the number of cycles required to hit the BC (**Figure 3C**) as well as the extent of their exploration of the simulation box. Exit point locations were determined for both the highest and lowest probability exit points for both OA-G6 (10^{-7} and 10^{-12}) and OA-G3 (10^{-6} and 10^{-12}). For both systems, it was found that for high probability exit points, most guest molecules reach the BC directly above the host, whereas the low probability exit points hit the BC at a wide distribution of points surrounding the host molecule (**Figure 6**).

4 DISCUSSION

In summary, we find that location-dependent ion densities play a significant role in the unbinding process for the OA-G6 and OA-G3 systems. These systems are widely used for both the testing and development of force fields and numerous computational methods Rizzi et al. (2020, 2018); Dixon et al. (2018); Papadourakis et al. (2018); Song et al. (2018); Yin et al. (2016) necessitating a thorough understanding of the mechanics of their unbinding. It is likely that ion densities play such a prominent role due to the charged nature of these systems (−8 for the host and −1 for the guest). Similar effects might also be observed in biological systems with even more significant charge densities such as calsequestrin Yano et al. (2009), a protein necessary for muscle relaxation/contraction, with a net charge of -64, as well as systems with nucleic acids, which have a charge of −1 per nucleotide.

Further exploration and utilization of the effects of ion densities on ligand (un)binding could be done via various methods. Constraints on spatial densities of ions could be included in simulations to further examine the relationship between ion densities and unbinding rates or free energies. One possible strategy would be to conduct 2D Umbrella Sampling Park and Im (2013); Dickson et al. (2015) simulations that include a direct descriptor of (un)binding, such as the host-guest center-of-mass distance, and the ion density added as a second collective variable. Ion densities (and other features of interest) could also be utilized for resampling purposes for weighted ensemble simulations for the determination of distances between trajectories. This could encourage cloning operations of trajectories with ions in desirable locations, potentially allowing for more efficient generation of high probability unbinding events.

In weighted ensemble sampling, the equilibrium probability of a state is obtained by summing over the weights of all trajectories that have visited that state. This is similarly true for reactive paths: the overall probability of a path is determined by a weighted sum of trajectories. The analysis above breaks down a reactive trajectory set by weight, but it is important to note that relationship between the weight of a trajectory and the probability of the corresponding reaction path is not one-to-one. While high-weight trajectories in general sample from high-probability regions of space, it is possible that a low-weight trajectory could visit a high-probability reaction path. For this reason, we should consider the low-probability trajectories (e.g. $p = 10^{-12}$) as a heterogeneous group that could contain observations of high-probability reaction paths. However, the high-weight trajectories (by definition) correspond only to high-probability paths.

Overall, these results suggest that greater attention may be required for ligand-ion interactions across various simulation methods, including those that require a predefined reaction coordinate. We find that there are many microscopic trajectories that contribute to the unbinding path ensemble, some of which are much more likely than others. Methods that only sample unlikely reactive paths could have difficulty computing accurate measurements of transition rates and free energies. In addition, incorrect transition states (including inaccuracies in solvent degrees of freedom) can lead to

incorrect hypotheses about the molecular interactions that govern kinetics. This work underscores the importance of proper consideration of ion densities along unbinding pathways, especially for charged systems.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The datasets analyzed in this study are available on request to the corresponding author. Requests to access these datasets should be directed to alexrd@msu.edu.

AUTHOR CONTRIBUTIONS

Both AD and NMR designed the project. NMR conducted the research and wrote the manuscript. AD and NMR revised the manuscript and prepared it for publication.

FUNDING

The authors acknowledge funding from the National Institutes of Health (R01GM130794) and the National Science Foundation (DMS1761320).

ACKNOWLEDGMENTS

The authors would like to acknowledge Tom Dixon, Samuel D. Lotz, and Robert Hall for generating the weighted ensemble data analyzed in this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.858316/full#supplementary-material

REFERENCES

Auffinger, P., Bielecki, L., and Westhof, E. (2004). Symmetric K+ and Mg2+ Ion-Binding Sites in the 5S rRNA Loop E Inferred from Molecular Dynamics Simulations. *J. Mol. Biol.* 335, 555–571. doi:10.1016/j.jmb.2003.10.057

Bonati, L., Piccini, G., and Parrinello, M. (2021). Deep Learning the Slow Modes for Rare Events Sampling. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2113533118. doi:10.1073/pnas.2113533118

Bonati, L., Rizzi, V., and Parrinello, M. (2020). Data-driven Collective Variables for Enhanced Sampling. *J. Phys. Chem. Lett.* 11, 2998–3004. doi:10.1021/acs.jpclett.0c00535

Chau, P.-L. (2004). Water Movement during Ligand Unbinding from Receptor Site. *Biophysical J.* 87, 121–128. doi:10.1529/biophysj.103.036467

Dickson, A., Ahlstrom, L. S., and Brooks, C. L. B., III (2015). Coupled Folding and Binding with 2d Window-Exchange Umbrella Sampling. *J. Comp. Chem.* 37, 587–594. doi:10.1002/jcc.24004

Dickson, A., and Brooks, C. L., III (2014). WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm. *J. Phys. Chem. B* 118, 3532–3542. doi:10.1021/jp411479c

Dixon, T., Lotz, S. D., and Dickson, A. (2018). Predicting Ligand Binding Affinity Using on- and Off-Rates for the Sampl6 Sampling challenge. *J. Comput. Aided. Mol. Des.* 32 (10), 1001–1012. doi:10.1007/s10822-018-0149-3

Donyapour, N., Roussey, N., and Dickson, A. (2019). Revo: Resampling of Ensembles by Variation Optimization. *J. Chem. Phys.* 150 (24), 244112. doi:10.1063/1.5100521

Erbas, A., de la Cruz, M. O., and Marko, J. F. (2018). Effects of Electrostatic Interactions on Ligand Dissociation Kinetics. *Phys. Rev. E.* 91 (2-1), 022405. doi:10.1103/PhysRevE.97.022405

Erdos, M., Frangou, M., Vlugt, T. J. H., and Moultos, O. A. (2021). Diffusivity of α-, β-, γ-cyclodextrin and the Inclusion Complex of β-cyclodextrin: Ibuprofen in Aqueous Solutions; a Molecular Dynamics Simulation Study. *J. Fluid Phase Equilib.* 528, 112842. doi:10.1016/j.fluid.2020.112842

Hall, R., Dixon, T., and Dickson, A. (2020). On Calculating Free Energy Differences Using Ensembles of Transition Paths. *Front. Mol. Biosci.* 7, 106. doi:10.3389/fmolb.2020.00106

Huber, G. A., and Kim, S. (1996). Weighted-ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys. J.* 70. doi:10.1016/S0006-3495(96)79552-8

Kolesnikov, E. S., Gushchin, I. Y., Zhilyaev, P. A., and Onufriev, A. V. (2021). Similarities and Differences between Na+ and K+ Distributions Around Dna Obtained with Three Popular Water Models. *J. Chem. Theor. Comput.* 17 (11), 7246–7259. doi:10.1021/acs.jctc.1c00332

Lotz, S. D., and Dickson, A. (2018a). Multiple Ligand Unbinding Pathways and Ligand-Induced Destabilization Revealed by Wexplore. *Biophys. J.* 112, 620–629. doi:10.1016/j.bpj.2017.01.006

Lotz, S. D., and Dickson, A. (2018b). Unbiased Molecular Dynamics of 11 Min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *J. Am. Chem. Soc.* 140, 618–628. doi:10.1021/jacs.7b08572

Lotz, S. D., and Dickson, A. (2020). Wepy: A Flexible Software Framework for Simulating Rare Events with Weighted Ensemble Resampling. *ACS Omega* 5 (49), 31608–31623. doi:10.1021/acsomega.0c03892

Mardt, A., Pasquali, L., Wu, H., and Noe, F. (2018). Vampnets for Deep Learning of Molecular Kinetics. *Nat. Comm.* 9, 1. doi:10.1038/s41467-017-02388-1

Maurer, M., and Oostenbrink, C. (2019). Water in Protein Hydration and Ligand Recognition. *J. Mol. Recognit.* 32, e2810. doi:10.1002/jmr.2810

Papadourakis, M., Bosisio, S., and Michel, J. (2018). Blinded Predictions of Standard Binding Free Energies: Lessons Learned from the Sampl6 challenge. *J. Comput. Aided. Mol. Des.* 32 (10), 1047–1058. doi:10.1007/s10822-018-0154-6

Park, S., and Im, W. (2013). Two Dimensional Window Exchange Umbrella Sampling for Transmembrane helix Assembly. *J. Chem. Theor. Comput.* 9, 13–17. doi:10.1021/ct3008556

Ravindra, P., Smith, Z., and Tiwary, P. (2020). Automatic Mutual Information Noise Omission (Amino): Generating Order Parameters for Molecular Systems. *Mol. Syst. Des. Eng.* 5 (1), 339–348. doi:10.1039/C9ME00115H

Rizzi, A., Jense, T., Slochower, D. R., Aldeghi, M., Gapsys, V., Ntekoumes, D., et al. (2020). The Sampl6 Sampling challenge: Assessing the Reliability and Efficiency of Binding Free Energy Calculations. *J. Comput. Aided. Mol. Des.* 34 (5), 601–633. doi:10.1007/s10822-020-00290-5

Rizzi, A., Murkli, S., McNeill, J. N., yao, W., Sullivan, M., Gilson, M. K., et al. (2018). Overview of the Sampl6 Host-Guest Binding Affinity Prediction challenge. *J. Comput. Aided. Mol. Des.* 32 (10), 937–963. doi:10.1007/s10822-018-0170-6

Rizzi, V., Bonati, L., Ansari, N., and Parrinello, M. (2021). The Role of Water in Host-Guest Interaction. *Nat. Comm.* 12, 93. doi:10.1038/s41467-020-20310-0

Song, L. F., Bansal, N., Zheng, Z., and Merz, K. M., Jr. (2018). Detailed Potential of Mean Force Studies on Host–Guest Systems from the Sampl6 challenge. *J. Comput. Aided. Mol. Des.* 32 (10), 1013–1026. doi:10.1007/s10822-018-0153-7

Tiwary, P., and Berne, B. J. (2016). Spectral gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Nat. Acad. Sci.* 113 (11), 2839–2844. doi:10.1073/pnas.1600917113

Tiwary, P., Mondal, J., Morrone, J. A., and Berne, B. J. (2015). Role of Water and Steric Constraints in the Kinetics of Cavity–Ligand Unbinding. *Proc. Nat. Acad. Sci.* 112 (39), 12015–12019. doi:10.1073/pnas.1516652112

Ward, M. D., Zimmerman, M. I., Miller, A., Chung, M., Swamidass, S. J., and Bowman, G. R. (2021). Deep Learning the Structural Determinants of Protein Biochemical Properties by Comparing Structural Ensembles with Diffnets. *Nat. Comm.* 12 (1), 3023. doi:10.1038/s41467-021-23246-1

Yano, M., Yamamoto, T., Kobayashi, S., and Matsuzaki, M. (2009). Role of Ryanodine Receptor as a Ca2+ Regulatory center in normal and Failing Hearts. *J. Cardio.* 53, 1–7. doi:10.1016/j.jjcc.2008.10.008

Yin, J., Henriksen, N. M., Slochower, D. R., Shirts, M. R., Chiu, M. W., Mobley, D. L., et al. (2016). Overview of the Sampl5 Host–Guest challenge: Are We Doing Better? *J. Comput. Aided. Mol. Des.* 31, 1–19. doi:10.1007/s10822-016-9974-4

Zhang, Y., and Voth, G. A. (2011). A Combined Metadynamics and Umbrella Sampling Method for the Calculation of Ion Permeation Free Energy Profiles. *J. Chem. Theor. Comput.* 7 (7), 2277–2283. doi:10.1021/ct200100e

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 29 April 2022
doi: 10.3389/fmolb.2022.878830

Practical Protocols for Efficient Sampling of Kinase-Inhibitor Binding Pathways Using Two-Dimensional Replica-Exchange Molecular Dynamics

Ai Shinobu[1], *Suyong Re*[1,2] *and Yuji Sugita*[1,3,4]*

[1]RIKEN Center for Biosystems Dynamics Research, Kobe, Japan, [2]Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health, and Nutrition, Ibaraki, Japan, [3]Theoretical Molecular Science Laboratory, RIKEN Cluster for Pioneering Research, Saitama, Japan, [4]RIKEN Center for Computational Science, Kobe, Japan

Molecular dynamics (MD) simulations are increasingly used to study various biological processes such as protein folding, conformational changes, and ligand binding. These processes generally involve slow dynamics that occur on the millisecond or longer timescale, which are difficult to simulate by conventional atomistic MD. Recently, we applied a two-dimensional (2D) replica-exchange MD (REMD) method, which combines the generalized replica exchange with solute tempering (gREST) with the replica-exchange umbrella sampling (REUS) in kinase-inhibitor binding simulations, and successfully observed multiple ligand binding/unbinding events. To efficiently apply the gREST/REUS method to other kinase-inhibitor systems, we establish modified, practical protocols with non-trivial simulation parameter tuning. The current gREST/REUS simulation protocols are tested for three kinase-inhibitor systems: c-Src kinase with PP1, c-Src kinase with Dasatinib, and c-Abl kinase with Imatinib. We optimized the definition of kinase-ligand distance as a collective variable (CV), the solute temperatures in gREST, and replica distributions and umbrella forces in the REUS simulations. Also, the initial structures of each replica in the 2D replica space were prepared carefully by pulling each ligand from and toward the protein binding sites for keeping stable kinase conformations. These optimizations were carried out individually in multiple short MD simulations. The current gREST/REUS simulation protocol ensures good random walks in 2D replica spaces, which are required for enhanced sampling of inhibitor dynamics around a target kinase.

Keywords: molecular dynamics simulations, multi-dimensional replica-exchange simulations, generalized replica exchange with solute tempering, replica-exchange umbrella sampling, kinase-inhibitor binding

OPEN ACCESS

Edited by:
Yinglong Miao,
University of Kansas, United States

Reviewed by:
Alex Dickson,
Michigan State University, United States
Chung Wong,
University of Missouri–St. Louis, United States

***Correspondence:**
Yuji Sugita
sugita@riken.jp

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *18 February 2022*
Accepted: *31 March 2022*
Published: *29 April 2022*

Citation:
Shinobu A, Re S and Sugita Y (2022) Practical Protocols for Efficient Sampling of Kinase-Inhibitor Binding Pathways Using Two-Dimensional Replica-Exchange Molecular Dynamics. Front. Mol. Biosci. 9:878830. doi: 10.3389/fmolb.2022.878830

1 INTRODUCTION

Ligand binding to a target protein or enzyme plays important roles in many biological processes which regulate protein functional activity (Du et al., 2016). Understanding of the binding processes directly contributes to the design of effective drugs which specifically bind to target proteins. Recently, the drug residence time on a protein has been attracting attention in the development of

effective drugs (Bernetti et al., 2017; Schuetz et al., 2017). For this purpose, understanding the molecular mechanisms underlying protein-ligand binding processes, namely, binding pathways, transition states, encounter complexes, and binding kinetics, are essential, as well as sampling stable ligand-bound structures. Unlike most stable bound poses, transient and dynamic information is hardly accessible by experiments.

Molecular dynamics (MD) simulations are widely used to investigate conformational dynamics of biomolecules at the atomic level and are applied to many biological processes including protein-ligand binding/unbinding (De Vivo et al., 2016; Dickson et al., 2017; Bruce et al., 2018; Zhang et al., 2022). All-atom MD simulations can easily simulate protein dynamics on the 1–10 ms timescales, while high-performance MD-specialized computers are necessary to explore 1 ms or slower dynamics (Dror et al., 2011; Shan et al., 2011). Thus, conventional MD simulations of a protein-ligand complex are not sufficient for observing multiple binding/unbinding events, which are necessary for obtaining converged thermodynamics or free-energy landscapes. To go beyond, parallel trajectory MD methods (Silva et al., 2011; Plattner and Noé, 2015; Dickson, 2018; Tran et al., 2020) perform multiple short simulations and provide us with large amount of structural data for predicting long timescale dynamics. Another approach is the use of enhanced sampling MD methods such as replica-exchange MD (Sugita and Okamoto, 1999), metadynamics (Valsson et al., 2016), and others (Meng et al., 2015; Miao and McCammon, 2017; Spitaleri et al., 2018; Gobbo et al., 2019; Hénin et al., 2022) to explore a wider conformational space of systems with rugged energy landscapes by overcoming high energy barriers between multiple minimum states. Replica-exchange MD (REMD) (Sugita and Okamoto, 1999; Sugita et al., 2000) effectively overcome energy barriers through the exchange of system parameters between independently running replicas. In temperature REMD, high temperature replicas sample various conformations including unfolded, extended, or other flexible ones, while low temperature replicas explore stable structures existing at different energy minima through the replica exchange. In replica exchange with solute tempering (REST REST/REST2) (Liu et al., 2005; Terakawa et al., 2011; Wang et al., 2011), a specific region of interest is selected as "solute," and "solute temperature" exchanges are attempted with a reduced number of replicas. Hence, REST/REST2 is applicable to larger biological systems than temperature-REMD owing to the reduced computational cost. Recently, we generalized the definition of "solute" in REST2 by selecting a part of the molecule of interest and/or a part of the potential energy function terms as "solute". This method, which we refer to as the generalized REST (gREST) (Kamiya and Sugita, 2018), can reduce the number of replicas even further while observing efficient conformational dynamics of proteins or protein-ligand complexes. For instance, in gREST simulations of protein-ligand binding, the solute is defined as a target ligand as well as amino-acid sidechains near the target protein binding site, which accelerates ligand dynamics more than in REST2 simulations, where only the ligand molecule is selected as "solute". The gREST method was applied for the prediction of the correct binding pose (Niitsu et al., 2019) and affinities, when combined with absolute binding free energy calculations (Oshima et al., 2020). The replica-exchange umbrella sampling (REUS) method (Sugita et al., 2000; Fukunishi et al., 2002) exchanges geometrical parameters along a predefined collective variable (CV). This method is also applicable to large biological systems, if a good CV is used for describing the target conformational motion.

It is noteworthy that different parameters can be exchanged in a multidimensional fashion to further enhance conformational sampling of various biological systems (Sugita et al., 2000). Multidimensional REMD was first applied in protein-ligand binding simulations by Kokubo et al.(2013) where they combined REST2 with REUS (the REST2/REUS method). In their study, a target ligand was selected as solute in REST2 and the protein-ligand distance was used as a CV in REUS. After the success of this approach, we replaced REST2 with gREST and applied the gREST/REUS method to inhibitor binding/unbinding in the c-Src kinase/PP1 complex (gREST/REUS) (Re et al., 2019). We briefly describe the gREST/REUS method in the **Supplementary Text** and **Figure S1**. The simulations could enhance inhibitor dynamics around c-Src kinase and we observed a total of about 100 binding/unbinding events for all replicas. Using the well-converged free-energy landscapes of protein-ligand binding processes, multiple binding pathways, transition states, encounter complex structures, and other atomistic insights were obtained for the c-Src kinase-PP1 complex in solution.

The gREST/REUS method is theoretically applicable to any biological system for studying molecular mechanisms of protein-ligand binding/unbinding processes. However, the size and flexibilty of the ligand increase the computational difficulty. Here, we re-examine the practical protocols of the two-dimensional (2D) gREST/REUS protein-ligand binding simulations and apply them for three kinase-inhibitor systems: c-Src kinase with PP1 (Src-PP1), c-Src kinase with Dasatinib (Src-Dasatinib), and c-Abl kinase with Imatinib (Abl-Imatinib) (**Figure 1**). As the size and flexibility of the ligand increases in the aforementioned order, binding simulations are expected to be more challenging. Kinase-inhibitor binding processes have been subjected to both long-time conventional MD (Shan et al., 2011; Morando et al., 2016; Paul et al., 2020; Sohraby et al., 2020) and enhanced sampling MD simulations (Yang et al., 2009; Lin et al., 2013; Tiwary et al., 2017; Gobbo et al., 2019; Koneru et al., 2019; Narayan et al., 2020; Spitaleri et al., 2020; Narayan et al., 2021; Shekhar et al., 2021). However, to gain more atomistic insight on protein-ligand binding processes, better computational algorithms and practical protocols are necessary. In this paper, we describe how to optimize parameters and procedures for the setup of gREST/REUS simulations and target biomolecular systems. The role of flexible inhibitor binding in c-Src/c-Abl kinases will be discussed in a separate paper, thus here we focus on the practical issues and the protocols, which are non-trivial when performing gREST/REUS simulations with more than a hundred replicas. The protocols presented here can be useful for carrying out ligand binding/unbinding simulations of various biomolecular systems with the gREST/REUS method on massively parallel supercomputers or GPU clusters.

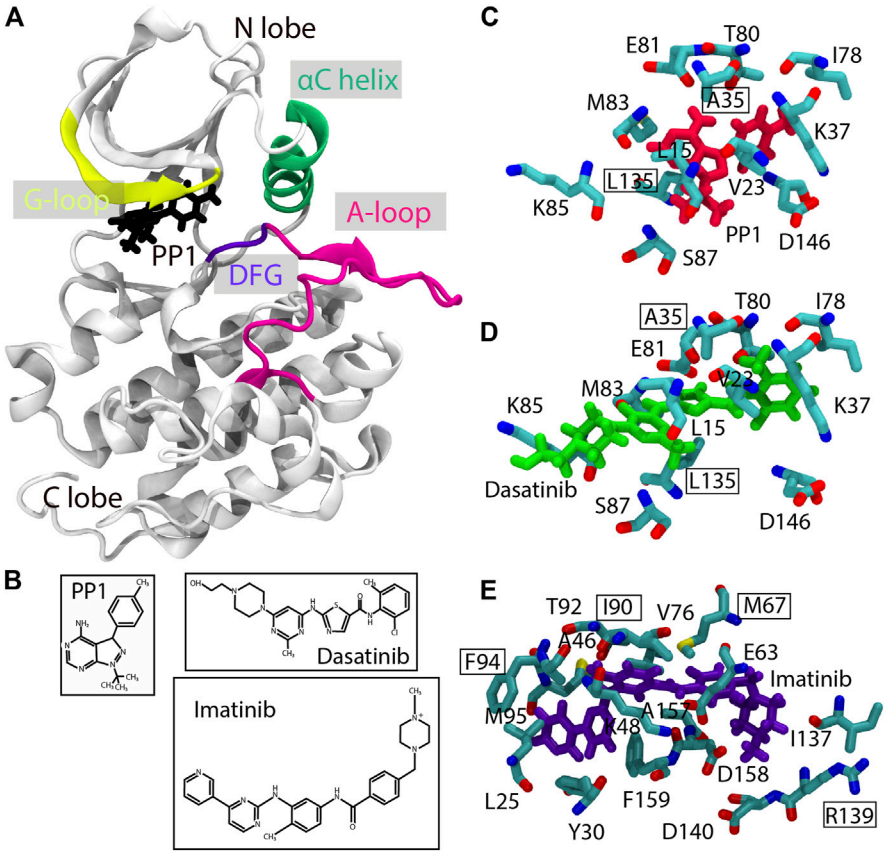


FIGURE 1 | Structures of the Src-PP1, Src-Dasatinib, and Abl-Imatinib complexes. **(A)** Src-PP1 model from X-ray structures (PDB ID: 1Y57/1QCF). **(B)** chemical structures of PP1, Dasatinib, and Imatinib. **(C)–(E)** Binding site of Src-PP1 **(C)**, Src-Dasatinib **(D)** and Abl-Imatinib **(E)** from X-ray structures (PDB ID: 1Y57/1QCF, 1Y57/3G5D and 1IEP/2OIQ for protein/ligand, respectively). PP1, Dasatinib, and Imatinib are colored red, green, and purple, respectively. Protein residues used as gREST solute regions are also shown. Residues used as protein COM for REUS CV are outlined.

2 METHODS

2.1 The gREST/REUS Simulation Protocols

The 2D-REMD methods such as gREST/REUS typically require a large number of replicas (i.e., more than 100 replicas), while they can realize better random walks in replica space including the bound, intermediate, and unbound states of the protein/ligand complexes. The preparation of replicas and the choice of solute temperatures in gREST and/or collective variables in REUS directly affects the conformational sampling efficiency. For instance, if there exist large distribution gaps between replicas, we cannot observe good random walks in replica space. This situation is equivalent to performing multiple independent REMD simulations with smaller number of replicas, which might lead to missing important intermediate structures and slow convergences of thermodynamic data. Initial setups of the gREST/REUS simulations are thus, essential for successful gREST/REUS calculations and for obtaining reliable simulation results.

In gREST/REUS, replica random walks are necessary in both the gREST and REUS dimensions. The former is realized only when the solute region and replica temperatures are defined appropriately, and we can observe sufficient overlaps in potential energies between replicas at neighboring solute temperatures. In REUS simulations, the choice of CVs, replica distributions along the CV, and proper force constants in US potentials are all important. There are many parameters and choices of procedures in gREST/REUS simulations with more than 100 replicas. For simplifying the parameter optimization, we tuned the parameters in each dimension separately using multiple short MD or gREST/REUS simulations, as described below.

2.1.1 Definition of the Protein-Ligand Distance as a CV for REUS

The protein-ligand distance is commonly used in binding MD simulation studies. The distance is usually measured as that between the centers of mass (COMs) of the backbone atoms of the selected binding site residues (protein anchor sites) and ligand heavy atoms (ligand COM). For Src-PP1 and Src-Dasatinib, the backbone atoms of Ala35 and Leu135 in c-Src kinase are used as the protein anchor site. All the heavy atoms in

PP1 and Dasatinib were used for obtaining the ligand COM, since they are small compounds with less conformational flexibilities than Imatinib, which is composed of five rings. There are multiple choices for Abl-Imatinib for the protein anchor sites and the ligand COM. As the former, we tested four choices: "2 sites" (Ile90 and Arg139), "3 sites" (Ile90, Arg139, and Phe94), "4 sites" (Ile90, Arg139, Phe94, and Met67), and "5 sites" (Ile90, Arg139, Phe94, Met67, and Phe159), respectively. For the ligand COM, four definitions including a single ring ("Ring3"), three rings ("Ring135" and "Ring 234"), and all rings ("Ring all") were examined. We expect that Imatinib flexibility is important not only near the binding site but also in the intermediate or unbound structures. A good combination of the protein anchor sites and the ligand COM may reduce the number of possible protein-ligand complex structures near the binding sites. In our protocols, the ligand COM definition was first examined for Imatinib and then, multiple choices of the protein anchor sites were tested for Abl-Imatinib simulations.

2.1.2 Preparation of Initial Structures in REUS

Thirty replicas were used for covering the protein-ligand distance in the range of 3–18 Å for Src-PP1, and 3–23 Å for Src-Dasatinib and Abl-Imatinib. We obtained the initial structure of each replica using two US simulations: In the "forward pull" simulation, the ligand was gradually pulled away from the protein binding site, while it was subsequently pulled back to the bound pose in the "reverse pull" simulation. Each replica was simulated for 300 ps with a force constant of 4 kcal/mol/Å^2. Positional restraints on the protein Cα atoms with a force constant of 1 kcal/mol/Å^2 were necessary during the pulling simulations to prevent artificial deformations of the protein. In this stage, the 30 initial structures were set in equidistance in the REUS dimension.

2.1.3 Tuning of Solute Temperatures in gREST

The solute region in gREST was defined as the dihedral angle and the nonbonded energy terms of the ligands and binding-site residues of the proteins (ca. 10 residues defined as SITE residues in the X-ray structure as shown in **Figures 1C–E**, and listed in **Supplementary Table S1**). We determined the solute temperatures using the automatic parameter tuning tool in the GENESIS MD program (Kobayashi et al., 2017). Given initial temperatures and desired acceptance ratio as inputs, the tool finds a set of solute temperatures which satisfies the desired acceptance ratio. The initial temperatures and the target acceptance ratio were set in the range of 310–663 K and 0.2, respectively. We performed five rounds of the tuning simulations (1.1 ns for each replica), by gradually increasing the frequency of exchange attempts (from every 0.21 ps in the first round to every 2.1 ps for final round), until temperature values were converged. The tuning was performed in 1D-gREST simulations at the bound (protein-ligand distance of 3.0 Å), intermediate (10.3 Å for Src-PP1, and 15.0 Å for Src-Dasatinib and Abl-Imatinib), and the unbound states (18.1 Å for Src-PP1, and 23 Å for Src-Dasatinib and Abl-Imatinib). The final temperature values were taken as the average of those obtained at the above three states.

2.1.4 Determination of REUS Parameters

To ensure sufficient potential energy overlaps between adjacent replicas, which is a pre-requisite for good REUS performance, we conducted several short trial simulations, while manually tweaking the location and force constants. At each round, we assessed the distribution overlaps between replicas and the acceptance ratios, and accordingly modified the REUS parameters, namely, the center position and the force constant of each harmonic umbrella potential. The tuning procedures were repeated in 1D-REUS simulations at three solute temperatures: 310 K (at lowest) for all three systems, 478 K, 471 K, and 440 K (at middle), and 692 K, 663 K, 590 K (at highest) for Src-PP1, Src-Dasatinib, and Abl-Imatinib, respectively.

2.2 System Preparation

The initial structure of Src-PP1 was taken from our previous work (Re et al., 2019). In brief, we extracted the kinase domain (residues 260–533, renumbered 2–275 in this work) from the X-ray crystal structure of the active-like c-Src kinase (PDB ID: 1Y57) (Cowan-Jacob et al., 2005) and replaced the co-crystallized ligand with PP1 bound to c-Src (PDB ID: 1QCF) (Schindler et al., 1999). The initial structures of Src-Dasatinib and Abl-Imatinib were constructed with the same modeling protocol used for Src-PP1. For Src-Dasatinib, we used the kinase domain of c-Src kinase (PDB ID: 1Y57) (Cowan-Jacob et al., 2005) and replaced the co-crystallized ligand with Dasatinib from an X-ray structure (PDB ID: 3G5D) (Getlik et al., 2009). Similarly, for Abl-Imatinib, we used the kinase domain (residues 225–498, renumbered 2–275 in this work) of c-Abl kinase (PDB ID: 1IEP) (Nagar et al., 2002) and the ligand structure from an X-ray structure (PDB ID: 2OIQ) (Seeliger et al., 2007). Each kinase-inhibitor complex was solvated with water molecules, where the number of water molecules was 7,698, 13,992, and 17,485 for Src-PP1, Src-Dasatinib, and Abl-Imatinib, respectively. The size of the simulation boxes for Src-Dasatinib and Abl-Imatinib was larger than for Src-PP1 because the farthest REUS replica (created by the US pulling simulations) was placed farther from the binding site (23 Å vs 18Å). The systems were neutralized by adding sodium counter ions (six for Src-PP1 and Src-Dasatinib and eight for Abl-Imatinib). Each system was minimized for 1,000 steps while applying a positional restraint of 10.0 kcal/mol/Å^2 on protein backbone atoms. Then it was gradually heated to 310 K in the NVT ensemble for 105 ps, followed by equilibration in the NPT ensemble for 105 ps. Finally, the system was equilibrated for 1.05 ns in the NPT ensemble without restraining the protein atoms. Modeling was performed using AmberTools16 (Case et al., 2021).

2.3 MD Simulation

Simulations were performed using the GENESIS MD program (Jung et al., 2015; Kobayashi et al., 2017) version 2.0 beta (Jung et al., 2021). The AMBER ff99SB-ILDN (Hornak et al., 2006; Lindorff-Larsen et al., 2010) force field was used for the proteins, GAFF (Wang et al., 2004) (with AM1-BCC) for the ligands, and the TIP3P (Jorgensen et al., 1983) was used for water molecules.

TABLE 1 | System models and simulation details.

System	Src-PP1 (Forward/Reverse[a])	Src-dasatinib	Abl-imatinib
Protein structure	1Y57 Cowan-Jacob et al. (2005)	1Y57 Cowan-Jacob et al. (2005)	1IEP Nagar et al. (2002)
Ligand structure	1QCF Schindler et al. (1999)	3G5D Getlik et al. (2009)	2OIQ Seeliger et al. (2007)
Number of atoms	27,549 (7,698 waters)	46,240 (13,992)	56,952 (17,485)
gREST solute temperature range, K	310–692	310–663	310–590
REUS distance range, Å	3.0–17.9/3.0–18.05[b]	3.0–23.1	2.7–23.0
Simulation time per replica, ns	500	750	1,000

[a]*For simulations that were initiated from REUS, replicas obtained from the forward and reverse simulations.*
[b]*Range of values for forward simulations/range of values for reverse simulations.*

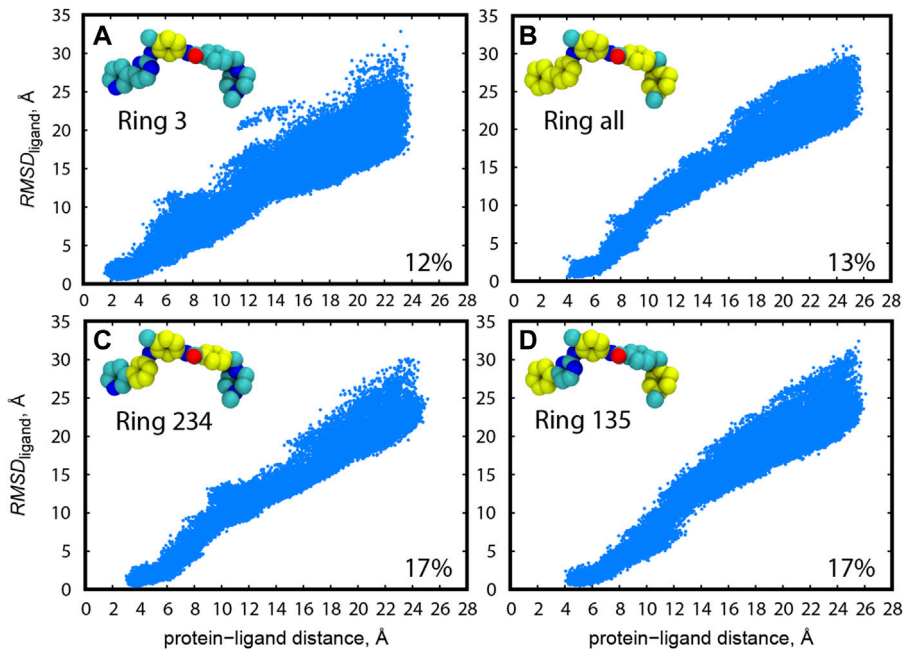


FIGURE 2 | Distribution of $RMSD_{ligand}$ along the protein-ligand distance for trial simulations (10 ns) for all replicas (1–240) of Abl-Imatinib for different definition of ligand COM atoms: **(A)** "Ring 3", **(B)** "Ring all", **(C)** "Ring 234", and **(D)** "Ring 135". Atoms used for ligand COM definitions are colored yellow. Ligand rings are numbered from 1 to 5, starting from the left. Protein atoms used for COM are backbone atoms of I90 and R139. The percentage of replicas that reached the bound pose is written on the bottom right of each panel.

Bonds involving hydrogen atoms were constrained using the SHAKE algorithm (Ryckaert et al., 1977). Water molecules were kept rigid using the SETTLE algorithm (Miyamoto and Kollman, 1992). Long-range electrostatic interactions were evaluated using the Particle-mesh Ewald summation (Darden et al., 1993; Essmann et al., 1995). The cutoff distance for non-bonded interaction was 8 Å. The NVT ensemble was used with the Bussi thermostat (Bussi et al., 2007) for keeping the temperature at 310 K, with a temperature coupling time of 5 ps. A timestep of 3.5 fs was used with the RESPA integrator (Tuckerman et al., 1992) and hydrogen mass repartitioning (HMR) (Feenstra et al., 1999) was applied on solute atoms with an HMR ratio of 3.0 (Jung et al., 2021).

Eight gREST replicas and 30 REUS replicas were used in the 2D-gREST/REUS simulation. In total, the number of replicas in each run was 240. All replicas were equilibrated for 1.05 ns without exchange attempts, followed by production runs. Exchanges were attempted every 2.1 ps alternatively in the gREST and the REUS dimensions. The gREST/REUS simulations were performed for 500 ns per replica for Src-PP1, 750 ns per replica for Src-Dasatinib, and 1,000 ns per replica for Abl-imatinib. For Src-PP1, two simulations using initial replicas from either the "forward pull" or "reverse pull" US simulations were performed (referred to as Src-PP1 and Src-PP1-Rev, respectively). All simulations steps (except the US pulling simulations for creating the initial REUS replicas) were performed without any restraints on protein atoms. The total simulation time in the current work was 660 μs. Frames for analysis were written every 10.5 ps. Simulations were performed on the supercomputer Fugaku[1] using 480 nodes.

[1]https://www.r-ccs.riken.jp/en/fugaku/.

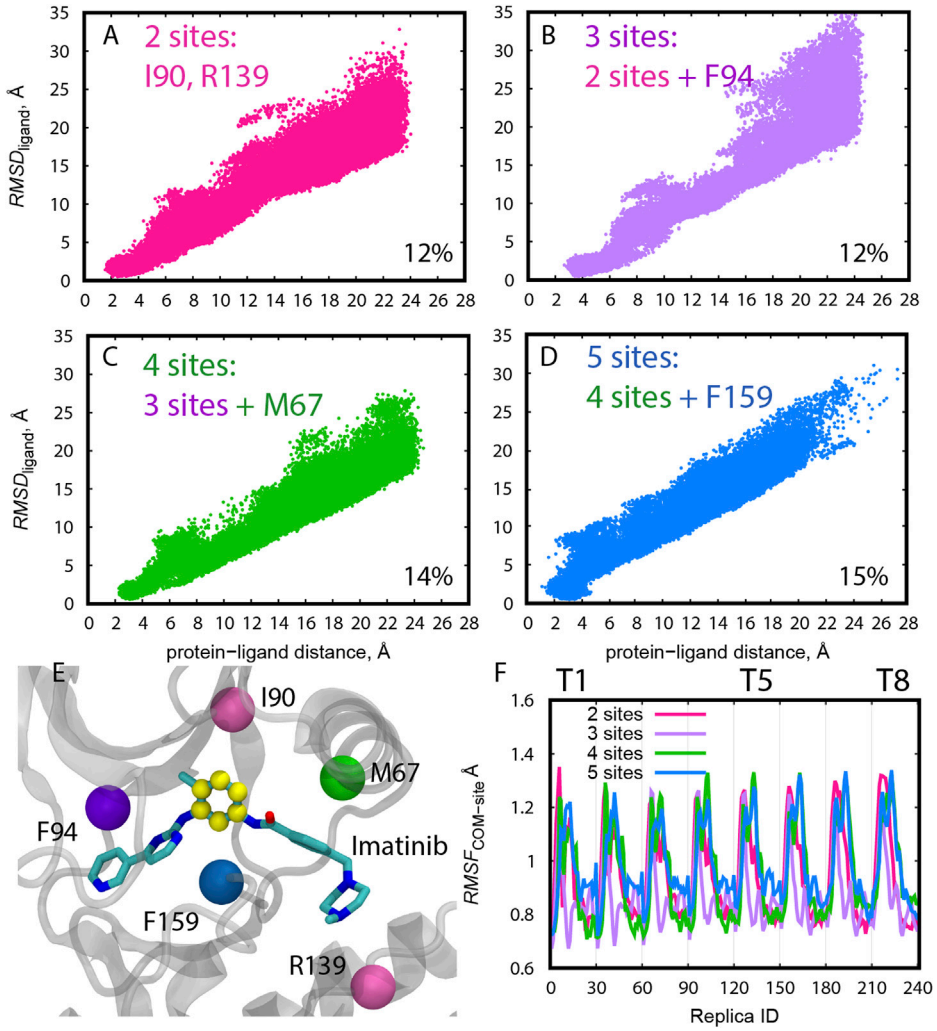


FIGURE 3 | (A–D) Distribution of $RMSD_{ligand}$ along the protein-ligand distance for trial simulations (10 ns) for all replicas (1–240) of Abl-Imatinib for different definition of protein COM atoms. The percentage of replicas that reached the bound pose is written on the bottom right of each panel. **(E)** Definition of COM atoms. Cα atoms of the residues used for COM definition of the protein are shown as colored balls. Atoms used for ligand COM definition ("Ring3") are colored yellow. **(F)** Root-mean-square-fluctuation (RMSF) of the COM of the protein anchor site atoms calculated for the 10 ns trial simulations. The reference structure used for calculating the RMSF was the initial X-ray structure. For the purpose of RMSF calculations replicas were sorted according to their REUS and gREST parameters as follows. Each group of 30 replicas belong to a single solute temperature, where replicas 1–30 represent T1 (lowest temperature), and replicas 211–240 represent T8 (highest temperature). Within each temperature, replicas are ordered according to increasing protein-ligand distances such that replicas 1 and 30 represent the smallest and largest distances, respectively.

The GENESIS 2.0 beta version was optimized to run on Fugaku and obtained a speed of >50 ns/day. The details of the systems and the simulation conditions are summarized in **Table 1**.

3 RESULTS

3.1 Tuning the Definition of Protein-Ligand Distance in the REUS Dimension

As for Src-PP1 and Src-Dasatinib, we defined the protein-ligand distance using the "2 site" model (Ala 35 and Leu135) in c-Src kinase for the protein anchor sites and all the heavy atoms for calculating the ligand COM. Due to the inhibitor size and flexibility, we tested multiple choices of the protein anchor sites and the ligand COMs for Abl-Imatinib by short (10 ns) gREST/REUS trial simulations. **Figure 2** shows the distribution of the ligand RMSD ($RMSD_{ligand}$) with respect to the bound pose of the X-ray crystal structure (2OIQ) (Seeliger et al., 2007) along the protein-ligand distance. As for the ligand COM, three rings ("Ring 135" and "Ring 234"), a single ring ("Ring 3") and all rings ("Ring all") were tested when we used "2 site" (Ile90 and Arg139) as the protein anchor site in c-Abl kinase. The probability of finding the bound pose, which we defined as the percentage of replicas that reached $RMSD_{ligand} < 1$ Å at least once during the simulation, is also shown. An efficient pose sampling should give a linear correlation with narrow distribution. CVs

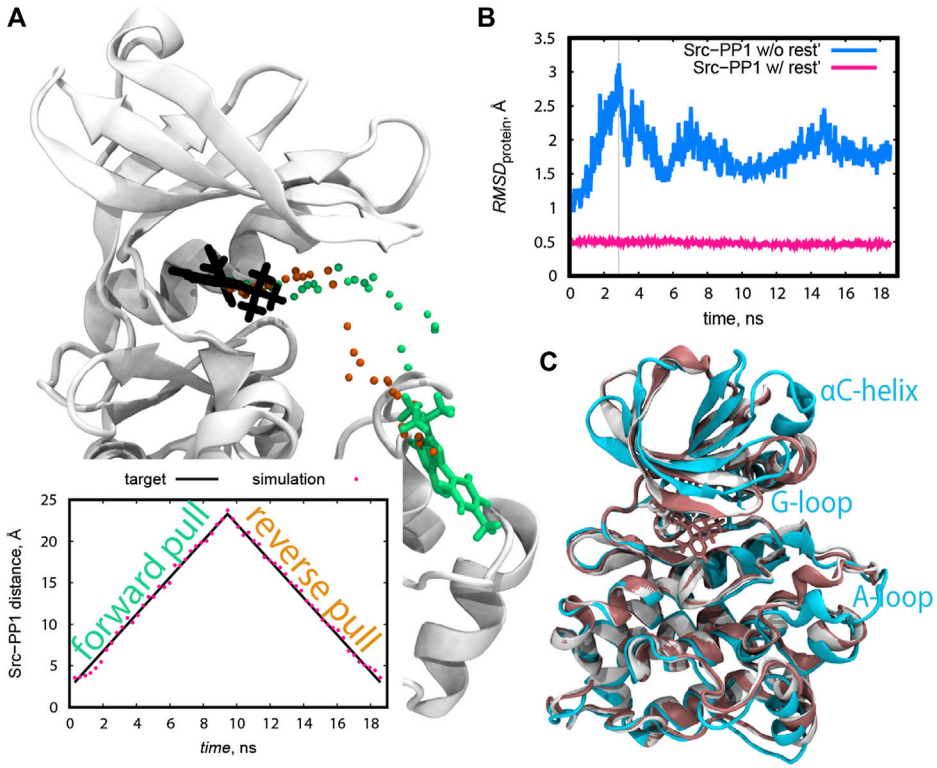


FIGURE 4 | (A) The course of the ligand PP1 during the pulling simulations represented as the coordinates of the C9 atom of PP1 at the end of each pulling step. The forward and reverse pulling directions are represented by green and orange dots, respectively. The locations of the ligand in the X-ray structure and in the last forward pulling simulation are shown in licorice representation in black and green, respectively. Inline plot: Src-PP1 distance for pulling simulations using a force constant of 4 kcal/mol/Å^2. Target distances are shown as black lines. **(B)** RMSD of protein backbone atoms during pulling simulations of PP1 from c-Src kinase with and without 1 kcal/mol/Å^2 positional restraints on the protein Cα atoms. The pulling was performed by applying the force of 4 kcal/mol/Å^2 over the protein-ligand COM distance. The *X*-axis represents the total time of concatenated consecutive pulling simulations. **(C)** Snapshots of the protein from the simulations described in **(B)** with (pink) and without (cyan) positional restraint on protein atoms. The snapshots were taken at the time marked by the grey vertical line in **(B)**.

with multiple rings in Imatinib ("Ring all", "Ring 234", and "Ring 135", **Figures 2B–D**) display linear and narrow distributions overall, compared to the single ring ("Ring 3", **Figure 2A**). The latter could possess various conformations at the same distance, likely worsening the efficiency. "Ring 234" (**Figure 2C**) and "Ring 135" (**Figure 2D**) both have higher probabilities of finding the bound pose, while the latter shows slightly narrower distribution in the range of short protein-ligand distances. These results suggest that three anchor sites (a molecular center and both edges, "Ring 135") is the practical choice for Abl-Imatinib.

At the same time, we tested four choices for the protein anchor sites using "Ring 3" as the ligand COM in Abl-Imatinib simulations (**Figure 3**). The overall distribution of $RMSD_{\mathrm{ligand}}$ becomes narrow with increasing number of protein anchor sites. The probability of finding the bound pose is higher for "4 sites" and "5 sites" (14–15%, **Figures 3C,D**) than "2 sites" and "3 sites" (12%, **Figures 3A,B**), suggesting that two or three anchor sites are not sufficient to resolve the bound conformations of a ligand as large as Imatinib. "5 sites" produces a relatively wide distribution compared to "4 sites" at the bound region (~4 Å, **Figure 3D**). Increasing the number of residues in the protein anchor sites (**Figure 3E**) is effective for resolving the ligand position and orientation but bears the risk of making protein anchor sites unstable. The COMs with "5 sites" indeed fluctuated more than the others through the replicas (**Figure 3F**). Consequently, the "4 sites–Ring135" pair was chosen as the best combination for Abl-Imatinib.

3.2 Preparation of Initial Structures in REUS From the Pulling Simulations

The initial structures along the protein-ligand distance CV were prepared from the following pulling simulations. Both forward (pulling away from the bound pose) and reverse directions (pulling back to the bound pose) were examined in the case of Src-PP1. The resulting initial pathways differed from each other (**Figure 4A**), suggesting that "dual direction pulling" could reduce the initial structure dependence and improve the convergence of the simulation results. To prepare the initial coordinates at each of the desired protein-ligand distances along the path in short simulations (9 ns per each), a rather strong force constant (4 kcal/mol/Å^2) of the umbrella potential was required (**Figure 4A**). Note that the pulling simulations can introduce an artificial structure change in the protein. In the case of Src-PP1, the structures

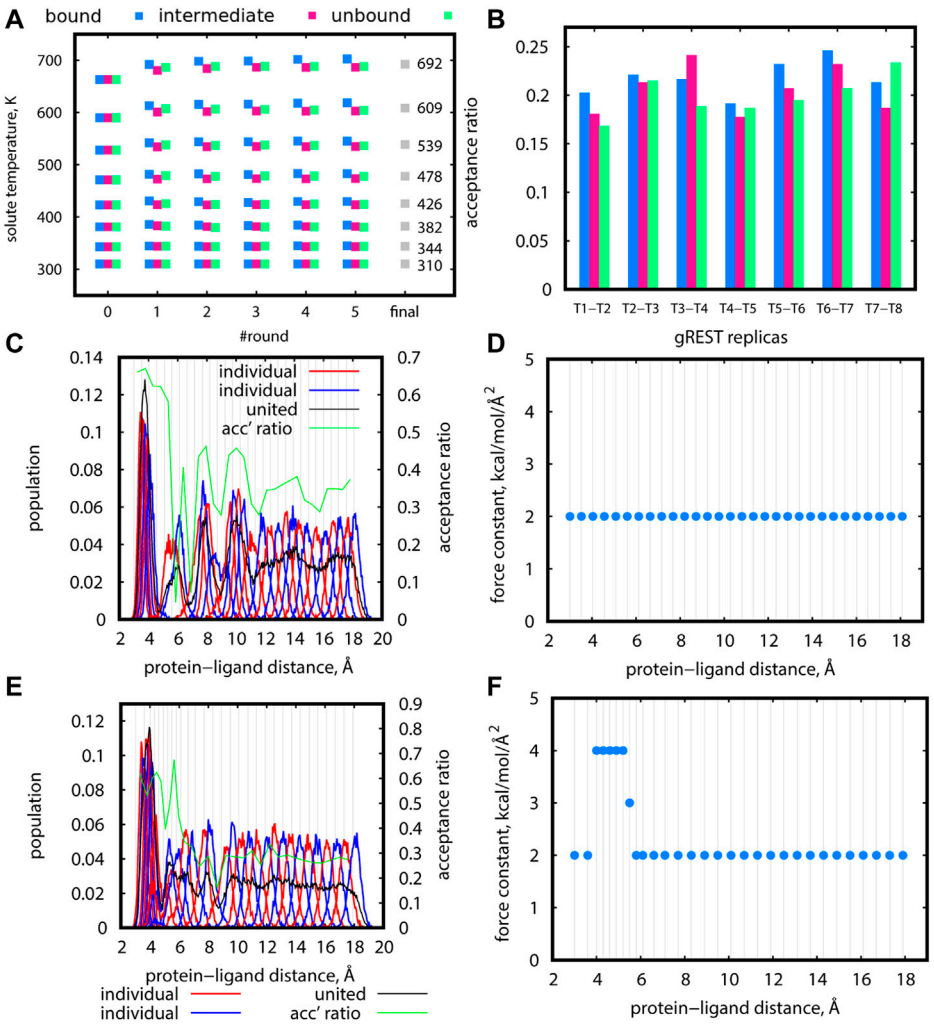


FIGURE 5 | (A) Final gREST temperatures after each automatic tuning round at three protein-ligand distances for the Src-PP1 system with a target acceptance ratio of 0.20. Round "0" specifies the initially guessed temperatures. Round "final" is the final temperature obtained from averaging the final temperatures for the three distances. **(B)** Acceptance ratios between adjacent replicas in simulations using the temperatures obtained in round "5" of the gREST tuning procedure described in **(A)**, at three protein-ligand distances. **(C)**, **(E)** Distribution of replicas according to their REUS distance for short trial simulations of 5.3 ns at 310 K for Src-PP1 (using initial replicas from the forward pulling simulations). Distributions of adjacent individual replicas ("individual") are shown in alternating red/blue lines for better visibility. Distributions of all replicas ("united") are shown in black lines. Population values for "united" data were scaled to match the "individual" populations. Acceptance ratios between adjacent REUS replicas are shown in green lines. **(D)**, **(F)** Force constants used for the simulations **(C)** and **(E)**, respectively. Vertical lines mark the protein-ligand COM distance at each replica. Blue dots mark the value of the force constant used at each REUS distance.

around the αC-helix, the G-loop, and the A-loop region significantly deviated from the X-ray crystal structure (**Figures 4B,C**). Since these regions directly relate to the binding mechanism, we also applied 1 kcal/mol/Å^2 restraints on the protein Cα atoms to avoid the artificial structure changes.

3.3 Tuning of Solute Temperatures in gREST

Solute temperatures in gREST could be determined rather effortlessly using the automatic tuning tool in GENESIS (Kobayashi et al., 2017). For Src-PP1, we set the initial temperatures in the range of 310–663 K, which is much narrower than our previous work (Re et al., 2019). This change markedly improved the sampling along the solute temperature space. In addition, there are two key points in determining the temperatures. First, multiple rounds of tuning are desired. **Figure 5A** shows the solute temperatures determined at each of the five tuning rounds, where we set the final temperature at each round as the initial temperature for the subsequent round. The temperature values changed for the first few rounds and converged. Second, tuning at different protein-ligand distances are desired. For Src-PP1, we performed the tunings at protein-ligand distances of 3.0 Å ("bound"), 10.3 Å ("intermediate"), and 18.1 Å ("unbound") distances. The resulting temperatures slightly differ in the three states (**Figure 5A**), and we therefore took their average at the final round to obtain the final set of solute temperatures. The resulting

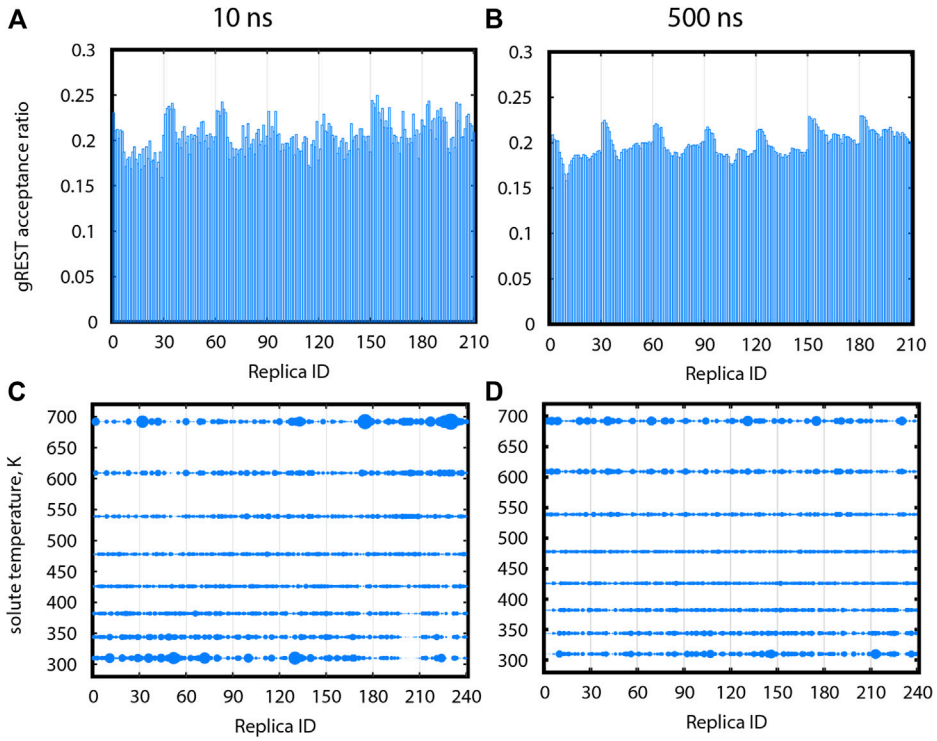


FIGURE 6 | Sampling in gREST dimension after 10 ns **(A,C)** and after 500 ns **(B,D)** for gREST/REUS simulations of Src-PP1. **(A)**, **(B)** Acceptance ratios between each replica and the gREST replica adjacent and above it. **(C)**, **(D)** Relative population for each replica, at each gREST replica. Sphere size is proportional to the population. Replicas assigned different initial solute temperatures are separated by vertical lines, where replicas 1–30 were assigned the initial temperature of 310 K (T1), replicas 31–60 were assigned the initial temperature of T2, etc.

solute temperatures provided uniform acceptance ratios along the gREST replicas (**Figure 5B**). Temperature tuning for Src-Dasatinib and Abl-Imatinib was performed using the same scheme. In practice, we manually changed the value of target acceptance ratios at each round for obtaining the final acceptance ratio of 0.2. The final sets of solute temperatures are listed in **Supplementary Table S1**.

3.4 Tuning of the REUS Parameters

The tuning of distance replicas and force constants in REUS simulations was more challenging, and several trial rounds were required for obtaining proper values. For Src-PP1, we started with even-spaced distance replicas and a uniform force constant of 2 kcal/mol/Å^2 at 310 K (**Figures 5C,D**). The sampled distance distribution was uneven. For example, the regions around 4 Å, 6 Å, and 9 Å are poorly covered, while there is an overlap in the region under 4 Å giving a large population in that region. The acceptance ratios around 6 Å drops to nearly zero, indicating almost no exchanges between replicas in that region. Accordingly, we put more replicas in poorly covered regions and set the force constants in those replicas to larger values (3 and 4 kcal/mol/Å^2) (**Figures 5E,F**). This modification resulted in better coverage of the REUS space and acceptance ratios of above 0.2 for most replicas, ensuring the occurrence of replica exchanges throughout the REUS dimension. Nevertheless, we still observed an overly large population of replicas in the bound region of under 4 Å alongside regions with poor coverage. The gap cannot be eliminated altogether since poorly covered regions represent high energy regions on the free energy landscape. This demonstrates the necessity of performing replica exchanges in two dimensions where increasing the temperature will facilitate the crossing of high energy barriers. The final REUS replica placements and force constant values are given in **Supplementary Table S1**.

3.5 Sampling Efficiency of the gREST/REUS Simulations After Parameters Tuning

As production runs, gREST/REUS simulations with 240 replicas were executed on the three systems, Src-PP1 (500 ns), Src-Dasatinib (750 ns), and Abl-Imatinib (1,000 ns), using the optimal parameters determined as described in previous sections. In the following sections, we quantify their sampling efficiencies in replica space and in the conformational space of the kinase-inhibitor complexes.

3.5.1 Random Walks in the gREST Dimension

Proper exchanges in the gREST dimension will allow replicas to go back and forth between low and high solute temperatures to sample high energy conformations. **Figures 6A,B** show acceptance ratios between adjacent gREST replicas for Src-PP1 after 10 and 500 ns, respectively. Acceptance ratios average around 0.2 as early as 10 ns.

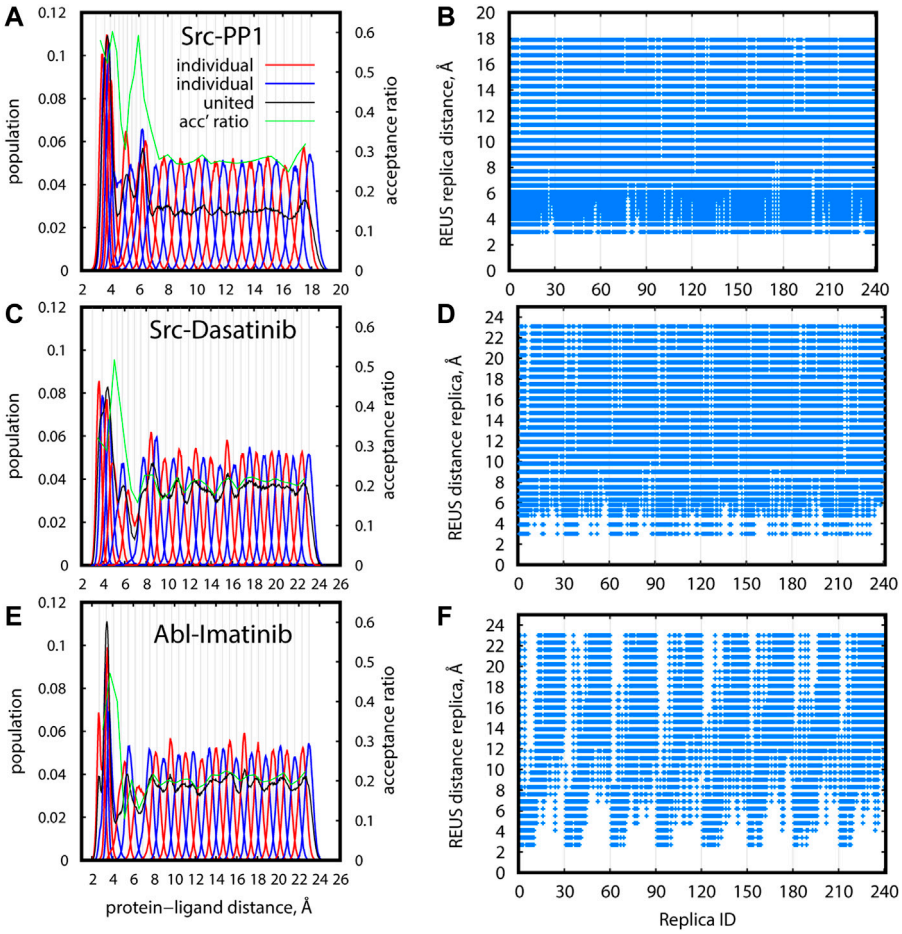


FIGURE 7 | Efficiency of sampling in REUS space for gREST/REUS simulations at 310 K for 500-ns Src-PP1 **(A,B)** 750-ns Src-Dasatinib **(C,D)**, and 1,000-ns Abl-Imatinib **(E, F)**. **(A)**, **(C)**, **(E)** Distribution of replicas according to their REUS distance. Distributions of adjacent individual replicas ("individual") are shown in alternating red/blue lines for better visibility. Distributions of all replicas ("united") are shown in black lines. Population values for "united" data were scaled to match the "individual" populations. Acceptance ratios between adjacent REUS replicas are shown in green lines. **(B)**, **(D)**, **(F)** REUS replicas visited at least once by individual replicas.

Figures 6C,D show the relative population at each solute temperature visited by individual replicas. Ideally, a uniform sampling is desired, in which each replica visits each temperature evenly. In a short simulation time (10 ns), all solute temperatures were already visited, although the replicas preferred the lowest and highest temperatures. After 500 ns, the sampling becomes more uniform, where the excessive population at the lowest and highest temperatures for some of the replicas flattens out. We found similar trends in the other two systems (**Supplementary Figure S2**), while the convergence becomes slower with increasing ligand size from PP1 to Dasatinib to Imatinib.

3.5.2 Random Walks in the REUS Dimension

The distribution of distance replicas along the REUS dimension at 310 K is shown in **Figures 7A,C,E** for Src-PP1, Src-Dasatinib, and Abl-Imatinib, respectively (**Supplementary Figure S3A** for Src-PP1-rev). All systems show a similar trend, where the population is large at short distances, drops once as the distance increases, and then converges to a constant value. Despite the drop in population, owing to the intensive tuning of the replica parameters, all regions were sampled to an acceptable extent, maintaining a constant overlap and good acceptance ratios between adjacent replicas. For example, in the case of Src-Dasatinib, the initial lack of population in the region of protein-ligand distances 4–8 Å was gradually filled with increasing sampling time, and converged at 250 ns (**Supplementary Figures S4A,C,E,G**). The lack of population is much less significant at higher temperature replicas (**Supplementary Figure S5**), indicating that two-dimensional replica exchanges improve the sampling at 310 K.

Figures 7B,D,F (**Supplementary Figure S3B**) demonstrate the random walks along the REUS dimension. Each of the 240 replicas visited all REUS distances almost perfectly for Src-PP1 and moderately for Src-Dasatinib. In contrast, for Abl-Imatinib, random walks in the vicinity of each region are rather good but the overall random walks are not as efficient, namely, replicas which started at small distances could not reach far distances and vice versa (**Figure 7F**). This suggests that a large and flexible ligand can be trapped in the vicinity of its starting configuration due to either specific or non-specific interactions with the protein.

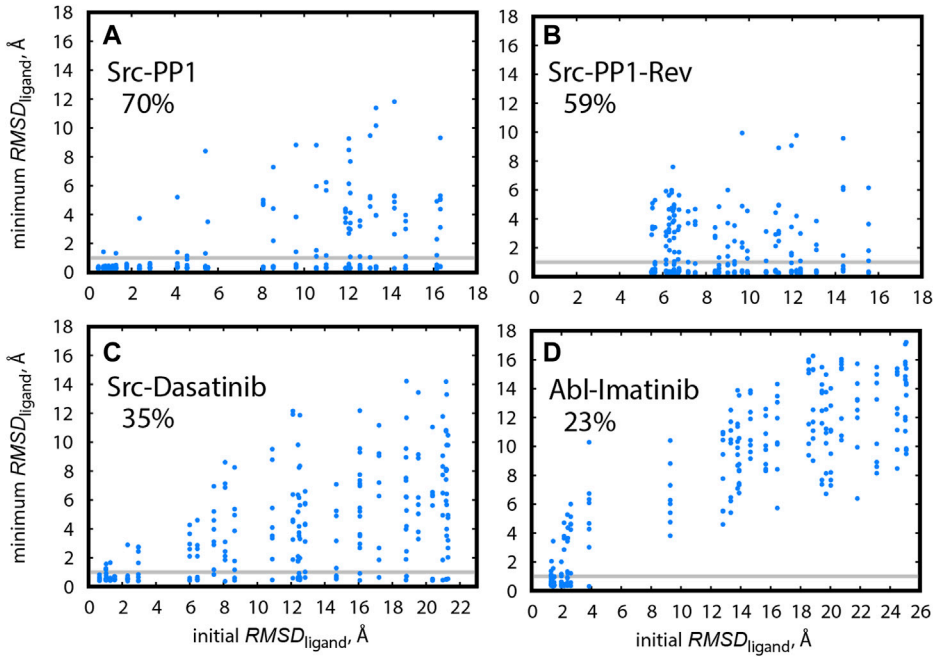


FIGURE 8 | Minimum $RMSD_{ligand}$ for replicas during the simulation, plotted against their initial $RMSD_{ligand}$, in **(A)** Src-PP1, **(B)** Src-PP1-Rev, **(C)** Src-Dasatinib, and **(D)** Abl-Imatinib. Grey horizontal lines mark $RMSD_{ligand}$ = 1 Å. The percentage of replicas that reached the bound pose is written for each system.

3.5.3 Finding the X-Ray Bound Pose in gREST/REUS Simulations

Finally, we compared the efficiencies of finding the X-ray bound pose for Src-PP1, Src-Dasatinib, and Abl-Imatinib. **Figure 8** shows the minimum $RMSD_{ligand}$ for individual replicas as a function of the initial $RMSD_{ligand}$ for all simulated systems. We define that a replica reached the bound pose if it had a $RMSD_{ligand}$ < 1 Å at least once during the simulation. The hit ratios along the sampling time are also summarized in **Supplementary Table S2**. For Src-PP1, 70% of the replicas, including those starting from far distances (large initial $RMSD_{ligand}$), found the X-ray bound pose (**Figure 8A**). Notably, the hit ratio was slightly low (59%) in the reverse pulling simulation (Src-PP1-Rev). We find that the initial $RMSD_{ligand}$ values are larger than 5 Å, indicating that the Src-PP1-Rev simulation did not include the bound pose, which is nearly identical to the X-ray crystal structure, in its initial structures. Nevertheless, many replicas starting from large $RMSD_{ligand}$ values found the bound pose within 500 ns simulations, demonstrating that the gREST/REUS method can efficiently find an unknown bound pose.

The hit ratio drops to 35% for Src-Dasatinib (**Figure 8C**). However, a fraction of replicas with an initial $RMSD_{ligand}$ of ~6 Å and above still finds the bound pose. For Abl-Imatinib (**Figure 8D**), which is the most challenging case, the hit ratio was only 23% even though its simulation time (1,000 ns) was the longest among the three systems. There is a gap in $RMSD_{ligand}$ values between ~4 Å and ~9 Å. Unlike the case of Src-Dasatinib, the replicas above ~9 Å cannot even reach the vicinity of the binding site (**Figure 8D**). Therefore, the hit ratio stays around 20% after 250 ns and until 1,000 ns (**Supplementary Table S2**). These results suggest that Imatinib binding is a considerably rare event and that Imatinib can be trapped at various locations in the vicinity of the binding region before fully entering deep inside the binding pocket. **Supplementary Movies S1–S3** show binding events for a single replica for Src-PP1, Src-Dasatinib, and Abl-Imatinib, respectively, and demonstrate the difference in the efficiency of finding the bound pose. Whereas for Src-PP1 the ligand binds and unbinds several times during 500 ns, for Src-Dasatinib, a single binding event of a replica that started from a far distance is observed after ~550 ns, and for Abl-Imatinib, a replica that started from an intermediate distance binds after ~250 ns and does not leave the binding site during the rest of the simulation time.

Here we followed the definition of Re et al. (2019) for hitting the bound pose, who deliberately set a strict cutoff of $RMSD_{ligand}$ < 1Å. We could set the cutoff slightly larger (for example 1.5 Å) to consider ligand fluctuations around the bound pose. In this case, we obtain hit ratios of 75, 65, 38, and 27% for Src-PP1, Src-PP1-rev, Src-Dasatinib, and Abl-Imatinib, respectively.

4 DISCUSSION AND CONCLUSION

In this work, we described a step-by-step procedure for obtaining the optimal parameter settings for efficient gREST/REUS simulations of protein-ligand binding. The protocol, which was demonstrated here for three kinase-inhibitor systems, was

validated through an extensive analysis of sampling efficiency based on a total of 660 μs of simulation time and can be applied to protein-ligand systems in general. We demonstrated that while the determination of gREST parameters is rather straightforward and nearly automatic, a particular care is needed in the determination of REUS parameters. First, a proper definition of the protein-ligand distance as the REUS CV, and second, careful tuning of replica space and force constants. Both of these practices can enhance the sampling efficiency. Taking care of these points, gREST/REUS simulations can sample binding events with high statistical accuracy and the obtained trajectories can be used to characterize binding poses and pathways on the free-energy landscape.

The use of protein-ligand distance as CV is a common practice for simulating binding events. Typically, the distance is determined using the COMs of the binding site and the ligand. For flexible ligands with molecular weight of few hundreds, as in the case of Imatinib, the determination of the CV significantly affects replica exchanges in REUS dimension. A lesson from this work is that each COM of the binding site and the ligand should be determined using multiple anchor sites for taking the flexibilities and orientation into account. This is because the flexible ligand can interact with the protein in different conformations and at different parts of the molecule. Even with a proper definition of the protein-ligand distance and well-tunned REUS parameters (replica spacing and force constants), the realization of constant acceptance ratios throughout the REUS dimension is quite dificult as shown for Imatinib. Here, we must add that applying too stiff umbrella potentials during the pulling simulations for obtaining the initial REUS replicas or during the REUS simulation may affect the obtained binding pathways. Thus, we must find the right balance of parameters that will not excessively bias the simulation but will still result in efficient sampling.

We showed that gREST/REUS can fill this gap with the aid of solute temperature exchange. Our results justify performing exchanges in two dimensions while using non-negligible computational resources. Using this protocol, protein-ligand binding simulations, in particular ligands or inhibitors of small or medium sizes, would be successfully performed on massively parallel supercomputers or GPU clusters.

Although good random walks in the replica space were observed in all three cases, simulation results of Abl-Imatinib suggest that efficient conformational sampling of Imatinib around the binding site of c-Abl kinase is still challenging. Unlike for Src-PP1 and Src-Dasatinib simulations, we could not observe many binding/unbinding events for Imatinib, especially of replicas initiated from far distances. Observing efficient random walks along the whole REUS range is important for visualizing the binding pathway. We learned that the problem is harder as the ligand size increased from PP1 (easy) to Dasatinib (moderate) to Imatinib (difficult), especially for obtaining the whole binding pathway. To further enhance the sampling for flexible ligands, consideration of a CV other than protein-ligand distance or an extension of the current scheme would be necessary. Considering the very slow unbinding rate of Imatinib, more drastic acceleration, such as simulations at higher solvent temperatures or enhancement of the c-Abl kinase domain motions might be introduced in the gREST/REUS simulations.

Another practical drawback of the gREST/REUS ligand-binding simulations is that huge computational resources are required for them. In this study, we used 240 replicas in the 2D-REMD for each of the three cases. Without the use of Fugaku or other massively parallel supercomputers, it is not easy to access such huge resources. One way to overcome the problem is to replace gREST or REUS with other enhanced sampling methods with less computational costs. We previously developed GaREUS (Gaussian accelerated replica-exchange umbrella sampling) (Oshima et al., 2019) by replacing gREST in gREST/REUS into GaMD (Miao and McCammon, 2017). We were able to significantly reduce the number of replicas using GaREUS while keeping the sampling strategy and efficiency, because GaREUS requires the same number of replicas as 1D-REUS. The use of such low-cost enhanced sampling methods is necessary for investigating molecular mechanisms for many other kinase-inhibitor binding processes.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

Conceptualization, YS; Methodology, AS, SR and YS; Investigation, AS; Resources, YS; Writing–Original Draft, AS; Writing–Review and Editing, AS, SR and YS; Funding Acquisition, YS; Supervision, YS.

FUNDING

We used the computational resources provided by the HPCI System Research Project (Project ID: hp200129, hp200135, hp210172, and hp210177) and those in RIKEN Advanced Center for Computing and Communication (HOKUSAI BigWaterfall). This work was supported by MEXT/JSPS KAKENHI Grant Number 19H05645 (to YS), 21H05249 (to YS), 19K12229 (to SR), RIKEN pioneering projects in "Biology of Intracellular Environment," "Dynamic Structural Biology," and "Glycolipidlogue" (to YS), MEXT "Program for Promoting Research on the Supercomputer Fugaku (Biomolecular dynamics in a living cell (JPMXP1020200101)/ MD-driven Precision Medicine (JPMXP1020200201))."

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.878830/full#supplementary-material

REFERENCES

Bernetti, M., Cavalli, A., and Mollica, L. (2017). Protein-Ligand (Un)binding Kinetics as a New Paradigm for Drug Discovery at the Crossroad between Experiments and Modelling. *Med. Chem. Commun.* 8 (3), 534–550. doi:10.1039/c6md00581k

Bruce, N. J., Ganotra, G. K., Kokh, D. B., Sadiq, S. K., 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

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* 126 (1), 014101. doi:10.1063/1.2408420

Case, D. A., Aktulga, H. M., Belfon, K., Ben-Shalom, I., Brozell, S. R., Cerutti, D. S., et al. (2021). *Amber 2021*. San Francisco: University of California.

Cowan-Jacob, S. W., Fendrich, G., Manley, P. W., Jahnke, W., Fabbro, D., Liebetanz, J., et al. (2005). The crystal Structure of a C-Src Complex in an Active Conformation Suggests Possible Steps in C-Src Activation. *Structure.* 13 (6), 861–871. doi:10.1016/j.str.2005.03.012

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 (12), 10089–10092. doi:10.1063/1.464397

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 (9), 4035–4061. doi:10.1021/acs.jmedchem.5b01684

Dickson, A., Tiwary, P., and Vashisth, H. (2017). Kinetics of Ligand Binding through Advanced Computational Approaches: a Review. *Curr. Top. Med. Chem.* 17 (23), 2626–2641. doi:10.2174/1568026617666170414142908

Dickson, A. (2018). Mapping the Ligand Binding Landscape. *Biophysical J.* 115 (9), 1707–1719. doi:10.1016/j.bpj.2018.09.021

Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., et al. (2011). Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc. Natl. Acad. Sci. U.S.A.* 108 (32), 13118–13123. doi:10.1073/pnas.1104614108

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., et al. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* 17 (2), 144. doi:10.3390/ijms17020144

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* 103 (19), 8577–8593. doi:10.1063/1.470117

Feenstra, K. A., Hess, B., and Berendsen, H. J. C. (1999). Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *J. Comput. Chem.* 20 (8), 786–798. doi:10.1002/(sici)1096-987x(199906)20:8<786::aid-jcc5>3.0.co;2-b

Fukunishi, H., Watanabe, O., and Takada, S. (2002). On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction. *J. Chem. Phys.* 116 (20), 9058–9067. doi:10.1063/1.1472510

Getlik, M., Grütter, C., Simard, J. R., Klüter, S., Rabiller, M., Rode, H. B., et al. (2009). Hybrid Compound Design to Overcome the Gatekeeper T338M Mutation in cSrc. *J. Med. Chem.* 52 (13), 3915–3926. doi:10.1021/jm9002928

Gobbo, D., Piretti, V., Di Martino, R. M. C., Tripathi, S. K., Giabbai, B., Storici, P., et al. (2019). Investigating Drug-Target Residence Time in Kinases through Enhanced Sampling Simulations. *J. Chem. Theor. Comput.* 15 (8), 4646–4659. doi:10.1021/acs.jctc.9b00104

Hénin, J., Lelièvre, T, Shirts, M. R., Valsson, O., and Delemotte, L. (2022). Enhanced Sampling Methods for Molecular Dynamics Simulations. arXiv preprint arXiv:2202.04164v1. doi:10.48550/arXiv.2202.04164

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of Multiple Amber Force fields and Development of Improved Protein Backbone Parameters. *Proteins.* 65 (3), 712–725. doi:10.1002/prot.21123

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 (2), 926–935. doi:10.1063/1.445869

Jung, J., Kasahara, K., Kobayashi, C., Oshima, H., Mori, T., and Sugita, Y. (2021). Optimized Hydrogen Mass Repartitioning Scheme Combined with Accurate Temperature/Pressure Evaluations for Thermodynamic and Kinetic Properties of Biological Systems. *J. Chem. Theor. Comput.* 17 (8), 5312–5321. doi:10.1021/acs.jctc.1c00185

Jung, J., Mori, T., Kobayashi, C., Matsunaga, Y., Yoda, T., Feig, M., et al. (2015). GENESIS: a Hybrid-Parallel and Multi-Scale Molecular Dynamics Simulator with Enhanced Sampling Algorithms for Biomolecular and Cellular Simulations. *Wires Comput. Mol. Sci.* 5 (4), 310–323. doi:10.1002/wcms.1220

Kamiya, M., and Sugita, Y. (2018). Flexible Selection of the Solute Region in Replica Exchange with Solute Tempering: Application to Protein-Folding Simulations. *J. Chem. Phys.* 149 (7), 072304. doi:10.1063/1.5016222

Kobayashi, C., Jung, J., Matsunaga, Y., Mori, T., Ando, T., Tamura, K., et al. (2017). *GENESIS 1.1: A Hybrid-parallel Molecular Dynamics Simulator with Enhanced Sampling Algorithms on Multiple Computational Platforms*. Wiley Online Library.

Kokubo, H., Tanaka, T., and Okamoto, Y. (2013). Two-dimensional Replica-Exchange Method for Predicting Protein-Ligand Binding Structures. *J. Comput. Chem.* 34 (30), 2601–2614. doi:10.1002/jcc.23427

Koneru, J. K., Sinha, S., and Mondal, J. (2019). In Silico reoptimization of Binding Affinity and Drug-Resistance Circumvention Ability in Kinase Inhibitors: a Case Study with RL-45 and Src Kinase. *J. Phys. Chem. B.* 123 (31), 6664–6672. doi:10.1021/acs.jpcb.9b02883

Lin, Y.-L., Meng, Y., Jiang, W., and Roux, B. (2013). Explaining Why Gleevec Is a Specific and Potent Inhibitor of Abl Kinase. *Proc. Natl. Acad. Sci. U.S.A.* 110 (5), 1664–1669. doi:10.1073/pnas.1214330110

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 (8), 1950–1958. doi:10.1002/prot.22711

Liu, P., Kim, B., Friesner, R. A., and Berne, B. J. (2005). Replica Exchange with Solute Tempering: A Method for Sampling Biological Systems in Explicit Water. *Proc. Natl. Acad. Sci. U.S.A.* 102 (39), 13749–13754. doi:10.1073/pnas.0506346102

Meng, Y., Lin, Y.-l., and Roux, B. (2015). Computational Study of the "DFG-Flip" Conformational Transition in C-Abl and C-Src Tyrosine Kinases. *J. Phys. Chem. B.* 119 (4), 1443–1456. doi:10.1021/jp511792a

Miao, Y., and McCammon, J. A. (2017). "Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications," in *Annual Reports in Computational Chemistry* (Elsevier), 231–278. doi:10.1016/bs.arcc.2017.06.005

Miyamoto, S., and Kollman, P. A. (1992). Settle: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* 13 (8), 952–962. doi:10.1002/jcc.540130805

Morando, M. A., Saladino, G., D'Amelio, N., Pucheta-Martinez, E., Lovera, S., Lelli, M., et al. (2016). Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the C-Src Kinase. *Sci. Rep.* 6 (1), 24439–9. doi:10.1038/srep24439

Nagar, B., Bornmann, W. G., Pellicena, P., Schindler, T., Veach, D. R., Miller, W. T., et al. (2002). Crystal Structures of the Kinase Domain of C-Abl in Complex with the Small Molecule Inhibitors PD173955 and Imatinib (STI-571). *Cancer Res.* 62 (15), 4236–4243.

Narayan, B., Buchete, N.-V., and Elber, R. (2021). Computer Simulations of the Dissociation Mechanism of Gleevec from Abl Kinase with Milestoning. *J. Phys. Chem. B.* 125 (22), 5706–5715. doi:10.1021/acs.jpcb.1c00264

Narayan, B., Fathizadeh, A., Templeton, C., He, P., Arasteh, S., Elber, R., et al. (2020). The Transition between Active and Inactive Conformations of Abl Kinase Studied by Rock Climbing and Milestoning. *Biochim. Biophys. Acta (Bba) - Gen. Subjects.* 1864 (4), 129508. doi:10.1016/j.bbagen.2019.129508

Niitsu, A., Re, S., Oshima, H., Kamiya, M., and Sugita, Y. (2019). De Novo prediction of Binders and Nonbinders for T4 Lysozyme by gREST Simulations. *J. Chem. Inf. Model.* 59 (9), 3879–3888. doi:10.1021/acs.jcim.9b00416

Oshima, H., Re, S., and Sugita, Y. (2020). Prediction of Protein-Ligand Binding Pose and Affinity Using the gREST+FEP Method. *J. Chem. Inf. Model.* 60 (11), 5382–5394. doi:10.1021/acs.jcim.0c00338

Oshima, H., Re, S., and Sugita, Y. (2019). Replica-exchange Umbrella Sampling Combined with Gaussian Accelerated Molecular Dynamics for Free-Energy Calculation of Biomolecules. *J. Chem. Theor. Comput.* 15 (10), 5199–5208. doi:10.1021/acs.jctc.9b00761

Paul, F., Thomas, T., and Roux, B. (2020). Diversity of Long-Lived Intermediates along the Binding Pathway of Imatinib to Abl Kinase Revealed by MD

Simulations. *J. Chem. Theor. Comput.* 16 (12), 7852–7865. doi:10.1021/acs.jctc.0c00739

Plattner, N., and Noé, F. (2015). Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* 6 (1), 7653–7710. doi:10.1038/ncomms8653

Re, S., Oshima, H., Kasahara, K., Kamiya, M., and Sugita, Y. (2019). Encounter Complexes and Hidden Poses of Kinase-Inhibitor Binding on the Free-Energy Landscape. *Proc. Natl. Acad. Sci. U.S.A.* 116 (37), 18404–18409. doi:10.1073/pnas.1904707116

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 23 (3), 327–341. doi:10.1016/0021-9991(77)90098-5

Schindler, T., Sicheri, F., Pico, A., Gazit, A., Levitzki, A., and Kuriyan, J. (1999). Crystal Structure of Hck in Complex with a Src Family-Selective Tyrosine Kinase Inhibitor. *Mol. Cell.* 3 (5), 639–648. doi:10.1016/s1097-2765(00)80357-3

Schuetz, D. A., de Witte, W. E. A., Wong, Y. C., Knasmueller, B., Richter, L., Kokh, D. B., et al. (2017). Kinetics for Drug Discovery: an Industry-Driven Effort to Target Drug Residence Time. *Drug Discov. Today.* 22 (6), 896–911. doi:10.1016/j.drudis.2017.02.002

Seeliger, M. A., Nagar, B., Frank, F., Cao, X., Henderson, M. N., and Kuriyan, J. (2007). c-Src Binds to the Cancer Drug Imatinib with an Inactive Abl/c-Kit Conformation and a Distributed Thermodynamic Penalty. *Structure.* 15 (3), 299–311. doi:10.1016/j.str.2007.01.015

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 (24), 9181–9183. doi:10.1021/ja202726y

Shekhar, M., Smith, Z., Seeliger, M., and Tiwary, P. (2021). Protein Flexibility and Dissociation Pathway Differentiation Can Explain Onset of Resistance Mutations in Kinases. *BioRxiv.*. doi:10.1101/2021.07.02.450932

Silva, D.-A., Bowman, G. R., Sosa-Peinado, A., and Huang, X. (2011). A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the Lao Protein. *Plos Comput. Biol.* 7 (5), e1002054. doi:10.1371/journal.pcbi.1002054

Sohraby, F., Javaheri Moghadam, M., Aliyar, M., and Aryapour, H. (2020). A Boosted Unbiased Molecular Dynamics Method for Predicting Ligands Binding Mechanisms: Probing the Binding Pathway of Dasatinib to Src-Kinase. *Bioinformatics.* 36 (18), 4714–4720. doi:10.1093/bioinformatics/btaa565

Spitaleri, A., Decherchi, S., Cavalli, A., and Rocchia, W. (2018). Fast Dynamic Docking Guided by Adaptive Electrostatic Bias: The MD-binding Approach. *J. Chem. Theor. Comput.* 14 (3), 1727–1736. doi:10.1021/acs.jctc.7b01088

Spitaleri, A., Zia, S. R., Di Micco, P., Al-Lazikani, B., Soler, M. A., and Rocchia, W. (2020). Tuning Local Hydration Enables a Deeper Understanding of Protein-Ligand Binding: The PP1-Src Kinase Case. *J. Phys. Chem. Lett.* 12 (1), 49–58. doi:10.1021/acs.jpclett.0c03075

Sugita, Y., Kitao, A., and Okamoto, Y. (2000). Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J. Chem. Phys.* 113 (15), 6042–6051. doi:10.1063/1.1308516

Sugita, Y., and Okamoto, Y. (1999). Replica-exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* 314 (1-2), 141–151. doi:10.1016/s0009-2614(99)01123-9

Terakawa, T., Kameda, T., and Takada, S. (2011). On Easy Implementation of a Variant of the Replica Exchange with Solute Tempering in GROMACS. *J. Comput. Chem.* 32 (7), 1228–1234. doi:10.1002/jcc.21703

Tiwary, P., Mondal, J., and Berne, B. J. (2017). How and when Does an Anticancer Drug Leave its Binding Site? *Sci. Adv.* 3 (5), e1700014. doi:10.1126/sciadv.1700014

Tran, D. P., Hata, H., Ogawa, T., Taira, Y., and Kitao, A. (2020). PaCS-MD/MSM: Parallel Cascade Selection Molecular Dynamic Simulation in Combination with Markov State Model as an Efficient Non-bias Sampling Method. *Ensemble.* 22 (2), 151–156. doi:10.11436/mssj.22.151

Tuckerman, M., Berne, B. J., and Martyna, G. J. (1992). Reversible Multiple Time Scale Molecular Dynamics. *J. Chem. Phys.* 97 (3), 1990–2001. doi:10.1063/1.463137

Valsson, O., Tiwary, P., and Parrinello, M. (2016). Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* 67, 159–184. doi:10.1146/annurev-physchem-040215-112229

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and Testing of a General Amber Force Field. *J. Comput. Chem.* 25 (9), 1157–1174. doi:10.1002/jcc.20035

Wang, L., Friesner, R. A., and Berne, B. J. (2011). Replica Exchange with Solute Scaling: a More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B.* 115 (30), 9431–9438. doi:10.1021/jp204407d

Yang, L.-J., Zou, J., Xie, H.-Z., Li, L.-L., Wei, Y.-Q., and Yang, S.-Y. (2009). Steered Molecular Dynamics Simulations Reveal the Likelier Dissociation Pathway of Imatinib from its Targeting Kinases C-Kit and Abl. *PLoS One.* 4 (12), e8470. doi:10.1371/journal.pone.0008470

Zhang, Q., Zhao, N., Meng, X., Yu, F., Yao, X., and Liu, H. (2022). The Prediction of Protein-Ligand Unbinding for Modern Drug Discovery. *Expert Opin. Drug Discov.* 17 (2), 191–205. doi:10.1080/17460441.2022.2002298

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 05 May 2022
doi: 10.3389/fmolb.2022.877170

Unexpected Dynamic Binding May Rescue the Binding Affinity of Rivaroxaban in a Mutant of Coagulation Factor X

Zhi-Li Zhang[1†], *Changming Chen*[2†], *Si-Ying Qu*[1], *Qiulan Ding*[2,3*] *and Qin Xu*[1*]

[1]State Key Laboratory of Microbial Metabolism & Joint International Research Laboratory of Metabolic and Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2]Department of Laboratory Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [3]Collaborative Innovation Center of Hematology, Shanghai Jiao Tong University School of Medicine, Shanghai, China

A novel coagulation factor X (FX) Tyr319Cys mutation (Y99C as chymotrypsin numbering) was identified in a patient with severe bleeding. Unlike the earlier reported Y99A mutant, this mutant can bind and cleave its specific chromogenetic substrate at a normal level, suggesting an intact binding pocket. Here, using molecular dynamics simulations and MM-PBSA calculations on a FX-rivaroxaban (RIV) complex, we confirmed a much stronger binding of RIV in Y99C than in Y99A on a molecular level, which is actually the average result of multiple binding poses in dynamics. Detailed structural analyses also indicated the moderate flexibility of the 99-loop and the importance of the flexible side chain of Trp215 in the different binding poses. This case again emphasizes that binding of ligands may not only be a dynamic process but also a dynamic state, which is often neglected in drug design and screening based on static X-ray structures. In addition, the computational results somewhat confirmed our hypothesis on the activated Tyr319Cys FX (Y99C FXa) with an impaired procoagulant function to bind inhibitors of FXa and to be developed into a potential reversal agent for novel oral anticoagulants (NOAC).

Keywords: coagulation factors, rivaroxaban, molecular dynamics simulations, molecular flexibility, structure-based drug design

OPEN ACCESS

Edited by:
Yinglong Miao,
University of Kansas, United States

Reviewed by:
Ferran Feixas,
Universitat de Girona, Spain
Yu-ming Huang,
Wayne State University, United States

***Correspondence:**
Qiulan Ding
qiulan_ding@shsmu.edu.cn
Qin Xu
xuqin523@sjtu.edu.cn

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

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *16 February 2022*
Accepted: *06 April 2022*
Published: *05 May 2022*

Citation:
Zhang Z-L, Chen C, Qu S-Y, Ding Q and Xu Q (2022) Unexpected Dynamic Binding May Rescue the Binding Affinity of Rivaroxaban in a Mutant of Coagulation Factor X. Front. Mol. Biosci. 9:877170. doi: 10.3389/fmolb.2022.877170

INTRODUCTION

In the coagulation cascade, the key position of coagulation factor X (FX) where the intrinsic and the extrinsic pathway merge into the common pathway makes it an ideal target to develop anticoagulants (Davie et al., 1991; Al-Obeidi and Ostrem, 1999; Lee and Player, 2011). Advances in crystallography have boosted the screening and design of synthetic inhibitors targeting on activated FX (FXa), which successfully resulted into novel oral anticoagulants (NOACs) approved by FDA, such as apixaban, rivaroxaban, and edoxaban (Perzborn et al., 2011; Wong et al., 2011; Wang et al., 2016). The X-ray structures of the FXa-inhibitor complexes showed that this type of anticoagulant can directly bind to the S1 pocket and S4 pocket at the same time (Maignan et al., 2003; Nazare et al., 2005). The former is a conserved pocket in the catalytic serine protease domains of coagulation factors, which accommodates the P1 residue of the peptide bond to be cleaved. The conserved Asp189 deep in the bottom of the pocket could provide strong electrostatic interaction with substrates and inhibitors (Katz et al., 2000; Hedstrom, 2002). At the same time, different from other coagulation factors with a

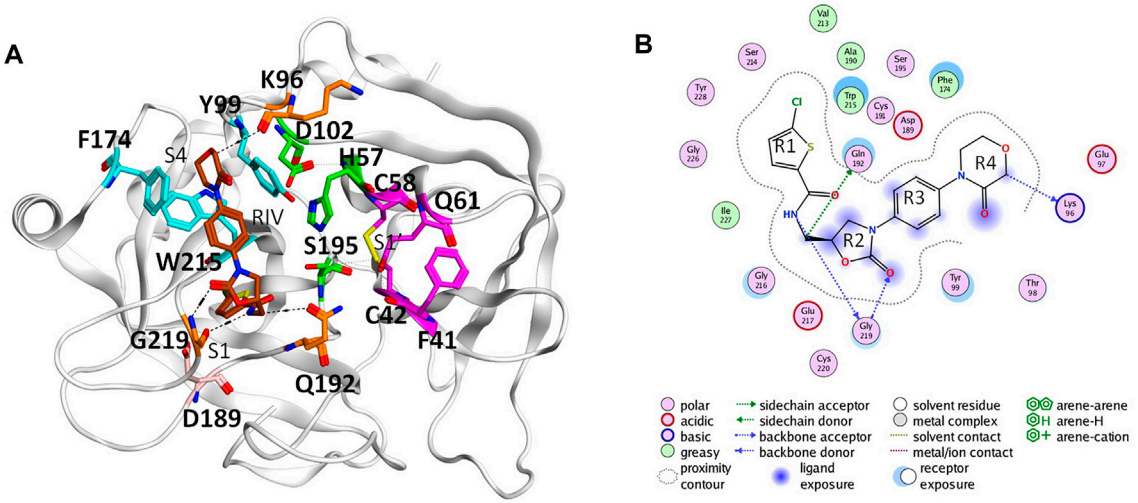


FIGURE 1 | Binding of rivaroxaban (RIV) in the active site of coagulation factor X. **(A)** Three-dimensional structure from the Protein Data Bank (2W26). The carbon atoms of RIV are colored in brown, while those of the key residues in the S1 pocket and S1' pocket are colored in magenta and those of the key residues in the S4 pocket are colored in cyan. Other residues concerning RIV-FX binding are colored in orange, with the catalytic triad in green and D189 in pink. **(B)** Two-dimensional plot of the binding of RIV with the interactions with the surrounding residues illustrated later. For convenience, the four rings of RIV are named as R1 to R4 from the chlorothiophene moiety to the morpholinone moiety.

serine protease domain, its distinctive hydrophobic S4 pocket provides an ideal target site for inhibitors to bind specifically (Wang et al., 2012). For example, the widely used NOAC rivaroxaban (RIV) has its chlorothiophene moiety (R1) and the morpholinone moiety (R4) binding to the S1 and S4 pocket of FXa, respectively (Roehrig et al., 2005) (**Figure 1**).

As the first approved NOAC, rivaroxaban has been increasingly used in clinical practice for treatment or prevention of thromboembolism. However, patients taking NOACs may present with major bleeding or need for management of an urgent unplanned bleeding challenge, and the best method to stop bleeding is the use of NOACs reversal agents to bind excessive NOACs (Kaatz et al., 2012; Samama, 2013; Samama et al., 2013). Until 2018, only one reversal agent for apixaban and rivaroxaban was approved by the Food and Drug Administration (FDA) (Escolar et al., 2017; Sartori and Cosmi, 2018; Carpenter et al., 2019), which is a recombinant FXa named andexanet alfa. However, in the ANNEXA-4 study, thrombotic events occurred in 18% of the patients in the safety population (Connolly et al., 2016), which is likely related to the binding of andexanet alfa to tissue factor pathway inhibitor (TFPI) (Ersayin et al., 2017). In the same way, different mutants of FXa were in development for more efficient reversal agents. For example, Verhoef et al. (2017) designed recombinants of FXa with either point mutation in the S4 pocket or fragment modifications on the 99-loop and compared their potentialities as reverse agents for NOACs with combined computational and biochemistry approach.

In addition to the fancy crystal structures of coagulation factors, molecular dynamics (MD) simulation also provided important understandings of the dynamics in FXa and its complex with different binding substrate/ligands. Daura et al. (2000) simulated the catalytic domain of FXa in aqueous solution and suggested possible hydrogen bonding of the active site residues with a substrate or inhibitor. Later, more simulations unveiled possible conformational changes in zymogen activation (Camire, 2002; Venkateswarlu et al., 2002) and in conformational transitions of open/closed states of the binding pocket (Singh and Briggs, 2010; Wang et al., 2012). MD simulations also suggested the importance of flexibilities to the catalytic activity impacted by mutations in the S4 pocket and S1 pocket (Abdel-Azeim et al., 2014) as well as the N-terminus of the serine protease domain (Li et al., 2019). The dynamics lying behind drugs targeting the S1 and S4 binding sites of FX were also implemented, such as edoxaban, betrixaban (Du et al., 2019), and rivaroxaban (Qu et al., 2019).

In this work, a novel FX Tyr319Cys mutation identified in a patient with severe bleeding diathesis was reported, whose activity for prothrombin activation is completely impaired but the cleavage on specific chromogenetic substrate is normal. A molecular model of the activated FX (FXa) with corresponding Y99C mutation in complex with rivaroxaban was then analyzed by molecular dynamics simulations. Unlike the Y99A mutant in which RIV is quickly released, this Y99C mutant is more like F174A and can have RIV dynamically bound with multiple patterns and unexpected strong affinity (Qu et al., 2019; Qu and Xu, 2019). Therefore, we presume that the activated Tyr319Cys FX (Y99C FXa) mutant may not have a procoagulant function but may have the ability to bind the NOAC rivaroxaban and the potential to be developed into a novel reversal agent.

MATERIALS AND METHODS

Blood Sampling

The peripheral blood was collected *via* venipuncture into tubes containing sodium citrate (final concentration 0.38%), followed by double centrifugation at 3,000 g for 15 min to obtain platelet-

TABLE 1 | Primers for *F10* amplification.

Exon	Forward 5'-3'	Backward 5'-3'
Exon 1	GTGGTCACTCCCCTGCCTCG	TGCTGTGCCCCTCGTCCTG
Exon 2	TGAGGGTGACCAGAGCTTTT	CTGTGGCCTGAGCTCCTTAC
Exon 3	TAAGATGACTGAAGCCACAT	CTATTATGGAAACACCCTGA
Exon 4	GAAACAGCTTGCAGACTCCAG	CTTCAGGGGCATCTGATCT
Exon 5	CCTTTGCTCAACCCAATGGC	TGGTGTCACTGTTACCTGCC
Exon 6	TATGGGGAGCCTCTCTCTGT	CAGGTGGTCTCTCCAGCAG
Exon 7	TGGCACAGGCAGAGAAAAGA	CCTCTGTGAAATGCCCCTAA
Exon 8	GATGTGCGAGAGCATGTCC	GGCAATCGAGAGACAAACCA

poor plasma (PPP). Normal pooled plasma (NPP) was prepared from 30 healthy donors. The study was approved by the Institutional Review Board of Ruijin Hospital. All related individuals gave their informed consent to participate.

Hemostatic Assays

The FX clotting activity (FX:C) was measured using the activated partial thromboplastin time (aPTT) and pro-thrombin time (PT) pathway-based coagulation function assays on the ACL-TOP automatic coagulometer (Instrumentation Laboratory). The plasma FX was activated to activate FX (FXa) by Russell's viper venom (RVV-X) (Haematologic Technologies Inc., Essex Junction, VT, United States), and its enzymatic activity was determined using specific chromogenic substrate S2765 (Hyphen-Biomed, Neuville-Sur-Oise, France). The antigen level of FX (FX:Ag) was measured using an enzyme-linked immunosorbent assay (ELISA) kit (Enzyme Research Laboratories, South Bend, United States).

Genetic Analysis of F10

Genomic DNA was extracted from peripheral whole blood using the QIAamp DNA blood purification kit (Qiagen, Hilden, Germany). The coding sequences and flank regions of F10 gene were amplified by polymerase chain reaction and sequenced. The primers used are listed in **Table 1**.

Modeling

The original molecular model of wild type FXa-RIV complex was based on the structure with ID as 2W26 in the Protein Data Bank (Roehrig et al., 2005), in which only the heavy chain of FXa, the ions, and crystallographic waters close to it were kept so as to reduce the size of the system (Abdel-Azeim et al., 2014). Tyr99 was selectively mutated into Cys and Ala by PyMOL as the initial mutant models. PyMOL (Schrodinger, 2015) was also used for visualization of the 3-D structures for both the initial models and the representative conformations from the simulations given later, while the 2-D plot of the RIV binding to critical residues was obtained by ProteinPlus (Fricker et al., 2004).

Molecular Dynamics Simulations

The procedure of molecular dynamics simulations is fairly same as that of our earlier studies (Li et al., 2019; Qu et al., 2019; Qu and Xu, 2019). The package of GROMACS5.1.2 was used for the molecular dynamics simulations (Abraham et al., 2015). The force field for FXa was using CHARMM36 (Klauda et al., 2010), that for water was TIP3P (Jorgensen et al., 1983), and the parameters of RIV bound to the protein were generated by CHARMM General Force Field (CGenFF) (Vanommeslaeghe et al., 2010; Vanommeslaeghe et al., 2011; Vanommeslaeghe et al., 2012). The two calcium ions in 2W26 were retained in the topology and described by the default parameters of CHARMM36. A total of four pairs of disulfide bridges (Cys22-Cys27, Cys42-Cys58, Cys168-Cys182, and Cys191-Cys220) were defined as linked in the topology. The protonated states of all the residues were determined by H++ (Anandakrishnan et al., 2012) at PH value = 7.0 with the water model of TIP3P followed by manual checking (sequence of Y99C mutant is shown as an example in **Supplementary Figure S1**), from which His83 was set as the "HID" in the topology while other histidines (His57, His91, His145, and His199) were set as the "HIE." The models were solvated in a cubic SPC216 water box (Berendsen et al., 1981) with the dimensions as 6.90 nm × 6.96 nm × 5.91 nm. The solvated systems were neutralized by adding five chloride ions, with the total number of atoms about 28,200. These systems were then energy-minimized by the steepest descent algorithm until the maximum force was lower than 1,000.0 kJ/mol/nm and then equilibrated with 100 ps NVT ensemble at 310 K and 100 ps NPT ensemble at 310 K and 1 atm, where the Nosé–Hoover weak coupling algorithm (Hoover, 1985) was used for temperature

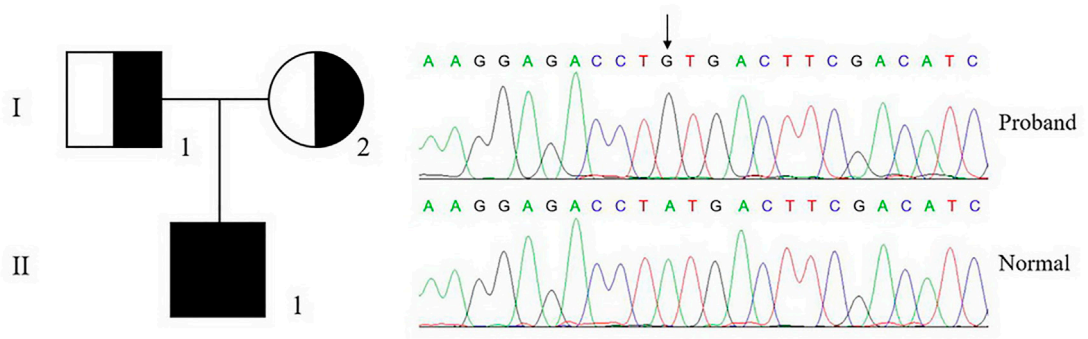


FIGURE 2 | Genetic analysis of proband and pedigree. The genetic analysis shows that the proband (II-1) carries a homozygous c.956a > g, p.Tyr319Cys mutation in F10, which is inherited from his parents.

maintenance and the Parrinello–Rahman barostat methodology (Parrinello and Rahman, 1981) was used to keep the pressure. At the same time, the particle mesh Ewald (PME) (Perera et al., 1995) method was used for long-range electrostatic interactions and the Linear Constraint Solver algorithm (LINCS) (Hess et al., 1997) was used to allow the integration step as 2 fs. After the NVT and NPT equilibrations, at least 200 ns further simulations were performed and collected three times under the same NPT condition.

Calculation of the Binding Free Energy

The g_mmpbsa (Kumari et al., 2014) package of GROMACS was used to calculate the binding free energy between RIV and FXa using the Molecular Mechanistic Poisson–Boltzmann Surface Area (MM-PBSA) method. In this work, the molecular mechanistic (MM) energies were only considered the electrostatic interactions and the van der Waals interactions. The solvation energies including the polar interactions were calculated by the Poisson–Boltzmann method and the nonpolar interactions were empirically estimated by the exposed area (SASA) using the surface tension coefficient γ = 0.0072 kcal/(mol•Å2). The entropy contribution was not included in this calculation for simplicity (Chen et al., 2013). In this way, the total binding free energy is:

$$\Delta G_{bind} = \Delta E_{vdw} + \Delta E_{coulomb} + \Delta G_{polar} + \Delta G_{nonpolar}$$

Due to the highly dynamic behavior of RIV observed in the mutants, we performed the calculation of the binding free energy in different methods for different purposes. For the overall comparison between the binding energies between the three systems (WT, Y99C, and Y99A), we concatenated the three trajectories of 200 ns MD simulations in each system to obtain a sampling set of 600 ns in total and picked out frames by intervals of 0.2 ns (3,000 frames in total for each system) for further calculation on the binding affinity of RIV. For calculation of the binding energies of the multiple binding modes in the Y99C mutant, the 3,000 frames were clustered as described later, and the frame sets of the top three clusters were collectively used for calculation of the RIV binding energy. Similarly, the MM-PBSA calculations were also applied to the frame sets of the top three clusters obtained separately from each 1,000 frames of the 200 ns trajectories of the Y99C mutant so as to explore the possible influence from different samplings between the three repeats.

Analyses on Simulation Trajectories

The g_rms and g_rmsf package of GROMACS were used for calculation of the root-mean-square deviations (RMSDs) of selected atoms and the root-mean-square fluctuations (RMSFs) of the backbone atoms of the selected residues from Asn92 to Ile103 as the 99-loop, respectively. The RMSD of RIV or of the 99-loop backbone was also used for clustering the RIV binding modes or the conformations of the 99-loop, respectively, by the g_cluster package using the method illustrated by Daura et al. (1999). The frames for clustering were obtained from the trajectories with 0.2 ns interval. The cut-off for clustering was empirically selected as 0.17 nm based on the results of all the three systems. The package of "hydrogen bonds" in VMD (visual molecular dynamics) (Humphrey et al., 1996) was used to count possible hydrogen bonds between RIV and FXa, where any time a donor atom and an acceptor atom is less than 3.5 Å in distance and the angle of donor-H. . .acceptor is less than 30°, a hydrogen bond is counted. For example, when a residue forms two hydrogen bonds with RIV at the same time, no matter with side chain or backbone, the frequency is counted twice, and its overall occupancy is the sum of frequency involving this residue divided by the sum of frames from all three repeats of simulations. At last, the dynamic distribution of the R4 group of RIV was visualized by the positions of the C3 atom on it for simplicity.

RESULTS

Clinical Results

The genetic analysis of a patient (II-1) with severe bleeding diathesis identified a homozygous c.956a > g mutation in F10, which was inherited from his heterozygous father (I-1) and mother (I-2), who carries only one copy of this mutation on this allele (**Figure 2**).

This mutation is leading to a p.Tyr319Cys mutation in FX or the Y99C mutation in FXa by chymotrypsin numbering. The plasma FX antigen levels (FX:Ag) of both the proband (II-1) and his parents (I-1 and I-2) are almost normal as about 100% of normal control. The clotting time was prolonged in both APTT and PT tests. The FX clotting activity (FX:C) was determined by activated partial thromboplastin time (APTT) and prothrombin time (PT), which was completely impaired in the proband as 1.1%–1.4% of normal control, and more importantly, lowered to around 50% in the heterozygotes. However, the proteolytic activity to the chromogenic substrate S2765 was almost unchanged, all around 100% of the normal control (**Table 2**). It is reasonable to assume an intact active site in the Y99C mutant, at least for small substrates/inhibitors.

Comparison Between the Rivaroxaban Bindings in Different FXa Systems

Combining the three repeats of 200 ns simulations into one trajectory, the overall binding free energies of RIV to wild type (WT), Y99C, and Y99A mutants were calculated by MM-PBSA, with the contributions from van der Waals interactions, electrostatic interactions, and polar and nonpolar interactions with solutions compared in **Table 3**. From the total binding energies, it was found that the overall binding of RIV–Y99C (−83.018 ± 1.770 kJ/mol) is surprisingly even stronger than that of wild type FXa (−59.450 ± 2.011 kJ/mol), while the binding in Y99A mutant is much weaker. The unexpected stronger binding in Y99C is mainly from van der Waal interactions, where the binding is improved from −132.144 ± 2.761 to −162.162 ± 2.531 kJ/mol, although the polar interactions with solutions offset this improvement by about 10 kJ/mol. On the other side, the binding of RIV to Y99A is weakened in all the four terms of interactions since RIV was released from the binding site to the solution in a major part of the simulations.

TABLE 2 | Clotting function and genetic profile of the pedigree.

	FX:C (%)—Clotting activity		FX:C (%)—Chromogenic activity	FX:Ag (%)	*F10* mutation
	APTT	PT			
Proband	1.1	1.4	98.0	97.8	Tyr99Cys
I-1	55.2	49.8	101.2	103.9	Tyr99Cys (Het)
I-2	52.7	51.6	99.5	97.3	Tyr99Cys (Het)
Reference	50–150	50–150	50–150	50–150	

APTT, activated partial thromboplastin time; PT, pro-thrombin time.

TABLE 3 | Comparison between the binding free energies of RIV in the wild type (WT), Y99C, and Y99A mutants of coagulation factor X.

Energy (KJ/mol)	WT	Y99C	Y99A
van der Waal energy	−132.144 ± 2.761	−162.162 ± 2.531	−75.749 ± 2.435
Electrostatic energy	−28.991 ± 1.134	−28.291 ± 0.924	−16.009 ± 0.778
Polar solvation energy	115.871 ± 2.565	124.552 ± 2.347	72.579 ± 2.596
Nonpolar solvation energy	−14.250 ± 0.294	−17.069 ± 0.256	−8.598 ± 0.266
Total binding energy	−59.450 ± 2.011	−83.018 ± 1.770	−27.605 ± 2.513

Residual contributions to the RIV binding were also analyzed. As shown in **Supplementary Table S1**, in wild type, only the three key residues of S4 pocket: Tyr99, Phe174, and Trp215 contribute more than 2 kJ/mol to the binding of RIV, and in all these three residues, Trp215 may contribute the most, same as our earlier analyses (Qu and Xu, 2019). In the Y99C mutant, the side chain of residue 99 is changed from aromatic tyrosine into a polar but much shorter cysteine. It was expected that the aromatic cage of the S4 pocket is damaged and its hydrophobic interactions with RIV would be weakened, as observed in Y99A. However, although the contribution from Cys99 (−2.916 ± 0.100 kJ/mol) is relatively a little lower than Tyr99 in wild type (−3.205 ± 0.149 kJ/mol), the stronger interaction from Trp215 (−8.497 ± 0.174 kJ/mol in Y99C versus −6.453 ± 0.178 kJ/mol in wild type) resulted in even stronger binding of RIV. In addition, another residue also contributes more than 2 kJ/mol to RIV binding: Val213 turns to contribute −2.204 ± 0.076 kJ/mol. This residue is close to Trp215 but on the edge between the S4 and S1 pocket, which may suggest more fluctuations in RIV binding pose in Y99C. On the other hand, in the Y99A mutant, the contributions from all the three key residues are much lowered, partly because RIV is released from the binding site. However, even in our earlier study where only frames before RIV were released and used from binding free energy calculations, the contributions of Ala99 and Trp215 are both about 5 kJ/mol lower than those in wild type. The importance of residues in the three trajectories are quite different, which is consistent with the observation that RIV is released from the binding pocket and may transiently bind to different positions on the surface of FXa.

According to the binding free energy analyses, the main difference between the RIV binding in wild type and the mutants comes from the hydrophobic interactions, where deformation of the S4 pocket was expected to be caused by the mutations (Qu and Xu, 2019). Comparisons of the global RMSF clearly show that the difference in backbone flexibilities mainly happens in the 99-loop, where the flexibility in Y99A mutant is much higher than that in the other two systems (**Supplementary Figure S2**). As compared in **Figure 3B**, the average RMSF of all the residues on the 99-loop is a little increased in the Y99C mutant from about 0.5 Å to about 0.8 Å, but greatly increased to about 1.5 Å in the Y99A mutant. More details of each trajectory are shown in **Supplementary Figure S3**. The RMSD and RMSF of the 99-loop in wild type are not only stable, but also similar in all three trajectories. In Y99C, only those of trajectory III are as stable as wild type. In trajectory II, RMSD of the 99-loop is increased to 3 Å after 100 ns of simulations, resulting in a little higher RMSF as 0.5–0.8 Å. However, in trajectory I, the value of RMSD is quickly increased to about 3 Å after a short time, and part of the 99-loop (Phe94) could have RMSF as high as 1.8 Å. Consistently, the cross-correlation analysis of Cα atoms of the Y99C mutant indicates anti-correlated movements of the 99-loop with the two segments surrounding the S4 pocket, while no notable correlations are detected with other regions (**Supplementary Figure S4**), suggesting that the S4 pocket is not totally deformed. In all the three trajectories of Y99A, RMSD is quickly increased and fluctuates acutely from 2.5 to 4.5 Å, and RMSF varies from 1.0 to 2.0 Å, which is quite different both between the residues and between the trajectories. The representative structure of the top cluster of the trajectories is shown in **Figure 3A**. In two trajectories of WT and Y99C, the orientations of the side chains of Y99, Phe174, and Trp215 are roughly similar around the S4 pocket, and the backbone of the 99-loop is in comparable positions. However, in the trajectories II of WT and I of Y99C, the side chain of Trp215 flips to the side of S1, and the side chain of Phe174 in the former and the backbone of the 99-loop in the latter deviates away, leaving the S4 pocket open. At last, in Y99A, the side chains of Trp215 flip to the side of the S1 pocket in all three trajectories, while Phe174 and the 99-loop are rather deviated from the S4 pocket, which is too loose to bind RIV stably.

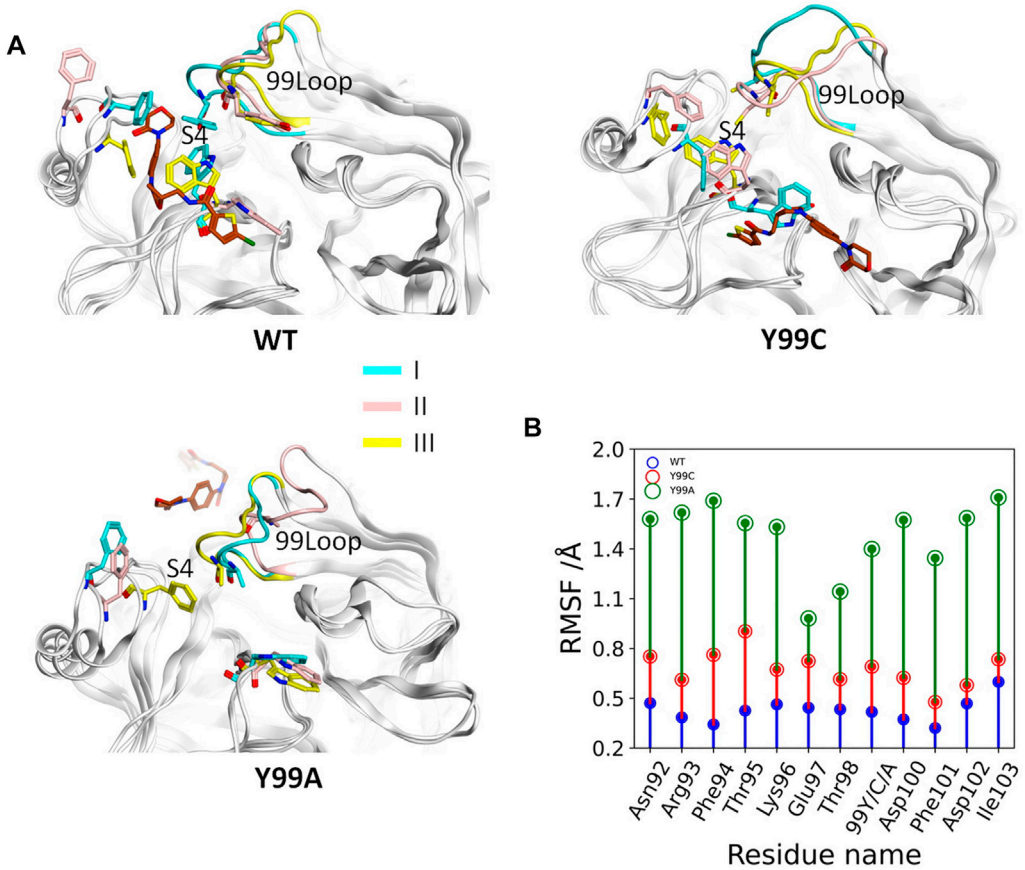


FIGURE 3 | Comparison of the 99-loop fluctuation in WT, Y99C, and Y99A mutant of coagulation factor X. **(A)** Representative conformations of the top cluster from the three trajectories in three systems, with RIV (brown), the backbone of 99-loop, and the residues 99, Phe174, and Trp215 shown and colored. **(B)** Time evolution of average RMSF of 99-loop residues in WT (blue), Y99C (red), and Y99A (green) mutant of FX.

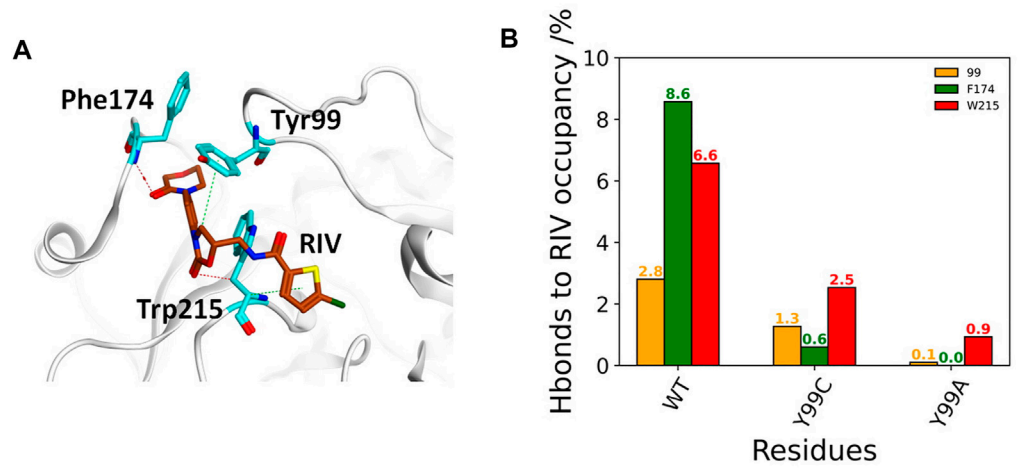


FIGURE 4 | Comparison of FX residue hydrogen-bonding with RIV in WT, Y99C, and Y99A mutant. **(A)** Illustration of the H-bonding (red dash) and H-π interaction (green dash) of RIV with the key residues in the S4 pocket of the wild-type system, where RIV has the carbon atoms colored in brown while those of the residues were colored in cyan. **(B)** Different H-bonding occupancies of the three key residues Y/C/A99 (in orange), F174 (in green), and W215 (in red) with RIV.

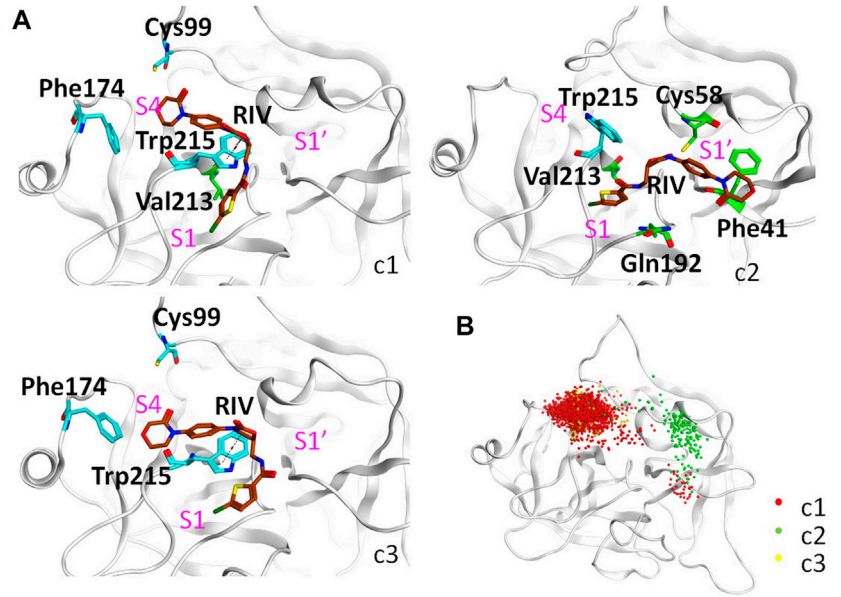


FIGURE 5 | Detailed analysis on the binding of RIV in the Y99C mutant. **(A)** Representative binding modes of the top three clusters. RIV carbon atoms are colored in brown, the S4 pocket residues are shown in cyan, other residues important to RIV binding are shown in green, and the positions of the three pockets are labeled in magenta. **(B)** Distributions of the R4 group are visualized by the positions of the C3 atom of RIV.

The deformation in the S4 pocket may affect not only the hydrophobic interactions but also the electrostatic interactions between RIV and FXa, as illustrated in **Figure 4A**. The total occupancies of H-bonds involving the key S4 pocket residues in all simulations are summarized in **Figure 4B**. In the stable S4 pocket of wild type FXa, RIV may form H-bonds with Tyr99, Phe174, and Trp215 as high as 2.8%, 8.6%, and 6.6% of the time, respectively. It is surprising that rearrangement of the S4 pocket induced by the Y99C mutation only reduces the possibility for RIV to form H-bonds with the residue 99 and Trp215 by about 60%, although the most stable H-bonds with Phe174 in WT are almost lost. At last, the occupancies of H-bonds with Trp215 is greatly reduced and hardly found with Ala99 or Phe174. The residues possibly H-bonding with RIV with more than 2% of the simulation time are all listed in **Supplementary Table S2**. In all three systems, the residues Gln192, Gly216, and Gly219 at the entry of the S1 pocket are always of the top ones, and their occupancies are the only ones more than 10% in both WT and Y99C. In WT, the next important residues with the occupancies higher than 5% are Phe174 and Trp215 in the S4 pocket. However, their occupancies are much lowered in Y99C, with only Trp215 as low as 2.5%. Instead, residues around the catalytic pocket, such as Ser195, Lys96, and His57, have elevated occupancies as 8.2%, 5.8%, and 5.3%, respectively. This difference may suggest a shift of RIV binding poses. In addition to increased frequency of H-bonding between RIV and His57, the introduced Cys99 may also proximate to His57 and interfere with its H-bond with Asp102 considering the high flexibility of the 99-loop or even act as a nucleophile as in a cysteine protease. However, at least in this RIV-bound model, the time evolution of the distance between H57 [ND1] and C99 [HG1] shows that the average distances (9.7, 8.4, and 6.4 Å) in the three trajectories of Y99C are too far for the proton transfer between these two residues accompanied with a possible nucleophilic attack by Cys99 (**Supplementary Figure S5**). At last, almost all residues in Y99A have relatively lower possibility to form H-bonds with RIV, which is consistent with the fact that RIV is rather released. However, Lys62 on the edge of the S1' pocket is the only residue with occupancy increased to 8.1%, which may suggest that the half-released RIV is easier to bind to the half-exposed S1' pocket, similar to what we discussed on the F174A mutant earlier (Qu et al., 2019).

Analyses on the Dynamic Binding in the Y99C Mutant

As compared previously, the binding of RIV in the Y99C mutant is neither as stable as in wild type FXa nor as totally released as in Y99A. The deformed S4 pocket, fluctuated 99-loop, and different interactions may suggest a dynamic binding in Y99C. Therefore, all the simulations of the Y99C mutant were combined together for clustering on the RMSD of RIV so as to figure out its most possible binding poses. MM-PBSA calculations were applied to the frame sets of the top three clusters to estimate the affinities in different binding modes.

As shown in **Figure 5A** and **Table 4**, the top three clusters have already exhibited alternative binding patterns. The top cluster (c1) occupies 83% of the population, where the R1 and R4 group of RIV still stay around the S1 and S4 pocket, respectively. However, this binding pattern is quite dynamic and not as stable as that in the rigid S1S4 binding in the crystal structure of wild type FXa. A major difference is the flip of the

TABLE 4 | Top three binding modes in the Y99C mutant.

	R1	R4	Frequencies (%)	Total binding energy FX:RIV	Key residues (Contribution ≤ −2 kJ/mol)	Residue contributions (kJ/mol)
c1	S1	S4	82.77	−96.505 ± 0.454	CYS-99	−3.4787
					PHE-174	−4.8023
					VAL-213	−2.3387
					TRP-215	−10.0666
c2	S1	S1′	7.06	−72.622 ± 1.554	PHE-41	−2.2875
					CYS-58	−2.0859
					GLN-192	−2.0565
					VAL-213	−2.7228
					TRP-215	−6.3439
c3	S1	S4	4.33	−83.085 ± 1.847	CYS-99	−2.2025
					PHE-174	−3.4795
					TRP-215	−10.5807

Binding modes are based on the representative structures of the top three clusters of the Y99C trajectories, where R1 and R4 refer to the positions of the chlorothiophene and the morpholinone rings of RIV, and S1, S1', and S4 refer to the RIV groups that are roughly close to the positions of the corresponding FXa pockets. Sampled conformations belonging to c1, c2, and c3 were picked out for MM-PBSA, calculations of the RIV, and binding energy, respectively. The key contributed residues and their contributions of each cluster are also listed in the last two columns of the table.

side chain of Trp215 between the S4 and S1 pocket, which not only reshapes the aromatic cage of the S4 pocket but also brings a shift of RIV outward by the hydrophobic interactions between the rings of Trp215 and the R2 & R3 rings of RIV. This shift gives RIV more flexibility to interact with surrounding residues with different binding poses, such as the hydrophobic interactions with Val213 and hydrogen bonds with His57 and Ser195, which may compensate some of the loss of binding energy in the S4 pocket. The second cluster (c2) is of only 7% of all the simulation time, in which RIV is relatively stable bound in a totally unexpected binding pattern, with R1 in the S1 pocket and R4 in the S1' pocket. In this pattern, RIV totally loses its interaction with Phe174 and Cys99 but still maintains hydrophobic interactions with Trp215 and Val213 in the S1 pocket and gains additional interactions from the residues in the S1' pocket, such as Phe41, Cy58, and Gln61. However, the S1' pocket is relatively exposed to the solution, making it easy for RIV to move away. At last, the third cluster (c3) is of only 4% of the total sampling, whose representative structure is somewhat like the stable S1S4 binding in wild type, but Trp215 is flipped and R4 of RIV is more likely sandwiched between the rings of Phe174 and Trp215. According to the MM-PBSA calculations on the concatenated trajectory of Y99C mutant, the c1 (−96.505 ± 0.454 kJ/mol) and c3 (−83.085 ± 1.847 kJ/mol) show relatively stronger binding affinity than c2 (−72.622 ± 1.554 kJ/mol) featured as the S1S1' binding pose (**Supplementary Table S3**).

It should be noted that a single representative structure of the clusters may not fully describe the dynamics in different binding poses of RIV. Therefore, the distribution of the R4 group (**Figure 5B**) was compared by superposition of the coordinates of RIV's C3 atom. In the first cluster, although R4 stays in the S4 pocket for the most time, it may move a little toward the catalytic pocket or even far away to the S1' pocket. In the second cluster, the distribution of R4 mainly assembles in the S1' pocket, with only minor frames diffused into the S4 pocket. At last, the distribution of the R4 group is more concentrated in S4 of the third cluster.

Possibly because of the variations between the trajectories of Y99C simulations, the difference in binding pattern is more obvious if we clustered the three repeats separately with a lower cut-off (**Supplementary Figure S6**). In the first trajectory Y99C_I, although there is some deformation of the S4 pocket, most of the time R4 stays around S4 to interact with Phe174 or the residues His57, Cys58, or Tyr60 close to the catalytic pocket. At the same time, the flip of Trp215 to the S1 pocket destabilizes the binding or R1 in the S1 pocket. In the second trajectory, most of the time the side chain of Trp215 is not flipped; thus, R1 can stay stably in the S1 pocket. However, the ring of R4 loses its hydrophobic interaction with Phe174 and has an astonishing shift toward the catalytic pocket or even to the S1' pocket, although most of the time, the electrostatic interaction with the thiol of Cys99 is kept, and the distribution of R4 is quietly diffused around S4 in cluster one (c1) or around S1' in c2 & c3. As a result, the binding pattern is relatively shifted from the typical S1S4 to the atypical S1S1'. At last, in the third trajectory, the S4 pocket is stronger and relatively stable, and the binding of RIV is quite similar to the typical S1S4 binding in wild type.

DISCUSSION

To explain the unexpected results of the Y99C mutant, the molecular dynamics simulations in this work provide a great help to explore the microscopic details of RIV-FXa binding in the dynamic condition of dilute solution, although we are aware that limited resources of computations and requirements of accuracy may lead to some deviations from reality.

At first, the great flexibilities in both the protein and the drug indicate that our analyses are not enough to sufficiently reflect the whole landscape of the dynamic binding. On one side, the

flexibilities may be the reason for RIV in equilibration of multiple binding poses, such as the fluctuation of the 99-loop and the flipping of Trp215's side chain. On the other side, they made it quite complex to compare, analyze, summarize, and visualize the binding in an easy and intuitive manner because the shape and the positions of the drug and the binding pockets are always changing. For example, RIV is an extended flexible compound with four rings, where the chlorothiophene (R1) and the morpholinone (R4) rings tend to bind with different pockets somewhat independent to each other. In structural clustering, the RMSD of all atoms of RIV may include all the structural information, but the distribution of different binding patterns with different pockets could not be clearly visualized by the superposed mass center of whole RIV. Considering that the R1 group should be more stably bound as discussed later, we chose to focus on the more dynamic distribution of the R4 group by superposition of the coordinates of the C3 atom on it. Conversely, a similar NOAC apixaban is much more rigid and compact, making it easier to be described, but at the same time, it may lose some possibility to bind FXa in alternative patterns as RIV, which may partially explain its lower binding affinity in some mutants, as discussed earlier (Qu et al., 2019). From this example, we would again emphasize the recognition of the important influence of the flexibility of both the target protein and the drug molecule, which was often neglected in structure-based drug design and screening using static crystal structures but more and more considered in recent studies, especially in protein and antibody design (Corbeil et al., 2021).

Second, here, we used a popular method of clustering to find a representative structure of different binding patterns for illustration. However, the representative structures may not always describe the difference between the bindings very well since the results of clustering may be importantly affected by the distribution of samples. Due to the limited resources, we chose to repeat the simulations thrice for 200 ns each instead of an extended simulation to obtain more diffused sampling. However, the sampling in the three trajectories of the Y99C mutant is quite distinctive with each other, as described previously. For comparison with wild type and Y99A, we used the clustering on all samples of the three trajectories combined together, which not only enlarged the pool of samples but also the variations within one cluster. Therefore, we supplemented the clustering of separate trajectories so as to visualize the different binding patterns more clearly. Adoption of alternative algorithms or strategies of clustering may improve the recognition of different binding patterns (Fraley and Raftery, 2002; Xu and Wunsch, 2005) but may also introduce more artificial bias or differences with earlier analyses.

At last, another source of error may come from the force field that we used. The simulations on the three FXa systems were performed as early as 2018, using the popular CHARMM36 force field for FXa, TIP3P model for water, and CGenFF to generate parameters for RIV as we did not evaluate the effect of force field using alternative ones. However, according to our discussions with experts in MD simulations, we would expect some improvement in the validity of our simulations if some conditions could be applied further. First, the general CHARMM36 force fields in GROMACS were fit for folded structures. However, according to the results of our simulations and earlier studies (Wang et al., 2012; Qu et al., 2019; Qu and Xu, 2019), the region of the 99-loop and the 174-loop may have much higher flexibilities in dilute solution rather than those of a compact aromatic cage as in the crystallization condition, especially when the hydrophobic environment of the S4 pocket is somewhat damaged by mutations. Therefore, if the loop regions were described by force fields specifically adjusted for an intrinsic disordered region, such as a folded-IDP balanced force field ff03CMAP that we used recently (Zhang et al., 2019; Mu et al., 2022), the dynamics of the critical loop regions might be more realistic, especially when we did not have enough resources for fully sufficient equilibrations. Second, the point charge generated by CGenFF for the chloride of RIV is too unsophisticated to describe the anisotropy of the electron density around it, which may form a σ-hole leading to a cation-π interaction with Tyr228 or a halogen bond with the conserved Asp189 deep in the bottom of the S1 pocket. The halogen bond might provide one of the reasons for the chlorothiophene as a critical pharmacophore for the specific binding of RIV in FXa. Although, in recent years, the importance of halogen bond has been highly recognized in biology and drug design, and great advances has been made to apply it in simulations, the cases in classical MD simulations are still too limited for us to find a way to integrate it into our simulations successfully (Franchini et al., 2018; Zhu et al., 2019; Dong et al., 2020; Zhang et al., 2020; Zhu et al., 2020). The inaccurate description of the chloride interactions may underestimate the binding of RIV in the S1 pocket, leading to overestimation of clusters with R1 out of the S1 pocket. However, this deviation happens in both wild type and the mutants and may not directly affect the deformation of the S4 pocket and the shift of R4 to the S1' pocket. Therefore, in our opinion, the qualitative conclusion on the dynamic binding of RIV in the Y99C mutant should be still valid.

CONCLUSION

In this work, we report a novel mutant of coagulation factor X, Tyr319Cys (Y99C of FXa), identified in a patient with severe bleeding, which can bind and cleave specific chromogenetic substrates at a normal level. Consistently, molecular dynamics simulations confirmed that its S4 pocket was much less deformed than that of the Y99A mutant and may maintain the binding affinity with rivaroxaban through dynamic binding between multiple poses. Detailed structural analyses indicated that the backbone of the 99-loop is only in minor fluctuation compared with Y99A and kept part of the hydrophobic and H-bonds with the S4 pocket. At the same time, the flexible flipping of Trp215's side chain may help stabilize alternative binding poses with the R4 group in the catalytic pocket or even in the S1' pocket. This result again emphasizes the importance to consider the flexibilities of both target protein and drug compound in structure-based drug design and may support the hypothesis

to develop similar recombinant FXa as a potential reversal agent for novel oral anticoagulants.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of Ruijin Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, QX and QD. Methodology, QX and QD. Investigation, Z-LZ, CC, and S-YQ. Software and data curation, Z-LZ. Writing—original draft, Z-LZ, CC, and QX. Writing—review & editing, QX and QD. Funding acquisition, QX and QD. Supervision, QX and QG.

FUNDING

This work is supported by the National Natural Science Foundation of China (31770772 to QX and 81770135 to QD) and the Shanghai Sailing Program (19YF1429600).

ACKNOWLEDGMENTS

We thank Prof. Dong-Qing Wei of Shanghai Jiao Tong University for support of partial computational resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.877170/full#supplementary-material

REFERENCES

Abdel-Azeim, S., Oliva, R., Chermak, E., De Cristofaro, R., and Cavallo, L. (2014). Molecular Dynamics Characterization of Five Pathogenic Factor X Mutants Associated with Decreased Catalytic Activity. *Biochemistry* 53 (44), 6992–7001. doi:10.1021/bi500770p

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

Al-Obeidi, F., and Ostrem, J. (1999). Factor Xa Inhibitors. *Expert Opin. Ther. Patents* 9 (7), 931–953. doi:10.1517/13543776.9.7.931

Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012). H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* 40, W537–W541. doi:10.1093/nar/gks375

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. (1981). "Interaction Models for Water in Relation to Protein Hydration," in *Intermolecular Forces* (Dordrecht: Springer), 331–342. doi:10.1007/978-94-015-7658-1_21

Camire, R. M. (2002). Prothrombinase Assembly and S1 Site Occupation Restore the Catalytic Activity of FXa Impaired by Mutation at the Sodium-Binding Site. *J. Biol. Chem.* 277 (40), 37863–37870. doi:10.1074/jbc.m203692200

Carpenter, E., Singh, D., Dietrich, E., and Gums, J. (2019). Andexanet Alfa for Reversal of Factor Xa Inhibitor-Associated Anticoagulation. *Ther. Adv. Drug Saf.* 10, 2042098619888133. doi:10.1177/2042098619888133

Chen, Q., Buolamwini, J. K., Smith, J. C., Li, A., Xu, Q., Cheng, X., et al. (2013). Impact of Resistance Mutations on Inhibitor Binding to HIV-1 Integrase. *J. Chem. Inf. Model.* 53 (12), 3297–3307. doi:10.1021/ci400537n

Connolly, S. J., Milling, T. J., Jr., Eikelboom, J. W., Gibson, C. M., Curnutte, J. T., Gold, A., et al. (2016). Andexanet Alfa for Acute Major Bleeding Associated with Factor Xa Inhibitors. *N. Engl. J. Med.* 375 (12), 1131–1141. doi:10.1056/nejmoa1607887

Corbeil, C. R., Manenda, M. S., Sulea, T., Baardsnes, J., Picard, M.-È., Hogues, H., et al. (2021). Redesigning an Antibody H3 Loop by Virtual Screening of a Small Library of Human Germline-Derived Sequences. *Sci. Rep.* 11 (1), 21362. doi:10.1038/s41598-021-00669-w

Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999). Peptide Folding: When Simulation Meets experiment. *Angew. Chem. Int. Ed.* 38 (1–2), 236–240. doi:10.1002/(sici)1521-3773(19990115)38:1/2<236::aid-anie236>3.0.co;2-m

Daura, X., Haaksma, E., and van Gunsteren, W. F. (2000). Factor Xa: Simulation Studies with an Eye to Inhibitor Design. *J. Computer Aided Mol. Des.* 14 (6), 507–529. doi:10.1023/a:1008120005475

Davie, E. W., Fujikawa, K., and Kisiel, W. (1991). The Coagulation Cascade: Initiation, Maintenance, and Regulation. *Biochemistry* 30 (43), 10363–10370. doi:10.1021/bi00107a001

Dong, T.-g., Peng, H., He, X.-f., Wang, X., and Gao, J. (2020). Hybrid Molecular Dynamics for Elucidating Cooperativity between Halogen Bond and Water Molecules during the Interaction of P53-Y220c and the PhiKan5196 Complex. *Front. Chem.* 8, 344. doi:10.3389/fchem.2020.00344

Du, Q. Q., Yan, Q., Yao, X. J., and Xue, W. W. (2019). Elucidating the Tight-Binding Mechanism of Two Oral Anticoagulants to Factor Xa by Using Induced-Fit Docking and Molecular Dynamics Simulation. *J. Biomol. Struct. Dyn.* 38, 625–633. doi:10.1080/07391102.2019.1583605

Ersayin, A., Thomas, A., Seyve, L., Thielens, N., Castellan, M., Marlu, R., et al. (2017). Catalytically Inactive Gla-Domainless Factor Xa Binds to TFPI and Restores *Ex Vivo* Coagulation in Hemophilia Plasma. *Haematologica* 102 (12), e483–e485. doi:10.3324/haematol.2017.174037

Escolar, G., Diaz-Ricart, M., and Arellano-Rodrigo, E. (2017). Andexanet Alpha: A Recombinant Mimetic of Human Factor Xa for the Reversal of Anticoagulant Therapies. *Drugs Today* 53 (5), 271–282. doi:10.1358/dot.2017.53.5.2630780

Fraley, C., and Raftery, A. E. (2002). Model-based Clustering, Discriminant Analysis, and Density Estimation. *J. Am. Stat. Assoc.* 97 (458), 611–631. doi:10.1198/016214502760047131

Franchini, D., Dapiaggi, F., Pieraccini, S., Forni, A., and Sironi, M. (2018). Halogen Bonding in the Framework of Classical Force fields: The Case of Chlorine. *Chem. Phys. Lett.* 712, 89–94. doi:10.1016/j.cplett.2018.09.052

Fricker, P. C., Gastreich, M., and Rarey, M. (2004). Automated Drawing of Structural Molecular Formulas under Constraints. *J. Chem. Inf. Comput. Sci.* 44 (3), 1065–1078. doi:10.1021/ci049958u

Hedstrom, L. (2002). Serine Protease Mechanism and Specificity. *Chem. Rev.* 102 (12), 4501–4524. doi:10.1021/cr000033x

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* 18 (12), 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h

Hoover, W. G. (1985). Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A.* 31 (3), 1695–1697. doi:10.1103/physreva.31.1695

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graph* 14 (133-8), 33–38. doi:10.1016/0263-7855(96)00018-5

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 (2), 926–935. doi:10.1063/1.445869

Kaatz, S., Kouides, P. A., Garcia, D. A., Spyropolous, A. C., Crowther, M., Douketis, J. D., et al. (2012). Guidance on the Emergent Reversal of Oral Thrombin and Factor Xa Inhibitors. *Am. J. Hematol.* 87, S141–S145. doi:10.1002/ajh.23202

Katz, B. A., Mackman, R., Luong, C., Radika, K., Martelli, A., Sprengeler, P. A., et al. (2000). Structural Basis for Selectivity of a Small Molecule, S1-Binding, Submicromolar Inhibitor of Urokinase-type Plasminogen Activator. *Chem. Biol.* 7 (4), 299–312. doi:10.1016/s1074-5521(00)00104-6

Klauda, J. B., Venable, R. M., Freites, J. A., O'Connor, J. W., Tobias, D. J., Mondragon-Ramirez, C., et al. (2010). Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J. Phys. Chem. B* 114 (23), 7830–7843. doi:10.1021/jp101759q

Kumari, R., Kumar, R., and Lynn, A. (2014). Open Source Drug Discovery Cg_mmpbsa-A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* 54 (7), 1951–1962. doi:10.1021/ci500020m

Lee, Y.-K., and Player, M. R. (2011). Developments in Factor Xa Inhibitors for the Treatment of Thromboembolic Disorders. *Med. Res. Rev.* 31 (2), 202–283. doi:10.1002/med.20183

Li, F., Chen, C., Qu, S. Y., Zhao, M. Z., Xie, X., Wu, X., et al. (2019). The Disulfide Bond between Cys22 and Cys27 in the Protease Domain Modulate Clotting Activity of Coagulation Factor X. *Thromb. Haemost.* 119 (6), 871–881. doi:10.1055/s-0039-1683442

Maignan, S., Guilloteau, J.-P., Choi-Sledeski, Y. M., Becker, M. R., Ewing, W. R., Pauls, H. W., et al. (2003). Molecular Structures of Human Factor Xa Complexed with Ketopiperazine Inhibitors: Preference for a Neutral Group in the S1 Pocket. *J. Med. Chem.* 46 (5), 685–690. doi:10.1021/jm0203837

Mu, J., Zhou, J., Gong, Q., and Xu, Q. (2022). An Allosteric Regulation Mechanism of Arabidopsis Serine/Threonine Kinase 1 (SIK1) through Phosphorylation. *Comput. Struct. Biotechnol. J.* 20, 368–379. doi:10.1016/j.csbj.2021.12.033

Nazaré, M., Will, D. W., Matter, H., Schreuder, H., Ritter, K., Urmann, M., et al. (2005). Probing the Subpockets of Factor Xa Reveals Two Binding Modes for Inhibitors Based on a 2-carboxyindole Scaffold: A Study Combining Structure-Activity Relationship and X-ray Crystallography. *J. Med. Chem.* 48 (14), 4511–4525. doi:10.1021/jm0490540

Parrinello, M., and Rahman, A. (1981). Polymorphic Transitions in Single Crystals: a New Molecular Dynamics Method. *J. Appl. Phys.* 52 (12), 7182–7190. doi:10.1063/1.328693

Perera, L., Essmann, U., and Berkowitz, M. L. (1995). Effect of the Treatment of Long-range Forces on the Dynamics of Ions in Aqueous Solutions. *J. Chem. Phys.* 102 (1), 450–456. doi:10.1063/1.469422

Perzborn, E., Roehrig, S., Straub, A., Kubitza, D., and Misselwitz, F. (2011). The Discovery and Development of Rivaroxaban, an Oral, Direct Factor Xa Inhibitor. *Nat. Rev. Drug Discov.* 10 (1), 61–75. doi:10.1038/nrd3185

Qu, S., and Xu, Q. (2019). Different Roles of Some Key Residues in the S4 Pocket of Coagulation Factor Xa for Rivaroxaban Binding. *Chem. J. Chin. Univ. Chin.* 40 (9), 1918–1925. doi:10.7503/cjcu20190261

Qu, S. Y., Xu, Q., Wu, W., Li, F., Li, C. D., Huang, R., et al. (2019). An Unexpected Dynamic Binding Mode between Coagulation Factor X and Rivaroxaban Reveals Importance of Flexibility in Drug Binding. *Chem. Biol. Drug Des.* 94 (3), 1664–1671. doi:10.1111/cbdd.13568

Roehrig, S., Straub, A., Pohlmann, J., Lampe, T., Pernerstorfer, J., Schlemmer, K.-H., et al. (2005). Discovery of the Novel Antithrombotic Agent 5-Chloro-N-({(5s)-2-Oxo-3- [4-(3-Oxomorpholin-4-Yl)phenyl]-1,3-Oxazolidin-5-Yl} methyl)thiophene- 2-carboxamide (BAY 59-7939): An Oral, Direct Factor Xa Inhibitor. *J. Med. Chem.* 48 (19), 5900–5908. doi:10.1021/jm050101d

Samama, M.-M., Amiral, J., Guinet, C., Flem, L. L., and Seghatchian, J. (2013). Monitoring Plasma Levels of Factor Xa Inhibitors: How, Why and when? *Expert Rev. Hematol.* 6 (2), 155–164. doi:10.1586/ehm.13.11

Samama, M. M. (2013). Which Test to Use to Measure the Anticoagulant Effect of Rivaroxaban: the Anti-factor Xa Assay. *J. Thromb. Haemost.* 11 (4), 579–580. doi:10.1111/jth.12165

Sartori, M., and Cosmi, B. (2018). Andexanet Alfa to Reverse the Anticoagulant Activity of Factor Xa Inhibitors: a Review of Design, Development and Potential Place in Therapy. *J. Thromb. Thrombolysis* 45 (3), 345–352. doi:10.1007/s11239-018-1617-2

Schrodinger, L. (2015). *The PyMOL Molecular Graphics System*. San Carlos, CA, United States: DeLano Scientific LLC.

Singh, N., and Briggs, J. M. (2010). Molecular Dynamics Simulations of Factor Xa: Insight into Conformational Transition of its Binding Subsites. *Biopolymers* 89 (12), 1104–1113. doi:10.1002/bip.21062

Vanommeslaeghe, K., Ghosh, J., Polani, N. K., Sheetz, M., Pamidighantam, S. V., Connolly, J. W. D., et al. (2011). Automation of the Charmm General Force Field for Drug-like Molecules. *Biophys. J.* 100 (3), 611. doi:10.1016/j.bpj.2010.12.3519

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., et al. (2010). CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* 31 (4), 671–690. doi:10.1002/jcc.21367

Vanommeslaeghe, K., Raman, E. P., and MacKerell, A. D., Jr (2012). Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* 52 (12), 3155–3168. doi:10.1021/ci3003649

Venkateswarlu, D., Perera, L., Darden, T., and Pedersen, L. G. (2002). Structure and Dynamics of Zymogen Human Blood Coagulation Factor X. *Biophysical J.* 82 (3), 1190–1206. doi:10.1016/s0006-3495(02)75476-3

Verhoef, D., Visscher, K. M., Vosmeer, C. R., Cheung, K. L., Reitsma, P. H., Geerke, D. P., et al. (2017). Engineered Factor Xa Variants Retain Procoagulant Activity Independent of Direct Factor Xa Inhibitors. *Nat. Commun.* 8, 528. doi:10.1038/s41467-017-00647-9

Wang, J.-F., Hao, P., Li, Y.-X., Dai, J.-L., and Li, X. (2012). Exploration of Conformational Transition in the Aryl-Binding Site of Human FXa Using Molecular Dynamics Simulations. *J. Mol. Model.* 18 (6), 2717–2725. doi:10.1007/s00894-011-1295-x

Wang, W., Yuan, J., Fu, X., Meng, F., Zhang, S., Xu, W., et al. (2016). Novel Anthranilamide-Based FXa Inhibitors: Drug Design, Synthesis and Biological Evaluation. *Molecules* 21 (4), 491. doi:10.3390/molecules21040491

Wong, P. C., Pinto, D. J. P., and Zhang, D. (2011). Preclinical Discovery of Apixaban, a Direct and Orally Bioavailable Factor Xa Inhibitor. *J. Thromb. Thrombolysis* 31 (4), 478–492. doi:10.1007/s11239-011-0551-3

Xu, R., and WunschII, D. (2005). Survey of Clustering Algorithms. *IEEE Trans. Neural Netw.* 16 (3), 645–678. doi:10.1109/tnn.2005.845141

Zhang, Q., Smalley, A., Zhu, Z., Xu, Z., Peng, C., Chen, Z., et al. (2020). Computational Study of the Substituent Effect of Halogenated Fused-Ring Heteroaromatics on Halogen Bonding. *J. Mol. Model.* 26 (10), 270. doi:10.1007/s00894-020-04534-x

Zhang, Y., Liu, H., Yang, S., Luo, R., and Chen, H.-F. (2019). Well-Balanced Force Field ff03CMAP for Folded and Disordered Proteins. *J. Chem. Theor. Comput.* 15 (12), 6769–6780. doi:10.1021/acs.jctc.9b00623

Zhu, Z., Wang, G., Xu, Z., Chen, Z., Wang, J., Shi, J., et al. (2019). Halogen Bonding in Differently Charged Complexes: Basic Profile, Essential Interaction Terms and Intrinsic σ-hole. *Phys. Chem. Chem. Phys.* 21 (27), 15106–15119. doi:10.1039/c9cp01379b

Zhu, Z., Xu, Z., and Zhu, W. (2020). Interaction Nature and Computational Methods for Halogen Bonding: A Perspective. *J. Chem. Inf. Model.* 60 (6), 2683–2696. doi:10.1021/acs.jcim.0c00032

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fmolb.2022.901603

Understanding the P-Loop Conformation in the Determination of Inhibitor Selectivity Toward the Hepatocellular Carcinoma-Associated Dark Kinase STK17B

Chang Liu[1†], Zhizhen Li[2†], Zonghan Liu[1], Shiye Yang[1], Qing Wang[3] and Zongtao Chai[1,4]**

[1]Department of Hepatic Surgery VI, Eastern Hepatobiliary Surgery Hospital, The Second Military Medical University (Navy Medical University), Shanghai, China, [2]Department of Biliary Surgery I, Eastern Hepatobiliary Surgery Hospital, The Second Military Medical University (Navy Medical University), Shanghai, China, [3]Oncology Department, Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China, [4]Department of Hepatic Surgery, Shanghai Geriatric Center, Shanghai, China*

OPEN ACCESS

Edited by:
Weiliang Zhu,
Shanghai Institute of Materia Medica (CAS), China

Reviewed by:
Jianzhong Chen,
Shandong Jiaotong University, China
Zhongjie Liang,
Soochow University, China

***Correspondence:**
Qing Wang
jushi1984@163.com
Zongtao Chai
runout@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: 22 March 2022
Accepted: 22 April 2022
Published: 10 May 2022

Citation:
Liu C, Li Z, Liu Z, Yang S, Wang Q and Chai Z (2022) Understanding the P-Loop Conformation in the Determination of Inhibitor Selectivity Toward the Hepatocellular Carcinoma-Associated Dark Kinase STK17B. Front. Mol. Biosci. 9:901603. doi: 10.3389/fmolb.2022.901603

As a member of the death-associated protein kinase family of serine/threonine kinases, the STK17B has been associated with diverse diseases such as hepatocellular carcinoma. However, the conformational dynamics of the phosphate-binding loop (P-loop) in the determination of inhibitor selectivity profile to the STK17B are less understood. Here, a multi-microsecond length molecular dynamics (MD) simulation of STK17B in the three different states (ligand-free, ADP-bound, and ligand-bound states) was carried out to uncover the conformational plasticity of the P-loop. Together with the analyses of principal component analysis, cross-correlation and generalized correlation motions, secondary structural analysis, and community network analysis, the conformational dynamics of the P-loop in the different states were revealed, in which the P-loop flipped into the ADP-binding site upon the inhibitor binding and interacted with the inhibitor and the C-lobe, strengthened the communication between the N- and C-lobes. These resulting interactions contributed to inhibitor selectivity profile to the STK17B. Our results may advance our understanding of kinase inhibitor selectivity and offer possible implications for the design of highly selective inhibitors for other protein kinases.

Keywords: protein kinase, STK17B, p-loop, molecular dynamics simulation, conformational dynamics

INTRODUCTION

Protein kinases transfer the γ-phosphate group of ATP to serine, threonine, or tyrosine residues of their substate proteins. This physiological process is also called as phosphorylation. Protein phosphorylation provokes cellular signal transduction cascades associated with cell differentiation, growth, homeostasis, and death (Pearce et al., 2010). Aberrant protein kinase function by either activating mutations or translocations is related with numerous disease states, including cancer, Alzheimer disease, Parkinson's disease, inflammation, and metabolic disease (Attwood et al., 2021; Cohen et al., 2021). Protein kinase are thus important therapeutic targets for drug discovery. Until now, 71 small-molecule kinase inhibitors have been approved by the FDA in the treatment of cancer and other diseases (Roskoski, 2021).

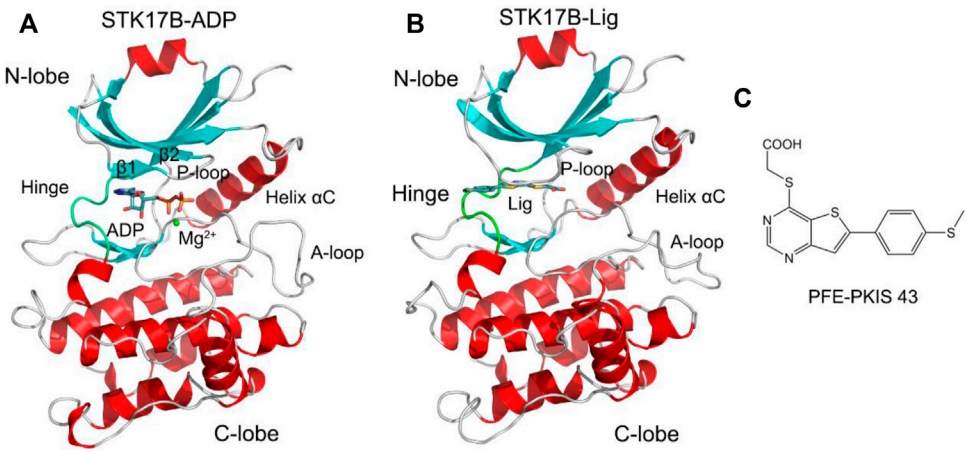


FIGURE 1 | Cartoon representation of STK17B in complex with ADP (PDB ID: 6QF4) **(A)** and the inhibitor PFE-PKIS 43 (PDB ID: 6Y6F) **(B)**. The secondary structural elements of α-helices and β-strands are colored by red and cyan, respectively. The loop including the phosphate-binding loop (P-loop) and the activation loop (A-loop) is colored by gray. The hinge domain is colored by green. ADP and inhibitor are depicted by stick representation. Mg^{2+} ion is shown by a green sphere. **(C)** Chemical structure of the inhibitor PFE-PKIS 43.

Despite the inspiring clinical benefits, kinase inhibitors are still encountered an unsurmountable challenge hallmarked by kinase selectivity profile. This is because that the vast majority of protein kinase inhibitors bind to the conserved ATP-binding site, leading to the poor selectivity of kinase inhibitors towards a unique kinase (Wu et al., 2015; Chen et al., 2020; Li C. et al., 2020). For example, Davis et al. (2011) have previously explored the interaction of 72 kinase inhibitors with 442 kinases representing >80% of the human catalytic protein kimome and found that the kinase inhibitor selectivity profile is relatively narrow, with 10%–40% of inhibitors interacting with >60% of kinases, and each inhibitor interacting with more than one kinase. Therefore, developing a promising strategy to discover highly selective inhibitors is an area of intensive research in kinase kinome (Lu et al., 2018, Lu et al., 2019a; Lu and Zhang, 2019).

To achieve inhibitor selectivity, several successful strategies have been reported. Covalent kinase inhibitors are a class of compounds that harbour a reactive, electrophilic warhead, reacting with a nucleophilic cysteine residue at the target site and then forming a stable covalent adduct (Nussinov and Tsai, 2015; Lu and Zhang, 2017; Ni et al., 2020). These covalent inhibitors have pharmacological advantages of high potency and selectivity. For instance, in the double mutant T790M/L858R epidermal growth factor receptor (EGFR), the FDA-approved Osimertinib engages with Cys797 at the ATP-binding site through a covalent bond (Jia et al., 2016; Nussinov et al., 2022). However, in the ATP-binding site, the availability of cysteine residues at the proper position is scarce for most of kinases, rendering the design of covalent inhibitors remaining a challenging task.

Harnessing the sequence differences of ATP-binding site that control inhibitor selectivity has emerged as an alternative. One quintessential example is STK17B, a member of the death-associated protein kinase family of serine/threonine kinases (Pearce et al., 2010). Overexpression of STK17B plays a crucial role in hepatocellular carcinoma and thus, inhibition of STK17B catalytic activity in cells implies clinical utility in the treatment of this malignancy (Lan et al., 2018). The crystal structure of ADP-bound STK17B contains a small N-lobe and a large C-lobe (**Figure 1A**). The N-lobe is mainly consisted of five β-strands and one catalytic helix αC. The phosphate-binding loop (P-loop) connecting the β1 to the β2 adopts a "U" shape. The C-lobe is largely constituted by helices. The activation loop (A-loop) that control catalytic activity runs along the substrate binding groove. The flexible hinge domain connects the N-lobe to the C-lobe. ADP binds to the cleft between the two lobes located under the P-loop. There are several reported STK17B inhibitors, including quercetin **1**, dovitinib **2**, and benzofuranone **3** (**Supplementary Figure S1**). However, these are non-selective or modest selective inhibitors toward STK17B. Recently, Picado et al. (2020) reported a cell active STK17B inhibitor, thieno[3,2-d] pyrimidine PFE-PKIS 43 (**Figure 1B**), which had remarkable potency and selectivity toward STK17B against other homologous protein kinases. A crystal structure of PFE-PKIS 43 complexed with STK17B highlights a unique P-loop flip that interacts with the inhibitor. In addition to the crystal structure of STK17B–PEF-PRIS 43 complex, there are five co-crystal structures of STK17B in complex with different inhibitors previously reported, including EBD (PDB ID: 3LMO), quercetin (PDB ID: 3LM5), UNC-AP-194 probe (PDB ID: 6Y6H), AP-229 (PDB ID: 6ZJF), and dovitinib (PDB ID: 7AKG). Structural superimposition of the five co-crystal structures shows that the P-loop conformation in these structures adopts the ordered β-strands (**Supplementary Figure S2**), which is different from that in the crystal structure of STK17B–PEF-PRIS 43 complex. However, the conformational dynamics of the P-loop in the STK17B–PEF-PRIS 43 complex remain unexplored.

Here, we performed a multi-microsecond length molecular dynamics (MD) simulation of STK17B in the ligand-free, ADP-bound, or ligand-bound states, to characterize the conformational

plasticity of the P-loop and its interplay with the ligand over long time-scales. We collected an overall simulated trajectories of 27 μs, which were conducted in multiple replicates in different states. Coupled with the analyses of principal component analysis (PCA), cross-correlation and generalized correlation motions, secondary structural elements, and community networks, the distinct conformational dynamics of the P-loop in the different states were presented. Our results will advance our understanding of kinase inhibitor selectivity and provide hits for the design of selective inhibitors for other protein kinases.

RESULTS AND DISCUSSION

System Stability

Based on the available X-ray crystal structures of STK17B, we collected conformational ensembles of μs-length MD simulations. We simulated STK17B in various states (i.e., ligand-free, ATP-bound, or ligand-bound) to explore differences and similarities during MD simulations. For each system, MD simulations were performed in explicit water environment, collecting multiple μs-length trajectories (i.e., 3 replicates of 3 μs each) and yielding a total of sampling of 27 μs. Such a multiple and independent μs-length MD trajectory has been proved efficient for investigating the interdependent conformational plasticity of the kinase domains (i.e., P-loop and A-loop) and their interactions with the ADP or the ligand (Lu et al., 2019b; Zhang et al., 2019; Lu et al., 2021a; Lu et al., 2021b; Maloney et al., 2021; Ni et al., 2021; Hu et al., 2022).

We first monitored the root mean square deviation (RMSD) of the kinase Cα atoms averaged over three replicates for each system. As shown in **Supplementary Figure S3**, the kinase backbone reached a similar stability in the apo (ligand-free), ADP-bound, and ligand-bound states (i.e., the RMSD reaches 1–1.5 Å). This suggested that upon ADP or ligand binding, the overall stability of the kinase has no significant conformational differences during the simulations.

Coupled Motions of Kinase Intradomains

The dynamic correlation analysis was carried out to probe the interdependent dynamics among different kinase domains. Two distinct methods, including the traditional Pearson cross-correlation (CC*ij*) and the generalized correlation (GC*ij*), were used to calculate the correlation analysis (Shibata et al., 2020; Liang et al., 2021; Zhang et al., 2022a), which was conducted and averaged over all MD trajectories. The CC*ij* analysis describes the collinear correlation between the two residue Cα atoms (*i* and *j*), reflecting whether they move in the correlated motions (CC*ij* > 0) or in the anti-correlated (CC*ij* < 0) motions. The GC*ij* analysis monitors the degree of correlation between the two residue Cα atoms (*i* and *j*), reflecting how much information of one atom's positions is provided by that of another atom. The GC*ij* analysis cannot identify correlated or anticorrelated motions of the two atoms, ignoring the elucidation of atom's motions.

The CC*ij* matrix of STK17B that is represented by a two-by-two plot of the Cα CC*ij* coefficients reveals a conserved pattern of correlated/anticorrelated motions in all apo, ADP-bound and ligand-bound states (**Figure 2**). The N-lobe containing the P-loop (residues 40–47) and C-lobe shows anticorrelated motions, which is also observed on other protein kinases such as anaplastic lymphoma kinase (ALK) (Liang et al., 2021), BCR-ABL (Zhang et al., 2022a) and epidermal growth factor receptor (EGFR) (Qiu et al., 2021). This suggests that the opposite movement of the N- and C-lobes favours the "open or closed" conformational transition of the nucleotide binding site underlying ADP/ATP and substrate binding. In addition, the difference matrix of ADP- and ligand-bound states using the apo state as the reference indicates that the opposite movement of the N- and C-lobes was stronger in the ADP-bound state than that in the ligand-bound state (**Supplementary Figure S4**). The GC*ij* analysis was further used to unravel the global dependencies of the protein kinase domain motions (**Figure 3**). Like the CC*ij* matrix, the GC*ij* matrix of the STK17B in the apo, ADP-bound and ligand-bound states showed a high degree of correlations between the N-lobe and the C-lobe. However, the protein in the ligand-bound system had a slightly higher correlations than that in the ADP-bound and apo systems, which was further supported by the difference matrix of ADP- and ligand-bound states using the apo state as the reference (**Supplementary Figure S5**). This result indicated that ligand binding induced an enhanced motions of protein kinase domains.

Local Motions and Conformational Dynamics

In order to unravel the predominant collective motions of different STK17B states and capture their essential degrees of freedom, we conducted principal component analysis (PCA) of STK17B in the apo, ADP-bound, and ligand-bound states. Based on the PCA, the first two principal modes of motion (i.e., principal components 1 and 2, PC1 and PC2) provide information regarding to the large-amplitude motions of different STK17B states, which represent their functional dynamics (Masterson et al., 2011; Chen et al., 2019; Chen et al., 2021; He et al., 2021; Okeke et al., 2021; Rehman et al., 2021). In PCA, we selected all simulated trajectories for each system and subjected to RMS-fit to the same initial structure to rule out the translational and rotational motions of the protein.

As shown in **Figure 4A**, the apo protein sampled a confined distribution of conformations. Addition of ADP largely changed PC1, but did not change PC2 (**Figure 4B**), indicating that the protein kinase had increased dynamics in response to ADP binding. More remarkably, in the ligand-bound system (**Figure 4C**), both PC1 and PC2 were enlarged compared to the apo and ADP-bound systems. This observation suggested that the ligand binding induced more enhanced conformational dynamics of STK17B, which was consistent with the GC*ij* analysis. We further extracted the most represented conformation from each cluster in the ligand-bound state (L1–L3). As shown in **Supplementary Figure S6**, structural overlapping of the three most represented conformations showed that the P-loop and A-loop in the ligand-bound STK17B underwent obvious conformational changes. Indeed,

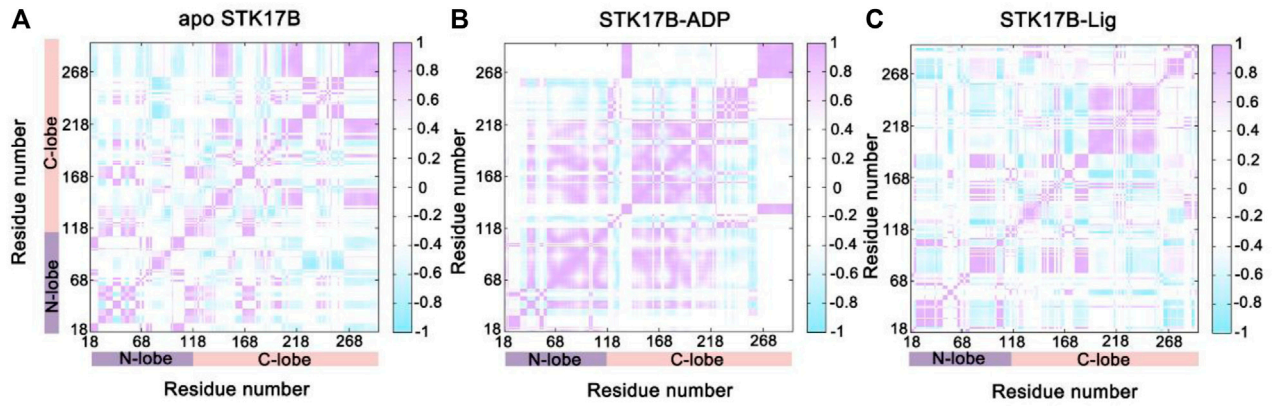


FIGURE 2 | The cross-correlation (CC_{ij}) matrix of STK17B for the apo **(A)**, ADP-bound **(B)**, and ligand-bound **(C)** systems. The correlated motions are colored by violent ($CC_{ij} > 0$), while the anti-correlated motions are colored by cyan ($CC_{ij} < 0$). Color scales are shown at the right. The CC_{ij} values with an absolute correlation coefficient of <0.4 are colored by white for clarity.

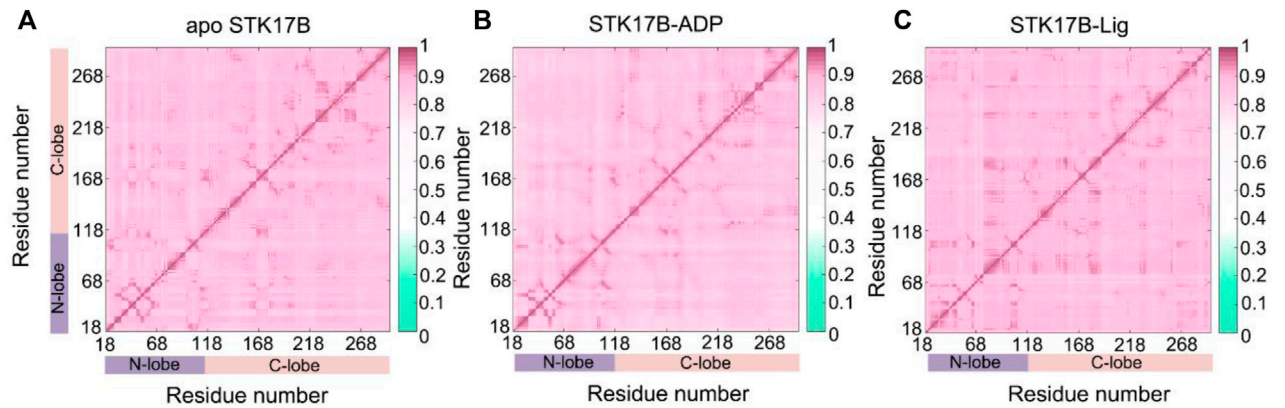


FIGURE 3 | The generalized correlation (GC_{ij}) matrix of STK17B for the apo **(A)**, ADP-bound **(B)**, and ligand-bound **(C)** systems. Color scales are shown at the right.

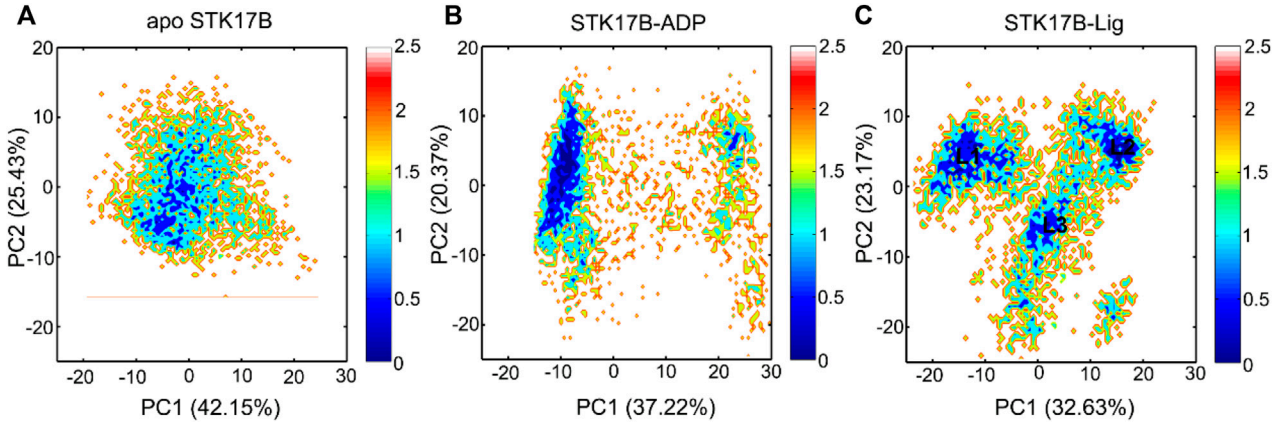


FIGURE 4 | The free energy landscape of the first and second principal components (PC1 and PC2) for the apo **(A)**, ADP-bound **(B)**, and ligand-bound **(C)** systems. The unit of free-energy values is kcal/mol.

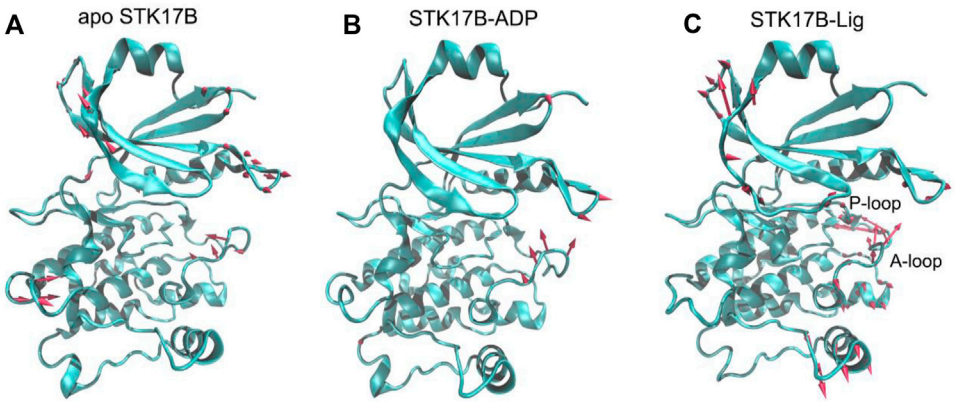


FIGURE 5 | The motion of the first principal component (PC1) for the apo **(A)**, ADP-bound **(B)**, and ligand-bound **(C)** systems. The red arrows represent the direction, with length proportional to the intensity of the motion.

previous MD simulations of protein kinase A (PKA) also indicated that ligand binding induced global transitions in the catalytic domain of PKA (Hyeon et al., 2009), supporting our MD simulation results of ligand-bound STK17B.

The conformational landscapes of different STK17B states based on the PCA results implied that STK17B was more dynamics in the presence of ligand. To further validate this hypothesis, the PC1 of the STK17B in the three different states was visualized on the 3D structure (**Figure 5**). The red arrows show the direction of residue motions, with the length proportional to the intensity of the motion. Remarkably, the ligand binding (**Figure 5C**) triggered more dynamic movement of P-loop and A-loop than the apo (**Figure 5A**) and the ADP-bound (**Figure 5B**) systems. For instance, no motion of the P-loop, but a weak motion of the A-loop was observed in both the apo and ADP-bound systems. In agreement with the PCA results, both the P-loop and the A-loop of STK17B in the presence of ligand were highly flexible, which may determine the selectivity profile of ligand to the STK17B.

Secondary Structural Analysis of the Phosphate-Binding Loop

To further reveal the different secondary structures of the P-loop in the three different STK17B states, the defined secondary structure of proteins (DSSP) (Lei et al., 2019) method was used to analyse the secondary structural elements of residues Tyr32–Ser55. **Figure 6** shows the secondary structural profile of residues Tyr32–Ser55 for the three systems. In both the apo (**Figure 6A**) and ADP-bound (**Figure 6B**) systems, the residues Ile33–Arg41 and Val46–Ile51 formed two extended strands (β1 and β2) and residues Gly42–Ala45 at the P-loop adopted the bend conformation. These secondary structural elements of the β1, P-loop and β2 in the apo and ADP-bound states are consistent with the typical protein kinases at the corresponding position. In sharp contrast, in the ligand-bound state (**Figure 6C**), the secondary structural conformation of the β-strand in the residues Ile33–Arg41 and Val46–Ile51 was disturbed, especially the residues Ile33–Arg41 in the disordered conformation. Together, DSSP results indicated that the conformational changes of residues Ile33–Arg41 induced by the ligand binding may have an important role in the control of inhibitor selectivity to the STK17B.

Community Network Analysis

We next performed community network analysis to reveal the altered community networks of STK17B in the apo, ADP-bound, and ligand-bound states. The whole simulated trajectories were selected for community network analysis. The two Cα atoms within a cut-off distance of 4.5 Å that has an occupation time >75% of simulation time were classified into the same community (Sethi et al., 2009; Liang et al., 2020; Li et al., 2021a; Foutch et al., 2021; Tian et al., 2021). Each community was represented by coloured circles whose size is related to the number of residues it includes. The strength of the two communities was represented by the width of sticks that connect inter-communities.

Figure 7 shows the communities of different STK18B states. In the apo system (**Figure 7A**), there has nine communities. The community **1** contains the P-loop, the helix αC, and the β3-β5. The community **2** consists of the helix αD and the β6-β7. The community **9** largely includes the A-loop. There was the existence of strong connection between the community **1** and community **2** and between the community **1** and community **9**. In contrast, the communication between the community **1** and community **9** was weak. This observation indicated that there was no information flow between the P-loop and the A-loop in the apo system. In the ADP-bound system (**Figure 7B**), the community **1** diminished, which only consists of the helix αC. The sizes of the community **2** and community **9** in the ADP-bound system were similar to those in the apo system. However, the information flow that connects between the community **1** and community **2** and between the community **1** and community **9** was markedly weaker in the ADP-bound system than in the apo system. This indicated that upon ADP binding to the STK17B, the inter-domain interaction between the P-loop in the N-lobe and the helix αD in the C-lobe

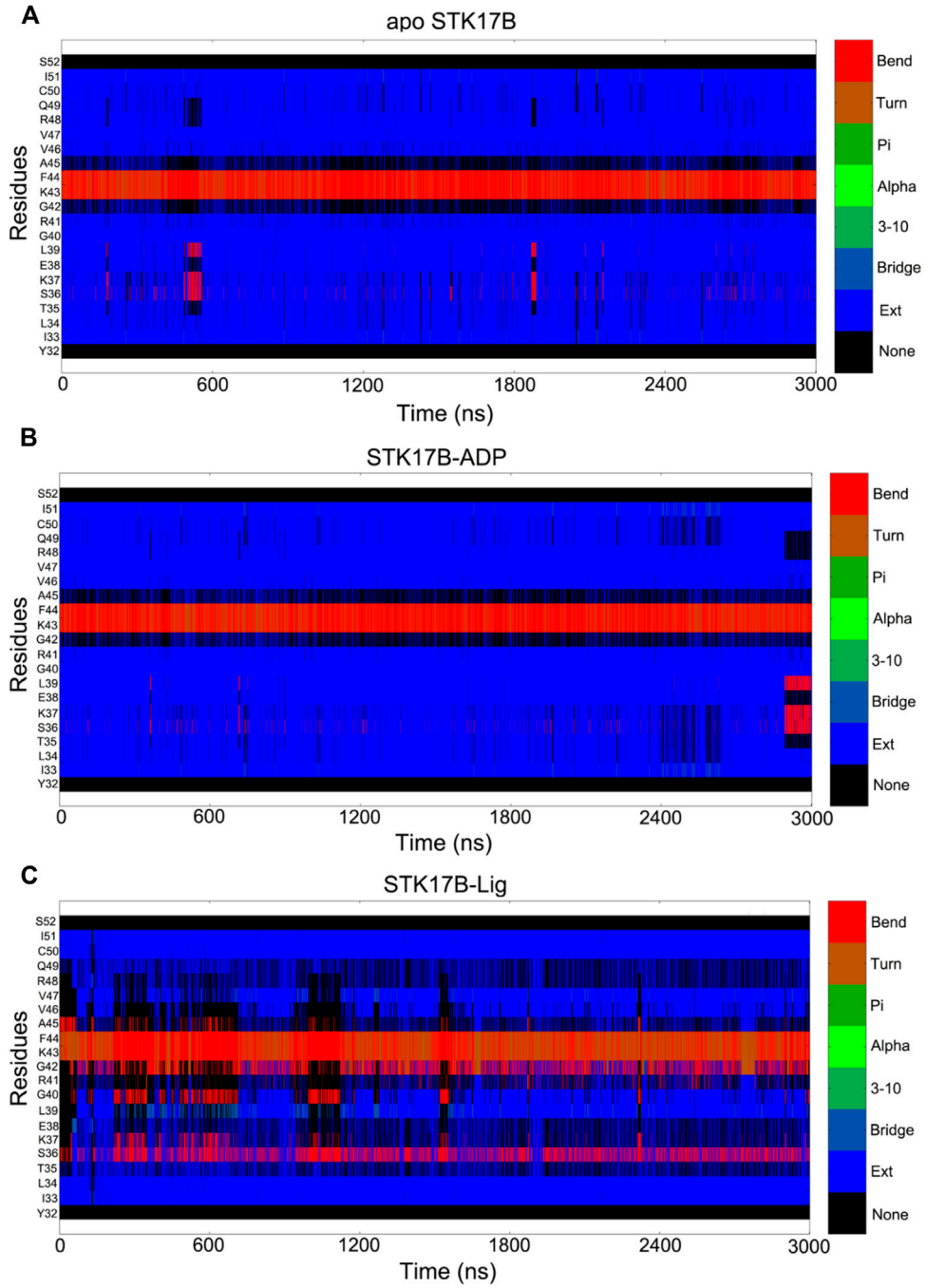


FIGURE 6 | Secondary structural element analysis as a function of simulation time for residues Tyr32 to Ser52 in the apo **(A)**, ADP-bound **(B)**, and ligand-bound **(C)** systems as calculated using the defined secondary structure of proteins (DSSP) method.

became weaken compared to the apo system. In the ligand-bound system (**Figure 7C**), the community **1** was enlarged compared to the ADP-bound systems, which was the same with the apo system. The community **1** in the ligand-bound systems consists of the P-loop, the helix αC, and the β3-β5. More significantly, the communication between the community 1 and community **2** in the ligand-bound system was enhanced compared to the ADP-bound system, with the strength resembling to the apo system. This observation suggested that upon ligand binding to the ADP-bound site, the information flow between the P-loop in the N-lobe and the helix αD in the C-lobe became stronger compared to the ADP-bound system. This enhanced interactions between the two lobes may promote inhibitor binding and selectivity to the STK17B.

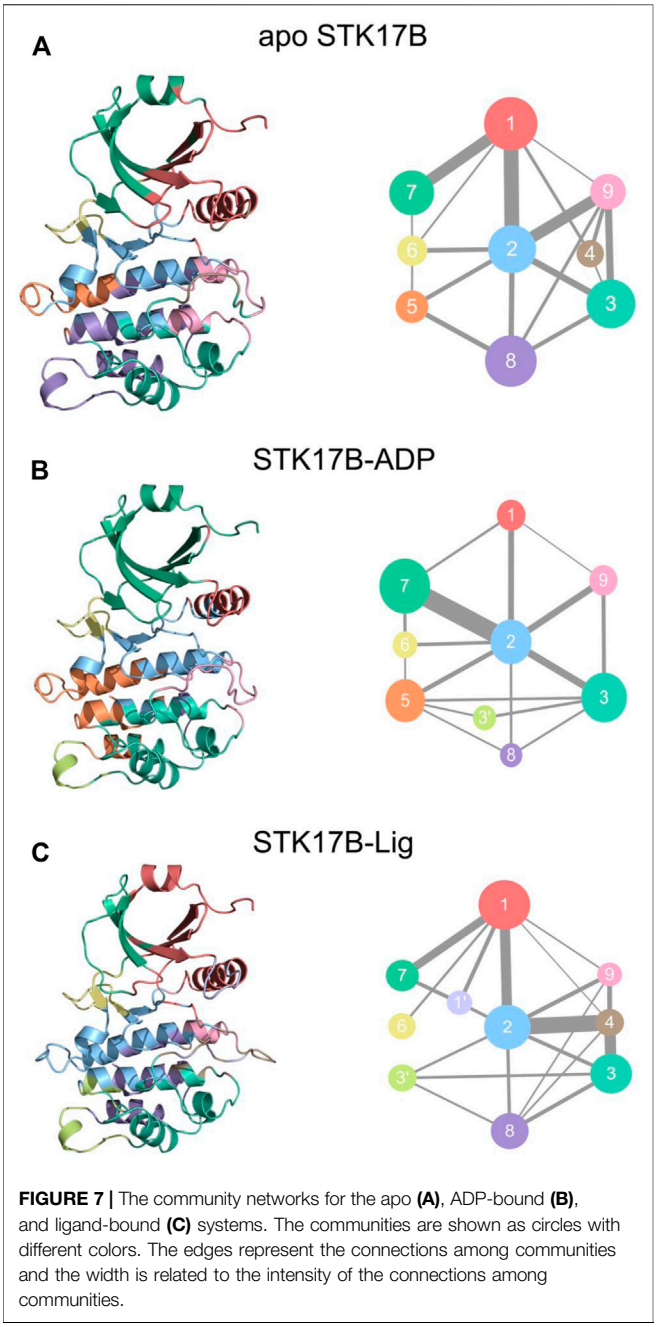


FIGURE 7 | The community networks for the apo **(A)**, ADP-bound **(B)**, and ligand-bound **(C)** systems. The communities are shown as circles with different colors. The edges represent the connections among communities and the width is related to the intensity of the connections among communities.

Comparative Binding Modes

Community network analysis implied the strong interactions between the N- and C-lobes in response to the ligand binding. To further elucidate the conformational arrangement of the two lobes of the protein kinase and the detailed interactions of ADP and the ligand with the STK17B, the most representative conformation of the STK17B-ligand and STK17B-ADP complexes was obtained using the cluster analysis of the three simulated trajectories (Liu et al., 2018; Xie et al., 2019). As shown in **Figure 8A**, in the ligand-bound state, there was a significantly disordered conformation of the P-loop, especially the β1, which was in good agreement with the DSSP results. Owing to the disordered P-loop conformation, the Arg41 at the β1 was flipped into the ADP-binding site and formed hydrogen bonding or salt bridge interactions with the residues Glu117 and Asn163 at the C-lobe and the carboxylic acid of the ligand. The hydrogen bonding occupation percentage was summarized in the **Supplementary Table S1**. These interactions promoted the strong communication between the N- and C-lobes, which contributed to increase the selectivity profile of the ligand to the STK17B. Simultaneously, the carboxylic acid of the ligand also interacted with the catalytic residue Lys62 through a salt bridge. Lys62 in turn formed salt bridge interactions with the Glu80 at the helix αC. In addition, the N1 of the thieno[3,2-d]pyrimidine formed a hydrogen bond with the amide backbone of Ala113 at the hinge domain. In contrast, in the ADP-bound state (**Figure 8B**), the β1 and β2 formed two anti-paralleled strands, which was consistent with the DSSP results. Owing to the ordered P-loop conformation, the Arg41 at the β1 was protruded into the solvent and had no interactions with the C-lobe, which was markedly different from that in the ligand-bound state. In the hinge domain, the backbone of residues Glu111 and Ala113 formed two hydrogen bonds with the adenine moiety of ADP. The hydrogen bonding occupation percentage was summarized in the **Supplementary Table S2**. The catalytic residue Lys62 formed salt bridges with the α- and β-phosphate moieties of ADP and the Mg^{2+} ion was coordinated with the α- and β-phosphate moieties, the carboxylic moiety of Asp179, and the carbonyl moiety of Asn163. Collectively, the comparative binding modes of ADP and the ligand with the STK17B highlighted that the unique *p* conformation induced by the ligand binding played a determined role in the increased selectivity of the ligand to the protein kinase. Given that the important role of the salt bridge interactions between the carboxylic acid moiety of the ligand and Arg41, it is advisable to retain the carboxylic acid moiety in the future drug design toward STK17B.

CONCLUSION

In the present study, the collective sampling of 27 μs MD simulations, coupled with the PCA, correlated motion analysis, DSSP, and community network analysis, revealed the effect of the conformational dynamics of the P-loop on the inhibitor selectivity profile to the STK17B. Ligand binding contributed to the increase of the conformational plasticity of the STK17B. Compared to the apo and ADP-bound STK17B, the P-loop, especially the β1, adopted the disordered conformation in the presence of the ligand. This unusual P-loop conformation rendered the residue Arg41 at the β1 flipping into the ADP-binding site and interacted with the carboxylic acid moiety of the ligand and residues Glu117 and Asn163 the C-lobe. These interactions in the ligand-bound state enhanced the information flow between the N- and C-lobes as observed by the community network analysis, which played an essential role in the control of the inhibitor selectivity to the STK17B. Owing to the importance of the salt bridge interactions between the carboxylic acid moiety of the ligand and Arg41 in the maintenance of the unique, disordered P-loop conformation,

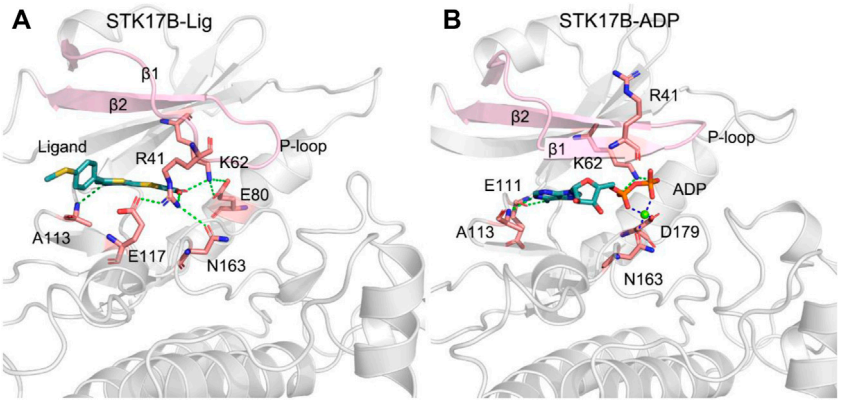


FIGURE 8 | The most representative structural complexes of ligand-bound **(A)** and ADP-bound **(B)** STK17B. The β1 and β2 and the P-loop are colored by pink. Hydrogen bonds or salt bridges are shown by green dotted lines. Coordinated bonds are shown by blue dotted lines.

the carboxylic acid moiety is suggested to retain in the future drug design toward STK17B. These results shed light on the structural basis of the selectivity of the inhibitor to the STK17B, which may be useful for the design of highly selective inhibitors to other protein kinases.

MATERIALS AND METHODS

System Preparation

The co-crystal structures of STK17B in complex with ADP (PDB ID: 6QF4) (Lieske et al., 2019) or PFE-PKIS 43 (PDB ID: 6Y6F) (Picado et al., 2020) were respectively downloaded from the Protein Data Bank (PDB). The missing residues E191–E194 in the 6QF4 and C187–I195 in the 6Y6F at the A-loop were modelled using the MODELLER program (Webb and Sali, 2014). The ADP molecule in the 6QF4 was removed to serve as the ligand-free STK17B (apo STK17B).

The force field parameters for ADP and Mg^{2+} were obtained from the AMBER parameter database (www.amber.manchester.ac.uk) and the generalized AMBER force field (GAFF) (Wang et al., 2004) was used for PFE-PKIS 43. Partial changes for PFE-PKIS 43 were computed using the RESP HF/6-31G* method (Bayly et al., 1993) through the antechamber module in AMBER 18 (Case et al., 2005) and Gaussian 09 program. The AMBER ff14SB (Maier et al., 2015) force field was used for the protein and the TIP3P model was used for water molecules (Jorgensen et al., 1983). The three simulated systems were embedded in a truncated octahedron TIP3P explicit water box with a boundary of 10 Å, while counterions Na^+ were added to neutralize the total charge. Then, 0.15 mol/L NaCl were added to simulate the physiological environment.

Molecular Dynamics Simulations

MD simulations were carried out using the AMBER 18 program (Case et al., 2005). Two rounds of minimizations of the three simulated systems were performed, including the steepest descent and conjugate gradient algorithms. This simulation protocol has also been employed in recent studies of protein conformational dynamics (Lu et al., 2019c; An et al., 2021; Liu et al., 2021; Zhang et al., 2022b). Then, each system was heated up from 0 to 300 K within 1 ns of MD simulations in the canonical ensemble (NVT), imposing position restraints of 100 kcal/mol·A^2 on the solute atoms. Finally, three replicas of independent 3 μs simulations were performed with random velocities under isothermal isobaric (NPT) conditions. An integration time step of 2 fs was used. The SHAKE algorithm was used to constrain all bond lengths involving hydrogen atoms (Ryckaert et al., 1977). The particle mesh Ewald (PME) method was used to treat with the long-range electrostatic interactions (Darden et al., 1993), while a 10 Å non-bonded cut-off was used for the short-range electrostatics and van der Waals interactions.

Principal Component Analysis

Principal component analysis (PCA) has been widely used to elucidate large-scale collective motions of biological macromolecules during MD simulations (Li et al., 2020b; Li et al., 2021b; Feng et al., 2021), which can transform a series of potentially coordinated observations into orthogonal vectors to capture large-amplitude motions. Among these vectors, the first two principal component (named PC1 and PC2) provide the dominant motions during MD simulations. In PCA, PCs were generated based on coordinate covariance matrix of Cα atoms in the STK17B protein and these collected frames were all projected on the PC1 and PC2.

Generalized Correlation Analysis

Generalized correlation (GC_{ij}) analysis was performed to monitor the correlated motions of residues (He et al., 2022; Wang et al., 2022; Zhuang et al., 2022). To describe that how much information of one atom was provided by another atom, Mutual Information (MI) was calculated using the **Eq. 1**:

$$MI\left[x_i, x_j\right] = \iint p\left(x_i, x_j\right) ln \frac{p\left(x_i, x_j\right)}{p\left(x_i\right) p\left(x_j\right)} dx_i dx_j \qquad (1)$$

The equation can be calculated using the known measure of entropy as the **Eq. 2**:

$$H[x] = \int p(x) \ln p(x) dx \quad (2)$$

The correlation between pairs of atoms x_i and x_j can be calculated using the marginal Shannon entropy $H[x_i]$, $H[x_j]$, and the joint entropy term $H[x_i, x_j]$ as the **Eq. 3**:

$$MI[x_i, x_j] = H[x_i] + H[x_j] - H[x_i, x_j] \quad (3)$$

The $MI[x_i, x_j]$ values can be further normalised to obtain the normalised generalised correlation coefficients (GC_{ij}) as the **Eq. 4**:

$$GC_{ij} = \left\{ 1 - e^{-\frac{2MI[x_i, x_j]}{d}} \right\}^{-\frac{1}{2}} \quad (4)$$

where d represents the dimensionality of x_i and x_j.

Cross-Correlation Analysis

Based on Pearson coefficients between the fluctuations of the Cα atoms, the cross-correlation matrix (CC_{ij}) was calculated to describe the coupling of the motions between the protein residues (Li et al., 2020b; Aledavood et al., 2021; Hernández-Alvarez et al., 2021; Wang et al., 2021). CC_{ij} was computed using the following **Eq. 5**,

$$C(i, j) = \frac{c(i, j)}{c(i, i)^{1/2} c(j, j)^{1/2}} \quad (5)$$

The positive CC_{ij} values indicate the two atoms i and j moving in the same direction, whereas the negative CC_{ij} values indicate the anti-correlated motions between the two atoms i and j.

Community Network Analysis

Community network was analyzed to uncover the inter-community interactions using the Network View plugin in VMD (Sethi et al., 2009; Marasco et al., 2021). In this analysis, the Cα atoms in the STK17B were selected as nodes to represent their corresponding residues. Edges were described between nodes whose distances are within a cut-off of 4.5 Å occupying >75% of simulation time. The edge between nodes was calculated using the **Eq. 6**:

$$d_{i,j} = -\log\left(|C_{i,j}|\right) \quad (6)$$

where i and j represent the two nodes.

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

Conceptualization: ZC and QW. Methodology: CL and ZhL. Validation: CL, ZhL, and ZoL. Formal analysis: CL, ZhL, and ZoL. Investigation: CL, ZhL, ZoL, and SY. Writing—Original draft preparation: CL. Writing—review and editing: ZC. Visualization: CL, ZhL, and ZC. Supervision: ZC and QW. Project administration, ZC and QW. Funding acquisition: ZC All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by National Natural Science Foundation of China (No. 82172846).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.901603/full#supplementary-material

REFERENCES

Aledavood, E., Forte, A., Estarellas, C., and Javier Luque, F. (2021). Structural Basis of the Selective Activation of Enzyme Isoforms: Allosteric Response to Activators of β1- and β2-containing AMPK Complexes. *Comput. Struct. Biotechnol. J.* 19, 3394–3406. doi:10.1016/j.csbj.2021.05.056

An, X., Bai, Q., Bing, Z., Liu, H., and Yao, X. (2021). Insights into the Molecular Mechanism of Positive Cooperativity Between Partial Agonist MK-8666 and Full Allosteric Agonist AP8 of hGPR40 by Gaussian Accelerated Molecular Dynamics (GaMD) Simulations. *Comput. Struct. Biotechnol. J.* 19, 3978–3989. doi:10.1016/j.csbj.2021.07.008

Attwood, M. M., Fabbro, D., Sokolov, A. V., Knapp, S., and Schiöth, H. B. (2021). Trends in Kinase Drug Discovery: Targets, Indications and Inhibitor Design. *Nat. Rev. Drug Discov.* 20, 839–861. doi:10.1038/s41573-021-00252-y

Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* 97, 10269–10280. doi:10.1021/j100142a004

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

Chen, J., Wang, X., Pang, L., Zhang, J. Z. H., and Zhu, T. (2019). Effect of Mutations on Binding of Ligands to Guanine Riboswitch Probed by Free Energy Perturbation and Molecular Dynamics Simulations. *Nucleic Acids Res.* 47, 6618–6631. doi:10.1093/nar/gkz499

Chen, J., Zhang, S., Wang, W., Pang, L., Zhang, Q., and Liu, X. (2021). Mutation-Induced Impacts on the Switch Transformations of the GDP- and GTP-Bound K-Ras: Insights from Multiple Replica Gaussian Accelerated Molecular Dynamics and Free Energy Analysis. *J. Chem. Inf. Model.* 61, 1954–1969. doi:10.1021/acs.jcim.0c01470

Chen, X., Li, C., Wang, D., Chen, Y., and Zhang, N. (2020). Recent Advances in the Discovery of CK2 Allosteric Inhibitors: From Traditional Screening to Structure-Based Design. *Molecules* 25, 870. doi:10.3390/molecules25040870

Cohen, P., Cross, D., and Jänne, P. A. (2021). Kinase Drug Discovery 20 Years after Imatinib: Progress and Future Directions. *Nat. Rev. Drug Discov.* 20, 551–569. doi:10.1038/s41573-021-00195-4

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

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* 29, 1046–1051. doi:10.1038/nbt.1990

Feng, L., Lu, S., Zheng, Z., Chen, Y., Zhao, Y., Song, K., et al. (2021). Identification of an Allosteric Hotspot for Additive Activation of PPARγ in Antidiabetic Effects. *Sci. Bull.* 66, 1559–1570. doi:10.1016/j.scib.2021.01.023

Foutch, D., Pham, B., and Shen, T. (2021). Protein Conformational Switch Discerned via Network Centrality Properties. *Comput. Struct. Biotechnol. J.* 19, 3599–3608. doi:10.1016/j.csbj.2021.06.004

He, X., Du, K., Wang, Y., Fan, J., Li, M., Ni, D., et al. (2022). Autopromotion of K-Ras4B Feedback Activation through an SOS-Mediated Long-Range Allosteric Effect. *Front. Mol. Biosci.* 9. doi:10.3389/fmolb.2022.860962

He, X., Huang, N., Qiu, Y., Zhang, J., Liu, Y., Yin, X.-L., et al. (2021). Conformational Selection Mechanism Provides Structural Insights into the Optimization of APC-Asef Inhibitors. *Molecules* 26, 962. doi:10.3390/molecules26040962

Hernández-Alvarez, L., Oliveira Jr, A. B., Hernández-González, J. E., Chahine, J., Pascutti, P. G., de Araujo, A. S., et al. (2021). Computational Study on the Allosteric Mechanism of Leishmania Major IF4E-1 by 4E-Interacting Protein-1: Unravelling the Determinants of m7GTP Cap Recognition. *Comput. Struct. Biotechnol. J.* 19, 2027–2044. doi:10.1016/j.csbj.2021.03.036

Hu, X., Pang, J., Zhang, J., Shen, C., Chai, X., Wang, E., et al. (2022). Discovery of Novel GR Ligands toward Druggable GR Antagonist Conformations Identified by MD Simulations and Markov State Model Analysis. *Adv. Sci.* 9, 2102435. doi:10.1002/advs.202102435

Hyeon, C., Jennings, P. A., Adams, J. A., and Onuchic, J. N. (2009). Ligand-induced Global Transitions in the Catalytic Domain of Protein Kinase A. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3023–3028. doi:10.1073/pnas.0813266106

Jia, Y., Yun, C.-H., Park, E., Ercan, D., Manuia, M., Juarez, J., et al. (2016). Overcoming EGFR(T790M) and EGFR(C797S) Resistance with Mutant-Selective Allosteric Inhibitors. *Nature* 534, 129–132. doi:10.1038/nature17960

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869

Lan, Y., Han, J., Wang, Y., Wang, J., Yang, G., Li, K., et al. (2018). STK17B Promotes Carcinogenesis and Metastasis via AKT/GSK-3β/Snail Signaling in Hepatocellular Carcinoma. *Cell Death Dis.* 9, 236. doi:10.1038/s41419-018-0262-1

Lei, J., Qi, R., Tang, Y., Wang, W., Wei, G., Nussinov, R., et al. (2019). Conformational Stability and Dynamics of the Cancer-associated Isoform Δ133p53β Are Modulated by P53 Peptides and P53-specific DNA. *FASEB J.* 33, 4225–4235. doi:10.1096/fj.201801973R

Li, C., Zhang, X., Zhang, N., Zhou, Y., Sun, G., Zhao, L., et al. (2020). Identification and Biological Evaluation of CK2 Allosteric Fragments through Structure-Based Virtual Screening. *Molecules* 25, 237. doi:10.3390/molecules25010237

Li, X., Dai, J., Ni, D., He, X., Zhang, H., Zhang, J., et al. (2020a). Insight into the Mechanism of Allosteric Activation of PI3Kα by Oncoprotein K-Ras4B. *Int. J. Biol. Macromol.* 144, 643–655. doi:10.1016/j.ijbiomac.2019.12.020

Li, X., Qi, Z., Ni, D., Lu, S., Chen, L., and Chen, X. (2021a). Markov State Models and Molecular Dynamics Simulations Provide Understanding of the Nucleotide-dependent Dimerization-Based Activation of LRRK2 ROC Domain. *Molecules* 26, 5647. doi:10.3390/molecules26185647

Li, X., Wang, C., Peng, T., Chai, Z., Ni, D., Liu, Y., et al. (2021b). Atomic-scale Insights into Allosteric Inhibition and Evolutional Rescue Mechanism of Streptococcus Thermophilus Cas9 by the Anti-CRISPR Protein AcrIIA6. *Comput. Struct. Biotechnol. J.* 19, 6108–6124. doi:10.1016/j.csbj.2021.11.010

Li, X., Ye, M., Wang, Y., Qiu, M., Fu, T., Zhang, J., et al. (2020b). How Parkinson's Disease-Related Mutations Disrupt the Dimerization of WD40 Domain in LRRK2: A Comparative Molecular Dynamics Simulation Study. *Phys. Chem. Chem. Phys.* 22, 20421–20433. doi:10.1039/D0CP03171B

Liang, S., Wang, Q., Qi, X., Liu, Y., Li, G., Lu, S., et al. (2021). Deciphering the Mechanism of Gilteritinib Overcoming Lorlatinib Resistance to the Double Mutant I1171N/F1174I in Anaplastic Lymphoma Kinase. *Front. Cell Dev. Biol.* 9, 808864. doi:10.3389/fcell.2021.808864

Liang, Z., Verkhivker, G. M., and Hu, G. (2020). Integration of Network Models and Evolutionary Analysis into High-Throughput Modeling of Protein Dynamics and Allosteric Regulation: Theory, Tools and Applications. *Brief. Bioinform.* 21, 815–835. doi:10.1093/bib/bbz029

Lieske, J., Cerv, M., Kreida, S., Komadina, D., Fischer, J., Barthelmess, M., et al. (2019). On-Chip Crystallization for Serial Crystallography Experiments and On-Chip Ligand-Binding Studies. *Int. Union Crystallogr. J.* 6, 714–728. doi:10.1107/S2052252519007395

Liu, N., Zhou, W., Guo, Y., Wang, J., Fu, W., Sun, H., et al. (2018). Molecular Dynamics Simulations Revealed the Regulation of Ligands to the Interactions Between Androgen Receptor and its Coactivator. *J. Chem. Inf. Model.* 58, 1652–1661. doi:10.1021/acs.jcim.8b00283

Liu, Q., Wang, Y., Leung, E. L.-H., and Yao, X. (2021). In Silico study of Intrinsic Dynamics of Full-Length Apo-ACE2 and RBD-ACE2 Complex. *Comput. Struct. Biotechnol. J.* 19, 5455–5465. doi:10.1016/j.csbj.2021.09.032

Lu, S., Chen, Y., Wei, J., Zhao, M., Ni, D., He, X., et al. (2021a). Mechanism of Allosteric Activation of SIRT6 Revealed by the Action of Rationally Designed Activators. *Acta Pharm. Sin. B* 11, 1355–1361. doi:10.1016/j.apsb.2020.09.010

Lu, S., He, X., Ni, D., and Zhang, J. (2019a). Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *J. Med. Chem.* 62, 6405–6421. doi:10.1021/acs.jmedchem.8b01749

Lu, S., He, X., Yang, Z., Chai, Z., Zhou, S., Wang, J., et al. (2021b). Activation Pathway of a G Protein-Coupled Receptor Uncovers Conformational Intermediates as Targets for Allosteric Drug Design. *Nat. Commun.* 12, 4721. doi:10.1038/s41467-021-25020-9

Lu, S., Ji, M., Ni, D., and Zhang, J. (2018). Discovery of Hidden Allosteric Sites as Novel Targets for Allosteric Drug Design. *Drug Discov. Today* 23, 359–365. doi:10.1016/j.drudis.2017.10.001

Lu, S., Ni, D., Wang, C., He, X., Lin, H., Wang, Z., et al. (2019b). Deactivation Pathway of Ras GTPase Underlies Conformational Substates as Targets for Drug Design. *ACS Catal.* 9, 7188–7196. doi:10.1021/acscatal.9b02556

Lu, S., Shen, Q., and Zhang, J. (2019c). Allosteric Methods and Their Applications: Facilitating the Discovery of Allosteric Drugs and the Investigation of Allosteric Mechanisms. *Acc. Chem. Res.* 52, 492–500. doi:10.1021/acs.accounts.8b00570

Lu, S., and Zhang, J. (2017). Designed Covalent Allosteric Modulators: an Emerging Paradigm in Drug Discovery. *Drug Discov. Today* 22, 447–453. doi:10.1016/j.drudis.2016.11.013

Lu, S., and Zhang, J. (2019). Small Molecule Allosteric Modulators of G-Protein-Coupled Receptors: Drug-Target Interactions. *J. Med. Chem.* 62, 24–45. doi:10.1021/acs.jmedchem.7b01844

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. Theory Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Maloney, R. C., Zhang, M., Jang, H., and Nussinov, R. (2021). The Mechanism of Activation of Monomeric B-Raf V600E. *Comput. Struct. Biotechnol. J.* 19, 3349–3363. doi:10.1016/j.csbj.2021.06.007

Marasco, M., Kirkpatrick, J., Nanna, V., Sikorska, J., and Carlomagno, T. (2021). Phosphotyrosine Couples Peptide Binding and SHP2 Activation via a Dynamic Allosteric Network. *Comput. Struct. Biotechnol. J.* 19, 2398–2415. doi:10.1016/j.csbj.2021.04.040

Masterson, L. R., Shi, L., Metcalfe, E., Gao, J., Taylor, S. S., and Veglia, G. (2011). Dynamically Committed, Uncommitted, and Quenched States Encoded in Protein Kinase A Revealed by NMR Spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6969–6974. doi:10.1073/pnas.1102701108

Ni, D., Li, Y., Qiu, Y., Pu, J., Lu, S., and Zhang, J. (2020). Combining Allosteric and Orthosteric Drugs to Overcome Drug Resistance. *Trends Pharmacol. Sci.* 41, 336–348. doi:10.1016/j.tips.2020.02.001

Ni, D., Wei, J., He, X., Rehman, A. U., Li, X., Qiu, Y., et al. (2021). Discovery of Cryptic Allosteric Sites Using Reversed Allosteric Communication by a Combined Computational and Experimental Strategy. *Chem. Sci.* 12, 464–476. doi:10.1039/D0SC05131D

Nussinov, R., and Tsai, C.-J. (2015). The Design of Covalent Allosteric Drugs. *Annu. Rev. Pharmacol. Toxicol.* 55, 249–267. doi:10.1146/annurev-pharmtox-010814-124401

Nussinov, R., Zhang, M., Maloney, R., Tsai, C. J., Yavuz, B. R., Tuncbag, N., et al. (2022). Mechanism of Activation and the Rewired Network: New Drug Design Concepts. *Med. Res. Rev.* 42, 770–799. doi:10.1002/med.21863

Okeke, C. J., Musyoka, T. M., Sheik Amamuddy, O., Barozi, V., and Tastan Bishop, Ö. (2021). Allosteric Pockets and Dynamic Residue Network Hubs of Falcipain 2 in Mutations Including Those Linked to Artemisinin Resistance. *Comput. Struct. Biotechnol. J.* 19, 5647–5666. doi:10.1016/j.csbj.2021.10.011

Pearce, L. R., Komander, D., and Alessi, D. R. (2010). The Nuts and Bolts of AGC Protein Kinases. *Nat. Rev. Mol. Cell Biol.* 11, 9–22. doi:10.1038/nrm2822

Picado, A., Chaikuad, A., Wells, C. I., Shrestha, S., Zuercher, W. J., Pickett, J. E., et al. (2020). A Chemical Probe for Dark Kinase STK17B Derives its Potency and High Selectivity through a Unique P-Loop Conformation. *J. Med. Chem.* 63, 14626–14646. doi:10.1021/acs.jmedchem.0c01174

Qiu, Y., Yin, X., Li, X., Wang, Y., Fu, Q., Huang, R., et al. (2021). Untangling Dual-Targeting Therapeutic Mechanism of Epidermal Growth Factor Receptor (EGFR) Based on Reversed Allosteric Communication. *Pharmaceutics* 13, 747. doi:10.3390/pharmaceutics13050747

Rehman, A. U., Zhen, G., Zhong, B., Ni, D., Li, J., Nasir, A., et al. (2021). Mechanism of Zinc Ejection by Disulfiram in Nonstructural Protein 5A. *Phys. Chem. Chem. Phys.* 23, 12204–12215. doi:10.1039/d0cp06360f

Roskoski, R. (2021). Properties of FDA-Approved Small Molecule Protein Kinase Inhibitors: A 2021 Update. *Pharmacol. Res.* 165, 105463. doi:10.1016/j.phrs.2021.105463

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

Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical Networks in tRNA:protein Complexes. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6620–6625. doi:10.1073/pnas.0810961106

Shibata, T., Iwasaki, W., and Hirota, K. (2020). The Intrinsic Ability of Double-Stranded DNA to Carry Out D-Loop and R-Loop Formation. *Comput. Struct. Biotechnol. J.* 18, 3350–3360. doi:10.1016/j.csbj.2020.10.025

Tian, X., Liu, H., and Chen, H. F. (2021). Catalytic Mechanism of Butane Anaerobic Oxidation for Alkyl-coenzyme M Reductase. *Chem. Biol. Drug Des.* 98, 701–712. doi:10.1111/cbdd.13931

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, Y., Ji, D., Lei, C., Chen, Y., Qiu, Y., Li, X., et al. (2021). Mechanistic Insights into the Effect of Phosphorylation on Ras Conformational Dynamics and its Interactions with Cell Signaling Proteins. *Comput. Struct. Biotechnol. J.* 19, 1184–1199. doi:10.1016/j.csbj.2021.01.044

Wang, Y., Li, M., Liang, W., Shi, X., Fan, J., Kong, R., et al. (2022). Delineating the Activation Mechanism and Conformational Landscape of a Class B G Protein-Coupled Receptor Glucagon Receptor. *Comput. Struct. Biotechnol. J.* 20, 628–639. doi:10.1016/j.csbj.2022.01.015

Webb, B., and Sali, A. (2014). Protein Structure Modeling with MODELLER. *Methods Mol. Biol.* 1137, 1–15. doi:10.1007/978-1-4939-0366-5_1

Wu, P., Nielsen, T. E., and Clausen, M. H. (2015). FDA-approved Small-Molecule Kinase Inhibitors. *Trends Pharmacol. Sci.* 36, 422–439. doi:10.1016/j.tips.2015.04.005

Xie, T., Yu, J., Fu, W., Wang, Z., Xu, L., Chang, S., et al. (2019). Insight into the Selective Binding Mechanism of DNMT1 and DNMT3A Inhibitors: a Molecular Simulation Study. *Phys. Chem. Chem. Phys.* 21, 12931–12947. doi:10.1039/C9CP02024A

Zhang, H., Zhu, M., Li, M., Ni, D., Wang, Y., Deng, L., et al. (2022a). Mechanistic Insights into Co-Administration of Allosteric and Orthosteric Drugs to Overcome Drug-Resistance in T315I BCR-ABL1. *Front. Pharmacol.* 13, 862504. doi:10.3389/fphar.2022.862504

Zhang, M., Jang, H., and Nussinov, R. (2019). The Mechanism of PI3Kα Activation at the Atomic Level. *Chem. Sci.* 10, 3671–3680. doi:10.1039/c8sc04498h

Zhang, Q., Chen, Y., Ni, D., Huang, Z., Wei, J., Feng, L., et al. (2022b). Targeting a Cryptic Allosteric Site of SIRT6 with Small-Molecule Inhibitors that Inhibit the Migration of Pancreatic Cancer Cells. *Acta Pharm. Sin. B* 12, 876–889. doi:10.1016/j.apsb.2021.06.015

Zhuang, H., Fan, X., Ji, D., Wang, Y., Fan, J., Li, M., et al. (2022). Elucidation of the Conformational Dynamics and Assembly of Argonaute-RNA Complexes by Distinct yet Coordinated Actions of the Supplementary microRNA. *Comput. Struct. Biotechnol. J.* 20, 1352–1365. doi:10.1016/j.csbj.2022.03.001

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 30 May 2022
doi: 10.3389/fmolb.2022.922839

Uncovering the Mechanism of Drug Resistance Caused by the T790M Mutation in EGFR Kinase From Absolute Binding Free Energy Calculations

Huaxin Zhou[1,2], *Haohao Fu*[1,2], *Han Liu*[1,2], *Xueguang Shao*[1,2]* *and Wensheng Cai*[1,2]*

[1]Research Center for Analytical Sciences, Frontiers Science Center for New Organic Matter, College of Chemistry, Tianjin Key Laboratory of Biosensing and Molecular Recognition, State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin, China, [2]Haihe Laboratory of Sustainable Chemical Transformations, Tianjin, China

OPEN ACCESS

Edited by:
Weiliang Zhu,
Shanghai Institute of Materia Medica (CAS), China

Reviewed by:
Shan Chang,
Jiangsu University of Technology, China
Xuemei Pu,
Sichuan University, China
Xiao Jun Yao,
Macau University of Science and Technology, Macao SAR, China

***Correspondence:**
Xueguang Shao
xshao@nankai.edu.cn
Wensheng Cai
wscai@nankai.edu.cn

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: 18 April 2022
Accepted: 16 May 2022
Published: 30 May 2022

Citation:
Zhou H, Fu H, Liu H, Shao X and Cai W (2022) Uncovering the Mechanism of Drug Resistance Caused by the T790M Mutation in EGFR Kinase From Absolute Binding Free Energy Calculations. Front. Mol. Biosci. 9:922839. doi: 10.3389/fmolb.2022.922839

The emergence of drug resistance may increase the death rates in advanced non-small cell lung cancer (NSCLC) patients. The resistance of erlotinib, the effective first-line antitumor drug for NSCLC with the L858R mutation of epidermal growth factor receptor (EGFR), happens after the T790M mutation of EGFR, because this mutation causes the binding of adenosine triphosphate (ATP) to EGFR more favorable than erlotinib. However, the mechanism of the enhancement of the binding affinity of ATP to EGFR, which is of paramount importance for the development of new inhibitors, is still unclear. In this work, to explore the detailed mechanism of the drug resistance due to the T790M mutation, molecular dynamics simulations and absolute binding free energy calculations have been performed. The results show that the binding affinity of ATP with respect to the L858R/T790M mutant is higher compared with the L858R mutant, in good agreement with experiments. Further analysis demonstrates that the T790M mutation significantly changes the van der Waals interaction of ATP and the binding site. We also find that the favorable binding of ATP to the L858R/T790M mutant, compared with the L858R mutant, is due to a conformational change of the αC-helix, the A-loop and the P-loop of the latter induced by the T790M mutation. This change makes the interaction of ATP and P-loop, αC-helix in the L858R/T790M mutant higher than that in the L858R mutant, therefore increasing the binding affinity of ATP to EGFR. We believe the drug-resistance mechanism proposed in this study will provide valuable guidance for the design of drugs for NSCLC.

Keywords: absolute binding free energy calculation, Epidermal Growth Factor Receptor (EGFR), T790M mutation, drug resistance, molecular dynamics simulation, BFEE2

INTRODUCTION

Lung cancer is the leading cause of cancer-related deaths worldwide (Jemal et al., 2011). The most common form of lung cancer is non-small cell lung cancer (NSCLC), which accounts for about 80–85% of lung cancer (Sharma et al., 2007; Inamura, 2017). In NSCLC, overexpression of epidermal growth factor receptor (EGFR) or hyper-activating mutations in its kinase domain have been observed in at least 50% of cases (Normanno et al., 2006). EGFR is a transmembrane receptor protein

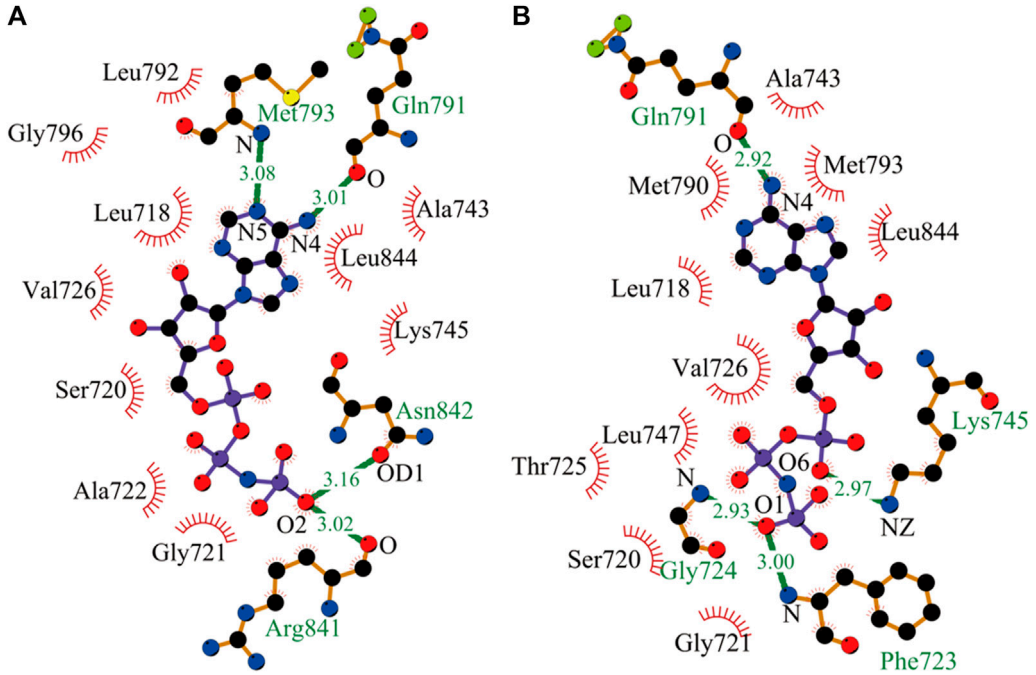


FIGURE 1 | Interactions of AMP with its nearby residues in EGFR mutants, plotted by LIGPLOT. **(A)** Analysis of hydrogen bond and hydrophobic interaction of binding interface in the L858R mutant. **(B)** The L858R/T790M mutant (Wallace et al., 1995; Laskowski and Swindells, 2011).

that has an essential role in cancer cell proliferation, survival, adhesion, migration, and differentiation by activating RAS/RAF/MEK/ERK and PI3K/AKT key downstream signaling pathways (Hirsch et al., 2003; Nagano et al., 2018; Zhou et al., 2022). In addition, among the currently marketed drugs, about 50–60% of drugs use membrane proteins to exert their effects (Santos et al., 2017). Therefore, EGFR and its mutations are one of the most valuable clinically validated drug targets for NSCLC treatment (Liao et al., 2010; Wee and Wang, 2017). A large number of small-molecule inhibitors acting on EGFR were developed to inhibit the kinase domain of EGFR and disrupt the oncogenic cell signaling by competing with adenosine triphosphate (ATP) for the binding site on the intracellular tyrosine kinase domain of EGFR. For example, first-generation EGFR inhibitor gefitinib or erlotinib is widely employed as first-line therapy for NSCLC with EGFR L858R mutation or exon 19 deletions. However, the secondary EGFR mutation T790M detected in NSCLC patients, can induce clinical resistance to gefitinib or erlotinib, greatly limiting the efficacy of these drugs in clinical use (Pao et al., 2005; Kosaka et al., 2006).

Understanding the mechanism of the T790M-induced drug resistance is important for further drug design. To this end, Kobayashi et al. proposed that the source of the acquired drug resistance was steric hindrance produced by the bulky methionine replaced the residue of threonine at position 790 (Kobayashi et al., 2005; Kwak et al., 2005; Pao et al., 2005). Interestingly, a later study demonstrated that the T790M resistance mutation increased the affinity of the receptor for ATP, which in turn diminished the potency of these ATP-competitive inhibitors (Yun et al., 2008). Several theoretical studies have been performed to explain the structural and energetic analyses of drug resistance conferred by the T790M mutation. Saldaña-Rivero and co-workers used the MM-GBSA approach to explain how L858R, T790M and L858R/T790M mutations impact the binding mechanism of ATP (Saldaña-Rivera et al., 2019). The popular MM/GBSA approach has been used to obtain a rough estimate of the binding free energy for a variety of complexes to explicate drug resistance (Zhang et al., 2019; Tan et al., 2022). A mechanistic explanation linking the mutations of the protein induce changes in the conformational free-energy landscape was also investigated by using massive molecular dynamics (MD) simulations together with parallel tempering, metadynamics, and one of the best force-fields available, showing a clear shift toward the active conformation for the T790M mutant and the L858R/T790M mutant (Sutto and Gervasio, 2013). The reason for the different binding affinities of ATP with respect to the L858R mutant and the L858R/T790M mutant, however, is still unclear. In addition, the relationship of the conformation changes of A-loop, αC-helix and P-loop and the difference of binding affinity remains to be further explored.

In this article, the standard binding free energies of ATP with respect to two EGFR mutants (L858R, L858R/T790M) have been calculated to investigate the mechanism of the drug resistance induced by the T790M mutation. Pair interaction calculations have been performed to reveal the driving force underlying the change of binding affinity of ATP to EGFR due to the T790M mutation and structural analysis has been carried out to capture the conformational change of the complex. The present study shows the essential reason for the drug resistance induced by the

TABLE 1 | Absolute binding free energies (in kcal/mol) for the ligand to EGFR mutants.

Contribution	L858R	Simulation time (ns)	L858R/T790M	Simulation time (ns)
$\triangle G_c^{site}$	−9.52 ± 0.66	20	−9.57 ± 0.28	30
$\triangle G_\Theta^{site}$	−0.58 ± 0.07	10	−0.42 ± 0.04	30
$\triangle G_\Phi^{site}$	−0.40 ± 0.04	20	−0.48 ± 0.08	30
$\triangle G_\Psi^{site}$	−0.35 ± 0.02	10	−0.45 ± 0.07	30
$\triangle G_\theta^{site}$	−0.11 ± 0.02	30	−0.23 ± 0.04	30
$\triangle G_\varphi^{site}$	−0.13 ± 0.01	30	−0.17 ± 0.02	30
$-\frac{1}{\beta}\ln(S*/*C^\circ)$	−11.01 ± 0.38	530	−10.43 ± 0.96	500
$\triangle G_c^{bulk}$	9.77 ± 0.11	20	8.36 ± 0.32	30
$\triangle G_o^{bulk}$	6.63	-	6.67	-
$\triangle G_{bind}^{o}$	−5.69 ± 0.48	670	−6.72 ± 0.91	710
$\triangle G_{bind}^{o}$ (exp)[a]	−5.25	-	−6.96	-

[a]*Experimental binding free energies [$\triangle G_{bind}^{o}$(exp)[a]] for L858R and L858R/T790M come from (Yun et al., 2008).*

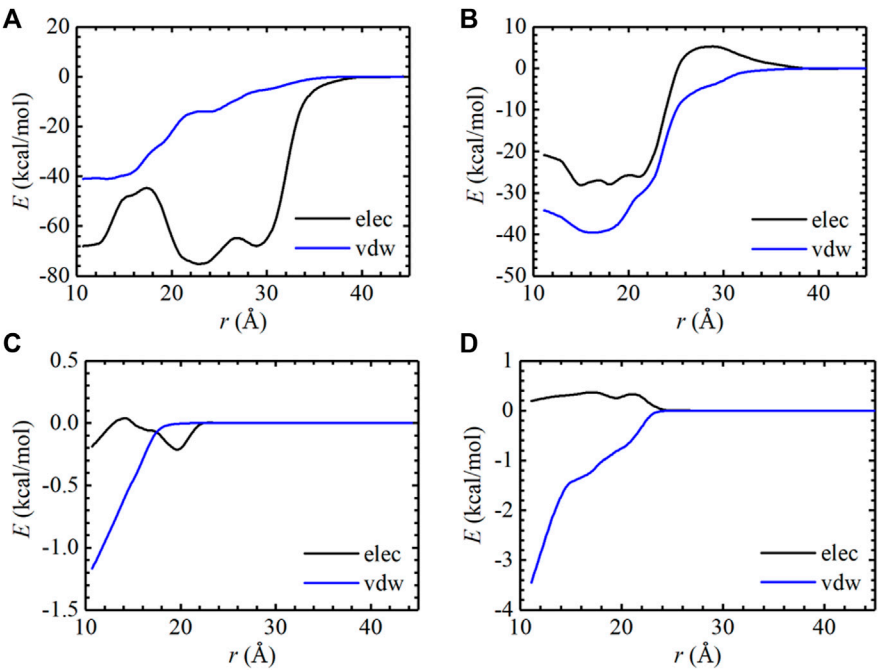


FIGURE 2 | Pair interaction energy for the separation of the L858R mutant: AMP **(A)** and the L858R/T790M mutant: AMP **(B)** were decoupled into electrostatic and van der Waals contributions. The pair interaction energy for the separation of the Thr790 residue: AMP **(C)** and the Met790 residue: AMP **(D)** were decoupled into electrostatic and van der Waals contributions.

T790M mutation, which can provide useful guidance for the further drug design against drug resistance.

METHODS

Structural Modeling

As the cocrystallized structure of EGFR or its mutants in a complex with ATP has not yet been solved, here, we adopted nonhydrolyzable analog AMP of ATP to carry out this research. The crystal structure of an EGFR L858R mutant kinase domain bound with the AMP molecule (PDB: 2EB3) as the structure template to model the EGFR L858R/T790M-AMP complex by CHARMM-GUI (Jo et al., 2008). Neither the protein nor the ligand was protonated. Missing residues in the retrieved structures were also examined and reconstructed using CHARMM-GUI. The atomic coordinates of the EGFR conformations were obtained from the Protein Data Bank (PDB) (http://www.pdb.org).

Molecular Dynamics Simulations

MD simulations for all EGFR models were performed using explicit-solvent periodic boundary conditions using NAMD (Phillips et al., 2020). Each model was solvated in a cubic box of TIP3P water, keeping a distance of 15 Å between the protein and the sides of the solvent box (Jorgensen et al., 1983). Each of

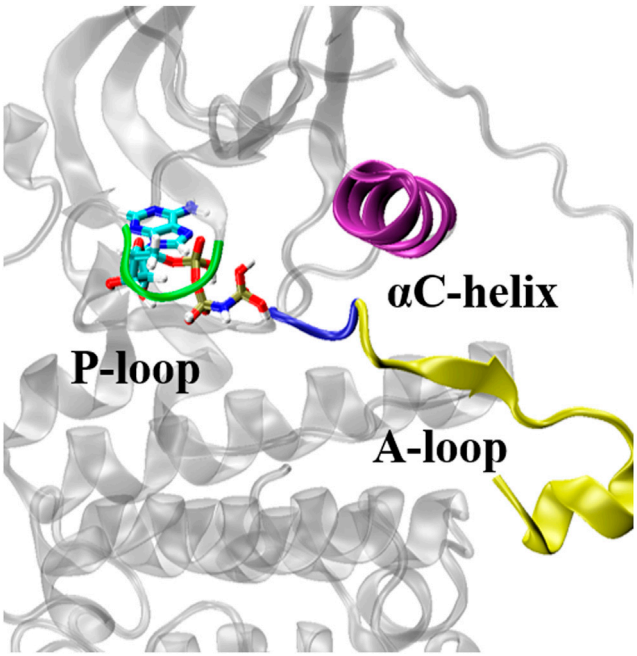


FIGURE 3 | Key structural elements structures of EGFR. The key structural elements are highlighted in purple (αC helix), yellow (A-loop), blue (the Asp-Phe-Gly motif) and green (P-loop).

the solvated systems was neutralized by adding enough chloride and sodium ions to give a concentration of 250 nM. The CHARMM36m protein force field was used to simulate all protein structures (Huang et al., 2017). The CHARMM General Force Field (CGenFF) force field was used to model the organic molecules (Vanommeslaeghe et al., 2010). All heavy atoms were restrained at the first stage of minimization. After that, the heavy atoms of ligand were fixed in the second step. Finally, all atoms in the system were minimized without any restraint. Production simulations were subsequently performed under the NPT condition at 300 K and 1.013 bar of the system. Temperature and pressure were held constant using Langevin dynamics and the Langevin piston (Uhlenbeck and Ornstein, 1930; Feller et al., 1995). All the trajectories were visualized using the VMD software (Humphrey et al., 1996).

Calculation of Standard Binding Free Energy

The binding free energy acts as a useful index to evaluate the binding affinity between mutants and drugs, and can be used as an important indicator of drug resistance (Zhou et al., 2013; Ma et al., 2015; Khan et al., 2020). In this article, the standard binding free-energy calculations of all systems were performed employing BFEE2 and following a geometrical route (Gumbart et al., 2013; Fu et al., 2021; Fu et al., 2022). BFEE2, which is a graphical user interface-based software, can automatically set up and analyze absolute binding free-energy calculations carried out with the popular MD engine NAMD (Fu et al., 2021; Fu et al., 2022). The calculation process of each protein-ligand complex was divided into eight independent subprocesses. Seven collective variables of geometrical restraints, that is, the root-mean-square deviation (RMSD) for describing the conformational change of the ligand in its bound state with respect to its native conformation, three Euler angles (Θ, Φ, Ψ) for describing the relative orientation of the ligand, the polar and azimuthal angles (θ, φ), together with the distance (r) between the center of mass of the ligand and that of the protein for describing its relative position (Fu et al., 2017), were introduced to accelerate the convergence of free-energy calculations. The contributions of the geometric restraints were evaluated by means of one-dimensional potential of mean force calculations carried out using the well-tempered meta-eABF (WTM-ABF) algorithm (Fu et al., 2016; Lesage et al., 2017; Fu et al., 2018; Fu et al., 2019).

RESULTS AND DISCUSSION

Structural Analysis of Ligand-Protein Complexes

Here, after the equilibrated simulations of all systems were completed, the intermolecular interactions of AMP with EGFR mutants were analyzed by the LIGPLOT program. As shown in **Figure 1**, AMP forms four specific hydrogen bonds with kinase polar residues Gln791, Met793, Arg841, and Asn842 of the L858R mutant and presents a wide hydrophobic contact interface with a number of kinase nonpolar residues, Leu792, Gly796, Val726, Leu718, Ser720, Ala722, Gly721, Ala743, Leu844, and Lys745. Interestingly, AMP also forms four hydrogen bonds with the L858R/T790M mutant, with an average distance shorter than those formed between AMP and the L858R mutant. However, these structural results may not completely explain the experimental observation from kinase assays that AMP has a higher binding affinity with the L858R/T790M mutant compared to the L858R one.

Absolute Binding Free Energy of AMP to EGFR Mutants

To evaluate the binding affinity of AMP with EGFR mutants, standard binding free-energy calculations were carried out on two complexes, i.e., AMP-L858R and AMP-L858R/T790M using the CHARMM36m force fields. The computed binding free-energy between AMP and EGFR kinase domains, with the contributions of geometric restraints acting on each degree of freedom, are reported in **Table 1**. The calculated standard binding free energies of AMP with respect to the L858R and the L858R/T790M double mutant are −5.69 kcal/mol and −6.72 kcal/mol, respectively. These estimates are in good agreement with the experimental values, namely, −5.25 kcal/mol and −6.96 kcal/mol, respectively, suggestive of a remarkable accuracy of BFEE2-based streamlined free-energy calculations. As expected, the binding affinity of AMP to EGFR increased by approximately 1.03 kcal/mol due to the T790M mutation. This result explains that the T790M substitution confers resistance by increasing the affinity for ATP, which was also demonstrated by (Yun et al., 2008). The one-dimensional free-energy profiles for the different

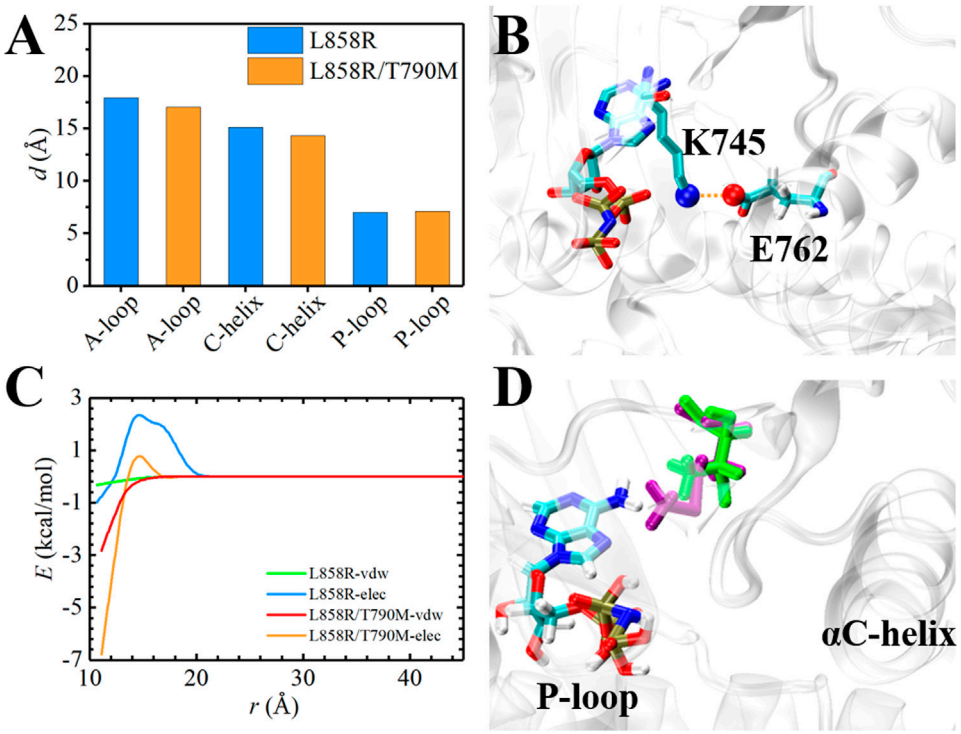


FIGURE 4 | Structural analysis of EGFR mutants and AMP. **(A)** Time-evolution of the average distance between A-loop, αC-helix and P-loop and AMP, respectively. **(B)** The interaction of E762 and K745 in L858R/T790M mutant. **(C)** The pair interaction energy for the separation between the ligand and αC-helix of the EGFR mutants. **(D)** Superposition and comparison between the structures of Thr790-AMP pair and mutant Met790-AMP pair. Thr790 and Met790 are colored in green and purple, respectively.

contributions are presented in **Supplementary Figures S1, S2**. As described in **Table 1**, the major contribution of the absolute binding free energies of AMP with the L858R mutant and L858R/T790M mutant was the -1/β ln ($S^*I^*C^\circ$) term in **Table 1**, which characterizes the separation of the protein and the ligand. The pair interaction energy for the separation was further decoupled into the van der Waals and electrostatic terms, as depicted in **Figures 2A,B**. It is apparent that electrostatic interactions constitute the driving force for the binding of AMP to the L858R mutant. Both van der Waals and electrostatic interactions, however, are critical to the binding of AMP to the L858R/T790M mutant. In addition, the energy profile characterizing AMP and residue 790 was analyzed. As shown in **Figures 2C,D**, the T790M mutation increases van der Waals interactions of AMP to EGFR. Based on these results, we conclude that the higher binding affinity of AMP to the L858R/T790M mutant, compared to the L858R one, probably because the T790M mutation increases the van der Waals interaction between AMP and EGFR.

Analysis of the Structural Conformational Changes Underlying the Increase of Binding Free Energy

The ATP-binding pocket is composed of a hinge region, A-loop, αC-helix, and P-loop (**Figure 3**), which are known to be crucial for their conformational stabilities and functional interactions with ATP (Johnson et al., 1996). The conformational changes of A-loop, P-loop, and αC-helix are important events occurring during kinase activation. In this section, we investigated the relationship between the structural changes of these key elements and binding affinity. We characterized the conformational changes of these key elements of EGFR by measuring the distance between these critical elements and ligand (Hu et al., 2022). As shown in **Figure 4A**, the location of AMP relative to A-loop, αC-helix and P-loop have a shorter distance in the L858R/T790M mutant, contributing to the favorable interactions that existed in the complex. Moreover, **Figure 4B** shows that the αC-helix is kept in place by a salt bridge formed by E762 and K745 in the L858R/T790M mutant, which is more stable than that observed in the L858R mutant (the average distance between N2 (Lys745) and CD (Glu762) of 3.05 Å vs. 7.86 Å, respectively). Additionally, the pair interaction energy for the separation was further decoupled into the van der Waals and electrostatic terms. As can be seen in **Figure 4C**, electrostatic interactions constitute the driving force for the binding of αC-helix to AMP in the L858R mutant. Although both electrostatic interactions and van der Waals interactions contribute to the binding of αC-helix to AMP in the L858R/T790M mutant, it is apparent that the effects of electrostatic interactions in higher than van der Waals interactions. The interactions of AMP and A-loop and P-loop are provided in **Supplementary Figure S3**. Further analysis revealed that the Met790 residue possesses a longer side

chain that can has a favorable contact with AMP compared with the Thr790 residue during the conformational change process (**Figure 4D**). This phenomenon is in agreement with the results of **Figures 2C,D**. Based on the discussion above, after the T790M mutation, the structural changes of αC-helix and P-loop mainly improve electrostatic interactions and van der Waals interactions, respectively. These are profitable to the binding affinity of AMP to EGFR.

CONCLUSION

Here, a powerful tool, BFEE2, was used to calculate the standard binding free energies of AMP to EGFR mutants. The results are well-consistent with the experiment. We found that the kinase affinity for AMP increased after the T790M mutation. In addition, our results indicate that electrostatic interaction plays a leading role in the binding of AMP to the L858R mutant, while both electrostatic interaction and van der Waals interaction are equally important for the binding of AMP to the L858R/T790M mutant. The present work emphasizes that the increased affinity of AMP to the L858R/T790M mutant compared with the L858R mutant is due to better stabilization of the active state for the mutant. This change may increase the interactions of AMP and P-loop, αC helix after the T790M mutation, therefore enhancing the binding affinity of AMP to EGFR. Although the calculated standard binding free energies are in good agreement with experimental values, there are challenges in the calculation of the standard binding free energies of EGFR inhibitors, especially for some of the fourth generation EGFR inhibitors without accurate binding sites. Still, the present work offers a perspective of the binding affinity of AMP to EGFR mutants and opens an avenue for further exploration of anticancer drugs acting on the EGFR to overcome drug resistance caused by the T790M mutation.

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

HZ, HF, XS, and WC designed the project. HZ performed all the MD simulations and analyses. HZ and HF wrote the manuscript. HL participated in the writing and discussion of the manuscript. All authors participated in editing the manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (Grants 22073050, 22174075, and 22103041), the China Post-doctoral Science Foundation (Grants bs6619012).

ACKNOWLEDGMENTS

We thank the Haihe Laboratory of Sustainable Chemical Transformations for financial support (ZYTS202105).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.922839/full#supplementary-material

REFERENCES

Feller, S. E., Zhang, Y., Pastor, R. W., and Brooks, B. R. (1995). Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* 103, 4613–4621. doi:10.1063/1.470648

Fu, H., Cai, W., Hénin, J., Roux, B., and Chipot, C. (2017). New Coarse Variables for the Accurate Determination of Standard Binding Free Energies. *J. Chem. Theory Comput.* 13, 5173–5178. doi:10.1021/acs.jctc.7b00791

Fu, H., Chen, H., Blazhynska, M., de Lacam, E. G. C., Szczepaniak, F., Pavlova, A., et al. (2022). Accurate Determination of Protein:ligand Standard Binding Free Energies from Molecular Dynamics Simulations. *Nat. Protoc.* 17, 1114–1141. doi:10.1038/s41596-021-00676-1

Fu, H., Chen, H., Cai, W., Shao, X., and Chipot, C. (2021). BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *J. Chem. Inf. Model.* 61, 2116–2123. doi:10.1021/acs.jcim.1c00269

Fu, H., Shao, X., Cai, W., and Chipot, C. (2019). Taming Rugged Free Energy Landscapes Using an Average Force. *Acc. Chem. Res.* 52, 3254–3264. doi:10.1021/acs.accounts.9b00473

Fu, H., Shao, X., Chipot, C., and Cai, W. (2016). Extended Adaptive Biasing Force Algorithm. An On-The-Fly Implementation for Accurate Free-Energy Calculations. *J. Chem. Theory Comput.* 12, 3506–3513. doi:10.1021/acs.jctc.6b00447

Fu, H., Zhang, H., Chen, H., Shao, X., Chipot, C., and Cai, W. (2018). Zooming across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys. *J. Phys. Chem. Lett.* 9, 4738–4745. doi:10.1021/acs.jpclett.8b01994

Gumbart, J. C., Roux, B., and Chipot, C. (2013). Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* 9, 794–802. doi:10.1021/ct3008099

Hirsch, F. R., Varella-Garcia, M., Bunn, P. A., Di Maria, M. V., Veve, R., Bremnes, R. M., et al. (2003). Epidermal Growth Factor Receptor in Non-small-cell Lung Carcinomas: Correlation between Gene Copy Number and Protein Expression and Impact on Prognosis. *Jco* 21, 3798–3807. doi:10.1200/Jco.2003.11.069

Hu, X., Pang, J., Zhang, J., Shen, C., Chai, X., Wang, E., et al. (2022). Discovery of Novel GR Ligands toward Druggable GR Antagonist Conformations Identified by MD Simulations and Markov State Model Analysis. *Adv. Sci.* 9, 2102435. doi:10.1002/advs.202102435

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., et al. (2017). CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods.* 14, 71–73. doi:10.1038/Nmeth.4067

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

Inamura, K. (2017). Lung Cancer: Understanding its Molecular Pathology and the 2015 WHO Classification. *Front. Oncol.* 7, 193. doi:10.3389/fonc.2017.00193

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global Cancer Statistics. *CA A Cancer J. Clin.* 61, 69–90. doi:10.3322/caac.20107

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* 29, 1859–1865. doi:10.1002/jcc.20945

Johnson, L. N., Noble, M. E. M., and Owen, D. J. (1996). Active and Inactive Protein Kinases: Structural Basis for Regulation. *Cell* 85, 149–158. doi:10.1016/s0092-8674(00)81092-2

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

Khan, M. T., Ali, S., Zeb, M. T., Kaushik, A. C., Malik, S. I., and Wei, D.-Q. (2020). Gibbs Free Energy Calculation of Mutation in PncA and RpsA Associated with Pyrazinamide Resistance. *Front. Mol. Biosci.* 7, 52. doi:10.3389/fmolb.2020.00052

Kobayashi, S., Boggon, T. J., Dayaram, T., Jänne, P. A., Kocher, O., Meyerson, M., et al. (2005). EGFRMutation and Resistance of Non-small-cell Lung Cancer to Gefitinib. *N. Engl. J. Med.* 352, 786–792. doi:10.1056/NEJMoa044238

Kosaka, T., Yatabe, Y., Endoh, H., Yoshida, K., Hida, T., Tsuboi, M., et al. (2006). Analysis of Epidermal Growth Factor Receptor Gene Mutation in Patients with Non-small Cell Lung Cancer and Acquired Resistance to Gefitinib. *Clin. Cancer Res.* 12, 5764–5769. doi:10.1158/1078-0432.CCR-06-0714

Kwak, E. L., Sordella, R., Bell, D. W., Godin-Heymann, N., Okimoto, R. A., Brannigan, B. W., et al. (2005). Irreversible Inhibitors of the EGF Receptor May Circumvent Acquired Resistance to Gefitinib. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7665–7670. doi:10.1073/pnas.0502860102

Laskowski, R. A., and Swindells, M. B. (2011). LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* 51, 2778–2786. doi:10.1021/ci200227u

Lesage, A., Lelièvre, T., Stoltz, G., and Hénin, J. (2017). Smoothed Biasing Forces Yield Unbiased Free Energies with the Extended-System Adaptive Biasing Force Method. *J. Phys. Chem. B* 121, 3676–3685. doi:10.1021/acs.jpcb.6b10055

Liao, P. L., Cheng, Y. W., Li, C. H., Wang, Y. T., and Kang, J. J. (2010). 7-Ketocholesterol and Cholesterol-5α,6α-Epoxide Induce Smooth Muscle Cell Migration and Proliferation through the Epidermal Growth Factor Receptor/phosphoinositide 3-kinase/Akt Signaling Pathways. *Toxicol. Lett.* 197, 88–96. doi:10.1016/j.toxlet.2010.05.002

Ma, L., Wang, D. D., Huang, Y., Yan, H., Wong, M. P., and Lee, V. H. (2015). EGFR Mutant Structural Database: Computationally Predicted 3D Structures and the Corresponding Binding Free Energies with Gefitinib and Erlotinib. *BMC Bioinforma.* 16, 85. doi:10.1186/s12859-015-0522-3

Nagano, T., Tachihara, M., and Nishimura, Y. (2018). Mechanism of Resistance to Epidermal Growth Factor Receptor-Tyrosine Kinase Inhibitors and a Potential Treatment Strategy. *Cells* 7, 212. doi:10.3390/cells7110212

Normanno, N., De Luca, A., Bianco, C., Strizzi, L., Mancino, M., Maiello, M. R., et al. (2006). Epidermal Growth Factor Receptor (EGFR) Signaling in Cancer. *Gene* 366, 2–16. doi:10.1016/j.gene.2005.10.018

Pao, W., Miller, V. A., Politi, K. A., Riely, G. J., Somwar, R., Zakowski, M. F., et al. (2005). Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. *Plos Med.* 2, 225–235. doi:10.1371/journal.pmed.0020073

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* 153, 044130. doi:10.1063/5.0014475

Saldaña-Rivera, L., Bello, M., and Méndez-Luna, D. (2019). Structural Insight into the Binding Mechanism of ATP to EGFR and L858R, and T790M and L858R/T790 Mutants. *J. Biomol. Struct. Dyn.* 37, 4671–4684. doi:10.1080/07391102.2018.1558112

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., et al. (2017). A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discov.* 16, 19–34. doi:10.1038/nrd.2016.230

Sharma, S. V., Bell, D. W., Settleman, J., and Haber, D. A. (2007). Epidermal Growth Factor Receptor Mutations in Lung Cancer. *Nat. Rev. Cancer.* 7, 169–181. doi:10.1038/nrc2088

Sutto, L., and Gervasio, F. L. (2013). Effects of Oncogenic Mutations on the Conformational Free-Energy Landscape of EGFR Kinase. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10616–10621. doi:10.1073/pnas.1221953110

Tan, S., Zhang, Q., Wang, J., Gao, P., Xie, G., Liu, H., et al. (2022). Molecular Modeling Study on the Interaction Mechanism between the LRRK2 G2019S Mutant and Type I Inhibitors by Integrating Molecular Dynamics Simulation, Binding Free Energy Calculations, and Pharmacophore Modeling. *ACS Chem. Neurosci.* 13, 599–612. doi:10.1021/acschemneuro.1c00726

Uhlenbeck, G. E., and Ornstein, L. S. (1930). On the Theory of the Brownian Motion. *Phys. Rev.* 36, 823–841. doi:10.1103/physrev.36.823

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., et al. (2010). CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* 31, 671–690. doi:10.1002/jcc.21367

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995). LIGPLOT: A Program to Generate Schematic Diagrams of Protein-Ligand Interactions. *Protein Eng. Des. Sel.* 8, 127–134. doi:10.1093/protein/8.2.127

Wee, P., and Wang, Z. (2017). Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers* 9, 52. doi:10.3390/cancers9050052

Yun, C.-H., Mengwasser, K. E., Toms, A. V., Woo, M. S., Greulich, H., Wong, K.-K., et al. (2008). The T790M Mutation in EGFR Kinase Causes Drug Resistance by Increasing the Affinity for ATP. *Proc. Natl. Acad. Sci. U.S.A.* 105, 2070–2075. doi:10.1073/pnas.0709662105

Zhang, Q., An, X., Liu, H., Wang, S., Xiao, T., and Liu, H. (2019). Uncovering the Resistance Mechanism of *Mycobacterium tuberculosis* to Rifampicin Due to RNA Polymerase H451D/Y/R Mutations from Computational Perspective. *Front. Chem.* 7, 819. doi:10.3389/fchem.2019.00819

Zhou, H., Fu, J., Jia, Q., Wang, S., Liang, P., Wang, Y., et al. (2022). Magnetic Nanoparticles Covalently Immobilizing Epidermal Growth Factor Receptor by SNAP-Tag Protein as a Platform for Drug Discovery. *Talanta* 240, 123204. doi:10.1016/j.talanta.2021.123204

Zhou, W., Wang, D. D., Yan, H., Wong, M., and Lee, V. (2013). "Prediction of Anti-EGFR Drug Resistance Base on Binding Free Energy and Hydrogen Bond Analysis," in 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Singapore, 16-19 April 2013, 193–197. doi:10.1109/CIBCB.2013.6595408

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 08 June 2022
doi: 10.3389/fmolb.2022.898838

Investigating Intrinsically Disordered Proteins With Brownian Dynamics

Surl-Hee Ahn[1*†], *Gary A. Huber*[1,2†] *and J. Andrew McCammon*[1,2]

[1]*Department of Chemistry and Biochemistry, University of California, San Diego, San Diego, CA, United States,* [2]*Department of Pharmacology, University of California, San Diego, San Diego, CA, United States*

Intrinsically disordered proteins (IDPs) have recently become systems of great interest due to their involvement in modulating many biological processes and their aggregation being implicated in many diseases. Since IDPs do not have a stable, folded structure, however, they cannot be easily studied with experimental techniques. Hence, conducting a computational study of these systems can be helpful and be complementary with experimental work to elucidate their mechanisms. Thus, we have implemented the coarse-grained force field for proteins (COFFDROP) in Browndye 2.0 to study IDPs using Brownian dynamics (BD) simulations, which are often used to study large-scale motions with longer time scales and diffusion-limited molecular associations. Specifically, we have checked our COFFDROP implementation with eight naturally occurring IDPs and have investigated five $(Glu\text{-}Lys)_{25}$ IDP sequence variants. From measuring the hydrodynamic radii of eight naturally occurring IDPs, we found the ideal scaling factor of 0.786 for non-bonded interactions. We have also measured the entanglement indices (average C_α distances to the other chain) between two $(Glu\text{-}Lys)_{25}$ IDP sequence variants, a property related to molecular association. We found that entanglement indices decrease for all possible pairs at excess salt concentration, which is consistent with long-range interactions of these IDP sequence variants getting weaker at increasing salt concentration.

Keywords: Brownian dynamics simulation, molecular associations, intrinsically disordered proteins, COFFDROP force field, Browndye

OPEN ACCESS

Edited by:
Alexey V. Onufriev,
Virginia Tech, United States

Reviewed by:
Stefano Piana-Agostinetti,
D. E. Shaw Research, United States
Jung-Hsin Lin,
Academia Sinica, Taiwan

****Correspondence:***
Surl-Hee Ahn
s3ahn@ucsd.edu

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *17 March 2022*
Accepted: *17 May 2022*
Published: *08 June 2022*

Citation:
Ahn S-H, Huber GA and McCammon JA (2022) Investigating Intrinsically Disordered Proteins With Brownian Dynamics. Front. Mol. Biosci. 9:898838. doi: 10.3389/fmolb.2022.898838

1 INTRODUCTION

One of the main determinants of biological structure and function is the interaction of two or more molecules, especially protein molecules. Understanding the dynamics of these bimolecular interactions is important for the understanding of such cellular structures as the cytoskeleton (actin and tubulin, for example), ribosomes, chromosomes, and polymerases, as well as processes such as cell signaling and cell motility (Alberts et al., 2002; Pollard and Earnshaw, 2007). Furthermore, the encounter stages of such reactions, which are often the rate-limiting steps, are diffusion-limited (Elcock, 2004). Therefore, the use of Brownian dynamics (BD) is appropriate for such systems [see Huber and McCammon (2019) for a review]. For several decades, BD has found use in polymer and peptide simulations, simulations of enzyme-substrate reactions, and protein–protein association reactions. More recently BD has found use in studies of large-scale cytoplasm simulations, microtubule dynamics, assembly of protein complexes, retroviral infectivity, molecular motors, chromosome organization, the nuclear pore complex, synapses, and endocytosis. The previous version of the Browndye software package (Browndye 1.0), which was limited to two

rigid bodies, has been used in enzyme kinetics and channeling (Huang et al., 2018), as well as protein-protein interactions (Grant et al., 2011).

The Browndye 2.0 software package, successor to the previous simulation package, consists of two simulation programs and about 38 auxiliary programs for processing data. Like the previous version, Browndye 2.0 can compute the second-order rate constants of the encounter of two bodies moving according to BD, compute the probabilities of the two bodies moving from one binding mode to another, and output the molecules' trajectories. The main addition is the ability to model each molecule as a collection of large rigid cores with flexible connectors and loops. In its original two-rigid-body model, Browndye has functionality very similar to the packages SDA (Martinez et al., 2015), MacroDox (Northrup et al., 1993), and GeomBD (Roberts and Chang, 2016), and is intended primarily for simulations of large biological molecules like those three other packages. Its current limitations arise mainly from the structural rigidity approximations and the nature of the force computations between the molecules.

Using Browndye 2.0, we have investigated intrinsically disordered proteins (IDPs), which are proteins that do not have a stable, folded structure and instead take on various structures depending on their current tasks in modulating biological processes. Conducting a computational study of these systems will be critical to elucidate their mechanisms. Specifically, we have implemented the coarse-grained force field for proteins (COFFDROP) (Andrews and Elcock, 2014; Frembgen-Kesner et al., 2015) in Browndye 2.0 to study eight naturally occurring IDPs and five $(Glu\text{-}Lys)_{25}$ IDP sequence variants. We have measured their structural properties, including radius of gyration (R_g), interresidue distances (R_{ij}), and hydrodynamic radius (R_h), and a property related to molecular association, namely the entanglement index (average C_α distance to the other chain).

2 MATERIALS AND METHODS

2.1 Structure Preparation

The Alphafold Colab (Jumper et al., 2021) was used to prepare the starting structures for the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants and eight naturally occurring IDPs that were used in Frembgen-Kesner et al. (2015), which are Alzheimer amyloid $\beta_{(1-40)}$ ($A\beta_{(1-40)}$) (Danielsson et al., 2002), suppressor of Mec1 lethality (Sml1) (Danielsson et al., 2008), *Lotus japonicas* intrinsically disordered protein 1 (*Lj*IDP1) (Haaning et al., 2008), prothymosin α (ProTα) (Yi et al., 2007), abscisic acid stress ripening 1 (ASR1) (Goldgur et al., 2007), yeast nucleoporin 116 (Nup116) (Krishnan et al., 2008), α-synuclein (Uversky et al., 2001), and cystic fibrosis transmembrane conductance regulator regulatory region (CFTR R) (Baker, 2009). We have used Alphafold to prepare the starting structures for the naturally occurring IDPs since they have conditionally folded regions that have confident per-residue confidence scores (pLDDT) (above 70 in a range from 0 to 100), which are expected to be accurately predicted by Alphafold (Alderson et al., 2022). We have also used Alphafold for the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants since peptides composed of many Glu and Lys residues favor forming α-helical structures (Marqusee and Baldwin, 1987; Iqbalsyah and Doig, 2005; Meuzelaar et al., 2016; Wolny et al., 2017), and Alphafold had yielded α-helical structures for all five IDP sequence variants.

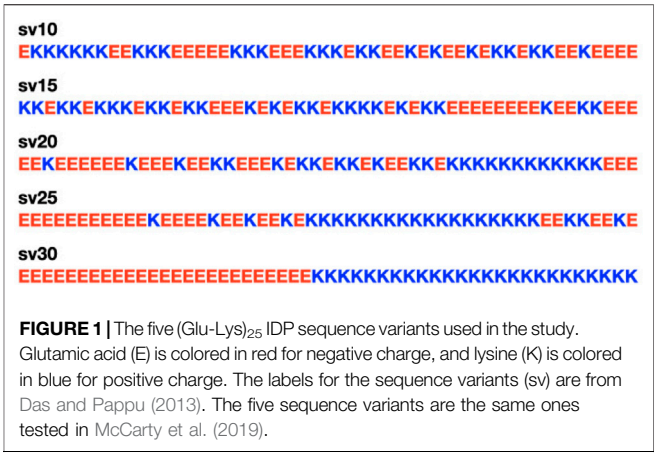


FIGURE 1 | The five $(Glu\text{-}Lys)_{25}$ IDP sequence variants used in the study. Glutamic acid (E) is colored in red for negative charge, and lysine (K) is colored in blue for positive charge. The labels for the sequence variants (sv) are from Das and Pappu (2013). The five sequence variants are the same ones tested in McCarty et al. (2019).

Figures 1, **2** show the amino acid sequences of these systems, respectively, and **Table 1** summarizes the various characteristics of the systems obtained from the classification of intrinsically disordered ensemble regions (CIDER) program (Holehouse et al., 2015).

The protonation states were assigned using PROPKA 3 (Olsson et al., 2011; Søndergaard et al., 2011) at pH 7.0 for the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants and at appropriate pH's for the eight IDPs as done in Frembgen-Kesner et al. (2015), which are listed in the **Supplementary Material**. PDB2PQR 3.4 (Jurrus et al., 2018; Unni et al., 2011; Dolinsky et al., 2007, 2004) was used to convert the PDB files to PQR format for the BD simulations. The temperature T was set to 298 K, and the dielectric constant was set to 78.4 for all systems. For the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants, the ionic concentration was set to NaCl 15 mM (reference concentration) or NaCl 125 mM (excess salt concentration) as done in Das and Pappu (2013) by setting the appropriate Debye length λ_D using **Equation 1**

$$\lambda_D = \left(\frac{\epsilon_0 \epsilon_r k_B T}{2e^2 N_A C}\right)^{1/2}, \tag{1}$$

where ϵ_0 is the permittivity of the free space, ϵ_r is the dielectric constant (of water in this case), k_B is the Boltzmann constant, T is the temperature (298 K in this case), e is the elementary charge, N_A is Avogadro's constant, and C is the ionic strength in mol/m^3 units. The Debye length λ_D was set to be 7.85 $\mathring{A}$ for NaCl 15 mM (reference concentration) and 2.72 $\mathring{A}$ for NaCl 125 mM (excess salt concentration).

2.2 Brownian Dynamics Simulations

The BD simulations were run using Browndye 2.0 (Huber and McCammon, 2010) with the spline-based potential coarse-grained force field for proteins (COFFDROP) (Andrews and

FIGURE 2 | The eight naturally occurring IDPs used in the study. Glutamic acid (E) and aspartic acid (D) are colored in red for negative charge, and lysine (K) and arginine (R) are colored in blue for positive charge. The eight IDPs are from Frembgen-Kesner et al. (2015).

TABLE 1 | Summary of classification of intrinsically disordered ensemble regions (CIDER) (Holehouse et al., 2015) results for the IDPs used in the study. NCPR denotes the net charge per residue, FCR denotes the fraction of charged residues, and κ denotes the measure of charge segregation from Das and Pappu (2013). Hydrophathy measures how hydrophobic the sequence is (0–9 with 0 being least hydrophobic and nine being most hydrophobic) (Kyte and Doolittle, 1982) and disorder measures the fraction of disorder promoting residues (Uversky, 2002). The categorization of each IDP is determined from the Das-Pappu phase diagram (Das and Pappu, 2013; Holehouse et al., 2015).

IDP	Length	NCPR	FCR	κ	Hydropathy	Disorder	Category
sv10	50	0.000	1.000	0.083	0.800	1.000	Strong polyampholytes
sv15	50	0.000	1.000	0.135	0.800	1.000	Strong polyampholytes
sv20	50	0.000	1.000	0.272	0.800	1.000	Strong polyampholytes
sv25	50	0.000	1.000	0.528	0.800	1.000	Strong polyampholytes
sv30	50	0.000	1.000	1.000	0.800	1.000	Strong polyampholytes
$A\beta_{(1-40)}$	40	-0.075	0.225	0.211	4.558	0.600	Weak polyampholytes
Sml1	104	-0.048	0.221	0.143	3.712	0.635	Weak polyampholytes
*Lj*IDP1	107	0.009	0.271	0.174	3.890	0.729	Janus sequences
ProTα	109	-0.394	0.578	0.424	2.507	0.881	Strong polyelectrolytes
ASR1	115	-0.017	0.383	0.100	3.326	0.809	Strong polyampholytes
Nup116	126	0.040	0.040	0.278	3.709	0.762	Weak polyampholytes
α-Synuclein	140	-0.064	0.279	0.172	4.097	0.729	Janus sequences
CFTR R	190	-0.026	0.289	0.285	3.743	0.679	Janus sequences

Elcock, 2014; Frembgen-Kesner et al., 2015), which was newly implemented for Browndye 2.0. In COFFDROP, each amino acid is represented as a "bead" so that a protein sequence can be represented as a flexible "chain" composed of beads. In addition, since the scaling of non-bonded interactions improved COFFDROP's ability to reproduce experimental results (Frembgen-Kesner et al., 2015), this feature was also implemented for Browndye 2.0. Moreover in Browndye 2.0, interactions can be computed less frequently, which is useful since computing these interactions take up most of the simulation time for longer chains. Finally in Browndye 2.0, a constant time step size can be set, and the recommended value is 0.05 ps for COFFDROP chains, unless bond constraints are used in which case a larger constant time step size is allowed, which can make the simulations run faster.

For the eight naturally occurring IDPs, the maximum number of BD simulation steps was set to 80,000,000, and a constant time step size of 0.05 ps was used (no bond constraints used). To calculate the hydrodynamic radius (R_h) for each COFFDROP potential with a scaling factor (0.5–1.0 in intervals of 0.1) for non-bonded interactions, ten trajectories were run for each system and potential, and simulation snapshots were recorded every 200,000 steps. Hydrodynamic interactions were updated every 400 steps. The specific parameter values follow the parameter values from Frembgen-Kesner et al. (2015) since these IDPs were used to check the COFFDROP implementation for Browndye 2.0.

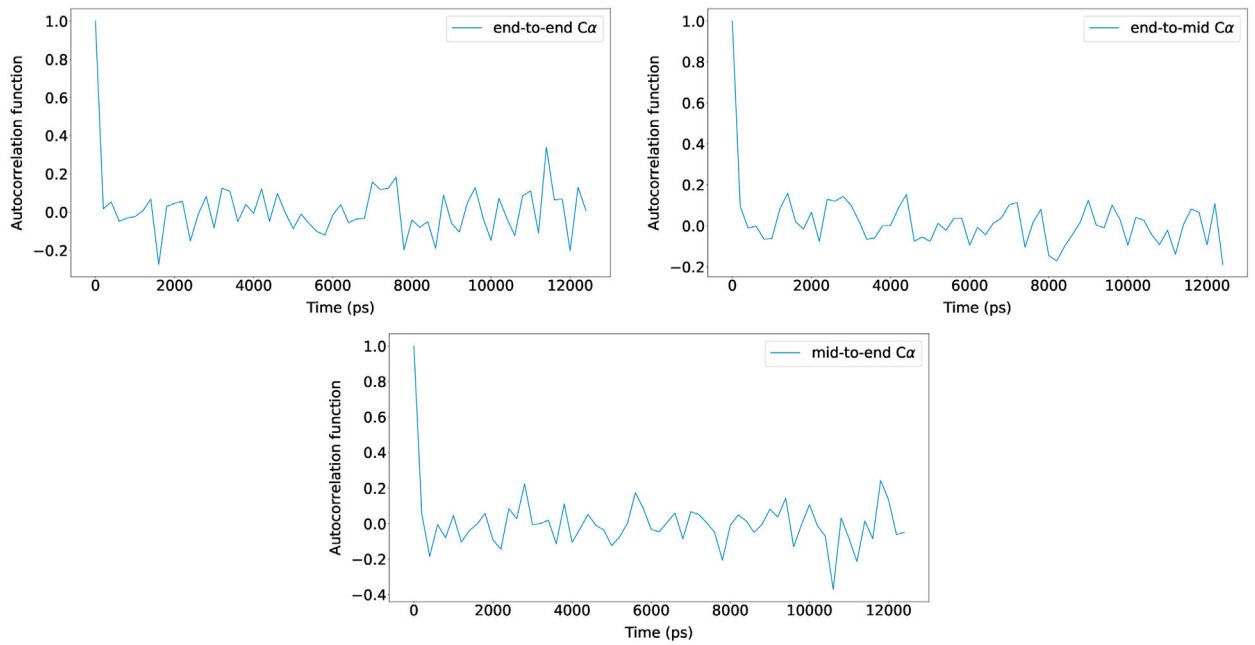


FIGURE 3 | Autocorrelation functions of N-terminus C_α to C-terminus C_α distance, N-terminus C_α to middle C_α distance, and middle C_α to C-terminus C_α distance for $(Glu\text{-}Lys)_{25}$ sv = 10 at NaCl 15 mM (reference concentration) with scaling factor 0.786. The autocorrelation functions from one of the ten trajectories are shown for clarity and the rest are included in the **Supplementary Material**.

For the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants, the maximum number of BD simulation steps was set to 125,000,000, and by using bond constraints, a constant time step size of 0.2 ps was used (25 μs total). The autocorrelation functions of N-terminus C_α to C-terminus C_α distance, N-terminus C_α to middle C_α distance, and middle C_α to C-terminus C_α distance were measured and plotted with new Browndye 2.0 functions chain_atom_distances and autocor to check whether the simulation time was sufficiently long enough to obtain converged properties. As seen in **Figure 3**, the three autocorrelation functions converge, and the simulation time was regarded to be sufficiently long enough. The rest of the autocorrelation functions are included in the **Supplementary Material**. As seen from the **Supplementary Material**, the shortest simulation was 12 ns, whereas the longest simulation was 25 µs. To calculate structural properties such as radius of gyration (R_g), ten trajectories were run for each system, and simulation snapshots were recorded every 100,000 steps. Hydrodynamic interactions were updated every 400 steps.

3 RESULTS

3.1 Eight Naturally Occurring IDPs

We first investigated the eight naturally occurring IDPs to see if Browndye 2.0 can reproduce the COFFDROP results in Frembgen-Kesner et al. (2015). In particular, we measured the hydrodynamic radius (R_h) for each system and COFFDROP potential with a scaling factor (0.5–1.0 in intervals of 0.1) for non-bonded interactions. R_h is the radius of a hard-sphere that diffuses at the same rate as solute and is dependent on the size and hydration of protein. The Kirkwood definition (Kirkwood, 1996) was used to calculate R_h as stated in **Equation 2**

$$\frac{1}{R_h} = \left\langle \frac{1}{r_{ij}} \right\rangle_{i \neq j}, \tag{2}$$

where r_{ij} denotes pairwise distances between C_α of amino acids i and j, as done in Nygaard et al. (2017). R_h was calculated for each simulation snapshot (every 200,000 steps), and the final R_h value for each simulation was obtained by averaging the R_h values from the simulation. The average R_h values, along with standard error bars (95% confidence interval), from ten independent simulations, are plotted in **Figure 4**. To match up with the COFFDROP results that used the HYDROPRO program (Ortega et al., 2011), the average R_h values and standard error bars were multiplied by 1.186 and added by 1.03 as done in Nygaard et al. (2017). The R_h values are in good agreement with those in Frembgen-Kesner et al. (2015), which are marked as dashed lines with square markers in **Figure 4**, indicating that the COFFDROP implementation in Browndye 2.0 is reliable. The small discrepancies between the two results could be from the long-range electrostatic interactions being computed differently, i.e., Frembgen-Kesner et al. (2015) used a treecode algorithm (Li et al., 2009) that involves Taylor expansion to compute particle-cluster interactions, whereas this study used pairwise summations of potentials evaluated by a cubic spline using tabulated COFFDROP potential data. Except for ProTα, the ideal scaling factor for the naturally occurring IDPs is between 0.7 and 0.8, which allows the BD simulation results to match up with

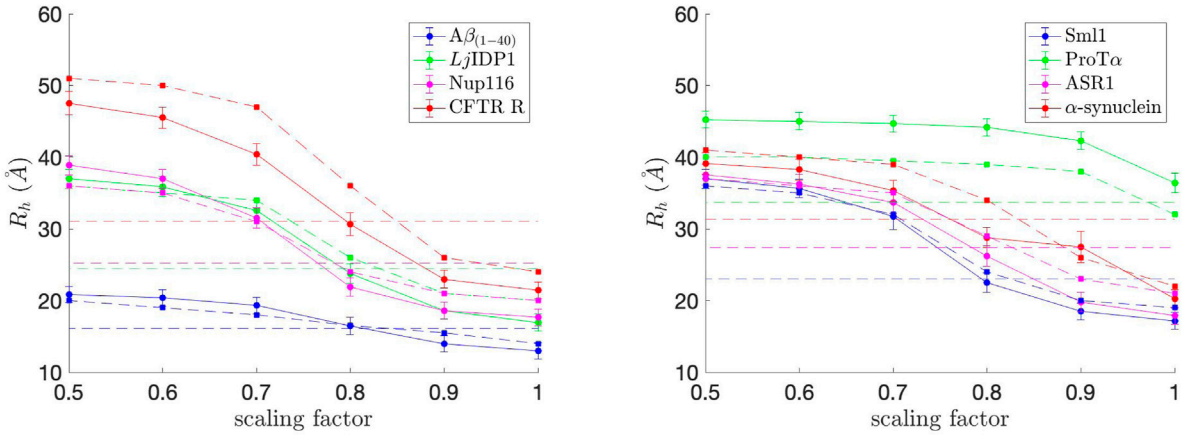


FIGURE 4 | Hydrodynamic radius (R_h) values for each of the eight naturally occurring IDPs with different scaling factors for non-bonded interactions. The experimental R_h values are marked as dashed straight lines and correspond to the IDP with the same color in each graph. The approximate R_h values from Frembgen-Kesner et al. (2015) are marked as dashed lines with square markers and correspond to the IDP with the same color in each graph. The ideal scaling factor value would be where R_h matches with the experimental value.

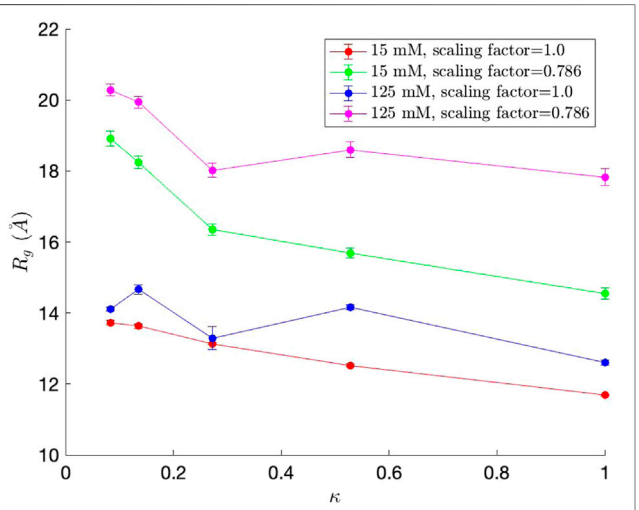


FIGURE 5 | Radius of gyration (R_g) values for each of the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants. Since each $(Glu\text{-}Lys)_{25}$ IDP sequence variant has a different κ value, which represents the measure of charge segregation (Das and Pappu, 2013), κ was used as one of the axes of the graph.

experimental values. We can consider ProTα to be an outlier among the eight naturally occurring IDPs since it substantially has more like-charged residues (i.e., positively charged residues aspartic acid (D) and glutamic acid (E)) as seen in **Figure 2** and as noted in Frembgen-Kesner et al. (2015). The averaged ideal scaling factor, after leaving ProTα out as an outlier, is 0.786, which is slightly different from the scaling factor in Frembgen-Kesner et al. (2015) (0.825). This scaling factor was used for subsequent COFFDROP BD simulations of the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants. Finally, **Figure 4** shows that R_h generally increases with sequence length, except for ProTα that is shorter than ASR1, Nup116, α-synuclein, and CFTR R.

3.2 Five $(Glu\text{-}Lys)_{25}$ IDP Sequence Variants

We then investigated five $(Glu\text{-}Lys)_{25}$ IDP sequence variants for rates of association, which were model IDP systems in Das and Pappu (2013), Sawle and Ghosh (2015), and McCarty et al. (2019). These block polymers of glutamate and lysine residues with different patterns serve as model IDPs since IDPs mostly consist of oppositely charged residues (i.e., they are polyampholytes) and do not have significant secondary structures.

We first measured the radius of gyration (R_g), which serves as an indicator of protein structure compactness, i.e., the smaller the R_g, the tighter the packing of the protein is. R_g was calculated for each simulation snapshot (every 100,000 steps), and the final R_g value for each simulation was obtained by averaging the R_g values from the simulation. The average R_g values, along with standard error bars (95% confidence interval), from ten independent simulations, are plotted in **Figure 5**. As observed in Das and Pappu (2013), R_g generally decreases as κ, which represents the measure of charge segregation (Das and Pappu, 2013), increases. The R_g values are smaller than those from Das and Pappu (2013), all within the value for classical Flory random coils (~18 Å) and compact globules (~11 Å). The R_g values never reach near the value for self-avoiding random walks (~28 Å), which is expected for well-mixed sequence variants or those with low κ values. This is most likely attributed from using different force fields and potentially shows the limitation for the COFFDROP potential in modeling highly charged systems. However, when using the averaged ideal scaling factor for IDPs (0.786), the R_g values increase, show closer to expected R_g values, and its minimum R_g range match with that in Das and Pappu (2013). As $\kappa \rightarrow 1$, the R_g values get closer to the value for compact globules (~11 Å) (Dima and Thirumalai, 2004). Finally, the R_g values increase as the salt concentration increases due to long-range interactions getting weaker, which is consistent with the results from Das and Pappu (2013). Overall, we were able to observe correct trends for

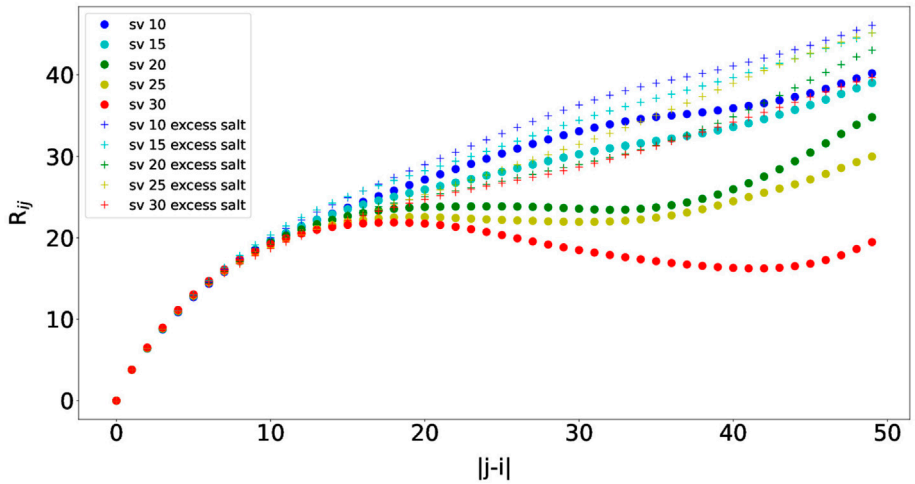


FIGURE 6 | Interresidue distances between residue *i* and residue *j* (R_{ij}) against residue separations $|j - i|$ for each of the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants.

R_g for the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants using the COFFDROP potential.

We then measured the interresidue distances between residue *i* and residue *j* (R_{ij}) against residue separations $|j - i|$, which can characterize local concentrations of chain segments within the IDP (Das and Pappu, 2013). Specifically, the distance between residue *i*'s C_α and residue *j*'s C_α was measured. The scaling factor was set to the averaged ideal scaling factor of 0.786. R_{ij} was calculated for each simulation snapshot (every 100,000 steps), and the final R_{ij} value for each simulation was obtained by averaging the R_{ij} values from the simulation. The average R_{ij} values from ten independent simulations are plotted in **Figure 6**. The R_{ij} values follow similar trends as observed in Das and Pappu (2013), and the concave upward parts show indications of long-range interactions between oppositely charged blocks. As observed for the R_g values, the R_{ij} values were also smaller than those from Das and Pappu (2013), which could be attributed from using different force fields. Finally, the R_{ij} values also increase as the salt concentration increases due to long-range interactions getting weaker, which is consistent with the results from Das and Pappu (2013). The effects of the salt concentration are the smallest for sv10, which has the most well-mixed sequence in comparison with the rest and can counterbalance electrostatic repulsions and attractions (Das and Pappu, 2013). Overall, we were also able to observe correct trends with R_{ij} for the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants using the COFFDROP potential.

Finally, we measured a property related to molecular association, namely the entanglement indices, or the average C_α distances to the other chain, between the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants. Since all possible pair combinations were tested, 15 simulations were run for each salt concentration (15 and 125 mM, respectively). The scaling factor was set to the averaged ideal scaling factor of 0.786. The pairwise simulations start with two IDP sequence variants oriented crosswise and translated 15 Å apart. All five IDP sequence variants have their

TABLE 2 | Summary of entanglement index values of all possible pair combinations between the five $(Glu\text{-}Lys)_{25}$ IDP sequence variants.

IDP #1	IDP #2	Entanglement index (Å) at 15 mM	Entanglement index (Å) at 125 mM
sv10	sv10	26.72 ± 0.98	26.38 ± 0.58
sv10	sv15	25.93 ± 0.63	27.12 ± 0.39
sv10	sv20	26.54 ± 0.98	26.70 ± 0.49
sv10	sv25	26.72 ± 0.71	26.92 ± 0.50
sv10	sv30	25.43 ± 1.61	26.70 ± 1.21
sv15	sv15	26.72 ± 0.53	27.68 ± 0.44
sv15	sv20	25.72 ± 0.69	27.29 ± 0.18
sv15	sv25	26.58 ± 0.75	26.87 ± 1.10
sv15	sv30	26.17 ± 0.28	27.98 ± 0.81
sv20	sv20	25.48 ± 0.31	25.78 ± 0.56
sv20	sv25	26.89 ± 0.83	25.85 ± 0.57
sv20	sv30	26.07 ± 0.31	26.38 ± 0.70
sv25	sv25	27.00 ± 0.25	27.44 ± 0.58
sv25	sv30	25.07 ± 3.05	27.15 ± 0.40
sv30	sv30	24.17 ± 2.38	26.59 ± 0.81

middle C_α centered at (0.0, 0.0, 0.0). The middle C_α's of the two IDP sequence variants in the pairwise simulation were restrained to be less than 20 $\AA$ apart. The entanglement indices were measured using the entanglement_index, which is a new implementation in Browndye 2.0. The entanglement index was calculated for each simulation snapshot (every 100,000 steps), and the final entanglement index for each simulation was obtained by averaging the entanglement indices from the simulation. The average entanglement indices, along with standard error bars (95% confidence interval), from ten independent simulations are listed in **Table 2**. The entanglement indices were similar across all possible IDP sequence variant pairs, indicating that there is no direct relation between entanglement indices and charge segregation κ. This may be from all sequence variants having relatively similar R_g values or degrees of compactness, however, which could be from using the COFFDROP potential. All entanglement indices, with the exception for two pairs, increased, however, at excess salt concentration. This is consistent with long-range interactions getting weaker at increasing salt concentration, and published work demonstrating that long-range interactions accelerate protein-protein encounter for IDPs (Chu et al., 2012; Ganguly et al., 2013; Pang and Zhou, 2016; Tsai et al., 2016; Chu et al., 2017; Yang et al., 2019).

4 DISCUSSION AND CONCLUSION

We have presented our new COFFDROP force field implementation on Browndye 2.0 that enabled us to study IDPs computationally. BD simulations are ideal to study large-scale motions with longer time scales and diffusion-limited molecular associations, including the aggregation of IDPs. We have presented results that show that our COFFDROP implementation is reliable to study naturally occurring IDPs. We have also studied model $(Glu\text{-}Lys)_{25}$ IDPs using our COFFDROP implementation and found that there is no relation between entanglement indices and how well the charges are mixed and segregated within the IDPs. However, this may be from the limitation of the COFFDROP potential in studying highly charged systems, which was also noted in Frembgen-Kesner et al. (2015). The COFFDROP potential was derived from MD simulations of all possible amino acid pairs (Andrews and Elcock, 2014; Frembgen-Kesner et al., 2015), but the simulations did not include salt so the COFFDROP potential may be limiting in modeling systems with strong charge-charge interactions.

For future work, we plan to implement a program to measure the rates of association with an appropriate reaction criterion as done in Ganguly et al. (2013), Liu et al. (2019). Then we plan to measure the rates of association of a highly positive IDP binding to a highly negative IDP (i.e., oppositely charged IDPs), an interaction that may be abundant in eukaryotes for regulation (e.g., cellular localization) (Borgia et al., 2018). We also plan to look at the rates of association between IDPs and folded proteins with secondary structures (Ruff et al., 2019). However, since COFFDROP is meant to model IDPs or systems without significant secondary or tertiary structures, the secondary structural elements would need to have constraints to have them fixed throughout the simulation, and the folded protein would be treated as a rigid body. The IDP would still be modeled as a flexible chain, and the scaling factor of 0.786 would be used for the simulation.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

S-HA: Conceptualization, investigation, methodology, data curation, validation, writing- original draft preparation. GH: Investigation, methodology, data curation, validation, writing-reviewing and editing. JM: Conceptualization, methodology, writing- reviewing and editing, supervision, funding acquisition.

FUNDING

S-HA and GH acknowledge support from NIH GM31749 and University of California, San Diego.

ACKNOWLEDGMENTS

All simulations were done using the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputing Center (SDSC). The authors thank Adrian Elcock for giving helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.898838/full#supplementary-material

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Fourth Edition. New York: Garland Science.

Alderson, T. R., Pritišanac, I., Moses, A. M., and Forman-Kay, J. D. (2022). Systematic Identification of Conditionally Folded Intrinsically Disordered Regions by Alphafold2. *bioRxiv*. doi:10.1101/2022.02.18.481080

Andrews, C. T., and Elcock, A. H. (2014). Coffdrop: a Coarse-Grained Nonbonded Force Field for Proteins Derived from All-Atom Explicit-Solvent Molecular

Dynamics Simulations of Amino Acids. *J. Chem. Theory Comput.* 10, 5178–5194. doi:10.1021/ct5006328

Baker, J. M. R. (2009). *Structural Characterization and Interactions of the CFTR Regulatory Region*. Toronto, Canada: University of Toronto.

Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., et al. (2018). Extreme Disorder in an Ultrahigh-Affinity Protein Complex. *Nature* 555, 61–66. doi:10.1038/nature25762

Chu, W.-T., Clarke, J., Shammas, S. L., and Wang, J. (2017). Role of Non-native Electrostatic Interactions in the Coupled Folding and Binding of Puma with Mcl-1. *PLoS Comput. Biol.* 13, e1005468. doi:10.1371/journal.pcbi.1005468

Chu, X., Wang, Y., Gan, L., Bai, Y., Han, W., Wang, E., et al. (2012). Importance of Electrostatic Interactions in the Association of Intrinsically Disordered Histone Chaperone Chz1 and Histone H2A.Z-H2b. *PLoS Comput. Biol.* 8, e1002608. doi:10.1371/journal.pcbi.1002608

Danielsson, J., Jarvet, J., Damberg, P., and Gräslund, A. (2002). Translational Diffusion Measured by PFG-NMR on Full Length and Fragments of the Alzheimer Aβ(1-40) Peptide. Determination of Hydrodynamic Radii of Random Coil Peptides of Varying Length. *Magn. Reson. Chem.* 40, S89–S97. doi:10.1002/mrc.1132

Danielsson, J., Liljedahl, L., Bárány-Wallje, E., Sønderby, P., Kristensen, L. H., Martinez-Yamout, M. A., et al. (2008). The Intrinsically Disordered Rnr Inhibitor Sml1 Is a Dynamic Dimer. *Biochemistry* 47, 13428–13437. doi:10.1021/bi801040b

Das, R. K., and Pappu, R. V. (2013). Conformations of Intrinsically Disordered Proteins Are Influenced by Linear Sequence Distributions of Oppositely Charged Residues. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13392–13397. doi:10.1073/pnas.1304749110

Dima, R. I., and Thirumalai, D. (2004). Asymmetry in the Shapes of Folded and Denatured States of Proteins. *J. Phys. Chem. B* 108, 6564–6570. doi:10.1021/jp037128y

Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., et al. (2007). Pdb2pqr: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic Acids Res.* 35, W522–W525. doi:10.1093/nar/gkm276

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004). PDB2PQR: an Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* 32, W665–W667. doi:10.1093/nar/gkh381

Elcock, A. H. (2004). "Molecular Simulations of Diffusion and Association in Multimacromolecular Systems," in *Numerical Computer Methods, Part D. Vol. 383 of Methods in Enzymology* (Cambridge, MA, USA: Academic Press), 166–198. doi:10.1016/s0076-6879(04)83008-8

Elcock, A. H. (2013). Molecule-centered Method for Accelerating the Calculation of Hydrodynamic Interactions in Brownian Dynamics Simulations Containing Many Flexible Biomolecules. *J. Chem. Theory Comput.* 9, 3224–3239. doi:10.1021/ct400240w

Frembgen-Kesner, T., Andrews, C. T., Li, S., Ngo, N. A., Shubert, S. A., Jain, A., et al. (2015). Parametrization of Backbone Flexibility in a Coarse-Grained Force Field for Proteins (Coffdrop) Derived from All-Atom Explicit-Solvent Molecular Dynamics Simulations of All Possible Two-Residue Peptides. *J. Chem. Theory Comput.* 11, 2341–2354. doi:10.1021/acs.jctc.5b00038

Ganguly, D., Zhang, W., and Chen, J. (2013). Electrostatically Accelerated Encounter and Folding for Facile Recognition of Intrinsically Disordered Proteins. *PLoS Comput. Biol.* 9, e1003363. doi:10.1371/journal.pcbi.1003363

Goldgur, Y., Rom, S., Ghirlando, R., Shkolnik, D., Shadrin, N., Konrad, Z., et al. (2007). Desiccation and Zinc Binding Induce Transition of Tomato Abscisic Acid Stress Ripening 1, a Water Stress- and Salt Stress-Regulated Plant-specific Protein, from Unfolded to Folded State. *Plant Physiol.* 143, 617–628. doi:10.1104/pp.106.092965

Grant, B. J., M. Gheorghe, D., Zheng, W., Alonso, M., Huber, G., Dlugosz, M., et al. (2011). Electrostatically Biased Binding of Kinesin to Microtubules. *PLoS Biol.* 9, e1001207. doi:10.1371/journal.pbio.1001207

Haaning, S., Radutoiu, S., Hoffmann, S. V., Dittmer, J., Giehm, L., Otzen, D. E., et al. (2008). An Unusual Intrinsically Disordered Protein from the Model Legume lotus Japonicus Stabilizes Proteins *In Vitro*. *J. Biol. Chem.* 283, 31142–31152. doi:10.1074/jbc.m805024200

Holehouse, A. S., Ahad, J., Das, R. K., and Pappu, R. V. (2015). Cider: Classification of Intrinsically Disordered Ensemble Regions. *Biophysical J.* 108, 228a. doi:10.1016/j.bpj.2014.11.1260

Huang, Y.-m. M., Huber, G. A., Wang, N., Minteer, S. D., and McCammon, J. A. (2018). Brownian Dynamic Study of an Enzyme Metabolon in the Tca Cycle: Substrate Kinetics and Channeling. *PROTEIN Sci.* 27, 463–471. doi:10.1002/pro.3338

Huber, G. A., and McCammon, J. A. (2010). Browndye: a Software Package for Brownian Dynamics. *Comput. Phys. Commun.* 181, 1896–1905. doi:10.1016/j.cpc.2010.07.022

Huber, G. A., and McCammon, J. A. (2019). Brownian Dynamics Simulations of Biological Molecules. *Trends Chem.* 1, 727–738. doi:10.1016/j.trechm.2019.07.008

Iqbalsyah, T. M., and Doig, A. J. (2005). Anticooperativity in a Glu−Lys−Glu Salt Bridge Triplet in an Isolated α-Helical Peptide. *Biochemistry* 44, 10449–10456. doi:10.1021/bi0508690

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

Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., et al. (2018). Improvements to the Apbs Biomolecular Solvation Software Suite. *Protein Sci.* 27, 112–128. doi:10.1002/pro.3280

Kirkwood, J. G. (1996). The General Theory of Irreversible Processes in Solutions of Macromolecules. *J. Polym. Sci. B Polym. Phys.* 34, 597–610. doi:10.1002/polb.1996.897

Krishnan, V. V., Lau, E. Y., Yamada, J., Denning, D. P., Patel, S. S., Colvin, M. E., et al. (2008). Intramolecular Cohesion of Coils Mediated by Phenylalanine-Glycine Motifs in the Natively Unfolded Domain of a Nucleoporin. *PLoS Comput. Biol.* 4, e1000145. doi:10.1371/journal.pcbi.1000145

Kyte, J., and Doolittle, R. F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* 157, 105–132. doi:10.1016/0022-2836(82)90515-0

Li, P., Johnston, H., and Krasny, R. (2009). A Cartesian Treecode for Screened Coulomb Interactions. *J. Comput. Phys.* 228, 3858–3868. doi:10.1016/j.jcp.2009.02.022

Liu, X., Chen, J., and Chen, J. (2019). Residual Structure Accelerates Binding of Intrinsically Disordered Actr by Promoting Efficient Folding upon Encounter. *J. Mol. Biol.* 431, 422–432. doi:10.1016/j.jmb.2018.12.001

Marqusee, S., and Baldwin, R. L. (1987). Helix Stabilization by Glu-.Lys+ Salt Bridges in Short Peptides of De Novo Designlys+ Salt Bridges in Short Peptides of De Novo Design. *Proc. Natl. Acad. Sci. U.S.A.* 84, 8898–8902. doi:10.1073/pnas.84.24.8898

Martinez, M., Bruce, N. J., Romanowska, J., Kokh, D. B., Ozboyaci, M., Yu, X., et al. (2015). Sda 7: A Modular and Parallel Implementation of the Simulation of Diffusional Association Software. *J. Comput. Chem.* 36, 1631–1645. doi:10.1002/jcc.23971

McCarty, J., Delaney, K. T., Danielsen, S. P. O., Fredrickson, G. H., and Shea, J.-E. (2019). Complete Phase Diagram for Liquid-Liquid Phase Separation of Intrinsically Disordered Proteins. *J. Phys. Chem. Lett.* 10, 1644–1652. doi:10.1021/acs.jpclett.9b00099

Meuzelaar, H., Vreede, J., and Woutersen, S. (2016). Influence of Glu/Arg, Asp/Arg, and Glu/Lys Salt Bridges on α -Helical Stability and Folding Kinetics. *Biophysical J.* 110, 2328–2341. doi:10.1016/j.bpj.2016.04.015

Northrup, S. H., Thomasson, K. A., Miller, C. M., Barker, P. D., Eltis, L. D., Guillemette, J. G., et al. (1993). Effects of Charged Amino Acid Mutations on the Bimolecular Kinetics of Reduction of Yeast Iso-1-Ferricytochrome C by Bovine Ferrocytochrome B5. *Biochemistry* 32, 6613–6623. doi:10.1021/bi00077a014

Nygaard, M., Kragelund, B. B., Papaleo, E., and Lindorff-Larsen, K. (2017). An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations. *Biophysical J.* 113, 550–557. doi:10.1016/j.bpj.2017.06.042

Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* 7, 525–537. doi:10.1021/ct100578z

Ortega, A., Amorós, D., and García de la Torre, J. (2011). Prediction of Hydrodynamic and Other Solution Properties of Rigid Proteins from

Atomic- and Residue-Level Models. *Biophysical J.* 101, 892–898. doi:10.1016/j.bpj.2011.06.046

Pang, X., and Zhou, H.-X. (2016). Mechanism and Rate Constants of the Cdc42 Gtpase Binding with Intrinsically Disordered Effectors. *Proteins* 84, 674–685. doi:10.1002/prot.25018

Pollard, T., and Earnshaw, W. (2007). *Cell Biology*. New York: W.B. Saunders.

Roberts, C. C., and Chang, C.-e. A. (2016). Analysis of Ligand-Receptor Association and Intermediate Transfer Rates in Multienzyme Nanostructures with All-Atom Brownian Dynamics Simulations. *J. Phys. Chem. B* 120, 8518–8531. doi:10.1021/acs.jpcb.6b02236

Ruff, K. M., Pappu, R. V., and Holehouse, A. S. (2019). Conformational Preferences and Phase Behavior of Intrinsically Disordered Low Complexity Sequences: Insights from Multiscale Simulations. *Curr. Opin. Struct. Biol.* 56, 1–10. doi:10.1016/j.sbi.2018.10.003

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

Sawle, L., and Ghosh, K. (2015). A Theoretical Method to Compute Sequence Dependent Configurational Properties in Charged Polymers and Proteins. *J. Chem. Phys.* 143, 085101. doi:10.1063/1.4929391

Søndergaard, C. R., Olsson, M. H., Rostkowski, M., and Jensen, J. H. (2011). Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* 7, 2284–2295. doi:10.1021/ct200133y

Tsai, M.-Y., Zheng, W., Balamurugan, D., Schafer, N. P., Kim, B. L., Cheung, M. S., et al. (2016). Electrostatics, Structure Prediction, and the Energy Landscapes for Protein Folding and Binding. *Protein Sci.* 25, 255–269. doi:10.1002/pro.2751

Unni, S., Huang, Y., Hanson, R. M., Tobias, M., Krishnan, S., Li, W. W., et al. (2011). Web Servers and Services for Electrostatics Calculations with Apbs and Pdb2pqr. *J. Comput. Chem.* 32, 1488–1491. doi:10.1002/jcc.21720

Uversky, V. N., Lee, H.-J., Li, J., Fink, A. L., and Lee, S.-J. (2001). Stabilization of Partially Folded Conformation during α-Synuclein Oligomerization in Both Purified and Cytosolic Preparations. *J. Biol. Chem.* 276, 43495–43498. doi:10.1074/jbc.c100551200

Uversky, V. N. (2002). Natively Unfolded Proteins: a Point where Biology Waits for Physics. *Protein Sci.* 11, 739–756. doi:10.1110/ps.4210102

Wolny, M., Batchelor, M., Bartlett, G. J., Baker, E. G., Kurzawa, M., Knight, P. J., et al. (2017). Characterization of Long and Stable De Novo Single Alpha-Helix Domains Provides Novel Insight into Their Stability. *Sci. Rep.* 7, 44341–44414. doi:10.1038/srep44341

Yang, J., Gao, M., Xiong, J., Su, Z., and Huang, Y. (2019). Features of Molecular Recognition of Intrinsically Disordered Proteins via Coupled Folding and Binding. *Protein Sci.* 28, 1952–1965. doi:10.1002/pro.3718

Yi, S., Boys, B. L., Brickenden, A., Konermann, L., and Choy, W.-Y. (2007). Effects of Zinc Binding on the Structure and Dynamics of the Intrinsically Disordered Protein Prothymosin α: Evidence for Metalation as an Entropic Switch. *Biochemistry* 46, 13120–13130. doi:10.1021/bi7014822

Zuk, P. J., Wajnryb, E., Mizerski, K. A., and Szymczak, P. (2014). Rotne-prager-yamakawa Approximation for Different-Sized Particles in Application to Macromolecular Bead Models. *J. Fluid Mech.* 741, 668. doi:10.1017/jfm.2013.668

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.

frontiers | Frontiers in Molecular Biosciences

REVIEW
published: 08 June 2022
doi: 10.3389/fmolb.2022.899805

Enhanced-Sampling Simulations for the Estimation of Ligand Binding Kinetics: Current Status and Perspective

Katya Ahmad[1], *Andrea Rizzi*[1,2], *Riccardo Capelli*[3], *Davide Mandelli*[1], *Wenping Lyu*[4,5] *and Paolo Carloni*[1,6]*

[1]Computational Biomedicine (IAS-5/INM-9), Forschungszentrum Jülich, Jülich, Germany, [2]Atomistic Simulations, Istituto Italiano di Tecnologia, Genova, Italy, [3]Department of Applied Science and Technology (DISAT), Politecnico di Torino, Torino, Italy, [4]Warshel Institute for Computational Biology, School of Life and Health Sciences, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China, [5]School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, China, [6]Molecular Neuroscience and Neuroimaging (INM-11), Forschungszentrum Jülich, Jülich, Germany

The dissociation rate (k_{off}) associated with ligand unbinding events from proteins is a parameter of fundamental importance in drug design. Here we review recent major advancements in molecular simulation methodologies for the prediction of k_{off}. Next, we discuss the impact of the potential energy function models on the accuracy of calculated k_{off} values. Finally, we provide a perspective from high-performance computing and machine learning which might help improve such predictions.

Keywords: kinetics, drug discovery, QM/MM, parallel computing, machine learning, enhanced sampling, molecular dynamics

OPEN ACCESS

Edited by:
Weiliang Zhu,
Shanghai Institute of Materia Medica (CAS), China

Reviewed by:
Ariane Nunes Alves,
Technical University of Berlin, Germany
Jinan Wang,
University of Kansas, United States
Zhijian Xu,
Shanghai Institute of Materia Medica (CAS), China

***Correspondence:**
Paolo Carloni
p.carloni@fz-juelich.de

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *19 March 2022*
Accepted: *09 May 2022*
Published: *08 June 2022*

Citation:
Ahmad K, Rizzi A, Capelli R, Mandelli D, Lyu W and Carloni P (2022) Enhanced-Sampling Simulations for the Estimation of Ligand Binding Kinetics: Current Status and Perspective. Front. Mol. Biosci. 9:899805. doi: 10.3389/fmolb.2022.899805

1 INTRODUCTION

The kinetics of drugs unbinding from proteins is an important parameter for the drugs' efficacy. (Pan et al., 2013; Copeland, 2021). Indeed, the drug-target residence time (Copeland, Pompliano and Meek, 2006) defined as the inverse of the dissociation rate k_{off}, has emerged as an effective surrogate measure of *in vivo* target occupancy, and it has been shown to correlate with clinical efficacy (Guo et al., 2012; Lee et al., 2019; Van Der Velden et al., 2020) along with other factors (e.g., association rates (Folmer, 2018; Lee et al., 2019) and target saturation (de Witte et al., 2018)). Residence time has been related not only to long-lasting pharmacodynamics but also to the reduced toxicity of specific inhibitors (Vauquelin et al., 2012).

Experimental approaches (most often combined with computations) measure ligand affinities and provide ligand binding poses for structure-based drug design campaigns (Durrant and McCammon, 2011; De Vivo et al., 2016; Proudfoot et al., 2017; Emwas et al., 2020; Mazzorana et al., 2020). They routinely also measure k_{off} values (Pollard, 2010). However, they cannot usually access the structural determinants of the transition states associated with ligand unbinding. This information would be crucial to eventually design ligands with longer residence times. In contrast, all-atom molecular simulations (in particular molecular dynamics (MD)) can provide a detailed map of protein-ligand interactions and the atomic rearrangements that drive ligand unbinding. However, the residence time of tight binders can be as long as several hours (Li et al., 2014), much longer than the timescales reached by plain MD (milliseconds on dedicated, specialized machines) (Pan et al., 2019; Shaw et al., 2021). Thus, k_{off} predictions based on such a straightforward approach so far have been few in number (Pan et al., 2017) or limited to model systems (Tang and Chang, 2018).

Enhanced sampling is a more general approach to the estimation of k_{off}, regardless of the timescale of the unbinding event. One group of methods (including metadynamics, Gaussian Accelerated MD, scaled MD, and dissipation-corrected targeted MD) employs biasing potentials designed to reduce the free energy barrier determining the frequency of dissociations. Because the bias affects the dynamics, correction terms are required to recover the unbiased k_{off} from the biased rates. A second group is represented by path sampling approaches such as weighted ensemble and milestoning. These rigorously generate an ensemble of trajectories by iteratively restarting the (unbiased) simulations from selected configurations (typically closer to the transition state than expected from the equilibrium distribution) with the aim of increasing the likelihood of observing dissociations. Finally, Markov state models (MSMs) can provide a complete picture of the metastable states of the system and transition rates between them by analyzing molecular simulation data.

In this review, we summarize principles and applications of the three approaches outlined above (**Sections 2–4**). Next, we discuss the impact of force fields on the accuracy of the calculations (**Section 5**). Finally, we provide a perspective on how machine learning, along with exascale computing, could constitute one way to address these challenges (**Section 6**).

1.1 Scope

Many methods have been developed for the calculation of rate constants in biomolecular simulations. Here, we review methodologies that have been applied to the calculation of binding dissociation rates (k_{off}) of protein-ligand complexes with a focus on the effect of the potential energy function. In particular, for the sake of conciseness, we do not cover methods that have been applied only to other types of systems/problems (e.g., supramolecular host-guest dissociations, peptide folding rates) and methods that enable relative comparisons of k_{off} between different ligands.[1] For these methods, we refer the reader to the other excellent resources on the topic (Chong et al., 2017; Bruce et al., 2018; Nunes-Alves et al., 2020).

2 BIASED MD METHODS

In this class of methods, the system is biased (by adding a potential term to the Hamiltonian, or adding external forces) to favor the observation of unbinding events. The bias is designed to enhance the exploration along low-dimensional collective variables (CV), which are represented as differentiable functions $s(\mathrm{x})$ of the atomic coordinates x. These describe the slow degrees of freedom governing the unbinding process. The CV must be able to distinguish the metastable states involved in the process i.e., configurations in different states should correspond to different values of the CV. The identification of optimal CVs (whenever they are not intrinsic in the technique) is a complicated task, and their identification is at the center of a heated debate that is still open (Sittel and Stock, 2018). Because biasing terms alter the dynamics, methods which recover the kinetic parameters of the unbiased system from its free energy surface have been devised. The majority of biased methods adopt specific corrections based on Kramers' rate theory

$$k_{AB} = \omega_A \kappa_A \frac{Z^*}{Z_A} \tag{1}$$

where k_{AB} is the rate of transition from state A to B (in this case the bound and unbound states), ω_A is typically associated with the curvature of the free energy surface, κ_A is the transmission coefficient, and Z^* and Z_A are the configurational partition functions of the transition state and state A, respectively. These methods require the identification of the transition state ensemble, defined as the set of conformations of highest free energy along the (un)binding pathway. This is in general a challenging task for drug binding processes, which can involve multiple dissociation pathways due to the conformational flexibility of the protein (Plattner and Noé, 2015). Approaches of this kind have been developed for Gaussian accelerated molecular dynamics (Miao et al., 2020) (see **Section 2.1**), dissipation-corrected Targeted Molecular Dynamics (Wolf and Stock, 2018) (see **Section 2.2**), and τ-random acceleration molecular dynamics (Kokh et al., 2018) (see **Section 2.3**)[2]. If no bias is deposited on the region of the transition state(s), the kinetic correction can be assumed not to depend on κ_A and Z^* (Voter and Doll, 1985; Hänggi et al., 1990; Truhlar et al., 1996). This simplifies dramatically the rate estimation, and it is used for ligand unbinding in the kinetics-oriented flavors of metadynamics (Tiwary and Parrinello, 2013; Wang et al., 2018) (see **Section 2.4**).

2.1 Ligand Gaussian Accelerated MD

2.1.1 Basic principles

In this approach (Miao, 2018), two harmonic potentials are added to the non-bonded component of the potential energy so as to lower the binding/unbinding free energy barrier (**Figure 1**). These potentials act on the following CVs: 1) the ligand-environment potential energy and (optionally) 2) the rest of the system potential energy. Both biasing potentials are capped at user-defined thresholds. Computing the correction to recover the unbiased transition rate requires the estimation of the potential of mean force (PMF) profile and free energy barrier as a function of a separate CV describing the binding process e.g., a distance between ligand and protein atoms (Miao, 2018). In the

[1]These techniques include, among others, scaled MD (Sinko et al., 2013; Bernetti et al., 2018), steered MD (Paci and Karplus, 2000; Potterton et al., 2019; Spiriti and Wong, 2021), targeted MD (Schlitter et al., 1994; Wolf et al., 2019), GAMBES (Debnath and Parrinello, 2020) path-reweighting methods (Chodera et al., 2011; Donati et al., 2017; Kieninger and Keller, 2021) metadynamics of paths (Mandelli et al., 2020) and many transition path sampling-derived methods (Pratt, 1986; Dellago et al., 1998; Van Erp et al., 2003). A brief review of some of these methods (namely scaled MD, targeted MD and GAMBES) is given in the supplementary material.

[2]A Kramers' rate theory correction has been developed for scaled MD as well but it has not been applied to the calculation of full dissociation rates (see **Supplementary Material S3.3**).

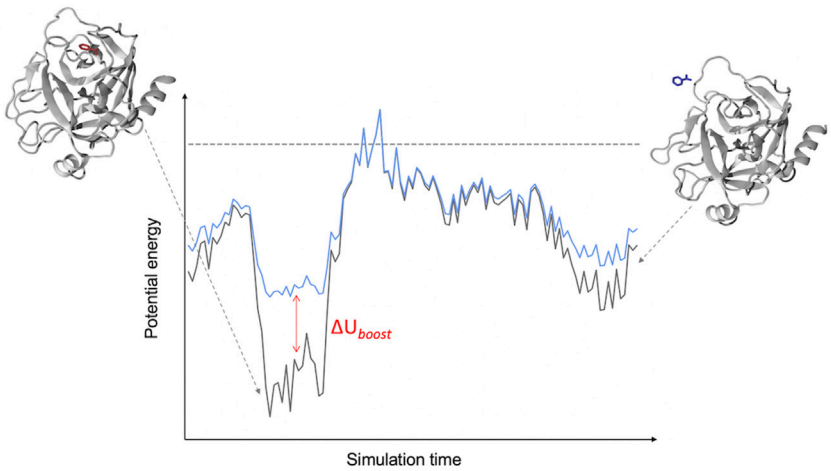


FIGURE 1 | Schematic of a LiGaMD Simulation. The LiGaMD potential (ΔU_{boost}) acts when the potential energy of a protein-ligand complex (black line) is below a predefined threshold (dashed line), adding a harmonic potential to raise the energy of the system (cyan line) and favor the exploration of the conformational space of the ligand-protein complex.

closely related Pep-GaMD method, developed specifically for simulating peptides unbinding from their target protein, the harmonic "boost" potentials are applied to the total potential (both non-bonded and bonded components) along the CVs (Wang and Miao, 2020). The application of the additional boost potential to the bonded component of the peptide potential energy accelerates the sampling of its conformational flexibility.

2.1.2 Applications

So far, the approach has been successfully applied to the ligand benzamidine targeting the trypsin enzyme (Miao, Bhattarai and Wang, 2020), using the AMBER14SB (Maier et al., 2015) and GAFF (Wang et al., 2004) force-fields. The calculated $k_{off} = 3.53 \pm 1.41\ s^{-1}$ was two orders of magnitude smaller than the experimental value of $600 \pm 300\ s^{-1}$ (Guillian and Thusias, 1970). The simulations required a cumulative 5 μs of MD for the estimation of k_{off}. Pep-GaMD has been used to investigate the un/binding of three model peptides that target the SH3 domain—one of which (PDB:1CKB) has an experimentally determined k_{off} available for comparison to the computed value. Employing the AMBER14SB (Maier et al., 2015) force field and an aggregate simulation time of 3 μs, a k_{off} of $1.45 \pm 1.17 \cdot 10^3\ s^{-1}$ was computed for 1CKB; a result that is in close agreement with the experimental value of $8.9 \cdot 10^3\ s^{-1}$ (Xue et al., 2014).

2.2 Dissipation-Corrected Targeted Molecular Dynamics (dcTMD)

2.2.1 Basic principles

This method (Wolf and Stock, 2018) assumes that unbinding processes (along with binding processes) can be described by the 1-dimensional Langevin dynamics of a suitable CV. The approach requires the determination of the free energy profile and the Langevin friction coefficient as a function of such a CV. These can be calculated from a nonequilibrium targeted molecular dynamics simulation (Schlitter et al., 1994) (see **Supplementary Material S4**), where a pulling force drives the system at a constant speed along the CV. Dissociation rates could then be obtained by performing the unbiased 1-dimensional Langevin dynamics simulations (Wolf and Stock, 2018). Despite the simplification, the timescales of ligand unbinding processes at room temperature still lead to prohibitively expensive simulations. To tackle this problem, the authors later introduced an approach that uses Kramers' theory to correct the rates obtained from Langevin simulations performed at higher temperatures. (Wolf et al., 2020).

2.2.2 Applications

The method has been successfully applied to the calculation of k_{off} of the trypsin-benzamidine complex, and the complex between a resorcinol scaffold-based inhibitor and the HSP90 protein. The calculated values $270 \pm 40\ s^{-1}$ and $1.6 \pm 0.2 \cdot 10^{-3}\ s^{-1}$ respectively, agree well with the experimental values of $600 \pm 300\ s^{-1}$ (Guillian and Thusias, 1970) and $3.4 \pm 0.2 \cdot 10^{-2}\ s^{-1}$ (Amaral et al., 2017) These predictions required an aggregate of $\sim 1.5 \times 10^4$ ms of Langevin simulations and used the AMBER99SB* force-field (Best and Hummer, 2009).

2.3 τRAMD

2.3.1 Basic principles

The τ-random acceleration molecular dynamics (τRAMD) (Kokh et al., 2018) protocol is a quasi-biased method that evolved from random acceleration molecular dynamics (RAMD) (Lüdemann et al., 2000). τRAMD simulations of ligand-protein systems proceed similarly to standard MD simulations, without the need for any prior parameter fitting, characterization of CVs or binding pathways. The user specifies the magnitude of a randomly oriented force that is applied to the ligand to accelerate its dissociation from the binding pocket at each

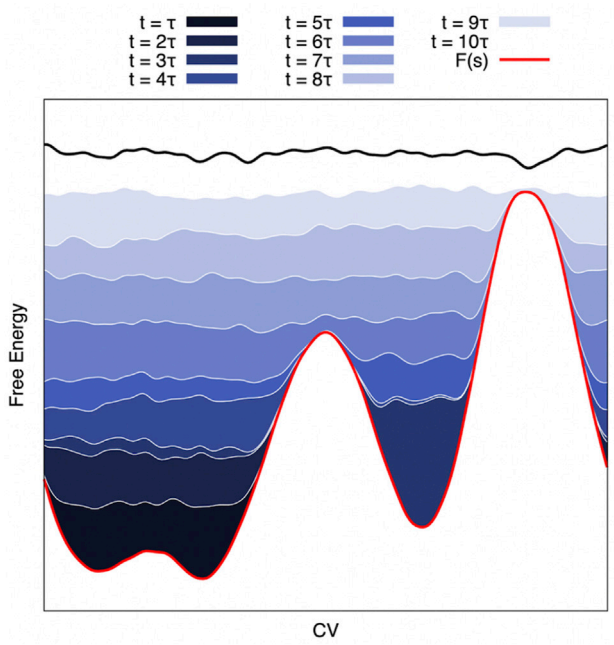


FIGURE 2 | Schematic of a Metadynamics Simulation. On the CV-projected FES (red line), MetaD deposits a series of gaussians that sum up (from dark blue to white) until the system becomes diffusive in the CV space. This approach can be exploited to reduce the barrier height to have a reasonable transition time and reweight it a posteriori for an estimation of the kinetic constants (see Text).

checkpoint, allowing for the observation of dissociation pathways within several nanoseconds of simulation time. The magnitude of the force dictates the duration of simulation time that is required and is reported to have a minimal effect on the accuracy of computed residence times. The direction of the force is reassigned after each checkpoint until the ligand COM moves past a certain distance threshold from its previous position. If the deviation of the ligand COM meets or exceeds this threshold after the force is applied, the direction of the force is retained until the following checkpoint. An ensemble of unbinding simulations is spawned from different starting configurations and velocities, and the ensemble-averaged residence time is calculated from the bootstrapped distribution of the time taken for dissociation to occur.

2.3.2 Applications

The earliest applications of τRAMD for unbinding kinetics focused on qualitatively ranking ligands according to their computed k_{off} values (see **Supplementary Material S5**) (Kokh et al., 2018, 2019, 2020). Recently, the first quantitative τRAMD application was demonstrated by Maximova and co-workers (Maximova et al., 2021), who formulated a Kramers' rate theory-based rescaling factor to correct for the influence of the applied force on the receptor-ligand coupling (which previously limited the method to qualitative ranking) to obtain a quantitative k_{off} estimate for the drug Isoniazid unbinding from the catalase enzyme. Using seven trajectories (with applied forces of different magnitudes), and the CHARMM36 forcefield (Best et al., 2012), a k_{off} of $2.8 \pm 3.7 \cdot 10^{-2}\ s^{-1}$ was computed—a result which agreed very well with the experimental equivalent of $2.0 \pm 0.3 \cdot 10^{-2}$ (Singh et al., 2008).

2.4 Metadynamics-Derived Methods

2.4.1 Basic principles

Well-tempered Metadynamics (MetaD) (Laio and Parrinello, 2002) is an exact free-energy method (Barducci et al., 2008; Dama et al., 2014). It draws inspiration from earlier CV-based enhanced sampling techniques such as local elevation (Huber et al., 1994), Wang-Landau (Wang and Landau, 2001), conformational flooding (Grubmüller, 1995), and adaptive umbrella sampling techniques (Hooft et al., 1992; Bartels and Karplus, 1997). In MetaD, a history-dependent bias potential $B_t(s)$ is built iteratively by adding Gaussian functions (as approximations of CV histograms) to the potential at regular intervals throughout the simulations. Several different bias-deposition schemes have been devised (Bussi and Laio, 2020). Ultimately, convergence is achieved when the sum of the free energy surface and the bias potential produces a flat landscape that results in diffusive dynamics in CV space (see **Figure 2**). It is then possible to compute the free energy surface along the CV *via* reweighting methods, such as Weighted Histogram Analysis Method (WHAM) (Kumar et al., 1992), Multistate Bennet Acceptance Ratio (MBAR) (Shirts and Chodera, 2008), or other estimators (Tiwary and Parrinello, 2015; Schäfer and Settanni, 2020).

MetaD has been extended to allow recovery of the kinetics of the unbiased ensemble. The method speeds up the calculation of kinetic rates by filling up the starting free energy basin so as to reduce the activation free energy barrier to ~ k_BT. This way, the biased residence time of the system in the initial state is small enough to allow multiple observations of the transition. Transition times obtained in the biased ensemble are then scaled to recover the unbiased kinetics. Following the approaches of Grubmüller (Conformational flooding (Grubmüller, 1995)) and Voter (Hyperdynamics (Voter, 1997)) the unbiased transition time is connected to the biased time by:

$$t_{\text{unbiased}} = \alpha\, t_{\text{biased}} = \sum_{t=0}^{t_{\text{biased}}} \exp\left(\beta B_t\left(s(t)\right)\right)\Delta t \qquad (2)$$

where $\beta = (k_BT)^{-1}$, $Bt(s)$ is the history-dependent bias potential, and Δt is the time step. For this last equation to be valid, no bias should be present on the transition state. In the so-called infrequent MetaD variant (Tiwary and Parrinello, 2013), the Gaussians are deposited less frequently in barrier regions than they are in standard MetaD, thus lowering the probability of adding bias to the transition state. In frequency-adaptive (FA) MetaD (Wang et al., 2018), the time interval between bias depositions is gradually increased as the system approaches the transition state. After an initial fast filling of the free energy minimum, the same deposition rate as infrequent MetaD is achieved. This way, results are obtained at a lower

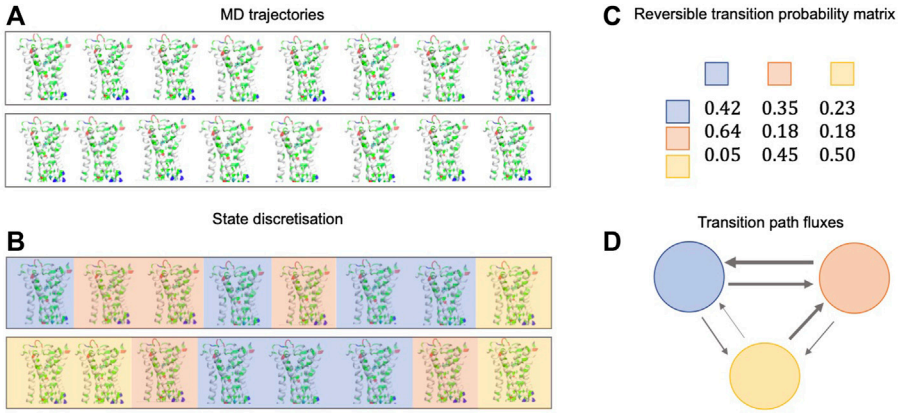


FIGURE 3 | Simplified schematic depiction of the MSM construction pipeline. **(A)** Several continuous MD trajectories are simulated in parallel. **(B)** The trajectories are discretized. **(C)** A reversible transition probability matrix is calculated from a matrix of state-to-state transition counts **(D)** Probability fluxes between states (gray arrows, with line thickness representing the magnitude of the flux) indicate the highest likelihood transition paths and can be used to calculate the mean first passage time (MFPT) between states.

computational cost compared to standard infrequent MetaD. Recently, an alternative method to infrequent and frequency-adapted MetaD has been presented. (Ansari et al., 2022) This method builds on a variant of MetaD called on-the-fly probability enhanced sampling (OPES). (Invernizzi and Parrinello, 2020) In the new approach, called OPES-flooding, the bias is constructed in a fast but controlled manner to fill the starting metastable basin up to a user-defined threshold value to automatically avoid depositing bias on the transition state. Usually, the standard protocol adopted in infrequent, FA-MetaD and OPES-flooding consists of running multiple independent simulations that yield an empirical distribution of residence times. A statistical analysis based on the Kolmogorov-Smirnov (KS) test (Salvalaglio et al., 2014), details in **Supplementary Material S1**) is then used to verify *a posteriori* that the transition state was indeed untainted.

2.4.2 Applications

Infrequent MetaD simulations based on the OPLS force-field (Kaminski et al., 2001) were used to study the unbinding of the ligand dasatinib from its target c-Src kinase (Tiwary et al., 2017). The CVs were the distance between the ligand and the binding pocket and the solvation state of the binding pocket. The calculated k_{off} of 4.8 ± 2.4 $\cdot 10^{-2}\,s^{-1}$ of dasatinib to c-Src obtained from 12 unbinding trajectories agreed well with an experimental value ($5.6 \cdot 10^{-2}\,s^{-1}$, measured indirectly from k_{on}) published by (Shan et al., 2009), but differed from a second value obtained for a fluorophore-tagged analogue ($1.8 \cdot 10^{-4}$ to $7.9 \cdot 10^{-4}\,s^{-1}$) (Kwarcinski et al., 2016). A similar protocol was used to calculate k_{off} for 1-(3-(tert-butyl)-1-(p-tolyl)-1H-pyrazol-5-yl) urea), an inhibitor of p38 MAP II kinase belonging to the BIRB-796 family, this time using AMBER99SB-ILDN (Hornak et al., 2006; Lindorff-Larsen et al., 2010) and GAFF force-fields (Wang et al., 2004; Wang J. et al., 2006). After 17 independent unbinding events, the calculated k_{off} (0.020 ± 0.011 s^{-1} (Casasnovas et al., 2017)) was almost one order of magnitude lower than the experimental value of 0.14 s^{-1} (Regan et al., 2003). Two other CVs yielded very similar results simulating 10 unbinding events each, suggesting that the discrepancy between the calculated and experimental values most likely arises from uncertainty in the force field rather than the choice of CVs.

FA-MetaD and Infrequent MetaD were used by Wang and co-authors (Wang et al., 2018) to obtain k_{off} values for benzene and indole ligands from the binding pocket of the L99A mutant of T4 lysozyme using CHARMM22 (MacKerell et al., 1998; MacKerell et al., 2004) and CGenFF (Vanommeslaeghe et al., 2011). The calculated k_{off} for benzene lay within the range of 4–10 s^{-1}, around 100-fold lower than the experimental value of 950 ± 20 s^{-1} (Feher et al., 1996). Both MetaD protocols used the same force-field, sample size (20 replicas), and path-CVs (Branduardi et al., 2007; Wang et al., 2017). CHARMM36-based (Best et al., 2012) infrequent MetaD simulations (Mondal et al., 2018) yielded a k_{off} for benzene (270 ± 100 s^{-1}) that was considerably closer to the experimental value. Although only the displacement between binding pocket and ligand centers-of-mass was used as the CV, and the sample size was smaller than that of the previous study by Wang et al, it is tempting to conclude that even a different version of the same force-field (CHARMM in this case) may significantly impact the result.

More recently, AMBER14SB-based (Maier et al., 2015) FA-MetaD simulations were applied to study the unbinding kinetics of a radioligand, iperoxo, from the M_2 human muscarinic acetylcholine receptor (Capelli et al., 2020). The calculated k_{off} ($3.7 \pm 0.7 \cdot 10^{-4}\,s^{-1}$) was two orders of magnitude smaller than the experimental value ($1.0 \pm 0.2 \cdot 10^{-2}\,s^{-1}$). Density Functional Theory (DFT)-based QM/MM calculations suggested that this estimation discrepancy may be ascribed, at least in part, to the lack of polarization and charge transfer effects lacking in standard biomolecular force fields (Capelli et al., 2020).

OPES-flooding simulations based on AMBER14SB (Maier et al., 2015) and GAFF (Wang et al., 2004) were recently applied to study the unbinding kinetics of the trypsin-benzamidine complex, unveiling the role of water in regulating the residence time. Notably, the authors identified two different unbinding pathways and were able to calculate the corresponding rates separately. The slowest rate of 687 s^{-1} that is supposed to dominate the residence time is in good agreement with the experimental value of 600 ± 300 s^{-1} (Guillian and Thusias, 1970).

3 MARKOV STATE MODELS

3.1 Basic Principles

Markov state models (MSMs) (Singhal et al., 2004) are discrete models describing the dynamics of a system in terms of transition probabilities between a finite set of metastable states. The fundamental ingredients of the method are 1) a discretization of the conformational space into (kinetically fast) microstates and 2) a transition matrix that describes the probability of observing the system in another microstate after a fixed lag time *t*. An interpretable, coarse-grained model is then built by defining kinetically metastable macrostates as collections of microstates, and this model can provide k_{off} values. **Figure 3** shows a simplified schematic depiction of the MSM construction pipeline. The lag time *t* must be long enough to ensure that transitions between states are approximately Markovian[4] and short enough for the model to represent all relevant fast processes. It should be chosen to be faster than association events to avoid systematic overestimation of the residence time (Paul et al., 2017). When this is not possible, k_{off} can still be estimated from rate matrices rather than transition matrices (Kalbfleisch and Lawless, 1985; Crommelin and Vanden-Eijnden, 2009). However, rate matrix estimation is not unique and different approaches can result in residence times that differ even by an order of magnitude (Paul et al., 2017).

The input data to build MSMs can come from an ensemble of unbiased MD trajectories that sample dissociation events. However, generating this data is usually prohibitively expensive. Hence, several powerful schemes have been designed to enable the estimation of second-long residence times from relatively short MD simulations. These include adaptive restarting strategies (Bowman et al., 2010; Wan and Voelz, 2020) and/or biased simulations (Wu et al., 2016; Paul et al., 2017; Stelzl et al., 2017). In particular, recently developed estimators such as transition-based reweighting analysis TRAM (Wu et al., 2016) and its MBAR variant TRAMMBAR (Paul et al., 2017) require only irreversible visits to metastable states in the unbiased MD (as long as these states are sampled reversibly in the biased ones) and can greatly alleviate the sampling problem.

[4]i.e., the probability of observing the system in a state y after the lag time given that it was in state *x* does not depend on the states of the system before *x*.

3.2 Applications

MSM calculations on the trypsin-benzamidine complex (Plattner and Noé, 2015) (methodological details in **Supplementary Material S6**) yielded a k_{off} of 131 ± 109 × 10^2 s^{-1}, which compares fairly well with experiments (k_{off} = 600 ± 300 s^{-1}) (Guillian and Thusias, 1970). However, the high level of uncertainty suggests that sampling of unbinding events might be insufficient despite the large amount of aggregate simulation time (149.1 μs in this case). The dissociation of benzene from the L99A mutant T4 Lysozyme was investigated in a hybrid MSM/infrequent MetaD study (Mondal et al., 2018) using the CHARMM36 force-field (Best et al., 2012). The MSM was constructed from unbiased MD trajectories, and gave a k_{off} of 310 ± 130 s^{-1}, which was marginally closer to the experimental k_{off} (950 ± 20 s^{-1}) (Feher et al., 1996) than the value reported by the accompanying infrequent MetaD simulations (k_{off} = 270 ± 100 s^{-1}) (Mondal et al., 2018) and considerably closer than the previous FA-MetaD-based predictions (see **Table 1**) (Wang et al., 2018) However, the statistical uncertainty in the MSM-derived k_{off} was quite large, and the calculation required more simulation time (60 μs) compared biased MD approaches to obtain similar uncertainties: FA-MetaD/Infrequent MetaD studies typically require 6–12 μs (Casasnovas et al., 2017; Wang et al., 2018; Capelli et al., 2020) and LiGaMD (Miao et al., 2020) required ~ 5 μs.

The use of biased simulations can greatly reduce the sampling requirements. Wu et al., (2016) showed that by integrating unbiased MD with umbrella sampling simulation data, only 5%–10% of the unbiased data was necessary to estimate the dissociation rate of the trypsin-benzamidine complex up to statistical significance (k_{off} = 1170s^{-1} [617s^{-1}, 2120s^{-1}]). A combination of 500 μs of unbiased MD and 1 μs of Hamiltonian replica exchange simulation was used to create an MSM model describing the binding of the oncoprotein fragment Mdm2 and a peptide inhibitor PMI. Estimates based on two different post-processing schemes yielded values of k_{off} = 0.125 s^{-1} [0.025 s^{-1}, 0.66 s^{-1}] and k_{off} = 1.13 s^{-1} [0.48 s^{-1}, 1.33 s^{-1}], corresponding to a 10–30-fold overestimation relative to experiments (k_{off} = 0.037 s^{-1} [0.029 s^{-1}, 0.04 s^{-1}]) (Paul et al., 2017).

4 PATH SAMPLING METHODS

Path sampling methods focus on generating an ensemble of transition pathways between bound and unbound states. Typically, this class of methods accelerates the unbinding event by exploiting restarting strategies to favor the sampling of short trajectories in the vicinity of the transition state, which are then used to reconstruct the full unbinding process. Weighted Ensemble (WE) (Huber and Kim, 1996), milestoning (Cho et al., 2006; Elber, 2007), transition state-partial path interface sampling (TS-PPTIS) (Juraszek et al., 2013), and adaptive multilevel splitting (AMS) (Cérou and Guyader, 2007; Cérou et al., 2011) are path sampling methods that were employed in calculations of k_{off} for ligand/protein complexes.

TABLE 1 | Quantitative in silico calculations (we highlighted in boldface the simulations that are below one order of magnitude for the predicted results with respect to the experimental ones)

Target	Technique	T [K]	Force field	k_{off} (sim) [s^{-1}]	k_{off} (Exp) [s^{-1}]	Simulation time [μs]	Ref	Year
Trypsin/Benzamidine	SEEKR	298	Amber14SB + GAFF	83 ± 14	600 ± 300	19	10.1021/acs.jpcb.6b09388	2017
Trypsin/Benzamidine	SEEKR	298	Amber14SB + GAFF	174 ± 9	600 ± 300	4.4	10.1021/acs.jctc.0c00495	2020
Trypsin/Benzamidine	SEEKR2	298	Amber14SB + GAFF	990 ± 70	600 ± 300	5	10.26434/chemrxiv-2021-pplfs	2021
Trypsin/Benzamidine	M-WEM	298	Amber14SB + GAFF	791 ± 197	600 ± 300	0.48	10.1021/acs.jctc.1c00803	2022
Trypsin/Benzamidine	Inf-MetaD	300	Amber99SB-ILDN	9.1 ± 2.5	600 ± 300	5	10.1073/pnas.1424461112	2015
Trypsin/Benzamidine	Inf-MetaD	300	Amber14SB + GAFF	4176 ± 324	600 ± 300	—	10.1021/acs.jctc.8b00934	2019
Trypsin/Benzamidine	MSM	298	Amber99SB + GAFF	$(9.5 \pm 3.3)\cdot10^{4}$	600 ± 300	50	10.1073/pnas.1103547108	2011
Trypsin/Benzamidine	MSM	—	—	$2.8 \cdot10^{4}$	600 ± 300	7.7	10.1021/ct400919u	2014
Trypsin/Benzamidine	MSM	—	Amber99SB + GAFF	131 ± 109	600 ± 300	149.1	10.1038/ncomms8653	2015
Trypsin/Benzamidine	MSM	298	Amber99SB + GAFF	1170 [617, 2120]	600 ± 300	58.28	10.1073/pnas.1525092113	2016
Trypsin/Benzamidine	WExplore	300	Charmm36 + CGenFF	$5.56 \cdot10^{4}$	600 ± 300	4.1	10.1016/j.bpj.2017.01.006	2017
Trypsin/Benzamidine	REVO	300	Charmm36 + CGenFF	2660	600 ± 300	8.75	10.1063/1.5100521	2019
Trypsin/Benzamidine	LiGaMD	300	Amber14SB + GAFF	3.53 ± 1.41	600 ± 300	5	10.1021/acs.jctc.0c00395	2020
Trypsin/Benzamidine	dcTMD	290	Amber99SB*	270 ± 40	600 ± 300	10000[c]	10.1038/s41467-020-16655-1	2020
Trypsin/Benzamidine	AMS	298	Charmm36 + CGenFF	260 ± 240	600 ± 300	2.3	10.1021/acs.jctc.6b00277	2016
Trypsin/Benzamidine	OPES	300	Amber14SB + GAFF	687	600 ± 300	3.2	arXiv:2204.05572	2022
T4L L99A-Benzene	In-MetaD	300	Charmm22*	6.0 ± 2.2	950 ± 200[a]	6.7	10.1039/c7sc01627a	2017
T4L L99A-Benzene	FA-MetaD	300	Charmm22*	5.7 ± 2.3	950 ± 200[a]	5.5	10.1063/1.5024679	2018
T4L L99A-Benzene	In-MetaD	303	Charmm36	270 ± 100	950 ± 200	—	10.1371/journal.pcbi.1006180	2018
T4L L99A-Benzene	MSM	303	Charmm36	310 ± 130	950 ± 200	60	10.1371/journal.pcbi.1006180	2018
T4L L99A-Indole	In-MetaD	300	Charmm22* + CGenFF	9.8 ± 10.2	325 ± 75[b]	4.5	10.1063/1.5024679	2018
T4L L99A-Indole	FA-MetaD	300	Charmm22* + CGenFF	6.0 ± 3.7	325 ± 75[b]	2.0	10.1063/1.5024679	2018
μOpioid receptor-morphine	In-MetaD	300	Charmm36 + CGenFF	$(5.7 \pm 0.5)\cdot10^{-2}$	$(2.3 \pm 0.2)\cdot10^{-2}$	6	10.1063/5.0019100	2020
μOpioid receptor-bruprenorphine	In-MetaD	300	Charmm36 + CGenFF	$(2.1 \pm 0.3)\cdot10^{-2}$	$(1.8 \pm 0.3)\cdot10^{-3}$	19	10.1063/5.0019100	2020
μOpioid receptor-Fentanyl	In-MetaD	310	Charmm36m + CGenFF	$(2.6 \pm 0.8)\cdot10^{-2}$ (HID) $(3.8 \pm 1.4)\cdot10^{-1}$ (HIE) 1.1 ± 0.3 (HIP)	$4.2 \cdot 10^{-3}$	6	10.1021/jacsau.1c00341	2021
TSPO-PK11195	REVO	300	Charmm36 + CGenFF	(D1)$6.4 \cdot 10^{-5}$ (D2)$6.67\cdot10^{1}$ (D3)$6.4 \cdot 10^{-3}$ (D4)$4.1 \cdot 10^{-3}$ (4RYI)$6.0 \cdot 10^{-4}$ (D1-D4 different docked poses)	$4.9 \cdot 10^{-4}$	40	10.1016/j.bpj.2020.11.015	2021
c-Src kinase-dasatinib	In-MetaD	300	OPLS	$(4.8 \pm 2.4)\cdot10^{-2}$	$5.6 \cdot 10^{-2}$ $1.1 \cdot 10^{-3}$	~7–8	10.1126/sciadv.1700014	2017
Src kinase - imatinib	TS-PPTIS	305	Amber99SB*-ILDN + GAFF (QM/MM)	0.026	0.11 ± 0.08	—	10.1021/acs.jctc.8b00687	2018

(Continued on following page)

TABLE 1 | (*Continued*) Quantitative in silico calculations (we highlighted in boldface the simulations that are below one order of magnitude for the predicted results with respect to the experimental ones)

Target	Technique	T [K]	Force field	k_{off} (sim) [s^{-1}]	k_{off} (Exp) [s^{-1}]	Simulation time [μs]	Ref	Year
Epoxide Hydrolase-TPPU	WExplore	300	Charmm36 + CGenFF	$2.4 \cdot 10^{-2}$ [$3.6 \cdot 10^{-3}$ s^{-1}, $4.4 \cdot 10^{-2}$ s^{-1}]	$1.5 \cdot 10^{-3}$	6	10.1021/jacs.7b08572	2018
p38 kinase/1-(3-(tert-butyl)-1- (p-tolyl)-1H-pyrazol-5-yl)urea	In-MetaD	300	Amber99SB-ILDN + GAFF	0.020 ± 0.011	0.14	6.8	10.1021/jacs.6b12950	2017
M2 muscarinic receptor/ iperoxo	FA-MetaD	310	Amber14SB + GAFF	$(3.7 \pm 0.7) \cdot 10^{-4}$	$(1.0 \pm 0.2) \cdot 10^{-2}$	8	10.1021/ acs.jpclett.0c00999	2020
HSP90-inhibitor	dcTMD	300	Amber99SB + GAFF	$(1.6 \pm 0.2) \cdot 10^{-3}$	$(3.4 \pm 0.2) \cdot 10^{-2}$	5000[c]	10.1038/s41467-020-16655-1	2020
Mdm2/PMI	MSM	300	Amber99SB-ILDN	0.125 [0.025, 0.66] 1.13 [0.48, 1.33] (Different rate matrix estimators)	0.037 [0.029, 0.04]	500	10.1038/s41467-017-01163-6	2017
Mdm2/p53	MSM	300	Amber99SB-ILDN-NMR	$1.9 \cdot 10^{5}$	2.1	831	10.1016/ j.bpj.2017.07.009	2017
SH3 Domain—1CKB	Pep-GaMD	300	Amber14SB	$(1.45 \pm 1.17) \cdot 10^{-3}$	$8.9 \cdot 10^{-3}$	3	10.1063/5.0021399	2020
MtKatG—Isonazid	τRAMD + extrapolation	300	CHARMM36 + SwissParam	$(2.8 \pm 3.7) \cdot 10^{-2}$	$(2.0 \pm 0.3) \cdot 10^{-2}$	—	10.1021/ acs.jpclett.1c02952	2021

[a]*The Authors in the original work considered the experimental k_{off} at 293 K (800 ± 200 s^{-1}), while they simulated the system at 300 K. Here we choose to put the value at the closest temperature available in experiments (303K—950 ± 200 s^{-1}). Both the experimental values come from (Feher et al., 1996).*

[b]*The experimental value has been measured at 293 K.*

[c]*For dcTMD, computational time is referred to 1D Langevin simulator, and the authors says that "1 ms of simulation time at a 5 fs time step take ~6 h of wall-clock time on a single CPU".*

4.1 Weighted Ensemble Methods

4.1.1 Basic principles

A set of unbiased molecular dynamics trajectories with equivalent statistical weights are spawned in parallel from a ligand/protein complex in the ground-state configuration (Huber and Kim, 1996). The configuration space is then subdivided into bins, which the trajectories/walkers navigate through. The weighted ensemble (WE) method aims to maintain a fixed number (N) of walkers per bin. Thus, the occupancy of the bins is calculated at specific resampling intervals τ_{int}. If the number of walkers in a given bin is lower than N, one or more of the walkers are cloned, with each daughter trajectory receiving a fraction of the weight of the original. Conversely, in regions populated by a number of walkers exceeding N, two or more trajectories are merged, with the resulting trajectory inheriting the weights of its constituents (Zuckerman and Chong, 2017). This process results in a resampled trajectory space spanning the bound, intermediate and unbound states from which k_{off} values can be obtained (Zhang et al., 2010).

Notably, the method does not require a detailed a prori definition of differentiable collective variables, and it is embarrassingly parallel. Given that the availability of Tier-0 and Tier-1 machines has grown significantly since the method was first formulated, several scalable open-source implementations have emerged. These include WExplore (Dickson and Brooks, 2014), Wepy (Lotz and Dickson, 2020), and REVO (Resampling of Ensembles by Variation Optimization) (Donyapour et al., 2019). The latter is a method featuring a novel resampling algorithm replacing bins in configurational space with a system-specific all-to-all pairwise distance matrix between walkers, thereby decreasing the correlation between trajectories. The novel concurrent adaptive sampling (CAS) algorithm (Ahn et al., 2017) builds on the traditional WE method by adaptively constructing macrostates (represented by n-dimensional Voronoi cells) while approximating the committor function of each macrostate, and clustering the macrostates according to their committor functions as the simulation progresses. This improves the efficiency of WE simulations in high-dimensional systems, by directing computational power to sampling portions of configuration space that are closer to the "product" configuration.

4.1.2 Applications

The k_{off} of the trypsin-benzamidine complex as calculated by WExplore (5560 s^{-1}) (Dickson and Lotz, 2017) overestimated by one order of magnitude the experimental value (600 s^{-1}) (Guillian and Thusias, 1970). This value was calculated from five independent WExplore runs, corresponding to an aggregate simulation time of 4.1 μs. Using clustering-based confirmation space network analysis techniques (Dickson and Brooks, 2013), three distinct ligand exit pathways were unearthed from the trajectories. The trypsin-benzamidine system was later investigated again with REVO (Donyapour et al., 2019). Based on five independent REVO runs, giving a total of 8.75 μs, a k_{off} of 2660 s^{-1} was predicted—a minor improvement on WExplore, but an overestimation nonetheless. WExplore was also employed to estimate the dissociation rate of the TPPU inhibitor from soluble epoxide hydrolase. The calculated k_{off} ($2.4 \cdot 10^{-2}$ s^{-1} [$3.6 \cdot 10^{-3}$ s^{-1}, $4.4 \cdot 10^{-2}$ s^{-1}]) was one order of magnitude greater than the experimental value of $3.6 \cdot 10^{-3}$ s^{-1}. (Lotz and Dickson, 2018), and required 6 μs of simulation time to compute. However, the reason for the systematic overestimations of k_{off} is not explicitly addressed. REVO was recently employed (Dixon et al., 2021) to quantify k_{off} values for five distinct bound poses of the PK-11195 radioligand in complex with TSPO (see **Table 1**), using a cumulative 5.18 μs of simulation time per pose. The calculated

values for the poses spanned five orders of magnitude, and the pose with the most favorable docking score (pose D1, $k_{off} = 6.4 \times 10^{-5}\ s^{-1}$) exhibited the closest agreement with the experimental value ($4.9 \times 10^{-4}\ s^{-1}$) out of all the docked poses. All the of the studies described here made use of the CHARMM36 (Best et al., 2012) and CGenFF (Vanommeslaeghe et al., 2011) force fields. At present, the CAS method described in **Section 4.1.1** has been successfully applied to host-guest systems only (Ahn et al., 2020).

4.2 Milestoning

4.2.1 Basic principles

Here, the configuration space is treated as a coarse mesh characterized by slowly relaxing variables, such as native contacts and/or distances between chemical groups that describe the ligand unbinding process (Cho et al., 2006; Elber, 2007). The mesh must be coarse enough for distinct long-lived metastable states to emerge, but fine enough to ensure that transitions between the interfaces between the mesh's cells or "milestones" are accessible in MD simulations. Equilibrium configurations for each milestone are usually generated with "pulling" SMD simulations, or with a series of MD simulations in which the diffusing group is harmonically restrained to the milestone surface. Afterward, a set of trajectories is spawned from each milestone, and whenever a trajectory reaches a new neighboring milestone, it is terminated. In practice, the criteria for termination of trajectories vary depending on the implementation. The lifetime and flux (i.e., number of trajectories passing through the milestone per unit time) associated with each milestone may be used to compute the ligand residence time (Elber, 2020).

Practical implementations of milestoning in ligand unbinding studies fall into two categories: 1) The Simulation Enabled Estimation of Kinetic Rates (SEEKR) (Votapka et al., 2017) approach, which exploits milestoning theory in a multiscale framework based on MD and Brownian Dynamics (BD) simulations (Luty et al., 1993). The milestones are nested spherical shells surrounding the binding pocket. Transitions between milestones close to the binding pocket are simulated using all-atom MD. Meanwhile, transitions between the more diffuse milestones further away are described by cheaper BD simulations—where fast sampling of rigid body interactions is more important than detailed sampling of ligand conformations. An updated implementation of SEEKR, named MMVT SEEKR, has been subsequently proposed (Jagger et al., 2020): it circumvents the need to compute the equilibrium distribution for all the milestones, reducing the computational time needed to compute kinetics constants. 2) The recently formulated weighted ensemble milestoning (WEM) methods combine milestoning theory with WE methods. Here, the configurational space between the milestones is split into bins, and WE simulations are run in between milestones to achieve faster convergence (Ray and Andricioaei, 2020).

4.2.2 Applications

All applications to kinetics of biological systems so far are based on AMBER14SB (Maier et al., 2015) and GAFF (Wang et al., 2004) force fields and applied to the trypsin-benzamidine complex. SEEKR yielded a k_{off} of 83 ± 14 s^{-1} for the trypsin-benzamidine system using 19 μs of aggregate MD and ten spherical milestones. These results underestimate the experimental value (600 ± 300 s^{-1}) (Guillian and Thusias, 1970). MMVT SEEKR improved the k_{off} estimate (174 ± 9 s^{-1}), with only a quarter of the aggregate simulation time (4.4 μs) used in the prior SEEKR study. WEM (Ray and Andricioaei, 2020) gave a further improvement k_{off} = 791 ± 197 s^{-1}, using a mere 0.5 μs of simulation time (Ray et al., 2022).

4.3 Transition State-Partial Path Transition Interface Sampling

4.3.1 Basic principles

In transition state-partial path transition interface sampling (TS-PPTIS) (Juraszek et al., 2013) an initial metadynamics calculation is performed to determine the transition state and the free energy barrier along a given CV. Then, the transmission coefficient is estimated, similarly to the PPTIS method (Van Erp et al., 2003; Moroni et al., 2004) by foliating the barrier region along the CV with interfaces and sampling short trajectories spanning three consecutive interfaces. These trajectories are sampled using transition path sampling (Pratt, 1986; Dellago et al., 1998). Under the assumptions that the dynamics in the barrier region is diffusive and there are no memory effects for travelled distances beyond two interfaces, the kinetic rates are independent of the CV.

4.3.2 Applications

TS-PPTIS was used to compute the k_{off} of the imatinib-Src kinase complex (Morando et al., 2016). The calculation used 5 CVs: 2 path collective variables (Branduardi et al., 2007), a CV counting the number of water molecules interacting with the ligand and the binding cavity, and two distances between key residues of Src characterizing the motion of the kinase A-loop. Using AMBER99SB*-ILDN and GAFF, the authors computed a value of $k_{off} = 0.0114\ s^{-1}$ [$0.001\ s^{-1}$, $0.139\ s^{-1}$], which is slow (but within statistical significance) compared to experiments (k_{off} = 0.11 ± 0.08 s^{-1}). In a separate work (Haldar et al., 2018), the authors refined the prediction by computing a free energy correction from the MM to a hybrid quantum mechanics/molecular mechanics Hamiltonian using a replica exchange thermodynamic integration scheme (Woods et al., 2003) and Metropolis-Hastings Monte Carlo sampling (Woods et al., 2008). This correction does not account for dynamical effects but only for changes in the free energy. The computed correction to k_{off} was small but consistent with faster dissociation dynamics obtaining a corrected value of $k_{off} = 0.026\ s^{-1}$.

4.4 Adaptive Multilevel Splitting

4.4.1 Basic principles

Similarly to WE, adaptive multilevel splitting (AMS) (Cérou and Guyader, 2007; Cérou et al., 2011) relies on a set of trajectories that are systematically cloned or killed. However, AMS does not require bins. Instead, the algorithm is initialized by generating a set of "loop" trajectories starting and ending in the bound state. At each iteration, the replica that travelled the least distance *d*

from the bound state (measured through a CV) is killed, and a new loop is created by restarting a simulation from a point at the same distance d previously visited by one of the remaining replicas. This is repeated until all loops travelled a distance above a threshold value (defining the unbound state) before returning to the bound state. The dissociation rate is then estimated from this collection of trajectories.

4.4.2 Applications

AMS was used to calculate the dissociation rate of trypsin-benzamidine using the CHARMM36 force field (Best et al., 2012) for trypsin and CGenFF (Vanommeslaeghe et al., 2011) for the ligand (Teo et al., 2016). As the CV, the authors used the distance between the center of mass of benzamidine and the alpha carbons of the amino acids close to the binding site. A suitable value for the threshold value of the CV was obtained through a steered MD simulation. Furthermore, 130 ns unbiased MD simulation was run to estimate the average time of a looping trajectory under the assumption that the short loops thus sampled represented the large majority of loops and thus dominated the average loop duration. In total, 2.3 μs of simulations were used to obtain a k_{off} = 260 s^{-1} ± 240 s^{-1}, in good agreement with experimental measurements.

5 LIMITATIONS ASSOCIATED WITH FORCE FIELDS

Table 1 summarizes the k_{off} predictions of various ligand/protein systems obtained using the methods discussed in previous sections. For completeness, we also report the temperature, total simulation time, and force field used. In about one-third of the cases, spanning all different classes of methodologies and force fields, the theoretical predictions are in the same order of magnitude as the experimental values, and in a few cases (shown in boldface in **Table 1**) reproduce them within statistical error. In most cases, however, calculated values show discrepancies from 1 to 2 orders of magnitude, regardless of the method and force field. Similarly, the only predictive study reported so far (Paul et al., 2017) reports values with an error of 1–2 orders of magnitude (albeit with large statistical errors) relative to experimental data performed afterwards.[5] All these results, taken together, lead us to suggest that regular force fields may be, at times, not accurate enough to predict k_{off} values.

Determining the source of the observed errors is a difficult task without dedicated studies as the accuracy of the predictions depends on methodological aspects, sampling accuracy, and the potential energy function, which are subject to mutual cancellation (or amplification) of error. In this and the next section, we discuss the literature focusing on the effect of the potential.

[5]All the other studies in **Table 1** are instead retrospective, and large-scale benchmarks in prospective settings (which are now common for binding affinity calculations) (Parks et al., 2020; Schindler et al., 2020) are missing.

5.1 Force Field Dependence of the Results

The published data indicate that careful parametrization of the force fields is essential to obtain k_{off} predictions. Comparison between the results obtained from brute force MD calculations on a set of ligands binding to a β-cyclodextrin (βCD) host showed that k_{off} predictions parametrizing βCD with the Q4MD force field (Cézard et al., 2011) were consistently more accurate (within one order of magnitude of experimental values) than the GAFF-based (Wang et al., 2004) estimates (Tang and Chang, 2018). On the other hand, the k_{on} estimates were consistently better for the GAFF model, which points to the difficulty of obtaining transferable potentials. In the case of benzene unbinding from L99A T4 lysozyme, infrequent MetaD simulations using CHARMM22 (MacKerell et al., 1998; MacKerell Jr. et al., 2004) yielded a significantly underestimated k_{off} in the range of 4–10 s^{-1} (Wang et al., 2018), while the same method combined with CHARMM36 (Best et al., 2012) produced a k_{off} (270 ± 100 s^{-1}) (Mondal et al., 2018) considerably closer to the experimental value of 950 ± 20 s^{-1} (Feher et al., 1996). Although different CVs were used in these two works (see **Section 2.4.2**), the effect of the force field cannot be ruled out. Indeed, the two force fields differ only in a few dihedral potential terms (Best et al., 2012) that control the rigidity of secondary structures, and in particular two helices of T4 which control benzene's access to the binding pocket. Finally, we mention here the work of (Vitalini et al., 2015), where it was shown that slow relaxation timescales of two small peptides using five protein force fields (AMBER99SB-ILDN (Lindorff-Larsen et al., 2010), AMBERff03 (Duan et al., 2003), OPLS-AA/L (Kaminski et al., 2001), CHARMM27 (MacKerell et al., 2000), and GROMOS43a1 (Daura et al., 1998)) differ up to two orders of magnitude. Given the importance of slow protein conformational changes in unbinding kinetics (Plattner and Noé, 2015), this result further highlights the role of force fields for accurate rate calculations.

5.2 Polarization and Charge Transfer Effects

Traditional force fields describe electrostatics using fixed point charges. This representation is extremely efficient and works remarkably well, even in the case of systems with high electric fields (Mironenko et al., 2021). However, such a scheme cannot adapt to changes in the electrostatic environment observed during ligand unbinding. Recently, some of us (Capelli et al., 2020) found that electrostatic effects contribute significantly to the force field misrepresentation of protein-ligand interactions at the transition state of the M2-iperoxo complex. Furthermore, the work of Haldar and coworkers (Haldar et al., 2018) showed that accounting for changes in charge distribution resulted in free energy corrections ranging from 1.9 to 4.7 kcal/mol as the ligand progressed from the hydrophobic binding pocket to the solvated state. Metalloenzymes (representing 40%–50% of all proteins in the PDB database (Chen et al., 2019)) and highly charged protein-ligand systems are also quite challenging to describe with traditional force fields (Li and Merz, 2017). Indeed, for the

latter systems, FF-based binding free energy calculations resulted in significant systematic errors (Rocklin et al., 2013). Overall, these results show that going beyond standard fixed-charged models is in many cases desirable to improve accuracy.

6 PERSPECTIVES: FROM POLARIZABLE FORCE FIELDS TO QM/MM CALCULATIONS TOWARDS THE EXASCALE

Force fields have been overwhelmingly successful in predicting equilibrium properties such as free energies of binding (Karplus and McCammon, 2002; Wang et al., 2015; Robustelli et al., 2022). Indeed, force fields are traditionally fitted to reproduce equilibrium experimental measurements (ensemble averages) and geometries obtained with quantum mechanical methods. As a result, their performance is expected to peak in the regions near the free energy minima (e.g., the bound state) rather than near the kinetically relevant transition states, where small errors are exponentially amplified in k_{off} predictions.[6] After observing discrepancies of two orders of magnitude in the kinetic predictions of several force fields, Vitalini and coworkers (Vitalini et al., 2015) suggested that kinetic information should be included in the fitting process. In general, designing new parametrization strategies for force fields is still a very active area of research (He et al., 2020; Giannos et al., 2021; Qiu et al., 2021). This is not surprising, given the issues discussed in **Section 5**. For example, methods to include polarization effects within a fixed-charge scheme (Kelly and Smith, 2020) and multisite models for transition metal ions have been developed (Li and Merz, 2017).

A different direction pursued by the modeling community is instead to use potential energy functions that go beyond the biomolecular force fields' simple representation of electrostatics (e.g., polarizable force fields, hybrid quantum mechanics/molecular mechanics (QM/MM) calculations, machine learning potentials). Without any claim of being comprehensive, here we provide a brief perspective on the role of these methods in the upcoming era of exascale computing.

6.1 Polarizable Force Fields

Polarizable force fields for biomolecules (Jing et al., 2019) including AMBER ff02pol (Wang Z. X. et al., 2006), AMOEBA (Ponder et al., 2010) CHARMM Drude (Baker et al., 2010), CHARMM-FQ (Patel and Brooks, 2003; Patel and Brooks, 2004), SIBFA (Piquemal et al., 2007), and ABEEMσπ (Liu et al., 2017) aim at providing an empirical description electronic polarizability. Simulations based on these potentials could dramatically improve the modeling of transition states in cases where electronic polarization and charge transfer may be linked to non-trivial rearrangements of hydrogen bonds and hydrophobic interactions (Schmidtke et al., 2011; Schiebel et al., 2018). Although polarizable force fields have recently shown excellent accuracy in prospective predictions of binding affinities in model systems (Amezcua et al., 2022), to the best of our knowledge, they have not been used for protein-ligand k_{off} predictions yet. Notably, in a very recent paper (Yue et al., 2022), it was shown how using a polarizable force field improved the accuracy of the predictions of anion permeation rates in fluoride channels compared to predictions based on standard fixed charge schemes, highlighting the necessity of using polarizable models for treating such processes. Although this is not a ligand/protein system, this work further showcases the limitations of conventional force fields in treating electrostatic interactions as well as the potential of polarizable models.

6.2 QM/MM Simulations

DFT-based QM/MM simulations treat a small region of interest (in our case this could be a ligand and the protein residues interacting with it) at the DFT level, while the overall computational cost is balanced by MM treatment of other regions (Kulik, 2018). The form of the potential energy is a hybrid model between classical mechanics and quantum chemistry:

$$U = U_{QM} + U_{MM} + U_{QM/MM} \tag{3}$$

where $U_{QM/MM}$ denotes the interaction between atom groups assigned to the QM region and MM region. DFT-based QM/MM simulations include both electronic polarizability and charge transfer effects (Blumberger, 2008; Capelli et al., 2020), and they address the problem of transferability, as they do not rely on optimizations against predefined training data sets. These approaches can tackle important biomedicine problems such as the study of transition-metal-based drugs binding to proteins (Calandrini et al., 2015) or the description of enzymatic reactions. (Carloni et al., 2002; Liao and Thiel, 2013; Roston et al., 2016; Caldararu et al., 2018; Kulik, 2018; Piniello et al., 2021) However, these simulations are orders of magnitude more expensive than any of the potentials described so far, and hence achieving high statistical accuracy with such an approach is obviously extremely challenging.

6.2.1 Parallel Computing in DFT-Based QM/MM

Modern supercomputers are currently breaching the exascale limit in the United States (Schneider, 2022), Japan, and China.[7] Exascale calculations however remain one of the major challenges in molecular simulations (Hospital et al., 2015; Páll et al., 2015). Recent advances in massively scalable QM/MM codes, such as that developed in Juelich in collaboration with European universities (Olsen et al., 2019; Bolnykh et al., 2020a) (see

[6]The k_{off} values depend exponentially on the height of the dissociation free energy barrier, so even small inaccuracies in the potential energy may impact dramatically kinetics calculations.

[7]According to the Top500 list (https://www.top500.org/), which ranks computers based on their performance on the HPLinpack benchmark (Dongarra et al., 2003), currently there are no machines that have exceeded the exascale limit. The Fugaku supercomputer in Japan showed performances above one EFlop/s, but on a different benchmark (Kudo et al., 2020). In China, one or two exascale supercomputers might be already operating(Ma et al., 2022; Schneider, 2022).

Supplementary Material S7) and their successful applications to predict free energy landscapes associated with biological processes (Bolnykh et al., 2020a; Chiariello et al., 2020) brings us to suggest that in a not-too-far future QM/MM calculations may exploit the unprecedented power of exascale computing for direct MD simulations of ligand (un)binding (Bolnykh et al., 2020b, Bolnykh et al., 2021).

6.2.2 Machine Learning in QM/MM

Neural network models of the potential energy function have emerged as a promising route to obtaining near-DFT accuracies (Unke et al., 2021; Kocer et al., 2022) at a computational cost only 1–2 orders of magnitude slower than force fields. Applications to the kinetics of chemical reactions have been published (Stocker et al., 2020; Yang et al., 2022) and in principle, they could be used to model DFT-based QM/MM predictions of ligand poses during the unbinding process. However, ML potentials are currently still limited to small molecule applications and robust solutions to model long-range interactions have yet to emerge (Yue et al., 2021). The advent of exascale computing could dramatically expand the domain of applicability of such approaches (Lu et al., 2021). Moreover, several approaches to solve these issues have been proposed based on hybrid machine learning/molecular mechanics models (Shen and Yang, 2018; Rufa et al., 2020; Böselt et al., 2021; Gastegger et al., 2021) (see **Supplementary Material S8** for details).

7 CONCLUSION

We have reviewed an array of rather diverse methods able to predict unbinding kinetics constants using atomistic representations of the biomolecules involved. These techniques have shown tremendous progress in the last years: considering trypsin-benzamidine as a benchmark system (as seen in **Table 1**), we start from 2–3 orders of magnitude in k_{off} error in the pioneering MSMs of De Fabritiis and co-workers (Buch et al., 2011) to an error of less than 1 order of magnitude in some of the most recent calculations (Plattner and Noé, 2015; Votapka et al., 2017; Brotzakis et al., 2019; Wolf et al., 2020). Despite these impressive methodological advances, the domain of applicability and accuracy appears to be still limited by current force fields. Better parametrization and polarizable force fields (Lin and MacKerell, 2019) promise to improve the quality of the potential energy model at a reasonable cost at a reasonable computational cost (Lemkul et al., 2016). Another possibility is the use of massively parallel DFT-QM/MM complemented by ML techniques, which include electronic polarizability as well as charge transfer. This approach could address the issue of transferability of current biomolecular force fields. However, the accuracy of these approaches is yet to be established.

Traditionally, computational drug discovery has used a combination of methods such as docking (Ferreira et al., 2015), quantitative structure-activity relationship (QSAR) modeling (Dossetter et al., 2013), free-energy methods (Cournia et al., 2017), and (recently) ML-based approaches (Zhao et al., 2020) to improve the binding affinity of a compound during lead optimization. Computer-aided ligand design campaigns could enormously profit from the design of so-called transition state analogues which, in the case of enzyme inhibitors, have been correlated with release rates that are orders of magnitude slower than product release (Schramm, 2013; Schramm 2015; Svensson et al., 2015). We hope that approaches beyond the use of standard force fields, such as those discussed here, will lead in a not-too-distant future to the accurate description of the energetics and structural determinants of the unbinding transition states, giving an unprecedented boost to the discovery of promising new small molecules and the optimization of known drugs.

AUTHOR CONTRIBUTIONS

KA is the first author. AR and RC made equal contributions as second authors. DM is third author, WL is the fourth. Prof. PC is the PI and the last author.

FUNDING

AR, PC, and DM acknowledge support from the Helmholtz European Partnering program ("Innovative high-performance computing approaches for molecular neuromedicine"). KA and PC acknowledge funding from the Human Brain Project (EU Horizon 2020). W.L. appreciates the National Natural Science Foundation of China No. 21505134.

ACKNOWLEDGMENTS

The authors are indebted to Prof. Michele Parrinello for many discussions on the topic of kinetics of ligand binding to proteins.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.899805/full#supplementary-material

REFERENCES

Ahn, S.-H., Grate, J. W., and Darve, E. F. (2017). Efficiently Sampling Conformations and Pathways Using the Concurrent Adaptive Sampling (CAS) Algorithm. *J. Chem. Phys.* 147, 074115. doi:10.1063/1.4999097

Ahn, S.-H., Jagger, B. R., and Amaro, R. E. (2020). Ranking of Ligand Binding Kinetics Using a Weighted Ensemble Approach and Comparison with a Multiscale Milestoning Approach. *J. Chem. Inf. Model.* 60, 5340–5352. doi:10.1021/acs.jcim.9b00968

Amaral, M., Kokh, D. B., Bomke, J., Wegener, A., Buchstaller, H. P., Eggenweiler, H. M., et al. (2017). Protein Conformational Flexibility Modulates Kinetics and Thermodynamics of Drug Binding. *Nat. Commun.* 8, 2276. doi:10.1038/s41467-017-02258-w

Amezcua, M., Setiadi, J., Ge, Y., and Mobley, D. L. (2022). An Overview of the SAMPL8 Host-Guest Binding Challenge. *ChemRxiv*, 1–42. doi:10.26434/chemrxiv-2022-lwd0h

Ansari, N., Rizzi, V., and Parrinello, M. (2022). Water Regulates the Residence Time of Benzamidine in Trypsin. *ArXiv*, 1–16. Available at: http://arxiv.org/abs/2204.05572. (Accessed April 16, 2022).

Baker, C. M., Lopes, P. E. M., Zhu, X., Roux, B., and MacKerell, A. D. (2010). Accurate Calculation of Hydration Free Energies Using Pair-specific Lennard-Jones Parameters in the CHARMM Drude Polarizable Force Field. *J. Chem. Theory Comput.* 6, 1181–1198. doi:10.1021/ct9005773

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* 100, 1–4. doi:10.1103/PhysRevLett.100.020603

Bartels, C., and Karplus, M. (1997). Multidimensional Adaptive Umbrella Sampling: Applications to Main Chain and Side Chain Peptide Conformations. *J. Comput. Chem.* 18, 1450–1462. doi:10.1002/(sici)1096-987x(199709)18:12<1450::aid-jcc3>3.0.co;2-i

Bernetti, M., Rosini, E., Mollica, L., Masetti, M., Pollegioni, L., Recanatini, M., et al. (2018). Binding Residence Time through Scaled Molecular Dynamics: A Prospective Application to hDAAO Inhibitors. *J. Chem. Inf. Model.* 58, 2255–2265. doi:10.1021/acs.jcim.8b00518

Best, R. B., and Hummer, G. (2009). Optimized Molecular Dynamics Force Fields Applied to the Helix–Coil Transition of Polypeptides. *J. Phys. Chem. B* 113, 9004–9015. doi:10.1021/jp901540t

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., et al. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* 8, 3257–3273. doi:10.1021/ct300400x

Blumberger, J. (2008). Free Energies for Biological Electron Transfer from QM/MM Calculation: Method, Application and Critical Assessment. *Phys. Chem. Chem. Phys.* 10, 5651–5667. doi:10.1039/B807444E

Bolnykh, V., Olsen, J. M. H., Meloni, S., Bircher, M. P., Ippoliti, E., Carloni, P., et al. (2020a). MiMiC: Multiscale Modeling in Computational Chemistry. *Front. Mol. Biosci.* 7, 1–4. doi:10.3389/fmolb.2020.00045

Bolnykh, V., Rossetti, G., Rothlisberger, U., and Carloni, P. (2021). Expanding the Boundaries of Ligand–Target Modeling by Exascale Calculations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 11. doi:10.1002/wcms.1535

Bolnykh, V., Rothlisberger, U., and Carloni, P. (2020b). Biomolecular Simulation: A Perspective from High Performance Computing. *Isr. J. Chem.* 60, 694–704. doi:10.1002/ijch.202000022

Böselt, L., Thürlemann, M., and Riniker, S. (2021). Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-phase Systems. *J. Chem. Theory Comput.* 17, 2641–2658. doi:10.1021/acs.jctc.0c01112

Bowman, G. R., Ensign, D. L., and Pande, V. S. (2010). Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* 6, 787–794. doi:10.1021/ct900620b

Branduardi, D., Gervasio, F. L., and Parrinello, M. (2007). From A to B in Free Energy Space. *J. Chem. Phys.* 126, 054103. doi:10.1063/1.2432340

Brotzakis, Z. F., 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., Sadiq, S. K., 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

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

Bussi, G., and Laio, A. (2020). Using Metadynamics to Explore Complex Free-Energy Landscapes. *Nat. Rev. Phys.* 2, 200–212. doi:10.1038/s42254-020-0153-0

Calandrini, V., Rossetti, G., Arnesano, F., Natile, G., and Carloni, P. (2015). Computational Metallomics of the Anticancer Drug Cisplatin. *J. Inorg. Biochem.* 153, 231–238. doi:10.1016/j.jinorgbio.2015.10.001

Caldararu, O., Feldt, M., Cioloboc, D., Van Severen, M.-C., Starke, K., Mata, R. A., et al. (2018). QM/MM Study of the Reaction Mechanism of Sulfite Oxidase. *Sci. Rep.* 8, 1–15. doi:10.1038/s41598-018-22751-6

Capelli, R., Lyu, W., Bolnykh, V., Meloni, S., Olsen, J. M. H., Rothlisberger, U., et al. (2020). Accuracy of Molecular Simulation-Based Predictions of Koff Values: A Metadynamics Study. *J. Phys. Chem. Lett.* 11, 6373–6381. doi:10.1021/acs.jpclett.0c00999

Carloni, P., Rothlisberger, U., and Parrinello, M. (2002). The Role and Perspective of Ab Initio Molecular Dynamics in the Study of Biological Systems. *Acc. Chem. Res.* 35, 455–464. doi:10.1021/ar010018u

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

Cérou, F., and Guyader, A. (2007). Adaptive Multilevel Splitting for Rare Event Analysis. *Stoch. Analysis Appl.* 25, 417–443. doi:10.1080/07362990601139628

Cérou, F., Guyader, A., Lelièvre, T., and Pommier, D. (2011). A Multiple Replica Approach to Simulate Reactive Trajectories. *J. Chem. Phys.* 134, 054108. doi:10.1063/1.3518708

Cézard, C., Trivelli, X., Aubry, F., Djedaïni-Pilard, F., and Dupradeau, F.-Y. (2011). Molecular Dynamics Studies of Native and Substituted Cyclodextrins in Different Media: 1. Charge Derivation and Force Field Performances. *Phys. Chem. Chem. Phys.* 13, 15103–15121. doi:10.1039/C1CP20854C

Chen, A. Y., Adamek, R. N., Dick, B. L., Credille, C. V., Morrison, C. N., and Cohen, S. M. (2019). Targeting Metalloenzymes for Therapeutic Intervention. *Chem. Rev.* 119, 1323–1455. doi:10.1021/acs.chemrev.8b00201

Chiariello, M. G., Bolnykh, V., Ippoliti, E., Meloni, S., Olsen, J. M. H., Beck, T., et al. (2020). Molecular Basis of CLC Antiporter Inhibition by Fluoride. *J. Am. Chem. Soc.* 142, 7254–7258. doi:10.1021/jacs.9b13588

Cho, S. S., Levy, Y., and Wolynes, P. G. (2006). P versus Q : Structural Reaction Coordinates Capture Protein Folding on Smooth Landscapes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 586–591. doi:10.1073/pnas.0509768103

Chodera, J. D., Swope, W. C., Noé, F., Prinz, J.-H., Shirts, M. R., and Pande, V. S. (2011). Dynamical Reweighting: Improved Estimates of Dynamical Properties from Simulations at Multiple Temperatures. *J. Chem. Phys.* 134, 244107. doi:10.1063/1.3592152

Chong, L. T., Saglam, A. S., and Zuckerman, D. M. (2017). Path-sampling Strategies for Simulating Rare Events in Biomolecular Systems. *Curr. Opin. Struct. Biol.* 43, 88–94. doi:10.1016/j.sbi.2016.11.019

Copeland, R. A. (2021). Evolution of the Drug-Target Residence Time Model. *Expert Opin. Drug Discov.* 16, 1441–1451. doi:10.1080/17460441.2021.1948997

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

Cournia, Z., Allen, B., and Sherman, W. (2017). Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* 57, 2911–2937. doi:10.1021/acs.jcim.7b00564

Crommelin, D., and Vanden-Eijnden, E. (2009). Data-Based Inference of Generators for Markov Jump Processes Using Convex Optimization. *Multiscale Model. Simul.* 7, 1751–1778. doi:10.1137/080735977

Dama, J. F., Parrinello, M., and Voth, G. A. (2014). Well-Tempered Metadynamics Converges Asymptotically. *Phys. Rev. Lett.* 112, 240602. doi:10.1103/PhysRevLett.112.240602

Daura, X., Mark, A. E., and Van Gunsteren, W. F. (1998). Parametrization of Aliphatic CHn United Atoms of GROMOS96 Force Field. *J. Comput. Chem.* 19, 535–547. doi:10.1002/(sici)1096-987x(19980415)19:5<535::aid-jcc6>3.0.co;2-n

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

de Witte, W. E. A., Danhof, M., van der Graaf, P. H., and de Lange, E. C. M. (2018). The Implications of Target Saturation for the Use of Drug-Target Residence Time. *Nat. Rev. Drug Discov.* 18, 84. doi:10.1038/nrd.2018.234

Debnath, J., and Parrinello, M. (2020). Gaussian Mixture-Based Enhanced Sampling for Statics and Dynamics. *J. Phys. Chem. Lett.* 11, 5076–5080. doi:10.1021/acs.jpclett.0c01125

Dellago, C., Bolhuis, P. G., Csajka, F. S., and Chandler, D. (1998). Transition Path Sampling and the Calculation of Rate Constants. *J. Chem. Phys.* 108, 1964–1977. doi:10.1063/1.475562

Dickson, A., and Brooks, C. L. (2013). Native States of Fast-Folding Proteins Are Kinetic Traps. *J. Am. Chem. Soc.* 135, 4729–4734. doi:10.1021/ja311077u

Dickson, A., and Brooks, C. L. (2014). WExplore: Hierarchical Exploration of High-Dimensional Spaces Using the Weighted Ensemble Algorithm. *J. Phys. Chem. B* 118, 3532–3542. doi:10.1021/jp411479c.WExplore

Dickson, A., and Lotz, S. D. (2017). Multiple Ligand Unbinding Pathways and Ligand-Induced Destabilization Revealed by WExplore. *Biophysical J.* 112, 620–629. doi:10.1016/j.bpj.2017.01.006

Dixon, T., Uyar, A., Ferguson-Miller, S., and Dickson, A. (2021). Membrane-Mediated Ligand Unbinding of the PK-11195 Ligand from TSPO. *Biophysical J.* 120, 158–167. doi:10.1016/j.bpj.2020.11.015

Donati, L., Hartmann, C., and Keller, B. G. (2017). Girsanov Reweighting for Path Ensembles and Markov State Models. *J. Chem. Phys.* 146, 244112. doi:10.1063/1.4989474

Dongarra, J. J., Luszczek, P., and Petitet, A. (2003). The LINPACK Benchmark: Past, Present and Future. *Concurr. Comput. Pract. Exper.* 15, 803–820. doi:10.1002/cpe.728

Donyapour, N., Roussey, N. M., and Dickson, A. (2019). REVO: Resampling of Ensembles by Variation Optimization. *J. Chem. Phys.* 150, 244112–12. doi:10.1063/1.5100521

Dossetter, A. G., Griffen, E. J., and Leach, A. G. (2013). Matched Molecular Pair Analysis in Drug Discovery. *Drug Discov. Today* 18, 724–731. doi:10.1016/j.drudis.2013.03.003

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., et al. (2003). A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-phase Quantum Mechanical Calculations. *J. Comput. Chem.* 24, 1999–2012. doi:10.1002/jcc.10349

Durrant, J. D., and McCammon, J. A. (2011). Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* 9, 71. doi:10.1016/B978-0-12-809633-8.20154-410.1186/1741-7007-9-71

Elber, R. (2007). A Milestoning Study of the Kinetics of an Allosteric Transition: Atomically Detailed Simulations of Deoxy Scapharca Hemoglobin. *Biophysical J.* 92, L85–L87. doi:10.1529/biophysj.106.101899

Elber, R. (2020). Milestoning: An Efficient Approach for Atomically Detailed Simulations of Kinetics in Biophysics. *Annu. Rev. Biophys.* 49, 69–85. doi:10.1146/annurev-biophys-121219-081528

Emwas, A.-H., Szczepski, K., Poulson, B. G., Chandra, K., Mckay, R. T., Dhahri, M., et al. (2020). NMR as a "Gold Standard" Method in Drug Design and Discovery. *Molecules* 25, 4597. doi:10.3390/molecules25204597

Feher, V. A., Baldwin, E. P., and Dahlquist, F. W. (1996). Access of Ligands to Cavities within the Core of a Protein Is Rapid. *Nat. Struct. Mol. Biol.* 3, 516–521. doi:10.1038/nsb0696-516

Ferreira, L., Dos Santos, R., Oliva, G., and Andricopulo, A. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* 20, 13384–13421. doi:10.3390/molecules200713384

Folmer, R. H. A. (2018). Drug Target Residence Time: a Misleading Concept. *Drug Discov. Today* 23, 12–16. doi:10.1016/j.drudis.2017.07.016

Gastegger, M., Schütt, K. T., and Müller, K.-R. (2021). Machine Learning of Solvent Effects on Molecular Spectra and Reactions. *Chem. Sci.* 12, 11473–11483. doi:10.1039/d1sc02742e

Gelpi, J., Hospital, A., Goñi, R., and Orozco, M. (2015). Molecular Dynamics Simulations: Advances and Applications. *Aabc* 8, 37–47. doi:10.2147/AABC.S70333

Giannos, T., Lešnik, S., Bren, U., Hodošček, M., Domratcheva, T., and Bondar, A.-N. (2021). CHARMM Force-Field Parameters for Morphine, Heroin, and Oliceridine, and Conformational Dynamics of Opioid Drugs. *J. Chem. Inf. Model.* 61, 3964–3977. doi:10.1021/acs.jcim.1c00667

Grubmüller, H. (1995). Predicting Slow Structural Transitions in Macromolecular Systems: Conformational Flooding. *Phys. Rev. E* 52, 2893–2906. doi:10.1103/physreve.52.2893

Guillian, F., and Thusias, D. (1970). The Use of Proflavin as an Indicator in Temperature-Jump Studies of the Binding of a Competitive Inhibitor to Trypsin. *J. Am. Chem. Soc.* 92 (18), 5534–5536.

Guo, D., Mulder-Krieger, T., IJzerman, A. P., and Heitman, L. H. (2012). Functional Efficacy of Adenosine A2A Receptor Agonists Is Positively Correlated to Their Receptor Residence Time. *Br. J. Pharmacol.* 166, 1846–1859. doi:10.1111/j.1476-5381.2012.01897.x

Haldar, S., Comitani, F., Saladino, G., Woods, C., Van Der Kamp, M. W., Mulholland, A. J., et al. (2018). A Multiscale Simulation Approach to Modeling Drug-Protein Binding Kinetics. *J. Chem. Theory Comput.* 14, 6093–6101. doi:10.1021/acs.jctc.8b00687

Hänggi, P., Talkner, P., and Borkovec, M. (1990). Reaction-rate Theory: Fifty Years after Kramers. *Rev. Mod. Phys.* 62, 251–341. doi:10.1103/RevModPhys.62.251

He, X., Man, V. H., Yang, W., Lee, T.-S., and Wang, J. (2020). A Fast and High-Quality Charge Model for the Next Generation General AMBER Force Field. *J. Chem. Phys.* 153, 114502. doi:10.1063/5.0019056

Hooft, R. W. W., Van Eijck, B. P., and Kroon, J. (1992). An Adaptive Umbrella Sampling Procedure in Conformational Analysis Using Molecular Dynamics and its Application to Glycol. *J. Chem. Phys.* 97, 6690–6694. doi:10.1063/1.463947

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* 65, 712–725. doi:10.1002/prot.21123

Huber, G. A., and Kim, S. (1996). Weighted-ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophysical J.* 70, 97–110. doi:10.1016/S0006-3495(96)79552-8

Huber, T., Torda, A. E., and van Gunsteren, W. F. (1994). Local Elevation: A Method for Improving the Searching Properties of Molecular Dynamics Simulation. *J. Computer-Aided Mol. Des.* 8, 695–708. doi:10.1007/BF00124016

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

Invernizzi, M., and Parrinello, M. (2020). Rethinking Metadynamics: From Bias Potentials to Probability Distributions. *J. Phys. Chem. Lett.* 11, 2731–2736. doi:10.1021/acs.jpclett.0c00497

Jagger, B., Ojha, A. A., and Amaro, R. (2020). Predicting Ligand Binding Kinetics Using a Markovian Milestoning with Voronoi Tessellations Multiscale Approach. *ChemRxiv* 16 (8), 5348–5357. doi:10.26434/chemrxiv.12275165.v1

Jing, Z., Liu, C., Cheng, S. Y., Qi, R., Walker, B. D., Piquemal, J.-P., et al. (2019). Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.* 48, 371–394. doi:10.1146/annurev-biophys-070317-033349

Juraszek, J., Saladino, G., van Erp, T. S., and Gervasio, F. L. (2013). Efficient Numerical Reconstruction of Protein Folding Kinetics with Partial Path Sampling and Pathlike Variables. *Phys. Rev. Lett.* 110, 108106. doi:10.1103/PhysRevLett.110.108106

Kalbfleisch, J. D., and Lawless, J. F. (1985). The Analysis of Panel Data under a Markov Assumption. *J. Am. Stat. Assoc.* 80, 863–871. doi:10.1080/01621459.1985.10478195

Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* 105, 6474–6487. doi:10.1021/jp003919d

Karplus, M., and McCammon, J. A. (2002). Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* 9, 646–652. doi:10.1299/jsmemag.116.1131_7810.1038/nsb0902-646

Kelly, B. D., and Smith, W. R. (2020). Alchemical Hydration Free-Energy Calculations Using Molecular Dynamics with Explicit Polarization and Induced Polarity Decoupling: An On-The-Fly Polarization Approach. *J. Chem. Theory Comput.* 16, 1146–1161. doi:10.1021/acs.jctc.9b01139

Kieninger, S., and Keller, B. G. (2021). Path Probability Ratios for Langevin Dynamics-Exact and Approximate. *J. Chem. Phys.* 154, 094102. doi:10.1063/5.0038408

Kocer, E., Ko, T. W., and Behler, J. (2022). Neural Network Potentials: A Concise Overview of Methods. *Annu. Rev. Phys. Chem.* 73, 163–186. doi:10.1146/annurev-physchem-082720-034254

Kokh, D. B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H.-P., et al. (2018). Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* 14, 3859–3869. doi:10.1021/acs.jctc.8b00230

Kokh, D. B., Doser, B., Richter, S., Ormersbach, F., Cheng, X., and Wade, R. C. (2020). A Workflow for Exploring Ligand Dissociation from a Macromolecule: Efficient Random Acceleration Molecular Dynamics Simulation and Interaction Fingerprint Analysis of Ligand Trajectories. *J. Chem. Phys.* 153, 125102. doi:10.1063/5.0019088

Kokh, D. B., Kaufmann, T., Kister, B., and Wade, R. C. (2019). Machine Learning Analysis of τRAMD Trajectories to Decipher Molecular Determinants of Drug-Target Residence Times. *Front. Mol. Biosci.* 6, 1–17. doi:10.3389/fmolb.2019.00036

Kudo, S., Nitadori, K., Ina, T., and Imamura, T. (2020). "Prompt Report on Exa-Scale HPL-AI Benchmark," in 2020 IEEE International Conference on Cluster Computing (CLUSTER), Kobe, Japan. St. Louis, Missouri, USA, 418–419. doi:10.1109/CLUSTER49012.2020.00058

Kulik, H. J. (2018). Large-scale QM/MM Free Energy Simulations of Enzyme Catalysis Reveal the Influence of Charge Transfer. *Phys. Chem. Chem. Phys.* 20, 20650–20660. doi:10.1039/c8cp03871f

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* 13, 1011–1021. doi:10.1002/jcc.540130812

Kwarcinski, F. E., Brandvold, K. R., Phadke, S., Beleh, O. M., Johnson, T. M., Meagher, J. L., et al. (2016). Conformation-Selective Analogues of Dasatinib Reveal Insight into Kinase Inhibitor Binding and Selectivity. *ACS Chem. Biol.* 11, 1296–1304.

Laio, A., and Parrinello, M. (2002). Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* 99 (20), 12562–12566. doi:10.1073/pnas.202427399

Lee, K. S. S., Yang, J., Niu, J., Ng, C. J., Wagner, K. M., Dong, H., et al. (2019). Drug-Target Residence Time Affects *In Vivo* Target Occupancy through Multiple Pathways. *ACS Cent. Sci.* 5, 1614–1624. doi:10.1021/acscentsci.9b00770

Lemkul, J. A., Huang, J., Roux, B., and MacKerell, A. D. (2016). An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* 116, 4983–5013. doi:10.1021/acs.chemrev.5b00505

Li, H.-J., Lai, C.-T., Pan, P., Yu, W., Liu, N., Bommineni, G. R., et al. (2014). A Structural and Energetic Model for the Slow-Onset Inhibition of the mycobacterium Tuberculosis ENoyl-ACP Reductase InhA. *ACS Chem. Biol.* 9, 986–993. doi:10.1021/cb400896g

Li, P., and Merz, K. M. (2017). Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* 117, 1564–1686. doi:10.1021/acs.chemrev.6b00440

Liao, R.-Z., and Thiel, W. (2013). Convergence in the QM-Only and QM/MM Modeling of Enzymatic Reactions: A Case Study for Acetylene Hydratase. *J. Comput. Chem.* 34, a–n. doi:10.1002/jcc.23403

Lin, F.-Y., and MacKerell, A. D. (2019). Improved Modeling of Halogenated Ligand-Protein Interactions Using the Drude Polarizable and CHARMM Additive Empirical Force Fields. *J. Chem. Inf. Model.* 59, 215–228. doi:10.1021/acs.jcim.8b00616

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

Liu, C., Li, Y., Han, B.-Y., Gong, L.-D., Lu, L.-N., Yang, Z.-Z., et al. (2017). Development of the ABEEMσπ Polarization Force Field for Base Pairs with Amino Acid Residue Complexes. *J. Chem. Theory Comput.* 13, 2098–2111. doi:10.1021/acs.jctc.6b01206

Lotz, S. D., and Dickson, A. (2018). Unbiased Molecular Dynamics of 11 Min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *J. Am. Chem. Soc.* 140, 618–628. doi:10.1021/jacs.7b08572

Lotz, S. D., and Dickson, A. (2020). Wepy: A Flexible Software Framework for Simulating Rare Events with Weighted Ensemble Resampling. *ACS Omega* 5, 31608–31623. doi:10.1021/acsomega.0c03892

Lu, D., Wang, H., Chen, M., Lin, L., Car, R., E, W., et al. (2021). 86 PFLOPS Deep Potential Molecular Dynamics Simulation of 100 Million Atoms with Ab Initio Accuracy. *Comput. Phys. Commun.* 259, 107624. doi:10.1016/j.cpc.2020.107624

Lüdemann, S. K., Lounnas, V., and Wade, R. C. (2000). How Do Substrates Enter and Products Exit the Buried Active Site of Cytochrome P450cam? 2. Steered Molecular Dynamics and Adiabatic Mapping of Substrate Pathways 1 1Edited by J. Thornton. *J. Mol. Biol.* 303, 813–830. doi:10.1006/jmbi.2000.4155

Luty, B. A., El Amrani, S., and McCammon, J. A. (1993). Simulation of the Bimolecular Reaction between Superoxide and Superoxide Dismutase: Synthesis of the Encounter and Reaction Steps. *J. Am. Chem. Soc.* 115, 11874–11877. doi:10.1021/ja00078a027

Ma, Z., He, J., Qiu, J., Cao, H., Wang, Y., Sun, Z., et al. (2022). "BaGuaLu: Targeting Brain Scale Pretrained Models with over 37 Million Cores," in *ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*.

MacKerell, A. D., Banavali, N., and Foloppe, N. (2000). Development and Current Status of the CHARMM Force Field for Nucleic Acids. *Biopolymers* 56, 257–265. doi:10.1002/1097-0282(2000)56:4<257::aid-bip10029>3.0.co;2-w

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, 3586–3616. doi:10.1021/jp973084f

MacKerell, A. D., Jr., Feig, M., and Brooks, C. L., III (2004). Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* 25, 1400–1415. doi:10.1002/jcc.20065

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. Theory Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Mandelli, D., Hirshberg, B., and Parrinello, M. (2020). Metadynamics of Paths. *Phys. Rev. Lett.* 125, 26001. doi:10.1103/PhysRevLett.125.026001

Mardt, A., Pasquali, L., Wu, H., and Noé, F. (2018). VAMPnets for Deep Learning of Molecular Kinetics. *Nat. Commun.* 9, 1–14. doi:10.1038/s41467-017-02388-1

Maximova, E., Postnikov, E. B., Lavrova, A. I., Farafonov, V., and Nerukh, D. (2021). Protein-Ligand Dissociation Rate Constant from All-Atom Simulation. *J. Phys. Chem. Lett.* 12, 10631–10636. doi:10.1021/acs.jpclett.1c02952

Mazzorana, M., Shotton, E. J., and Hall, D. R. (2020). A Comprehensive Approach to X-Ray Crystallography for Drug Discovery at a Synchrotron Facility - the Example of Diamond Light Source. *Drug Discov. Today Technol.* 37, 83–92. doi:10.1016/j.ddtec.2020.10.003

Miao, Y. (2018). Acceleration of Biomolecular Kinetics in Gaussian Accelerated Molecular Dynamics. *J. Chem. Phys.* 149, 072308. doi:10.1063/1.5024217

Miao, Y., Bhattarai, A., and Wang, J. (2020). Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): Characterization of Ligand Binding Thermodynamics and Kinetics. *J. Chem. Theory Comput.* 16, 5526–5547. doi:10.1021/acs.jctc.0c00395

Mironenko, A., Zachariae, U., de Groot, B. L., and Kopec, W. (2021). The Persistent Question of Potassium Channel Permeation Mechanisms. *J. Mol. Biol.* 433, 167002. doi:10.1016/j.jmb.2021.167002

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, e1006180–20. doi:10.1371/journal.pcbi.1006180

Morando, M. A., Saladino, G., D'Amelio, N., Pucheta-Martinez, E., Lovera, S., Lelli, M., et al. (2016). Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the C-Src Kinase. *Sci. Rep.* 6, 1–9. doi:10.1038/srep24439

Moroni, D., Bolhuis, P. G., and van Erp, T. S. (2004). Rate Constants for Diffusive Processes by Partial Path Sampling. *J. Chem. Phys.* 120, 4055–4065. doi:10.1063/1.1644537

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

Olsen, J. M. H., Bolnykh, V., Meloni, S., Ippoliti, E., Bircher, M. P., Carloni, P., et al. (2019). MiMiC: A Novel Framework for Multiscale Modeling in Computational Chemistry. *J. Chem. Theory Comput.* 15, 3810–3823. doi:10.1021/acs.jctc.9b00093

Páll, S., Abraham, M. J., Kutzner, C., Hess, B., and Lindahl, E. (2015). Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS BT *Solving Software Challenges for Exascale*. in, eds. S. Markidis and E. Laure (Cham: Springer International Publishing), 3–27.doi:10.1007/978-3-319-15976-8_1

Paci, E., and Karplus, M. (2000). Unfolding Proteins by External Forces and Temperature: The Importance of Topology and Energetics. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6521–6526. doi:10.1073/pnas.100124597

Pan, A. C., Borhani, D. W., Dror, R. O., and Shaw, D. E. (2013). Molecular Determinants of Drug-Receptor Binding Kinetics. *Drug Discov. Today* 18, 667–673. doi:10.1016/j.drudis.2013.02.007

Pan, A. C., Jacobson, D., Yatsenko, K., Sritharan, D., Weinreich, T. M., and Shaw, D. E. (2019). Atomic-level Characterization of Protein-Protein Association. *Proc. Natl. Acad. Sci. U.S.A.* 116, 4244–4249. doi:10.1073/pnas.1815431116

Pan, A. C., Xu, H., Palpant, T., and Shaw, D. E. (2017). Quantitative Characterization of the Binding and Unbinding of Millimolar Drug

Fragments with Molecular Dynamics Simulations. *J. Chem. Theory Comput.* 13, 3372–3377. doi:10.1021/acs.jctc.7b00172

Parks, C. D., Gaieb, Z., Chiu, M., Yang, H., Shao, C., Walters, W. P., et al. (2020). D3R Grand Challenge 4: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided. Mol. Des.* 34, 99–119. doi:10.1007/s10822-020-00289-y

Patel, S., and Brooks, C. L., 3rd (2004). CHARMM Fluctuating Charge Force Field for Proteins: I Parameterization and Application to Bulk Organic Liquid Simulations. *J. Comput. Chem.* 25, 1–16. doi:10.1002/jcc.10355

Patel, S., and Brooks, C. L., III (2003). CHARMM Fluctuating Charge Force Field for Proteins: I Parameterization and Application to Bulk Organic Liquid Simulations. *J. Comput. Chem.* 25, 1–16. doi:10.1002/jcc.10355

Paul, F., Wehmeyer, C., Abualrous, E. T., Wu, H., Crabtree, M. D., Schöneberg, J., et al. (2017). Protein-peptide Association Kinetics beyond the Seconds Timescale from Atomistic Simulations. *Nat. Commun.* 8, 1–9. doi:10.1038/s41467-017-01163-6

Piniello, B., Lira-Navarrete, E., Takeuchi, H., Takeuchi, M., Haltiwanger, R. S., Hurtado-Guerrero, R., et al. (2021). Asparagine Tautomerization in Glycosyltransferase Catalysis. The Molecular Mechanism of Protein O-Fucosyltransferase 1. *ACS Catal.* 11, 9926–9932. doi:10.1021/acscatal.1c01785

Piquemal, J.-P., Chevreau, H., and Gresh, N. (2007). Toward a Separate Reproduction of the Contributions to the Hartree−Fock and DFT Intermolecular Interaction Energies by Polarizable Molecular Mechanics with the SIBFA Potential. *J. Chem. Theory Comput.* 3, 824–837. doi:10.1021/ct7000182

Plattner, N., and Noé, F. (2015). Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* 6. doi:10.1038/ncomms8653

Pollard, T. D. (2010). A Guide to Simple and Informative Binding Assays. *MBoC* 21, 4061–4067. doi:10.1091/mbc.E10-08-0683

Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., et al. (2010). Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* 114, 2549–2564. doi:10.1021/jp910674d

Potterton, A., Husseini, F. S., Southey, M. W. Y., Bodkin, M. J., Heifetz, A., Coveney, P. V., et al. (2019). Ensemble-Based Steered Molecular Dynamics Predicts Relative Residence Time of A2A Receptor Binders. *J. Chem. Theory Comput.* 15, 3316–3330. doi:10.1021/acs.jctc.8b01270

Pratt, L. R. (1986). A Statistical Method for Identifying Transition States in High Dimensional Problems. *J. Chem. Phys.* 85, 5045–5048. doi:10.1063/1.451695

Proudfoot, A., Bussiere, D. E., and Lingel, A. (2017). High-Confidence Protein-Ligand Complex Modeling by NMR-Guided Docking Enables Early Hit Optimization. *J. Am. Chem. Soc.* 139, 17824–17833. doi:10.1021/jacs.7b07171

Qiu, Y., Smith, D. G. A., Boothroyd, S., Jang, H., Hahn, D. F., Wagner, J., et al. (2021). Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field. *J. Chem. Theory Comput.* 17, 6262–6280. doi:10.1021/acs.jctc.1c00571

Ray, D., and Andricioaei, I. (2020). Weighted Ensemble Milestoning (WEM): A Combined Approach for Rare Event Simulations. *J. Chem. Phys.* 152, 234114. doi:10.1063/5.0008028

Ray, D., Stone, S. E., and Andricioaei, I. (2022). Markovian Weighted Ensemble Milestoning (M-WEM): Long-Time Kinetics from Short Trajectories. *J. Chem. Theory Comput.* 18, 79–95. doi:10.1021/acs.jctc.1c00803

Regan, J., Pargellis, C. A., Cirillo, P. F., Gilmore, T., Hickey, E. R., Peet, G. W., et al. (2003). The Kinetics of Binding to p38MAP Kinase by Analogues of BIRB 796. *Bioorg. Med. Chem. Lett.* 13, 3101–3104. doi:10.1016/S0960-894X(03)00656-5

Robustelli, P., Ibanez-de-Opakua, A., Campbell-Bezat, C., Giordanetto, F., Becker, S., Zweckstetter, M., et al. (2022). Molecular Basis of Small-Molecule Binding to α-Synuclein. *J. Am. Chem. Soc.* 144, 2501–2510. doi:10.1021/jacs.1c07591

Rocklin, G. J., Boyce, S. E., Fischer, M., Fish, I., Mobley, D. L., Shoichet, B. K., et al. (2013). Blind Prediction of Charged Ligand Binding Affinities in a Model Binding Site. *J. Mol. Biol.* 425, 4569–4583. doi:10.1016/j.jmb.2013.07.030

Roston, D., Demapan, D., and Cui, Q. (2016). Leaving Group Ability Observably Affects Transition State Structure in a Single Enzyme Active Site. *J. Am. Chem. Soc.* 138, 7386–7394. doi:10.1021/jacs.6b03156

Rufa, D. A., Bruce Macdonald, H. E., Fass, J., Wieder, M., Grinaway, P. B., Roitberg, A. E., et al. (2020). Towards Chemical Accuracy for Alchemical Free Energy Calculations with Hybrid Physics-Based Machine Learning/Molecular Mechanics Potentials. *bioRxiv*, 1–21. doi:10.1101/2020.07.29.227959

Salvalaglio, M., Tiwary, P., and Parrinello, M. (2014). Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. *J. Chem. Theory Comput.* 10, 1420–1425. doi:10.1021/ct500040r

Schäfer, T. M., and Settanni, G. (2020). Data Reweighting in Metadynamics Simulations. *J. Chem. Theory Comput.* 16, 2042–2052. doi:10.1021/acs.jctc.9b00867

Schiebel, J., Gaspari, R., Wulsdorf, T., Ngo, K., Sohn, C., Schrader, T. E., et al. (2018). Intriguing Role of Water in Protein-Ligand Binding Studied by Neutron Crystallography on Trypsin Complexes. *Nat. Commun.* 9, 3559. doi:10.1038/s41467-018-05769-2

Schindler, C. E. M., Baumann, H., Blum, A., Böse, D., Buchstaller, H.-P., Burgdorf, L., et al. (2020). Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *J. Chem. Inf. Model.* 60, 5457–5474. doi:10.1021/acs.jcim.0c00900

Schlitter, J., Engels, M., and Krüger, P. (1994). Targeted Molecular Dynamics: A New Approach for Searching Pathways of Conformational Transitions. *J. Mol. Graph.* 12, 84–89. doi:10.1016/0263-7855(94)80072-3

Schmidtke, P., Luque, F. J., Murray, J. B., and Barril, X. (2011). Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics: Application in Drug Design. *J. Am. Chem. Soc.* 133, 18903–18910. doi:10.1021/ja207494u

Schneider, D. (2022). The Exascale Era Is upon Us: The Frontier Supercomputer May Be the First to Reach 1,000,000,000,000,000,000 Operations Per Second. *IEEE Spectr.* 59, 34–35. doi:10.1109/MSPEC.2022.9676353

Schramm, V. L. (2015). Transition States and Transition State Analogue Interactions with Enzymes. *Acc. Chem. Res.* 48, 1032–1039. doi:10.1021/acs.accounts.5b00002

Schramm, V. L. (2013). Transition States, Analogues, and Drug Development. *ACS Chem. Biol.* 8, 71–81. doi:10.1021/cb300631k

Shan, Y., Seeliger, M. A., Eastwood, M. P., Frank, F., Xu, H., Jensen, M., et al. (2009). A Conserved Protonation-Dependent Switch Controls Drug Binding in the Abl Kinase. *Proc. Natl. Acad. Sci.* 106, 139–144. doi:10.1073/pnas.0811223106

Shaw, D. E., Adams, P. J., Azaria, A., Bank, J. A., Batson, B., Bell, A., et al. (2021). "Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch," in SC '21: The International Conference for High Performance Computing, Networking, Storage and Analysis. New York, NY, United States: Association for Computing Machinery. doi:10.1145/3458817.3487397

Shen, L., and Yang, W. (2018). Molecular Dynamics Simulations with Quantum Mechanics/Molecular Mechanics and Adaptive Neural Networks. *J. Chem. Theory Comput.* 14, 1442–1455. doi:10.1021/acs.jctc.7b01195

Shirts, M. R., and Chodera, J. D. (2008). Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* 129, 124105. doi:10.1063/1.2978177

Singh, R., Wiseman, B., Deemagarn, T., Jha, V., Switala, J., and Loewen, P. C. (2008). Comparative Study of Catalase-Peroxidases (KatGs). *Archives Biochem. Biophysics* 471, 207–214. doi:10.1016/j.abb.2007.12.008

Singhal, N., Snow, C. D., and Pande, V. S. (2004). Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin. *J. Chem. Phys.* 121, 415–425. doi:10.1063/1.1738647

Sinko, W., Miao, Y., de Oliveira, C. A. F., and McCammon, J. A. (2013). Population Based Reweighting of Scaled Molecular Dynamics. *J. Phys. Chem. B* 117, 12759–12768. doi:10.1021/jp401587e

Sittel, F., and Stock, G. (2018). Perspective: Identification of Collective Variables and Metastable States of Protein Dynamics. *J. Chem. Phys.* 149, 150901. doi:10.1063/1.5049637

Spiriti, J., and Wong, C. F. (2021). Qualitative Prediction of Ligand Dissociation Kinetics from Focal Adhesion Kinase Using Steered Molecular Dynamics. *Life* 11, 74–19. doi:10.3390/life11020074

Stelzl, L. S., Kells, A., Rosta, E., and Hummer, G. (2017). Dynamic Histogram Analysis to Determine Free Energies and Rates from Biased Simulations. *J. Chem. Theory Comput.* 13, 6328–6342. doi:10.1021/acs.jctc.7b00373

Stocker, S., Csányi, G., Reuter, K., and Margraf, J. T. (2020). Machine Learning in Chemical Reaction Space. *Nat. Commun.* 11, 1–11. doi:10.1038/s41467-020-19267-x

Suárez, E., Wiewiora, R. P., Wehmeyer, C., Noé, F., Chodera, J. D., and Zuckerman, D. M. (2021). What Markov State Models Can and Cannot Do: Correlation versus Path-Based Observables in Protein-Folding Models. *J. Chem. Theory Comput.* 17, 3119–3133. doi:10.1021/acs.jctc.0c01154

Svensson, F., Engen, K., Lundbäck, T., Larhed, M., and Sköld, C. (2015). Virtual Screening for Transition State Analogue Inhibitors of IRAP Based on Quantum Mechanically Derived Reaction Coordinates. *J. Chem. Inf. Model.* 55, 1984–1993. doi:10.1021/acs.jcim.5b00359

Tang, Z., and Chang, C.-e. A. (2018). Binding Thermodynamics and Kinetics Calculations Using Chemical Host and Guest: A Comprehensive Picture of Molecular Recognition. *J. Chem. Theory Comput.* 14, 303–318. doi:10.1021/acs.jctc.7b00899

Teo, I., Mayne, C. G., Schulten, K., and Lelièvre, T. (2016). Adaptive Multilevel Splitting Method for Molecular Dynamics Calculation of Benzamidine-Trypsin Dissociation Time. *J. Chem. Theory Comput.* 12, 2983–2989. doi:10.1021/acs.jctc.6b00277

Tiwary, P., Mondal, J., and Berne, B. J. (2017). How and when Does an Anticancer Drug Leave its Binding Site? *Sci. Adv.* 3, e1700014. doi:10.1126/sciadv.1700014

Tiwary, P., and Parrinello, M. (2015). A Time-independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B* 119, 736–742. doi:10.1021/jp504920s

Tiwary, P., and Parrinello, M. (2013). From Metadynamics to Dynamics. *Phys. Rev. Lett.* 111, 230602. doi:10.1103/PhysRevLett.111.230602

Truhlar, D. G., Garrett, B. C., and Klippenstein, S. J. (1996). Current Status of Transition-State Theory. *J. Phys. Chem.* 100, 12771–12800. doi:10.1021/jp953748q

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., et al. (2021). Machine Learning Force Fields. *Chem. Rev.* 121, 10142–10186. doi:10.1021/acs.chemrev.0c01111

Van Der Velden, W. J. C., Heitman, L. H., and Rosenkilde, M. M. (2020). Perspective: Implications of Ligand-Receptor Binding Kinetics for Therapeutic Targeting of G Protein-Coupled Receptors. *ACS Pharmacol. Transl. Sci.* 3, 179–189. doi:10.1021/acsptsci.0c00012

Van Erp, T. S., Moroni, D., and Bolhuis, P. G. (2003). A Novel Path Sampling Method for the Calculation of Rate Constants. *J. Chem. Phys.* 118, 7762–7774. doi:10.1063/1.1562614

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., et al. (2011). CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* 31, 671–690. doi:10.1002/jcc.21367.CHARMM

Vauquelin, G., Bostoen, S., Vanderheyden, P., and Seeman, P. (2012). Clozapine, Atypical Antipsychotics, and the Benefits of Fast-Off D2 Dopamine Receptor Antagonism. *Schmiedeb. Arch. Pharmacol.* 385, 337–372. doi:10.1007/s00210-012-0734-2

Vitalini, F., Mey, A. S. J. S., Noé, F., and Keller, B. G. (2015). Dynamic Properties of Force Fields. *J. Chem. Phys.* 142, 084101. doi:10.1063/1.4909549

Votapka, L. W., Jagger, B. R., Heyneman, A. L., and Amaro, R. E. (2017). SEEKR: Simulation Enabled Estimation of Kinetic Rates, A Computational Tool to Estimate Molecular Kinetics and its Application to Trypsin-Benzamidine Binding. *J. Phys. Chem. B* 121, 3597–3606. SEEKR. doi:10.1021/acs.jpcb.6b09388

Voter, A. F., and Doll, J. D. (1985). Dynamical Corrections to Transition State Theory for Multistate Systems: Surface -Self-Diffusion in the Rare-Event Regime. *J. Chem. Phys.* 82, 80–92. doi:10.1063/1.448739

Voter, A. F. (1997). Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Phys. Rev. Lett.* 78, 3908–3911. doi:10.1103/PhysRevLett.78.3908

Wan, H., and Voelz, V. A. (2020). Adaptive Markov State Model Estimation Using Short Reseeding Trajectories. *J. Chem. Phys.* 152, 024103. doi:10.1063/1.5142457

Wang, F., and Landau, D. P. (2001). Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.* 86, 2050–2053. doi:10.1103/PhysRevLett.86.2050

Wang, J., and Miao, Y. (2020). Peptide Gaussian Accelerated Molecular Dynamics (Pep-GaMD): Enhanced Sampling and Free Energy and Kinetics Calculations of Peptide Binding. *J. Chem. Phys.* 153, 154109. doi:10.1063/5.0021399

Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006). Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* 25, 247–260. doi:10.1016/j.jmgm.2005.12.005

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, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., et al. (2015). Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* 137, 2695–2703. doi:10.1021/ja512751q

Wang, Y., Martins, J. M., and Lindorff-Larsen, K. (2017). Biomolecular Conformational Changes and Ligand Binding: from Kinetics to Thermodynamics. *Chem. Sci.* 8, 6466–6473. doi:10.1039/c7sc01627a

Wang, Y., Valsson, O., Tiwary, P., Parrinello, M., and Lindorff-Larsen, K. (2018). Frequency Adaptive Metadynamics for the Calculation of Rare-Event Kinetics. *J. Chem. Phys.* 149, 072309. doi:10.1063/1.5024679

Wang, Z. X., Zhang, W., Wu, C., Lei, H., Cieplak, P., and Duan, Y. (2006). Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force Field for Simulations of Proteins and Peptides. *J. Comput. Chem.* 27, 781–790. doi:10.1002/jcc.20386

Wolf, S., Amaral, M., Lowinski, M., Vallée, F., Musil, D., Güldenhaupt, J., et al. (2019). Estimation of Protein-Ligand Unbinding Kinetics Using Non-equilibrium Targeted Molecular Dynamics Simulations. *J. Chem. Inf. Model.* 59, 5135–5147. doi:10.1021/acs.jcim.9b00592

Wolf, S., Lickert, B., Bray, S., and Stock, G. (2020). Multisecond Ligand Dissociation Dynamics from Atomistic Simulations. *Nat. Commun.* 11, 1–8. doi:10.1038/s41467-020-16655-1

Wolf, S., and Stock, G. (2018). Targeted Molecular Dynamics Calculations of Free Energy Profiles Using a Nonequilibrium Friction Correction. *J. Chem. Theory Comput.* 14, 6175–6182. doi:10.1021/acs.jctc.8b00835

Woods, C. J., Essex, J. W., and King, M. A. (2003). Enhanced Configurational Sampling in Binding Free-Energy Calculations. *J. Phys. Chem. B* 107, 13711–13718. doi:10.1021/jp036162+

Woods, C. J., Manby, F. R., and Mulholland, A. J. (2008). An Efficient Method for the Calculation of Quantum Mechanics/molecular Mechanics Free Energies. *J. Chem. Phys.* 128, 014109. doi:10.1063/1.2805379

Wu, H., Paul, F., Wehmeyer, C., and Noé, F. (2016). Multiensemble Markov Models of Molecular Thermodynamics and Kinetics. *Proc. Natl. Acad. Sci. U.S.A.* 113, E3221–E3230. doi:10.1073/pnas.1525092113

Xue, Y., Yuwen, T., Zhu, F., and Skrynnikov, N. R. (2014). Role of Electrostatic Interactions in Binding of Peptides and Intrinsically Disordered Proteins to Their Folded Targets. 1. NMR and MD Characterization of the Complex between the C-Crk N-SH3 Domain and the Peptide Sos. *Biochemistry* 53, 6473–6495. doi:10.1021/bi500904f

Yang, M., Bonati, L., Polino, D., and Parrinello, M. (2022). Using Metadynamics to Build Neural Network Potentials for Reactive Events: the Case of Urea Decomposition in Water. *Catal. Today* 387, 143–149. doi:10.1016/j.cattod.2021.03.018

Yue, S., Muniz, M. C., Calegari Andrade, M. F., Zhang, L., Car, R., and Panagiotopoulos, A. Z. (2021). When Do Short-Range Atomistic Machine-Learning Models Fall Short? *J. Chem. Phys.* 154, 034111. doi:10.1063/5.0031215

Yue, Z., Wang, Z., and Voth, G. A. (2022). Ion Permeation, Selectivity, and Electronic Polarization in Fluoride Channels. *Biophysical J.* 121, 1336–1347. doi:10.1016/j.bpj.2022.02.019

Zhang, B. W., Jasnow, D., and Zuckerman, D. M. (2010). The "Weighted Ensemble" Path Sampling Method Is Statistically Exact for a Broad Class of Stochastic Processes and Binning Procedures. *J. Chem. Phys.* 132, 054107. doi:10.1063/1.3306345

Zhao, L., Ciallella, H. L., Aleksunes, L. M., and Zhu, H. (2020). Advancing Computer-Aided Drug Discovery (CADD) by Big Data and Data-Driven Machine Learning Modeling. *Drug Discov. Today* 25, 1624–1638. doi:10.1016/j.drudis.2020.07.005

Zuckerman, D. M., and Chong, L. T. (2017). Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu. Rev. Biophys.* 46, 43–57. Weighted. doi:10.1146/annurev-biophys-070816-033834

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 15 June 2022
doi: 10.3389/fmolb.2022.904445

Statistical Analysis of Protein-Ligand Interaction Patterns in Nuclear Receptor RORγ

*Bill Pham[1], Ziju Cheng[1], Daniel Lopez[1], Richard J. Lindsay[2], David Foutch[2], Rily T. Majors[1] and Tongye Shen[1]**

[1]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN, United States, [2]UT-ORNL Graduate School of Genome Science and Technology, Knoxville, TN, United States

The receptor RORγ belongs to the nuclear receptor superfamily that senses small signaling molecules and regulates at the gene transcription level. Since RORγ has a high basal activity and plays an important role in immune responses, inhibitors targeting this receptor have been a focus for many studies. The receptor-ligand interaction is complex, and often subtle differences in ligand structure can determine its role as an inverse agonist or an agonist. We examined more than 130 existing RORγ crystal structures that have the same receptor complexed with different ligands. We reported the features of receptor-ligand interaction patterns and the differences between agonist and inverse agonist binding. Specific changes in the contact interaction map are identified to distinguish active and inactive conformations. Further statistical analysis of the contact interaction patterns using principal component analysis reveals a dominant mode which separates allosteric binding vs. canonical binding and a second mode which may indicate active vs. inactive structures. We also studied the nature of constitutive activity by performing a 100-ns computer simulation of apo RORγ. Using constitutively active nuclear receptor CAR as a comparison, we identified a group of conserved contacts that have similar contact strength between the two receptors. These conserved contact interactions, especially a couple key contacts in H11–H12 interaction, can be considered essential to the constitutive activity of RORγ. These protein-ligand and internal protein contact interactions can be useful in the development of new drugs that direct receptor activity.

Keywords: protein-ligand interaction, statistical analysis, nuclear receptor, constitutive activity, inverse agonist

OPEN ACCESS

Edited by:
Weiliang Zhu,
Shanghai Institute of Materia Medica (CAS), China

Reviewed by:
Ho Leung Ng,
Atomwise Inc., United States
Jian Zhang,
Shanghai Jiao Tong University, China

****Correspondence:***
Tongye Shen
tshen@utk.edu

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *25 March 2022*
Accepted: *16 May 2022*
Published: *15 June 2022*

Citation:
Pham B, Cheng Z, Lopez D, Lindsay RJ, Foutch D, Majors RT and Shen T (2022) Statistical Analysis of Protein-Ligand Interaction Patterns in Nuclear Receptor RORγ. Front. Mol. Biosci. 9:904445. doi: 10.3389/fmolb.2022.904445

INTRODUCTION

The nuclear receptor (NR) superfamily is a group of important transcription factors that detect the presence of specific compounds using their ligand-binding domain (LBD) and respond by modulating gene transcription, which is directed through interaction between specific DNA response elements and the DNA-binding domain (DBD) and interaction between co-activators and the LBD (Helsen et al., 2012; Weikum et al., 2018). Well-known examples of the NR superfamily include estrogen receptor, androgen receptor, glucocorticoid receptor, vitamin D receptor, peroxisome proliferator-activated receptor, retinoid receptor, thyroid hormone receptors, and many others. While the structures of the DBD and the LBD are highly conserved, there is a highly varied and largely unstructured N-terminal domain (NTD) that also plays an important role in the function of these transcription factors (Kumar and Thompson, 2003; Simons et al., 2014).

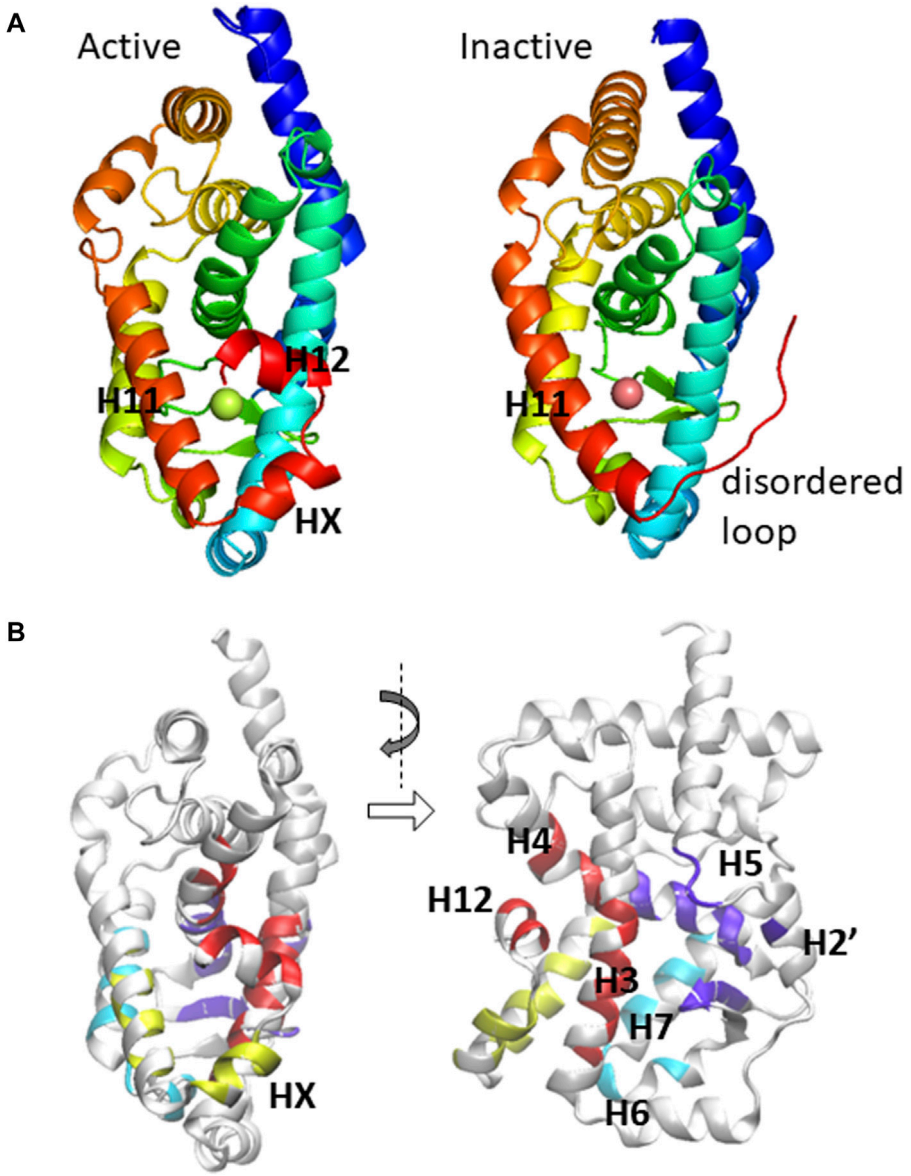


FIGURE 1 | (A) The ligand-binding domain of RORγ is displayed in active (PDB: 3L0L) and inactive (PDB: 4QM0) conformations. The important secondary structural elements that undergo conformational change are located at the C-terminal region and are explicitly labeled. **(B)** The residues that potentially form direct contact with ligands are shown in a canonical front view and a side view.

One of the important NRs is called RAR-related orphan receptor (ROR), since initially ROR was discovered as an orphan receptor that is related to retinoid acid receptor (RAR) (Solt and Burris, 2012; Zhang et al., 2015). ROR was found to play important roles in regulating immune responses and circadian rhythm (Takeda et al., 2012; Cook et al., 2015). Furthermore, ROR was also found to be one of the few NRs that are constitutively active, meaning the receptor exhibits high basal activity (active without ligand). Since a hyperactive ROR can be tied to autoimmune diseases such as multiple sclerosis and rheumatoid arthritis, identifying potent inverse agonists to regulate ROR is of interest (Zhang et al., 2015). One essential question arises as to how one can efficiently obtain details of the protein-ligand interaction and predict how ligands affect the protein conformation, i.e., turn on or off ROR activity. As detailed below, this remains a puzzle as the ligand-protein pairing for NRs is highly sensitive.

Various structural biology and chemical biology studies have focused on ROR-ligand interactions. Among the three subtypes of ROR (RORα, RORβ, and RORγ), RORγ appears to be very important with the most structural data available, and thus we focus on examining the LBD of the γ-subtype in this study. Indeed, there have been more than 100 X-ray crystallography structures reported in the Protein Data Bank

(PDB), all of which are in the monomer form having the identical protein sequence while the only differences are the unique identity of ligand(s) that forms a complex with RORγ. In many previous reports, a set of similar ligands was used to probe the cellular activity and/or biophysical properties of ROR induced by ligand binding. Similar to other NRs, binding of an agonist to the LBD leads to a conformational change that facilitates a more favorable interaction with the co-activators at the activation function 2 (AF2) region of the LBD (Weikum et al., 2018). Alternatively, when an inverse agonist binds to the LBD, co-activator binding becomes inhibited due to (at least in part) the structural changes in helices H10, H11, and/or H12 (Li et al., 2017; Noguchi et al., 2017; Gong et al., 2018). Two mechanisms of inverse agonism that have been observed include: 1) a disorder of helix H12, which would otherwise form part of the binding pocket, reduces available agonist interaction sites and 2) the formation of a "kink" between helices H10 and H11 consequently obstructs the co-activator binding site formed by helix H12, as shown in **Figure 1**.

Often, researchers found that whether specific ligands can turn ROR on or off is quite sensitive. For example, several studies showed that a slight modification of a known agonist or inverse agonist can switch its properties. Specifically, there were reported pairs of ligands (obtained from tertiary sulfonamides, biaryl amides, tertiary amines, benzoxazinones, and other families) that bind at the same binding site; however, the shorter of the two is an agonist while the longer ligand is an inverse agonist or vice versa, exemplified by PDB pairs: 4WPF/4WQP, 5IZ0/5IXK, 6NWU/6NWS, and 6R7K/6R7J (agonist/inverse agonist-bound structures) (Yang et al., 2014; René et al., 2015; Wang et al., 2015a; Marcotte et al., 2016; Gong et al., 2018; Wang et al., 2018; Strutzenberg et al., 2019; von Berg et al., 2019). Using 5IZ0/5IXK as an example, M358 was reported to interact with an inverse agonist ligand Bio399 (synthetic benzoxazinone) and consequently, it affects residue F506 and changes the protein conformation into an inactive form (PDB: 5IXK). In contrast, a similar ligand Bio592 has nearly identical contact interactions with the rest of the binding pocket while lacking the contact with M358, which in turn keeps RORγ in an active conformation (PDB: 5IZ0) (Marcotte et al., 2016). Meanwhile, another intriguing study found that lengthening or shortening modifications of a specific agonist (biphenyl-ethylsulfonyl-phenyl-acetamide) leads to inverse agonism (Wang et al., 2018). As different studies reported different local trigger spots for inverse agonists, one may want to consolidate ROR-ligand interactions and rethink the canonical view that the ligand-directed action comes from a fixed chemical group of the compound with a specific residue of the binding pocket. Instead, the ligands examined in these studies are diverse and distinct from one study to another. It appears that a specific chemical group is not enough to determine the effect of a molecule, and yet the ligand identity clearly affects the interactions with the binding pocket and subsequently the protein activities. For nuclear receptors, there is a challenge on how to connect the ligand identity with directed structural changes and subsequent activities.

We think that a statistical analysis of extensive structural information where one collectively examines protein-ligand contact interaction patterns may provide insight into this challenge. For this work, we studied 136 RORγ structures: 132 with one ligand (or ligand fragments) bound (100+ distinct ligands) from X-ray crystallography experiments (Jin et al., 2010; Fujita-Sato et al., 2011; Fauber et al., 2013; Fauber et al., 2014; van Niel et al., 2014; Yang et al., 2014; Chao et al., 2015; Muegge et al., 2015; René et al., 2015; Santori et al., 2015; Scheepstra et al., 2015; Wang et al., 2015b; Wang et al., 2015c; Enyedy et al., 2016; Hirata et al., 2016; Hintermann et al., 2016; Marcotte et al., 2016; Olsson et al., 2016; Ouvry et al., 2016; Xue et al., 2016; Kallen et al., 2017; Kummer et al., 2017; Li et al., 2017; Noguchi et al., 2017; Carcache et al., 2018; Fukase et al., 2018; Gege et al., 2018; Gong et al., 2018; Kono et al., 2018; Narjes et al., 2018; Noguchi et al., 2018; Sasaki et al., 2018; Schnute et al., 2018; Shirai et al., 2018; Wang et al., 2018; Amaudrut et al., 2019; Duan et al., 2019; Hoegenauer et al., 2019; Kotoku et al., 2019; Lu et al., 2019; Marcoux et al., 2019; Sato et al., 2019; Strutzenberg et al., 2019; Tanis et al., 2019; von Berg et al., 2019; Yukawa et al., 2019; Zhang et al., 2019; Cherney et al., 2020; Duan et al., 2020; Gege et al., 2020; Harikrishnan et al., 2020; Jiang et al., 2020; Liu et al., 2020; Meijer et al., 2020; Nakajima et al., 2020; Shi et al., 2020; Vries et al., 2020; Zhang et al., 2020; Lugar et al., 2021; Meijer et al., 2021; Nakajima et al., 2021; Ruan et al., 2021; Yang et al., 2021). Additionally, there was also a report of 12 structures (with two ligands bound) (de Vries et al., 2021). The full list of PDBs and their associated properties can be found in the **Supplementary Data File S1**. We did not include the double-liganded structures in the analysis since almost all the ligands in those structures have been crystalized with RORγ previously, and thus these ligand-protein contacts have already been included in our analysis. There might be new RORγ structures deposited in the PDB since the time of our structural bioinformatics research, and any newly reported RORγ structures after that time would not be included in the current analysis. However, we expect that the results of our statistical analyses and the conclusions drawn should still hold.

The current work has two main focuses. The first one is the statistical analyses of the protein-ligand interactions, obtained from previous experimental studies in which each examined ligand or multiple ligands interact with the binding pocket. The comparison across all ligands will provide a more comprehensive picture of the molecular interactions that differentiate between agonists and inverse agonists, and potentially illustrate the mechanism (structural change) by which each ligand imposes its effect. The second focus is the nature of NR constitutive activity. The RORs have a high basal activity and thus they are considered to be active without any ligands. Such constitutive activity of receptors, including many prominent examples from the NR and GPCR families, are difficult to study experimentally at times. Often, receptors, including RORγ, do not have structures resolved experimentally in the absence of ligand. Computational study of an apo conformation may provide clues on how they function (Pham et al., 2019a; Rosenberg et al., 2019). Within the nuclear receptor superfamily, only a few wild-type receptors display

constitutive activity. Constitutive androstane receptor (CAR) is also deemed to be constitutively active as suggested by its name (Dussault et al., 2002; Xu et al., 2004). The CAR protein functions as a xenobiotic sensor, which detects foreign substances such as drug molecules and metabolizes them primarily in the liver (Xie et al., 2003; Wang et al., 2012). Additionally, a couple other NRs were also suggested to have a high basal activity, such as ERR and SF-1/LRH-1 (Schimmer and White, 2010; Huss et al., 2015). Often, they have a relatively small binding pocket. A previous computational study performed on CAR has shown some essential protein contacts contributing to the constitutive stability of the unliganded CAR (Pham et al., 2019a). By comparing CAR with RORγ, we may gain insights into the important protein interactions that help facilitate the constitutive activity of nuclear receptors in general.

METHODS AND SYSTEMS

Crystal Structure Ensemble of RORs With Various Ligands Bound

The statistical analysis includes a total of 136 X-ray crystal structures. Only four of the structures (PDB: 5K38, 5VB3, 5X8U, and 5X8W) are absent of a ligand and the other 132 structures contain a single ligand at the ligand-binding pocket. The binding pocket mentioned refers to either a canonical, largely enclosed ligand-binding pocket or an adjacent, more exposed allosteric binding site. We did not include the 12 structures with double ligands (one each at canonical and allosteric binding sites).

For the protein component of the complex, all 136 structures contain a single chain of LBD of RORγ. Note that we also include RORγt, an isoform of RORγ that is selectively expressed in the thymus. Although the sequence of RORγt is 21-residues shorter than RORγ at the N-terminal domain (NTD) due to alternative splicing, both RORγt and RORγ have an identical LBD. Among 136 structures, only a few of them (PDBs: 4NB6, 6O98, 6XFV, and 7JH2) were reported using RORγt indices for their residues while the rest used RORγ. Six of them are from gibbon ROR, which only contains a double substitution (K469A/R473A) from human UNP P51449. Another 34 PDBs are single point mutants at C455 (mostly C455S, occasionally C–H or C–E mutations were reported). By visual inspection, these mutations or substitutions are far from the ligand binding pocket, e.g., C455 is at helix H9, thus none of them are directly involved with the protein-ligand contact interaction. Therefore, we do not treat them separately from the wild-type RORγt.

For the 132 structures containing only one ligand, there are a total of 125 distinct ligands. Notably, four PDB pairs (4YPQ and 5C4O, 5K3M and 5X8S, 5NI5 and 5NU1, 5APJ and 5APK) share the same ligands and four additional PDBs (4NB6, 5EJV, 5K3L, and 5NTQ) all share the same synthetic ligand T0901317 (also an agonist for LXR). Even though these pairs and groups may share the same ligand, the protein conformations are not necessarily the same. For example, despite sharing the same inverse agonist, 5APJ is active (due to a fused coactivator) while 5APK is in an inactive conformation (Olsson et al., 2016). The structures of 4YPQ and 5C4O are in different space groups (Scheepstra et al., 2015), whereas 5NI5 and 5NU1 are bound to different coactivators. A unique case, 5G44, contains three ligand fragments in the binding pocket as it was obtained from a cosolvent engineering study (Xue et al., 2016). We treated this three-ligand "cocktail" as a single ligand. One of the 116 PDBs, 6W9J, was removed from the structure database and replaced by 6XAE after we started our study. Since 6W9J has an identical ligand as the one from another structure already in this database, 6W9H, we have included 6XAE in the below analysis and excluded 6W9J. As mentioned in the Introduction section, a few additional structures of ROR-ligand complexes were reported after our search but we did not include them in the study.

Structural Ensemble of Apo ROR From MD Simulation

To construct the initial conformation of the unliganded ROR system, we used a crystal structure of a coactivator-fused ROR (PDB: 5VB3) that is absent of any ligands (Li et al., 2017). The structure is deemed to be in an active conformation, and it was selected from a set of four apo crystal structures (PDB: 5K38, 5VB3, 5X8U, and 5X8W) which are void of agonist ligand binding (Li et al., 2017; Noguchi et al., 2017). It is worth noting that three structures (PDB: 5VB3, 5X8U, and 5X8W) within that set are not fully unliganded since they are either bound to or fused with the coactivator peptide (CoA), while the other structure (PDB: 5K38) without CoA binding has an incomplete C-terminus. As a fused protein complex of the ligand-binding domain of ROR and CoA, the CoA component is believed to assist ROR in remaining in the active conformation. Interestingly, the CoA effect is so strong that the inverse agonist-bound form of this fused protein is still in the active conformation as the PDB 4YMQ shows (Muegge et al., 2015). To simulate our fully apo system, we removed the fused CoA segment from the ROR protein of the crystal structure. The protein contains 243 residues with internal indices (1–243) corresponding to the standard RORγ (UniProt: P51449) A265–S507.

The AMBER14SB forcefield was used for the protein molecules of the simulation, whereas the TIP3P model was used to solvate the system with 11,622 water molecules in a rectangular box. The protonation status of the residues was determined by H++ (Anandakrishnan et al., 2012) at pH of 7.0 and assigned accordingly: Asp and Glu are deprotonated, Arg and Lys are protonated, and all His are singly protonated at the ε position, except His452 and His479 which are protonated at the δ position. Two Cl^- counterions were added to neutralize the system.

After the initial setup of the system, we conducted minimization, heating, and production runs using NAMD. NPT simulations were used for the system with T = 300 K and p = 1 atm. The production run time was 100 ns after an initial 5 ns equilibration. The time step was 2 fs, and snapshots were collected every 1 ps.

Statistical Analysis

Contact matrices are calculated to render the structure information at residue-residue contact resolution (Johnson

et al., 2015; Clark et al., 2016; Johnson et al., 2018). For the calculation of residue-residue and residue-ligand contacts of these 136 PDBs, hydrogen atoms are excluded from all but the 10 ligands for which they were explicitly reported. Since the hydrogen atoms of proteins were not explicitly reported either, we remove hydrogens from all of the PDBs to obtain a uniform resolution (heavy atom only) of the protein complex systems. A contact a_{ij} between two components, i and j (a pair of amino acid residues or a residue and a ligand), is considered formed $a_{ij} = 1$ if the minimum distance between heavy atoms from the two components is within 4.5 Å, otherwise $a_{ij} = 0$. For the corresponding processing of simulation results, the distance cutoff is 4.2 Å using an all-atom resolution (hydrogen included).

Several analysis methods are used to further render the information contained in the contact matrix ensemble. Besides the principal component analysis (PCA) (Jolliffe, 2002; Brunton and Kutz, 2019) of contacts, statistical analyses such as hierarchical clustering and construction of a dendrogram can be used to classify the protein structures (contact maps), the residues of binding pocket, and the ligands. The distance score between two protein contact matrices a_{ij} and b_{ij} is defined as $N_{10} = \sum_{ij} (a_{ij} - b_{ij})^2$. An alternative definition based on $1/\sum_{ij}(a_{ij} \cdot b_{ij}) = 1/N_{11}$ provides similar clustering results. Here, N_{10} is the number of the elements that are different (i.e., logic gate XOR performed) and N_{11} is the number of elements that are 1 (contact formed) in both cases. Note that sequentially neighboring contacts (those between residue i and i+1, i+2) are not counted. Since different proteins have different starting and ending positions for their contact maps, we use a common region (between 276 and 475) and thus a contact matrix size of 200×200 for this distance calculation.

For protein-ligand interactions, we used an I-PCA style of contact statistical analysis (Lindsay et al., 2018). The I-PCA method was initially developed to reveal internal domain structures of semi-structured biopolymers, from large-scale chromosome structures to intrinsically disordered proteins (Das et al., 2020; Lindsay et al., 2021). In those cases, each row (or column) of the mean contact map of the structure ensemble is treated as a linear set of contact variables (the number of rows is the number of monomer units of the system) that symbolize the contact interaction with other unlabeled monomers. Here, we generalize this idea to protein-ligand contacts, i.e., protein residues have a contact variable L that forms unspecified contact with ligands, i.e., $L_i = 1$ if residue i is in contact with the ligand or 0 otherwise. Thus, we emphasize the correlation of contact formation between residue i with the ligand while residue j simultaneously forms contact with the ligand. The covariance matrix is $C_{ij} = <\delta L_i \cdot \delta L_j> = \sum_{\alpha=1}^{K} \delta L_i^{\alpha} \cdot \delta L_j^{\alpha}/K$. Here, $\delta L_i^{\alpha} = L_i^{\alpha} - <L_i>$ is the protein-ligand contact fluctuation of residue i and symbol the $< >$ indicates an average performed over $K = 132$ different protein-ligand contact patterns. The emphasis is on which residue makes contact, while the details of ligand structure are not emphasized here. One can expect that applying I-PCA may sort out the dominant contact interaction patterns between residues and ligands.

RESULTS AND DISCUSSION

Conformations With Various Ligand Binding Status Expressed by Contact Matrices

We first use contact interaction matrices to compare the conformations of RORγ structures reported in the PDB database. As mentioned in the previous section, a distance (dissimilarity score N_{10}) between two conformations measured by the similarity between the corresponding contact matrices is calculated. This contact-based similarity measurement provides a simple way of grouping similar conformations. Practically, the value of N_{10} ranges from 0 to 142. The hierarchical clustering of structures characterized by these contact maps (based on the pairwise dissimilarity scores) is represented as a dendrogram in **Figure 2**.

With the exception of some isolated structures that branch off earlier, the bulk of the structures form two main branches in the dendrogram. As seen in **Figure 2**, the active structures form the majority of one branch, the active branch, whereas the inactive ones concentrate in the other branch. Note that none of the nine structures complexed with allosteric inhibitors (PDBs: 4YPQ, 5C4O, 5C4S, 5C4T, 5C4U, 5LWP, 6SAL, 6TLM, and 6UCG) is in the active branch and most of them are in the inactive branch. Two structure outliers (5K3M and 5K3N) branch off the earliest, since they have distinct conformations compared to the rest of the structures with less contacts being formed. In addition, three of the four apo structures can be found within the "active branch" of the cluster tree and the fourth is found within the "inactive branch."

We use color labeling to demonstrate the conformation being active (blue) or inactive (red) in **Figure 2**. Active structures are largely located in one branch of the cluster tree. Note that our definition of active conformation is based on the features of contact matrices as stated below. We could not solely rely on the self-reported status from literature associated with the PDBs, because not all ligands were self-reported as an agonist or an inverse agonist. Additionally, the active or inactive conformation is not always linked to the ligand being reported as an agonist or an inverse agonist. In certain cases, such as a coactivator-fused ROR or due to ROR-coactivator interaction, a known inverse agonist can be "trapped" within the active conformation of the protein (e.g., PDB: 5APJ). However, by visual inspection of the contact interactions, one can clearly observe two distinct patterns of contact maps being formed. The first group of contact maps is predominantly associated with self-identified active conformations and features two regions of contacts that are absent from the second (inactive) group. One of the two regions represents contacts between H3-H4 and HX-H12, whereas the other region contains contacts between HX and H12. Examples of the active and inactive contact maps are displayed in **Figure 2**.

In practice, we summed the total number of contacts from two regions on the contact maps where Region one is defined as any contacts between residues i and j, i.e., (i, j) satisfy $300 < i < 340$ and $j > 475$, and Region two by any contacts satisfying $470 < i < 495$, $i < j$, and $j > 490$. We further applied a cutoff

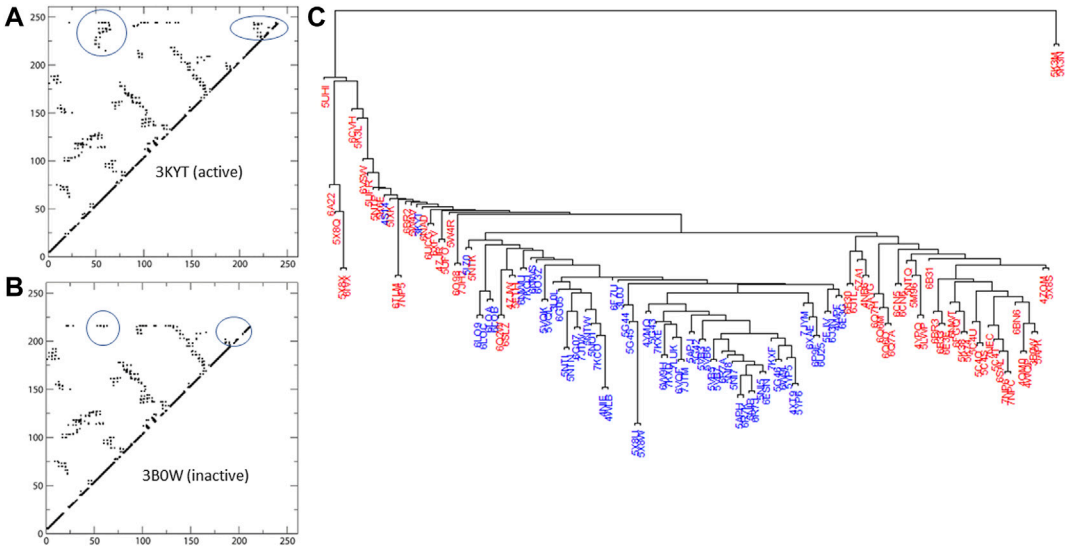


FIGURE 2 | Active and inactive structures of RORγ are rendered as contact maps in **(A,B)**, respectively. Based on the similarity matrix, a dendrogram of the structures (active labeled in blue, inactive in red) at the contact level is displayed in **(C)**.

value of 60 contacts in these two regions combined to determine whether a specific structure is active or inactive, with 60 and above considered as active. Based on this *ad hoc* cutoff criterion, we can separate all the structures into two camps of roughly equal size: 67 of 136 structures are considered active and the remaining 69 are considered inactive. This cutoff selection and the ensuing definition of structure (active vs. inactive) are also proven to be largely consistent with most self-reported or presumed classification of ligand status (agonist vs. inverse agonist). Out of these 132 ligand-bound structures, 113 have a consistent ligand identity and structure identity. There are 19 structures with a presumed or self-reported inverse agonist that yield a value slightly greater than our cutoff of 60 (mostly around 65–70), which makes them active structures by our definition. Several factors can contribute to this result. Besides the factor that a structure can be influenced by elements other than the ligand's nature (e.g., 5APJ vs. 5APK), different experimental tests to determine the nature of the ligand being an inverse agonist or not are inconsistent. Besides, the action of the ligand binding is not a discrete value, but rather the level of effect is on a continuous spectrum.

Ligand Binding Patterns Revealed by Statistical Analysis of Protein-Ligand Contacts

In general, the ligand binding pockets of NR can be quite large and complex. For most cases, ligands are considered to be completely enclosed inside the LBD of the receptor. A unique aspect of ROR is that it contains an allosteric binding site besides the canonical (orthosteric) binding site (Scheepstra et al., 2015). It was reported further that both sites can be occupied by ligands and exhibit a degree of communication between them (de Vries et al., 2021). Specifically, even when an agonist ligand binds to the canonical site and stabilizes the binding pocket structure, the presence of an allosteric inverse agonist can negate the agonistic effect and turn off the receptor activity (de Vries et al., 2021). Although such complex multivalent interaction is interesting to study and can have deep implications on controlling how the protein functions *via* allostery (Pham et al., 2019b), our study is limited to only single ligand-protein interaction.

A basic property of ligands which we can investigate is their size and its relationship with the ligand identity as an agonist or an inverse agonist. Here, we chose to characterize each ligand by its total number of atoms. The mean ligand size is $n = 58.4$ with a standard deviation of 17.5, whereas the active structures have a mean ligand size of $n_a = 55.5$ and the inactive structures have $n_i = 61.9$. Although ligands in the inactive structures are slightly heavier, the difference is much smaller compared to the standard deviation of size distribution. Thus, we conclude that ligand size is not a determining factor as to whether the ligand is an agonist or not. The conclusions drawn here are insensitive to alternative definitions using molecular weight or number of heavy atoms, as these three definitions are highly correlated.

With the same noise filtering cutoff, we define the binding-pocket residues as those forming contacts with ligands in at least 10 out of 132 ligand-bound structures. As a result, we found a total of 55 residues forming the binding pocket. For comparison, a slightly more relaxed, alternative cutoff of 8 hits yields a total of 56 residues.

As shown in **Figure 3A**, the total number of ligands N_i^T, is shown as a function of protein residue index i. Particularly, residues at the binding pocket form constant contacts with

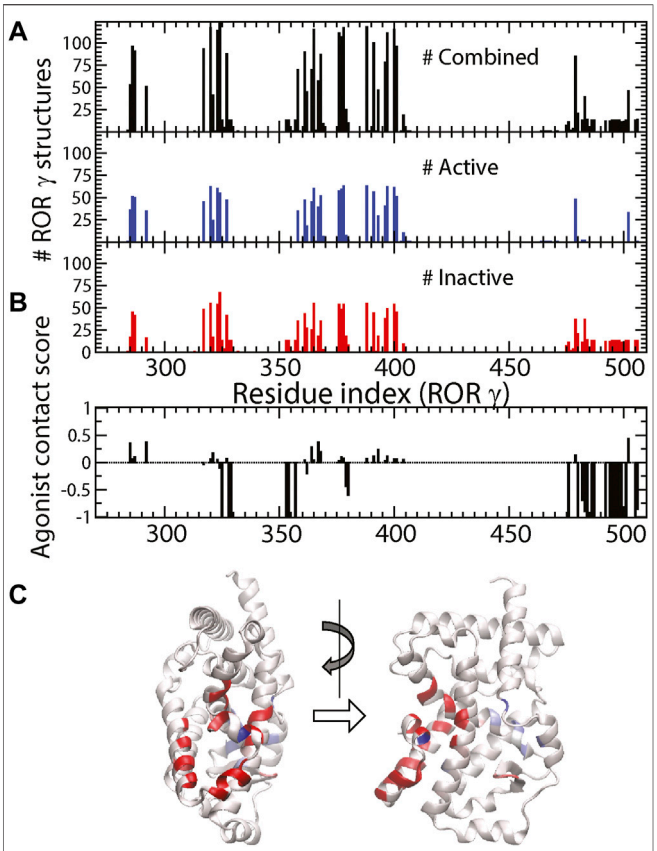


FIGURE 3 | (A) The total number of ligand-protein contacts made by a specific residue (black), the number of ligand-protein contacts in active and inactive structures (blue and red). **(B)** The agonist contact score is displayed as a function of residue. **(C)** The corresponding values from **(B)** are color-labeled onto the 3D structure of RORγ (front and side views).

ligands are 320, 323–324, 365, 376–378, 388, 397, and 400 as they form contacts with a ligand in more than 80% of the structure ensemble. One can separately list the number of residue-ligand contacts N_i^T formed in active ($N_i^A = \sum' {}^A L_i$) vs. inactive ($N_i^I = \sum' {}^I L_i$) conformations based on the (in)active definitions defined in the previous subsection. Here $\Sigma'^{A,I}$ is a restricted sum for active and inactive conformations respectively and $N_i^T = N_i^A + N_i^I$. Furthermore, we have defined the preference of each residue as the agonist contact score, $x = (N^A - N^I)/(N^A + N^I)$, as seen in **Figure 3B**. The value of x is in the range of $[-1, 1]$, where -1 indicates residue-ligand contacts only formed in the inactive case and +1 are those only formed in the active case. Note that we filtered out residues that have minimal contacts with the ligand ($N_i^T \leq 9$) to avoid poor statistics.

It might be useful to point out that the residues with strongest negative agonist contact scores (highlighted in red in **Figure 3C**) are located at five specific spots (near residues 325, 328–329, 353–354, 357, and 379–380) and the C-terminus (476–506). Note that three of those spots belong to the allosteric binding pocket and the other two spots (residues 357 and 379–380) reside in the boundary between the allosteric and canonical binding pockets. These inverse agonist "hot spots" can be important for ligand design and some of them have been reported as the "trigger" for inducing inactive conformations of RORγ. For example, one of the most potent inverse agonists from isoxazole family was reported to have contact interaction with Q329, L353, and K354 (Meijer et al., 2020). Another example on the boundary is the M358 trigger (Marcotte et al., 2016) mentioned in the Introduction. Combined with molecular docking (Trott and Olson, 2010), this contact score (**Figure 3B**) can be further developed and applied to high throughput screening for selecting new inverse agonist ligands for RORγ. This statistical approach may also be generalized to study ligand recognition by other receptors.

Besides obtaining independent statistics on the ligand contact tendency of each residue, we further investigated the concerted pattern of the residue-ligand contacts, i.e., whether residues *i* and *j* form protein-ligand contacts in sync. Various statistical analyses can be used to achieve this correlation analysis, and we use contact PCA as described in the Method section. The contact PCA on the covariance matrix of residue-ligand contacts provides the dominant patterns of residue-ligand interaction. The top eigenvectors PC1 and PC2 were presented in **Figures 4A,B**. We also analyzed the PC projection for PC1 and PC2, which is shown in **Figure 4C**. Each PC mode indicates a specific binding pattern: all residues with positive values form contacts with the ligand (i.e., not necessarily the same ligand) in sync and the same goes for residues with all negative values. Additionally, there is an anti-correlation between positive residues and negative residues. One can see that the dominant mode, PC1, largely divides residues into two groups. As ligand contacts from conformations of PDB structures (4YPQ, 5C4O, 5C4S, 5C4T, 5C4U, 5LWP, 6SAL, 6TLM, and 6UCG) mostly come from the positive group, they show up as positive PC projections, whereas the remaining conformations comprise the negative group. Overall, we found that PC1 distinguishes two binding modes: allosteric binding for the positive group and canonical binding for the negative one. The position of the allosteric binding pocket is distal to the traditional canonical binding pocket, and the ligands that interact with the allosteric binding pocket have been found to be a class of inverse agonists (Meijer et al., 2020; Vries et al., 2020; Zhang et al., 2020; de Vries et al., 2021; Meijer et al., 2021). Function-wise, these allosteric inverse agonists induce another orientation of helix H12 such that it prevents the binding of a coactivator. The second dominant interaction pattern, PC2, shows another prominent binding feature, which seems to weakly separate agonist vs. inverse agonist binding. It is interesting to point out that most conformations with an extreme positive PC2 projection are inactive conformations (red) and vice versa, an extreme negative PC2 for active conformations (black).

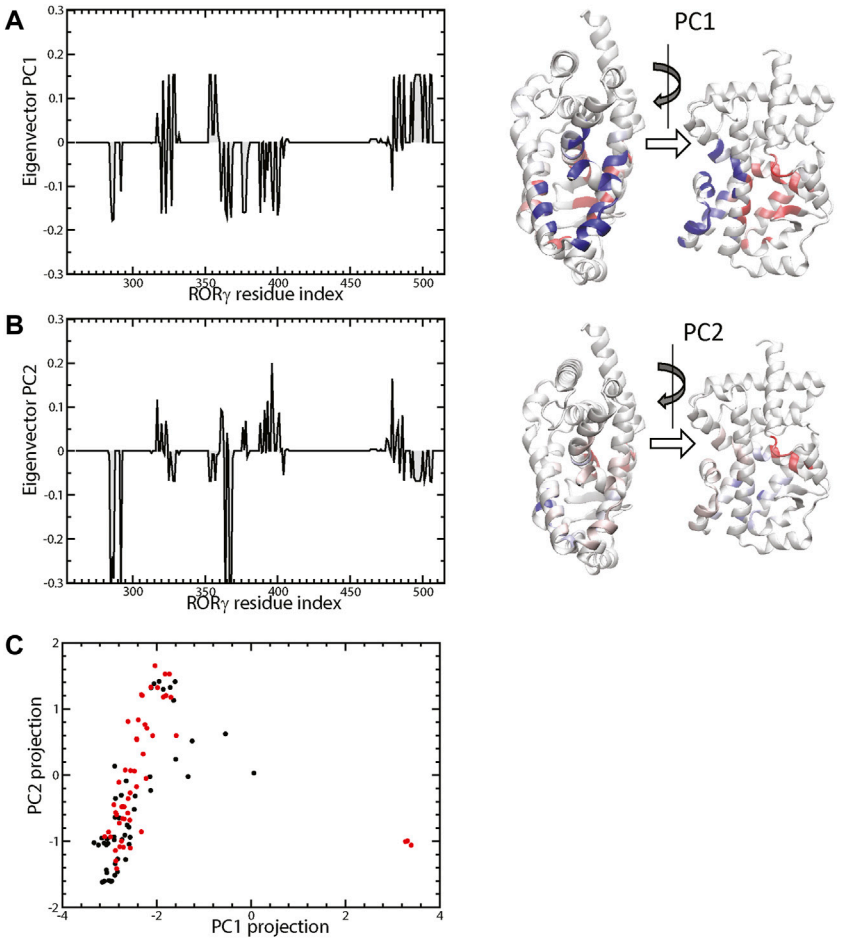


FIGURE 4 | The top two eigenvectors PC1 and PC2 are shown in **(A,B)**, respectively. The 3D representations are colored by the elements of the corresponding eigenvectors (blue+ and red–). The projection of protein-ligand interaction from the top two PCs of the 132 structures (active labeled in black and inactive in red) is shown in **(C)**.

Essential Contacts for Constitutive Activity Revealed by Molecular Dynamics Simulation

As mentioned in the Introduction, the high basal activity of certain NRs makes finding the mechanism of constitutive activity and inhibitors to these NRs important. Since the apo structure of NRs in general is difficult to obtain and most existing NR structures are complexed with ligand(s), we use computer simulation to sample the apo structural ensemble. Furthermore, we use statistics of residue-residue contacts to characterize the mechanism of high basal activity at the residue-residue interaction resolution.

We performed a 100-ns simulation on the unliganded RORγ and analyzed the snapshots using contact analysis, which focuses on residue-residue contact interaction during the simulation. We then examined the contact interactions within the RORγ receptor and focused on identifying contacts with high interaction strength in the apo ensemble. The mean contact matrix of apo RORγ is shown in **Figure 5A** (upper triangle). Each element of the mean contact matrix (also termed contact frequency) is displayed on a contact map using spectrum color-labeling, ranging from rarely formed with contact ratio at 0.1–0.2 (red) to nearly always formed at 0.9–1.0 (dark gray), which largely reflects the contact interaction strength during the simulation.

There are many ways of selecting essential contact interactions that are responsible for high basal activity. Here, we focus on two aspects of conserved highly-formed contact interactions. One aspect is the conservation across different nuclear receptors and the other is the conservation between ligand-bound and apo forms. Thus, we emphasize that the essential contacts are the contacts that not only consistently show up regardless of the ligand binding status but also persist across different NRs. To address how constitutive activity can be conserved across nuclear receptors, we compare the essential contact interactions of RORγ with those of a prominent constitutively active receptor, CAR. The LBDs of both receptors are similar in size and structure. The LBD of RORγ contains 243 residues compared to 242 for CAR. Structure-wise, these two LBDs share a similar fold and both display a short helix, HX, which is unique among the LBDs of nuclear receptors. The presence of helix HX in CAR has been suggested to stabilize the active conformation of the apo form leading to the constitutive

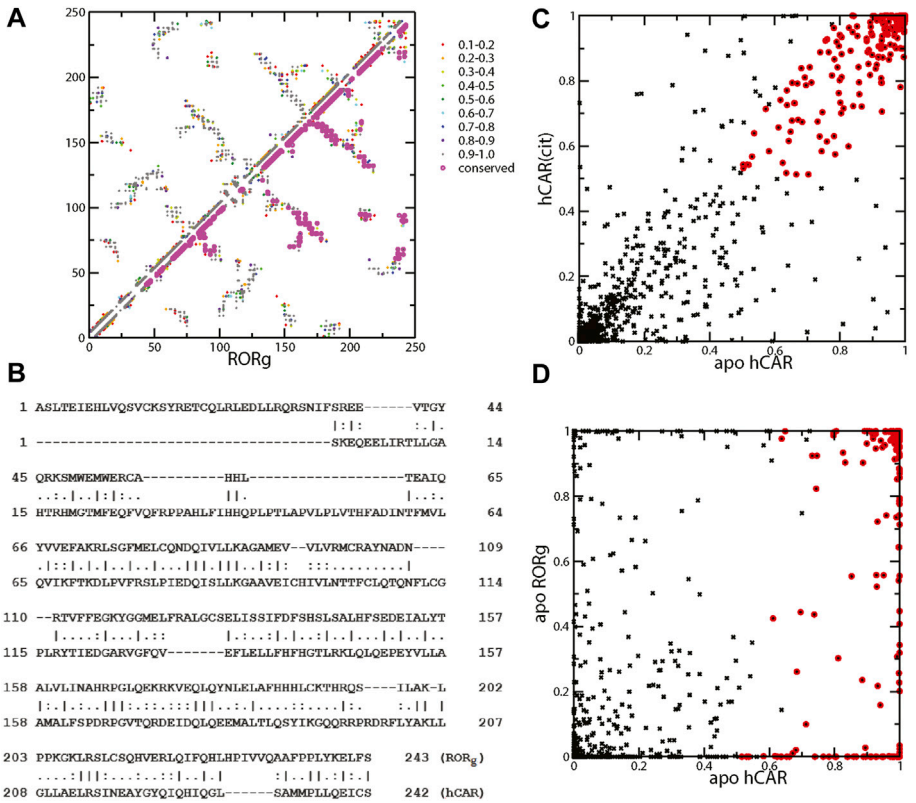


FIGURE 5 | (A) Mean contact interaction of unliganded RORγ during simulation (upper-left triangle). The contact strength values are color-coded, and the conserved contacts are highlighted in magenta. **(B)** Sequence alignment between RORγ (UniProt: P51449) and hCAR (UniProt: Q14994) for a direct comparison between contacts. The listed index can be converted to the standard index by +264 and +106 for RORγ and hCAR, respectively, i.e., the first Ala is A265 for RORγ and the first Ser is S107 for hCAR. **(C)** The scatter plot of contact interaction between unliganded hCAR and hCAR with the agonist CITCO. The conserved contacts are labeled using red circles. **(D)** The scatter plot of contact interaction between unliganded hCAR and unliganded RORγ.

activity of CAR (Pham et al., 2019a). The sequence alignment on RORγ and CAR shows a good alignment and conserved residues in **Figure 5B**, especially after the first 50 residues. The sequences of the two receptors share 59 identical residues (~20%). This alignment facilitates our comparison of residues between RORγ and CAR and the comparison of residue-residue contact interactions between the two receptors.

Before we locate the conserved contacts between NRs, we first identify the contacts that are conserved between apo and agonist ligand-bound forms. Since we only performed the apo RORγ simulation and we have previously obtained both ligand-bound and apo simulations for CAR, we use the CAR system to select the contacts between apo and agonist ligand-bound forms. Specifically, we use an *ad hoc* selection criterion of contact formation that is larger than 50% for both forms to identify the conserved contacts between apo and ligand-bound forms, and these conserved contacts are annotated with red circles as seen in **Figure 5C**. Furthermore, to locate the essential contacts that are conserved between different receptors, we directly compared the contact interactions of RORγ with the CAR receptor, as shown in **Figure 5D**. Again, the red circles are annotated for the conserved contacts selected (based on the high contact conservation between apo and ligand-bound) from **Figure 5C**. Finally, a subset of annotated contacts, which are the conserved contacts across NRs (defined as larger than 50% for both apo forms), is highlighted in magenta in **Figure 5A** (lower triangle). The conservation between CAR and RORγ contacts is quite extensive especially at the C-terminal half of the LBD, which leads to the conclusion that the mechanism of constitutive activity is similar between them. It may be of interest to investigate whether we can apply the inverse agonist ligand design of ROR to another system, e.g., to explore a potential allosteric binding site of CAR.

Based on the mean contact strength (**Figure 5A**) and the sequence alignment (**Figure 5B**), we found that the contacts between helices H11 and H12 are preserved for the two apo receptors. Specifically, the contact pairs H479-Y502 (H11–H12) and Y502-F506 (H12) of RORγ have a high contact strength and they are similar to Y326-L343 and L343-C347 in CAR. These three residues H479-Y502-F506 are collectively known as the HYF triplet, which forms a contact interaction network that is important for RORγ activity (Li et al., 2017; Ma et al., 2021). For example, inverse agonist may function by interacting with residue M358 and further disrupting contact interaction involving F506 (Marcotte et al., 2016). Previously, Y326-L343 in CAR was found

to be critical to the agonist activity in the active conformation for CAR. In both CAR and RORγ, the His-Tyr lock stabilizes the position of helix H12 and contributes to the formation of the AF2 region. The disruption of H479-Y502 (H11–H12) through mutagenesis can prevent the coactivator from binding, thus reducing RORγ transcriptional activity (Kurebayashi et al., 2004). This is also supported by a high number of ligands forming contacts with both residues His479 and Tyr502 in the active conformation of RORγ in **Figure 3A**. In a previous study, the equivalent contact to His479-Tyr502 in CAR (Tyr326-Leu343) has been shown to be present in the apo conformation and strengthened by the binding of an agonist ligand (Pham et al., 2019a). Analogous to CAR, the His-Tyr lock is also present in our apo RORγ simulation with the average contact strength of 96.2%.

CONCLUDING REMARKS

We studied the existing crystal structures of nuclear receptor RORγ, where various ligands (100+) interact with the binding pocket differently and result in an active or inactive conformation. By characterizing the protein conformation and protein-ligand interaction using residue contact interactions, we further performed a statistical analysis on these contact patterns. We identified the important residues at the binding pocket(s) that may be essential for interacting with potential inverse agonists. Besides studying the experimental data on a protein-ligand complex, we also used simulation to examine the apo structure ensemble and compared the high basal activity between RORγ and CAR. We found that the mechanism of constitutive activity is highly similar between them. These efforts lead to the understanding of the structure ensemble and protein-ligand interaction from a contact viewpoint, and they may facilitate future designs of inverse agonists for nuclear receptors.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

BP, RL, DF, and TS conceptualized and designed the research. All authors involved extracting data from PDB files and calculated contact interactions. BP set up the computer simulation and performed computer simulation. BP and TS performed simulation data analysis and wrote the initial draft. All authors participated in the revision.

FUNDING

This work was also supported in parts by NIH R15 GM123469.

ACKNOWLEDGMENTS

We thank Dr. E. J. Fernandez for the helpful discussion on NR constitutive activity. We also acknowledge the computational support provided by the allocations of advanced computing resources XSEDE (STAMPEDE2 at TACC) on the simulation of unliganded ROR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.904445/full#supplementary-material

REFERENCES

Amaudrut, J., Argiriadi, M. A., Barth, M., Breinlinger, E. C., Bressac, D., Broqua, P., et al. (2019). Discovery of Novel Quinoline Sulphonamide Derivatives as Potent, Selective and Orally Active RORγ Inverse Agonists. *Bioorg. Med. Chem. Lett.* 29, 1799–1806. doi:10.1016/j.bmcl.2019.05.015

Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012). H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* 40, W537–W541. doi:10.1093/nar/gks375

Brunton, S. L., and Kutz, J. N. (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press.

Carcache, D. A., Vulpetti, A., Kallen, J., Mattes, H., Orain, D., Stringer, R., et al. (2018). Optimizing a Weakly Binding Fragment into a Potent RORγt Inverse Agonist with Efficacy in an *In Vivo* Inflammation Model. *J. Med. Chem.* 61, 6724–6735. doi:10.1021/acs.jmedchem.8b00529

Chao, J., Enyedy, I., Van Vloten, K., Marcotte, D., Guertin, K., Hutchings, R., et al. (2015). Discovery of Biaryl Carboxylamides as Potent RORγ Inverse Agonists. *Bioorg. Med. Chem. Lett.* 25, 2991–2997. doi:10.1016/j.bmcl.2015.05.026

Cherney, R. J., Cornelius, L. A. M., Srivastava, A., Weigelt, C. A., Marcoux, D., Duan, J. J.-W., et al. (2020). Discovery of BMS-986251: A Clinically Viable, Potent, and Selective RORγt Inverse Agonist. *ACS Med. Chem. Lett.* 11, 1221–1227. doi:10.1021/acsmedchemlett.0c00063

Clark, A. K., Wilder, J. H., Grayson, A. W., Johnson, Q. R., Lindsay, R. J., Nellas, R. B., et al. (2016). The Promiscuity of Allosteric Regulation of Nuclear Receptors by Retinoid X Receptor. *J. Phys. Chem. B* 120, 8338–8345. doi:10.1021/acs.jpcb.6b02057

Cook, D. N., Kang, H. S., and Jetten, A. M. (2015). Retinoic Acid-Related Orphan Receptors (RORs): Regulatory Functions in Immunity, Development, Circadian Rhythm, and Metabolism. *Nucl. Recept. Res.* 2, 101185. doi:10.11131/2015/101185

Das, P., Golloshi, R., McCord, R. P., and Shen, T. (2020). Using Contact Statistics to Characterize Structure Transformation of Biopolymer Ensembles. *Phys. Rev. E* 101, 012419. doi:10.1103/PhysRevE.101.012419

de Vries, R. M. J. M., Meijer, F. A., Doveston, R. G., Leijten-van de Gevel, I. A., and Brunsveld, L. (2021). Cooperativity Between the Orthosteric and Allosteric Ligand Binding Sites of RORγt. *Proc. Natl. Acad. Sci.* 118, e2021287118. doi:10.1073/pnas.2021287118

Duan, J. J.-W., Jiang, B., Lu, Z., Stachura, S., Weigelt, C. A., Sack, J. S., et al. (2020). Discovery of 2,6-difluorobenzyl Ether Series of Phenyl ((R)-3-phenylpyrrolidin-3-yl)sulfones as Surprisingly Potent, Selective and Orally Bioavailable RORγt Inverse Agonists. *Bioorg. Med. Chem. Lett.* 30, 127441. doi:10.1016/j.bmcl.2020.127441

Duan, J. J.-W., Lu, Z., Jiang, B., Stachura, S., Weigelt, C. A., Sack, J. S., et al. (2019). Structure-based Discovery of Phenyl (3-Phenylpyrrolidin-3-Yl) Sulfones as

Selective, Orally Active RORγt Inverse Agonists. *ACS Med. Chem. Lett.* 10, 367–373. doi:10.1021/acsmedchemlett.9b00010

Dussault, I., Lin, M., Hollister, K., Fan, M., Termini, J., Sherman, M. A., et al. (2002). A Structural Model of the Constitutive Androstane Receptor Defines Novel Interactions that Mediate Ligand-independent Activity. *Mol. Cell Biol.* 22, 5270–5280. doi:10.1128/mcb.22.15.5270-5280.2002

Enyedy, I. J., Powell, N. A., Caravella, J., van Vloten, K., Chao, J., Banerjee, D., et al. (2016). Discovery of Biaryls as RORγ Inverse Agonists by Using Structure-Based Design. *Bioorg. Med. Chem. Lett.* 26, 2459–2463. doi:10.1016/j.bmcl.2016.03.109

Fauber, B. P., de Leon Boenig, G., Burton, B., Eidenschenk, C., Everett, C., Gobbi, A., et al. (2013). Structure-based Design of Substituted Hexafluoroisopropanol-Arylsulfonamides as Modulators of RORc. *Bioorg. Med. Chem. Lett.* 23, 6604–6609. doi:10.1016/j.bmcl.2013.10.054

Fauber, B. P., René, O., de Leon Boenig, G., Burton, B., Deng, Y., Eidenschenk, C., et al. (2014). Reduction in Lipophilicity Improved the Solubility, Plasma-Protein Binding, and Permeability of Tertiary Sulfonamide RORc Inverse Agonists. *Bioorg. Med. Chem. Lett.* 24, 3891–3897. doi:10.1016/j.bmcl.2014.06.048

Fujita-Sato, S., Ito, S., Isobe, T., Ohyama, T., Wakabayashi, K., Morishita, K., et al. (2011). Structural Basis of Digoxin that Antagonizes RORγt Receptor Activity and Suppresses Th17 Cell Differentiation and Interleukin (IL)-17 Production. *J. Biol. Chem.* 286, 31409–31417. doi:10.1074/jbc.m111.254003

Fukase, Y., Sato, A., Tomata, Y., Ochida, A., Kono, M., Yonemori, K., et al. (2018). Identification of Novel Quinazolinedione Derivatives as RORγt Inverse Agonist. *Bioorg. Med. Chem.* 26, 721–736. doi:10.1016/j.bmc.2017.12.039

Gege, C., Albers, M., Kinzel, O., Kleymann, G., Schlüter, T., Steeneck, C., et al. (2020). Optimization and Biological Evaluation of Thiazole-Bis-Amide Inverse Agonists of RORγt. *Bioorg. Med. Chem. Lett.* 30, 127205. doi:10.1016/j.bmcl.2020.127205

Gege, C., Cummings, M. D., Albers, M., Kinzel, O., Kleymann, G., Schlüter, T., et al. (2018). Identification and Biological Evaluation of Thiazole-Based Inverse Agonists of RORγt. *Bioorg. Med. Chem. Lett.* 28, 1446–1455. doi:10.1016/j.bmcl.2018.03.093

Gong, H., Weinstein, D. S., Lu, Z., Duan, J. J.-W., Stachura, S., Haque, L., et al. (2018). Identification of Bicyclic Hexafluoroisopropyl Alcohol Sulfonamides as Retinoic Acid Receptor-Related Orphan Receptor Gamma (RORγ/RORc) Inverse Agonists. Employing Structure-Based Drug Design to Improve Pregnane X Receptor (PXR) Selectivity. *Bioorg. Med. Chem. Lett.* 28, 85–93. doi:10.1016/j.bmcl.2017.12.006

Harikrishnan, L. S., Gill, P., Kamau, M. G., Qin, L.-Y., Ruan, Z., O'Malley, D., et al. (2020). Substituted Benzyloxytricyclic Compounds as Retinoic Acid-Related Orphan Receptor Gamma T (RORγt) Agonists. *Bioorg. Med. Chem. Lett.* 30, 127204. doi:10.1016/j.bmcl.2020.127204

Helsen, C., Kerkhofs, S., Clinckemalie, L., Spans, L., Laurent, M., Boonen, S., et al. (2012). Structural Basis for Nuclear Hormone Receptor DNA Binding. *Mol. Cell. Endocrinol.* 348, 411–417. doi:10.1016/j.mce.2011.07.025

Hintermann, S., Guntermann, C., Mattes, H., Carcache, D. A., Wagner, J., Vulpetti, A., et al. (2016). Synthesis and Biological Evaluation of New Triazolo- and Imidazolopyridine RORγt Inverse Agonists. *ChemMedChem* 11, 2640–2648. doi:10.1002/cmdc.201600500

Hirata, K., Kotoku, M., Seki, N., Maeba, T., Maeda, K., Hirashima, S., et al. (2016). SAR Exploration Guided by LE and Fsp3: Discovery of a Selective and Orally Efficacious RORγ Inhibitor. *ACS Med. Chem. Lett.* 7, 23–27. doi:10.1021/acsmedchemlett.5b00253

Hoegenauer, K., Kallen, J., Jiménez-Núñez, E., Strang, R., Ertl, P., Cooke, N. G., et al. (2019). Structure-Based and Property-Driven Optimization of N-Aryl Imidazoles toward Potent and Selective Oral RORγt Inhibitors. *J. Med. Chem.* 62, 10816–10832. doi:10.1021/acs.jmedchem.9b01291

Huss, J. M., Garbacz, W. G., and Xie, W. (2015). Constitutive Activities of Estrogen-Related Receptors: Transcriptional Regulation of Metabolism by the ERR Pathways in Health and Disease. *Biochimica Biophysica Acta (BBA) - Mol. Basis Dis.* 1852, 1912–1927. doi:10.1016/j.bbadis.2015.06.016

Jiang, B., Duan, J. J.-W., Stachura, S., Karmakar, A., Hemagiri, H., Raut, D. K., et al. (2020). Discovery of (3S,4S)-3-Methyl-3-(4-Fluorophenyl)-4-(4-(1,1,1,3,3,3-Hexafluoro-2-Hydroxyprop-2-Yl)phenyl) Pyrrolidines as Novel RORγt Inverse Agonists. *Bioorg. Med. Chem. Lett.* 30, 127392. doi:10.1016/j.bmcl.2020.127392

Jin, L., Martynowski, D., Zheng, S., Wada, T., Xie, W., and Li, Y. (2010). Structural Basis for Hydroxycholesterols as Natural Ligands of Orphan Nuclear Receptor RORγ. *Mol. Endocrinol.* 24, 923–929. doi:10.1210/me.2009-0507

Johnson, Q. R., Lindsay, R. J., Nellas, R. B., Fernandez, E. J., and Shen, T. (2015). Mapping Allostery through Computational Glycine Scanning and Correlation Analysis of Residue-Residue Contacts. *Biochemistry* 54, 1534–1541. doi:10.1021/bi501152d

Johnson, Q. R., Lindsay, R. J., and Shen, T. (2018). CAMERRA: An Analysis Tool for the Computation of Conformational Dynamics by Evaluating Residue-Residue Associations. *J. Comput. Chem.* 39, 1568–1578. doi:10.1002/jcc.25192

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer.

Kallen, J., Izaac, A., Be, C., Arista, L., Orain, D., Kaupmann, K., et al. (2017). Structural States of RORγt: X-Ray Elucidation of Molecular Mechanisms and Binding Interactions for Natural and Synthetic Compounds. *ChemMedChem* 12, 1014–1021. doi:10.1002/cmdc.201700278

Kono, M., Ochida, A., Oda, T., Imada, T., Banno, Y., Taya, N., et al. (2018). Discovery of [cis-3-({(5r)-5-[(7-Fluoro-1,1-Dimethyl-2,3-Dihydro-1h-Inden-5-Yl)carbamoyl]-2-Methoxy-7,8-Dihydro-1,6-Naphthyridin-6(5h)-Yl} Carbonyl)cyclobutyl] Acetic Acid (TAK-828F) as a Potent, Selective, and Orally Available Novel Retinoic Acid Receptor-Related Orphan Receptor γt Inverse Agonist. *J. Med. Chem.* 61, 2973–2988. doi:10.1021/acs.jmedchem.8b00061

Kotoku, M., Maeba, T., Fujioka, S., Yokota, M., Seki, N., Ito, K., et al. (2019). Discovery of Second Generation RORγ Inhibitors Composed of an Azole Scaffold. *J. Med. Chem.* 62, 2837–2842. doi:10.1021/acs.jmedchem.8b01567

Kumar, R., and Thompson, E. B. (2003). Transactivation Functions of the N-Terminal Domains of Nuclear Hormone Receptors: Protein Folding and Coactivator Interactions. *Mol. Endocrinol.* 17, 1–10. doi:10.1210/me.2002-0258

Kummer, D. A., Cummings, M. D., Abad, M., Barbay, J., Castro, G., Wolin, R., et al. (2017). Identification and Structure Activity Relationships of Quinoline Tertiary Alcohol Modulators of RORγt. *Bioorg. Med. Chem. Lett.* 27, 2047–2057. doi:10.1016/j.bmcl.2017.02.044

Kurebayashi, S., Nakajima, T., Kim, S.-C., Chang, C.-Y., McDonnell, D. P., Renaud, J.-P., et al. (2004). Selective LXXLL Peptides Antagonize Transcriptional Activation by the Retinoid-Related Orphan Receptor RORγ. *Biochem. Biophysical Res. Commun.* 315, 919–927. doi:10.1016/j.bbrc.2004.01.131

Li, X., Anderson, M., Collin, D., Muegge, I., Wan, J., Brennan, D., et al. (2017). Structural Studies Unravel the Active Conformation of Apo RORγt Nuclear Receptor and a Common Inverse Agonism of Two Diverse Classes of RORγt Inhibitors. *J. Biol. Chem.* 292, 11618–11630. doi:10.1074/jbc.m117.789024

Lindsay, R. J., Mansbach, R. A., Gnanakaran, S., and Shen, T. (2021). Effects of pH on an IDP Conformational Ensemble Explored by Molecular Dynamics Simulation. *Biophys. Chem.* 271, 106552. doi:10.1016/j.bpc.2021.106552

Lindsay, R. J., Pham, B., Shen, T., and McCord, R. P. (2018). Characterizing the 3D Structure and Dynamics of Chromosomes and Proteins in a Common Contact Matrix Framework. *Nucleic Acids Res.* 46, 8143–8152. doi:10.1093/nar/gky604

Liu, Q., Batt, D. G., Weigelt, C. A., Yip, S., Wu, D.-R., Ruzanov, M., et al. (2020). Novel Tricyclic Pyroglutamide Derivatives as Potent RORγt Inverse Agonists Identified Using a Virtual Screening Approach. *ACS Med. Chem. Lett.* 11, 2510–2518. doi:10.1021/acsmedchemlett.0c00496

Lu, Z., Duan, J. J.-W., Xiao, H., Neels, J., Wu, D.-R., Weigelt, C. A., et al. (2019). Identification of Potent, Selective and Orally Bioavailable Phenyl ((R)-3-phenylpyrrolidin-3-yl) Sulfone Analogues as RORγt Inverse Agonists. *Bioorg. Med. Chem. Lett.* 29, 2265–2269. doi:10.1016/j.bmcl.2019.06.036

Lugar, C. W., Clarke, C. A., Morphy, R., Rudyk, H., Sapmaz, S., Stites, R. E., et al. (2021). Defining Target Engagement Required for Efficacy *In Vivo* at the Retinoic Acid Receptor-Related Orphan Receptor C2 (RORγt). *J. Med. Chem.* 64, 5470–5484. doi:10.1021/acs.jmedchem.0c01918

Ma, X., Sun, N., Li, X., and Fu, W. (2021). Discovery of Novel N-Sulfonamide-Tetrahydroisoquinolines as Potent Retinoic Acid Receptor-Related Orphan Receptor γt Agonists. *Eur. J. Med. Chem.* 222, 113585. doi:10.1016/j.ejmech.2021.113585

Marcotte, D. J., Liu, Y., Little, K., Jones, J. H., Powell, N. A., Wildes, C. P., et al. (2016). Structural Determinant for Inducing RORgamma Specific Inverse Agonism Triggered by a Synthetic Benzoxazinone Ligand. *BMC Struct. Biol.* 16, 7. doi:10.1186/s12900-016-0059-3

Marcoux, D., Duan, J. J.-W., Shi, Q., Cherney, R. J., Srivastava, A. S., Cornelius, L., et al. (2019). Rationally Designed, Conformationally Constrained Inverse Agonists of RORγt-Identification of a Potent, Selective Series with Biologic-

like *In Vivo* Efficacy. *J. Med. Chem.* 62, 9931–9946. doi:10.1021/acs.jmedchem.9b01369

Meijer, F. A., Doveston, R. G., de Vries, R. M. J. M., Vos, G. M., Vos, A. A. A., Leysen, S., et al. (2020). Ligand-Based Design of Allosteric Retinoic Acid Receptor-Related Orphan Receptor γt (RORγt) Inverse Agonists. *J. Med. Chem.* 63, 241–259. doi:10.1021/acs.jmedchem.9b01372

Meijer, F. A., Saris, A. O. W. M., Doveston, R. G., Oerlemans, G. J. M., de Vries, R. M. J. M., Somsen, B. A., et al. (2021). Structure-Activity Relationship Studies of Trisubstituted Isoxazoles as Selective Allosteric Ligands for the Retinoic-Acid-Receptor-Related Orphan Receptor γt. *J. Med. Chem.* 64, 9238–9258. doi:10.1021/acs.jmedchem.1c00475

Muegge, I., Collin, D., Cook, B., Hill-Drzewi, M., Horan, J., Kugler, S., et al. (2015). Discovery of 1,3-Dihydro-2,1,3-Benzothiadiazole 2,2-Dioxide Analogs as New RORC Modulators. *Bioorg. Med. Chem. Lett.* 25, 1892–1895. doi:10.1016/j.bmcl.2015.03.042

Nakajima, R., Oono, H., Kumazawa, K., Ida, T., Hirata, J., White, R. D., et al. (2021). Discovery of 6-Oxo-4-Phenyl-Hexanoic Acid Derivatives as RORγt Inverse Agonists Showing Favorable ADME Profile. *Bioorg. Med. Chem. Lett.* 36, 127786. doi:10.1016/j.bmcl.2021.127786

Nakajima, R., Oono, H., Sugiyama, S., Matsueda, Y., Ida, T., Kakuda, S., et al. (2020). Discovery of [1,2,4] Triazolo[1,5-A] Pyridine Derivatives as Potent and Orally Bioavailable RORγt Inverse Agonists. *ACS Med. Chem. Lett.* 11, 528–534. doi:10.1021/acsmedchemlett.9b00649

Narjes, F., Xue, Y., von Berg, S., Malmberg, J., Llinas, A., Olsson, R. I., et al. (2018). Potent and Orally Bioavailable Inverse Agonists of RORγt Resulting from Structure-Based Design. *J. Med. Chem.* 61, 7796–7813. doi:10.1021/acs.jmedchem.8b00783

Noguchi, M., Nomura, A., Doi, S., Yamaguchi, K., Hirata, K., Shiozaki, M., et al. (2018). Ternary Crystal Structure of Human RORγ Ligand-Binding-Domain, an Inhibitor and Corepressor Peptide Provides a New Insight into Corepressor Interaction. *Sci. Rep.* 8, 17374. doi:10.1038/s41598-018-35783-9

Noguchi, M., Nomura, A., Murase, K., Doi, S., Yamaguchi, K., Hirata, K., et al. (2017). Ternary Complex of Human RORγ Ligand-Binding Domain, Inverse Agonist and SMRT Peptide Shows a Unique Mechanism of Corepressor Recruitment. *Genes cells.* 22, 535–551. doi:10.1111/gtc.12494

Olsson, R. I., Xue, Y., von Berg, S., Aagaard, A., McPheat, J., Hansson, E. L., et al. (2016). Benzoxazepines Achieve Potent Suppression of IL-17 Release in Human T-Helper 17 (TH17) Cells Through an Induced-Fit Binding Mode to the Nuclear Receptor RORγ. *ChemMedChem* 11, 207–216. doi:10.1002/cmdc.201500432

Ouvry, G., Bouix-Peter, C., Ciesielski, F., Chantalat, L., Christin, O., Comino, C., et al. (2016). Discovery of Phenoxyindazoles and Phenylthioindazoles as RORγ Inverse Agonists. *Bioorg. Med. Chem. Lett.* 26, 5802–5808. doi:10.1016/j.bmcl.2016.10.023

Pham, B., Arons, A. B., Vincent, J. G., Fernandez, E. J., and Shen, T. (2019). Regulatory Mechanics of Constitutive Androstane Receptors: Basal and Ligand-Directed Actions. *J. Chem. Inf. Model.* 59, 5174–5182. doi:10.1021/acs.jcim.9b00695

Pham, B., Lindsay, R. J., and Shen, T. (2019). Effector-Binding-Directed Dimerization and Dynamic Communication Between Allosteric Sites of Ribonucleotide Reductase. *Biochemistry* 58, 697–705. doi:10.1021/acs.biochem.8b01131

René, O., Fauber, B. P., Boenig, G. L., Burton, B., Eidenschenk, C., Everett, C., et al. (2015). Minor Structural Change to Tertiary Sulfonamide RORc Ligands Led to Opposite Mechanisms of Action. *ACS Med. Chem. Lett.* 6, 276–281. doi:10.1021/ml500420y

Rosenberg, E. M., Harrison, R. E. S., Tsou, L. K., Drucker, N., Humphries, B., Rajasekaran, D., et al. (2019). Characterization, Dynamics, and Mechanism of CXCR4 Antagonists on a Constitutively Active Mutant. *Cell Chem. Biol.* 26, 662–673. doi:10.1016/j.chembiol.2019.01.012

Ruan, Z., Park, P. K., Wei, D., Purandare, A., Wan, H., O'Malley, D., et al. (2021). Substituted Diaryl Ether Compounds as Retinoic Acid-Related Orphan Receptor-Γt (RORγt) Agonists. *Bioorg. Med. Chem. Lett.* 35, 127778. doi:10.1016/j.bmcl.2021.127778

Santori, F. R., Huang, P., van de Pavert, S. A., Douglass, E. F., Jr., Leaver, D. J., Haubrich, B. A., et al. (2015). Identification of Natural RORγ Ligands that Regulate the Development of Lymphoid Cells. *Cell metab.* 21, 286–298. doi:10.1016/j.cmet.2015.01.004

Sasaki, Y., Odan, M., Yamamoto, S., Kida, S., Ueyama, A., Shimizu, M., et al. (2018). Discovery of a Potent Orally Bioavailable Retinoic Acid Receptor-Related Orphan Receptor-Gamma-T (RORγt) Inhibitor, S18-000003. *Bioorg. Med. Chem. Lett.* 28, 3549–3553. doi:10.1016/j.bmcl.2018.09.032

Sato, A., Fukase, Y., Kono, M., Ochida, A., Oda, T., Sasaki, Y., et al. (2019). Design and Synthesis of Conformationally Constrained RORγt Inverse Agonists. *ChemMedChem* 14, 1917–1932. doi:10.1002/cmdc.201900416

Scheepstra, M., Leysen, S., van Almen, G. C., Miller, J. R., Piesvaux, J., Kutilek, V., et al. (2015). Identification of an Allosteric Binding Site for RORγt Inhibition. *Nat. Commun.* 6, 8833. doi:10.1038/ncomms9833

Schimmer, B. P., and White, P. C. (2010). Minireview: Steroidogenic Factor 1: Its Roles in Differentiation, Development, and Disease. *Mol. Endocrinol.* 24, 1322–1337. doi:10.1210/me.2009-0519

Schnute, M. E., Wennerstål, M., Alley, J., Bengtsson, M., Blinn, J. R., Bolten, C. W., et al. (2018). Discovery of 3-Cyano- N-(3-(1-Isobutyrylpiperidin-4-yl)-1-methyl-4-(Trifluoromethyl)-1 H-Pyrrolo[2,3- B]pyridin-5-Yl) Benzamide: A Potent, Selective, and Orally Bioavailable Retinoic Acid Receptor-Related Orphan Receptor C2 Inverse Agonist. *J. Med. Chem.* 61, 10415–10439. doi:10.1021/acs.jmedchem.8b00392

Shi, Q., Xiao, Z., Yang, M. G., Marcoux, D., Cherney, R. J., Yip, S., et al. (2020). Tricyclic Sulfones as Potent, Selective and Efficacious RORγt Inverse Agonists - Exploring C6 and C8 SAR Using Late-Stage Functionalization. *Bioorg. Med. Chem. Lett.* 30, 127521. doi:10.1016/j.bmcl.2020.127521

Shirai, J., Tomata, Y., Kono, M., Ochida, A., Fukase, Y., Sato, A., et al. (2018). Discovery of Orally Efficacious RORγt Inverse Agonists, Part 1: Identification of Novel Phenylglycinamides as Lead Scaffolds. *Bioorg. Med. Chem.* 26, 483–500. doi:10.1016/j.bmc.2017.12.006

Simons, S. S., Jr, Edwards, D. P., and Kumar, R. (2014). Minireview: Dynamic Structures of Nuclear Hormone Receptors: New Promises and Challenges. *Mol. Endocrinol.* 28, 173–182. doi:10.1210/me.2013-1334

Solt, L. A., and Burris, T. P. (2012). Action of RORs and Their Ligands in (Patho) physiology. *Trends Endocrinol. Metabolism* 23, 619–627. doi:10.1016/j.tem.2012.05.012

Strutzenberg, T. S., Garcia-Ordonez, R. D., Novick, S. J., Park, H., Chang, M. R., Doebellin, C., et al. (2019). HDX-MS Reveals Structural Determinants for RORγ Hyperactivation by Synthetic Agonists. *eLife* 8, 47172. doi:10.7554/eLife.47172

Takeda, Y., Jothi, R., Birault, V., and Jetten, A. M. (2012). RORγ Directly Regulates the Circadian Expression of Clock Genes and Downstream Targets *In Vivo*. *Nucleic Acids Res.* 40, 8519–8535. doi:10.1093/nar/gks630

Tanis, V. M., Venkatesan, H., Cummings, M. D., Albers, M., Kent Barbay, J., Herman, K., et al. (2019). 3-Substituted Quinolines as RORγt Inverse Agonists. *Bioorg. Med. Chem. Lett.* 29, 1463–1470. doi:10.1016/j.bmcl.2019.04.021

Trott, O., and Olson, A. J. (2010). Auto Dock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* 31, 455–461. doi:10.1002/jcc.21334

van Niel, M. B., Fauber, B. P., Cartwright, M., Gaines, S., Killen, J. C., René, O., et al. (2014). A Reversed Sulfonamide Series of Selective RORc Inverse Agonists. *Bioorg. Med. Chem. Lett.* 24, 5769–5776. doi:10.1016/j.bmcl.2014.10.037

von Berg, S., Xue, Y., Collins, M., Llinas, A., Olsson, R. I., Halvarsson, T., et al. (2019). Discovery of Potent and Orally Bioavailable Inverse Agonists of the Retinoic Acid Receptor-Related Orphan Receptor C2. *ACS Med. Chem. Lett.* 10, 972–977. doi:10.1021/acsmedchemlett.9b00158

Vries, R. M. J. M., Doveston, R. G., Meijer, F. A., and Brunsveld, L. (2020). Elucidation of an Allosteric Mode of Action for a Thienopyrazole RORγt Inverse Agonist. *ChemMedChem* 15, 561–565. doi:10.1002/cmdc.202000044

Wang, T., Banerjee, D., Bohnert, T., Chao, J., Enyedy, I., Fontenot, J., et al. (2015). Discovery of Novel Pyrazole-Containing Benzamides as Potent RORγ Inverse Agonists. *Bioorg. Med. Chem. Lett.* 25, 2985–2990. doi:10.1016/j.bmcl.2015.05.028

Wang, Y.-M., Ong, S. S., Chai, S. C., and Chen, T. (2012). Role of CAR and PXR in Xenobiotic Sensing and Metabolism. *Expert Opin. Drug Metabolism Toxicol.* 8, 803–817. doi:10.1517/17425255.2012.685237

Wang, Y., Cai, W., Cheng, Y., Yang, T., Liu, Q., Zhang, G., et al. (2015). Discovery of Biaryl Amides as Potent, Orally Bioavailable, and CNS Penetrant RORγt Inhibitors. *ACS Med. Chem. Lett.* 6, 787–792. doi:10.1021/acsmedchemlett.5b00122

Wang, Y., Cai, W., Tang, T., Liu, Q., Yang, T., Yang, L., et al. (2018). From RORγt Agonist to Two Types of RORγt Inverse Agonists. *ACS Med. Chem. Lett.* 9, 120–124. doi:10.1021/acsmedchemlett.7b00476

Wang, Y., Yang, T., Liu, Q., Ma, Y., Yang, L., Zhou, L., et al. (2015). Discovery of N -(4-Aryl-5-Aryloxy-Thiazol-2-Yl)-Amides as Potent RORγt Inverse Agonists. *Bioorg. Med. Chem.* 23, 5293–5302. doi:10.1016/j.bmc.2015.07.068

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

Xie, W., Yeuh, M.-F., Radominska-Pandya, A., Saini, S. P. S., Negishi, Y., Bottroff, B. S., et al. (2003). Control of Steroid, Heme, and Carcinogen Metabolism by Nuclear Pregnane X Receptor and Constitutive Androstane Receptor. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4150–4155. doi:10.1073/pnas.0438010100

Xu, R. X., Lambert, M. H., Wisely, B. B., Warren, E. N., Weinert, E. E., Waitt, G. M., et al. (2004). A Structural Basis for Constitutive Activity in the Human CAR/RXRα Heterodimer. *Mol. Cell* 16, 919–928. doi:10.1016/j.molcel.2004.11.042

Xue, Y., Guo, H., and Hillertz, P. (2016). Fragment Screening of RORγt Using Cocktail Crystallography: Identification of Simultaneous Binding of Multiple Fragments. *ChemMedChem* 11, 1881–1885. doi:10.1002/cmdc.201600242

Yang, M. G., Beaudoin-Bertrand, M., Xiao, Z., Marcoux, D., Weigelt, C. A., Yip, S., et al. (2021). Tricyclic-Carbocyclic RORγt Inverse Agonists-Discovery of BMS-986313. *J. Med. Chem.* 64, 2714–2724. doi:10.1021/acs.jmedchem.0c01992

Yang, T., Liu, Q., Cheng, Y., Cai, W., Ma, Y., Yang, L., et al. (2014). Discovery of Tertiary Amine and Indole Derivatives as Potent RORγt Inverse Agonists. *ACS Med. Chem. Lett.* 5, 65–68. doi:10.1021/ml4003875

Yukawa, T., Nara, Y., Kono, M., Sato, A., Oda, T., Takagi, T., et al. (2019). Design, Synthesis, and Biological Evaluation of Retinoic Acid-Related Orphan Receptor γt (RORγt) Agonist Structure-Based Functionality Switching Approach from in House RORγt Inverse Agonist to RORγt Agonist. *J. Med. Chem.* 62, 1167–1179. doi:10.1021/acs.jmedchem.8b01181

Zhang, H., Lapointe, B. T., Anthony, N., Azevedo, R., Cals, J., Correll, C. C., et al. (2020). Discovery of N-(Indazol-3-yl) Piperidine-4-carboxylic Acids as RORγt Allosteric Inhibitors for Autoimmune Diseases. *ACS Med. Chem. Lett.* 11, 114–119. doi:10.1021/acsmedchemlett.9b00431

Zhang, Y., Luo, X.-y., Wu, D.-h., and Xu, Y. (2015). ROR Nuclear Receptors: Structures, Related Diseases, and Drug Discovery. *Acta Pharmacol. Sin.* 36, 71–87. doi:10.1038/aps.2014.120

Zhang, Y., Wu, X., Xue, X., Li, C., Wang, J., Wang, R., et al. (2019). Discovery and Characterization of XY101, a Potent, Selective, and Orally Bioavailable RORγ Inverse Agonist for Treatment of Castration-Resistant Prostate Cancer. *J. Med. Chem.* 62, 4716–4730. doi:10.1021/acs.jmedchem.9b00327

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 28 June 2022
doi: 10.3389/fmolb.2022.935375

Molecular Modeling of ABHD5 Structure and Ligand Recognition

Rezvan Shahoei[1], Susheel Pangeni[1], Matthew A. Sanders[2], Huamei Zhang[2], Ljiljana Mladenovic-Lucas[2], William R. Roush[3], Geoff Halvorsen[3], Christopher V. Kelly[1], James G. Granneman[2,4] and Yu-ming M. Huang[1]†*

[1]Department of Physics and Astronomy, Wayne State University, Detroit, MI, United States, [2]Center for Molecular Medicine and Genetics, School of Medicine, Wayne State University, Detroit, MI, United States, [3]Department of Chemistry, Scripps Florida, Jupiter, FL, United States, [4]Center for Integrative Metabolic and Endocrine Research, School of Medicine, Wayne State University, Detroit, MI, United States

Alpha/beta hydrolase domain-containing 5 (ABHD5), also termed CGI-58, is the key upstream activator of adipose triglyceride lipase (ATGL), which plays an essential role in lipid metabolism and energy storage. Mutations in ABHD5 disrupt lipolysis and are known to cause the Chanarin-Dorfman syndrome. Despite its importance, the structure of ABHD5 remains unknown. In this work, we combine computational and experimental methods to build a 3D structure of ABHD5. Multiple comparative and machine learning-based homology modeling methods are used to obtain possible models of ABHD5. The results from Gaussian accelerated molecular dynamics and experimental data of the apo models and their mutants are used to select the most likely model. Moreover, ensemble docking is performed on representative conformations of ABHD5 to reveal the binding mechanism of ABHD5 and a series of synthetic ligands. Our study suggests that the ABHD5 models created by deep learning-based methods are the best candidate structures for the ABHD5 protein. The mutations of E41, R116, and G328 disturb the hydrogen bonding network with nearby residues and suppress membrane targeting or ATGL activation. The simulations also reveal that the hydrophobic interactions are responsible for binding sulfonyl piperazine ligands to ABHD5. Our work provides fundamental insight into the structure of ABHD5 and its ligand-binding mode, which can be further applied to develop ABHD5 as a therapeutic target for metabolic disease and cancer.

Keywords: molecular dynamics, docking, structural modeling, ABHD5, lipid droplet, AlphaFold, ligand binding, protein mutation

OPEN ACCESS

Edited by:
Weiliang Zhu,
Shanghai Institute of Materia Medica (CAS), China

Reviewed by:
Haohao Fu,
Nankai University, China
Ferran Feixas,
Universitat de Girona, Spain
Antti Tapani Poso,
University of Eastern Finland, Finland

***Correspondence:**
Yu-ming M. Huang
ymhuang@wayne.edu

†***ORCID:***
Yu-ming M. Huang
orcid.org/0000-0003-3257-6170

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *03 May 2022*
Accepted: *10 June 2022*
Published: *28 June 2022*

Citation:
Shahoei R, Pangeni S, Sanders MA, Zhang H, Mladenovic-Lucas L, Roush WR, Halvorsen G, Kelly CV, Granneman JG and Huang Y-mM (2022) Molecular Modeling of ABHD5 Structure and Ligand Recognition. Front. Mol. Biosci. 9:935375. doi: 10.3389/fmolb.2022.935375

INTRODUCTION

Lipolysis requires the trafficking and activation of intracellular lipases, such as adipose triglyceride lipase (ATGL, officially known as patatin-like phospholipase domain-containing 2, PNPLA2), to the lipid droplet (LD) surface (Lass et al., 2011; Kintscher et al., 2020). Alpha/beta hydrolase domain-containing protein 5 (ABHD5), also known as comparative gene identification 58 (CGI-58), is a key regulator of the trafficking and activation of ATGL and related members of the patatin-like phospholipase (PNPLA) domain-containing family (Lass et al., 2006; Granneman et al., 2009; Vieyres et al., 2020). Despite its classification as an alpha/beta hydrolase, ABHD5 lacks hydrolase activity owing to the S155N substitution that occurred early in vertebrate evolution (Lass et al., 2006).

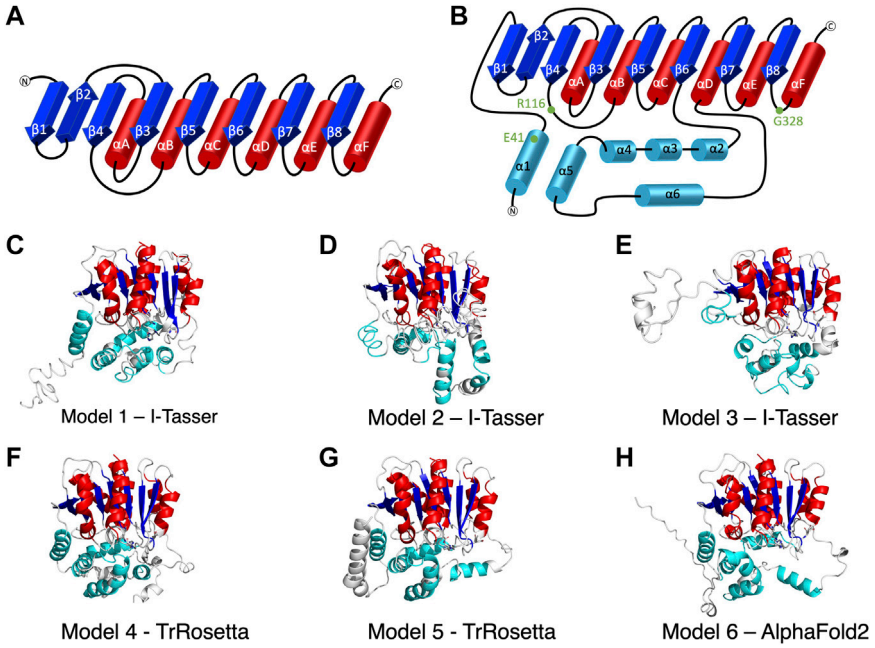


FIGURE 1 | The secondary structure of alpha/beta hydrolase domain-containing 5 (ABHD5). **(A)** The secondary structure diagram of the "canonical" α/β hydrolase fold characterized by six α helices (red) and eight β strands (blue). The catalytic triad in the α/β hydrolase fold family is made up of 1) a nucleophilic residue (Ser, Cys, or Asp) after β5, 2) an acidic residue after β7, and 3) a conserved His after β8. **(B)** The secondary structure diagram of ABHD5 with two insertion regions (cyan). The first insertion region is an α helix (α1), located before β1, and the second insertion region is composed of five α helices (α2–α6) between β6 and αD. **(C–H)** The six models of ABHD5 built from different homology modeling tools. Models 1, 2, and 3 are from I-TASSER; Models 4 and 5 are from TrRosetta; and Model 6 is from AlphaFold2. The "canonical" six α helices and eight β strands for all models are shown in red and blue cartoon representations, respectively. The six insertion α helices are shown in cyan. The secondary structures are defined as: insertion α1 (residues 33–46), β1 (residues 53–59), β2 (residues 64–71), β3 (residues 80–84), αA (residues 89–103), β4 (residues 106–111), αB (residues 126–143), β5 (residues 149–154), αC (residues 157–171), β6 (residues 172–181), insertion α2 (residues 198–207), insertion α3 (residues 221–229), insertion α4 (residues 232–240), insertion α5 (residues 245–257), insertion α6 (residues 259–270), αD (residues 278–288), β7 (residues 292–298), αE (residues 305–314), β8 (residues 319–325), and αF (residues 331–351). Asn155, Asp303, and H329 are drawn in sticks with the carbon atoms in white, nitrogen atoms in blue, and oxygen atoms in red. The hydrogen atoms are not shown.

Instead, ABHD5 evolved structural elements that allow activation of ATGL, as well as a binding site for endogenous and synthetic ligands that regulate interactions with repressor and effector proteins (Sanders et al., 2015). Thus, ABHD5 is a complex protein that contains presently unknown structural elements mediating important functions, including targeting to intracellular LDs, ligand binding, and ATGL activation.

Mutations in the ABHD5 gene cause Chanarin-Dorfman syndrome (Lefevre et al., 2001) wherein the ability of ABHD5 to activate members of the PNPLA family is disrupted and results in disrupted lipid metabolism in numerous organs throughout the body (Hirabayashi et al., 2017; Yang et al., 2019). Previous work has shown the importance of G328 in loss and gain of function assays (Sanders et al., 2017). Recently, a novel mutation of this amino acid in humans (G328E) was reported to produce fatty liver disease (Youssefian et al., 2019). We, therefore, investigated the impact of the G328E mutation in our study. Moreover, endogenous and synthetic ligands bind to ABHD5 protein, which regulates its interactions with protein activators and repressors of lipolysis (Sanders et al., 2015; Rondini et al., 2017).

Because of the importance of lipolysis, efforts have been made to determine ABHD5 structure. NMR experiments revealed the structure and flexibility of the tryptophan-rich N-terminal peptide (residues 10–43) of ABHD5 (Boeszoermenyi et al., 2015). Even though the experimental structure of the entire protein is still unknown, ABHD5 is considered to share the 3D features of the alpha/beta hydrolase fold superfamily. The "canonical" alpha/beta hydrolase fold (Ollis et al., 1992) and the variations inserted among these folds (Nardini and Dijkstra, 1999) have been suggested for the members of the ABHD family (**Figure 1A**). To enhance the structural understanding of ABHD5 activation of ATGL, a homology model of ABHD5 was built by Modeller (Sanders et al., 2017). The model identified R299, G328, and D334 as critical for ATGL activation, which was validated experimentally in both gain- and loss-of-function assays. In this model, G328 was suggested to interact with phospholipids to provide favorable interactions of R299 and D334 (Sanders et al., 2017).

Homology or comparative modeling is the most efficient computational method to predict 3D protein structures using 1D protein sequence data (Hameduh et al., 2020). This modeling approach is based on two assumptions: 1) the 3D structure of a protein is uniquely determined by its amino acid sequence, and 2) during evolution, the changes in the structure are much slower than the changes in the sequence; hence similar sequences adopt similar physical structures. Traditionally, in homology modeling, the structure of the target protein is determined based on another

experimentally derived structure, termed template. The 3D structure of the target protein is then built based on the alignment with the chosen template. For example, I-TASSER is one of the broadly used servers (Roy et al., 2010; Yang et al., 2015). With years of effort, homology modeling has become a major approach in structural prediction, benefiting from the ever-growing number of high-resolution experimental structures deposited in the protein data bank (PDB) and a multitude of new algorithms that improve target-template alignment. In recent years, deep learning-based methods have provided an innovative approach to structural modeling, even when no similar structures are available. The AlphaFold1 algorithm, developed by Alphabet/Google DeepMind, was first introduced in 2018 (Wei, 2019). The method applies convolutional neural networks to predict inter-residue distances, which inspired other deep learning-based methods such as trRosetta in protein structural prediction (Du et al., 2021). The new version, AlphaFold2, released in 2021 (Jumper et al., 2021; Jumper and Hassabis, 2022), uses an innovative network architecture to model atomistic positions with an average success rate greater than 90% (Marx., 2022).

Molecular dynamics (MD) simulations have been successful in numerous biomolecular simulations at the atomic level since they were introduced in 1977 (McCammon et al., 1977). However, despite recent advances, conventional MD (cMD) simulations of biomolecules remain limited to timescales of hundreds of nanoseconds to tens of microseconds, whereas most biological processes take place over milliseconds and longer timescales. To overcome this limitation, two types of enhanced sampling methods were developed. The methods, such as umbrella sampling (Torrie and Valleau, 1977) and metadynamics (Laio and Parrinello, 2002), require the definition of a set of collective variables. However, the algorithms, such as replica-exchange dynamics (Sugita and Okamoto, 1999), accelerated MD (Hamelberg et al., 2004), and Gaussian accelerated MD (GaMD) (Miao et al., 2015), obviate this requirement. GaMD is an enhanced sampling approach in which users may access large biomolecule conformational changes within hundreds of nanoseconds (Miao et al., 2015; Wang et al., 2021). By adding a harmonic boost potential to the original energy surface of the system, users do not need to provide any information about the boost reaction coordinates before executing the simulation, which avoids the simulation bias from pre-defined variables. GaMD has been successfully used in protein-ligand binding, protein folding, and ion channel dynamics studies (Wang et al., 2021). It has been applied in various types of biosystems, such as G-protein-coupled receptor (Miao and McCammon, 2018), HIV protease (Miao et al., 2018), CRISPR-Cas9 (Nierzwicki et al., 2021), ACE2 receptor (Bhattarai et al., 2021), androgen receptor (Zhan et al., 2021), and p38 kinase (Huang, 2021).

In this work, we aimed to construct the atomistic structure of ABHD5 to uncover its dynamics and ligand recognition using newly available computational tools together with experimental validation. First, traditional and deep learning-based homology modeling methods were used to build multiple 3D models for ABHD5. Then, the functional activation was used to select potential ABHD5 structures. The dynamics of ABHD5 under physiological conditions were studied using microsecond-long GaMD simulations. The ABHD5 systems with point mutations were also built and modeled. Finally, we reveal the binding mechanism of synthetic ligands to ABHD5, focusing on one of the independent chemical scaffolds, sulfonyl piperazines (SPZs) (**Figure 2**) that was shown to promote the lipase-activating state of ABHD5 by disrupting its interaction with perilipin 1 (PLIN1) and perilipin 5 (PLIN5) repressors (Sanders et al., 2015). Our work provides a framework for ABHD5 structural modeling and insights into its interaction surface with ligands and related partners, which helps the understanding of ABHD5 functional evolution and lipase regulation.

MATERIALS AND METHODS

Modeling Systems

This work focuses on the ABHD5 protein from *Mus musculus*. The protein sequence was obtained from UniProt with the entry ID: Q9DBL9. Three SPZ ligands, SR-01000604559 (CID: 2674365), SR-03000003133 (CID: 4827533), and SR-03000003134 (CID: 25701322), were investigated in this study. The abbreviations SR4559, SR3133, and SR3134 are applied to the ligands in this paper (**Figure 2**). The Maestro software was used to build the ligand structures for later docking and MD simulations.

Structural Modeling

I-TASSER (Yang et al., 2015), TrRosetta (Du et al., 2021), and AlphaFold2 (Jumper and Hassabis, 2022) were the three programs used to predict ABHD5 structures. I-TASSER requires a structural template, which can be chosen by the software (default mode) or provided by the user. Both TrRosetta and AlphaFold2 employ a deep learning algorithm for protein structure prediction and do not need a user input template. TrRosetta requires only the protein sequence, and the predicted structures by AlphaFold2 are available on AlphaFold Protein Structure Database. To obtain the structural templates for I-TASSER, the blastp-suite of Protein BLAST (Zhang and Madden, 1997; Altschul et al., 1998; Madden et al., 2019) and LALIGN (Madeira et al., 2019) programs were applied to search for potential templates according to the structural data presented in PDB. A total of five templates were selected, PDB code: 6I8W, 3NWO, 3SK0, 1S8O, and 6NY9. The first four templates have the highest overall alignment scores to ABHD5 (**Supplementary Table S1**). The last template, 6NY9, was selected as it is a high resolution (1.66 Å) X-ray crystallography structure of ABHD10, a protein in the same family and from the same organism (Cao et al., 2019). Using I-TASSER, we performed six calculations, including the default mode and five calculations for each template we selected above, which reports 30 models in total. TrRosetta offered five models, and AlphaFold2 provided only one model. To select models for later studies, we used the g_cluster program, an RMSD-based clustering method, in GROMACS 2021.2 (Pronk et al., 2013) to cluster the resulted models. The program reported the six most representative ABHD5 structures, shown in **Figures 1C–H**.

Simulation System Setup

To build hydrogen atoms on ABHD5, the protonation states of ABHD5 were assigned by the Adaptive Poisson Boltzmann Solver

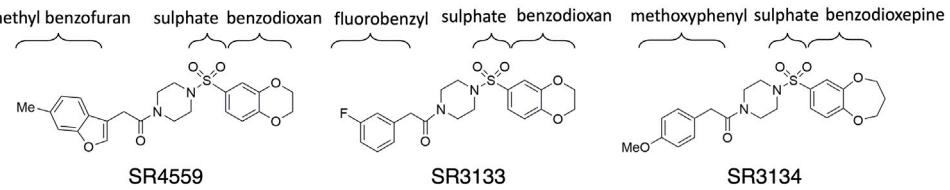


FIGURE 2 | The 2D sketches of the three sulfonyl piperazine (SPZ) ligands.

(APBS) webserver (Jurrus et al., 2018). We applied the Amber18 package (Case et al., 2015) for the GaMD simulation setup and production simulations with an efficient GPU implementation. The Amber ff14SB (Maier et al., 2015) and General Amber Force Field (GAFF) were applied to the protein and ligand, respectively (Ozpinar et al., 2010). Before solvation, energy minimization was performed on the hydrogen atoms, protein side chains, and the entire system for 500, 5,000, and 5,000 steps, respectively. The systems were then solvated in TIP3P water molecules (Jorgensen et al., 1983; Izadi and Onufriev, 2016) with ~12 Å between the box edge and the solutes to create a rectangular box. A salt concentration of 150 mM NaCl was added to the system (Machado and Pantano, 2020) using the ion parameters developed by Joung and Cheatham (2008). Additional minimizations of the water and ion molecules and the entire system (including water, ions, and protein) were performed for 2 and 10 ps, respectively. The equilibration of the system started from the solvent equilibration for 100 ps, then the complex system was gradually heated to 50, 100, 150, 200, 250, and 300 K for 10 ps at each temperature using the isochoric-isothermal (NVT) ensemble. To ensure the system reached equilibrium, a 5.0 ns cMD simulation was further performed at 300 K using the isobaric-isothermal (NPT) ensemble. The time step of the MD simulations was 2 fs. Periodic boundary conditions were applied for the simulation systems, and long-range electrostatics were accounted for using the particle mesh Ewald summation method (Essmann et al., 1995) with a cutoff of 12 Å. Bonds containing hydrogen atoms were restrained using the SHAKE algorithm (Krautler et al., 2001). The Langevin thermostat with a damping constant of 2 ps^{-1} was turned on to maintain a temperature of 300 K (Loncharich et al., 1992).

GaMD Simulations

GaMD (Miao et al., 2015) is an enhanced sampling method that can perform aggressive sampling of a biomolecule. By adding a harmonic boost potential (ΔV) on the original potential energy surface (V), the modified potential (V^*) can be written as $V^*(\vec{r}) = V(\vec{r}) + \Delta V(\vec{r})$

$$\Delta V\left(\vec{r}\right) = \begin{cases} \frac{1}{2}k\left(E - V\left(\vec{r}\right)\right)^2, & if\ V\left(\vec{r}\right) < E \\ 0, & if\ V\left(\vec{r}\right) \geq E \end{cases} \quad (1)$$

where k is the harmonic force constant and E is the threshold energy that should be lower than the system potential (V). To ensure that the boost potential does not alter the shape of the original potential surface, the threshold energy needs to satisfy the following relation:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k} \quad (2)$$

where V_{max} and V_{min} are the system maximum and minimum potential energies, respectively.

To start, a 2-ns cMD simulation was executed to collect the potential statistics, such as V_{max} and V_{min}, for calculating GaMD acceleration parameters. Then, we performed a 1-ns GaMD simulation with applied boost potential but no updating on V_{max} and V_{min} values. The second GaMD simulation was carried out with the updated boost potential for 50 ns. Finally, we performed 1,000 ns production GaMD simulations with a fixed boost potential for all systems. The upper bound for the boost acceleration was selected for all simulations (iE = 2). The average and standard deviation of the system potential energies were calculated every 500 ps. The boost potential was added to both the dihedral energy and the total potential energy. The upper limit of the standard deviation of the potential energy was set to 6.0 kcal/mol for both the dihedral and total potential energy terms. We saved the resulting trajectories every 10 ps for analysis. Note that, GaMD modifies the potential energy surface without considering the entropic contribution. However, the entropic effect would not affect the overall calculation in this study.

We performed 1-μs GaMD simulations for six wild-type ABHD5 models. Because Model 6 was selected as the most representative ABHD5 model, additional four 1-μs GaMD simulations were executed. We also performed 1-μs GaMD simulations for the three ABHD5 mutants (E41A, R116N, and G328E) and three ABHD5-ligand complexes.

Post-GaMD Analysis

All simulation trajectories were visualized by the VMD program (Humphrey et al., 1996). The CPPTRAJ program (Roe and Cheatham, 2013) from the Amber18 package (Case et al., 2015) was used to analyze the root mean square fluctuation (RMSF), residue-residue correlation, and hydrogen bonds. The RMSF was used to measure the average fluctuation of the Cα atom of a specific protein residue over time. The hydrogen bonds were considered when the donor-acceptor distance was less than 3.5 Å, and the donor-hydrogen-acceptor angle was less than 30°. The amino acids in the ligand binding pocket were defined as the residues that are within 3.5 Å of the ligands, and the occupancy is

over 75% of 1-μs GaMD trajectories. To obtain the representative conformation from the GaMD trajectories of the six models, the g_cluster program from GROMACS 2021.2 (Pronk et al., 2013) was used to cluster the 1-μs trajectory for each model into 10 conformations. The most representative conformation of each model was reported. The VMD (Humphrey et al., 1996) and PyMOL, programs were used to create images for the publication.

Ligand Docking

We applied the g_cluster program from GROMACS 2021.2 (Pronk et al., 2013) to cluster the structural ensemble from the 1-μs GaMD trajectories of the ABHD5 protein built by AlphaFold2 into 10 representative conformations. Because the protein is restrained during docking simulations, to avoid the structural bias, these 10 conformations were used for ligand docking. AutoDock Vina (Trott and Olson, 2010; Eberhardt et al., 2021) was applied to dock three Scripps Research (SR) ligands shown in **Figure 2**. The 3D structures of all ligands were prepared using Schrödinger Maestro software. The AutoDock Tools (ADT) (Morris et al., 2009) of the MGLTools (Forli et al., 2016) was used to prepare the proper file formats (pdbqt) for the ligands and the protein conformations and to determine the docking box sizes, which were set to 26, 28, and 30 Å. The docking box center was selected near the center of the protein and the exhaustiveness value was set to 64. For each ligand system, we performed a 1-μs GaMD simulation on the reported ABHD5-ligand complex structure with the best docking score.

Cell Culture, Imaging, Scoring, and Luciferase Complementation

ABHD5-dependent activation of ATGL was performed as described previously (Sanders et al., 2017). Briefly, Cos7 cells plated on coverslips in 12-well dishes were transfected with 0.5 μg each/well of mCherry-tagged ABHD protein, PLIN5, or ATGL using Lipofectamine and Plus reagent (Invitrogen) as described by the manufacturer. Cells were then lipid-loaded for 16–20 h with 200 μM oleic acid, then fixed with 4% paraformaldehyde. Images were acquired with an Olympus IX-81 microscope equipped with a spinning disc confocal unit. Images were captured using a 60x, 1.4 NA objective and a Hamamatsu ORCA Flash cooled CMOS camera. The following Chroma filter sets were used for the indicated fluorophores: mCherry, 41043; EYFP, 31044; ECFP, 41028. LD scoring was performed by an investigator blinded to transfection conditions. For each transfection condition in each experiment, 25 or more cells visibly expressing all three proteins were scored. Mutant ABHD5 proteins were made using standard molecular biological methods, and all PCR-derived proteins were confirmed by sequencing. ABHD5 proteins are from mice, and the numbering of amino acids refers to the mouse protein unless indicated otherwise. PLIN5 and ATGL were also from the mouse.

ABHD5 ligand binding was assessed by ligand-induced inhibition of luciferase complementation between ABHD5 and PLIN5, as previously described (Granneman et al., 2007). Briefly, cell lysates were prepared from 293T cells that were transiently transfection with ABHD5 or PLIN5 fused to the C- or N-terminal fragments of G. princeps luciferase, respectively. Lysates were mixed together in the presence of indicated concentration of ABHD5 ligands or DMSO (control) for 4 h at room temperature, after which coelenterazine substrate was added and the resulting luminescence was read in a Clariostar multiplate reader. Ligand affinity (IC_{50} values) was determined by nonlinear regression using GraphPad software.

RESULTS AND DISCUSSION

ABHD5 Structures

We report six potential ABHD5 structures (**Figures 1C–H**; **Supplementary Text S1**). Models 1, 2, and 3 were selected from the resulting structures created by I-TASSER, a template-based homology modeling method. Models 4 and 5 were built using TrRosetta, and the structure obtained from AlphaFold2 is shown as Model 6. Both TrRosetta and AlphaFold2 are deep learning-based methods for protein structure prediction. Despite using different computational methods in constructing these six ABHD5 models, all structures follow the "canonical" alpha/beta hydrolase fold (Ollis et al., 1992; Mindrebo et al., 2016), where the protein is composed of six α helices, eight β sheets, and two insertion regions—an α helix (α1) near the N-terminal before β1 and the helical insertions (α2–α6) between β6 and αD (**Figure 1B**). In the six models, the three catalytic triad residues, N155, H329, and D303, are all in proximity to each other. Although the lipase substrate/ligand site is preserved, it is not an active site as ABHD5 lacks hydrolase activity (Lass et al., 2006).

The largest variations among the six models come from the N-terminal (residues 1–52) and the insertion region between β6 and αD (residues 198–270). Our 1-μs GaMD simulations reveal that both the N-terminal and the insertion regions display high RMSD and RMSF values (**Supplementary Figures S1B,C, S2**), indicating these areas are flexible in the ABHD5 protein. Although the alpha helices predicted by the homology modeling in these regions maintain along the GaMD simulations, the loops connecting the helices fluctuate. This is consistent with the report from AlphaFold2 that the structural modeling on these areas has the lowest confidence. In Models 1, 4, 5, and 6, the N-terminal residues form α helices. Models 1, 4, and 6 show a helix from residues 33–46, while Model 5 contains three helices, residues 1–19, 21–25, and 33–46. These α helices are all close to the bulk of the protein in Model 5. In Model 2, although the N-terminal does not form a secondary structure, it still stays close to the rest of the protein. However, the N-terminal in Model 3 is different from other models, in which only a short helix (residues 16–26) is present, and the N-terminus does not form interactions with other parts of the protein (**Figure 1E**). The results from both GaMD and homology modeling agree well with the early NMR finding (Boeszoermenyi et al., 2015) that the N-terminal peptide is very flexible and may not form a stable secondary structure. The most representative conformation of each model from the GaMD simulations is shown in **Supplementary Figure S3**.

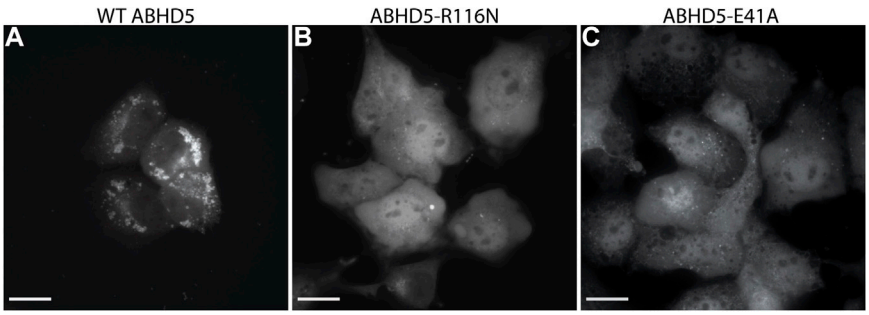


FIGURE 3 | Fluorescence imaging reveals the importance of R116 and E41 in the basal localization of ABHD5. **(A)** Wild-type (WT) ABHD5-mCherry localizes to the LDs within COS7 cells. Either the **(B)** R116N or **(C)** E41A mutations demonstrate a cytosolic distribution of ABHD5-mCherry and inhibited basal LD targeting in the absence of ATGL and PLIN5 expression. **(A–C)** The scale bars represent 20 μm.

Validation and Differentiation of ABHD5 Models

To validate the ABHD5 models created from homology modeling, we first executed point mutation experiments and GaMD simulations. A critical step in the activation of lipolysis is the association of ABHD5 to lipid membranes, and we found that a point mutation, R116N, is specifically defective in basal membrane binding (**Figure 3**). We hypothesized that R116 stabilizes an amphipathic helix that facilitates the membrane binding of ABHD5. To gain deeper insights into the structural basis, we examined the residue-residue interactions near R116 in the six models. Except in Models 2 and 3, R116 forms interactions with E41. The distance between the sidechain oxygen of E41 (OE1/2) and the sidechain nitrogen of R116 (NH1/2) is 2.77, 10.92, 34.21, 4.70, 4.96, and 2.72 Å for Models 1–6, respectively. We next examined the dynamics of all models in solution by GaMD simulations. In Models 1, 4, 5, and 6, our 1-μs simulations reveal that R116 forms stable electrostatic interactions with E41, within an α-helix encompassing residues 33–46, and R116N mutation disrupts the interaction between R116 and E41 and its associated helix. However, no R116-E41 interactions are found in Models 2 or 3. To evaluate if the R116-E41 interactions affect ABHD5 membrane binding, we performed a point mutation experiment on E41. We found that the E41A mutation of ABHD5 phenocopies the R116N mutant—both mutants were defective in basal membrane binding (**Figure 3**). The experimental results support the hypothesis from Models 1, 4, 5, and 6 that the interactions between R116 and E41 play a key role in stabilizing the ABHD5 structure and promoting an extended amphipathic helix to penetrate the phospholipid tails and bring about membrane binding. Taken together, Models 2 and 3 present unlikely structures of ABHD5 as the interactions between E41 and R116 are missing.

To further differentiate Models 1, 4, 5, and 6, we examined the interactions of R299, G328, and D334, which are conserved residues necessary for mediating PNPLA2 activation (Sanders et al., 2017). We hypothesized that a stable interaction of R299, G328, and D334 would occur in more probable models of ABHD5. In Models 4, 5, and 6, the interactions of R299, D328, and D334 are stable during 1-μs GaMD simulations.

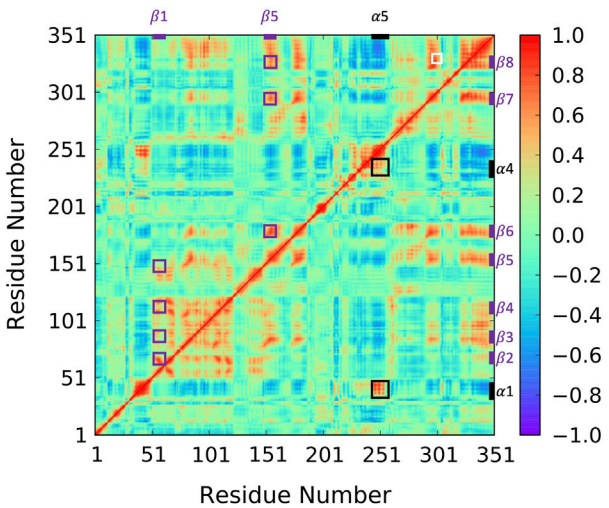


FIGURE 4 | Correlation map for 1-μs GaMD trajectory of Model 6. The correlation between β1 and β2–β5 as well as the correlation between α5 and β6–β8 are highlighted using purple boxes. The correlation between insertion helix α5 and insertion helices α1 and α4 are highlighted with black boxes. The correlation formed by R299, G328, and D334 is highlighted by a white box.

However, in Model 1, R299 moves away from G328 and D334, destabilizing their interactions during the GaMD simulation (**Supplementary Figure S4A**).

In Models 4 and 5, the N-terminal peptide (residues 1–32) either interacts with the insertion helices, α2, α3, and α6, or makes direct contact with D334 (**Supplementary Figures S4F,G**), both of which are not a preferred structure for an LD-bound conformation of ABHD5. It has been hypothesized that the N-terminal peptide forms interactions with the phospholipids on the LD surface (Boeszoermenyi et al., 2015), so the N-terminal peptide should be near E41 and R116. Although the goal of this work is not to obtain a membrane-bound structure of ABHD5, identifying a protein structure consistent with the membrane-bound conformation will assist in further experimental and computational studies (Sanders et al., 2015; Rondini et al.,

2017). Because the N-terminal peptide in ABHD5 is highly flexible, further simulations of Models 4 and 5 may reveal changes to the peptide position that are consistent with the membrane-binding conformation. However, to save computational efforts, we chose Model 6 for the following mutation and ligand binding studies because, in Model 6, we found that 1) E41 forms interactions with R116 that stabilize the N-terminal amphipathic helix, 2) the interactions of R299, G328, and D334 are stable, and 3) the N-terminal peptide is closer to the protein-membrane binding surface.

ABHD5 Protein Dynamics

To reveal ABHD5 protein dynamics and equilibrate the structure created from homology modeling, we performed five replicas of 1-μs GaMD simulations on Model 6 created by AlphaFold2. Our simulations show that the β strands (β1–β8) form strong interactions and correlations with each other (**Figure 4** purple labels), stabilizing the overall protein folding. The GaMD simulations also refine some helical structures, such as αD, which was reported with low confidence by the AlphaFold2 algorithm (Jumper and Hassabis, 2022). In this canonical alpha/beta hydrolase fold, the eight β strands and the six α helices demonstrate less fluctuations as indicated by the RMSF calculations (**Supplementary Figure S2**). However, the N-terminal (residues 1-52 including α1) and the region composed of insertion helices (residues 198–270 including α2-α6) are highly flexible, which is also the main variance within other alpha/beta-hydrolase proteins. The α5 insertion helix shows correlations with the α1 and α4 insertion helices (**Figure 4** black labels). No correlations were found between other insertion helices, suggesting that the motions of the secondary structures in the insertion area are mostly independent.

Since the mutation of R299 and G328 disrupt the ATGL activation by ABHD5 (Sanders et al., 2017), we closely examined the interactions near these two residues. Our model shows that R299 and G328 are in the vicinity of D334. The three residues, R299, G328, and D334, are located on the protein surface exposed to the solvent molecules. The residues form a stable hydrogen bond network and move together during the simulation (**Figure 4** white label; **Supplementary Figure S4**). These results are consistent with the previously reported homology model created by Modeller (Sanders et al., 2017; Tseng et al., 2022).

ABHD5 Mutations

Mutations of the ABHD5 protein can alter its membrane binding, protein binding, ligand binding, lipolysis activation, and consequently its function. In this study, three new ABHD5 mutations were identified—E41A and R116N affect the ABHD5 membrane binding (**Figure 3**), and G328E suppresses the ability of ABHD5 to activate ATGL (**Supplementary Figure S5**). These mutations alter the local interactions with nearby residues and further restrain the protein function.

In wild-type ABHD5, E41 forms hydrogen bonds with R116, K38, and K54 (**Supplementary Figure S6A**). The hydrogen-bonding network between these four residues stabilizes the secondary structure of α1, β1, the loop between β4 and αB, and a cavity that could facilitate the binding of negatively charged phospholipid heads as the binding site is composed of the positively charged K38, K54, and R116. The E41A mutation disrupts the hydrogen bonding network (**Figure 5**; **Supplementary Figure S6B**). Without this hydrogen bonding, the sidechains of K54 and R116 rotate away from the cavity and no longer move in concert (**Supplementary Figure S7**), which may contribute to the reduced membrane binding of E41A. With the E41A mutation, the smaller alanine sidechain contributes to the increased distance between E41A and R116, which results in structural changes of nearby insertion helices. For example, the α5 insertion moves closer to insertion α1. This motion of the insertion α5 alters the mobility of the neighboring insertion helices, α3, α4, and α6 (**Supplementary Figure S8**). Although the E41A mutation does not alter the overall folding of the eight canonical β strands, the correlation within the β strands reduces significantly (**Supplementary Figure S7**).

The R116N ABHD5 reveals very similar experimental phenotype and molecular dynamics results to the E41A mutant. Although E41 can still form hydrogen bonds with R116N, the occupancy of hydrogen bonding in the 1-μs GaMD simulation reduces from 71.7% to 56.3%, leading to deformation of the cavity formed by K38, E41, K54, and R116N. We hypothesized that the misorientation of the K38, E41, K54, and R116 sidechains disturbs phospholipid binding and membrane targeting (**Supplementary Figure S6C**). The R116N mutation alters the motions of neighboring insertion helices, similar to the E41A mutation.

In wide-type ABDH5, G328 is spatially close to R299 and D334 while both the R299 sidechain and the G328 mainchain form stable hydrogen bonds with the D334 sidechain (**Supplementary Figure S9A**). Both R299 and G328 are situated at loops—R299 is at the loop between β7 and αE, and G328 is at the loop between β8 and αF, whereas D334 is in the αF helix. The electrostatic interactions among the three residues connect the two loops and the αF helix, stabilizing the loop conformation. Five independent GaMD simulations of the wild-type ABHD5 show a pocket formed by the insertion helices, α2 and α4, and the two loops containing R299, G328, and D334 (**Supplementary Figure S9C**). The pocket could accommodate the binding of ABHD5 protein partners. In humans, G328E mutation results in neutral lipid storage disease (Youssefian et al., 2019), so it is of interest to determine how this mutation alters the shape of this critical pocket. We found that the positively charged sidechain of G328E forms strong electrostatic interactions with R299, eliminating the interactions between R299 and D334 and altering the conformation of the loop between β7 and αE (**Supplementary Figure S9B**). In addition, the charged sidechain of G328E also forms hydrogen bonds with V198 near the insertion helix α2, which further alters the structure of insertion α2 and results in closing of the pocket (**Supplementary Figure S9D**). For example, the distance between G328Cα and V198Cα reduces after the mutation (**Supplementary Figure S10**). In the G328E GaMD simulation, the RMSF measurement reduces in the region between amino acids 300 and 340 (**Supplementary Figure S2**), indicating that the reduced flexibility of the protein may affect the ATGL activation.

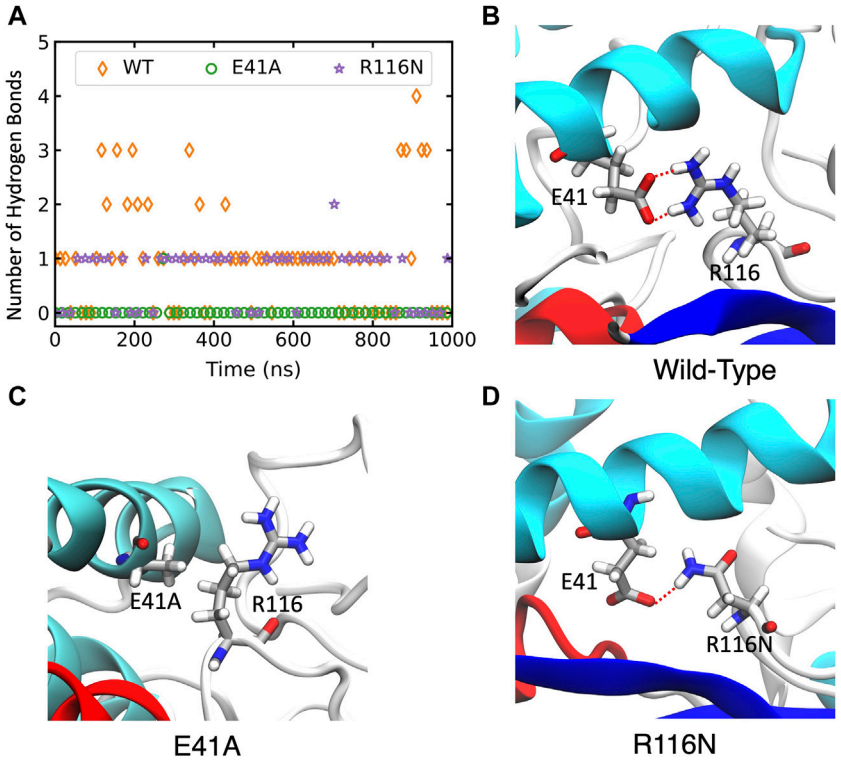


FIGURE 5 | The changes of hydrogen bonds between residues 41 and 116 in the wild-type (WT) ABHD5 and its mutants. **(A)** The number of hydrogen bonds between residues 41 and 116 over the course of 1-μs GaMD trajectories. The changes in hydrogen bonds of the WT, E41A, and R116N systems are indicated by orange, green, and purple, respectively. **(B–D)** The GaMD snapshots of the WT, E41A, and R116N system.

ABHD5 Ligand Complexes

To reveal the dynamics of ABHD5-ligand complexes, we explored the binding mode of three synthetic SPZ ligands, namely SR4559, SR3133, and SR3134. The three ligands are structurally analog with a similar 2D sketch (**Figure 2**). Both SR4559 and SR3133 have a benzodioxan and sulphate functional group. SR4559 includes a methyl benzofuran group, while SR3133 has a fluorobenzyl group. SR3134 includes a benzodioxepine group instead of a benzodioxan group of SR4559 and SR3133. SR3134 also has a sulphate group and displays a similar sketch to SR3133. Although some computational techniques, such as attach-pull-release (Velez-Vega and Gilson, 2013), confine-and-release (Cacelli and Prampolini, 2007), and BFEE2 (Fu et al., 2022) may estimate ligand binding affinities *in silico*, we measured the ABHD5-SPZ ligand binding by experiments. The results show that SR4559, SR3133, and SR3134 dissociate the ABHD5-PLIN5 complexes in Cos7 cell lysates with IC_{50} values of $3.44 \pm 0.50 \times 10^{-6}$ M, $1.59 \pm 0.66 \times 10^{-5}$, and $8.90 \pm 1.84 \times 10^{-6}$ M, respectively (**Figure 6**), indicating that all three ligands bind ABHD5.

Our docking and GaMD simulations show that the three ligands are located at the pocket formed by the insertion helices, α2-α6 (**Figure 7A**; **Supplementary Text S1**). The region of the insertion helices is typically flexible and challenging in protein structural prediction, thus the accuracy of the predicted ABHD5 structure directly corresponds to the docking results. The RMSF values for the residues 220–250 significantly reduce in the three ligand-bound models compared to the apo model (**Supplementary Figure S11**),

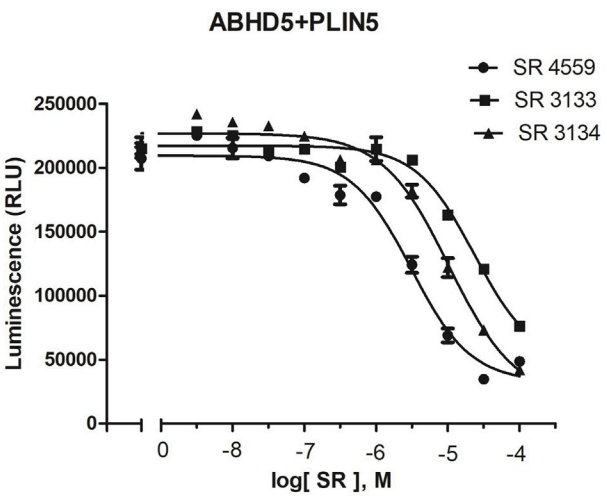


FIGURE 6 | The binding of SPZ ligands to ABHD5 was assessed by dose-response curves of ABHD5-PLIN5 binding. ABHD5-PLIN5 binding was assessed by luciferase complementation assays of Cos7 lysates. Each ligand has micromolar IC_{50} values with specific binding to ABHD5 (Sanders et al., 2015).

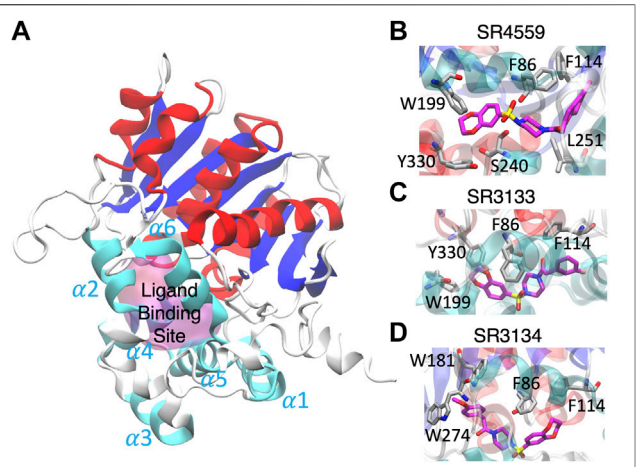


FIGURE 7 | The ligand binding modes of SPZ ligands. **(A)** SPZ ligands bind to the pocket (magenta) formed by insertion helices, α2-α6 (cyan). Red and blue indicate the canonical α helices and β strands, respectively. **(B–D)** reveal the residues (carbons in gray) interacting with SR4559, SR3133, and SR3134 (carbons in magenta), respectively.

suggesting that the ligand binding restrains the protein motion. Besides, each SPZ analog interacts with similar residues (F86, G87, F114, L203, A206, N211, Y238, S240, I251, N255, and E262) in the binding pocket (**Supplementary Table S2**).

Although the residues contributing to ligand binding are similar, the detailed interactions are different for each ligand. In the SR4559 binding, the methyl benzofuran group forms nonpolar interactions with F86, F114, and L251, and the benzofuran functional group interacts with W199, S240, and Y330 (**Figure 7B**). Although SR3133 displays a similar binding mode to SR4559, fewer contacts with protein residues were found. For example, the fluorophenyl group interacts with F86 and F114, and the benzofuran functional group interacts with W199 and Y330 (**Figure 7C**). SR3134 presents a different binding mode. The benzodioxepine group interacts with F86 and F114, while the methoxyphenyl functional group binds deeply into the pocket formed by W181 and W274 (**Figure 7D**).

CONCLUSION

In this work, we employed traditional and deep learning-based homology modeling tools to model the structure of the ABHD5 protein. The ABHD5 structures reported from the deep learning-based modeling, including TrRosetta and AlphaFold2, are most consistent with experimental analysis of E41, R116, R299, G328, D334, and the N-terminal mutants. We therefore selected the AlphaFold2 model for mutation and ligand docking studies, as the orientation of the N-terminal peptide is close to the residues required for the ABHD5 membrane binding. Our structural modeling and dynamics simulations show that the canonical α helices and β strands of the ABHD5 protein are highly stable. The main variance in the structure originates from the insertion of helical regions, which are correlated to essential ABHD5 functions, such as membrane and ligand binding. The E41A and R116N mutations disturb the ABHD5 membrane binding by disrupting the hydrogen bonding network of several nearby lysine residues (K38 and K54). The mutation of G328E changes the electrostatic interactions of the surrounding residues, thereby affecting the activation of ATGL. Our study also reveals the ligand binding modes of three SPZ ligands to ABHD5. The results show that the SPZ ligands bind stably in the hydrophobic pocket formed by the insertion helices, α2-α6.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

RS and Y-MH designed the project. RS, SP, and Y-MH performed computational work and analyzed data. MS, HZ, LM-L, WR, GH, and JG performed experimental work. RS, SP, CK, JG, and Y-MH wrote the manuscript.

FUNDING

This work was supported by the WSU startup grant to Y-MH, the WSU postdoctoral fellows grant to CK, JG, and Y-MH, and the NIH grants (R01DK076629 to JG and CK and R01DK105963 to JG).

ACKNOWLEDGMENTS

We thank Wayne State University high performance computing center for the support of GaMD simulations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.935375/full#supplementary-material

REFERENCES

Altschul, S., Madden, T., Schaffer, A., Zhang, J. H., Zhang, Z., Miller, W., et al. (1998). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *FASEB J.* 12, 1326.

Bhattarai, A., Pawnikar, S., and Miao, Y. (2021). Mechanism of Ligand Recognition by Human ACE2 Receptor. *J. Phys. Chem. Lett.* 12, 4814–4822. doi:10.1021/acs.jpclett.1c01064

Boeszoermenyi, A., Nagy, H. M., Arthanari, H., Pillip, C. J., Lindermuth, H., Luna, R. E., et al. (2015). Structure of a CGI-58 Motif Provides the Molecular Basis of Lipid Droplet Anchoring. *J. Biol. Chem.* 290, 26361–26372. doi:10.1074/jbc.M115.682203

Cacelli, I., and Prampolini, G. (2007). Parametrization and Validation of Intramolecular Force Fields Derived from DFT Calculations. *J. Chem. Theory Comput.* 3, 1803–1817. doi:10.1021/ct700113h

Cao, Y., Qiu, T., Kathayat, R. S., Azizi, S.-A., Thorne, A. K., Ahn, D., et al. (2019). ABHD10 Is an S-Depalmitoylase Affecting Redox Homeostasis through Peroxiredoxin-5. *Nat. Chem. Biol.* 15, 1232–1240. doi:10.1038/s41589-019-0399-y

Case, D. A., Berryman, J. T., Betz, R. M., Cerutti, D. S., Cheatham, T. E., Darden, T. A., et al. (2015). *AMBER*. San Francisco: University of California.

Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., et al. (2021). The trRosetta Server for Fast and Accurate Protein Structure Prediction. *Nat. Protoc.* 16, 5634–5651. doi:10.1038/s41596-021-00628-9

Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and python Bindings. *J. Chem. Inf. Model.* 61, 3891–3898. doi:10.1021/acs.jcim.1c00203

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* 103, 8577–8593. doi:10.1063/1.470117

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

Fu, H., Chen, H., Blazhynska, M., Goulard Coderc de Lacam, E., Szczepaniak, F., Pavlova, A., et al. (2022). Accurate Determination of Protein:ligand Standard Binding Free Energies from Molecular Dynamics Simulations. *Nat. Protoc.* 17, 1114–1141. doi:10.1038/s41596-021-00676-1

Granneman, J. G., Moore, H.-P. H., Granneman, R. L., Greenberg, A. S., Obin, M. S., and Zhu, Z. (2007). Analysis of Lipolytic Protein Trafficking and Interactions in Adipocytes. *J. Biol. Chem.* 282, 5726–5735. doi:10.1074/jbc.M610580200

Granneman, J. G., Moore, H.-P. H., Krishnamoorthy, R., and Rathod, M. (2009). Perilipin Controls Lipolysis by Regulating the Interactions of AB-Hydrolase Containing 5 (Abhd5) and Adipose Triglyceride Lipase (Atgl). *J. Biol. Chem.* 284, 34538–34544. doi:10.1074/jbc.M109.068478

Hameduh, T., Haddad, Y., Adam, V., and Heger, Z. (2020). Homology Modeling in the Time of Collective and Artificial Intelligence. *Comput. Struct. Biotechnol. J.* 18, 3494–3506. doi:10.1016/j.csbj.2020.11.007

Hamelberg, D., Mongan, J., and McCammon, J. A. (2004). Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* 120, 11919–11929. doi:10.1063/1.1755656

Hirabayashi, T., Anjo, T., Kaneko, A., Senoo, Y., Shibata, A., Takama, H., et al. (2017). PNPLA1 Has a Crucial Role in Skin Barrier Function by Directing Acylceramide Biosynthesis. *Nat. Commun.* 8, 14609. doi:10.1038/ncomms14609

Huang, Y. M. (2021). Multiscale Computational Study of Ligand Binding Pathways: Case of P38 MAP Kinase and its Inhibitors. *Biophys. J.* 120, 3881–3892. doi:10.1016/j.bpj.2021.08.026

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

Izadi, S., and Onufriev, A. V. (2016). Accuracy Limit of Rigid 3-point Water Models. *J. Chem. Phys.* 145, 074501. doi:10.1063/1.4960175

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

Joung, I. S., and Cheatham, T. E. (2008). Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* 112, 9020–9041. doi:10.1021/jp8001614

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

Jumper, J., and Hassabis, D. (2022). Protein Structure Predictions to Atomic Accuracy with AlphaFold. *Nat. Methods* 19, 11–12. doi:10.1038/s41592-021-01362-6

Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., et al. (2018). Improvements to the APBS Biomolecular Solvation Software Suite. *Protein Sci.* 27, 112–128. doi:10.1002/pro.3280

Kintscher, U., Foryst-Ludwig, A., Haemmerle, G., and Zechner, R. (2020). The Role of Adipose Triglyceride Lipase and Cytosolic Lipolysis in Cardiac Function and Heart Failure. *Cell. Rep. Med.* 1, 100001. doi:10.1016/j.xcrm.2020.100001

Krautler, V., Van Gunsteren, W. F., and Hunenberger, P. H. (2001). A Fast SHAKE: Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* 22, 501–508. doi:10.1002/1096-987x(20010415)22:5<501::aid-jcc1021>3.0.co;2-v

Laio, A., and Parrinello, M. (2002). Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* 99, 12562–12566. doi:10.1073/pnas.202427399

Lass, A., Zimmermann, R., Haemmerle, G., Riederer, M., Schoiswohl, G., Schweiger, M., et al. (2006). Adipose Triglyceride Lipase-Mediated Lipolysis of Cellular Fat Stores Is Activated by CGI-58 and Defective in Chanarin-Dorfman Syndrome. *Cell. Metab.* 3, 309–319. doi:10.1016/j.cmet.2006.03.005

Lass, A., Zimmermann, R., Oberer, M., and Zechner, R. (2011). Lipolysis - A Highly Regulated Multi-Enzyme Complex Mediates the Catabolism of Cellular Fat Stores. *Prog. Lipid Res.* 50, 14–27. doi:10.1016/j.plipres.2010.10.004

Lefèvre, C., Jobard, F., Caux, F., Bouadjar, B., Karaduman, A., Heilig, R., et al. (2001). Mutations in CGI-58, the Gene Encoding a New Protein of the Esterase/lipase/thioesterase Subfamily, in Chanarin-Dorfman Syndrome. *Am. J. Hum. Genet.* 69, 1002–1012. doi:10.1086/324121

Loncharich, R. J., Brooks, B. R., and Pastor, R. W. (1992). Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates ofN-Acetylalanyl-N?-Methylamide. *Biopolymers* 32, 523–535. doi:10.1002/bip.360320508

Machado, M. R., and Pantano, S. (2020). Split the Charge Difference in Two! A Rule of Thumb for Adding Proper Amounts of Ions in MD Simulations. *J. Chem. Theory Comput.* 16, 1367–1372. doi:10.1021/acs.jctc.9b00953

Madden, T. L., Busby, B., and Ye, J. (2019). Reply to the Paper: Misunderstood Parameters of NCBI BLAST Impacts the Correctness of Bioinformatics Workflows. *Bioinformatics* 35, 2699–2700. doi:10.1093/bioinformatics/bty1026

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

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. Theory Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Marx, V. (2022). Method of the Year 2021: Protein Structure Prediction. *Nat. Methods* 19, 1. doi:10.1038/s41592-021-01380-4

McCammon, J. A., Gelin, B. R., and Karplus, M. (1977). Dynamics of Folded Proteins. *Nature* 267, 585–590. doi:10.1038/267585a0

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

Miao, Y., Huang, Y. M. M., Walker, R. C., McCammon, J. A., and Chang, C. E. A. (2018). Ligand Binding Pathways and Conformational Transitions of the HIV Protease. *Biochemistry* 57, 1533–1541. doi:10.1021/acs.biochem.7b01248

Miao, Y., and McCammon, J. A. (2018). Mechanism of the G-Protein Mimetic Nanobody Binding to a Muscarinic G-Protein-Coupled Receptor. *Proc. Natl. Acad. Sci. U.S.A.* 115, 3036–3041. doi:10.1073/pnas.1800756115

Mindrebo, J. T., Nartey, C. M., Seto, Y., Burkart, M. D., and Noel, J. P. (2016). Unveiling the Functional Diversity of the Alpha/beta Hydrolase Superfamily in the Plant Kingdom. *Curr. Opin. Struct. Biol.* 41, 233–246. doi:10.1016/j.sbi.2016.08.005

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* 30, 2785–2791. doi:10.1002/jcc.21256

Nardini, M., and Dijkstra, B. W. (1999). α/β Hydrolase Fold Enzymes: the Family Keeps Growing. *Curr. Opin. Struct. Biol.* 9, 732–737. doi:10.1016/s0959-440x(99)00037-8

Nierzwicki, Ł., Arantes, P. R., Saha, A., and Palermo, G. (2021). Establishing the Allosteric Mechanism in CRISPR-Cas9. *WIREs Comput. Mol. Sci.* 11, e1503. doi:10.1002/wcms.1503

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., et al. (1992). The α/β Hydrolase Fold. *Protein Eng. Des. Sel.* 5, 197–211. doi:10.1093/protein/5.3.197

Özpınar, G. A., Peukert, W., and Clark, T. (2010). An Improved Generalized AMBER Force Field (GAFF) for Urea. *J. Mol. Model.* 16, 1427–1440. doi:10.1007/s00894-010-0650-7

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al. (2013). GROMACS 4.5: a High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* 29, 845–854. doi:10.1093/bioinformatics/btt055

Roe, D. R., and Cheatham, T. E. (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

Rondini, E. A., Mladenovic-Lucas, L., Roush, W. R., Halvorsen, G. T., Green, A. E., and Granneman, J. G. (2017). Novel Pharmacological Probes Reveal ABHD5 as a Locus of Lipolysis Control in White and Brown Adipocytes. *J. Pharmacol. Exp. Ther.* 363, 367–376. doi:10.1124/jpet.117.243253

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction. *Nat. Protoc.* 5, 725–738. doi:10.1038/nprot.2010.5

Sanders, M. A., Madoux, F., Mladenovic, L., Zhang, H., Ye, X., Angrish, M., et al. (2015). Endogenous and Synthetic ABHD5 Ligands Regulate ABHD5-Perilipin Interactions and Lipolysis in Fat and Muscle. *Cell. Metab.* 22, 851–860. doi:10.1016/j.cmet.2015.08.023

Sanders, M. A., Zhang, H., Mladenovic, L., Tseng, Y. Y., and Granneman, J. G. (2017). Molecular Basis of ABHD5 Lipolysis Activation. *Sci. Rep.* 7, 42589. doi:10.1038/srep42589

Sugita, Y., and Okamoto, Y. (1999). Replica-exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* 314, 141–151. doi:10.1016/s0009-2614(99)01123-9

Torrie, G. M., and Valleau, J. P. (1977). Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* 23, 187–199. doi:10.1016/0021-9991(77)90121-8

Trott, O., and Olson, A. J. (2009). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* 31, NA. doi:10.1002/jcc.21334

Tseng, Y. Y., Sanders, M. A., Zhang, H., Zhou, L., Chou, C.-Y., and Granneman, J. G. (2022). Structural and Functional Insights into ABHD5, a Ligand-Regulated Lipase Co-activator. *Sci. Rep.* 12, 2565. doi:10.1038/s41598-021-04179-7

Velez-Vega, C., and Gilson, M. K. (2013). Overcoming Dissipation in the Calculation of Standard Binding Free Energies by Ligand Extraction. *J. Comput. Chem.* 34, a–n. doi:10.1002/jcc.23398

Vieyres, G., Reichert, I., Carpentier, A., Vondran, F. W. R., and Pietschmann, T. (2020). The ATGL Lipase Cooperates with ABHD5 to Mobilize Lipids for Hepatitis C Virus Assembly. *PLoS Pathog.* 16, e1008554. doi:10.1371/journal.ppat.1008554

Wang, J., Arantes, P. R., Bhattarai, A., Hsu, R. V., Pawnikar, S., Huang, Y. M., et al. (2021). Gaussian Accelerated Molecular Dynamics: Principles and Applications. *WIREs Comput. Mol. Sci.* 11, e1521. doi:10.1002/wcms.1521

Wei, G.-W. (2019). Protein Structure Prediction beyond AlphaFold. *Nat. Mach. Intell.* 1, 336–337. doi:10.1038/s42256-019-0086-4

Yang, A., Mottillo, E. P., Mladenovic-Lucas, L., Zhou, L., and Granneman, J. G. (2019). Dynamic Interactions of ABHD5 with PNPLA3 Regulate Triacylglycerol Metabolism in Brown Adipocytes. *Nat. Metab.* 1, 560–569. doi:10.1038/s42255-019-0066-3

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* 12, 7–8. doi:10.1038/nmeth.3213

Youssefian, L., Vahidnezhad, H., Saeidian, A. H., Pajouhanfar, S., Sotoudeh, S., Mansouri, P., et al. (2019). Inherited Non-alcoholic Fatty Liver Disease and Dyslipidemia Due to Monoallelic A*BHD5* Mutations. *J. Hepatology* 71, 366–370. doi:10.1016/j.jhep.2019.03.026

Zhan, T., Cui, S., Liu, X., Zhang, C., Huang, Y. M., and Zhuang, S. (2021). Enhanced Disrupting Effect of Benzophenone-1 Chlorination Byproducts to the Androgen Receptor: Cell-Based Assays and Gaussian Accelerated Molecular Dynamics Simulations. *Chem. Res. Toxicol.* 34, 1140–1149. doi:10.1021/acs.chemrestox.1c00023

Zhang, J., and Madden, T. L. (1997). PowerBLAST: A New Network BLAST Application for Interactive or Automated Sequence Analysis and Annotation. *Genome Res.* 7, 649–656. doi:10.1101/gr.7.6.649

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 29 June 2022
doi: 10.3389/fmolb.2022.879212

Essential Dynamics Ensemble Docking for Structure-Based GPCR Drug Discovery

Kyle McKay[†], *Nicholas B. Hamilton*[†], *Jacob M. Remington, Severin T. Schneebeli and Jianing Li**

Department of Chemistry, University of Vermont, Burlington, VT, United States

The lack of biologically relevant protein structures can hinder rational design of small molecules to target G protein-coupled receptors (GPCRs). While ensemble docking using multiple models of the protein target is a promising technique for structure-based drug discovery, model clustering and selection still need further investigations to achieve both high accuracy and efficiency. In this work, we have developed an original ensemble docking approach, which identifies the most relevant conformations based on the essential dynamics of the protein pocket. This approach is applied to the study of small-molecule antagonists for the PAC1 receptor, a class B GPCR and a regulator of stress. As few as four representative PAC1 models are selected from simulations of a homology model and then used to screen three million compounds from the ZINC database and 23 experimentally validated compounds for PAC1 targeting. Our essential dynamics ensemble docking (EDED) approach can effectively reduce the number of false negatives in virtual screening and improve the accuracy to seek potent compounds. Given the cost and difficulties to determine membrane protein structures for all the relevant states, our methodology can be useful for future discovery of small molecules to target more other GPCRs, either with or without experimental structures.

Keywords: computer aided drug design, PAC1 receptor, antagonist, virtual screening, molecular dynamics, principal component analysis

OPEN ACCESS

Edited by:
Yinglong Miao,
University of Kansas, United States

Reviewed by:
Mihaly Mezei,
Icahn School of Medicine at Mount Sinai, United States
Shuguang Yuan,
Shenzhen Institutes of Advanced Technology (CAS), China

***Correspondence:**
Jianing Li
jianing.li@uvm.edu

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *19 February 2022*
Accepted: *18 May 2022*
Published: *29 June 2022*

Citation:
McKay K, Hamilton NB, Remington JM, Schneebeli ST and Li J (2022) Essential Dynamics Ensemble Docking for Structure-Based GPCR Drug Discovery. Front. Mol. Biosci. 9:879212. doi: 10.3389/fmolb.2022.879212

INTRODUCTION

Many G protein-coupled receptors (GPCRs) are being investigated as important therapeutic targets, but the success rate of structure-based drug design (SBDD) for GPCRs remains to be further improved (Hauser et al., 2017; Wootten et al., 2018; Odoemelam et al., 2020). One of the primary challenges is that the three-dimensional (3D) structures of most GPCRs have not been fully determined. Even with latest breakthroughs in protein structure prediction like AlphaFold (Jumper et al., 2021), the available structures may not represent the conformational states needed for accurate SBDD. The receptor (*ADCYAP1R1*, hereafter referred to as PAC1R) of the pituitary adenylate cyclase-activating peptide (PACAP), an emerging therapeutic target for stress-related disorders (Hammack et al., 2009; Ressler et al., 2011; Roman et al., 2014; Missig et al., 2017; Liao et al., 2019a), is a good example. Currently, the full-length PAC1R structures in the Protein Data Bank (PDB) are short isoforms (Uniprot ID: P41586-3) (Kobayashi et al., 2020; Liang et al., 2020), but the structures of the most prevalent long isoforms—PAC1null (Uniprot ID: P41586) or PAC1hop (Uniprot ID: P41586-2) — are still unavailable (Liao et al., 2019b). All the published structures of PAC1R are complexed with peptide agonists and a heterotrimeric G protein complex

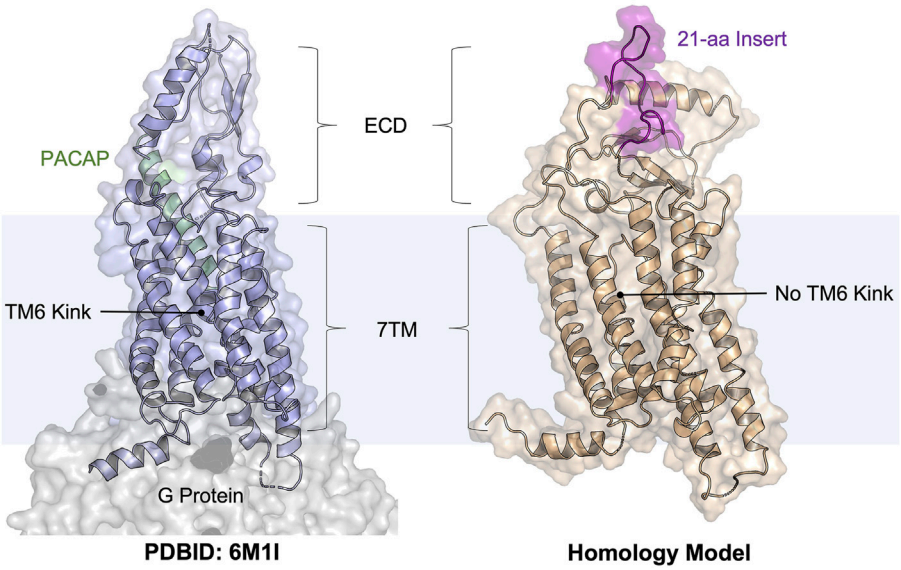


FIGURE 1 | Cartoon illustrations of the PACAP-bound PAC1R model (PDBID: 6M1I, PAC1very short) and our homology model (template PDBID: 4L6R, PAC1null, simulation snapshot at 500 ns). The PAC1null isoform is more biomedically relevant than the very short isoform. The PACAP peptide is shown as a helix cartoon (pale green); the 21-amino acid ECD insert (see the sequence in **Supplementary Figure S1**) is shown as a flexible coil (purple). This study focused on docking to the peptide-binding pocket.

(**Figure 1**), and thus do not represent the inactive conformations required for antagonist development. So far, it is thought that over 40% of GPCRs have more than one isoform (Marti-Solano et al., 2020), and each GPCR can adopt multiple conformational states which can be stabilized upon interactions with binding partners (Li et al., 2013; Vardy and Roth, 2013). For accurate SBDD, it is important to employ conformations of the most medically relevant isoform, as it is to this ensemble of 3D pocket structures that the drug must show affinity. Here, we used PAC1R as a model system and investigated how to improve modeling accuracy and to gain predictive power for SBDD with limited 3D structural information, using the method of Essential Dynamics Ensemble Docking (EDED). With the proof of principle, this method can be readily generalized to develop new therapeutic targets to target a wider range of GPCRs.

PAC1R and its endogenous peptide hormone PACAP play an important role in neural development, calcium homeostasis, glucose metabolism, circadian rhythm, thermoregulation, inflammation, feeding behavior, pain modulation, as well as stress, and related endocrine responses (Harmar et al., 2012; Bortolato et al., 2014; Culhane et al., 2015). For example, increased levels of PACAP in the blood have been reported in women diagnosed with post-traumatic stress disorder (Ressler et al., 2011), implicating chronic activation of the PAC1R in the disorder. Other studies (Boehr et al., 2009; Missig et al., 2017) have suggested that PAC1R activation mediates the adverse emotional consequences of chronic pain via downstream MAPK/ERK activation. Thus, these prior studies indicate that PAC1R antagonism, especially with small-molecule antagonists, represents a new strategy to treat stress, chronic pain, and related disorders (Ressler et al., 2011). Similar to other class B GPCRs, PAC1R possesses a heptahelical transmembrane domain (7TM) and an extracellular domain (ECD) (Odoemelam et al., 2020). Most of the neural and peripheral tissues known to date contain the PAC1null or PAC1hop isoforms that includes a 21-amino acid insert in the ECD (**Figure 1**), which is missing in available PAC1R structures in the PDB (May and Parsons, 2017). This ECD insert was found highly dynamic in our previous modeling studies (Liao et al., 2017; Liao et al., 2021), but its role in regulating PAC1R remains unknown. While PAC1R antagonists are being developed as potential treatments for stress-related disorders, the agonist-bound cryo-EM structures are not directly applicable to computational design or screening of PAC1R antagonists. GPCRs spontaneously adapt active and inactive signaling states, each of which are characterized by broad conformational ensembles. In the confirmational selection view, agonists and antagonists stabilize GPCR conformations of the active and inactive ensembles, respectively (Boehr et al., 2009; Abrol et al., 2013). It is now well accepted that to accurately design GPCR ligands as drug candidates, one should use active conformations for agonist design and inactive conformations for antagonist design. With the transition between active and inactive GPCR conformations occurring on the millisecond timescale (Vilardaga, 2010; Heyden et al., 2013; Weis and Kobilka, 2014; Scherer et al., 2015), it is computationally demanding to obtain the inactive PAC1R conformations from the agonist-bound cryo-EM structures via molecular dynamics (MD) simulations. Instead, we seek to use a homology model in this work and test with the EDED method.

Ensemble docking utilizes multiple receptor models for pocket sampling, obtained from clustering the conformations sampled by MD simulations for molecular docking, and displays noted

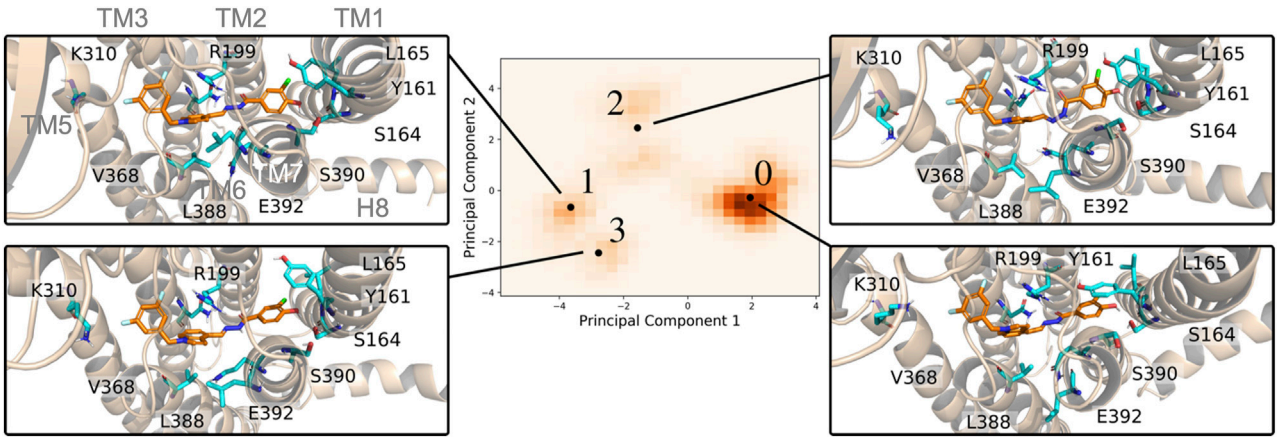


FIGURE 2 | Four representative PAC1R conformations using in EDED reveal important changes in the binding pocket. The protein is shown in cartoon and the reference ligand is shown as sticks. The histogram of all trajectory frames projected onto the first two principal components of residues within the ligand-binding pocket of PAC1R. Black dots labelled with numbers from 0 to 3 are the representative structures (S0, S1, S2, and S3) determined by the *K*-means clustering algorithm.

improvement at identifying GPCR ligands when compared to docking against a single experimental structure (Lin et al., 2002; Huang and Zou, 2007; Amaro et al., 2018; Velazquez et al., 2018; Li et al., 2019; Acharya et al., 2020; Bhattarai et al., 2020; Chandak et al., 2020; Jukič et al., 2020; Patel et al., 2021; Li et al., 2022). EDED is distinct from prior ensemble docking approaches, mainly in clustering and selection of receptor models. Global root mean square deviation (RMSD) is convenient to cluster similar structures, but the highly dynamic extracellular and intracellular loops (ECLs and ICLs) of GPCRs can significantly compromise the otherwise good similarity between the 7TM structures. Thus, clustering based on global RMSD can generate many models that, while representative of global changes, are irrelevant to the intricate differences within the local binding pocket of the GPCR. This additional overhead ultimately lowers both the efficiency and accuracy of ensemble docking when using the global RMSD approach for clustering. EDED avoids this issue by focusing on both local similarity and essential dynamics of the binding pocket. Although computational power is more accessible than ever, streamlined workflows which expend computational resources only on worthwhile calculations are always desirable. Herein, we applied EDED to PAC1R with as few as four receptor models, whose results show a reduced false negative rate and a good correlation between the small molecule efficacy and the predicted score. Our results provide the evidence for initial success to develop small-molecule antagonists for PAC1R and pave the way for future structure-based GPCR drug discovery.

RESULTS AND DISCUSSION

Inactive Conformations of PAC1null and Key Interactions With Small Molecules

Towards discovery of novel PAC1R antagonists, the inactive state conformational ensemble of PAC1R was estimated using an all-atom MD simulation of a ligand-bound PAC1R model from homology modeling (**Figure 1**). Our reference ligand is an analog of known PAC1 antagonists (Beebe et al., 2008) that were discovered previously using structure activity relationships. We created the antagonist-bound model by docking the reference compound into the PAC1R homology model. This complex model was simulated in the POPC membrane for 500 ns, and for the entire length of the simulation the ligand remained bound in roughly the starting conformation (**Supplementary Figure S2**). Other features like the closed ECD and straight transmembrane helix six (TM6), as well as short separation between TM3, TM5 and TM6, are consistent with a deactivated structure of a class B GPCR (Wu et al., 2020).

Despite the overall appearance of an inactivated receptor, there were critical changes within the orthosteric pocket during the MD simulations. Using EDED, four members of the inactive conformational ensemble (states S0, S1, S2, and S3 ordered by observed population) were extracted and reveal distinct conformations of the 7TM helices and different side chain orientations within the binding pocket (**Figure 2**). For one, bending of TM1 was observed to follow S0 < S2 < S3 < S1, where the most populated state (S0) was the most straightened helix. This correlated with local changes to residues Y161, L165, and S164 on TM1, and most significantly the stiffened TM1 in the S0 state enabled both π-stacking (with Y161) and hydrogen bonding (with S164) interactions. On the other hand, displacement of TM7 in the S1 state relative to S0 caused replacement of the hydrogen bond with S164 in the S0 state with a new hydrogen bond with S390. The interactions between the indole on the ligand and V368, L388, and E392 were modulated between the different receptor states with generally tighter interactions in the S1 and S3 states, in comparison with the S0 and S2 states. In addition, changes in TM5 affected the ability of K310 to form the stable interactions with the electron rich substituents on the ligand in states S0 and S3 which were diminished in the S1 and S2 states. Ultimately, this analysis reveals how EDED is able capture the subtle changes in pocket structure that are highly relevant for accurate modeling of ligand-receptor interactions when performing SBDD.

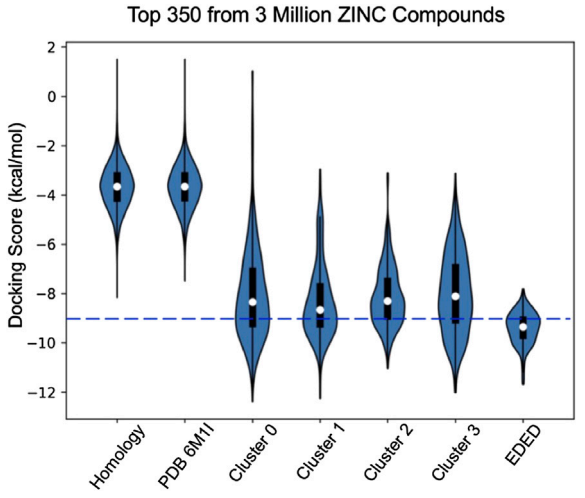


FIGURE 3 | Violin plots of the docking score distribution of the top 350 compounds to different receptor models. The dash line shows the –9.0 kcal/mol cutoff used to prioritize compounds for synthesis.

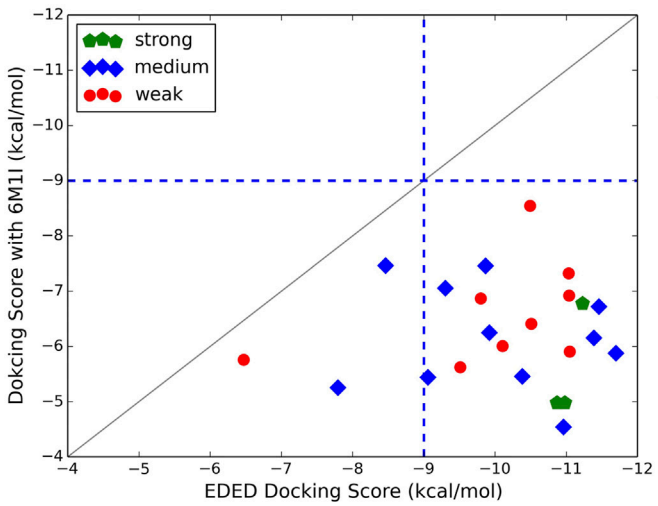
Comparison of Docking to a Single Receptor Model and to the Conformational Representatives

Compared with docking to the ligand-free homology model and ligand-free cryo-EM structure, EDED significantly improved the identification of candidate compounds (**Figure 3**). The average binding score of the top 350 (approximately 2.5%) of compounds docked to the ligand-free homology model improved from −5.9 to −9.4 kcal/mol when docked against the ensemble. Likewise, it improved from an average of −5.8 to −9.4 kcal/mol when compared to the PACAP-bound model (PDBID: 6M1I). This gives an average 3.6 kcal/mol improvement in average docking score of the top selected compounds. Additionally, EDED identified six compounds predicted to bind to PAC1R with comparable binding score (−11.2 kcal/mol) as our reference ligand**.**

To gain physical insight into the improvement of the docking scores, the binding pose(s) of the top compounds from both methodologies were examined. We have previously reported the key role of R199 in PACAP-induced activation of PAC1R (Liang et al., 2020). This is further corroborated by strong cation-pi interactions with the residue in our models. Interaction with R199 across all the ensemble conformations became a critical determining factor for which top ensemble docking compounds should be prioritized for synthesis and/or computational optimization. Examining the compounds which have ensemble docking scores close to or better than our reference ligand, this interaction is present for all six top scoring ligands in *at least* one of the docking poses. This is in contrast with the homology model and the PACAP-bound model where only relatively few of the top compounds from this methodology were able to engage in this key interaction. Also, of note are induced fit effects where the MD simulation of our reference ligand in the pocket may affect the binding pocket through subtle shifts in the backbone and the rotation of side chains. In the rigid receptor docking to the homology model, the 7TM helical bundle is closer together, defining a more compact orthosteric pocket. Thus, it is only accessible for small ligands to bind deep into the pocket below R199. In contrast, the conformations in the ensemble docking are more open, better allowing ligands to access the pocket. This can be seen by where most ligands found their best pose. Although both datasets were docked against a grid centered on R199, the ensemble docking results have the majority of top ligands below the residue, low in the pocket. When docked against the homology model, the top ligands are higher in the pocket at lowest in line with R199.

The new ligands examined within the orthosteric pocket showcased the ability of ensemble docking to provide integral confirmations omitted by static modelling, with the ensemble approach providing key ligand poses corresponding to interactions with new side chains revealed in the ensemble. Aside from R199, several key contacts were discovered from study of the top ligands bound to each receptor in the ensemble (**Supplementary Figure S3**). These contacts expand the understanding of the orthosteric pocket dynamics and can be exploited in small molecule rational design. In comparison with consistent interactions to the ligand-binding pocket of the homology model, these results suggests that EDED may reveal new crucial ligand-receptor interactions even from a rigid template.

A thermodynamically driven approach to scoring the binding poses of a given compound to multiple receptor structures was used to assess the binding affinity of the docked ligands. This approach quantitatively captures various physical phenomena that are often considered when computing overall docking scores: 1) the relative likelihood of the receptor obtaining the different conformations are explicitly included, and 2) the binding of the ligand to the receptor changes the energies of the complex differentially in the distinct conformations. Importantly, this model properly handles confounding cases that other approaches, such as a simple direct averaging of different docking scores, would not describe well. For instance, for any given ligand, a protein is hypothetically able to adopt an unlikely conformation ($\Delta G\_{\mathrm{conf}_{1,i}} \gg 0$, i.e., much higher relative free energy than the structure with lowest relative free energy) where the binding of the ligand to the protein could be quite favorable ($-\Delta G_{bind_i}$ approximately equal to 10 kT). Simply including this state in an average of docking scores would treat it as equivalently important as conformations that are far more relevant to the signaling states of the protein. Our approach avoids such errors, by including the energetics of the receptor confirmations, assuring that the overall energy of these rare states is indeed still relatively high and do not contribute significantly to the final score in **Eq. 1**. In sum, our docking score considers the difference in overall energies of the bound receptor conformations and is appropriate for comparison with a physical experiment that is unlikely to be able to distinguish between different bound conformations (**Eq. 1**).



Experimental Positives: 14
Experimental Negatives: 9

	EDED	6M1I
TP	12	0
TN	1	9
FP	8	0
FN	2	14
FPR	89%	0%
FNR	14%	100%
ACC	57%	0%

FIGURE 4 | Ensemble weighted glide scores (ΔG_{bind}) of 23 experimentally tested compounds. Compounds with strong, modest, and poor ERK inhibitive activity are depicted in green, blue, and red, respectively. Corresponding colored lines represent the average ensemble weighted glide score for that category. A cutoff of –9 kcal/mol was applied for predicted antagonists to be compared to their experimental results showing either strong or medium inhibition (active) or weak inhibition (inactive).

$$e^{\Delta G/kT} = e^{(G_{bound} - G_{unbound})/kT} = \frac{P_{unbound}}{P_{bound}} = \frac{\sum_{i=1}^{n} P_{cluster\ i}}{\sum_{i=1}^{n} P_{complex\ i}} \quad (1)$$

Where G_{bound} and $G_{unbound}$ are the energies associated with the ligand being bound or unbound to any receptor conformation, respectively, P_{bound} and $P_{unbound}$ are the total probabilities of the ligand being bound or unbound to any receptor conformation in the ensemble, respectively, $P_{cluster\ i}$ is the probability of a specific receptor conformation (calculated from the MD, see SI for more information), and $P_{complex\ i}$ is the probability of the ligand being bound to that specific ensemble conformation. We note that our model is still more appropriate than equal weighting for cases where one does not trust the relative energies of the different conformations obtained directly from the MD simulations. In such cases setting the $\Delta G\_conf_{1,i}$ to 0 for each conformation (i.e., each conformation is equally likely) reduces **Eqs 1–2**.

$$\Delta G_{bind,equal\ weighting} = \ln\left(\frac{n}{\sum_{i=1}^{n} e^{-\Delta G_{bind_i}/kT}}\right) kT \quad (2)$$

Clearly, **Eq. 2** is not a simple weighted average of the different binding scores, however to our knowledge this analysis is lacking in the literature.

Evaluation of EDED Predictions

Additional to testing EDED with compounds from ZINC, we also tested 23 small-molecule compounds which were classified as strong, moderate, and weak antagonists in PAC1R activity assays (unpublished data from Prof. Victor May). The design of these small molecules was based on previously published work outlining the structure-activity relationship between small molecules and the PAC1 receptor (Beebe et al., 2008). Ligand-based virtual screening was then performed and yielded the 23 compounds which were experimentally tested. Docking each analog against all four conformations in the ensemble and scoring them as previous described (**Eq. 1**) shows modest correlation to experimental results (**Figure 4**). The strong experimental antagonist had the highest predicted binding affinities with an average −10.4 kcal/mol, while the moderate and weak antagonists both had worse predicted binding affinities −9.8 kcal/mol and −8.5 kcal/mol, respectively.

It is worth noting that our EDED method is best used to identify potential antagonists from a collection of compounds, but the dockings scores (like Glide SP, XP, and our EDED score) to estimate binding energies should be interpreted with caution (Elokely and Doerksen, 2013; Pantsar and Poso, 2018; Pinzi and Rastelli, 2019). While we successfully reduced the false negative rate (FNR) with EDED, there is still a high false positive rate (FPR). A delicate balance between ensemble size and the FPR has previously been reported, inspiring us to select a relatively small ensemble for analysis (Mohammadi et al., 2022). Our FPR is comparable to prior studies employing both ensemble and static methods for virtual screening (Ferreira et al., 2010; Deng et al., 2015; Hou et al., 2018). Additionally, the experimental assays provided here are a measure of antagonistic ability, and not binding affinity. As quantitative binding assays remain to be performed, it is possible some of the false positives (compounds with poor experimental results but high ensemble docking scores) bind tightly but are not effective antagonists, i.e., they do not stabilize the inactive conformations or prevent cognate ligand binding in other ways. With the extended view provided by EDED, we envision that the chance of obtaining a false negative prediction is likely reduced in our model when compared with static Glide docking. This added width within the sampled energy

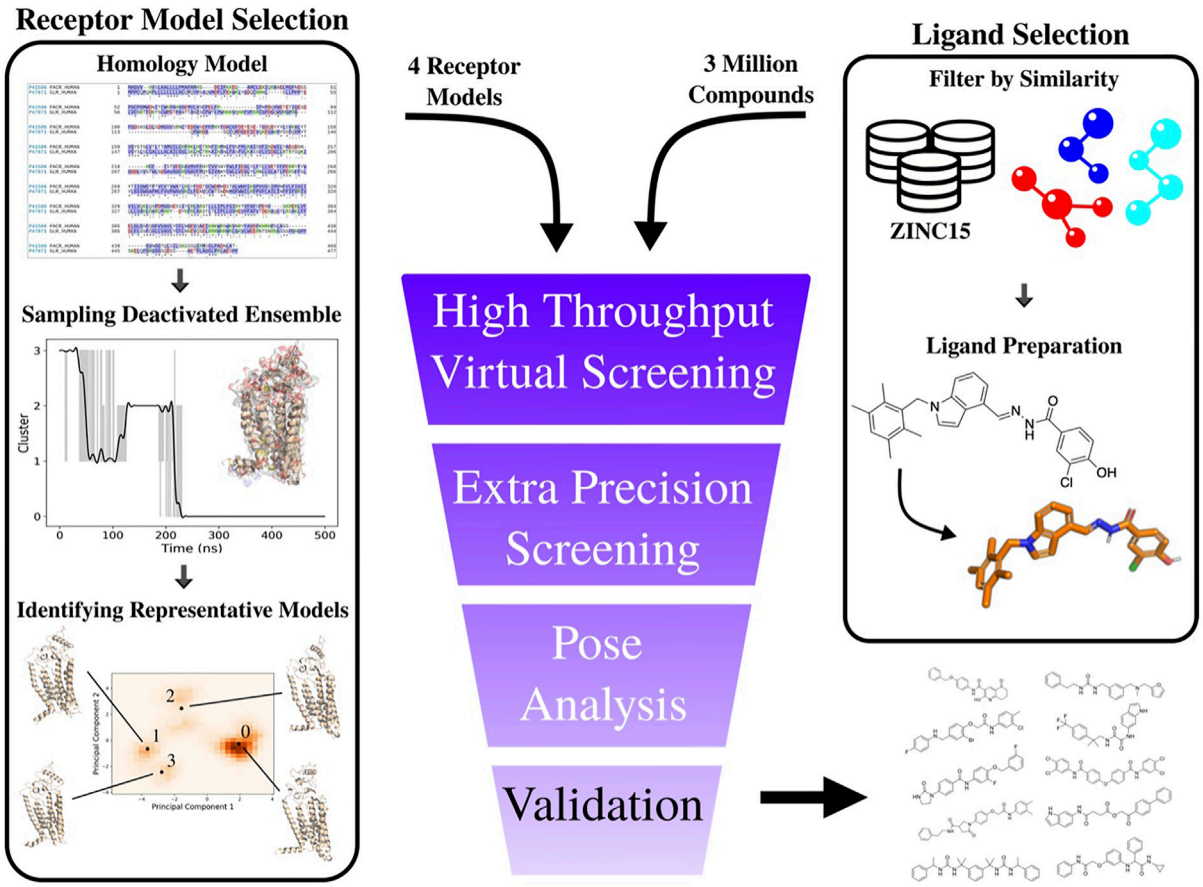


FIGURE 5 | Overview of computational workflow for development of PAC1R antagonists. Right Column: selection of input ligands from a structure database (in this example the ZINC15 (Sterling and Irwin, 2015) database). Custom filters were used to select raw structures with desirable properties (molecular weight, log*P*, etc.). These structures are then prepared using Schrödinger's *ligprep* software program. Left column): the PAC1null homology model is constructed from the protein's sequence, simulated for 500 ns, and the raw coordinates are analyzed. The representative structures are used in ensemble docking. Hit compounds are selected based on visual inspection of the results.

landscape (from the new side chain confirmations) allows our EDED method to achieve more accurate sampling of potential ligand-receptor interactions, thus increasing the chances of finding a hit compound otherwise overlooked in the static model. Overall, EDED displayed an accuracy of 57% in predicted binding affinity when compared to our experimental results, an increase when compared with Glide's empirical scoring function (Adeshina et al., 2020). Combined with the overall low variance in EDED docking scores for the top 350 compounds analyzed (**Figure 3**), we believe our methodology represents a robust route for the recognition of small molecules with high receptor affinity.

CONCLUSION

In conclusion, we have developed and implemented EDED, an ensemble docking inspired methodology for SBDD. By focusing on the essential dynamics of the ligand binding pocket, our method is distinct from many prior studies that built receptor clusters solely based on the root mean square deviation (RMSD) of the entire protein backbone (Kufareva and Abagyan, 2012). Further, the use of clustering within this reduced dimensionality conformational space directly considers the local structural similarity of the ligand-binding pocket. We demonstrate that EDED captures the critical changes in the 3D structure of the binding pocket that are known to correlate strongly with binding affinity of ligands. Our approach is partially based on the assumption that differences in the binding pocket itself (as opposed to the protein as a whole) predominately give rise to the different binding poses and energies that are the goal of any ensemble docking workflow. Using the EDED derived representative structures, we screened a large dataset of compounds and successfully identified novel small molecule antagonists of the PAC1 receptor. However, EDED is not specific to a single GPCR and will likely accelerate the design of small molecule drugs that target other GPCRs with currently unknown conformational states.

METHODS AND MODELS

Receptor Model Preparation in EDED

One key idea of EDED is to obtain chemically relevant receptor models for docking. Instead of using the agonist-bound PAC1R structure, we generated a homology model of inactive PAC1R (with the canonical variant sequence, Uniprot ID: P41586) with a template of the glucagon receptor (PDBID: 4L6R, ~40% similarity) (Boehr et al., 2009). This PAC1R model incorporated the inactive features of class B GPCRs such as a continuous helix along TM6 and a closed ECD. A small-molecule PAC1R antagonist, our reference ligand, was placed in the orthosteric pocket via molecular docking (Glide, Schrödinger Inc.). The complex model was later simulated to sample the inactive conformational ensemble.

Receptor Model Sampling in EDED

To sample inactive conformations for docking, the ligand-bound PAC1R model was simulated with the OPLS3 (Harder et al., 2016) force field in explicit SPC solvent in the NPT ensemble (300K, 1 atm, Martyna-Tuckerman Klein coupling scheme) using classical MD simulations. A POPC membrane was place around the 7TM using the Orientations of Proteins in Membrane (OPM) database (Lomize et al., 2012). The simulation was performed in the Maestro-Desmond program (Bowers et al., 2006) (GPU version 5.4) with a timestep of 2 fs, recording interval of 4.8 ps, and a total simulation time of 500 ns The Ewald technique was used for the electrostatic calculations. The van der Waals and short-range electrostatic interactions were cut off at 9 Å. Hydrogen atoms were constrained using the SHAKE algorithm. Two extended simulations were also examined to confirm the ligand poses and receptor confirmations. Once again, a POPC membrane was placed around the 7TM bundle using OPM. NAMD 2.11 was used as the simulation package for these replicates (Phillips et al., 2020). The CHARMM36 forcefield (Lee et al., 2016) was used with a TIP3 solvent model in a NPT ensemble (310 K, 1 atm) Force switching was utilized at the range of 10–12 Å to approximate the LJ interactions. Langevin piston/Nose-Hoover (Martyna et al., 1994; Feller et al., 1995) methods were utilized for the pressure control with a piston period of 50 fs and a decay time of 25 fs Langevin coupling of these simulations with a dampening coefficient of 1 ps^{-1} was also utilized. Long range electrostatic interactions were modeled with the particle mesh Ewald method (Essmann et al., 1995). These simulations were run with a 2 fs timestep and combined for 350 ns of data. MD trajectories were analyzed using in-house Python and TCL scripts as well as Visual Molecular Dynamics (VMD). (Humphrey et al., 1996).

Receptor Ensemble Selection in EDED

We first aligned the 7TM of PAC1R (residues 156–405) to the homology model to reduce noise due to translational movement. Next, the coordinates of the centers of mass for any residue whose side chain was within 3 Å of any ligand atom in the static model were collected and parsed using in-house designed TCL and python scripts. A dimension reduction based on principal component analysis (PCA) was used to determine which collective motions (termed principal components, PCs) contributed most to variations in the overall conformations of the binding pocket. The first fifteen PCs (accounting for 90% of the cumulative variance) were clustered using a *K*-means clustering algorithm implemented by PyEmma (Liao et al., 2021). Based on inspection of the first two PCs (**Figure 5**), four cluster centers were identified. As these cluster centers are not precise frames within the trajectory but are instead points in the PC space, the cluster centers' PC coordinates were approximately projected back to the original Cartesian coordinates. Frames from the trajectory which had PC values closest to the centers based on a RMSD measurement, were then selected as the ensemble docking receptor structures. This approach allowed a minimum of representative frames to capture the most variance of the binding pocket as opposed to other methodologies which often have many structures. Also, our physics-based approach is transferrable to other GPCRs and expanded clustering. In fact, our focus on the relevant receptor models likely requires less sampling in MD simulations and fewer clusters for subsequent docking, a practical advantage for large-scale screening.

Docking and Scoring of Potential PAC1R Antagonists

Receptor grid models were generated using the three-dimensional structures selected as detailed above with R199 selected as the center of the docking box with an 18-Å cutoff. Docking was carried out using Schrödinger Virtual Screening Workflow (Friesner et al., 2006) (VSW) at three consecutive levels of precision, both for small molecules docked to the static homology model and to the conformation ensemble. Small molecules docked to our PAC1null ensembled were given an overall score, Ensemble Δ G_{bind}, based on **Eq. 3**.

$$\text{Ensemble } \Delta G_{bind} = \ln\left(\frac{1+\sum_{i=2}^{n} e^{-\Delta G_{conf\,1,i}}}{\sum_{i=1}^{n} e^{-\Delta G_{conf\,(1,i)}} - \Delta G_{bind\,(1,i)/kT}}\right) kT \quad (3)$$

In **Eq. 3**, $\Delta G_{\text{conf}_{1,i}}$ is the difference in energy (in units of kT) between the lowest energy receptor conformation and each subsequent conformation calculated using the clustered trajectory, and $-\Delta G_{bind_i}$ is the corresponding Glide XP docking score to that same conformation. While $\Delta G_{\text{conf}_{1,i}}$ is representative of the apo receptor free energy, it is worth noting that simulation data used to generate these confirmations included the ligand bound within the pocket.

Docking was carried out against compounds 1) pseudo-randomly selected from the ZINC15 (Sterling and Irwin, 2015) database, 2) as analogs of known antagonists to the static ligand-free homology model, the cryo-EM structure, and the conformational ensemble. In total, a small test set of 10,000 drug-like compounds were selected and download from

the ZINC database and docked using Schrödinger's VSW as described previously.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

KM, JR, SS, and JL contributed to the conception and design of the study. KM performed the model preparation, simulation setup, docking, and docking analysis. KM and NH performed the simulation analysis. JR derived the EDED scoring function. KM, JR, and JL wrote the first draft of the manuscript. KM, JR, NH, SS, and JL wrote sections of the manuscript. All authors contributed to the revision, read, and approved the submitted version.

FUNDING

The work was mainly supported by the NIH grant R01-GM129431 to JL. SS. was partially supported by the Army Research Office (Grant 71015-CH-YIP). STS was supported by the U.S. Army Research Office (Grant 71015-CH-YIP). Part of the computational facilities was also supported by an NSF CAREER award (Grant CHE-1848444 awarded to STS).

ACKNOWLEDGMENTS

We thank Drs. Victor May and Matthias Brewer (UVM) for sharing the experimental data and for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.879212/full#supplementary-material

REFERENCES

Abrol, R., Kim, S.-K., Bray, J. K., Trzaskowski, B., and Goddard, W. A. (2013). "Chapter Two - Conformational Ensemble View of G Protein-Coupled Receptors and the Effect of Mutations and Ligand Binding," in *Methods in Enzymology*. Editor P. M. Conn (Academic Press), 520, 31–48. doi:10.1016/b978-0-12-391861-1.00002-2

Acharya, A., Agarwal, R., Baker, M. B., Baudry, J., Bhowmik, D., Boehm, S., et al. (2020). Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J. Chem. Inf. Model.* 60 (12), 5832–5852. doi:10.1021/acs.jcim.0c01010

Adeshina, Y. O., Deeds, E. J., and Karanicolas, J. (2020). Machine Learning Classification Can Reduce False Positives in Structure-Based Virtual Screening. *Proc. Natl. Acad. Sci. U.S.A.* 117 (31), 18477–18488. doi:10.1073/pnas.2000585117

Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J. A., Miao, Y., et al. (2018). Ensemble Docking in Drug Discovery. *Biophys. J.* 114 (10), 2271–2278. doi:10.1016/j.bpj.2018.02.038

Beebe, X., Darczak, D., Davis-Taber, R. A., Uchic, M. E., Scott, V. E., Jarvis, M. F., et al. (2008). Discovery and SAR of Hydrazide Antagonists of the Pituitary Adenylate Cyclase-Activating Polypeptide (PACAP) Receptor Type 1 (PAC1-R). *Bioorg. Med. Chem. Lett.* 18 (6), 2162–2166. doi:10.1016/j.bmcl.2008.01.052

Bhattarai, A., Wang, J., and Miao, Y. (2020). Retrospective Ensemble Docking of Allosteric Modulators in an Adenosine G-Protein-Coupled Receptor. *Biochim. Biophys. Acta (BBA) - General Subj.* 1864 (8), 129615. doi:10.1016/j.bbagen.2020.129615

Boehr, D. D., Nussinov, R., and Wright, P. E. (2009). The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* 5 (11), 789–796. doi:10.1038/nchembio.232

Bortolato, A., Doré, A. S., Hollenstein, K., Tehan, B. G., Mason, J. S., and Marshall, F. H. (2014). Structure of Class B GPCRs: New Horizons for Drug Discovery. *Br. J. Pharmacol.* 171 (13), 3132–3145. doi:10.1111/bph.12689

Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). "*Scalable Algorithms For Molecular Dynamics Simulations On Commodity Clusters*, SC '06," in Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, 11-17 Nov. 2006, 43.

Chandak, T., Mayginnes, J. P., Mayes, H., and Wong, C. F. (2020). Using Machine Learning to Improve Ensemble Docking for Drug Discovery. *Proteins* 88 (10), 1263–1270. doi:10.1002/prot.25899

Culhane, K. J., Liu, Y., Cai, Y., and Yan, E. C. (2015). Transmembrane Signal Transduction by Peptide Hormones via Family B G Protein-Coupled Receptors. *Front. Pharmacol.* 6, 264. doi:10.3389/fphar.2015.00264

Deng, N., Forli, S., He, P., Perryman, A., Wickstrom, L., Vijayan, R. S. K., et al. (2015). Distinguishing Binders from False Positives by Free Energy Calculations: Fragment Screening Against the Flap Site of HIV Protease. *J. Phys. Chem. B* 119 (3), 976–988. doi:10.1021/jp506376z

Elokely, K. M., and Doerksen, R. J. (2013). Docking Challenge: Protein Sampling and Molecular Docking Performance. *J. Chem. Inf. Model.* 53 (8), 1934–1945. doi:10.1021/ci400040d

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* 103 (19), 8577–8593. doi:10.1063/1.470117

Feller, S. E., Zhang, Y., Pastor, R. W., and Brooks, B. R. (1995). Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* 103 (11), 4613–4621. doi:10.1063/1.470648

Ferreira, R. S., Simeonov, A., Jadhav, A., Eidam, O., Mott, B. T., Keiser, M. J., et al. (2010). Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* 53 (13), 4891–4905. doi:10.1021/jm100488w

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., et al. (2006). Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes. *J. Med. Chem.* 49 (21), 6177–6196. doi:10.1021/jm0512560

Hammack, S. E., Cheung, J., Rhodes, K. M., Schutz, K. C., Falls, W. A., Braas, K. M., et al. (2009). Chronic Stress Increases Pituitary Adenylate Cyclase-Activating Peptide (PACAP) and Brain-Derived Neurotrophic Factor (BDNF) mRNA Expression in the Bed Nucleus of the Stria Terminalis (BNST): Roles for PACAP in Anxiety-Like Behavior. *Psychoneuroendocrinology* 34 (6), 833–843. doi:10.1016/j.psyneuen.2008.12.013

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. Theory Comput.* 12 (1), 281–296. doi:10.1021/acs.jctc.5b00864

Harmar, A. J., Fahrenkrug, J., Gozes, I., Laburthe, M., May, V., Pisegna, J. R., et al. (2012). Pharmacology and Functions of Receptors for Vasoactive Intestinal Peptide and Pituitary Adenylate Cyclase-Activating Polypeptide: IUPHAR Review 1. *Br. J. Pharmacol.* 166 (1), 4–17. doi:10.1111/j.1476-5381.2012.01871.x

Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B., and Gloriam, D. E. (2017). Trends in GPCR Drug Discovery: New Agents, Targets and Indications. *Nat. Rev. Drug Discov.* 16 (12), 829–842. doi:10.1038/nrd.2017.178

Heyden, M., Jardon-Valadez, H. E., Bondar, A.-N., and Tobias, D. J. (2013). GPCR Activation on the Microsecond Timescale in MD Simulations. *Biophysical J.* 104 (2Suppl. 1), 115a. doi:10.1016/j.bpj.2012.11.666

Hou, X., Rooklin, D., Yang, D., Liang, X., Li, K., Lu, J., et al. (2018). Computational Strategy for Bound State Structure Prediction in Structure-Based Virtual Screening: A Case Study of Protein Tyrosine Phosphatase Receptor Type O Inhibitors. *J. Chem. Inf. Model.* 58 (11), 2331–2342. doi:10.1021/acs.jcim.8b00548

Huang, S. Y., and Zou, X. (2007). Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins* 66 (2), 399–421. doi:10.1002/prot.21214

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graph.* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5

Jukič, M., Janežič, D., and Bren, U. (2020). Ensemble Docking Coupled to Linear Interaction Energy Calculations for Identification of Coronavirus Main Protease (3CLpro) Non-Covalent Small-Molecule Inhibitors. *Molecules* 25 (24). doi:10.3390/molecules25245808

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Kobayashi, K., Shihoya, W., Nishizawa, T., Kadji, F. M. N., Aoki, J., Inoue, A., et al. (2020). Cryo-EM Structure of the Human PAC1 Receptor Coupled to an Engineered Heterotrimeric G Protein. *Nat. Struct. Mol. Biol.* 27 (3), 274–280. doi:10.1038/s41594-020-0386-8

Kufareva, I., and Abagyan, R. (2012). Methods of Protein Structure Comparison. *Methods (Mol. Biol. Clift. N.J.)* 857, 231–257. doi:10.1007/978-1-61779-588-6_10

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 (1), 405–413. doi:10.1021/acs.jctc.5b00935

Li, D., Jiang, K., Teng, D., Wu, Z., Li, W., Tang, Y., et al. (2022). Discovery of New Estrogen-Related Receptor α Agonists via a Combination Strategy Based on Shape Screening and Ensemble Docking. *J. Chem. Inf. Model.* 62 (3), 486–497. doi:10.1021/acs.jcim.1c00662

Li, J., Jonsson, A. L., Beuming, T., Shelley, J. C., and Voth, G. A. (2013). Ligand-Dependent Activation and Deactivation of the Human Adenosine A2A Receptor. *J. Am. Chem. Soc.* 135 (23), 8749–8759. doi:10.1021/ja404391q

Li, X., Zhang, X. X., Lin, Y. X., Xu, X. M., Li, L., and Yang, J. B. (2019). Virtual Screening Based on Ensemble Docking Targeting Wild-Type P53 for Anticancer Drug Discovery. *Chem. Biodivers.* 16 (7), e1900170. doi:10.1002/cbdv.201900170

Liang, Y.-L., Belousoff, M. J., Zhao, P., Koole, C., Fletcher, M. M., Truong, T. T., et al. (2020). Toward a Structural Understanding of Class B GPCR Peptide Binding and Activation. *Mol. Cell* 77 (3), 656–668.e5. doi:10.1016/j.molcel.2020.01.012

Liao, C., de Molliens, M. P., Schneebeli, S. T., Brewer, M., Song, G., Chatenet, D., et al. (2019). Targeting the PAC1 Receptor for Neurological and Metabolic Disorders. *Curr. Top. Med. Chem.* 19, 1399–1417. doi:10.2174/1568026619666190709092647

Liao, C., Remington, J. M., May, V., and Li, J. (2021). Molecular Basis of Class B GPCR Selectivity for the Neuropeptides PACAP and VIP. *Front. Mol. Biosci.* 8, 644644. doi:10.3389/fmolb.2021.644644

Liao, C., May, V., and Li, J. (2019). PAC1 Receptors: Shapeshifters in Motion. *J. Mol. Neurosci.* 68 (3), 331–339. doi:10.1007/s12031-018-1132-0

Liao, C., Zhao, X., Brewer, M., May, V., and Li, J. (2017). Conformational Transitions of the Pituitary Adenylate Cyclase-Activating Polypeptide Receptor, a Human Class B GPCR. *Sci. Rep.* 7 (1), 5427. doi:10.1038/s41598-017-05815-x

Lin, J.-H., Perryman, A. L., Schames, J. R., and McCammon, J. A. (2002). Computational Drug Design Accommodating Receptor Flexibility: The Relaxed Complex Scheme. *J. Am. Chem. Soc.* 124 (20), 5632–5633. doi:10.1021/ja0260162

Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., and Lomize, A. L. (2012). OPM Database and PPM Web Server: Resources for Positioning of Proteins in Membranes. *Nucleic Acids Res.* 40, D370–D376. doi:10.1093/nar/gkr703

Marti-Solano, M., Crilly, S. E., Malinverni, D., Munk, C., Harris, M., Pearce, A., et al. (2020). Combinatorial Expression of GPCR Isoforms Affects Signalling and Drug Responses. *Nature* 587 (7835), 650–656. doi:10.1038/s41586-020-2888-2

Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* 101 (5), 4177–4189. doi:10.1063/1.467468

May, V., and Parsons, R. L. (2017). G Protein-Coupled Receptor Endosomal Signaling and Regulation of Neuronal Excitability and Stress Responses: Signaling Options and Lessons from the PAC1 Receptor. *J. Cell. Physiol.* 232 (4), 698–706. doi:10.1002/jcp.25615

Missig, G., Mei, L., Vizzard, M. A., Braas, K. M., Waschek, J. A., Ressler, K. J., et al. (2017). Parabrachial Pituitary Adenylate Cyclase-Activating Polypeptide Activation of Amygdala Endosomal Extracellular Signal-Regulated Kinase Signaling Regulates the Emotional Component of Pain. *Biol. Psychiatry* 81 (8), 671–682. doi:10.1016/j.biopsych.2016.08.025

Mohammadi, S., Narimani, Z., Ashouri, M., Firouzi, R., and Karimi-Jafari, M. H. (2022). Ensemble Learning from Ensemble Docking: Revisiting the Optimum Ensemble Size Problem. *Sci. Rep.* 12 (1), 410. doi:10.1038/s41598-021-04448-5

Odoemelam, C. S., Percival, B., Wallis, H., Chang, M.-W., Ahmad, Z., Scholey, D., et al. (2020). G-Protein Coupled Receptors: Structure and Function in Drug Discovery. *RSC Adv.* 10 (60), 36337–36348. doi:10.1039/d0ra08003a

Pantsar, T., and Poso, A. (2018). Binding Affinity via Docking: Fact and Fiction. *Molecules* 23 (8), 1899. doi:10.3390/molecules23081899

Patel, D., Athar, M., and Jha, P. C. (2021). Exploring Ruthenium-Based Organometallic Inhibitors Against Plasmodium Falciparum Calcium Dependent Kinase 2 (PfCDPK2): A Combined Ensemble Docking, QM/MM and Molecular Dynamics Study. *ChemistrySelect* 6 (32), 8189–8199. doi:10.1002/slct.202101801

Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* 153 (4), 044130. doi:10.1063/5.0014475

Pinzi, L., and Rastelli, G. (2019). Molecular Docking: Shifting Paradigms in Drug Discovery. *Int. J. Mol. Sci.* 20 (18), 4331. doi:10.3390/ijms20184331

Ressler, K. J., Mercer, K. B., Bradley, B., Jovanovic, T., Mahan, A., Kerley, K., et al. (2011). Post-Traumatic Stress Disorder Is Associated with PACAP and the PAC1 Receptor. *Nature* 470 (7335), 492–497. doi:10.1038/nature09856

Roman, C. W., Lezak, K. R., Hartsock, M. J., Falls, W. A., Braas, K. M., Howard, A. B., et al. (2014). PAC1 Receptor Antagonism in the Bed Nucleus of the Stria Terminalis (BNST) Attenuates the Endocrine and Behavioral Consequences of Chronic Stress. *Psychoneuroendocrinology* 47, 151–165. doi:10.1016/j.psyneuen.2014.05.014

Scherer, M. K., Trendelkamp-Schroer, B., Paul, F., Pérez-Hernández, G., Hoffmann, M., Plattner, N., et al. (2015). PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* 11 (11), 5525–5542. doi:10.1021/acs.jctc.5b00743

Sterling, T., and Irwin, J. J. (2015). ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* 55 (11), 2324–2337. doi:10.1021/acs.jcim.5b00559

Vardy, E., and Roth, B. L. (2013). Conformational Ensembles in GPCR Activation. *Cell* 152 (3), 385–386. doi:10.1016/j.cell.2013.01.025

Velazquez, H. A., Riccardi, D., Xiao, Z., Quarles, L. D., Yates, C. R., Baudry, J., et al. (2018). Ensemble Docking to Difficult Targets in Early-Stage Drug Discovery: Methodology and Application to Fibroblast Growth Factor 23. *Chem. Biol. Drug Des.* 91 (2), 491–504. doi:10.1111/cbdd.13110

Vilardaga, J.-P. (2010). Theme and Variations on Kinetics of GPCR Activation/Deactivation. *J. Recept. Signal Transduct.* 30 (5), 304–312. doi:10.3109/10799893.2010.509728

Weis, W. I., and Kobilka, B. K. (2014). The Molecular Basis of G Protein-Coupled Receptor Activation. *Annu. Rev. Biochem.* 87 (1), 897–919. doi:10.1146/annurev-biochem-060614-033910

Wootten, D., Christopoulos, A., Marti-Solano, M., Babu, M. M., and Sexton, P. M. (2018). Mechanisms of Signalling and Biased Agonism in G Protein-Coupled Receptors. *Nat. Rev. Mol. Cell Biol.* 19 (10), 638–653. doi:10.1038/s41580-018-0049-3

Wu, F., Yang, L., Hang, K., Laursen, M., Wu, L., Han, G. W., et al. (2020). Full-Length Human GLP-1 Receptor Structure Without Orthosteric Ligands. *Nat. Commun.* 11 (1), 1272. doi:10.1038/s41467-020-14934-5

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.

frontiers | Frontiers in Molecular Biosciences

BRIEF RESEARCH REPORT
published: 04 July 2022
doi: 10.3389/fmolb.2022.922361

Changes in Protonation States of In-Pathway Residues can Alter Ligand Binding Pathways Obtained From Spontaneous Binding Molecular Dynamics Simulations

*Helena Girame, Marc Garcia-Borràs and Ferran Feixas**

Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, Girona, Spain

Protein-ligand binding processes often involve changes in protonation states that can be key to recognize and orient the ligand in the binding site. The pathways through which (bio) molecules interplay to attain productively bound complexes are intricate and involve a series of interconnected intermediate and transition states. Molecular dynamics (MD) simulations and enhanced sampling techniques are commonly used to characterize the spontaneous binding of a ligand to its receptor. However, the effect of protonation state changes of in-pathway residues in spontaneous binding MD simulations remained mostly unexplored. Here, we used molecular dynamics simulations to reconstruct the trypsin-benzamidine binding pathway considering different protonation states of His57. This residue is part of the trypsin catalytic triad and is located more than 10 Å away from Asp189, which is responsible for benzamidine binding in the trypsin S1 pocket. Our MD simulations showed that the binding pathways that benzamidine follow to target the S1 binding site are critically dependent on the His57 protonation state. Binding of benzamidine frequently occurs when His57 is protonated in the delta nitrogen while the binding process is significantly less frequent when His57 is positively charged. Constant-pH MD simulations retrieved the equilibrium populations of His57 protonation states at trypsin active pH offering a clearer picture of benzamidine recognition and binding. These results indicate that properly accounting for protonation states of distal residues can be important in spontaneous binding MD simulations.

Keywords: ligand binding pathways, protonation states, spontaneous binding simulations, constant-pH molecular dynamics, trypsin-benzamidine complex

OPEN ACCESS

Edited by:
Chia-en A. Chang, University of California, Riverside, United States

Reviewed by:
Wenping Lyu, The Chinese University of Hong Kong, Shenzhen, China
Yandong Huang, Jimei University, China

****Correspondence:***
Ferran Feixas
ferran.feixas@udg.edu

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *17 April 2022*
Accepted: *14 June 2022*
Published: *04 July 2022*

Citation:
Girame H, Garcia-Borràs M and Feixas F (2022) Changes in Protonation States of In-Pathway Residues can Alter Ligand Binding Pathways Obtained From Spontaneous Binding Molecular Dynamics Simulations. Front. Mol. Biosci. 9:922361. doi: 10.3389/fmolb.2022.922361

INTRODUCTION

Characterizing the mechanisms of ligand binding and unbinding to a biomolecule is crucial to elucidate the molecular basis of biological processes and improve the potency of drugs (Bernetti et al., 2019). The pathways through which (bio)molecules interplay to attain stable (and often transient) bound complexes are intricate and involve a series of interconnected intermediate, misbound, and transition states. Molecular dynamics (MD) simulations and enhanced sampling techniques are frequently used to characterize the spontaneous binding pathways of drugs, substrates, or peptides to its biological receptors (Dror et al., 2011; Shan et al., 2011; Decherchi and Cavalli, 2020). In these

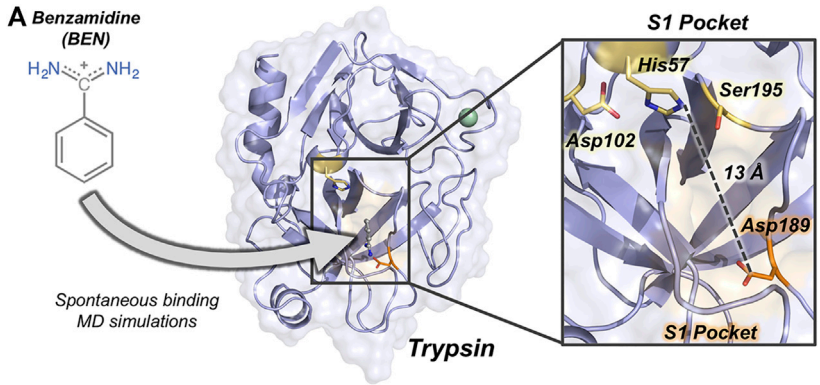


FIGURE 1 | Overview of trypsin structure and S1 pocket. **(A)** Trypsin (in purple) structure corresponding to PDB ID 3PTB with benzamidine inhibitor (in grey) bound to the S1 pocket (orange surface). His57 and Asp189 residues are highlighted in yellow and in orange sticks, respectively. Calcium ion is depicted as a green sphere. Overview of the S1 pocket with catalytic residues (Asp102, His57, and Ser195) shown in yellow sticks. The distance between the carbon of the side-chain of Asp189 and the side-chain of His57 is 13 Å.

simulations, one or more ligands are commonly placed in the solvent and are allowed to freely diffuse without biasing the MD simulation toward a particular protein region. Providing sufficient simulation time and an accurate description of the system, the ligand freely explores the dynamic protein surface until it spontaneously finds its presumed binding site (Betz and Dror, 2019). Markov-State Models were used to completely reconstruct ligand binding and unbinding pathways and the associated kinetics of enzyme-inhibitor complexes (Buch et al., 2011; Plattner and Noé, 2015). Unconstrained enhanced sampling methods were used to simulate binding of allosteric modulators into G-protein coupled receptors (Miao and McCammon, 2016) or substrate binding in allosterically regulated enzymes (Calvó-Tusell et al., 2022). The predictive power of spontaneous binding simulations relies on being able to access the timescale required to sample the binding event and critically depends on the accurate description of the simulated system.

Around 60% of protein-ligand binding events involve changes in protonation states (Aguilar et al., 2010; Onufriev and Alexov, 2013). Properly accounting for protonation states of protein residues is crucial to characterize ligand binding with MD simulations. The prediction of protonation states from rigid X-ray structures can lead to their incorrect assignment as even subtle structural fluctuations can affect each residue environment. With constant-pH molecular dynamics (CpH-MD) it is possible to model pH effects retrieving the protonation equilibriums of titratable residues coupled to protein conformational dynamics (Mongan et al., 2004; Khandogin and Brooks, 2005; Chen et al., 2014; Huang et al., 2016). Recently, Vo and co-workers reconstructed how fentanyl binds μ-opiod receptor with CpH-MD showing that the protonation of His257 at the binding pocket plays a crucial role to properly orient fentanyl (Vo et al., 2021). These results point out the importance of correctly accounting for protonation states of residues in the binding pocket to characterize the thermodynamics and kinetics of ligand-binding. However, as captured by spontaneous binding MD simulations, the ligand can establish contact with different protein residues in its pathway toward the binding site. The nature of these interactions will also be determinant for the kinetics of the ligand binding process. Despite the number of studies of protein-ligand pathways, the effect of protonation state changes of in-pathway residues in spontaneous binding simulations remains mostly unexplored.

The binding of benzamidine to trypsin has been commonly used as an enzyme-inhibitor model system for studying spontaneous binding and benchmarking enhanced sampling techniques due to the rapid formation of the trypsin-benzamidine complex (Betz and Dror, 2019). Trypsin is a serine protease responsible of hydrolyzing proteins through a catalytic triad formed by Ser195, His57, and Asp102 (*see* **Figure 1**). The positively charged inhibitor benzamidine is recognized in the specific S1 pocket which contains a negatively charged Asp189 located more than 10 Å away from the catalytic triad. In a landmark publication, Buch et al. reconstructed the free-energy landscape of benzamidine binding from a total of 495 MD simulations of 100 ns, observing productive binding in 38% of the simulations (Buch et al., 2011). By analyzing the independent trajectories, they observed that catalytic His57 and Ser195 residues were commonly found in the binding pathway of benzamidine in its way toward the S1 pocket. The binding of benzamidine have also been studied using unconstrained enhanced sampling methods by Miao and co-workers who reconstructed binding and unbinding pathways using Gaussian accelerated molecular dynamics (GaMD) (Miao et al., 2020). Interestingly, GaMD unbinding pathways showed that benzamidine passes next to His57 in its dissociation from the S1 pocket to the solvent. Therefore, His57 play a prominent role in both catalysis and binding.

Enzymes are sensitive to pH changes and catalytic residues commonly change their protonation states at different stages of the catalytic cycle. Trypsin is active in a pH range between 7.0 and

9.0 as His57 is required to alter between two protonation states along binding, acylation and deacylation steps of the hydrolysis reaction (Sipos and Merkel, 1970; Malthouse, 2020). Czodrowski and co-workers studied the protonation changes in ligand binding in trypsin concluding that His57 is responsible for the most relevant pKa shifts during binding and catalysis (Czodrowski et al., 2007). Most common software to assign protonation states from X-ray structures predict a positively charged His57 (both delta and epsilon nitrogens protonated, HIP) at pH = 7.0 while a less clear picture arises at pH = 8.0, where both HIP and neutral His57 with the delta nitrogen protonated (HID) are possible protonation states. Short-time scale MD simulations revealed that both HIP and HID were possible protonation states of His57 (Uranga et al., 2012). Spontaneous binding simulations of the benzamidine-trypsin system have been commonly performed with His57 in the HID state, which is the assumed protonation state when the Michaelis complex is formed (Wahlgren et al., 2011). The question is whether the protonation state of His57 can influence benzamidine binding.

Here, we use spontaneous binding MD simulations to reconstruct the trypsin-benzamidine binding pathway considering different protonation states of His57. This histidine is part of the trypsin catalytic triad and is located more than 10 Å away from Asp189 responsible for benzamidine binding in the S1 pocket (**Figure 1**). Our MD simulations show that the spontaneous binding pathways are critically dependent on the His57 protonation state. Binding of benzamidine frequently occurs in a few hundreds of nanoseconds when histidine is protonated in delta (HID) while productive binding is scarcely observed when His57 it is positively charged (HIP). CpH-MD simulations reflect that both HID and HIP forms are significantly populated at the pH range between 7.0 and 8.0 showing the displacement of the equilibrium toward the HID protonation state upon pH increase. These results indicate that properly accounting for protonation states of distal residues can be key to obtain reliable pathways in spontaneous binding simulations.

METHODS

System Preparation

We used the crystal structure of benzamidine-bound *Bos taurus* trypsin (PDB ID 3PTB) as starting point for our molecular dynamics (MD) simulations. First, benzamidine was removed from the S1 pocket to protonate the system. Second, protonation states of all protein residues were assigned based on 3.0 H++ webserver (http://biophysics.cs.vt.edu/H++) at pH 7.0 (Anandakrishnan et al., 2012). To explore the role of His57 protonation in benzamidine spontaneous binding, we manually assigned the protonation of His57 residue to either HID, HIE, or HIP. Once protonated, four benzamidine molecules were arbitrarily placed in the solvent, 30 Å away from Asp189 binding pocket, as described by Miao (Miao et al., 2020). Benzamidine parameters for MD simulations were obtained from the generalized AMBER force field (GAFF) (Wang et al., 2004), with partial charges set to fit the electrostatic potential generated at HF/6-31G* level of theory by restrained electrostatic potential model (Bayly et al., 1993). The atomic charges were calculated according to the Merz–Singh–Kollman (Singh and Kollman, 1984; Besler et al., 1990) scheme using Gaussian 09 (Frisch et al., 2016).

Conventional Molecular Dynamics Simulations

Spontaneous MD simulations starting from three different protonation states of His57 were performed in explicit water using AMBER18 package (Case et al., 2018). AMBER-ff14SB force field (Maier et al., 2015) was used to describe the protein, GAFF for benzamidine, and TIP3P for water molecules (Jorgensen et al., 1983). Each system was solvated in a cubic box with a 12 Å buffer of TIP3P water molecules and was neutralized by adding chloride counterions (Cl^-). Subsequently, a two-stage geometry optimization approach was performed: *1*) a short minimization of water molecules, with positional restraints on solute molecules; *2*) an unrestrained minimization of all the atoms in the simulation cell. Then, the systems were heated using six steps of 50 ps, incrementing the temperature 50 K each step (0–300 K) under constant-volume, periodic-boundary conditions, and the particle-mesh Ewald approach to introduce long-range electrostatic effects (Darden et al., 1993). A 10 Å cut-off was applied to Lennard-Jones and electrostatic interactions. Bonds involving hydrogen were constrained with the SHAKE algorithm (Ryckaert et al., 1977). The Langevin equilibration scheme is used to control and equalize the temperature (Wu and Brooks, 2003). The time step was kept at 2 fs during the heating stages. Each system was then equilibrated for 2 ns with a 2 fs timestep at a constant pressure of 1 atm to relax the density of the system. After the systems were equilibrated in the NPT ensemble, 50 replicas of 200 ns MD simulations for each protonation state of His57 (HID, HIE, HIP) were performed under the NVT ensemble and periodic-boundary conditions.

Constant-pH Molecular Dynamics Simulations

A total of 30 replicas of 200 ns of spontaneous binding constant-pH MD simulations (CpH-MD) were run at pH 7.0 and 8.0 considering all His residues as titratable (His40, His57, and His91). Here, discrete CpH-MD simulations have been carried out following the protocol described by Swails and co-workers as implemented in Amber (Swails et al., 2014): the MD is propagated in explicit solvent following the previously described protocol while the protonation state changes are carried out using a Generalized-Born (GB) implicit solvent model. A salt concentration of 0.1 M was also introduced to reproduce the same GB conditions the original algorithm was parametrized for (Mongan et al., 2004). Every 50 MD steps (100 fs) in explicit solvent, the protonation state of selected titratable residues in random order could change *via* Metropolis Monte Carlo attempts in an implicit solvent

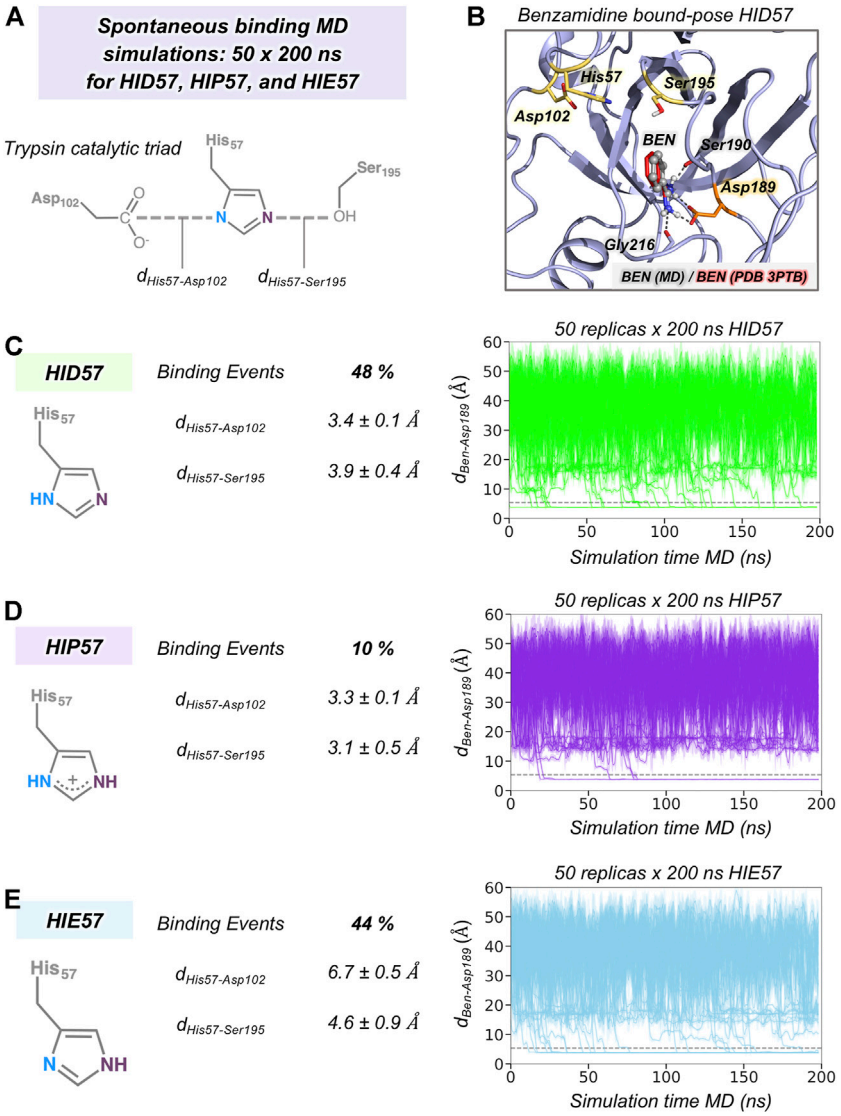


FIGURE 2 | The effect of His57 protonation in benzamidine binding. **(A)** Representation of the trypsin catalytic triad formed by Asp102, His57, and Ser195. The distance between Asp102 and His57 residues is calculated between the carbon of the carboxylate group of Asp102 and the delta nitrogen of His57. The distance between His57 and Ser195 is calculated between the epsilon nitrogen of His57 and the oxygen of the hydroxyl group of Ser195. **(B)** Representative conformation of the trypsin-benzamidine bound complex predicted from spontaneous binding HID57 MD simulations. The binding pose of benzamidine obtained from molecular dynamics (MD) simulations is shown in grey while the X-ray orientation (PDB 3PTB) is depicted in red. Catalytic residues and Asp189 are coloured in yellow and in orange, respectively. **(C)** Analysis of 50 replicas of 200 ns of HID57 MD simulations with the delta nitrogen of His57 protonated. **(D)** Analysis of 50 replicas of 200 ns of HIP57 MD simulations with His57 positively charged. **(E)** Analysis of 200 ns of HIE57 MD simulations with the epsilon nitrogen of His57 protonated. The percentage of binding events and the average distances (in Å) between catalytic residues is provided for each protonation state. Plot of the distance between the carbon atom of the amidine group of benzamidine and the carbon of the carboxylate group of Asp189 side chain along the 50 replicas of 200 ns MD simulations for each protonation state. The grey horizontal dashed line indicates when productive benzamidine binding takes place (distance below 5.4 Å).

framework. When the protonation state changes, a total of 100 relaxation steps (200 fs) were performed to relax the explicit solvent dynamics.

RESULTS

To evaluate the impact of His57 protonation state in the binding of benzamidine, we performed spontaneous binding molecular dynamics (MD) simulations placing four benzamidine molecules at least 30 Å away from the trypsin S1 pocket. From this starting point, benzamidine molecules are allowed to freely diffuse from the solvent and explore the trypsin surface in the predefined simulation time. As shown by Buch et al., spontaneous binding can readily occur in the nanosecond time scale (Buch et al., 2011). Spontaneous binding MD simulations were performed using two different strategies. First, a total of 50 replicas of 200 ns MD simulations were run for the three possible protonation states of

His57: *1*) delta nitrogen protonated (HID57); *2*) epsilon nitrogen protonated (HIE57); and *3*) positively charged histidine (HIP57). Second, we carried out a total of 30 replicas of 200 ns of constant-pH MD simulations at pH 7.0 and 8.0. From these simulations, the percentage of binding events and ligand pathways of benzamidine into trypsin were extracted. To evaluate the binding of benzamidine along the MD simulations, we monitored the distance between the amidine carbon of benzamidine and the carboxylate carbon of Asp189 side chain. We considered that productive binding is attained when this distance is below 5.4 Å, which is the minimum distance to retain the salt bridge interaction between the inhibitor and Asp189 in the S1 pocket.

The Effect of His57 Protonation States in Benzamidine Binding

Our MD simulations showed that spontaneous binding of benzamidine to the S1 pocket can occur in all protonation states of His57 (*see* **Figure 2**). The preferred binding mode obtained from MD simulations in the three protonation states is equivalent to the one observed in the PDB 3PTB (trypsin-benzamidine complex), which is also the principal binding pose predicted in previous computational studies (Buch et al., 2011; Plattner and Noé, 2015; Miao et al., 2020). However, a different number of binding events was observed for the three protonation states of His57. In particular, benzamidine binds the S1 pocket in 48% of HID57 simulations (24 out of 50 replicas of 200 ns), 44% of HIE57 (22 out of 50 replicas), and 10% of HIP57 (5 out of 50 replicas), as shown in **Figure 2**. Therefore, binding frequently occurs when His57 is found in its neutral form (either delta or epsilon nitrogen protonated) and becomes a less frequent event in its positively charged state (HIP57). Buch et al. described a total of 38% of productive binding events protonating His57 as HID but using shorter simulation time and only one molecule of benzamidine (Buch et al., 2011). Considering only the replicas that captured productive binding events, the average binding times are: 88, 80, and 55 ns for HID57, HIE57, and HIP57 respectively. These results point out that binding can readily occur in all cases irrespective of the lower probability of binding events observed for HIP57. Therefore, the open question is how the different protonation states of His57 alter the probability of binding of benzamidine.

In terms of global conformational dynamics, trypsin showed similar flexibility in the three protonation states of His57. Root-mean square fluctuations (RMSF) indicated that the flexibility of the S1 pocket is not significantly altered (*see* **Supplementary Figure S1**). The main differences were observed when analysing the stability of the catalytic triad formed by Asp102, His57, and Ser195 (*see* **Figure 2**). When His57 is positively charged (HIP57), the catalytic triad remains significantly stable, being 3.2 ± 0.5 Å and 3.3 ± 0.1 Å the Ser195-His57 and Asp102-His57 distances respectively. Additional flexibility is gained in the HID57 state with distances of 3.9 ± 0.4 Å (Ser195-His57) and 3.4 ± 0.1 Å (Asp102-His57). The intrinsic dynamism of the Ser195-His57 interaction in HID57 can be key to accommodate trypsin substrates triggering the formation of the Michaelis complex. Finally, the HIE57 protonation state is the least probable in trypsin because as shown in the MD simulations, when protonated in epsilon, His57 destabilizes the catalytic triad (*see* **Figure 2**). For this reason, all subsequent analyses will be focused on HID57 and HIP57 states.

Characterization of Benzamidine Binding Pathways

To gain insight into the molecular basis of the effect of His57 protonation changes in the binding process, we explored the binding pathways that benzamidine followed to get into trypsin S1 pocket. First, we collectively represented all spontaneous binding MD simulations corresponding to each protonation state using two coordinates (*see* **Figure 3** and **Supplementary Methods**): *1*) the binding distance ($d_{Ben\text{-}Asp189}$, *x* axis) between the amidine carbon of benzamidine and the carboxylate carbon of Asp189; and *2*) the distance between the amidine carbon of benzamidine and the epsilon nitrogen of His57, for either HID57 or HIP57 ($d_{Ben\text{-}His57}$, *y* axis). We selected the epsilon nitrogen because it is directly interacting with the hydroxyl group of catalytic Ser195. In both HID57 and HIP57, the most populated state of this 3D free-energy landscape (FEL) is the benzamidine-bound state (i.e. $d_{Ben\text{-}Asp189}$ below 5 Å and $d_{Ben\text{-}His57}$ around 10 Å, *see* **B** in **Figure 3**). Benzamidine also accumulates in both cases in a region of the FEL defined by a range of $d_{Ben\text{-}Asp189}$ [15,20] Å and $d_{Ben\text{-}His57}$ [10,15] Å, which corresponds to trypsin hydrophobic S3 pocket composed by Trp210 and surrounding residues. Interestingly, the FEL displays significant differences in the binding patterns of HID57 and HIP57 when benzamidine approaches the S1 pocket ($d_{Ben\text{-}Asp189}$ within [5,15] Å). For HID57, the FEL shows a metastable state where benzamidine directly interacts with His57 ($d_{Ben\text{-}His57}$ below 5 Å while $d_{Ben\text{-}Asp189}$ is still found between 10 and 15 Å, *see* $\mathbf{I_1}$ and $\mathbf{I_2}$ states in **Figure 3A**). This state is not visited in the FEL of HIP57 indicating that the interaction between His57 and benzamidine is not established in these MD simulations (*see* **Figure 3B**). These results are not surprising considering that both benzamidine and His57 are positively charged in HIP57 simulations resulting in repulsive interactions that prevent the interaction. From these simulations, we estimated the free-energy difference between the unbound conformation and the transition state that leads to productive binding (*see* **Supplementary Figure S2**). The free-energy difference is lower for HID57 (around 2.5 kcal/mol) than for HIP57 (around 4 kcal/mol), pointing out that the binding of benzamidine is globally slowed down when His57 is positively charged. Despite the free energy differences are not significant, the different distribution of binding events and the reshape of the FEL indicates that benzamidine binding is modulated by the protonation state of His57 which is located more than 10 Å away from Asp189.

To characterize the molecular basis of the ligand binding processes, we independently analyzed the MD trajectories projecting them into the corresponding FEL (*see* **Figure 3** and **Supplementary Figure S3**). The analysis of independent HID57 MD trajectories showed that benzamidine binding occurred mainly through two different pathways. In the major binding

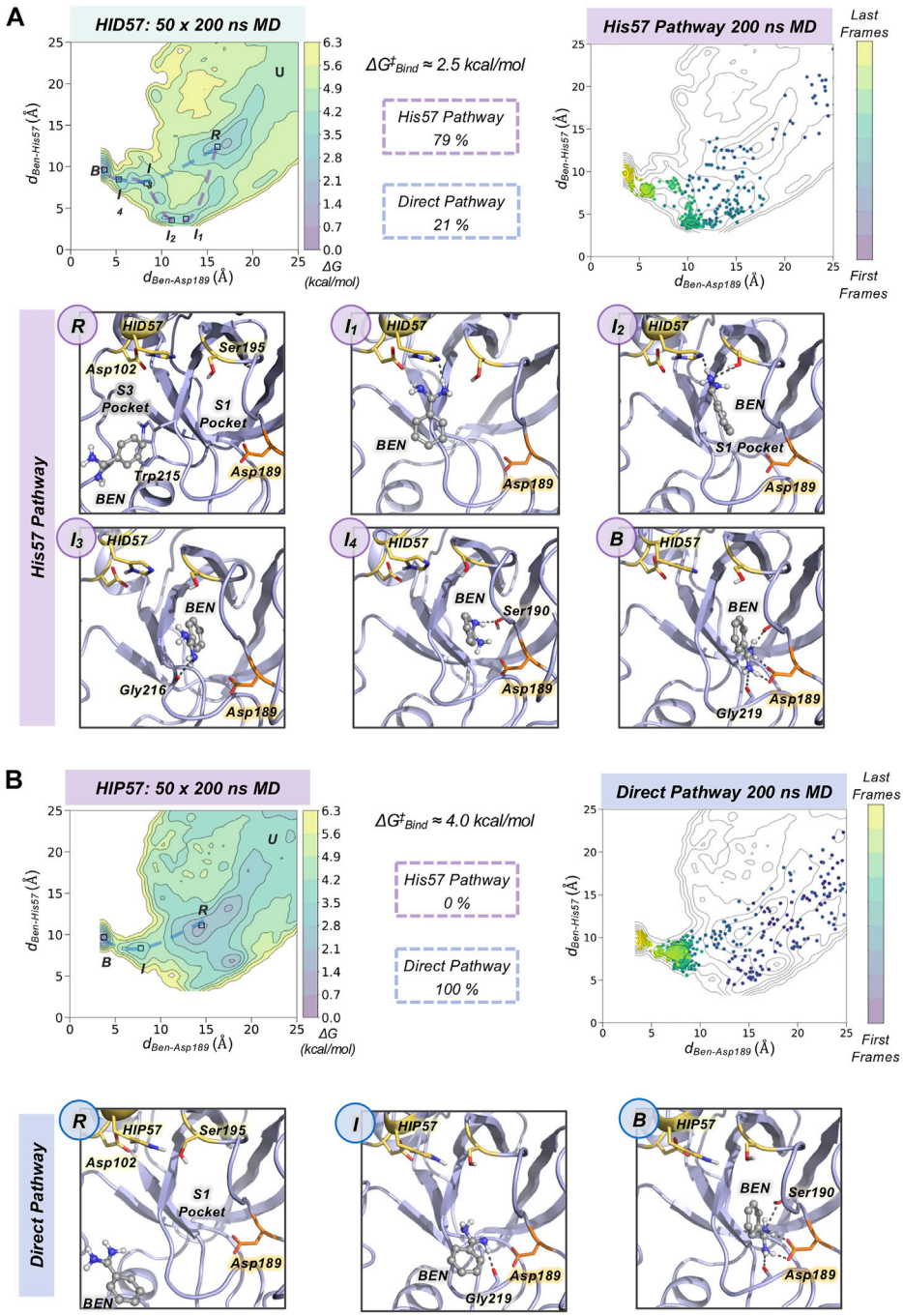


FIGURE 3 | Characterization of benzamidine binding pathways. Free energy landscape (FEL) reconstructed from 50 replicas of 200 ns of spontaneous binding MD simulations of HID57 **(A)** and HIP57 **(B)** using the binding distance ($d_{Ben\text{-}Asp189}$, *x* axis) between the carbon atom of the amidine group of benzamidine and the carbon of the carboxylate group of Asp189 and the distance between the carbon atom of the amidine group of benzamidine and the epsilon nitrogen of His57 ($d_{Ben\text{-}His57}$, *y* axis). The most relevant states of the FEL are highlighted in black boxes: *U* (unbound), *R* (recognition), *B* (bound) and *I* (intermediate) states. The free-energy difference between the unbound conformation and the transition state that leads to productive binding is given in kcal/mol. The purple dashed line indicates the trajectory of the His57 pathway while the blue dashed line indicates the trajectory of the direct pathway. The percentage of binding events (considering only productive binding simulations) that follow each pathway is provided. Projection of a representative spontaneous binding 200 ns MD trajectory on the FEL of each HID57 and HIP57. The time evolution of the ligand binding pathway is represented in a colour scale ranging from purple for the first frames to yellow for the last frames of the MD trajectory. Molecular representation of the most relevant states of the FEL corresponding to the His57 and direct pathways. Catalytic residues are shown in yellow, benzamidine in grey, and Asp189 in orange.

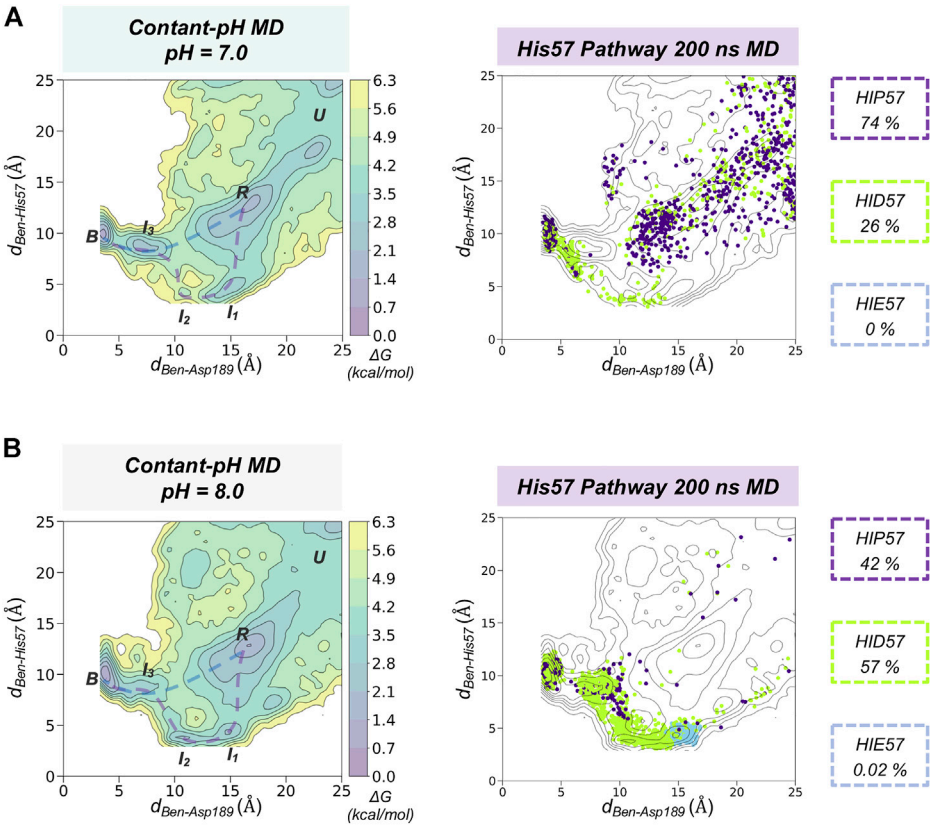


FIGURE 4 | Spontaneous Benzamidine Binding with Constant-pH Molecular Dynamics Simulations. Free energy landscape (FEL) reconstructed from 30 replicas of 200 ns of spontaneous binding constant-pH MD simulations at pH = 7.0 **(A)** and pH = 8.0 **(B)** using the binding distance ($d_{Ben\text{-}Asp189}$, *x* axis) between the carbon atom of the amidine group of benzamidine and the carbon of the carboxylate group of Asp189 and the distance between the carbon atom of the amidine group of benzamidine and the epsilon nitrogen of His57 ($d_{Ben\text{-}His57}$, *y* axis). The most relevant states of the FEL are highlighted in black boxes: *U* (unbound), *R* (recognition), *B* (bound) and *I* (intermediate) states. Projection of a representative 200 ns trajectory of the His57 pathway showing the protonation state of each frame in different colour: HID in green, HIP in purple, and HIE in blue. The equilibrium populations of each protonation state retrieved from the 30 replicas of 200 ns is provided for pH = 7.0 and 8.0.

pathway (observed in 79% of productive binding simulations), benzamidine first enters the S3 pocket being recognized by Trp210. Catalytic Asp102 may be involved in the attraction of positively charged benzamidine into the S3 pocket (*see* **Supplementary Figure S4**). Second, the inhibitor rolls along this pocket to establish a hydrogen bond interaction with His57 while keeping the phenyl moiety in the S3 pocket. Then, the aromatic ring of benzamidine repositions from the S3 to the S1 pocket maintaining the hydrogen bonding with catalytic His57 and Ser195. Finally, through a series of successive steps, benzamidine turns around to establish a salt bridge interaction with the carboxylate group of Asp189, attaining the benzamidine-bound pose. We term this predominant pathway as "His57 pathway" since establishing a hydrogen bond interaction with His57 is a requisite to access the S1 pocket. A similar binding pathway was described by Buch and co-workers using 495 100 ns MD simulations (Buch et al., 2011). A second, less frequent, pathway was observed in 21% of productive binding simulations. In this case, benzamidine directly evolves from the solvent to the S1 pocket without establishing a hydrogen bond interaction with HID57 (*see* **Figure 3**). The direct access to the S1 pocket from the solvent can take place from different sites but the common feature is that benzamidine directly enters the pocket establishing a hydrogen bond with Ser190 and then repositions to interact with Asp189 through a salt-bridge interaction. In this particular pathway, the binding of benzamidine is practically a pure diffusion from the solvent to the S1 pocket, for this reason we term it as the "direct pathway."

In HIP57 MD simulations, benzamidine binding only occurred through the direct pathway (5 out of 50 replicas), as shown in **Figure 3B**. The repulsion between the positive charges of both protonated His57 (HIP) and benzamidine prevents their approximation and interaction. Thus, the His57 pathway is not observed in HIP57 simulations, making the number of binding events significantly less frequent than in HID57. Despite benzamidine can be recognized also in the S3 pocket by Trp210, when it approaches the positively charged HIP57 a hydrogen bond with this catalytic residue cannot be established and benzamidine returns back to the solvent (*see* **Supplementary Figure S5**). Therefore, when His57 is positively charged, binding will preferentially occur through direct diffusion from the solvent. These results demonstrate that the protonation state of His57

determines which ligand binding pathways can be populated and, thus, the path that benzamidine preferentially takes to access the S1 pocket of trypsin.

Spontaneous Benzamidine Binding with Constant-pH Molecular Dynamics Simulations

Using fixed protonation states for His57 can offer a limited picture of benzamidine binding, considering that this residue is responsible for the most relevant pKa shifts during binding and catalysis in trypsin (Czodrowski et al., 2007). To account for the pH effects in the benzamidine binding process, we performed 30 replicas of 200 ns of spontaneous binding constant-pH MD simulations at pH 7.0 and 8.0 (*see* **Figure 4** and **Supplementary Figure S6** for a complete analysis and S7 for convergence of equilibrium populations). In these simulations, we allowed the three histidines of trypsin to change their protonation state. Only His57 is found along the binding pathway and can play a key role in the benzamidine recognition process. At pH 7.0, the populations of the different protonation states obtained from CpH-MD simulations were 74, 26, and 0% for HIP, HID, and HIE, respectively. From these CpH-MD simulations at pH 7.0, we obtained a FEL of the binding process that resembles the one captured for HIP57 protonation state (*see* **Figures 3B**, **4A**). However, the intermediate states corresponding to the interaction between benzamidine and His57 became moderately populated. This increases the number of binding events compared to HIP57 simulations up to 23% (7 out of 30 replicas). In this case, benzamidine accessed the S1 pocket through both direct and His57 pathways. Interestingly, CpH-MD simulations showed that the His57 pathway is activated only when His57 attains the HID protonation state (*see* **Figure 4A**). At pH 8.0, we observed equilibrium populations of protonation states for His57 of 57% for HID, 42% for HIP, while HIE is scarcely populated. The FEL obtained from CpH-MD simulations at pH 8.0 clearly highlights the stabilization of the interaction between His57 and benzamidine (*see* $\mathbf{I_1}$ and $\mathbf{I_2}$ states in **Figure 4B**). Again, binding of benzamidine proceeded through both the direct and His57 pathways. The number of binding events retrieved from CpH-MD simulations at both pH 7.0 and 8.0 is found between the values observed in HIP57 and HID57 MD simulations offering a clearer picture of the His57 protonation state ensemble. Therefore, properly accounting for the equilibrium of protonation states of His57 may be key to retrieve accurate kinetics for the binding of benzamidine to trypsin.

DISCUSSION

Spontaneous binding MD simulations showed that the protonation state of His57, which is located more than 10 Å away from the gorge of the S1 pocket, plays a key role in determining the binding pathway of benzamidine to trypsin. Binding is more favorable when His57 attains a neutral HID protonation state while is less probable in the positively charged HIP protonation. These results are in line with kinetic experiments that indicate that K_s of substrate N-α-benzyloxycarbonyl-l-lysine-p-nitroanilide increases more than 80 fold when His57 is protonated (Malthouse, 2020). Benzamidine can access the S1 pocket through two main pathways: the His57 pathway and the direct pathway from the solvent. We observed that His57 is found in the way of benzamidine to the S1 pocket through the most probable binding pathway in the HID protonation state, establishing a hydrogen bond that is key to drive benzamidine toward the binding site. Therefore, subsequent kinetic analysis of the binding process will provide different outcomes depending on how the protonation states are defined at the beginning of the simulation. Constant-pH MD simulations naturally account for the protonation state ensemble of His57 offering a more accurate description of the spontaneous binding of benzamidine at a fixed pH.

Based on these results, we suggest to always consider the impact of protonation changes of residues that are found along the ligand binding pathway (even distal residues) when performing spontaneous binding MD simulations. In particular, in systems with slow binding processes that follow complex pathways through different metastable intermediate and transition states. It is important to remark that in our study we have excluded protonation changes of other residues (e.g. Asp, Glu, . . .) which may further alter the binding pathways obtained. These observations are not limited to the study of drug-binding into their biological receptors. In protein folding studies, it was reported that changes protonation states of certain residues were important to describe the intrinsic dynamics of amyloid-β peptides (Li et al., 2017). In enzyme engineering, substrate access tunnels are commonly engineered through point mutations to evolve the enzyme toward a new function or broad its substrate scope. Thus, properly accounting for coupled protonation state changes between the residues conforming the access tunnel and the substrate will be important to evaluate the substrate binding pathways in enzymes. By unravelling the details of ligand binding and unbinding it is possible to gain insight into the detailed molecular mechanisms of relevant biochemical processes and then, harness this information to rationally improve the potency of drugs and/or evolve an enzyme toward novel functions.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

FF, HG, and MG-B designed research. HG performed research. HG, MG-B, and FF analysed data and wrote the manuscript.

FUNDING

This work was supported by the Spanish MICINN projects PID 2019-111300GA-I00 (MG-B), RTI 2018-101032-J100 (FF). We thank Spanish MICINN for Ramón y Cajal fellowships RYC

2020-028628-I (MG-B) and RYC 2020-029552-I (FF). We thank the Generalitat de Catalunya for the emerging group CompBioLab (2017 SGR-1707), the consolidated group DiMoCat (2017 SGR-39), for predoctoral FI fellowship 2022 FI_B 00615 (HG) and Beatriu de Pinós H2020 MSCA-Cofund 2018-BP-00204 project (to MG-B).

ACKNOWLEDGMENTS

We thank Carla Calvó-Tusell and Miguel A. Maria-Solano for fruitful discussions and Daniel Masó for technical support. We thank Yinglong Miao for helpful discussions on spontaneous binding simulations of the trypsin-benzamidine complex. This work has been recently published online as a preprint (Girame et al., 2022).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.922361/full#supplementary-material

REFERENCES

Aguilar, B., Anandakrishnan, R., Ruscio, J. Z., and Onufriev, A. V. (2010). Statistics and Physical Origins of pK and Ionization State Changes upon Protein-Ligand Binding. *Biophysical J.* 98, 872–880. doi:10.1016/J.BPJ.2009.11.016

Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012). H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* 40, W537–W541. doi:10.1093/NAR/GKS375

Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* 97, 10269–10280. doi:10.1021/j100142a004

Bernetti, M., Masetti, M., Rocchia, W., and Cavalli, A. (2019). Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem.* 70, 143–171. doi:10.1146/annurev-physchem-042018-052340

Besler, B. H., Merz, K. M., and Kollman, P. A. (1990). Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* 11, 431–439. doi:10.1002/JCC.540110404

Betz, R. M., and Dror, R. O. (2019). How Effectively Can Adaptive Sampling Methods Capture Spontaneous Ligand Binding? *J. Chem. Theory Comput.* 15, 2053–2063. doi:10.1021/acs.jctc.8b00913

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

Calvó-Tusell, C., Maria-Solano, M. A., Osuna, S., and Feixas, F. (2022). Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase. *J. Am. Chem. Soc.* 144, 7146–7159. doi:10.1021/jacs.1c12629

Case, D., Ben-Shalom, I., Brozell, S. R., Cerutti, D. S., Cheatham, T., Cruzeiro, V. W. D., et al. (2018). *AMBER 2018*. University of California, San Francisco.

Chen, W., Morrow, B. H., Shi, C., and Shen, J. K. (2014). Recent Development and Application of Constant pH Molecular Dynamics. *Mol. Simul.* 40, 830–838. doi:10.1080/08927022.2014.907492

Czodrowski, P., Sotriffer, C. A., and Klebe, G. (2007). Protonation Changes upon Ligand Binding to Trypsin and Thrombin: Structural Interpretation Based on pKa Calculations and ITC Experiments. *J. Mol. Biol.* 367, 1347–1356. doi:10.1016/J.JMB.2007.01.022

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

Decherchi, S., and Cavalli, A. (2020). Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chem. Rev.* 120, 12788–12833. doi:10.1021/acs.chemrev.0c00534

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

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2016). *Gaussian 09, Revision A.02*. Wallingford CT: Gaussian, Inc.

Girame, H., Garcia-Borràs, M., and Feixas, F. (2022). Changes in Protonation States of In-Pathway Residues Can Alter Ligand Binding Pathways Obtained from Spontaneous Binding Molecular Dynamics Simulations. *bioRxiv*. doi:10.1101/2022.04.30.490145

Huang, Y., Chen, W., Wallace, J. A., and Shen, J. (2016). All-Atom Continuous Constant pH Molecular Dynamics with Particle Mesh Ewald and Titratable Water. *J. Chem. Theory Comput.* 12, 5411–5421. doi:10.1021/acs.jctc.6b00552

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

Khandogin, J., and Brooks, C. L. (2005). Constant pH Molecular Dynamics with Proton Tautomerism. *Biophysical J.* 89, 141–157. doi:10.1529/BIOPHYSJ.105.061341

Li, J., Lyu, W., Rossetti, G., Konijnenberg, A., Natalello, A., Ippoliti, E., et al. (2017). Proton Dynamics in Protein Mass Spectrometry. *J. Phys. Chem. Lett.* 8, 1105–1112. doi:10.1021/acs.jpclett.7b00127

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. Theory Comput.* 11, 3696–3713. doi:10.1021/ACS.JCTC.5B00255

Malthouse, J. P. G. (2020). Kinetic Studies of the Effect of pH on the Trypsin-Catalyzed Hydrolysis of N-α-Benzyloxycarbonyl-L-Lysine-P-Nitroanilide: Mechanism of Trypsin Catalysis. *ACS Omega* 5, 4915–4923. doi:10.1021/acsomega.9b03750

Miao, Y., Bhattarai, A., and Wang, J. (2020). Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): Characterization of Ligand Binding Thermodynamics and Kinetics. *J. Chem. Theory Comput.* 16, 5526–5547. doi:10.1021/acs.jctc.0c00395

Miao, Y., and McCammon, J. A. (2016). Graded Activation and Free Energy Landscapes of a Muscarinic G-Protein-Coupled Receptor. *Proc. Natl. Acad. Sci. U.S.A.* 113, 12162–12167. doi:10.1073/PNAS.1614538113

Mongan, J., Case, D. A., and McCammon, J. A. (2004). Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* 25, 2038–2048. doi:10.1002/JCC.20139

Onufriev, A. V., and Alexov, E. (2013). Protonation and pK Changes in Protein-Ligand Binding. *Quart. Rev. Biophys.* 46, 181–209. doi:10.1017/S0033583513000024

Plattner, N., and Noé, F. (2015). Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* 6, 1–10. doi:10.1038/ncomms8653

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

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

Singh, U. C., and Kollman, P. A. (1984). An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* 5, 129–145. doi:10.1002/JCC.540050204

Sipos, T., and Merkel, J. R. (1970). Effect of Calcium Ions on the Activity, Heat Stability, and Structure of Trypsin. *Biochemistry* 9, 2766–2775. doi:10.1021/bi00816a003

Swails, J. M., York, D. M., and Roitberg, A. E. (2014). Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput.* 10, 1341–1352. doi:10.1021/ct401042b

Uranga, J., Mikulskis, P., Genheden, S., and Ryde, U. (2012). Can the Protonation State of Histidine Residues Be Determined from Molecular Dynamics Simulations? *Comput. Theor. Chem.* 1000, 75–84. doi:10.1016/j.comptc.2012.09.025

Vo, Q. N., Mahinthichaichan, P., Shen, J., and Ellis, C. R. (2021). How μ-opioid Receptor Recognizes Fentanyl. *Nat. Commun.* 12, 1–11. doi:10.1038/s41467-021-21262-9

Wahlgren, W. Y., Pál, G., Kardos, J., Porrogi, P., Szenthe, B., Patthy, A., et al. (2011). The Catalytic Aspartate Is Protonated in the Michaelis Complex Formed between Trypsin and an In Vitro Evolved Substrate-like Inhibitor. *J. Biol. Chem.* 286, 3587–3596. doi:10.1074/jbc.M110.161604

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

Wu, X., and Brooks, B. R. (2003). Self-guided Langevin Dynamics Simulation Method. *Chem. Phys. Lett.* 381, 512–518. doi:10.1016/j.cplett.2003.10.013

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 07 July 2022
doi: 10.3389/fmolb.2022.933676

The Role of Conformational Dynamics and Allostery in the Control of Distinct Efficacies of Agonists to the Glucocorticoid Receptor

Yuxin Shi[1,2†], *Shu Cao*[3†], *Duan Ni*[4], *Jigang Fan*[1], *Shaoyong Lu*[1,2*] *and Mintao Xue*[5*]

[1]Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Chinese Ministry of Education, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2]Medicinal Chemistry and Bioinformatics Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [3]Department of Urology, Ezhou Central Hospital, Hubei, China, [4]The Charles Perkins Centre, University of Sydney, Sydney, NSW, Australia, [5]Department of Orthopedics, Second Affiliated Hospital of Naval Medical University, Shanghai, China

Glucocorticoid receptor (GR) regulates various cellular functions. Given its broad influence on metabolic activities, it has been the target of drug discovery for decades. However, how drugs induce conformational changes in GR has remained elusive. Herein, we used five GR agonists (dex, AZ938, pred, cor, and dibC) with different efficacies to investigate which aspect of the ligand induced the differences in efficacy. We performed molecular dynamics simulations on the five systems (dex-, AZ938-, pred-, cor-, and dibC-bound systems) and observed a distinct discrepancy in the conformation of the cofactor TIF2. Moreover, we discovered ligand-induced differences regarding the level of conformational changes posed by the binding of cofactor TIF2 and identified a pair of essential residues D590 and T39. We further found a positive correlation between the efficacies of ligands and the interaction of the two binding pockets' domains, where D590 and T739 were involved, implying their significance in the participation of allosteric communication. Using community network analysis, two essential communities containing D590 and T739 were identified with their connectivity correlating to the efficacy of ligands. The potential communication pathways between these two residues were revealed. These results revealed the underlying mechanism of allosteric communication between the ligand-binding and cofactor-binding pockets and identified a pair of important residues in the allosteric communication pathway, which can serve as a guide for future drug discovery.

Keywords: glucocorticoid receptor, allosteric communication, allosteric site, molecular dynamics simulation, drug discovery

OPEN ACCESS

Edited by:
Weiliang Zhu,
Chinese Academy of Sciences (CAS), China

Reviewed by:
Jinan Wang,
University of Kansas, United States
Marcel Bermudez,
University of Münster, Germany

***Correspondence:**
Shaoyong Lu
lushaoyong@sjtu.edu.cn
Mintao Xue
xmt1984@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *01 May 2022*
Accepted: *30 May 2022*
Published: *07 July 2022*

Citation:
Shi Y, Cao S, Ni D, Fan J, Lu S and Xue M (2022) The Role of Conformational Dynamics and Allostery in the Control of Distinct Efficacies of Agonists to the Glucocorticoid Receptor. Front. Mol. Biosci. 9:933676. doi: 10.3389/fmolb.2022.933676

INTRODUCTION

Glucocorticoid receptor belongs to the nuclear receptor (NR) superfamily to transduce the signals triggered upon its ligand glucocorticoid (GC) binding (Veleiro et al., 2010; Kadmiel and Cidlowski, 2013; Cain and Cidlowski, 2015). It is broadly implicated in a variety of biological events such as metabolism, proliferation, and apoptosis. Given the critical significance of GR, its structures and related signaling pathways have been intensively investigated in detail. GR comprised three domains,

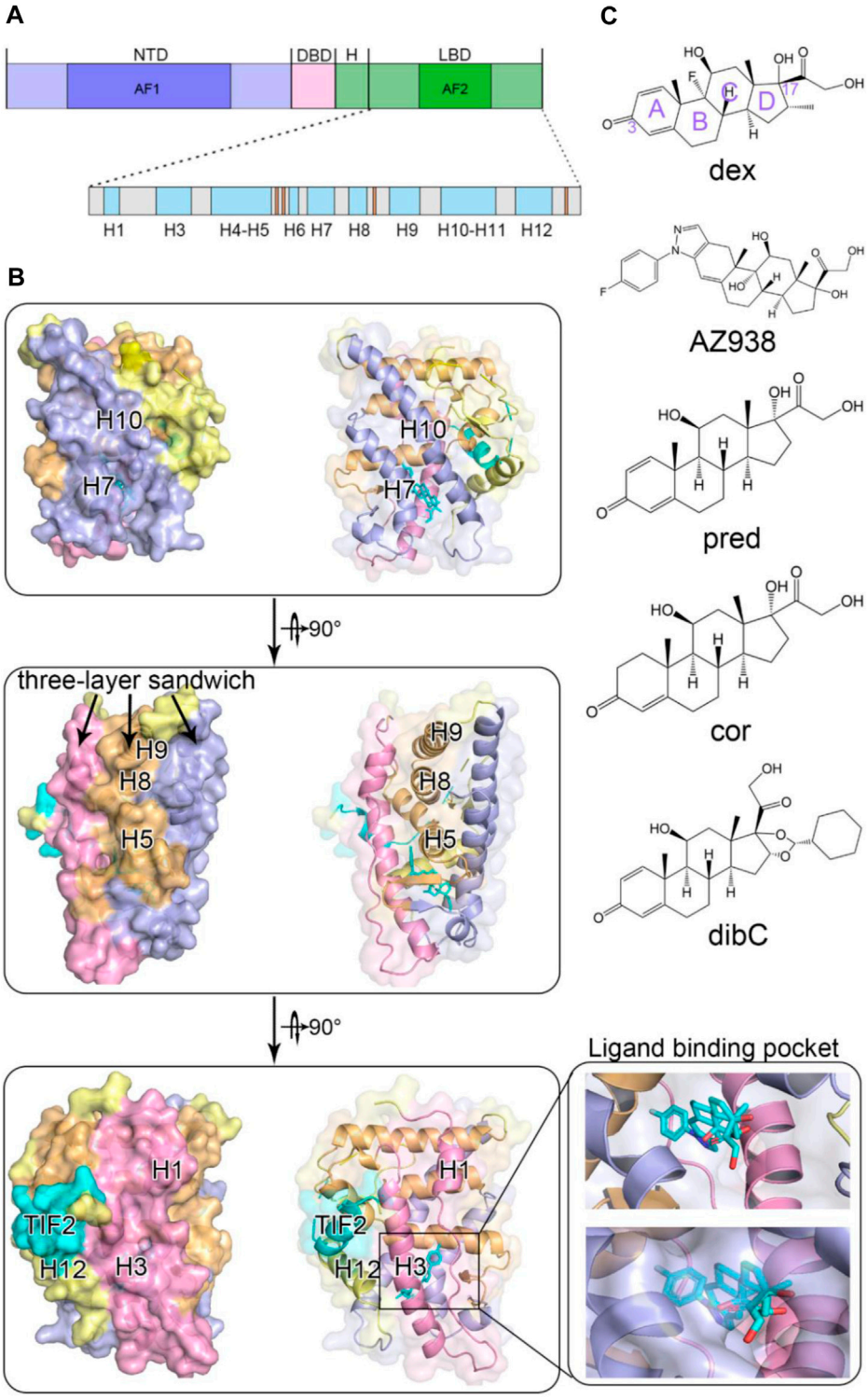


FIGURE 1 | Overall structure of glucocorticoid receptor with agonists and a cofactor. **(A)** domain organization of GR. **(B)** cartoon representation of GR ligand-binding domain with helices colored according to the three-layered sandwich structure. **(C)** chemical structures of five agonists.

including one N-terminal transactivation domain (NTD), one DNA-binding domain (DBD), and one ligand-binding domain (LBD) (**Figure 1A**) (Álvarez et al., 2008a). The NTD is intrinsically disordered and contains an activation function 1 (AF-1) transactivation domain, which is responsible for interacting with the coactivator and is responsible for GR's

transcriptional activities. Despite lacking a stable tertiary structure in its intrinsically disordered region (IDR), NTD is essential in the allosteric control of GR's activity (Li et al., 2017). Li et al. (2017) demonstrated that hGR tunes signaling from NTD by producing isoforms differing uniquely in the length of the disordered region. This IDR with a discrepancy in length was believed to propagate structural changes and influence the function of the receptor. On the other hand, the DBD possesses two distinguishable zinc finger regions where DNA anchors. The C-terminal region is where ligands bind, which is also involved in dimerization and interaction with the cofactor through the activation function-2 (AF2) domain (Carson-Jurica et al., 1990; Gronemeyer and Moras, 1995; Kumar and Thompson, 1999; Nagy and Schwabe, 2004). Upon agonists binding, the ligand-dependent AF2 induced conformational changes in GR and accomplished full transactivation function together with AF1 (Goto et al., 2003). The peculiarity of the LBD makes it the most relevant region for the potential interaction of ligand and receptor (Álvarez et al., 2008b).

Due to its critical implication in GR's functions, LBD structural biology receives considerable research interest. Although intense time has been invested toward this aspect, relatively limited success has been achieved. The first crystal structure of LBD was not successfully obtained until 2002, which formed a complex with its coactivator nuclear receptor coactivator 2 (TIF2) and ligand dexamethasone (Bledsoe et al., 2002). Since then, experimental studies and computational analyses have rapidly accumulated to focus on structural changes of LBD. It is now widely acknowledged that the LBD domain consists of 11 α-helices (H1, H3-H12) and four small β strands (**Figure 1A**). The protein folds into a canonical three-layer sandwich with a hydrophobic pocket in the shape of a one-side-opened box to accommodate the ligand (**Figure 1B**). The side of the box consists of three helices (H3, H7, and H11), and H4–H5 forms the top of the box (Edman et al., 2015). The C-terminal AF2 of the receptor has been found to be an important indicator of the ligand's efficacy. Since it adopts different conformations in distinct agonist-bound GR systems, AF2's plasticity suggested its contribution to the discrepancy of different agonists' efficacy (Buttgereit et al., 2018; Köhler et al., 2020; Hu et al., 2022).

GR executed an essential role in cells, bearing the responsibility of both transcriptional activation and non-genomic actions (Jiang et al., 2014; Meijer et al., 2018). In the absence of ligand, GR is predominantly localized in the cytoplasm and bound to either HSP70 or HSP90 and a tyrosine kinase-like c-Src to form a quaternary complex (Weikum et al., 2017; Lee et al., 2021; Karra et al., 2022). When an agonist binds to the GR and alters its structure, it stimulates downstream signaling pathways. The activated GR disassociates from the quaternary complex and moves into the nucleus in the form of homodimers, where it assembles and integrates with glucocorticoid-responsive elements (GREs) (O'Malley and Tsai, 1992; Pratt and Toft, 1997). GREs often sit at the promoters or exons of the target genes, and GR's binding leads to the recruitment of other factors required for transcription (Jenkins et al., 2001). By regulating different gene expressions, GR manipulates a wide range of cellular activities and thus possesses enormous potential for clinical applications (Darimont et al., 1998; Hu and Lazar, 1999).

GR is emerging as a critical factor for drug discovery especially in carbohydrate, protein, and fat metabolism (Buttgereit, 2020) and immunological disorder-related disease, such as asthma and dermatitis (Cato and Wade, 1996; Köhler et al., 2020). In 1995, there were ~6.6 million prescriptions relative to GR written in Germany. Until now, ~10 million drugs are prescribed just for oral corticosteroids each year merely in the United States (Van Staa et al., 2000; Schäcke et al., 2002). Large amounts of efforts have been dedicated over the last several decades by scientists and pharmaceutical companies to enhance the potency of drugs while minimizing side effects by modifying the chemical groups of natural glucocorticoid cortisol (Cain and Cidlowski, 2015). According to a long-standing hypothesis, the adverse effects were induced by dimer-mediated transcriptional activation since the involved genes participate in glucose synthesis and fat metabolism (Meijer et al., 2018). Based on this hypothesis, the goal of drug design is relatively unambiguous, which is to enhance the non-genomic effect and induce GR-protein interaction while impairing the genomic effect of GR-DNA binding (Heck et al., 1994; Reichardt et al., 2001; Meijer et al., 2018). Thitherto, the most common systemic glucocorticoids in clinical treatments are glucocorticoids with good oral bioavailability, which are eliminated mainly by hepatic metabolism and renal excretion of the metabolites. For instance, hydrocortisone (cortisone; cor), prednisolone (pred), methylprednisolone, and dexamethasone (dex) are all commonly used medicines (Thiessen, 1976; Musson et al., 1991). In addition to the traditional drugs on the market, scientists are inventing drugs with more innovative carbon backbones. One of the new compounds is AZ938, a cortivazol analog, which is currently under clinical trial (Styczynski et al., 2005). The chemical structure of AZ938 contains a bulky phenylpyrazole group replacing the C3 ketone of the steroid A ring. Previously, the 3-ketone was thought to be essential as it is conserved among steroid–receptor structures. However, the equivalent activity of cortivazol turned out to be 165-fold higher than prednisolone. Another notable compound is desisobutyryl-ciclesonide (dibC), which is the active metabolite of ciclesonide. It was proved to modulate *in vitro* allergen-driven activation of blood mononuclear cells and allergen-specific T-cell blasts (Czock et al., 2005). Unfortunately, despite the prosperity of drug design, a troublesome setback for drug design is that it is hard to separate the anti-inflammatory efficacy from side effects such as diabetes, muscle wasting, and osteoporosis (Schäcke et al., 2002; Gebhardt et al., 2013), which has become a huge disturbance to many people worldwide. Thus, it is becoming urgent to understand the structural mechanisms of GR–agonist interaction to better optimize drug design (Nussinov and Tsai, 2013). Even so, the underlying mechanism regarding interactions of GR and agonists is still unclear. In addition, the challenge of drug resistance requires an urgent design of new drugs (Fan et al., 2021; Liang et al., 2021). Without accurate comprehension of the relationship between ligands and GR as guidance, it will be difficult to optimize the current drugs and invent new ones with high efficacy and few side effects (Lu et al., 2016; Feng

et al., 2021; Lu et al., 2021a). Despite this, most of the studies currently are focusing on the allosteric discrepancy between agonist-bound and antagonist-bound GR systems, while few are focusing on the subtle changes that occurred in different agonist-bound GR systems. To tackle the long-standing setbacks of drug design, a study on the regulation of agonists on the GR is imminently needed (Liu and Nussinov, 2016; Lu et al., 2019c).

Here, we chose five typical GR agonists (dex, AZ938, pred, cor, and dibC) (**Figure 1C**) with different efficacies to investigate the mechanism underlying ligand–LBD interactions, accounting for different levels of GR function. The efficacies of the five ligands were previously measured using a transactivation reporter gene assay (Köhler et al., 2020). Compared with the highest effect of dex (100%), AZ938 ranked second with 90% of efficacy, which was followed by pred (86%). DibC and cor turned out to be the least effective (77%). Based on these results, we raised the question that what aspect of ligands induced the difference in efficacies. We carried out molecular dynamics (MD) simulations through a multiple microsecond timescale to explore the underlying allosteric effects and conformational dynamics of the LBD. We focused on the two pockets: the ligand-binding pocket and the cofactor-binding pocket, and their allosteric communication induced by different ligand binding to GR (Lu et al., 2019b). By aligning the representative structure of each system, we found different structural ensembles in the cofactor-binding pocket. Further dissection of conformational landscapes showed that induced by different ligands, dynamics in allosteric regulation was found in the response to cofactor TIF2. Moreover, using molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) calculation and distance analysis, we identified crucial residues that displayed preference for a more stable conformation in dex-bound and AZ938-bound systems (Zhang et al., 2019). On the other hand, dynamic cross-correlation matrices (DCCM) calculations also suggested that regions containing crucial residues exhibited significantly increased correlated motions in dex-bound systems compared to other systems. Finally, community network analysis and allosteric pathway analysis were carried out to reveal the potential communication pathways in each system (Ni et al., 2020). Together, this study investigated the allosteric dynamics between the five systems in detail, expounding the mechanism of interactions between agonists and GR. We expect this dynamic model of allostery will prove to be generally adopted in explaining signaling in all the other GR–agonist systems. Ultimately, we hope that this model can be a guide for chemical modification and optimization of drugs and give insights into novel treatments of concomitant drugs (Shen et al., 2016; Lu et al., 2019a; Lu and Zhang, 2019d; Zhang et al., 2022).

MATERIALS AND METHODS

System Preparation

Three co-crystal structures of GR complexed with agonists (dex–GR, PDB ID: 4UDC; cor–GR, PBD ID: 4P6X; and dibC–GR, PBD ID: 4UDD) were selected from the Protein Data Bank (PDB) as initial structures for MD simulations. The mutated residues were mutated back, and the missing residues were added using the Discovery Studio.

Molecular Docking

Due to the unavailability of co-crystal structures of GR–AZ938 and GR–pred complexes, molecular docking was performed to generate the 3D structure of these two complexes. The chemical structures of AZ938 and pred were built and pre-optimized using the ChemDraw software. The GR–NN7 complex (PBD ID: 4CSJ) and GR–dex complex (PBD ID: 4UDC) were used as templates for AZ938 and pred, respectively. The following docking procedures were accomplished using the Schrödinger program. The unnecessary water molecules beyond 5 Å and other cofactors were deleted from the template structure using the protein preparation module of Schrödinger. The H-bonds were optimized, and the system energy was minimized. The glide module was then used to generate boxes for docking. The target agonists were loaded into the software and processed by the ligPrep module. Finally, molecular docking was conducted using the Ligand Docking module in SP mode. All the above operations were carried out using default settings and parameters. The resulting docking poses were then analyzed with Pymol and Discovery Studio. Additional minimization of 10,000 steps using the steepest descent algorithm was performed by Discovery Studio to optimize the docking interface.

MD Simulations

MD simulations were performed on five systems (GR–dex, GR–AZ938, GR–pred, GR–cor, and GR–dibC) using the AMBER18 software (Jang et al., 2020; Li et al., 2020). First, we used Antechamber to create inpcrd and prmtop files for each agonist. Antechamber is a forcefield specifically designed to cover most pharmaceutical molecules and has excellent compatibility with the traditional AMBER forcefield. We loaded the ligand input PDB files and ran the *reduce* to add all the hydrogen to the systems. Then, we transformed the PDB files into Tripos Mol2 format. The AM1-BCC charge model was used to calculate the atomic charges. Utility *parmchk* was applied to create parameter files that can be loaded into LEaP. After loading the parameter files, we ran the LEaP and finally obtained the inpcrd and prmtop files (Bayly et al., 1993; Jakalian et al., 2000; Wang et al., 2004). Second, we obtained all the parameter files of the protein using ff14SB forcefield (Maier et al., 2015) and general Amber forcefield (GAFF). We added hydrogen to all the systems and created a truncated octahedron transferable intermolecular potential three-point (TIP3P) water box (Jorgensen et al., 1983) to approach the environment in physical conditions. We also added Na^+ and Cl^- atoms to neutralize the charge. After the preparation, we operated a protocol using four steps. We operated the minimization step two times. All the atoms in the complex were restrained at 500 kcal $mol^{-1}Å^{-2}$ using the steepest descent algorithms at the first time. Other ions and water molecules were minimized within 50,000 cycles (25,000 each for steepest descent and conjugate gradient cycles). At the second time, the systems underwent 50,000 cycles of steepest descent and conjugate gradient minimization each free of restrictions. Then, we heated up the system from 300 ps to 300 K in a canonical ensemble (NVT) with

a 700 ps equilibration step. Finally, a 1000 ns MD simulation was carried out in each system with random velocities in isothermal isobaric conditions (NPT) with periodic boundaries. The system was regulated by Langevin dynamics (Uberuaga et al., 2004; Sindhikara et al., 2009) with the collision frequency γ = 1.0. The random seeds were defined by the current time and date. The particle-mesh Ewald (PME) procedure was applied to the long-range electrostatic interaction. A cutoff of 10 Å was set for van der Waals interactions and short-range electrostatics. The SHAKE algorithm was used for the bond's interaction omitting the H-bonds. Every 5,000 steps, the coordinates would be written into the mdcrd file. The simulation was repeated three times for each complex.

Cluster Analysis

Cluster analysis was applied to MD trajectories to classify and make sense of information in trajectories. We used the k-means algorithm (Shao et al., 2007), which generated seed points at the start. Then, we iterated all the data points and assigned each of them to the closest seed point. Then, the most representative structures were generated in each cluster for further analysis.

Molecular Mechanics Poisson–Boltzmann Surface Area (MM/PBSA) Calculations

MM/PBSA was performed using the MMPBSA.py to evaluate the most essential residues in the complex between ligands and the receptor or the cofactors and the receptor with a large contribution to the free binding energy (Chong et al., 2009). The binding free energy was calculated as the total Gibbs free energy changes before and after the binding of ligands or cofactors.

$$\Delta G_{binding} = \Delta G_{complex} - \Delta G_{receptor} - \Delta G_{ligand}.$$

Gibbs free energy mainly consists of three parts: solvation energy (G_{solv}), molecular mechanical energy (E_{MM}), and the entropic compartments ($-TS$).

$$\begin{aligned}\Delta G_{binding} &= \left(E_{MM,\,complex} - E_{MM,\,ligand} - E_{mm,\,receptor}\right) \\ &+ \left(G_{solv,\,complex} - G_{solv,\,receptor} - G_{solv,\,ligand}\right) \\ &- \left(TS_{complex} - TS_{ligand} - TS_{receptor}\right).\end{aligned}$$

Thus, the equation can turn into this formation:

$$\Delta G_{binding} = \Delta E_{MM} + \Delta G_{solv} - \mathrm{TS}.$$

Furthermore, ΔE_{MM} can be divided as follows:

$$\Delta E_{MM} = \Delta E_{vdw} + \Delta E_{ele} + \Delta E_{int},$$

where ΔE_{vdw} is the van der Waals component, ΔE_{ele} is the electrostatic component, and ΔE_{int} is the internal component with angles, bonds, and torsional energies.

According to Poisson–Boltzmann continuum solvent model, ΔG_{solv} can be divided as:

$$\Delta G_{solv} = \Delta E_{PB} + \Delta E_{nonpolar},$$

where ΔE_{PB} stands for the polar part and $\Delta E_{nonpolar}$ stands for the nonpolar part using solvent-accessible surface area (SASA) for calculation.

$$\Delta E_{nonpolar} = \gamma SASA + b.$$

The surface tension parameter was set to 0.00542 $kcal \cdot mol^{-1} \cdot \AA^{-2}$ and the solvent parameter was 0.92 $kcal/mol$. Given that the five systems were similar with low RMSDs, the $-TS$ could be ignored in our calculations.

Dynamic Cross-Correlation Matrix (DCCM) Analysis

All trajectories were simplified using only the Cα atoms that were rotated and translated using a least-square fitting procedure (Hünenberger et al., 1995; Li et al., 2021). For the two Cα atoms i and j at time t, the position vectors are $\boldsymbol{r_i}(\boldsymbol{t})$ and $\boldsymbol{r_j}(\boldsymbol{t})$, respectively. Correspondingly, the covariance matrix element c_{ij} had the following equation:

$$\begin{aligned}\boldsymbol{C_{ij}} &= <(\boldsymbol{r_i} - <\boldsymbol{r_i}>)\left(\boldsymbol{r_j} - <\boldsymbol{r_j}>\right)> = <\boldsymbol{r_i}\,\boldsymbol{r_j}> - <\boldsymbol{r_i}><\boldsymbol{r_j}> \\ &= \frac{\Delta t}{t_{aver}}\left[\sum_{t=0}^{t_{aver}-\Delta t} r_i(t) r_j(t) - \frac{\Delta t}{t_{aver}}\left(\sum_{t=0}^{t_{aver}-\Delta t} r_i(t)) \times \left(\sum_{t=0}^{t_{aver}-\Delta t} r_j(t)\right)\right)\right],\end{aligned}$$

where Δt stands for the time interval between two frames and t_{aver} stands for average time. Covariance can be used in estimating systems' entropy (Karplus and Kushick, 1981; Swegat et al., 2003). The cross-correlation matrix element, c_{ij}, was defined as:

$$C_{ij} = \frac{c_{ij}}{c_{ii}^{\frac{1}{2}} c_{jj}^{\frac{1}{2}}} = \frac{<r_i r_j> - <r_i><r_j>}{\left[\left(<r_i^2> - <r_i>^2\right)\left(<r_j^2> - <r_j>^2\right)\right]^{\frac{1}{2}}}.$$

Dynamic Network Analysis

In order to reveal the underlying mechanisms of residue–residue interactions, we performed dynamic network analysis to calculate group constitution within the GR. According to this algorithm, the whole GR could be seen as a bunch of nodes. Nodes sitting within a threshold of 4.5Å for at least 75% throughout the trajectories could be seen as a group. We used $d_{ij} = -\log(|c_{i,j}|)$ to calculate the edges between each group. The i and j represented two nodes and C_{ij} could be calculated using the equation mentioned earlier. We also investigated the optimal and suboptimal pathways between two certain nodes using the Floyd–Warshall algorithm. All the procedures could be done using the NetworkView plugin in VMD (Hünenberger et al., 1995; Sethi et al., 2009).

RESULTS

Different Agonists' Binding Induces Distinct TIF2 Conformations

Three independent rounds of 1 μs MD simulations for five systems were conducted to probe into the dynamic conformational changes induced by different agonists. The

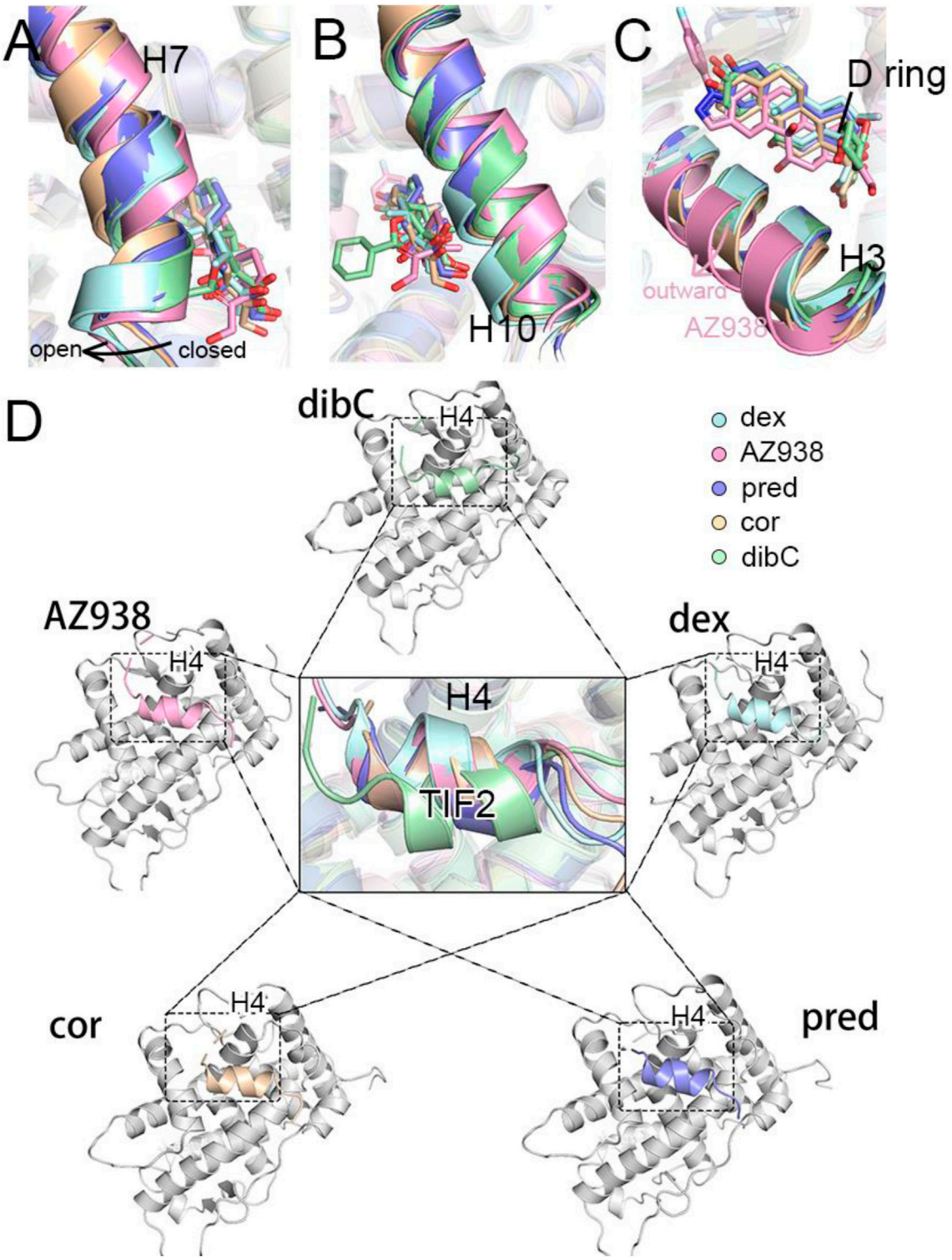


FIGURE 2 | Representative structures of five systems. **(A)** cartoon representations of H7 and ligands. **(B)** cartoon representations of H10 and ligands. **(C)** cartoon representations of H3 and ligands. **(D)** cartoon representations of TIF2 in the cofactor pocket.

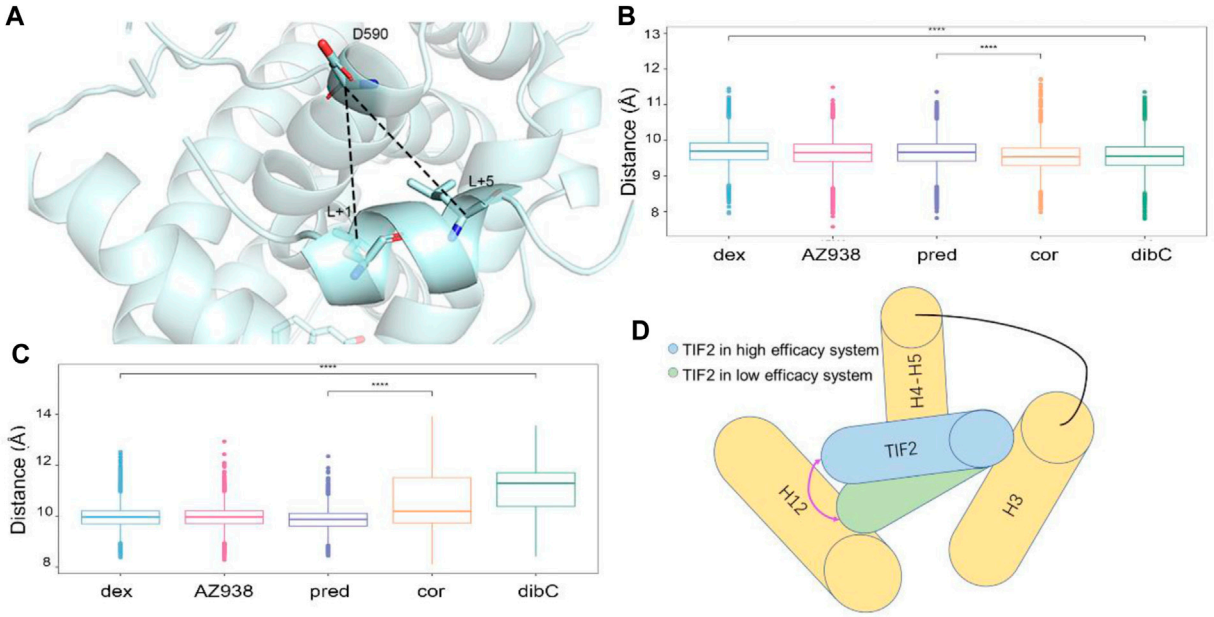


FIGURE 3 | Measurement of the direction of TIF2. **(A)** cartoon structure of the distance measured. Pivotal inter-residue distance reflecting the direction of TIF2 between **(B)** D590 and L+5 and **(C)** D590 and L+1. **(D)** model of TIF2-binding dynamics.

root-mean-square deviation (RMSD) of the C_α atoms was calculated relative to the initial structure to compare the overall conformational dynamics of the five systems. As shown in **Figure 1A**, all systems reached equilibrium after simulations. The RMSD fell into the range of 2.5–3 Å. Systems possessing ligands with higher efficacy had a slightly lower RMSD, suggesting that different ligands had induced subtle differences in the response of GR. This may indicate that the allosteric effects of ligands might differentially influence the overall energy landscape of GR. To uncover the domain-specific dynamics of GR, we calculated per-residue root mean square fluctuation (RMSF) of each system (**Figure 1B**). No significant domain-specific conformational differences between the five systems were observed during the simulations.

To identify the potential region that could contribute to conformational dynamics, we extracted representative structures for each system using cluster analysis. As shown in **Figure 2**, the representative structure of each system was superimposed on the dex-bound GR with no large structural deviation observed. However, in the systems with less bulky ligands (**Figure 1C**), the H7 regions formed a slightly "closed" conformation on the ligand-binding site (**Figure 2A**). In contrast, in the systems with bulky agonists such as dibC and dex, this region formed a more "open" conformation due to the steric hindrance of the bulky chemical groups at the tail of the D-ring. However, the conformation of the H10 appeared to be the opposite. The systems with a more "open" conformation at the H7 tended to be more "closed" at the H10, indicating that the distance between the C terminus of H10 and agonists exhibited a negative correlation with the distance between H7 and agonists (**Figure 2B**). Exceptionally, AZ938 lacks a conserved 3-ketone head but has a much bigger and electronegative fluorophenylpyrazole (**Figure 1C**). This unique structure of AZ938 resulted in the expansion of the top half of the ligand-binding pocket (**Figure 2C**). This expansion led to the outward movement of the H3 in the AZ938-bound GR (**Figure 2C**). More intriguingly, a much stronger correlation was found in the cofactor-binding pocket. By aligning all the representative structures, the combination direction of the cofactor TIF2 was found to be correlated with the efficacies of agonists. As shown in **Figure 2D**, the TIF2 in the dex-bound GR adopted a conformation closest to the H4. The TIF2 moved in an anticlockwise direction slowly in the sequence of the descending order of efficacy, which pulled the cofactor further away from the H4.

To verify whether this discrepancy in the direction of TIF2 was observed in all three independent replicas of simulations, we measured two pair-wise distances throughout the trajectories (**Figure 3A**). The distributions of distances between D590 and L+5 indicated that the structure of the C-terminal TIF2 was conserved among the five systems (**Figure 3B**). However, a distinct discrepancy could be found in the distance between D590 and L+1 (**Figure 3C**), indicating that within TIF2, the N terminus was dynamic while the C terminus was relatively stable. The distances between D590 and L+1 were roughly consistent with the order of efficacy, implying a significant role of D590 in the communication with the cofactor TIF2. In addition, no significance was found for distances between dex- and AZ938-bound systems, elucidating that the adopted conformation of these two was preferential for higher efficacy (**Figure 3B**). Interestingly, the fluctuation of distances in the dibC- and cor-bound systems was much larger than that in the rest of the

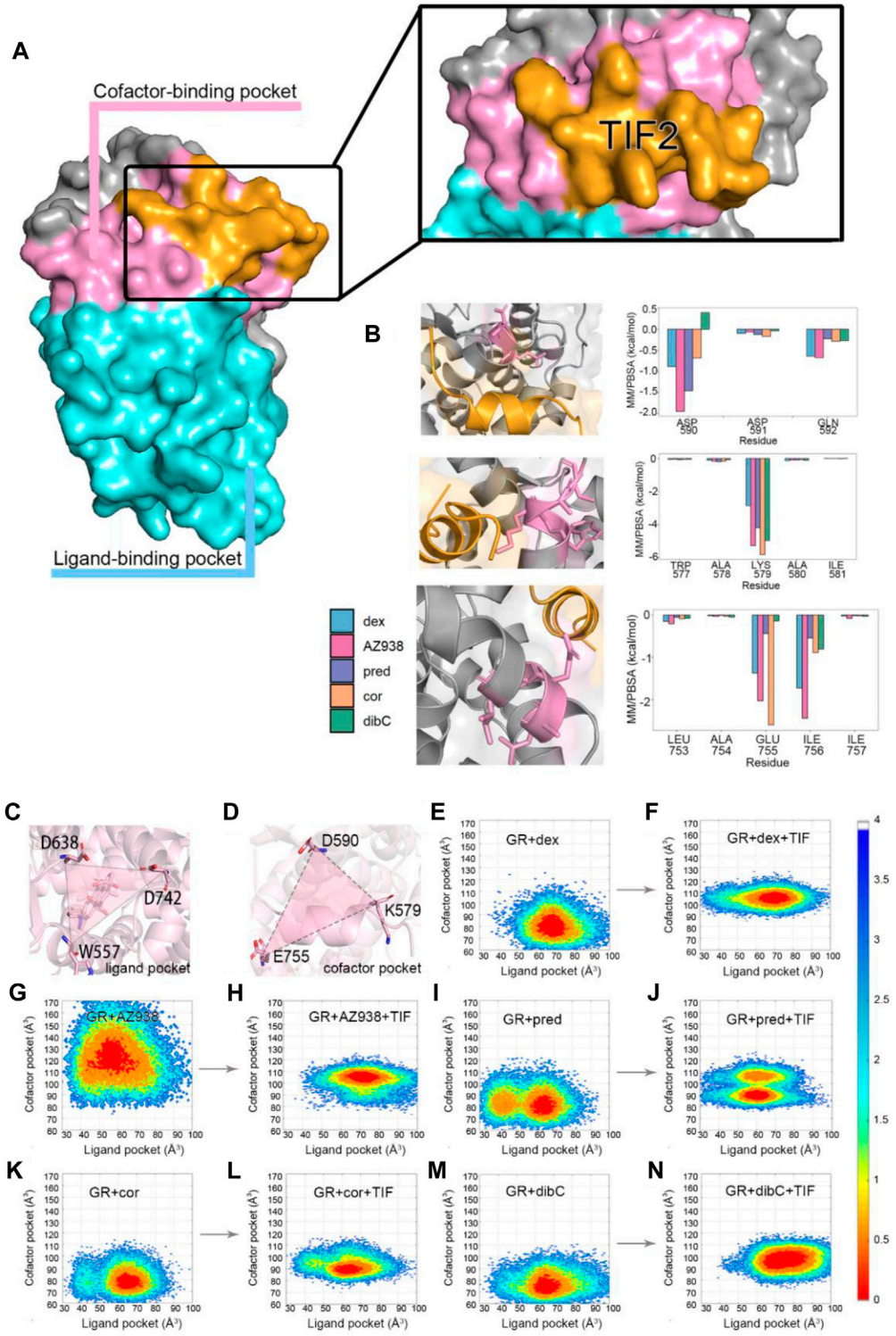


FIGURE 4 | (A) cartoon representation of two pockets. **(B)** MM/PBSA of three helices with crucial residues. **(C)** cartoon representation of parameter representing ligand pocket. **(D)** cartoon representation of parameter representing cofactor pocket. Conformation FEL of dex, AZ938, pred, cor, and dibC with or without TIF2 binding **(E–N)**. The landscape was generated with $\Delta_{D638\text{-}D742\text{-}W557}$ and $\Delta_{D590\text{-}K579\text{-}E755}$.

TABLE 1 | Free energy analysis (kcal/mol) for K579, D590, and E755.

K579[a]	Dex-bound system	AZ938-bound system	Pred-bound system	Cor-bound system	DibC-bound system
Δ*Evdw*	−2.21 (0.93)	−2.26 (1.13)	−2.06 (0.97)	−1.53 (1.02)	−1.56 (1.17)
Δ*Eele*	−88.93 (7.53)	−101.14 (9.71)	−85.06 (9.44)	−93.99 (5.62)	−101.54 (7.41)
Δ*Enonpolar*	−0.62 (0.08)	−0.66 (0.06)	−0.63 (0.08)	−0.71 (0.07)	−0.65 (0.06)
Δ*Esolv*	88.90 (6.76)	98.77 (7.66)	83.55 (7.97)	90.40 (4.49)	98.76 (6.20)
Δ*Ebinding*	−2.86 (1.31)	−5.29 (2.06)	−4.20 (1.75)	−5.83 (1.41)	−4.98 (1.56)
D590	**Dex-bound system**	**AZ938-bound system**	**Pred-bound system**	**Cor-bound system**	**DibC-bound system**
Δ*Evdw*	−0.09 (0.52)	0.17 (0.87)	0.13 (0.89)	−0.36 (0.61)	−0.43 (0.20)
Δ*Eele*	−3.50 (10.86)	−8.11 (3.49)	−23.41 (10.31)	−27.30 (8.66)	10.41 (4.57)
Δ*Enonpolar*	−0.08 (0.06)	−0.16 (0.02)	−0.09 (0.06)	−0.15 (0.06)	−0.05 (0.05)
Δ*Esolv*	2.79 (9.43)	6.16 (3.18)	21.92 (8.57)	27.14 (8.35)	−9.54 (4.56)
Δ*Ebinding*	−0.89 (1.37)	−1.95 (0.76)	−1.46 (1.75)	−0.68 (0.71)	0.39 (0.32)
E755	**Dex-bound system**	**AZ938-bound system**	**Pred-bound system**	**Cor-bound system**	**DibC-bound system**
ΔEvdw	−2.33 (0.73)	−2.54 (0.97)	−1.12 (0.68)	−2.40 (0.78)	−1.46 (0.88)
Δ*Eele*	−20.95 (5.65)	−32.31 (3.29)	−44.89 (11.18)	−62.71 (5.55)	−24.10 (3.85)
Δ*Enonpolar*	−0.51 (0.07)	−0.57 (0.03)	−0.40 (0.06)	−0.54 (0.06)	−0.41 (0.08)
Δ*Esolv*	22.40 (5.33)	33.38 (3.07)	45.97 (10.61)	63.05 (5.44)	25.82 (3.94)
Δ*Ebinding*	−1.37 (1.35)	−2.03 (1.22)	−0.44 (1.37)	−2.60 (1.02)	−0.14 (0.82)

[a]*Numbers in the parentheses are the standard deviations.*

systems, suggesting that they went through severe vibration during the simulation, which illustrated that interaction with D590 could also be important in the stability of TIF2. Previous studies had already investigated the important residues in the cofactor-binding pocket, which interacted with the conserved sequence (LXXLL) on the TIF2 (Necela and Cidlowski, 2003; Liu X. et al., 2019). The D590 on the H4 was proved to be one of the essential residues. This gave us insights into the importance of D590. Thus, we hypothesized that D590 could be an essential residue in the change of the TIF2 conformation.

Communication Between Ligand- and Cofactor-Binding Pockets Indicates Connection of the Regulation Between Two Pockets to Efficacy Discrepancy

The superposition of representative structures had important implications for allosteric communication in the GR. Consequently, the free energy landscape was projected onto the 2D space using parameters reflecting the situations of pockets (**Figure 4**) (Lu et al., 2021b). To quantify the influence of residues in the binding pockets on the energetics of TIF2 binding, molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) was employed to compute the binding free energy ($\Delta G_{binding}$) of TIF2 to GR, which was divided among each residue (**Table 1**). The lower binding free energy indicated a stronger interaction between TIF2 and the residue. In each of the three helices that surrounded the cofactor-binding pocket, we selected three residues with a large contribution to MM/PBSA. D590, K579, and E755 were selected, respectively, which was consistent with the results in previous studies (Suino-Powell et al., 2008; Veleiro et al., 2010; Alves et al., 2020) to mimic the area of the cofactor-binding pocket (**Figure 4A**).

Consequently, D590, K579, and E755 were chosen to be the three residues defining the parameter of the triangle that reflected the relative degree of openness of cofactor-binding pockets (**Figure 4B**). The other parameter reflecting the openness of the ligand-binding pocket was defined by the triangle representing the ligand-binding pocket, which was formed by three residues (W557, D638, and D742) shown in **Figure 4C**. They were all located at the terminus of the helices constituting the ligand-binding pocket, which reflected the fluctuation of the pocket sensitively. The two areas of triangles were used as the parameters to generate the two-dimensional landscape for each system, which reflected the correlation of the openness of the two pockets. Additional five systems without the cofactor TIF2 have also conducted simulations for the purpose of comparing the landscape before and after the binding of the cofactor. The same parameters were used for the two-dimensional landscape for the five systems without TIF2. By comparing the distribution of the area of the two pockets, we could profile the difference of each system in the response to TIF2's binding.

As shown in **Figure 4**, two distinct states were observed before and after the binding of TIF2. Before the binding of the cofactor, the five systems mutually exhibited a conformational state with the ligand-binding pocket area of approximately 65 $Å^2$. The area of the cofactor-binding pocket was around 80 $Å^2$ in the mutual state with AZ938 to be an exception. A trend for a second preferential conformation at the left of the original one was also discovered in the pred-, cor-, and dibC-bound systems. After cofactor binding, the parameters condensed into a state at the up-right of the plot, with both parameters enlarged. The binding of TIF2 not only influenced the area of the cofactor-binding pocket but also affected the ligand-binding pocket, which implied the allosteric communication between the ligand-binding pocket and

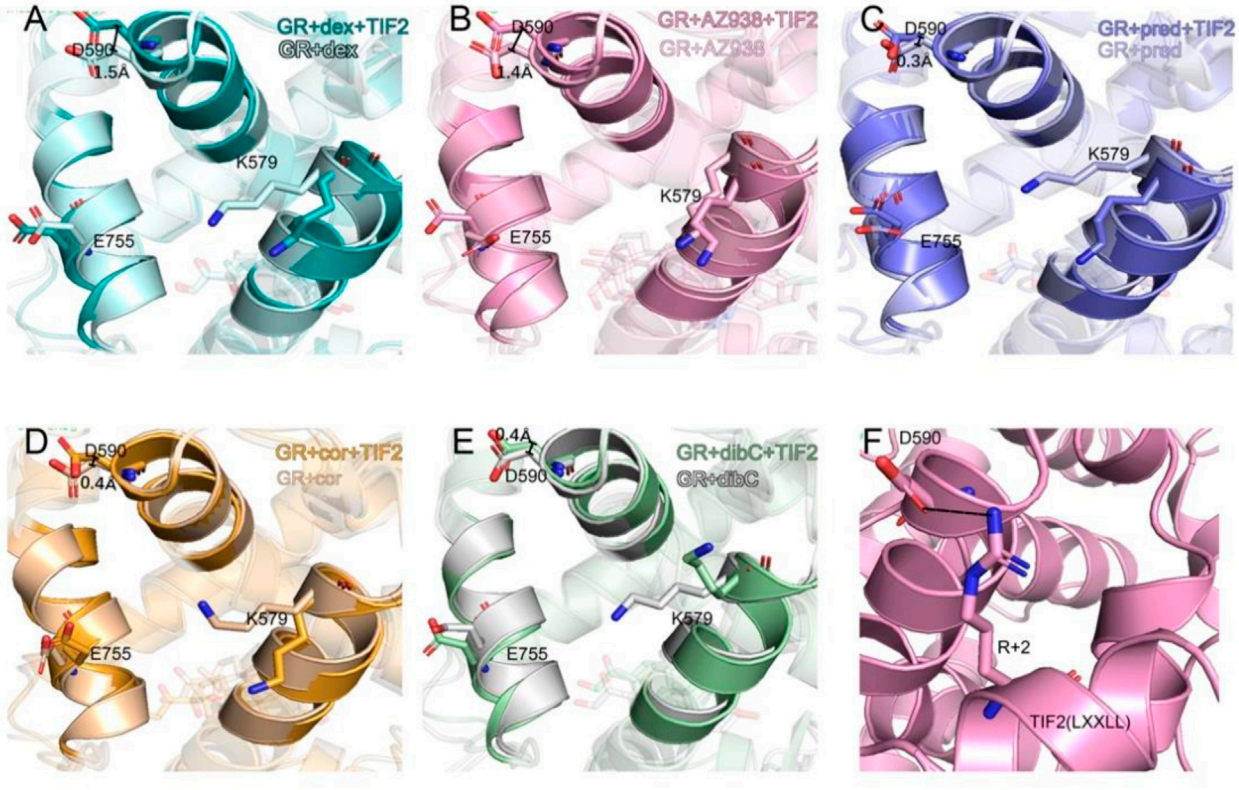


FIGURE 5 | Representative structures of the dex-bound system **(A)**, AZ938-bound system **(B)**, pred-bound system **(C)**, cor-bound system **(D)**, and dibC-bound system **(E)**. In each system, TIF2-bound and TIF2-unbound systems were aligned. Hydrogen bonds between D590 and R+2 are shown in **(F)**. The distance of CB atom before and after the binding of TIF2 was determined.

cofactor-binding pocket. Despite having inconsistent landscapes before the binding of TIF2, all systems converged their conformational landscape after TIF2's combination. Intriguingly, parameter $\Delta_{D590\text{-}K579\text{-}E755}$ in the dex- and AZ938-bound systems increased significantly to around 100–110 Å (C^{up}), which was 10 Å more than the increase in the cor-bound and dibC-bound systems (C^{down}). This illustrated that the level of conformational changes induced by the binding of TIF2 was different in each system, probably by influencing the interaction between the two pockets, which might result in the different efficacies of agonists. Interestingly, the coexistence of C^{up} and C^{down} was observed in the pred-bound system, which exhibited features of both agonists with high efficacy (dex-bound and AZ938-bound systems) and low efficacy (cor-bound and dibC-bound systems). These results further verified that these two features regarding the area of two pockets may sensitively reflect the efficacy of the agonists. After the binding of TIF2, the constriction of the ligand-binding pocket conformation was much stronger in the pred-bound system than that in the dibC- and cor-bound systems, implying a stronger response toward the binding of the cofactor in the pred-bound system. Altogether, the results indicated the pocket conformational changes induced by the TIF2 could reflect the efficacies of ligands.

Representative Structures Indicate That D590 May Be an Important Residue

To further investigate the conformation of the chosen residues in the cofactor pocket, the representative structures were extracted from each two-dimensional landscape (**Figure 5**). Obvious expansion of the three helices forming the cofactor-binding pocket (H3, H4, and H12) occurred in the dex- and AZ938-bound systems (**Figures 5A, B**). In the cor- and dibC-bound systems, no significant expansion was observed (**Figures 5D, E**). The outward movement of the H3 that contains K579 was the most distinct one among the three helices. After the binding of TIF2, K579 all rotated outward, except the one in the dibC-bound system, which flipped away and formed a weak interaction with TIF2. The expansion of H4 only occurred in dex-bound and AZ938-bound systems. In the systems of TIF2-bound and TIF2-unbound, no significant changes occurred in the representative structures of H4 in pred-bound, cor-bound, and dibC-bound systems, which only underwent slight rotation in D590. However, obvious outward movements of D590 and H4 were observed in dex and AZ938. The dynamic conformation of D590 induced a strong interaction between the O atoms in D590 and H atoms in the conserved sequence of LXXLL (**Figure 5F**), which participated in the stabilization of TIF2. The LXXLL was important in the binding of the AF2 and activation of

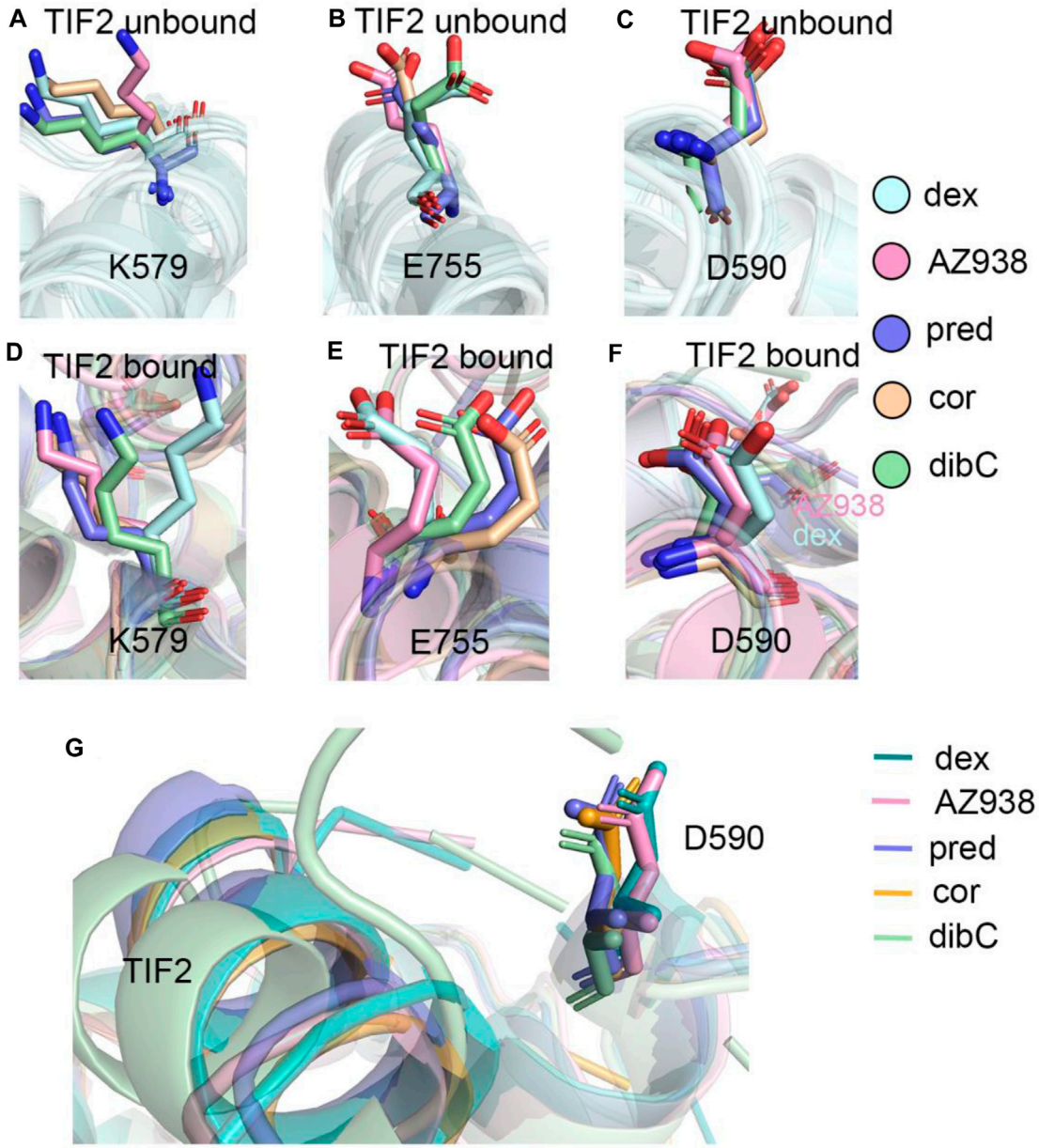


FIGURE 6 | (A) representative structures of K579 in the TIF2-unbound systems. **(B)** representative structures of E755 in the TIF2-unbound systems. **(C)** representative structures of D590 in the TIF2-unbound systems. **(D)** representative structures of K579 in TIF2-bound systems. **(E)** representative structure of E755 in TIF2-bound systems. **(F)** representative structure of D590 in TIF2-bound systems. **(G)** representative structure of TIF2 and D590 in TIF2-bound systems.

transcription, thereby having direct relationships with agonists' efficacy (Heery et al., 1997; Torchia et al., 1997; Plevin et al., 2005). The unique conformational dynamics in dex-bound and AZ938-bound systems implied an important conformation contributing to the higher efficacy of dex and AZ938 (Heery et al., 1997).

Given the unique expansion of D590 in dex-bound and AZ938-bound systems, the D590 was further investigated given that it might be a crucial residue in the allosteric communication between the ligand-binding pocket and cofactor-binding pocket. K579 and E755 were observed to have various conformations before and after the binding of TIF2 (**Figure 6**). No evidence showed that the pattern of K579 and E755 conformation had a relation with the order of efficacy between different systems (**Figures 6A, B**). However, the conformation of D590 was consistent among the five systems both before and after the binding of TIF2, respectively (**Figures 6C–F**). The conformation of D590 almost overlapped in the five systems of TIF-unbonded GR. However, a discrepancy was shown in **Figure 6F** after the binding of TIF2, with the D590 in dex and AZ938 moved slightly outward and separated from the rest of D590 in other systems, despite the overall conformation

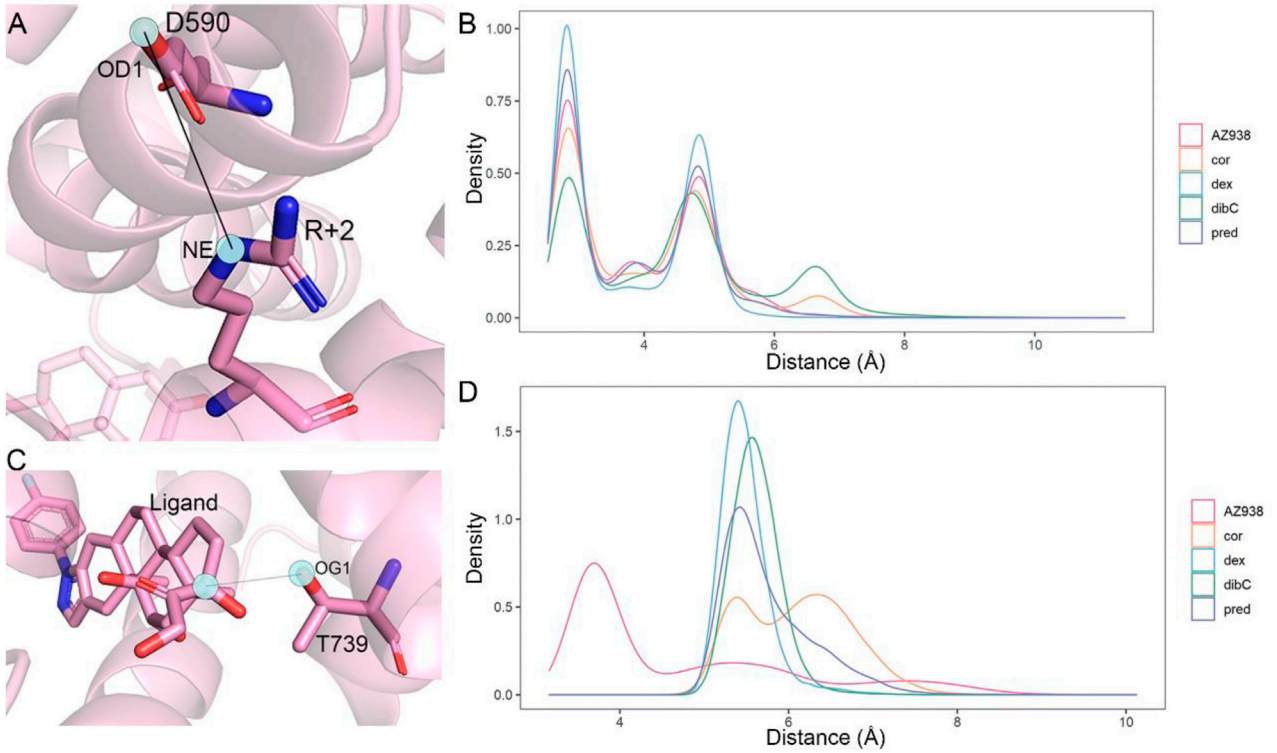


FIGURE 7 | Distance measurement of two pairwise atoms. **(A)** cartoon representation of the distance between OD1 in D590 and R+2 in NE. **(B)** distance distribution density of D1 in D590 and NE in R+2 in five systems. **(C)** cartoon representation of the distance between OG1 in T739 and the C atom in agonists. **(D)** distance distribution density of OG1 in T739 and corresponding C atoms in each agonist (C25 in the AZ938-bound system, C15 in the dibC-bound system, C18 in the pred-bound system, and C17 in dex- and cor-bound systems).

being consistent between the five systems. This was accompanied by the tight loading of TIF2, which pushed the D590 away from the original conformation (**Figure 6G**), suggesting an underlying mechanism that TIF2's binding may be related to the conformational dynamics of D590. Altogether, the representative structures between the 10 systems revealed a potential important residue for communications between the ligand-binding pocket and cofactor-binding pocket.

Identification of Two Important Residues in the Ligand-Binding Pocket and Cofactor-Binding Pocket

Comparative analyses of the representative structure of the cofactor-binding pocket emphasized the importance of the special relationship between D590 and TIF2. Previous crystal structure analysis also revealed that D590 formed vital hydrogen bonds with the conserved residue R+2 on TIF2 (Suino-Powell et al., 2008). Thus, we calculated the distance distribution of D590 and R+2. Since the oxygen atoms in D590 could form various hydrogen bonds with different N atoms in R+2, we analyzed one pair that could best represent the relationship between these two residues. As shown in **Figure 7A**, the atom OD1 and atom NE were selected from D590 and R+2, respectively, for distance measurement since these two atoms formed a stable hydrogen bond throughout the three replicas of simulations. The density distribution of distances was shown in **Figure 7B**. Dex-bound system, the system with the most obvious expansion of D590, had the highest peak of density distribution within 5 Å, while the dibC-bound system had the lowest distribution of distances in this region. The distribution of the dex-bound system rapidly fell to zero beyond 3.5 Å of the distance. The distribution peak of other systems was also significantly lower than that of the dex-bound system. This indicated that the dex-bound system was the most likely system to form the hydrogen bonds between OD1 in D590 and NE in R+2 since hydrogen bonds were considered unable to form in two atoms with a distance larger than 3.5 Å. The highest peaks of the dex-bound system might correspond to the preferential structures of hydrogen bonds, which persistently existed during simulations. Intriguingly, another small peak at a distance of around 7 Å was also observed in dibC-bound and cor-bound systems, where the hydrogen bonds were almost unlikely to form. This implied that dibC-bound and cor-bound systems had an additional sub-preferential conformation in a state that D590 would not form hydrogen bonds with R+2. All these properties of the density distribution illustrated that the dex-bound system might be the most suitable for the formation of hydrogen bonds between OD1 in D590 and NE in R+2, while cor- and dibC-bound systems were less favorable for the formation of hydrogen bonds. The conserved

TABLE 2 | Free energy contribution (kcal/mol) by residue and the corresponding free energy difference of H10.

Residue[a]	Dex-bound system	AZ938-bound system	Pred-bound system	Cor-bound system	DibC-bound system
L732	−1.54 (0.29)	−1.32 (0.20)	−1.11 (0.36)	−1.26 (0.25)	−1.30 (0.27)
L733	−0.18 (0.04)	−0.13 (0.04)	−0.12 (0.05)	−0.13 (0.04)	−0.12 (0.04)
N734	−0.06 (0.03)	−0.02 (0.02)	−0.05 (0.04)	−0.03 (0.03)	−0.03 (0.03)
Y735	−1.97 (0.37)	−0.98 (0.23)	−1.40 (0.61)	−1.41 (0.33)	−1.99 (0.52)
C736	−1.07 (0.32)	−1.06 (0.24)	−1.76 (0.63)	−1.06 (0.33)	−1.06 (0.23)
F737	−0.00 (0.03)	0.02 (0.02)	0.00 (0.04)	0.02 (0.02)	0.01 (0.02)
Q738	−0.00 (0.03)	0.01 (0.02)	0.03 (0.03)	−0.02 (0.03)	−0.02 (0.03)
T739	−2.35 (0.64)	−0.23 (0.13)	−0.93 (0.56)	−2.52 (0.52)	−2.65 (0.45)

[a]*Numbers in the parentheses are the standard deviations.*

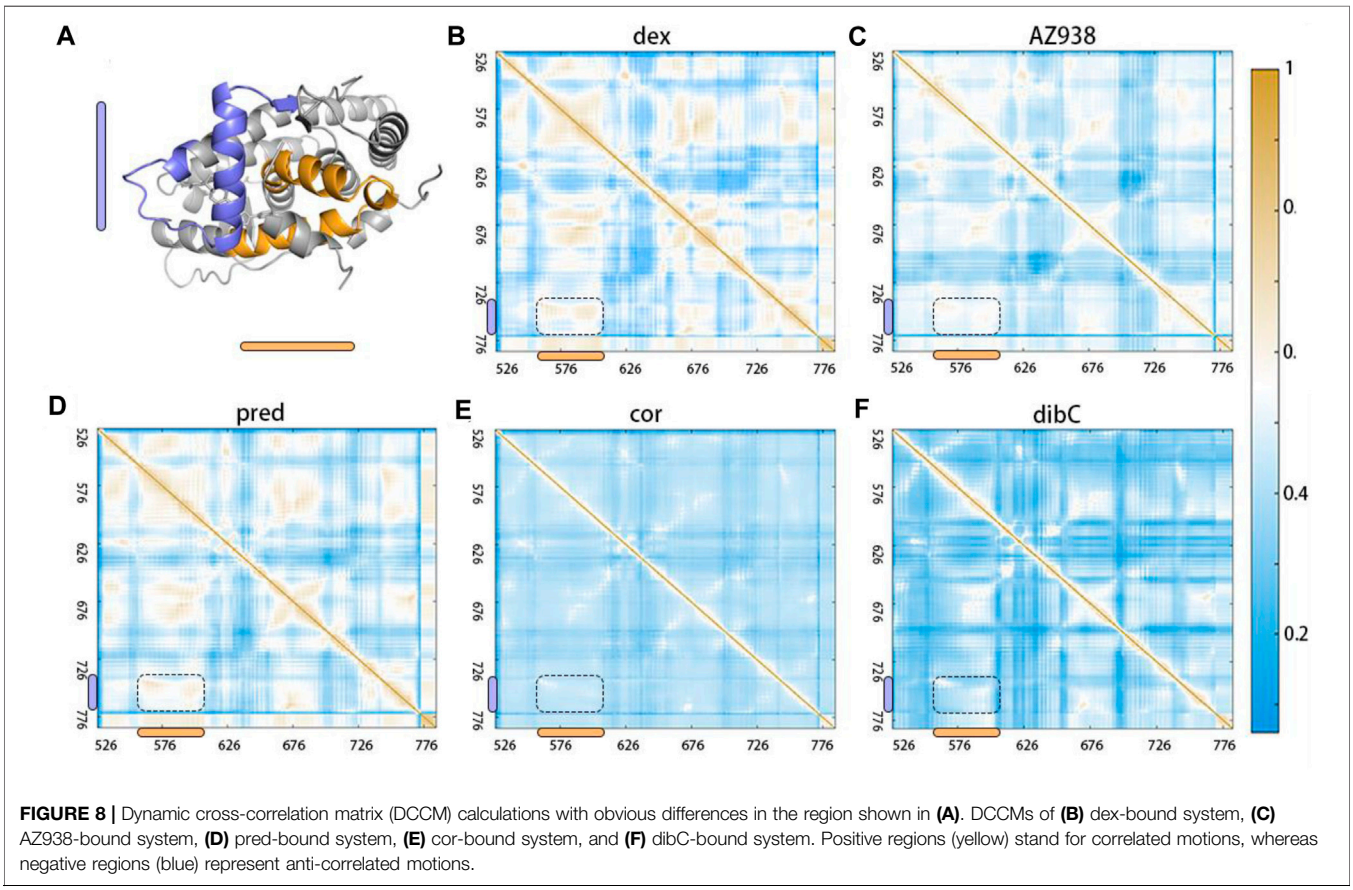


FIGURE 8 | Dynamic cross-correlation matrix (DCCM) calculations with obvious differences in the region shown in **(A)**. DCCMs of **(B)** dex-bound system, **(C)** AZ938-bound system, **(D)** pred-bound system, **(E)** cor-bound system, and **(F)** dibC-bound system. Positive regions (yellow) stand for correlated motions, whereas negative regions (blue) represent anti-correlated motions.

hydrogen bond served as a connection between the GR and TIF2 and was thought to have contributed to the efficacy of ligands. Therefore, this finding agreed to the efficacy order of the five systems.

We next attempt to investigate other residues within the ligand-binding pocket which might also be important for allosteric communication between two pockets. MM/PBSA analysis for the binding of ligand to the GR was carried out, and the result was further decomposed into every residue forming the ligand-binding pocket. Consistent with previous studies, T739 was identified as one of the most important residues for the binding of ligands (**Table 2**). Among the five systems, the T739 had a consistently large contribution to the binding free energy of ligand to the GR. Thus, we measured the distance between the T739 and the ligands in our five simulation systems, respectively, which was considered an important interaction between the ligand and its pocket. Given that different ligands had distinct structures and tended to have distinct preferences for oxygen atom to form the hydrogen bond with the T739, we selected a conserved atom that was related to all the oxygen possible in forming the hydrogen bond with the T739 (C25 in the AZ938-bound system shown in **Figure 7C**, C15 in the dibC-bound

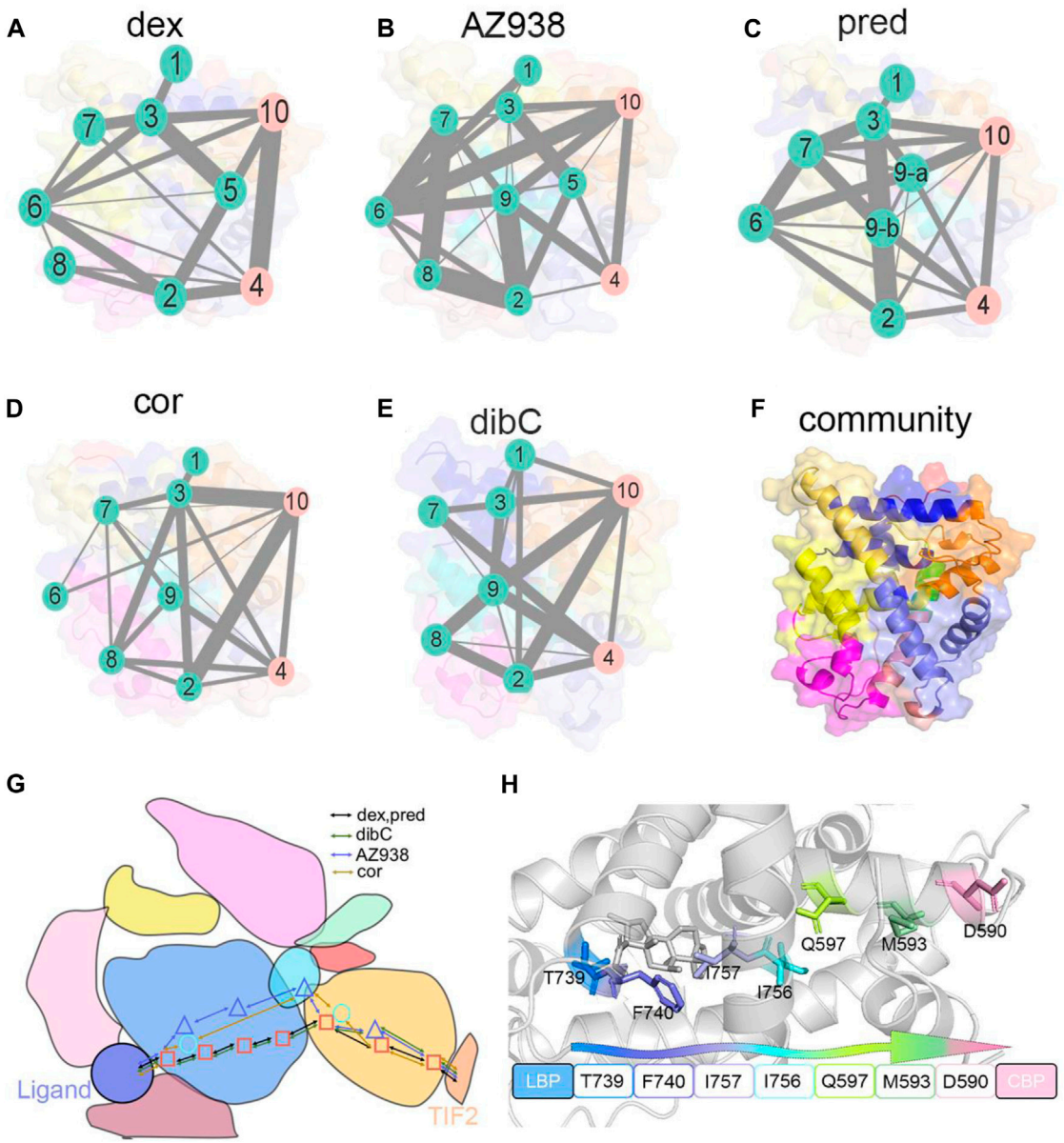


FIGURE 9 | Community networks and the allosteric signaling pathways in each simulation system. Community network representation of **(A)** dex-bound system, **(B)** AZ938-bound system, **(C)** pred-bound system, **(D)** cor-bound system, and **(E)** dibC-bound system. **(F)** cartoon representation of cluster distribution in GR. Each sphere represents an individual community, and the thickness of sticks connecting communities is proportional to the corresponding edge connectivity. Schematic representation **(G)** of the domain-level allosteric signaling pathway and cartoon representation of signal propagations **(H)** connecting D590 and T739 in five systems.

system, and C17 in dex-, pred-, and cor-bound systems). As shown in **Figure 7D**, the dex-bound system had the highest density distribution of the distance at around 5 Å. The crest of the pred-bound system was slightly farther than the dex-bound system located at around 4.5 Å. The cor-bound system presented two crests, both of which were farther than 5 Å, indicating a less favorable condition to form hydrogen bonds between the ligand and T739. AZ938 was an exception with an obvious smaller distance between the ligand and T739. This was due to the six rings of AZ938 which contributed to the elongated chemical structure. In order to fit into the ligand-binding pocket with this unusual structure, AZ938 folded its tail at the D ring toward the direction of H7, while the C25 in AZ938 was exposed to the T739. As a result, the distance distribution of the AZ938-bound system contributed to the decrease in the peak distance. However, from the position of the peaks in the five systems, we concluded that the distance between the T739 and the selected atom in the ligand was able to show the different characteristics of ligands.

Elucidation of Allosteric Communication Pathway in Two Chosen Residues

After identifying the critical T739 and D590 in the ligand-binding pocket and cofactor-binding pocket, we next tried to explore the potential allosteric pathways connecting them. Using dynamic cross-correlation matrix (DCCM) calculations, we provided an overview of the inter-residue correlations within the simulation systems (Wang et al., 2022). Residues distributed in regions representing two sets of residues located near the ligand-binding pocket and cofactor-binding pocket demonstrated the biggest changes in the whole system. As shown in **Figure 8**, compared to the cor-bound system and dibC-bound system, the dex-bound system exhibited significantly increase correlated motions among distant residues. In the dex-bound system, obvious correlations between around G568 and around D590 suggested communication between H3 and H4–H5 (**Figure 8B**), indicating a certain correlation within the cofactor-binding pocket. Particularly, in the dex-bound system, the correlation of inter-molecular motions among the region near the ligand-binding pocket and cofactor-binding pocket colored by yellow and blue bars (framed using black dash lines) was compellingly strengthened than the other four systems. Pred-bond and AZ938-bound systems possessed weaker correlated motions in this region than dex-bound systems (**Figures 8C, D**) but were relatively stronger than the dibC-bound and cor-bound systems (**Figures 8E, F**). Weakened correlative movements between the ligand-binding pocket and the cofactor-binding pocket in dibC-bound and cor-bound systems suggested impaired signal propagation pathways between the ligand-binding pocket and cofactor-binding pocket. The degree of correlated motion levels in five systems could also partly reflect the different allosteric regulations among the five systems. Notably, the two residues discussed before were also in this region, which served as another evidence for their role in allosteric communication between the two pockets.

Next, community network analysis and allosteric pathway analysis were carried out for the five systems to systematically investigate the allosteric networks (Wang et al., 2021). During the three replicas of simulations, residues that distanced within a 4.5 Å cut-off for at least 75% of the time were categorized into the same community, which could be seen as a congenerous unit within the systems (Qiu et al., 2021; Zhuang et al., 2022). As shown in **Figures 9A–E**, different systems were divided into different quantities of communities. Each community was represented by a colored circle and was superimposed on the 2D structure of the corresponding protein complex to reflect the relative positions with adjacent communities. Based on graph theory and topology, each community's structural information flow was calculated (Sethi et al., 2009). The width of lines connecting two communities was proportional to the corresponding edge connectivity which was defined by the number of shortest paths passing through the edging nodes. In general, the residual components of each community were similar in the five systems. However, discrepancies between different systems still occurred. In the AZ938-bound system (**Figure 9B**) and dibC-bound system (**Figure 9E**), the complex was divided into 10 groups and eight groups, respectively, while in the other four systems, the complexes were divided into nine communities. Some communities were not consistently existed in all the five systems. For instance, community 9 was absent in the dex-bound system and community 6 was absent in the dibC-bound system. However, community 4 and community 10 consistently existed in five systems. They contained domains regarding the ligand-binding pocket and cofactor-binding pocket and the constituent residues within were similar among the five systems, indicating a critical role of these domains in allosteric communication. In the dex-bound system (**Figure 9A**), the connection between communities 4 and 10 was direct and strong. In contrast, the connection of communities 4 and 10 was much weaker in dibC- and cor-bound systems (**Figures 9D, E**), suggesting less informational communication through these two communities in these two systems. The thickness of the lines in communities 4 and 10 was in positive correlation with the order of efficacies of five systems, indicating that the communication between these two parts of the ligand-binding pocket and cofactor-binding pocket might dominate the differences in the ligand's efficacy. However, the connection of communities 4 and 10 in dibC- and cor-bound systems were relatively weak, suggesting some structural impairment in these two systems. Such loosen connection in dibC-bound and cor-bound systems may due to the lack of community 5. In the dex-bound system, community 5 served as a major hub for information transduction. It connected communities 2 and 10, which indirectly strengthened the connection between communities 4 and 10. A similar impact was also observed in community 9 in AZ938-bound and pred-bound systems (**Figures 9B, C**). Notably, D590 and T739 were located at community 10 and community 4, respectively, suggesting that these two residues also participated in domains that drive the communication pathways in these two communities.

Additionally, by calculating the optimal and suboptimal pathways that link D590 in community 10 and T739 in community 4, we revealed the potential allosteric pathways between the chosen residues in the five systems. As shown in **Table 3**, the number of residues involved in the optimal pathways from T739 to D590 was similar in the five systems. However, the AZ938-bound system, pred-bound system, and dex-bound system displayed shorter optimal pathways, with a length of around 300, which indicated a stronger relationship between two chosen residues than in cor- and dibC-bound systems. (**Figures 9G, H**). Therefore, it could be concluded that the allosteric pathway between D590 and T739 was stronger in dex- and AZ938-bound systems than that in dibC- and cor-bound systems, which might also influence the efficacy of ligands. These results, together with DCCM analyses, collectively demonstrated that ligand-induced allosteric communications between the ligand-binding pocket and cofactor-binding pocket were one of the driving forces for the discrepancy of ligand's efficacy.

DISCUSSION

GR, as an essential nucleus receptor, controls a myriad of cellular functions and signal transduction (Fowden et al., 1998; Kumar

TABLE 3 | Allosteric pathway analysis between D590 and T739.

	Length	Residue	Pathway
Dex-bound system	309	7	590, 593, 597, 756, 757, 740, and 739
AZ938-bound system	272	7	590, 594, 597, 600, 733, 735, and 739
Pred-bound system	303	7	590, 593, 597, 756, 757, 740, and 739
Cor-bound system	362	6	590, 593, 596, 600, 736, and 739
DibC-bound system	451	7	590, 594, 597, 756, 757, 740, and 739

and Thompson, 1999; Meijer et al., 2018; Liu B. et al., 2019). Upon the binding of ligands, GR is activated and induces conformational changes, involving post-translation modifications such as acetylation and phosphorylation. GR then translocates into the nucleus, where GR exerts its actions through transactivation and transrepression mechanisms (Vandevyver et al., 2014), regulating various metabolic functions. Thus, GR has been used to treat various metabolism and immunological disorder-related disease. Despite its broad clinical application, the serious side effects have always bothered patients and doctors. The underlying mechanisms of allosteric communications in GR may be an instructor in GR drug designs. Allosteric communication in the N-terminal domain and DNA-binding domain of GR has been detailly elaborated by Hilser and coworkers (Li et al., 2017). However, how ligands drive the allosteric effects and influence signal transductions remain unknown. Herein, by using MD simulations, we provided structural insights into the different allosteric effects induced by different ligands, thereby motivating progress in targeting GR's ligand-binding domain for drug discovery.

By comparing the representative structures extracted from the three replicas of simulations, we revealed conformational dynamics in five systems bound to five different ligands (**Figure 2**). Conformational discrepancies in the ligand-binding pocket were largely due to the different chemical structures that ligands possessed, resulting in different degrees of openness in the ligand-binding region of H7 and H10. Conformational differences at the cofactor-binding pocket appeared much more magnificent. The directions of the TIF2's conserved LXXLL helix in different systems strictly follow the order of ligand's efficacy, with the dex-bound system having the closest distance between H4 and TIF2 and the dibC-bound system having the farthest one. This result was further testified by two pairwise distance measurements between D590 and the two ends of TIF2 (**Figure 3**). MM/PBSA analysis of residues near the cofactor-binding pocket and hydrogen bond analysis revealed that D590 on H4 was likely to be a potentially vital residue to have an impact on the conformation of TIF2.

Two-dimensional landscapes of two parameters relative to the ligand-binding pocket and cofactor-binding pocket separately were projected in five GR-ligand-TIF2 and five GR-ligand systems (**Figure 4**). The parameter representing the cofactor-binding pocket used the area of the triangle formed by three high MM/PBSA contribution residues. The parameter representing the cofactor-binding pocket used another three residues in the cofactor-biding pocket. By comparing landscapes from before and after the TIF2's binding, changes occurred both in the ligand-binding pockets and cofactor-binding pockets in the five pairs of systems, suggesting the influences of allosteric communication between two pockets in all the systems. Representative structures in the five pairs of two-dimensional landscapes were aligned and compared. Various degrees of expansion occurred in H3, while evident expansion of H4 only occurred in the dex-bound system and AZ938-bound system, which might be related to the exceeding efficacy of these two systems (**Figure 5**). Distance between two atoms in D590 and R+2, respectively, that formed a hydrogen interaction was also analyzed (**Figure 7**). Dex-bound system appeared to be the most preferential one for the formation of the hydrogen interaction, while dibC-bound and cor-bound systems had an extra peak at distance beyond 3.5 Å, suggesting less preferential conformations for hydrogen interaction. This hydrogen bond was believed to be a crucial interaction between the TIF2 and GR. Thus, the different abilities of forming the hydrogen bond in these systems might influence the efficacy of ligands. MM/PBSA analysis and distance measurements were conducted on residues around the ligand-binding pocket, and T739 was identified as an important residue with large MM/PBSA contribution and hydrogen interaction with the ligand. Distance analysis of T739 and the ligand was able to show the different qualities of ligands' binding in different systems. By applying DCCM, inter-residue correlations were investigated among the five ligands (**Figure 8**). A distinguishable discrepancy was found in correlations of the region relative to the ligand-binding pocket and cofactor-binding pocket. In the dex-bound system, the correlation was the strongest, while in dibC-bound and cor-bound systems, the correlation was much weaker, suggesting impaired allosteric communication in the two complexes. Notably, D590 and T739 were also in this region, implying their participation in the allosteric communication. To systematically investigate the allosteric networks, community network analysis and allosteric pathway analysis were carried out (**Figure 9**). We observed different levels of communication between group 4 and group 10, which was consistent with the ligands' efficacy (**Figure 9**). In addition, from community analyses and suboptimal pathway analysis, we found that the allosteric propagation pathway between two representative residues in the ligand-binding pocket and cofactor-binding pocket in five systems.

In view of the crucial role played by GR in clinical treatments (Van Staa et al., 2000), the development of new drug targeting GR has been the major focus over the past few decades. Thitherto, few accomplished design drugs with high efficacy and low side effects. This is largely due to the obstacles in the lack of knowledge of GR's allosteric effects (Ni et al., 2021). The underlying

mechanisms of what induces the discrepancy in agonists' efficacy remain elusive. Thus, our study focusing on the allosteric communications of GR's conformational dynamics is useful. Moreover, members of the NR family possess mutual structures with similar sequences. The TIF2 is the common cofactor that interacts with the AF2 interface of NRs. Thereby, it is presumable that the mechanism we unveiled in the GR also applies to others in the NR family and therefore has a more generalized value. Taken together, our study elucidated the driving force behind the ligands' efficacy induced by different agonists' binding as well as the detailed mechanism of allosteric communication between the ligand-binding pocket and cofactor-binding pocket. Our explorations of the conformational outcomes induced by the binding of different ligands have provided insights for new drug design by conditional genome manipulation or modifying ligand's interactions with its pocket.

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

SL and MX conceived and supervised the project, designed the experiments, and edited the manuscript; YS performed MD analysis and drafted the manuscript; JF and DN helped with the construction of simulated systems and the analysis of MD trajectories; SL acquired the data and revised the manuscript; MX was responsible for funding support. All authors discussed the results and reviewed the manuscript.

FUNDING

This work was supported by the Innovative Research Team of High-Level Local Universities in Shanghai.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.933676/full#supplementary-material

REFERENCES

Álvarez, L. D., Martí, M. A., Veleiro, A. S., Misico, R. I., Estrin, D. A., Pecci, A., et al. (2008a). Hemisuccinate of 21-Hydroxy-6,19-Epoxyprogesterone: A Tissue-specific Modulator of the Glucocorticoid Receptor. *ChemMedChem* 3, 1869–1877. doi:10.1002/cmdc.200800256

Álvarez, L. D., Martí, M. A., Veleiro, A. S., Presman, D. M., Estrin, D. A., Pecci, A., et al. (2008b). Exploring the Molecular Basis of Action of the Passive Antiglucocorticoid 21-Hydroxy-6,19-Epoxyprogesterone. *J. Med. Chem.* 51, 1352–1360. doi:10.1021/jm800007w

Alves, N. R. C., Pecci, A., and Alvarez, L. D. (2020). Structural Insights into the Ligand Binding Domain of the Glucocorticoid Receptor: A Molecular Dynamics Study. *J. Chem. Inf. Model.* 60, 794–804. doi:10.1021/acs.jcim.9b00776

Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: the RESP Model. *J. Phys. Chem.* 97, 10269–10280. doi:10.1021/j100142a004

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

Buttgereit, F., Bijlsma, J. W. J., and Strehl, C. (2018). Will We Ever Have Better Glucocorticoids? *Clin. Immunol.* 186, 64–66. doi:10.1016/j.clim.2017.07.023

Buttgereit, F. (2020). Can We Shift the Benefit–Risk Ratio of Glucocorticoids? *Lancet Rheumatol.* 2, E5–E6. doi:10.1016/s2665-9913(19)30138-9

Cain, D. W., and Cidlowski, J. A. (2015). Specificity and Sensitivity of Glucocorticoid Signaling in Health and Disease. *Best Pract. Res. Clin. Endocrinol. Metabolism* 29, 545–556. doi:10.1016/j.beem.2015.04.007

Carson-Jurica, M. A., Schrader, W. T., and O'Malley, B. W. (1990). Steroid Receptor Family: Structure and Functions. *Endocr. Rev.* 11, 201–220. doi:10.1210/edrv-11-2-201

Cato, A. C. B., and Wade, E. (1996). Molecular Mechanisms of Anti-inflammatory Action of Glucocorticoids. *BioEssays* 18, 371–378. doi:10.1002/bies.950180507

Chong, L. T., Pitera, J. W., Swope, W. C., and Pande, V. S. (2009). Comparison of Computational Approaches for Predicting the Effects of Missense Mutations on P53 Function. *J. Mol. Graph. Model.* 27, 978–982. doi:10.1016/j.jmgm.2008.12.006

Czock, D., Keller, F., Rasche, F. M., and Häussler, U. (2005). Pharmacokinetics and Pharmacodynamics of Systemically Administered Glucocorticoids. *Clin. Pharmacokinet.* 44, 61–98. doi:10.2165/00003088-200544010-00003

Darimont, B. D., Wagner, R. L., Apriletti, J. W., Stallcup, M. R., Kushner, P. J., Baxter, J. D., et al. (1998). Structure and Specificity of Nuclear Receptor-Coactivator Interactions. *Genes Dev.* 12, 3343–3356. doi:10.1101/gad.12.21.3343

Edman, K., Hosseini, A., Bjursell, M. K., Aagaard, A., Wissler, L., Gunnarsson, A., et al. (2015). Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints. *Structure* 23, 2280–2290. doi:10.1016/j.str.2015.09.012

Fan, J., Liu, Y., Kong, R., Ni, D., Yu, Z., Lu, S., et al. (2021). Harnessing Reversed Allosteric Communication: A Novel Strategy for Allosteric Drug Discovery. *J. Med. Chem.* 64, 17728–17743. doi:10.1021/acs.jmedchem.1c01695

Feng, L., Lu, S., Zheng, Z., Chen, Y., Zhao, Y., Song, K., et al. (2021). Identification of an Allosteric Hotspot for Additive Activation of PPARγ in Antidiabetic Effects. *Sci. Bull.* 66, 1559–1570. doi:10.1016/j.scib.2021.01.023

Fowden, A. L., Li, J., and Forhead, A. J. (1998). Glucocorticoids and the Preparation for Life after Birth: Are There Long-Term Consequences of the Life Insurance? *Proc. Nutr. Soc.* 57, 113–122. doi:10.1079/pns19980017

Gebhardt, J. C. M., Suter, D. M., Roy, R., Zhao, Z. W., Chapman, A. R., Basu, S., et al. (2013). Single-molecule Imaging of Transcription Factor Binding to DNA in Live Mammalian Cells. *Nat. Methods* 10, 421–426. doi:10.1038/nmeth.2411

Goto, K., Zhao, Y., Saito, M., Tomura, A., Morinaga, H., Nomura, M., et al. (2003). Activation Function-1 Domain of Androgen Receptor Contributes to the Interaction between Two Distinct Subnuclear Compartments. *J. Steroid Biochem. Mol. Biol.* 85, 201–208. doi:10.1016/s0960-0760(03)00196-1

Gronemeyer, H., and Moras, D. (1995). How to Finger DNA. *Nature* 375, 190–191. doi:10.1038/375190a0

Heck, S., Kullmann, M., Gast, A., Ponta, H., Rahmsdorf, H. J., Herrlich, P., et al. (1994). A Distinct Modulating Domain in Glucocorticoid Receptor Monomers in the Repression of Activity of the Transcription Factor AP-1. *EMBO J.* 13, 4087–4095. doi:10.1002/j.1460-2075.1994.tb06726.x

Heery, D. M., Kalkhoven, E., Hoare, S., and Parker, M. G. (1997). A Signature Motif in Transcriptional Co-activators Mediates Binding to Nuclear Receptors. *Nature* 387, 733–736. doi:10.1038/42750

Hu, X., and Lazar, M. A. (1999). The CoRNR Motif Controls the Recruitment of Corepressors by Nuclear Hormone Receptors. *Nature* 402, 93–96. doi:10.1038/47069

Hu, X., Pang, J., Zhang, J., Shen, C., Chai, X., Wang, E., et al. (2022). Discovery of Novel GR Ligands toward Druggable GR Antagonist Conformations Identified by MD Simulations and Markov State Model Analysis. *Adv. Sci.* 9, 2102435. doi:10.1002/advs.202102435

Hünenberger, P. H., Mark, A. E., and van Gunsteren, W. F. (1995). Fluctuation and Cross-Correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J. Mol. Biol.* 252, 492–503.

Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000). Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* 21, 132–146. doi:10.1002/(sici)1096-987x(20000130)21:2<132::aid-jcc5>3.0.co;2-p

Jang, H., Zhang, M., and Nussinov, R. (2020). The Quaternary Assembly of KRas4B with Raf-1 at the Membrane. *Comput. Struct. Biotechnol. J.* 18, 737–748. doi:10.1016/j.csbj.2020.03.018

Jenkins, B. D., Pullen, C. B., and Darimont, B. D. (2001). Novel Glucocorticoid Receptor Coactivator Effector Mechanisms. *Trends Endocrinol. Metab.* 12, 122–126. doi:10.1016/s1043-2760(00)00357-x

Jiang, C.-L., Liu, L., and Tasker, J. G. (2014). Why Do We Need Nongenomic Glucocorticoid Mechanisms? *Front. Neuroendocrinol.* 35, 72–75. doi:10.1016/j.yfrne.2013.09.005

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

Kadmiel, M., and Cidlowski, J. A. (2013). Glucocorticoid Receptor Signaling in Health and Disease. *Trends Pharmacol. Sci.* 34 (9), 518–530. doi:10.1016/j.tips.2013.07.003

Karplus, M., and Kushick, J. N. (1981). Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* 14, 325–332. doi:10.1021/ma50003a019

Karra, A. G., Sioutopoulou, A., Gorgogietas, V., Samiotaki, M., Panayotou, G., and Psarra, A.-M. G. (2022). Proteomic Analysis of the Mitochondrial Glucocorticoid Receptor Interacting Proteins Reveals Pyruvate Dehydrogenase and Mitochondrial 60 kDa Heat Shock Protein as Potent Binding Partners. *J. Proteomics* 257, 104509. doi:10.1016/j.jprot.2022.104509

Köhler, C., Carlström, G., Gunnarsson, A., Weininger, U., Tångefjord, S., Ullah, V., et al. (2020). Dynamic Allosteric Communication Pathway Directing Differential Activation of the Glucocorticoid Receptor. *Sci. Adv.* 6, eabb5277. doi:10.1126/sciadv.abb5277

Kumar, R., and Thompson, E. B. (1999). The Structure of the Nuclear Hormone Receptors. *Steroids* 64, 310–319. doi:10.1016/s0039-128x(99)00014-8

Lee, K., Thwin, A. C., Nadel, C. M., Tse, E., Gates, S. N., Gestwicki, J. E., et al. (2021). The Structure of an Hsp90-Immunophilin Complex Reveals Cochaperone Recognition of the Client Maturation State. *Mol. Cell* 81, 3496–3508. doi:10.1016/j.molcel.2021.07.023

Li, J., White, J. T., Saavedra, H., Wrabl, J. O., Motlagh, H. N., Liu, K., et al. (2017). Genetically Tunable Frustration Controls Allostery in an Intrinsically Disordered Transcription Factor. *Elife* 6, e30688. doi:10.7554/eLife.30688

Li, X., Dai, J., Ni, D., He, X., Zhang, H., Zhang, J., et al. (2020). Insight into the Mechanism of Allosteric Activation of PI3Kα by Oncoprotein K-Ras4B. *Int. J. Biol. Macromol.* 144, 643–655. doi:10.1016/j.ijbiomac.2019.12.020

Li, X., Wang, C., Peng, T., Chai, Z., Ni, D., Liu, Y., et al. (2021). Atomic-Scale Insights into Allosteric Mechanism Inhibition and Evolutional Rescue Mechanism of Streptococcus Thermophilus Cas9 by the Anti-CRISPR Protein AcrIIA6. *Comput. Struct. Biotechnol. J.* 19, 6108–6124. doi:10.1016/j.csbj.2021.11.010

Liang, S., Wang, Q., Qi, X., Liu, Y., Li, G., Lu, S., et al. (2021). Deciphering the Mechanism of Gilteritinib Overcoming Lorlatinib Resistance to the Double Mutant I1171N/F1174I in Anaplastic Lymphoma Kinase. *Front. Cell Dev. Biol.* 9, 808864. doi:10.3389/fcell.2021.808864

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

Liu, B., Zhang, T.-N., Knight, J. K., and Goodwin, J. E. (2019a). The Glucocorticoid Receptor in Cardiovascular Health and Disease. *Cells* 8, 1227. doi:10.3390/cells8101227

Liu, X., Wang, Y., and Ortlund, E. A. (2019b). First High-Resolution Crystal Structures of the Glucocorticoid Receptor Ligand-Binding Domain–Peroxisome Proliferator-Activated γ Coactivator 1-α Complex with Endogenous and Synthetic Glucocorticoids. *Mol. Pharmacol.* 96, 408–417. doi:10.1124/mol.119.116806

Lu, S., and Zhang, J. (2019d). Small Molecule Allosteric Modulators of G-Protein-Coupled Receptors: Drug-Target Interactions. *J. Med. Chem.* 62, 24–45. doi:10.1021/acs.jmedchem.7b01844

Lu, S., Jang, H., Muratcioglu, S., Gursoy, A., Keskin, O., Nussinov, R., et al. (2016). Ras Conformational Ensembles, Allostery, and Signaling. *Chem. Rev.* 116, 6607–6665. doi:10.1021/acs.chemrev.5b00542

Lu, S., Ni, D., Wang, C., He, X., Lin, H., Wang, Z., et al. (2019a). Deactivation Pathway of Ras GTPase Underlies Conformational Substates as Targets for Drug Design. *ACS Catal.* 9, 7188–7196. doi:10.1021/acscatal.9b02556

Lu, S., He, X., Ni, D., and Zhang, J. (2019b). Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *J. Med. Chem.* 62, 6405–6421. doi:10.1021/acs.jmedchem.8b01749

Lu, S., Shen, Q., and Zhang, J. (2019c). Allosteric Methods and Their Applications: Facilitating the Discovery of Allosteric Drugs and the Investigation of Allosteric Mechanisms. *Acc. Chem. Res.* 52, 492–500. doi:10.1021/acs.accounts.8b00570

Lu, S., Chen, Y., Wei, J., Zhao, M., Ni, D., He, X., et al. (2021a). Mechanism of Allosteric Activation of SIRT6 Revealed by the Action of Rationally Designed Activators. *Acta Pharm. Sin. B* 11, 1355–1361. doi:10.1016/j.apsb.2020.09.010

Lu, S., He, X., Yang, Z., Chai, Z., Zhou, S., Wang, J., et al. (2021b). Activation Pathway of a G Protein-Coupled Receptor Uncovers Conformational Intermediates as Targets for Allosteric Drug Design. *Nat. Commun.* 12, 4721. doi:10.1038/s41467-021-25020-9

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. Theory Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Meijer, O. C., Koorneef, L. L., and Kroon, J. (2018). Glucocorticoid Receptor Modulators. *Ann. Endocrinol.* 79, 107–111. doi:10.1016/j.ando.2018.03.004

Musson, D. G., Bidgood, A. M., and Olejnik, O. (1991). Assay Methodology for Prednisolone, Prednisolone Acetate and Prednisolone Sodium Phosphate in Rabbit Aqueous Humor and Ocular Physiological Solutions. *J. Chromatogr. B Biomed. Sci. Appl.* 565, 89–102. doi:10.1016/0378-4347(91)80373-k

Nagy, L., and Schwabe, J. W. R. (2004). Mechanism of the Nuclear Receptor Molecular Switch. *Trends Biochem. Sci.* 29, 317–324. doi:10.1016/j.tibs.2004.04.006

Necela, B. M., and Cidlowski, J. A. (2003). Crystallization of the Human Glucocorticoid Receptor Ligand Binding Domain: a Step towards Selective Glucocorticoids. *Trends Pharmacol. Sci.* 24, 58–61. doi:10.1016/s0165-6147(02)00046-9

Ni, D., Wei, J., He, X., Rehman, A. U., Li, X., Qiu, Y., et al. (2020). Discovery of Cryptic Allosteric Sites Using Reversed Allosteric Communication by a Combined Computational and Experimental Strategy. *Chem. Sci.* 12, 464–476. doi:10.1039/d0sc05131d

Ni, D., Chai, Z., Wang, Y., Li, M., Yu, Z., Liu, Y., et al. (2021). Along the Allostery Stream: Recent Advances in Computational Methods for Allosteric Drug Discovery. *WIREs Comput. Mol. Sci.* doi:10.1002/wcms.1585

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

O'Malley, B. W., and Tsai, M.-J. (1992). Molecular Pathways of Steroid Receptor Action. *Biol. Reprod.* 46, 163–167.

Plevin, M. J., Mills, M. M., and Ikura, M. (2005). The LxxLL Motif: a Multifunctional Binding Sequence in Transcriptional Regulation. *Trends Biochem. Sci.* 30, 66–69. doi:10.1016/j.tibs.2004.12.001

Pratt, W. B., and Toft, D. O. (1997). Steroid Receptor Interactions with Heat Shock Protein and Immunophilin Chaperones. *Endocr. Rev.* 18, 306–360. doi:10.1210/edrv.18.3.0303

Thiessen, J. J. (1976). Prednisolone, Bioavailability Monograph. *J. Am. Pharm. Assoc.* 16, 143–146.

Qiu, Y., Yin, X., Li, X., Wang, Y., Fu, Q., Huang, R., et al. (2021). Untangling Dual-Targeting Therapeutic Mechanism of Epidermal Growth Factor Receptor (EGFR) Based on Reversed Allosteric Communication. *Pharmaceutics* 13, 747. doi:10.3390/pharmaceutics13050747

Reichardt, H. M., Tuckermann, J. P., Göttlicher, M., Vujic, M., Weih, F., Angel, P., et al. (2001). Repression of Inflammatory Responses in the Absence of DNA Binding by the Glucocorticoid Receptor. *EMBO J.* 20, 7168–7173. doi:10.1093/emboj/20.24.7168

Schäcke, H., Döcke, W.-D., and Asadullah, K. (2002). Mechanisms Involved in the Side Effects of Glucocorticoids. *Pharmacol. Ther.* 96, 23–43.

Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical Networks in tRNA:protein Complexes. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6620–6625. doi:10.1073/pnas.0810961106

Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E. (2007). Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* 3, 2312–2334. doi:10.1021/ct700119m

Shen, Q., Wang, G., Li, S., Liu, X., Lu, S., Chen, Z., et al. (2016). ASD v3.0: Unraveling Allosteric Regulation with Structural Mechanisms and Biological Networks. *Nucleic Acids Res.* 44, D527–D535. doi:10.1093/nar/gkv902

Sindhikara, D. J., Kim, S., Voter, A. F., and Roitberg, A. E. (2009). Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules. *J. Chem. Theory Comput.* 5, 1624–1631. doi:10.1021/ct800573m

Styczynski, J., Kurylak, A., and Wysocki, M. (2005). Cytotoxicity of Cortivazol in Childhood Acute Lymphoblastic Leukemia. *Anticancer Res.* 25, 2253–2258.

Suino-Powell, K., Xu, Y., Zhang, C., Tao, Y.-g., Tolbert, W. D., Simons, S. S., et al. (2008). Doubling the Size of the Glucocorticoid Receptor Ligand Binding Pocket by Deacylcortivazol. *Mol. Cell. Biol.* 28, 1915–1923. doi:10.1128/mcb.01541-07

Swegat, W., Schlitter, J., Krüger, P., and Wollmer, A. (2003). MD Simulation of Protein-Ligand Interaction: Formation and Dissociation of an Insulin-Phenol Complex. *Biophys. J.* 84, 1493–1506. doi:10.1016/s0006-3495(03)74962-5

Torchia, J., Rose, D. W., Inostroza, J., Kamei, Y., Westin, S., Glass, C. K., et al. (1997). The Transcriptional Co-activator P/CIP Binds CBP and Mediates Nuclear-Receptor Function. *Nature* 387, 677–684. doi:10.1038/42652

Uberuaga, B. P., Anghel, M., and Voter, A. F. (2004). Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* 120, 6363–6374. doi:10.1063/1.1667473

Van Staa, T. P., Leufkens, H. G. M., Abenhaim, L., Begaud, B., Zhang, B., and Cooper, C. (2000). Use of Oral Corticosteroids in the United Kingdom. *QJM - Mon. J. Assoc. Physicians* 93, 105–111. doi:10.1093/qjmed/93.2.105

Vandevyver, S., Dejager, L., and Libert, C. (2014). Comprehensive Overview of the Structure and Regulation of the Glucocorticoid Receptor. *Endocr. Rev.* 35, 671–693. doi:10.1210/er.2014-1010

Veleiro, A. S., Alvarez, L. D., Eduardo, S. L., and Burton, G. (2010). Structure of the Glucocorticoid Receptor, a Flexible Protein that Can Adapt to Different Ligands. *ChemMedChem* 5, 649–659. doi:10.1002/cmdc.201000014

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, Y., Ji, D., Lei, C., Chen, Y., Qiu, Y., Li, X., et al. (2021). Mechanistic Insights into the Effect of Phosphorylation on Ras Conformational Dynamics and its Interactions with Cell Signaling Proteins. *Comput. Struct. Biotechnol. J.* 19, 1184–1199. doi:10.1016/j.csbj.2021.01.044

Wang, Y., Li, M., Liang, W., Shi, X., Fan, J., Kong, R., et al. (2022). Delineating the Activation Mechanism and Conformational Landscape of a Class B G Protein-Coupled Receptor Glucagon Receptor. *Comput. Struct. Biotechnol. J.* 20, 628–639. doi:10.1016/j.csbj.2022.01.015

Weikum, E. R., Okafor, C. D., D'Agostino, E. H., Colucci, J. K., and Ortlund, E. A. (2017). Structural Analysis of the Glucocorticoid Receptor Ligand-Binding Domain in Complex with Triamcinolone Acetonide and a Fragment of the Atypical Coregulator, Small Heterodimer Partner. *Mol. Pharmacol.* 92, 12–21. doi:10.1124/mol.117.108506

Zhang, M., Jang, H., and Nussinov, R. (2019). The Mechanism of PI3Kα Activation at the Atomic Level. *Chem. Sci.* 10, 3671–3680. doi:10.1039/c8sc04498h

Zhang, Q., Chen, Y., Ni, D., Huang, Z., Wei, J., Feng, L., et al. (2022). Targeting a Cryptic Allosteric Site of SIRT6 with Small-Molecule Inhibitors that Inhibit the Migration of Pancreatic Cancer Cells. *Acta Pharm. Sin. B* 12, 876–889. doi:10.1016/j.apsb.2021.06.015

Zhuang, H., Fan, X., Ji, D., Wang, Y., Fan, J., Li, M., et al. (2022). Elucidation of the Conformational Dynamics and Assembly of Argonaute-RNA Complexes by Distinct yet Coordinated Actions of the Supplementary microRNA. *Comput. Struct. Biotechnol. J.* 20, 1352–1365. doi:10.1016/j.csbj.2022.03.001

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 08 July 2022
doi: 10.3389/fmolb.2022.879000

A Curvilinear-Path Umbrella Sampling Approach to Characterizing the Interactions Between Rapamycin and Three FKBP12 Variants

Dhananjay C. Joshi[1], *Charlie Gosse*[2], *Shu-Yu Huang*[1] *and Jung-Hsin Lin*[1,3,4,5,6]*

[1]Research Center for Applied Sciences, Academia Sinica, Taipei, Taiwan, [2]Institut de Biologie de l'Ecole Normale Supérieure, ENS, CNRS, INSERM, PSL Research University, Paris, France, [3]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, [4]Biomedical Translation Research Center, National Biotechnology Research Park, Academia Sinica, Taipei, Taiwan, [5]School of Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan, [6]College of Engineering Sciences, Chang Gung University, Taoyuan, Taiwan

Rapamycin is an immunosuppressant macrolide that exhibits anti-proliferative properties through inhibiting the mTOR kinase. In fact, the drug first associates with the FKBP12 enzyme before interacting with the FRB domain of its target. Despite the availability of structural and thermodynamic information on the interaction of FKBP12 with rapamycin, the energetic and mechanistic understanding of this process is still incomplete. We recently reported a multiple-walker umbrella sampling simulation approach to characterizing the protein–protein interaction energetics along curvilinear paths. In the present paper, we extend our investigations to a protein-small molecule duo, the FKBP12•rapamycin complex. We estimate the binding free energies of rapamycin with wild-type FKBP12 and two mutants in which a hydrogen bond has been removed, D37V and Y82F. Furthermore, the underlying mechanistic details are analyzed. The calculated standard free energies of binding agree well with the experimental data, and the roles of the hydrogen bonds are shown to be quite different for each of these two mutated residues. On one hand, removing the carboxylate group of D37 strongly destabilizes the association; on the other hand, the hydroxyl group of Y82 is nearly unnecessary for the stability of the complex because some nonconventional, cryptic, indirect interaction mechanisms seem to be at work.

Keywords: rapamycin, FKBP12, umbrella sampling simulations, molecular dynamics, free energy calculation, hydrogen bond

OPEN ACCESS

Edited by:
J. Andrew McCammon, University of California, San Diego, United States

Reviewed by:
Jing Huang, Westlake University, China
Serdal Kirmizialtin, New York University Abu Dhabi, United Arab Emirates

***Correspondence:**
Jung-Hsin Lin
jhlin@gate.sinica.edu.tw

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *18 February 2022*
Accepted: *12 May 2022*
Published: *08 July 2022*

Citation:
Joshi DC, Gosse C, Huang S-Y and Lin J-H (2022) A Curvilinear-Path Umbrella Sampling Approach to Characterizing the Interactions Between Rapamycin and Three FKBP12 Variants. Front. Mol. Biosci. 9:879000. doi: 10.3389/fmolb.2022.879000

INTRODUCTION

Protein–ligand interactions are central in modern drug-discovery, and their characterization by various approaches is crucial for better drug development. In this regard, computational investigation is one of the ways to acquire a deeper understanding of the interactions of interest. In particular, molecular dynamics (MD) simulations provide the physical connection between the structure and the function of biomolecules (Karplus and McCammon, 2002), especially at the atomic level; therefore, MD simulation-based techniques can often cast insights into such interactions, especially in the early stage drug-discovery (Mobley and Gilson, 2017). In addition to conformational dynamics of the interacting molecules, MD simulations are also employed to estimate

thermodynamic properties such as the relative binding free energy and the binding affinity. Major ongoing challenges in this field are related to the reliability, the accuracy, and the rapidity of the estimation method/approach.

A reliable estimation of the free energy difference between thermodynamically well-defined end states of interacting molecules is one of the major goals of computational biophysics. Umbrella sampling along a chosen reaction coordinate, followed by potential of mean force calculations, is one of the ways widely used to estimate binding affinities (Torrie and Valleau, 1977; Kästner, 2011). However, umbrella sampling along predefined vectorial reaction coordinates followed by potential of mean force (PMF) profile constructions has some serious concerns in performing reliable energetic calculations (Doudou et al., 2009). Recently, we discussed them for the protein–protein systems and proposed a naïve multiple-walker approach, in which independent umbrella sampling simulations were conducted without predefined vectorial reaction coordinates. We observed similar large variations in the values of converged PMF profiles, resulting from different curvilinear paths. The variations were attributed as due to the different excessive dissipations in different paths taken, and therefore the lower-bound PMF was chosen, and by introducing a correction term derived from statistical mechanics, the standard free energy of binding was estimated for the protein–protein complex system (Joshi and Lin, 2019). The estimations were in good agreement with the experimental values for the barnase–barstar complex. Furthermore, the revealed mechanistic details from our simulations, e.g., the physical pathways/trajectories of dissociation/association, were quite consistent with two major physical pathways that were determined from several milliseconds-long adaptive molecular dynamics simulations reported by Plattner et al. (2017). Thus, the proposed approach is quite useful in maintaining a suitable balance between estimation of binding energetics and revealing underlying mechanistic details of the dissociation reaction within much less computational cost than the brute force approaches. Since the sampling is enhanced along a spontaneously evolved (i.e., non-predefined) curvilinear physical trajectories, the overall approach is referred to as curvilinear-path umbrella sampling (CPUS) approach. In the present paper, we extend our previous work to the interactions of a drug with its protein partner.

This implementation of CPUS will be validated on FKBP12, a model used in computational ligand binding studies for more than 20 years (Pearlman and Connelly, 1995; Lin et al., 2002; Lin et al., 2003; Sun et al., 2003; Swanson et al., 2004; Fujitani et al., 2005; Lee and Olson, 2006; Olivieri and Gardebien, 2011; Dickson and Lotz, 2016; Nerattini et al., 2016). This enzyme participates in protein folding by catalyzing the isomerization of amide bonds adjacent to proline residues (Hamilton and Steiner, 1998; Galat, 2013). Discovered because it could bind the immunosuppressant macrolide FK506, it was further shown that it could also complex with Rapamycin, a parent drug sharing the same pharmacophore and similar therapeutic indications (Van Duyne et al., 1993; Hamilton and Steiner, 1998; Galat, 2013). It later turned out that the inhibition of the FKBP12 enzymatic activity was irrelevant to account for the pharmaceutical properties of the two macrolides: In fact, this protein only potentializes the binding of each drug to its effective cellular target through the formation of a specific ternary complex, with calcineurin for FK506 and with mTOR for rapamycin (Choi et al., 1996; Banaszynski et al., 2005; Galat, 2013). Apart from its biological significance, the choice of FKBP12 as a model in the 1990s can be explained by its small size (107 residues) and its compact structure, both key factors at a time where computational power was limited. Moreover, several X-ray crystal structures were quickly released, in free form as well as in complex with FK506 and rapamycin (Van Duyne et al., 1991; Van Duyne et al., 1993; Wilson et al., 1995). Additionally, with time, a huge amount of biophysical data have been accumulated from thermodynamic and kinetic measurements on the complexation reaction (Bierer et al., 1990; Holt et al., 1993; Bossard et al., 1994; Connelly et al., 1994; Luengo et al., 1995; DeCenzo et al., 1996a; Schuler et al., 1997; Wagner et al., 1998; Graziani et al., 1999; Dickman et al., 2000; Banaszynski et al., 2005; Wear et al., 2007; Wear and Walkinshaw, 2007; Shor et al., 2008; Kozany et al., 2009; Wu et al., 2011; Tamura et al., 2013; Singh et al., 2015; Lu and Wang, 2017; Kostrz et al., 2019; Wang et al., 2019) to NMR investigations on the protein dynamics (Sapienza et al., 2011; Yang et al., 2015; Solomentsev et al., 2018).

As far as *in silico* studies are concerned, FKBP12 is extensively used as a model system to study protein–ligand interaction energetics. For instance, the relaxed complex scheme (Lin et al., 2002) was employed to estimate interaction energetics of FKBP12 with several different ligands (Lin et al., 2003). Another study using FKBP12 was performed to establish the groundwork for the end-point free energy methods, in which the theoretical framework was proposed to calculate the association free energy (Swanson et al., 2004). A study on estimation of absolute binding free energy calculations of FKBP12 and eight ligands was carried out, where the Bennett acceptance ration (BAR) method was employed in the direct calculations (Fujitani et al., 2005). The FKBP12•ligand system was used to estimate the binding free energies with two ligands, 4-hydroxy-2-butanone and FK506, respectively. Although the necessity of sampling along curvilinear paths was mentioned, the theoretical framework for the free energy estimation was developed only for linear/vectorial paths, and with some quadratic approximations for the variance along the principal axis (Swanson et al., 2004; Lee and Olson, 2006). The CPUS approach can be considered as an alternative approach without such approximations.

From a structural point of view, the macrolide binding site (Van Duyne et al., 1991; Van Duyne et al., 1993) significantly overlaps with the FKBP12 catalytic cleft, in agreement with the observed catalytic inhibition (DeCenzo et al., 1996a; Ikura and Ito, 2007). Rapamycin, the only ligand we will study here, binds in the cavity located between the short α-helix and the five-stranded β-sheet that is wrapped around it. More specifically, the drug pipecolinyl ring is deeply buried inside the protein (**Figure 1A**) and is involved in hydrophobic interactions with the aromatic side-chains of residues Y26, F46, W59, and F99 (**Figure 1B**; **Supplementary Figure S1**)—see (Van Duyne et al., 1993; Sun et al., 2003) for lists of the

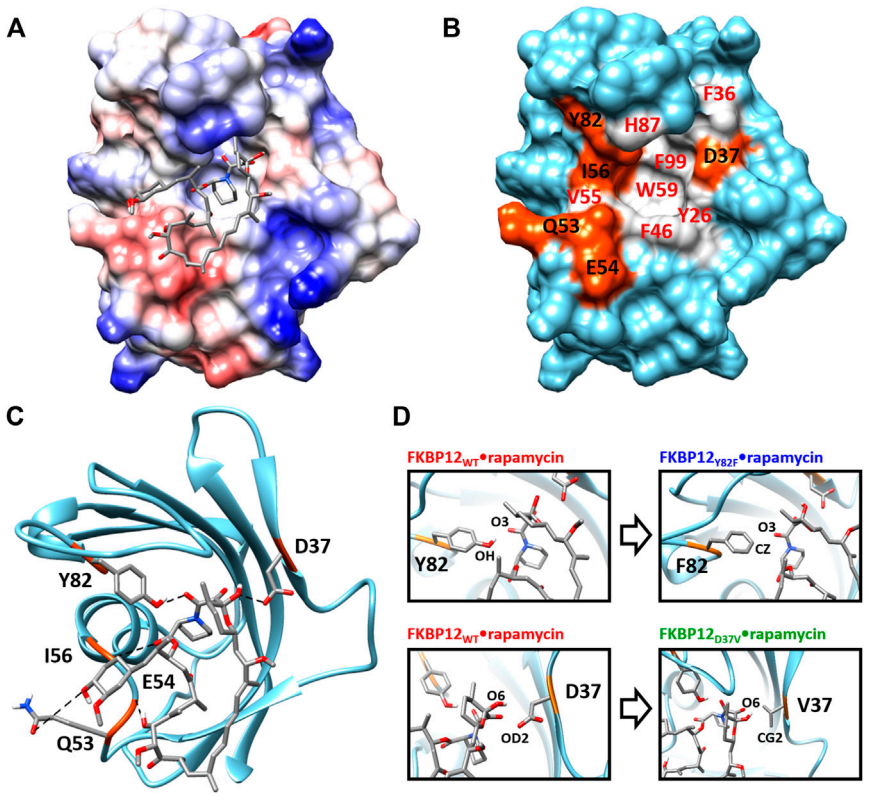


FIGURE 1 | Structure of the FKBP12•rapamycin complex as determined in PDB 1FKB. **(A)** Coulombic surface representation of the protein with the ligand as sticks. **(B)** Surface representation of the protein alone with the hydrogen-bond forming residues in orange and the hydrophobic residues in gray—coloring according to the LIGPLOT diagram provided as **Supplementary Figure S1** (Wallace et al., 1995). **(C)** Ribbon representation of the protein with the ligand as sticks. The five hydrogen bonds formed between FKBP12 and rapamycin are shown as dashed lines. **(D)** Close-up view on the Y82 and D37 residues and resulting starting conformation obtained after either the Y82F or the D37V substitution.

atom-pair contacts. In addition to auxiliary hydrophobic interactions with some of the amino acids surrounding the cavity (i.e., F36, V55, I56, H87, Y82—**Figure 1B**), rapamycin is retained at the FKBP12 surface through five hydrogen bonds. Of the hydrogen bond-implied residues (**Figures 1B,C**), D37, I56, and Y82 are also positioned on the rim, whereas Q53 and E54 are more distant. Since the first three amino acids are the most conserved (Van Duyne et al., 1993; Hamilton and Steiner, 1998; Galat, 2013), they seemed the most attractive for performing an *in silico* mutagenesis program aiming at demonstrating the strengths of the CPUS approach. More precisely, our goal was to test if we could predict the influence of H-bond removal on the stability of the FKBP12•rapamycin complex. We finally selected to target D37 and Y82 because they represent significantly contrasting examples of the contribution of hydrogen bonding to ligand binding. Indeed, experimental measurements have demonstrated that the D37V substitution strongly affects the protein–drug interaction (DeCenzo et al., 1996a; Singh et al., 2015; Kostrz et al., 2019), whereas the Y82F substitution does not (Connelly et al., 1994; Pearlman and Connelly, 1995; Kostrz et al., 2019) (**Figure 1D**). Incidentally, beyond the determination of the binding free energies for the wild-type protein and the two mutants, we expect that the MD trajectories produced by CPUS will enable us to understand the intriguing role of the H-bond formed between Y82 and rapamycin, a bond that has an atypical geometry (Van Duyne et al., 1993), which displays a clear signature in NMR spectroscopy (Yang et al., 2015), but whose elimination has nearly no impact on the affinity (Bossard et al., 1994; DeCenzo et al., 1996b).

MATERIALS AND METHODS

System Modeling

The $FKBP12_{WT}$•rapamycin binary complex was simulated using the PDB 1FKB crystal structure as a starting conformation (Van Duyne et al., 1991). For the complexes involving the Y82F and D37V mutants, we simply carried out *in silico* mutagenesis using the swapaa module of the Chimera molecular viewing platform (Pettersen et al., 2004) (**Figure 1D**). Thus, we here clearly assume that none of the two amino acid substitutions results in any significant

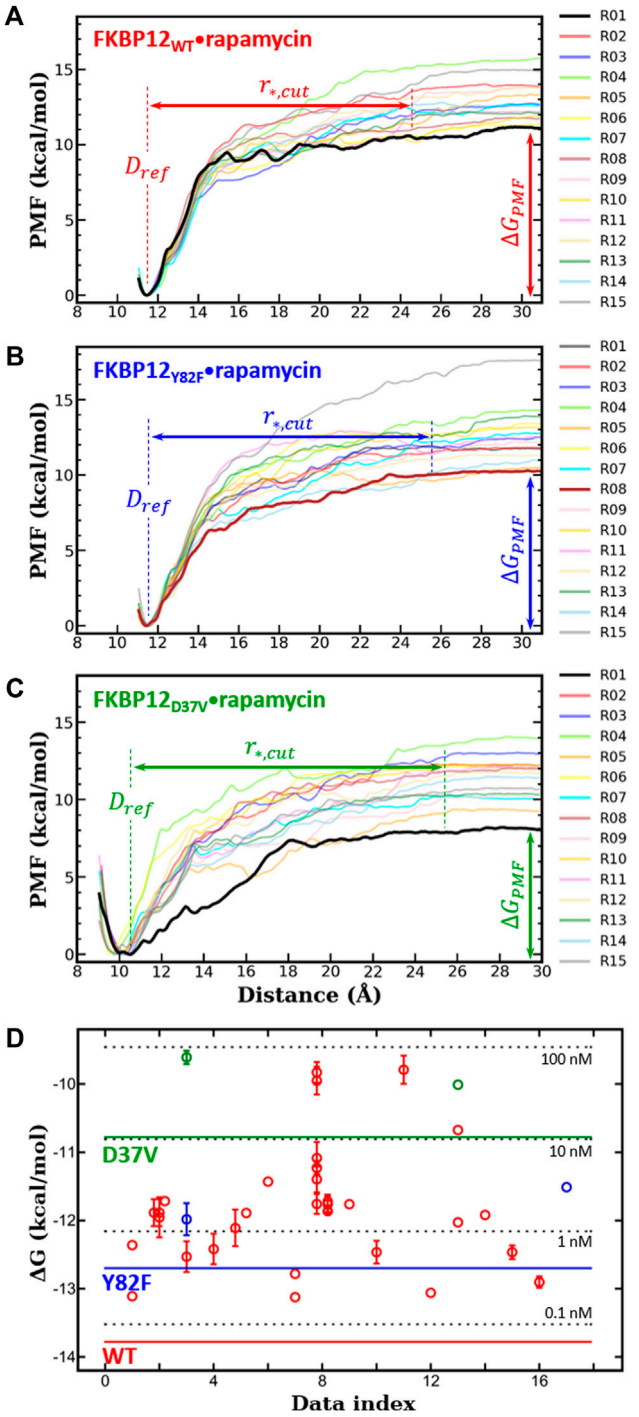


FIGURE 2 | *In silico* determination of the binding free energies for the three variants of FKBP12 + rapamycin ⇄ FKBP12•rapamycin reaction and comparison with experimental data. **(A)** PMF profiles issued from the 15 runs of CPUS MD simulations performed on the $FKBP12_{WT}$•rapamycin complex at T^{sim} = 21.85°C. The lower-bound PMF profile is highlighted in bold, and snapshots taken during this particular dissociation process are displayed in **Supplementary Figure S3**. Furthermore, we have indicated with double-headed arrows the associated cutoff separation distance, $r_{*,cut}$, and binding free energy, ΔG_{PMF} (see **Supplementary Table S1** for all 15 (*Continued*)

FIGURE 2 | numerical values). **(B)** Same plots and views for $FKBP12_{Y82F}$•rapamycin (see **Supplementary Figure S4** for the snapshots associated with the lower-bound PMF). **(C)** Same plots and views for $FKBP12_{D37V}$•rapamycin (see **Supplementary Figure S5** for the snapshots associated with the lower-bound PMF). **(D)** Collation of the corrected binding free energies obtained with the MD CPUS approach, ΔG^{0}_{bind} as horizontal lines, and of measurements retrieved from the literature and extrapolated at T^{sim}, ΔG^{corr}_{exp} as open circles (see **Table 1** and **Supplementary Table S2** for the corresponding numerical values). Red markers refer to the wild-type FKBP12, blue ones refer to the Y82F mutant, and green ones refer to the D37V mutant. Data have been sorted along the x-axis according to their publication year; moreover, to emphasize on possible biases due to individual practices, we have clustered together all measurements coming from a same laboratory. The four horizontal, black, and dashed lines indicate the ΔG values associated with the 0.1, 1, 10, and 100 nM dissociation equilibrium constants.

structural change with respect to the wild-type reference. Such an assertion has already been harnessed in the $FKBP12_{Y82F}$ case, crystallographic evidences being provided to support it (Pearlman and Connelly, 1995)—despite our efforts, we could not retrieve the original data from any of the usual repositories. As far as $FKBP12_{D37V}$ is concerned, we performed ^{1}H-^{15}N hetero-nuclear single quantum coherence NMR measurements and chemical shift perturbation analysis to back up our starting hypothesis (**Supplementary Figure S2**) (Williamson, 2018). Omitting the substituted amino acid, only three residues display changes larger than the mean values of the all changes plus two standard deviations: one is facing D37 in the adjacent β-strand and two are in the loop just downstream of the mutation site. As a consequence, we ruled out the possibility of any large-scale rearrangement.

The partial charges on the atoms of rapamycin were derived with the RESP scheme, calculated with Gaussian03 at the level of HF/6-31G basis set. The coordinates from the wild-type crystal structure and from the two mutant models were then processed using the antechamber and LEaP programs of the AMBER software suite (Case et al., 2005; Salomon-Ferrer et al., 2013). The well solvated system was built with 22926 TIP3P waters; 42 K^{+} and 43 Cl^{-} ions were added to neutralize the system and to mimic a 100 mM salt concentration. The box dimensions were 112.4 Å × 81.6 Å × 90.7 Å. The improved ff14SB force field along with the general amber force field (GAFF) (Wang et al., 2004) was used for all bonded and non-bonded parameters (Maier et al., 2015).

Molecular Dynamics Simulations

Production runs were conducted using the Graphical Processing Unit (GPU RTX 2080Ti) implementation of the AMBER pmemd program. All MD simulations were conducted under NPT conditions with SHAKE-enabled 2-fs time steps. The particle mesh Ewald (PME) algorithm of electrostatics was employed, and the non-bonded interaction cutoff was set to 10.0 Å. Prior to production run, all three systems were subjected to energy minimization, heated to 295 K, and then equilibrated for 100 ps. MD trajectories were analyzed using the cpptraj program of AmberTools-20 (Case et al., 2020) and several in-house shell and python scripts.

Multiple Walker Curvilinear-Path Umbrella Sampling Simulations

The successive steps necessary to implement multiple-walker umbrella sampling simulations along non-predefined curvilinear paths, as well as the corresponding theoretical framework and data treatment procedures, have been described in a previous article (Joshi and Lin, 2019). In brief, before each CPUS run, a short 10 ns-long unbiased MD simulation was conducted and the mean distance between the centers of geometry (CoG) of both FKBP12 and rapamycin was computed so as to provide a reference bound distance, D_{ref}. These CoG are defined by all Cα atoms for FKBP12 and all heavy atoms for rapamycin in the DISANG file, an input to the pmemd. The last frame of this unbiased MD simulation is used as the starting conformations for the CPUS simulation. In the umbrella sampling simulation, the reaction coordinate was chosen as the distance D between the CoGs. The reaction path was divided into N windows (i.e., the umbrella windows), and D was restrained using a biasing potential with a spring constant of k = 10.0 kcal/mol/Å^2 so as to keep the interacting molecules in the ξ^{th} distance window at distance D_ξ. The sampling in each umbrella window was enhanced sequentially so that the physical trajectory of dissociation could evolve spontaneously. To do so, the last conformation from each sampled window was chosen as the starting conformation for the next window for sampling enhancement. The distance sequence took the form $\{D_\xi\}_{\xi=1}^{\xi=N}$ with $D_\xi = D_{ref} + \delta(\xi - 1)$ and δ the distance per step, set to 0.1 Å. The PMF profile was constructed using well-established weighted histogram analysis methods (WHAMs) (Kumar et al., 1992; Grossfield, 2013), which removed the contribution of biased potential (**Figures 2A–C**). Finally, ΔG_{PMF} was chosen as the PMF value corresponding to the final distance of the constructed profile, D_{final}. Incidentally, the PMF profile was completed for distances smaller than D_{ref} by conducting the umbrella sampling in the backward direction for 10 or 20 windows (depending on the cases).

For each of the FKBP12 variants, 15 umbrella sampling simulations were carried out independently by assigning different starting velocities, i.e., setting ig = −1 in the pmemd input file. Sampling in each distance window was enhanced for 1.0 ns, and CPUS MD simulations were 3.15 μs-long in total.

PMF Correction to Standard Binding Free Energy

To determine the standard free energy of binding, ΔG^0_{bind}, each pair composed of a CPUS MD trajectory and of the associated PMF profile was further processed according to **Eq. 1** (Joshi and Lin, 2019),

$$\Delta G^0_{bind} = \Delta G_{PMF} - k_B T \ln\left[\left(\frac{4\pi r^2_{*,\,cut}}{V_0}\right)\int_{bound} dr\, e^{-\beta A_r}\right] \quad (1)$$

with $r = D - D_{ref}$ being the separation distance, V_0 being the standard state volume (1 663 Å^3), A_r being the PMF value at the separation distance r, and $\beta = 1/k_B T$. The evaluation of the second term on the right-hand side requires to know $r_{*,cut}$, i.e., the separation distance at which the interaction between the protein and its ligand vanished. To determine this cutoff, an interface interaction analysis was conducted using the linear interaction energy (LIE) module of the AMBER cpptraj program (Roe and Cheatham, 2013). The first cancellation of the van der Waals component provided $r_{*,cut}$. Then, using home-grown python and UNIX shell scripts, the bound integral included in the second term of **Eq. 1** was computed and ΔG^0_{bind} was finally determined (**Table 1**; **Supplementary Table S1**).

Curvilinear Path Tracing

The physical paths of dissociation were traced from the trajectories issued from the umbrella sampling simulations. First, for each run, the conformations were extracted at the 10-ps interval using the cpptraj module of AmberTools-20 and aligned with respect to FKBP12 only, the reference conformation being the one obtained after equilibration of the system. Doing so, the traversing of rapamycin from a bound to an unbound state could be plotted in the reference frame of the protein. Next, the aligned conformations were sorted with respect to the umbrella windows, which were sampled for 1 ns and thus contained a total of 100 conformations. For each window, the CoG of all 100 rapamycin molecules were computed and the geometric center of this ensemble (i.e., the window-center) was computed. The aligned conformation for which rapamycin was the nearest to this geometric center was chosen as the representative conformation for that umbrella window. All such umbrella window representative conformations were determined and their CoG, represented by colored spheres, used for tracing the physical path of dissociation. Finally, for each run, a black curve was drawn manually as a guide for the eye evidencing the separation process. Paths were only traced up to r = 10 Å, i.e., for the first 100 umbrella windows, so as to provide a clear vision of the initial steps of the dissociation reaction.

RESULTS

Binding Free Energy Computations

All 10 ns-long unbiased MD simulations prior to CPUS showed root-mean-square deviations (RMSDs) within 1.0 Å and root-mean-square fluctuations (RMSFs) in the 0.4–3.6 Å range, which indicates the absence of any large conformational transition (**Supplementary Figure S6**). Thus, the FKBP12 variants in complexes with rapamycin are stable while well equilibrated. If this result is not surprising for the FKBP12$_{WT}$•rapamycin complex, it also validates our simple modeling of FKBP12$_{Y82F}$•rapamycin and FKBP12$_{D37V}$•rapamycin.

Next, for each variant, 15 independent CPUS MD simulations were conducted and the corresponding PMF profiles were constructed. All curves flatten before reaching a CoG separation distance of 30.0 Å, which is a clear signature of complete dissociations (**Figures 2A–C**). The ΔG_{PMF} binding free energies could thus be identified as the last obtained PMF values. Additionally, in all cases, we could determine a cutoff

TABLE 1 | Comparison between the numerically and the experimentally determined binding free energies for the FKBP12 + rapamycin ⇄ FKBP12•rapamycin reaction. MD results, i.e., the PMF and the corrected standard free energies, ΔG_{PMF} and ΔG^0_{bind} respectively, were obtained thanks to the CPUS approach. For each of the three variants, only the values corresponding to the lower-bound PMF profiles were selected (Joshi and Lin, 2019). Measurements are issued from a publication reporting on an inhibition assay in which the rotamase activity of FKBP12 was spectroscopically monitored using succinyl-AlaLeuProPhe-*para*-nitroalinide as a substrate and α-chimotrypsin digestion as a development reaction (DeCenzo et al., 1996b). These original data were acquired at 15°C, and we extrapolated them at 21.85°C, the temperature at which simulations were performed, so as to obtain the corrected binding free energies and equilibrium constants, ΔG^{corr}_{exp} and K^{corr}_{exp} respectively (see **Supplementary Table S2** and **Supplementary Figure S8** for details).

FKBP12 variant	ΔG_{PMF} (kcal/mol)	ΔG^0_{bind} (kcal/mol)	ΔG^{corr}_{exp} (kcal/mol)	K^{corr}_{exp} (nM)
WT	–11.09	–13.98	–12.53 ± 0.23	0.54 ± 0.21
Y82F	–10.21	–13.22	–11.98 ± 0.23	1.37 ± 0.50
D37	–8.08	–10.78	–9.61 ± 0.09	77.70 ± 11.97

separation distance at which the van der Waals interactions had decreased to zero (**Supplementary Figure S7**): $r_{*,cut}$ roughly ranges between 13 and 19 Å (**Supplementary Table S1**). With these cutoffs in hand we could further correct the ΔG_{PMF} according to **Eq. 1**, which yielded the ΔG^0_{bind} standard free energies (**Supplementary Table S1**). In application of the variational principle for each of the three complexes, we retained the lower-bound PMF profile for comparison with experimental data (**Table 1**, **Figure 2D**, **Supplementary Figure S8**, and **Supplementary Table S2**).

We first compared our simulations results with the data contained in the sole article we found that experimentally evaluated, with the same technique, the binding of rapamycin to all of the three FKBP12 variants hereby considered (**Table 1**) (DeCenzo et al., 1996b). With our CPUS approach, the differences of binding free energies between the mutant proteins and the wild-type one are $\Delta\Delta G^0_{bind}$ = 0.76 and 3.20 kcal/mol for Y82F and D37V, respectively. With the enzymatic inhibition assay reported in the literature, we have $\Delta\Delta G^{corr}_{exp}$ = 0.55 ± 0.33 and 2.92 ± 0.25 kcal/mol. Thus, both datasets are consistent if we consider uncertainties: We could predict *in silico* that Y82F is a nearly neutral mutation whereas D37V significantly weakens the interaction with rapamycin. Interestingly, the thermodynamic integration technique applied to the FKBP12•FK506 complex could also account for the very small destabilizing effect of Y82F, in similar agreement with experimental measurements (Pearlman and Connelly, 1995).

In a second step, we aimed at testing if our CPUS strategy can also provide reliable estimates of binding free energies (and not only of their differences). Therefore, we collected dissociation equilibrium constants for the FKBP12•rapamycin complex from a little less than 20 original research articles, spanning more than a 30 years-long period and describing measurements ranging from enzymatic inhibition assays to single-molecule ones, from isothermal calorimetry to surface plasmon resonance (**Supplementary Table S2**). The experimental data were first temperature corrected using published binding enthalpies (**Supplementary Figure S8**) (Connelly et al., 1993; Connelly et al., 1994) and then gathered in **Figure 2D**. The results that could be anticipated from **Table 1** were confirmed: The *in silico* stability seems to be ≈1 kcal/mol higher than the one observed at the bench–although the reported values for $FKBP12_{WT}$ are spread over more than 3 kcal/mol, the less negative binding free energies are possibly artifactual. Since the same discrepancy between simulations and measurements is observed for all three variants, the chances are high that it is a systematic effect. It could be due to the force-field inaccuracy and/or to the inability to fully account for the experimental system.

Atom-Pair Distance Analysis

Even though both Y82 and D37 residues form a hydrogen bond with rapamycin, their role in ligand binding seems to differ: Removing the hydroxyl from the tyrosine has nearly no impact on the affinity (Connelly et al., 1994; Pearlman and Connelly, 1995; Kostrz et al., 2019) whereas exchanging the ethanoate side chain of the aspartate for an isopropyl one is significantly destabilizing (DeCenzo et al., 1996a; Singh et al., 2015; Kostrz et al., 2019). Since CPUS simulations could reproduce these energetic measurements, we next tried to see if the obtained trajectories could shed some light on the molecular mechanisms at work. More precisely, we performed a detailed analysis of both hydrogen bonded atom-pairs in the $FKBP12_{WT}$•rapamycin complex, during the starting 10 ns-long unbiased simulations and during the following 190 ns-long CPUS ones (**Figures 3**, **4**; **Supplementary Figures S9, S10**). Furthermore, the resulting patterns were compared with the ones obtained on the two mutant complexes, for atoms located at equivalent positions on the side chains.

During the first 50 ns simulating the wild-type complex (10 ns unbiased and 40 ns of CPUS), the length of the H-bond between Y82 and rapamycin remains unchanged, with ground level values around 3 Å for all 15 simulations. This fact is illustrated by the single position of the main peaks in the atom-pair distance histograms of **Figure 3A** and **Supplementary Figure S9A** (to illustrate our description different markers have been posted on the time-trace and on the histogram associated with the lower-bound PMF profile). Then, step-increases are observed and the atom-pair distances stabilize on plateaus located between 5 and 10 Å, depending on the simulation run. It thus yields secondary peaks with narrowly dispersed positions. As pulling progresses, these rapid transitions are sometimes followed by others, reflecting temporary interactions that last from ten to tens of ns and atom-pair distances that correspond to plateaus with higher average values. Such an evolution is evidenced by additional peaks in the histograms, peaks that are broader and that now lie in the 10 to 25 Å range. Incidentally, temporary back motions to a previous position are also possible, as exemplified by the R01 time-trace in **Figure 3A**. As one can see in **Figure 3B** and

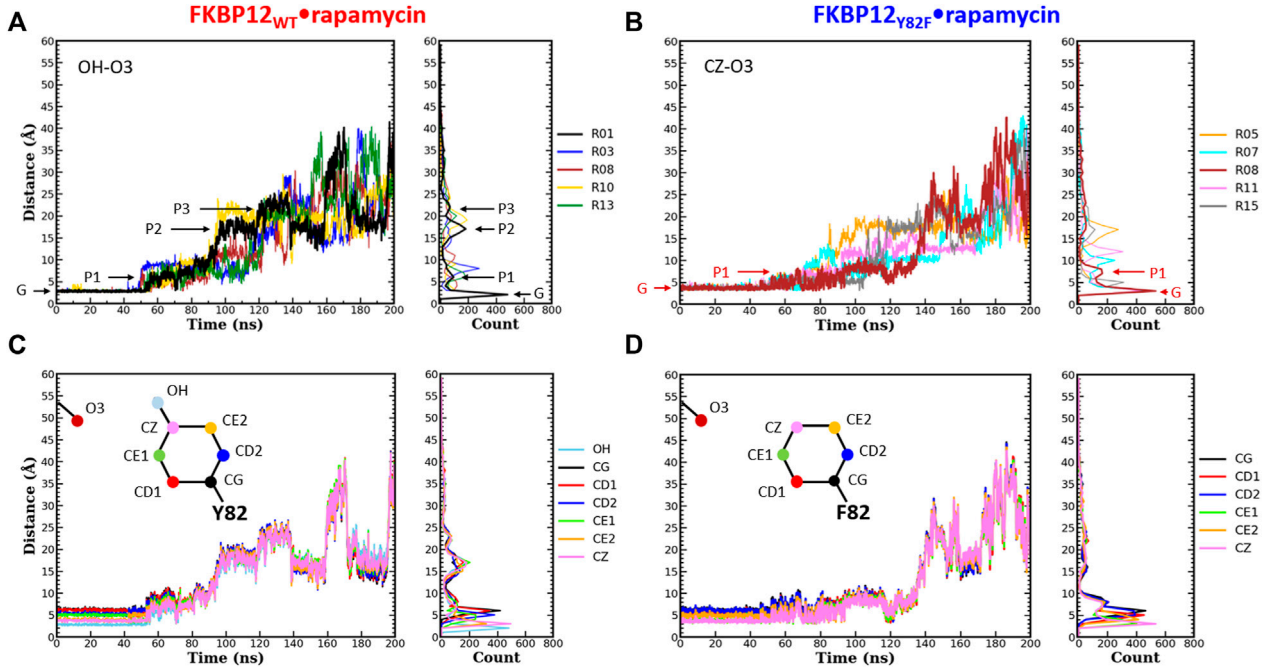


FIGURE 3 | Comparison of the separation distance patterns obtained during the dissociation of rapamycin from the wild-type FKBP12 and from the Y82F mutant. Pulling on the molecular partners CoG starts at 10 ns. **(A)** Evolution of the distance between the O3 carbonyl oxygen of rapamycin and the OH hydroxyl oxygen of residue Y82 in the $FKBP12_{WT}$•rapamycin complex. For the sake of clarity, data are only provided for the R01 lower-bound PMF profile and for four other ones. The G and Pi markers indicate, for the time-trace and the histogram associated with the lower-bound profile, the ground level and the different plateaus, respectively. The results for all 15 CPUS simulations are displayed in **Supplementary Figures S9A**. **(B)** Same plot for the distance between the O3 carbonyl oxygen of rapamycin and the CZ most distal carbon of residue F82 in the $FKBP12_{Y82F}$•rapamycin complex. The lower-bound PMF profile is now R08, and the data for all 15 CPUS simulations are available in **Supplementary Figures S9B**. **(C)** Evolution of the distance between the O3 carbonyl oxygen of rapamycin and the seven heavy atoms of residue Y82 phenol group in the $FKBP12_{WT}$•rapamycin complex. The data are only provided for the lower-bound PMF profile (see **Supplementary Figures S9A** for the whole set of curves). **(D)** Same plot for the distance between the O3 carbonyl oxygen of rapamycin and the six carbons of the residue F82 phenyl group in the $FKBP12_{Y82F}$•rapamycin complex. Once more data are only provided for the lower-bound PMF profile, see **Supplementary Figures S9B** for an exhaustive presentation of the results.

Supplementary Figure S9B, a similar behavior is observed for the carbon atom at the tip of the F82 phenyl ring in the $FKBP12_{Y82F}$•rapamycin complex; the only differences are that the first step-increases occur roughly 5–10 ns earlier and that the starting atom-pair distance is 1–2 Å shifted (because we are one more covalent bond apart from the oxygen of the ligand). Such resembling stepwise dissociations for both the wild-type and the mutant proteins indicate that the aromaticity of the side-chain may play a role here. To probe this assumption, we plotted the atom-pair distance for all six carbons of the ring (**Figures 3C,D** and **Supplementary Figure S9**) and indeed evidenced a concerted motion making us think that a persistent, long-range pseudo-bond could be relevant to anion-π interactions between the side-chain benzyl ring and the oxygen atoms of the rapamycin (Schottel et al., 2008; Chifotides and Dunbar, 2013; Giese et al., 2016; Lucas et al., 2016).

The same detailed approach was subsequently applied to study the role of D37 in rapamycin binding. In the wild-type complex, the length of the H-bond between this residue and the ligand remains nearly unchanged during the first 50–70 ns of simulation: It just fluctuates between 2.7 and 4.0 Å (**Figure 4A** and **Supplementary Figures S10A**). The origin of this fluctuation is clearly revealed when considering the other oxygen of the aspartate side-chain (**Figure 4B** and **Supplementary Figures S10A**): The two heteroatoms alternatively interact with rapamycin upon rotation of the carboxylate group, which results in two sharp peaks in the atom-pair distance histograms. Once the H-bond is broken, the oxygens of D37 and the one of the macrolide are torn in a continuous manner, showing very little step-like distance increments. Moreover, if a plateau is observed, it manifests itself in the histogram through a low and shallow peak, i.e., a somehow short and weak interaction. Comparatively, in the $FKBP12_{D37V}$•rapamycin complex the methyl groups of valine do not display any switching (**Figure 4C** and **Supplementary Figures S10B**). Furthermore, all peaks in the histograms are broader than in the wild-type case, the one corresponding to the complex at equilibrium as well as the ones corresponding to transient interactions occurring upon dissociation. These two observations are in line with the inability of the alkyl moieties to form strong, directional intermolecular bonds. Finally, the most obvious signature of the weakening of the association between rapamycin and FKBP12 after the D37V substitution is that the atom-pair distances start to increase earlier, nearly as soon as the bias is applied.

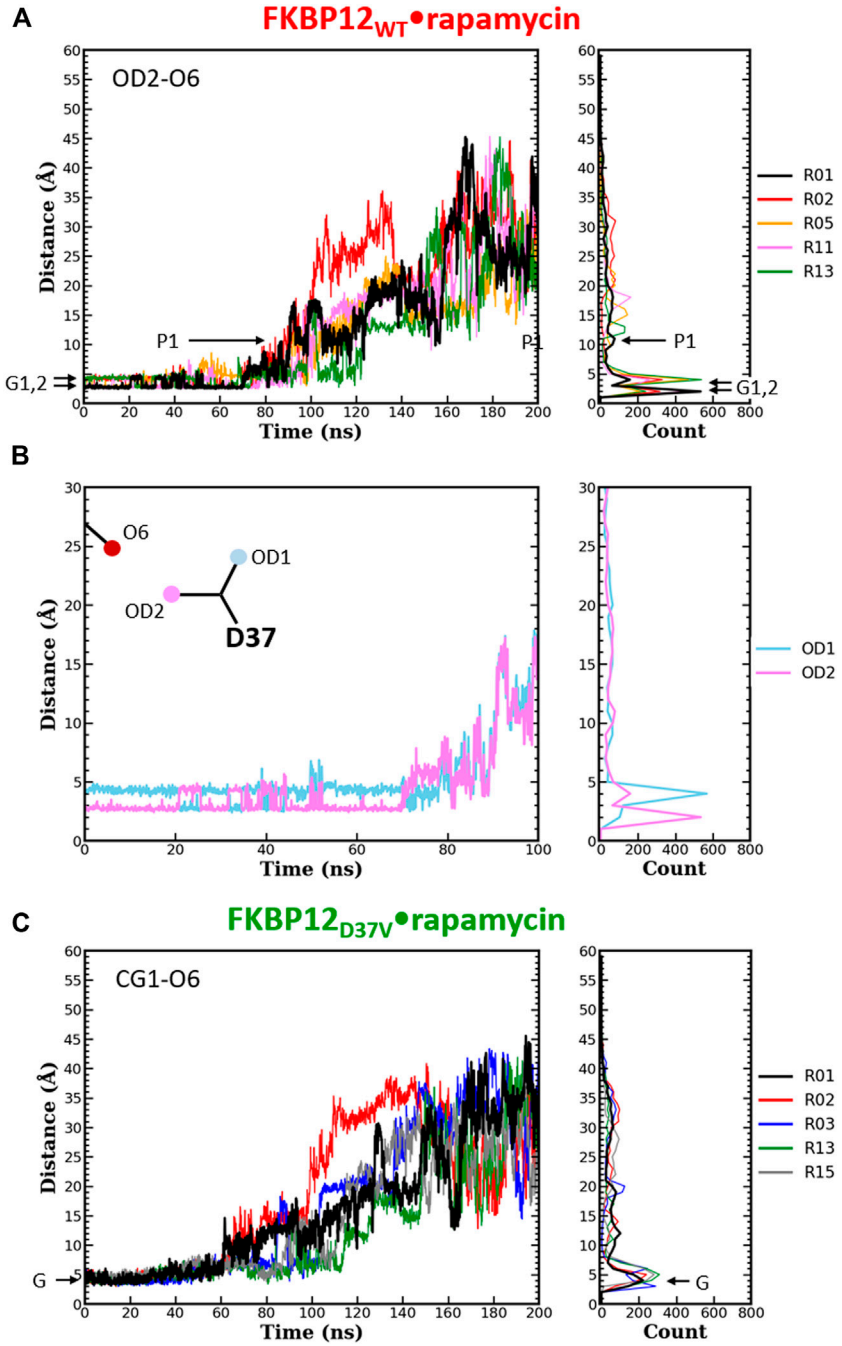


FIGURE 4 | Comparison of the separation distance patterns obtained during the dissociation of rapamycin from the wild-type FKBP12 and from the D37V mutant. Pulling on the molecular partners CoG starts at 10 ns. **(A)** Evolution of the distance between the O6 hydroxyl oxygen of rapamycin and the OD2 carboxylate oxygen of residue D37 in the $FKBP12_{WT}$•rapamycin complex. For the sake of clarity, the data are only provided for the R01 lower-bound PMF profile and for four other ones. The Gi and Pi markers indicate, for the time-trace and the histogram associated with the lower-bound profile, the ground levels and the different plateaus, respectively. The data for all 15 CPUS simulations are displayed in **Supplementary Figures S10A**. **(B)** Enlargement on the first tens of ns of the time-trace associated with the lower-bound PMF profile. The distance to the second carboxylate oxygen of residue D37 (OD1) has been added to evidence the H-bond switching between the two acceptor atoms. **Supplementary Figures S10A** provides similar data for the whole set of simulations. **(C)** Evolution of the distance between the O6 hydroxyl oxygen of rapamycin and one of the CG1 methyl carbon of residue V37 in the $FKBP12_{D37V}$•rapamycin complex. Once more data are only provided for the R01 lower-bound PMF profile and for four other ones, see **Supplementary Figures S10B** for an exhaustive presentation of the results.

Dissociation Path Tracing

The umbrella sampling simulation trajectories were processed to trace the physical paths of dissociation for the $FKBP12_{WT}$•rapamycin, $FKBP12_{Y82F}$•rapamycin, and $FKBP12_{D37V}$•rapamycin complexes (**Supplementary Figures S11–S13**, respectively). Rapamycin is always seen to traverse to the solvent in a curvilinear way, exiting through either the top left or the top right of the binding cavity. For the paths corresponding to the lower-bound PMF profiles, the macrolide initially slides on the surface of the protein smoothly and then wind a lot. In contrast, the paths associated with the highest PMF profiles are relatively less curvy (**Supplementary Figure S14**). The highly curved nature of lower-bound paths suggests that rapamycin does not get stuck on FKBP12 and tends to free from its surface towards the solvent quite early, or at least it does not get trapped into local energy minima. On the other hand, the highest PMF paths for $FKBP12_{WT}$ and $FKBP12_{Y82F}$ display accumulations of spheres that could be the signature of some interactions outside of the binding site.

DISCUSSION AND CONCLUSIONS

In this work, we present a curvilinear-path umbrella sampling simulation approach to estimating the free energy of binding for the FKBP12•rapamycin complex. The strategy consists of several crucial steps that includes sampling enhancement along multiple independent curvilinear paths, construction of PMFs, determination of the lower-bound profile and of the corresponding ΔG_{PMF} term, and then correction of this term to obtain the standard free energy of binding. The corresponding theoretical framework was recently developed and successfully implemented to estimate protein–protein interaction energetics (Joshi and Lin, 2019). Here, we extended the approach to a protein–ligand system, and our results are in good agreement with reported experimental data. Although overall the CPUS approach is quite effective to estimate reliably interaction energetics, there are a few things one needs to keep in mind before computing the correction term. The umbrella implementation needs to decide the distance range, i.e., starting reference bound distance and the end separation distance. Upon constructing PMF, the profile may depict a signature of complete dissociation in terms of converging of the curve to some value. However, in the computation of the correction term, we suggested interaction energetics (VDW component) criterion to judge the complete dissociation (see the Methods and Materials section). The simulations should be run to that separation distance where the VDW component converges to zero. In the case of protein–small ligand dissociation reaction, this distance may vary a bit depending upon the path that the ligand could take. Therefore, while setting the end separation distance, it is advised to set it to sufficiently large values, considering that the dissociating compound may slide/rebind on the protein surface after an initial dissociation.

Although Lee and Olson mentioned the necessity of sampling along a curvilinear physical trajectory, they estimated the absolute binding affinities using a vectorial path (Lee and Olson, 2006). Furthermore, their energetic results were in good agreement with the experiments but quadratic approximations were required. Doudou et al. (2009) discussed rigorously the problem associated with energetics obtained by sampling along a predefined vectorial reaction coordinate and pointed out related inconsistencies in the free energy estimation along different linear paths/directions. The CPUS approach is a naïve approach that tries to address these issues by conducting successive umbrella sampling simulations. The approach is applicable especially when a precise input, such as the detailed shape of the binding pocket or a predefined vector of confinement, is unavailable. It relies on the minimal use of restraint potentials (namely, only one biasing potential in the umbrella sampling simulation); hence, it fairly bypasses some major challenges one can face for appropriate de-biasing. However, the cost has to be paid in terms of conducting a sufficiently large number of umbrella sampling simulations to find the lower-bound of the PMF profiles. In addition, due to the complex nature of biomolecular systems it is rather difficult, if not impossible, to decide *a priori* the number of multiple-walker umbrella sampling simulations to be conducted. This difficulty is actually also shared by any variational-based approaches, e.g., quantum Monte Carlo (Ceperley, 1978; Ceperley and Alder, 1980). It would be always a good sign to find some robust lower-bound from the multiple PMF profiles. Thus, the current work delivers an important message that the PMF profile from any single umbrella sampling simulation (irrespective of whether predefined vector-based or non-predefined curved-path-based sampling enhancement) may likely be misleading. With ongoing advances in GPU architecture, the simulation time is expected to keep decreasing, enabling more simulations to be performed in parallel. Indeed, the implicit/continuum solvent model would rather be a more natural choice for rapid estimation and reducing the computation cost. However, some serious concerns are yet to be suitably resolved, such as too large numerical ranges of estimated energies and strong dependence of the energetics on the employed continuum solvation (such as the choice of protein dielectric constant, the definition of protein boundary, etc.) (Genheden and Ryde, 2015). Therefore, explicit solvent modes are highly preferred.

A related issue in the CPUS approach is to evaluate if the molecular behaviors associated with the lower-bound PMF profile significantly differ from the ones associated with the higher profiles. In our previous work, we traced the protein–protein dissociation paths and observed that there was a clear difference in the direction of traverse between the lower-bound case and the rest (Joshi and Lin, 2019). This distinction between the paths was made possible because the two prominent paths had been previously evidenced using an extensive brute force adaptive MD simulation approach (Plattner et al., 2017). In the case of the FKBP12•rapamycin system, we observed that the rapamycin initially moves up and is then released in the solvent by taking either a left or a right path (**Supplementary Figures S11–S14**). However, we could not find any supportive correlation between the plateau value of the PMF profile and the path taken. Similarly, we looked for correlations between the PMF final value and the presence of sphere clusters along the way to dissociation,

the latter signature being interpreted as the temporary immobilization of rapamycin in a local energy minimum present at the surface of FKBP12. Although few of such potential wells could be identified (see, for instance, the red circles in **Supplementary Figure S14**), these finding are at the moment only qualitative. As a consequence, the physical paths of dissociation do not enable us to differentiate one or a subset of simulations from the other ones. In fact, it would be interesting to see whether the paths traced from the CPUS approach are consistent with the ones revealed by other pathway sampling methods, such as the string method (Weinan et al., 2002), the minimum free energy path (Maragliano et al., 2006), the adaptive and unconstrained enhanced sampling (Miao and McCammon, 2016), the nudged elastic band (NEB) method (Jónsson et al., 1998), or stochastic difference equation (in length version) SDEL (Arora and Schlick, 2005).

In this work, we also performed a detailed analysis on the atom-pair distance distribution patterns (**Figures 3**, **4** and **Supplementary Figures S9, S10**). For all three variants the data that yielded the lower-bound PMF profile and the ones issued from the 14 other runs were rather similar in terms of global behavior, i.e., notwithstanding the variability one can expect when looking at single molecules. Since the ligand conformation may also play a crucial role in the dissociation mechanism, we plotted the temporal evolution of the RMSD for the heavy atoms (C, N, O) of rapamycin (**Supplementary Figures S15–S17**). Although some peaks or steps are present, the overall RMSD profiles are quite stable for all CPUS simulations and the dissociation process does not seem to be coupled with important conformational transitions in the macrolide. The two hydrogen-bonded atom-pairs investigated here, i.e., Y(F)82-O3 and residue D(V)37-O6, have significantly different patterns. Moreover, the comparison of distance distribution patters between atom-pairs of the side-chain benzyl ring of FKBP12 variants (Y82 and F82) and the O6 of rapamycin pointed out that they might be involved in the anion-π type of interactions. These atomistic-level factors could contribute to the excess dissipation along the curvilinear trajectory of dissociation, and further detailed investigations are needed to have a comprehensive understanding of the relationship between estimated energetics and atom-pair interactions.

Thus, the CPUS approach paves the way to reveal binding energetics along with mechanistic details. The approach is highly generalized and can be implemented to almost any protein–ligand systems, along with protein–protein systems, as well. The approach offers a strong platform to perform several types of in-depth analysis towards revealing the underlying mechanistic details. The drug discovery process often deals with known drug targets and with new or modified drug molecules. In the case of rapamycin-FRB interactions, one such study recently identified DL001 compound that reduces the side effects *in vivo* (Schreiber et al., 2019). The CPUS approach-based characterization could be helpful for further investigation and for the design of such effective drugs.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

Design of the study: J-HL and CG; NMR experiments: S-YH; MD simulations: DJ; data analysis: DJ, S-YH, CG, and J-HL; and manuscript writing: DJ, S-YH, CG, and J-HL.

FUNDING

J-HL and CG benefited from grant "DynaTOR" from French MAE/MESR and Taiwan MoST (Orchid PHC program). DJ and S-YH were supported by the grants of MoST [108-2112-M-001-037-MY3, 110-2811-M-001-580, and 110-2927-I-001-512–] and RCAS, Academia Sinica. The research of JHL was supported by the Investigator Award [AS-IA-108-M05] of Academia Sinica. The HSQC and CSP experiments were conducted at the NMR core facility of Academia Sinica. The Molecular Motors and Machines team at the Institut de Biologie de l'Ecole Normale Supérieure is an "Equipe Labellisée" by the Ligue Nationale Contre la Cancer.

ACKNOWLEDGMENTS

We would like to thank L. Catoire (Institut de Biologie Physico-Chimique, Paris) and E. Lescop (Institut de Chimie des Substances Naturelles, Gif-sur-Yvette) for helpful discussions and D. Kostrz and T. Strick (Institut de Biologie de l'Ecole Normale Supérieure, Paris) for the gift of the FKBP12 plasmid.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.879000/full#supplementary-material

REFERENCES

Arora, K., and Schlick, T. (2005). Conformational Transition Pathway of Polymerase β/DNA upon Binding Correct Incoming Substrate. *J. Phys. Chem. B* 109, 5358–5367. doi:10.1021/jp0446377

Banaszynski, L. A., Liu, C. W., and Wandless, T. J. (2005). Characterization of the FKBP·Rapamycin·FRB Ternary Complex. *J. Am. Chem. Soc.* 127, 4715–4721. doi:10.1021/ja043277y

Bierer, B. E., Mattila, P. S., Standaert, R. F., Herzenberg, L. A., Burakoff, S. J., Crabtree, G., et al. (1990). Two Distinct Signal Transmission Pathways in T Lymphocytes Are Inhibited by Complexes Formed between an Immunophilin

and Either FK506 or Rapamycin. *Proc. Natl. Acad. Sci. U.S.A.* 87, 9231–9235. doi:10.1073/pnas.87.23.9231

Bossard, M. J., Bergsma, D. J., Brandt, M., Livi, G. P., Eng, W. K., Johnson, R. K., et al. (1994). Catalytic and Ligand Binding Properties of the FK506 Binding Protein FKBP12: Effects of the Single Amino Acid Substitution of Tyr82 to Leu. *Biochem. J.* 297, 365–372. doi:10.1042/bj2970365

Case, D. A., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., Iii, et al. (2020). *AMBER 2020*. San Francisco: University of California.

Case, D. A., Cheatham, T. E., Iii, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr, et al. (2005). The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* 26, 1668–1688. doi:10.1002/jcc.20290

Ceperley, D. (1978). Ground State of the Fermion One-Component Plasma: A Monte Carlo Study in Two and Three Dimensions. *Phys. Rev. B* 18, 3126–3138. doi:10.1103/physrevb.18.3126

Ceperley, D. M., and Alder, B. J. (1980). Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* 45, 566–569. doi:10.1103/physrevlett.45.566

Chifotides, H. T., and Dunbar, K. R. (2013). Anion–π Interactions in Supramolecular Architectures. *Acc. Chem. Res.* 46, 894–906. doi:10.1021/ar300251k

Choi, J., Chen, J., Schreiber, S. L., and Clardy, J. (1996). Structure of the FKBP12-Rapamycin Complex Interacting with Binding Domain of Human FRAP. *Science* 273, 239–242. doi:10.1126/science.273.5272.239

Connelly, P. R., Aldape, R. A., Bruzzese, F. J., Chambers, S. P., Fitzgibbon, M. J., Fleming, M. A., et al. (1994). Enthalpy of Hydrogen Bond Formation in Aprotein-Ligand Binding Reaction. *Proc. Natl. Acad. Sci. U.S.A.* 91, 1964–1968. doi:10.1073/pnas.91.5.1964

Connelly, P. R., Thomson, J. A., Fitzgibbon, M. J., and Bruzzese, F. J. (1993). Probing Hydration Contributions to the Thermodynamics of Ligand Binding by Proteins. Enthalpy and Heat Capacity Changes of Tacrolimus and Rapamycin Binding to FK506 Binding Protein in Deuterium Oxide and Water. *Biochemistry* 32, 5583–5590. doi:10.1021/bi00072a013

Decenzo, M. T., Park, S. T., Jarrett, B. P., Aldape, R. A., Futer, O., Murcko, M. A., et al. (1996a). FK506-binding Protein Mutational Analysis: Defining the Active-Site Residue Contributions to Catalysis and the Stability of Ligand Complexes. *Protein Eng. Des. Sel.* 9, 173–180. doi:10.1093/protein/9.2.173

Decenzo, M. T., Park, S. T., Jarrett, B. P., Aldape, R. A., Futer, O., Murcko, M. A., et al. (1996b). FK506-binding Protein Mutational Analysis: Defining the Active-Site Residue Contributions to Catalysis and the Stability of Ligand Complexes. *Protein Eng. Des. Sel.* 9, 173–180. doi:10.1093/protein/9.2.173

Dickman, D. A., Ding, H., Li, Q., Nilius, A. M., Balli, D. J., Ballaron, S. J., et al. (2000). Antifungal Rapamycin Analogues with Reduced Immunosuppressive Activity. *Bioorg. Med. Chem. Lett.* 10, 1405–1408. doi:10.1016/s0960-894x(00)00184-0

Dickson, A., and Lotz, S. D. (2016). Ligand Release Pathways Obtained with WExplore: Residence Times and Mechanisms. *J. Phys. Chem. B* 120, 5377–5385. doi:10.1021/acs.jpcb.6b04012

Doudou, S., Burton, N. A., and Henchman, R. H. (2009). Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force. *J. Chem. Theory Comput.* 5, 909–918. doi:10.1021/ct8002354

Fujitani, H., Tanida, Y., Ito, M., Jayachandran, G., Snow, C. D., Shirts, M. R., et al. (2005). Direct Calculation of the Binding Free Energies of FKBP Ligands. *J. Chem. Phys.* 123, 084108. doi:10.1063/1.1999637

Galat, A. (2013). Functional Diversity and Pharmacological Profiles of the FKBPs and Their Complexes with Small Natural Ligands. *Cell. Mol. Life Sci.* 70, 3243–3275. doi:10.1007/s00018-012-1206-z

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

Giese, M., Albrecht, M., and Rissanen, K. (2016). Experimental Investigation of Anion-π Interactions - Applications and Biochemical Relevance. *Chem. Commun.* 52, 1778–1795. doi:10.1039/c5cc09072e

Graziani, F., Aldegheri, L., and Terstappen, G. C. (1999). High Throughput Scintillation Proximity Assay for the Identification of FKBP-12 Ligands. *SLAS Discov.* 4, 3–7. doi:10.1177/108705719900400102

Grossfield, A. (2013). WHAM: the Weighted Histogram Analysis Method, Version 2.0. 9. Available at membrane. urmc. rochester. edu/content/wham. Accessed November 15, 2013.

Hamilton, G. S., and Steiner, J. P. (1998). Immunophilins: Beyond Immunosuppression. *J. Med. Chem.* 41, 5119–5143. doi:10.1021/jm980307x

Holt, D. A., Luengo, J. I., Yamashita, D. S., Oh, H. J., Konialian, A. L., Yen, H. K., et al. (1993). Design, Synthesis, and Kinetic Evaluation of High-Affinity FKBP Ligands and the X-Ray Crystal Structures of Their Complexes with FKBP12. *J. Am. Chem. Soc.* 115, 9925–9938. doi:10.1021/ja00075a008

Ikura, T., and Ito, N. (2007). Requirements for Peptidyl-Prolyl Isomerization Activity: A Comprehensive Mutational Analysis of the Substrate-Binding Cavity of FK506-Binding Protein 12. *Protein Sci.* 16, 2618–2625. doi:10.1110/ps.073203707

Jónsson, H., Mills, G., and Jacobsen, K. W. (1998). Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions.

Joshi, D. C., and Lin, J. H. (2019). Delineating Protein-Protein Curvilinear Dissociation Pathways and Energetics with Naïve Multiple-Walker Umbrella Sampling Simulations. *J. Comput. Chem.* 40, 1652–1663. doi:10.1002/jcc.25821

Karplus, M., and Mccammon, J. A. (2002). Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* 9, 646–652. doi:10.1038/nsb0902-646

Kästner, J. (2011). Umbrella Sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1, 932–942. doi:10.1002/wcms.66

Kostrz, D., Wayment-Steele, H. K., Wang, J. L., Follenfant, M., Pande, V. S., Strick, T. R., et al. (2019). A Modular DNA Scaffold to Study Protein-Protein Interactions at Single-Molecule Resolution. *Nat. Nanotechnol.* 14, 988–993. doi:10.1038/s41565-019-0542-7

Kozany, C., März, A., Kress, C., and Hausch, F. (2009). Fluorescent Probes to Characterise FK506-Binding Proteins. *Chembiochem* 10, 1402–1410. doi:10.1002/cbic.200800806

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* 13, 1011–1021. doi:10.1002/jcc.540130812

Lee, M. S., and Olson, M. A. (2006). Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophysical J.* 90, 864–877. doi:10.1529/biophysj.105.071589

Lin, J.-H., Perryman, A. L., Schames, J. R., and Mccammon, J. A. (2002). Computational Drug Design Accommodating Receptor Flexibility: the Relaxed Complex Scheme. *J. Am. Chem. Soc.* 124, 5632–5633. doi:10.1021/ja0260162

Lin, J.-H., Perryman, A. L., Schames, J. R., and Mccammon, J. A. (2003). The Relaxed Complex Method: Accommodating Receptor Flexibility for Drug Design with an Improved Scoring Scheme. *Biopolymers* 68, 47–62. doi:10.1002/bip.10218

Lu, C., and Wang, Z.-X. (2017). Quantitative Analysis of Ligand Induced Heterodimerization of Two Distinct Receptors. *Anal. Chem.* 89, 6926–6930. doi:10.1021/acs.analchem.7b01274

Lucas, X., Bauzá, A., Frontera, A., and Quiñonero, D. (2016). A Thorough Anion-π Interaction Study in Biomolecules: on the Importance of Cooperativity Effects. *Chem. Sci.* 7, 1038–1050. doi:10.1039/c5sc01386k

Luengo, J. I., Yamashita, D. S., Dunnington, D., Beck, A. K., Rozamus, L. W., Yen, H.-K., et al. (1995). Structure-activity Studies of Rapamycin Analogs: Evidence that the C-7 Methoxy Group Is Part of the Effector Domain and Positioned at the FKBP12-FRAP Interface. *Chem. Biol.* 2, 471–481. doi:10.1016/1074-5521(95)90264-3

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. Theory Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255

Maragliano, L., Fischer, A., Vanden-Eijnden, E., and Ciccotti, G. (2006). String Method in Collective Variables: Minimum Free Energy Paths and Isocommittor Surfaces. *J. Chem. Phys.* 125, 024106. doi:10.1063/1.2212942

Miao, Y., and Mccammon, J. A. (2016). Unconstrained Enhanced Sampling for Free Energy Calculations of Biomolecules: a Review. *Mol. Simul.* 42, 1046–1055. doi:10.1080/08927022.2015.1121541

Mobley, D. L., and Gilson, M. K. (2017). Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* 46, 531–558. doi:10.1146/annurev-biophys-070816-033654

Nerattini, F., Chelli, R., and Procacci, P. (2016). II. Dissociation Free Energies in Drug-Receptor Systems via Nonequilibrium Alchemical Simulations: Application to the FK506-Related Immunophilin Ligands. *Phys. Chem. Chem. Phys.* 18, 15005–15018. doi:10.1039/c5cp05521k

Olivieri, L., and Gardebien, F. (2011). Molecular Dynamics Simulations of a Binding Intermediate between FKBP12 and a High-Affinity Ligand. *J. Chem. Theory Comput.* 7, 725–741. doi:10.1021/ct100394d

Pearlman, D. A., and Connelly, P. R. (1995). Determination of the Differential Effects of Hydrogen Bonding and Water Release on the Binding of FK506 to Native and Tyr82→Phe82 FKBP-12 Proteins Using Free Energy Simulations. *J. Mol. Biol.* 248, 696–717. doi:10.1006/jmbi.1995.0252

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, 1605–1612. doi:10.1002/jcc.20084

Plattner, N., Doerr, S., De Fabritiis, G., and Noé, F. (2017). Complete Protein-Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.* 9, 1005–1011. doi:10.1038/nchem.2785

Roe, D. R., and Cheatham, T. E., Iii (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

Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3, 198–210. doi:10.1002/wcms.1121

Sapienza, P. J., Mauldin, R. V., and Lee, A. L. (2011). Multi-Timescale Dynamics Study of FKBP12 along the Rapamycin-mTOR Binding Coordinate. *J. Mol. Biol.* 405, 378–394. doi:10.1016/j.jmb.2010.10.037

Schottel, B. L., Chifotides, H. T., and Dunbar, K. R. (2008). Anion-π Interactions. *Chem. Soc. Rev.* 37, 68–83. doi:10.1039/b614208g

Schreiber, K. H., Arriola Apelo, S. I., Yu, D., Brinkman, J. A., Velarde, M. C., Syed, F. A., et al. (2019). A Novel Rapamycin Analog Is Highly Selective for mTORC1 *In Vivo*. *Nat. Commun.* 10, 3194. doi:10.1038/s41467-019-11174-0

Schuler, W., Sedrani, R., Cottens, S., H?berlin, B., Schulz, M., Schuurman, H.-J., et al. (1997). Sdz Rad, a New Rapamycin Derivative. *Transplantation* 64, 36–42. doi:10.1097/00007890-199707150-00008

Shor, B., Zhang, W.-G., Toral-Barza, L., Lucas, J., Abraham, R. T., Gibbons, J. J., et al. (2008). A New Pharmacologic Action of CCI-779 Involves FKBP12-independent Inhibition of mTOR Kinase Activity and Profound Repression of Global Protein Synthesis. *Cancer Res.* 68, 2934–2943. doi:10.1158/0008-5472.can-07-6487

Singh, V., Nand, A., and Sarita, S. (2015). Universal Screening Platform Using Three-Dimensional Small Molecule Microarray Based on Surface Plasmon Resonance Imaging. *RSC Adv.* 5, 87259–87265. doi:10.1039/c5ra15637h

Solomentsev, G., Diehl, C., and Akke, M. (2018). Conformational Entropy of Fk506 Binding to Fkbp12 Determined by Nuclear Magnetic Resonance Relaxation and Molecular Dynamics Simulations. *Biochemistry* 57, 1451–1461. doi:10.1021/acs.biochem.7b01256

Sun, F., Li, P., Ding, Y., Wang, L., Bartlam, M., Shu, C., et al. (2003). Design and Structure-Based Study of New Potential FKBP12 Inhibitors. *Biophysical J.* 85, 3194–3201. doi:10.1016/s0006-3495(03)74737-7

Swanson, J. M. J., Henchman, R. H., and Mccammon, J. A. (2004). Revisiting Free Energy Calculations: a Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophysical J.* 86, 67–74. doi:10.1016/s0006-3495(04)74084-9

Tamura, T., Kioi, Y., Miki, T., Tsukiji, S., and Hamachi, I. (2013). Fluorophore Labeling of Native FKBP12 by Ligand-Directed Tosyl Chemistry Allows Detection of its Molecular Interactions *In Vitro* and in Living Cells. *J. Am. Chem. Soc.* 135, 6782–6785. doi:10.1021/ja401956b

Torrie, G. M., and Valleau, J. P. (1977). Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* 23, 187–199. doi:10.1016/0021-9991(77)90121-8

Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L., and Clardy, J. (1993). Atomic Structures of the Human Immunophilin FKBP-12 Complexes with FK506 and Rapamycin. *J. Mol. Biol.* 229, 105–124. doi:10.1006/jmbi.1993.1012

Van Duyne, G. D., Standaert, R. F., Schreiber, S. L., and Clardy, J. (1991). Atomic Structure of the Rapamycin Human Immunophilin FKBP-12 Complex. *J. Am. Chem. Soc.* 113, 7433–7434. doi:10.1021/ja00019a057

Wagner, R., Rhoades, T. A., Or, Y. S., Lane, B. C., Hsieh, G., Mollison, K. W., et al. (1998). 32-Ascomycinyloxyacetic Acid Derived Immunosuppressants. Independence of Immunophilin Binding and Immunosuppressive Potency. *J. Med. Chem.* 41, 1764–1776. doi:10.1021/jm960066y

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995). LIGPLOT: a Program to Generate Schematic Diagrams of Protein-Ligand Interactions. *Protein Eng. Des. Sel.* 8, 127–134. doi:10.1093/protein/8.2.127

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, Y., Barnett, S. F. H., Le, S., Guo, Z., Zhong, X., Kanchanawong, P., et al. (2019). Label-free Single-Molecule Quantification of Rapamycin-Induced FKBP-FRB Dimerization for Direct Control of Cellular Mechanotransduction. *Nano Lett.* 19, 7514–7525. doi:10.1021/acs.nanolett.9b03364

Wear, M. A., Patterson, A., and Walkinshaw, M. D. (2007). A Kinetically Trapped Intermediate of FK506 Binding Protein Forms *In Vitro*: Chaperone Machinery Dominates Protein Folding *In Vivo*. *Protein Expr. Purif.* 51, 80–95. doi:10.1016/j.pep.2006.06.019

Wear, M. A., and Walkinshaw, M. D. (2007). Determination of the Rate Constants for the FK506 Binding Protein/rapamycin Interaction Using Surface Plasmon Resonance: an Alternative Sensor Surface for Ni2+-Nitrilotriacetic Acid Immobilization of His-Tagged Proteins. *Anal. Biochem.* 371, 250–252. doi:10.1016/j.ab.2007.06.034

Weinan, E., Ren, W., and Vanden-Eijnden, E. (2002). String Method for the Study of Rare Events. *Phys. Rev. B* 66, 052301. doi:10.1103/physrevb.66.052301

Williamson, M. P. (2018). "Chemical Shift Perturbation," in *Chemical Shift perturbationModern Magnetic Resonance*. Editor G.A. Webb (Cham: Springer International Publishing), 995–1012. doi:10.1007/978-3-319-28388-3_76

Wilson, K. P., Yamashita, M. M., Sintchak, M. D., Rotstein, S. H., Murcko, M. A., Boger, J., et al. (1995). Comparative X-Ray Structures of the Major Binding Protein for the Immunosuppressant FK506 (Tacrolimus) in Unliganded Form and in Complex with FK506 and Rapamycin. *Acta Cryst. D.* 51, 511–521. doi:10.1107/s0907444994014514

Wu, X., Wang, L., Han, Y., Regan, N., Li, P.-K., Villalona, M. A., et al. (2011). Creating Diverse Target-Binding Surfaces on FKBP12: Synthesis and Evaluation of a Rapamycin Analogue Library. *ACS Comb. Sci.* 13, 486–495. doi:10.1021/co200057n

Yang, C.-J., Takeda, M., Terauchi, T., Jee, J., and Kainosho, M. (2015). Differential Large-Amplitude Breathing Motions in the Interface of FKBP12-Drug Complexes. *Biochemistry* 54, 6983–6995. doi:10.1021/acs.biochem.5b00820

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 11 July 2022
doi: 10.3389/fmolb.2022.879251

PASSer2.0: Accurate Prediction of Protein Allosteric Sites Through Automated Machine Learning

Sian Xiao, Hao Tian and Peng Tao**

Center for Research Computing, Center for Drug Discovery, Design and Delivery (CD4), Department of Chemistry, Southern Methodist University, Dallas, TX, United States

Allostery is a fundamental process in regulating protein activities. The discovery, design, and development of allosteric drugs demand better identification of allosteric sites. Several computational methods have been developed previously to predict allosteric sites using static pocket features and protein dynamics. Here, we define a baseline model for allosteric site prediction and present a computational model using automated machine learning. Our model, PASSer2.0, advanced the previous results and performed well across multiple indicators with 82.7% of allosteric pockets appearing among the top three positions. The trained machine learning model has been integrated with the Protein Allosteric Sites Server (PASSer) to facilitate allosteric drug discovery.

Keywords: allostery, machine learning, allosteric site prediction, automated machine learning (AutoML), deep learning

OPEN ACCESS

Edited by:
Chia-en A. Chang, University of California, Riverside, United States

Reviewed by:
Gennady Verkhivker, Chapman University, United States
Junmei Wang, University of Pittsburgh, United States

****Correspondence:***
Hao Tian
haot@smu.edu
Peng Tao
ptao@smu.edu

Specialty section:
This article was submitted to Molecular Recognition, a section of the journal Frontiers in Molecular Biosciences

Received: *19 February 2022*
Accepted: *23 May 2022*
Published: *11 July 2022*

Citation:
Xiao S, Tian H and Tao P (2022) PASSer2.0: Accurate Prediction of Protein Allosteric Sites Through Automated Machine Learning. Front. Mol. Biosci. 9:879251. doi: 10.3389/fmolb.2022.879251

1 INTRODUCTION

Allostery is a fundamental process that regulates protein functional activities and is known to play a key role in biology (Gunasekaran et al. 2004). In an allosteric process, an effector molecule binds to a protein at its allosteric site, often resulting in conformational and dynamical changes (Srinivasan et al. 2014; Huang et al. 2013). Allosteric drug development is promising for many reasons: the allosteric drugs could be more selective and less toxic with fewer side effects; they can either activate or inhibit proteins; they can be used in conjunction with orthosteric drugs. Due to these advantages, the development of allosteric drugs has gradually increased in recent years (Wagner et al. 2016; Nussinov et al. 2011; Nussinov and Tsai 2013).

Several methods have been developed to detect and predict allosteric sites in proteins, such as normal mode analysis (NMA) (Panjkovich and Daura 2012), molecular dynamics (MD) simulations (Laine et al. 2010), and machine learning (ML) models (Amor et al. 2016; Bian et al. 2019; Huang et al. 2013). Several current methods are available as web servers or open-source packages, such as Allosite (Huang et al. 2013), SPACER (Goncearenco et al. 2013), PARS (Panjkovich and Daura 2014), AlloPred (Greener and Sternberg 2015), AllositePro (Song et al. 2017), and PASSer (Tian et al. 2021a). These studies have demonstrated the feasibility of allosteric site prediction models which combine pocket features and protein dynamics. As summarized by Lu et al. (2014), these studies can be classified as structure-based, dynamics-based, NMA-based, or combined prediction approaches. In structure-based approaches, such as Allosite, site descriptors describing chemical and physical properties of protein pockets are calculated as features for prediction. NMA-based approaches, such as PARS, take the ability of NMA, which can provide global modes that bear functional significance, for discovering protein sites that can mediate or propagate allosteric signals. In dynamics-based approaches, MD simulations and a two-state Ga model are used to construct a conformational or

energy landscape, in which the latter can be used to calculate population distribution upon perturbation. SPACER combines dynamics-based and NMA-based approaches, which apply Monte Carlo simulations and normal mode evaluation to unravel latent allosteric sites.

The past decade has witnessed the rapid development of machine learning in chemistry and biology (Zhang et al. 2020; Chen L. et al. 2021; Tian et al. 2020; Tian et al. 2021b; Tian et al. 2022). ML methods have been shown to be superior in the classification of protein allosteric pockets. Allosite and AlloPred used support vector machine (SVM) (Suykens and Vandewalle 1999) with curated features. Chen et al. (2016) used random forest (RF) (Liaw and Wiener 2002) to construct a three-way predictive model. Our previous study (Tian et al. 2021a) used an ensemble learning method combining the results of eXtreme gradient boosting (XGBoost) (Chen and Guestrin 2016) and graph convolutional neural networks (GCNNs) (Kipf and Welling 2016).

Recently, automated machine learning (AutoML) has emerged as a novel strategy to implement machine learning methods to solve real-world problems Hutter et al. (2019). It has been widely applied in biomedical or chemistry fields like nucleic acid (Chen Z. et al. 2021), healthcare (Waring et al. 2020), and disease studies (Karaglani et al. 2020; Panagopoulou et al. 2021). As the name suggests, AutoML helps to automate the machine learning pipeline, from data processing, model selection, and ensemble to hyperparameter tuning. This saves human power from the time-consuming and iterative tasks of machine learning model development Yao et al. (2018). Also, AutoML offers the opportunities to produce simpler solutions with superior model performance (Elshawi et al. 2019).

In this study, we first defined the baseline for protein allosteric site prediction, an algorithm that identifies the pocket with the highest pocket score among all pockets detected by FPocket (Le Guilloux et al. 2009) as allosteric. This primitive baseline predictor has accuracy, precision, recall, and F1 score values of 0.968, 0.689, 0.571, and 0.624, respectively. Then, we applied two AutoML frameworks, AutoKeras (Jin et al. 2019) and AutoGluon (Erickson et al. 2020), for the prediction of protein allosteric sites. Our model is shown to be robust and powerful under various indicators with precision, recall, and F1 score values of 0.850, 0.616, and 0.701, respectively, on the test set, and 82.7% of allosteric sites in the test set are ranked among the top three positions. We also applied the well-trained model to predict allosteric sites from novel proteins that are not included in the training set and demonstrated their binding structures.

2 MATERIALS AND METHODS

2.1 Protein Database

The protein data used in this work were collected from the Allosteric Database (ASD) (Huang et al. 2011). Its newest version contains a total of 1,949 entries of allosteric sites, each with different proteins and modulators Liu et al. (2020). However, data need to be filtered from ASD under certain criteria to ensure the data quality Zha et al. (2022). To ensure protein quality and diversity, Huang et al. (2013) selected 90 proteins using the previous rules: protein structures with either resolution below 3 Å or missing residues in the allosteric sites were removed, and redundant proteins that have more than 30% sequence identity were filtered out. ASBench (Huang et al. 2015), an optimized selection of ASD data, includes a core set with 235 unique allosteric sites and a core-diversity set with 147 structurally diverse allosteric sites. Here, we use 90 proteins from ASD and 138 proteins in the core-diversity set from ASBench. A total of 204 proteins were used in this study, after removing the duplicate records. The selected proteins were stored in the GitHub repository for this study.

2.2 Site Descriptors

FPocket, a geometry-based algorithm to identify pockets, is used to detect pockets on the surface of the selected proteins. For each of the detected pockets, 19 numerical features are calculated from FPocket (**Supplementary Table S1**). Compared with other web servers and open-source pocket detection packages, FPocket is superior in execution time and the ease to be integrated with other models.

For the 90 proteins from ASD, a pocket is labeled as either 1 (positive) if it contains at least one residue identified as binding to allosteric modulators or 0 (negative) if it does not contain such residues. Therefore, a protein structure may have more than one positive label. A total of 2,123 pockets were detected with 133 pockets being labeled as allosteric sites. For the 138 proteins from ASBench, a total of 3,708 pockets were detected. A pocket is labeled as 1 (positive) only if its centroid is the closest to that of the allosteric modulator, otherwise 0 (negative).

2.3 Automated Machine Learning

The implementation of the state-of-the-art ML methods normally requires extensive domain knowledge and experience. This process includes data preparation and preprocessing, feature engineering, model selection, and hyperparameter tuning, which are time-consuming and challenging. Automated machine learning aims to free human effort from this process.

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras offers consistent and simple APIs and provides clear and actionable error messages. It also has extensive documentation and developer guides. AutoKeras (Jin et al. 2019) is an AutoML system based on Keras, enabling Bayesian optimization to guide the network morphism for efficient neural architecture parameter search. In the current study, AutoKeras v1.0.16 is applied.

Developed by Amazon Web Services, AutoGluon (Erickson et al. 2020) automates these ML tasks and achieves the best performance. Moreover, AutoGluon includes techniques for multi-layer stacking that can further boost ML performance. AutoGluon is advantageous in: (1) simplicity: straightforward and user-friendly APIs; (2) robustness: no data manipulation or feature engineering required; (3) predictable-timing: ML models are trained within the allocated time; (4) fault-tolerance: the training process can be resumed after interruption. Also,

TABLE 1 | Binary classification results in a confusion matrix.

	Real positive	Real negative
Predicted positive	True-positive (TP)	False-positive (FP)
Predicted negative	False-negative (FN)	True-negative (TN)

AutoGluon is an open-source library with transparency and extensibility. Another advantage is that the AutoGluon framework uses a multi-layer stacking with k-fold bagging to reduce the model's variance. The number of layers and the value of k are heuristically determined within the framework. AutoGluon v0.2.0 is applied in this study with 14 base models, including random forest, XGBoost, and neural network. The models are listed in **Supplementary Table S2**.

2.4 Performance Indicators

For binary classification, the results can be evaluated using a confusion matrix (**Table 1**).

Various indicators could be constructed based on the confusion matrix to quantify the model performance: (1) precision measures how well the model can predict real positive labels; (2) recall measures the ability to classify true-positive and true-negative; (3) F1 score is the weighted average of precision and recall. These indicators are calculated through **Eqs 1–3**. The higher the values of these indicators, the better the model's performance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

3 RESULTS AND DISCUSSION

3.1 Baseline With FPocket

FPocket detects pockets on the surface of the selected proteins and sorts them in the descending order of pocket scores, which reflect the putative capacity of the pocket to bind a small molecule. The scoring function formula in FPocket is shown in the supporting information. As described in FPocket, a training dataset containing 307 proteins was first generated to determine the weights of the five features in calculating the pocket score. These proteins are filtered based on a previous study for the evaluation of PocketFinder (An et al. 2005), which is trained on 5,616 protein–ligand complexes, including 4,711 unique proteins and 2,175 unique ligands. As proposed, PocketFinder can be used to predict ligand-binding pockets and suggest new allosteric pockets, leveraging the allosteric site prediction power to FPocket.

We notice that many positive pockets have relatively high pocket scores. For 70.6% of the total 204 proteins used in our study, the top-ranked pocket among the pockets detected is positive in our labeling method. For 84.3% of proteins in the test set, the positive pockets are among the top three ranked positions. Among all the positive pockets, nearly 90% of them appear in the first eight positions (**Figure 1**).

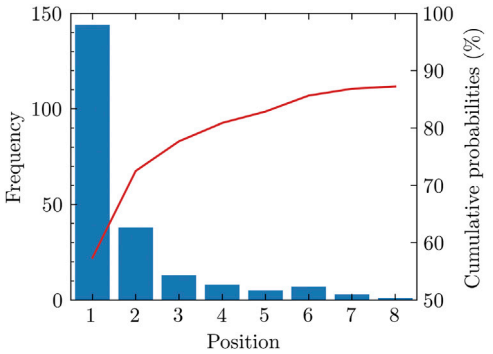


FIGURE 1 | Rank of positive pockets among all pockets. Nearly 90 percent of positive pockets appear among the first eight pockets sorted by the pocket score.

TABLE 2 | Confusion matrix of the baseline predictor.

	Real positive	Real negative
Predicted positive	144	60
Predicted negative	107	4844

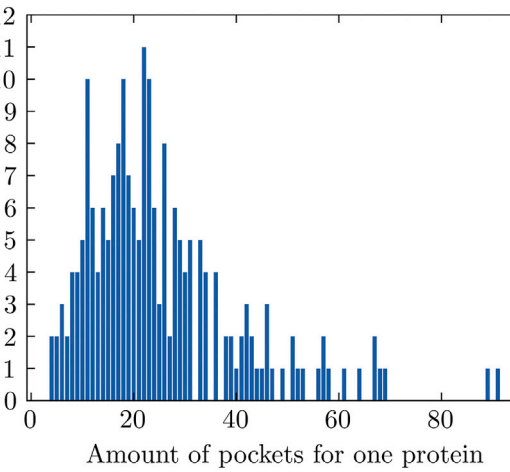


FIGURE 2 | Amounts of pockets for proteins. The amount varies from 4 to 91.

Here, we designed a baseline for allosteric site prediction: a predictor that predicts the pocket with the highest pocket score as positive, and others as negative. We applied this baseline model to the data and evaluated the performance. The confusion matrix is shown in **Table 2**. The accuracy, precision, recall, and F1 score values are 0.968, 0.706, 0.574, and 0.633, respectively.

A model could be evaluated as useful if it either has higher performance indicator values (classifying power) or higher top

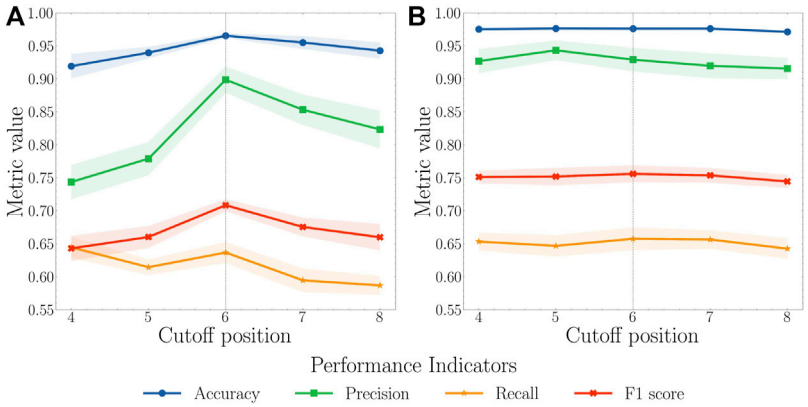


FIGURE 3 | (A) AutoKeras and **(B)** AutoGluon models performance for all pockets of the proteins in the validation set based on different cutoff positions. The cutoff value for the training set ranges from 4 to 8. Each model was trained in 10 independent runs for each value. The mean and standard deviation of each metric were calculated. A cutoff of 6 was considered reaching a balance between recall and precision with the highest F1 score.

three probabilities (ranking power) than this baseline predictor model.

3.2 Model Selection and Fine-Tuning on the Validation Set

The number of pockets that FPocket detects for each individual protein ranges between 4 and 91 for 204 proteins used in this study and has an average value of 25 (**Figure 2**). The pockets with positive labels only account for 4.87% (251 out of 5,155) in all pockets, making this dataset highly imbalanced. Data imbalance happens in a classification problem where the samples are not equally distributed among classes. This could lead to unsatisfactory model performance because the trained machine learning model might not learn sufficiently from the limited minority examples.

There are mainly two effective ways, over-sampling and under-sampling, to handle an imbalanced dataset (Lemaître et al. 2017). Over-sampling expands the size of the minority class by randomly duplicating existing examples or generating new but similar examples. However, this could result in overfitting for some machine learning models. Also, in the context of protein allosteric sites, the generated allosteric sites may not be biologically reasonable. Due to these reasons, under-sampling was applied to adjust the composition of the training data in the following procedure.

We first randomly split the selected 204 proteins into a training set with 122 proteins, a validation set with 41 proteins, and a test set with 41 proteins. To balance the training process, we only kept a certain number, referred to as the cutoff position, of top pockets based on their pocket scores generated by FPocket for each protein in the training set. For example, if the cutoff position is set to 5, only the first five pockets sorted by FPocket for proteins in the training set were used for the model training purpose. For cutoff positions from 4 to 8, both AutoKeras and AutoGluon models are trained and validated (**Figure 3**). The pocket descriptors generated by FPocket were used as features. Whether a pocket is allosteric or not according to ASD is represented as 1 for allosteric or 0 for nonallosteric. In the validation and test sets, a predicted value above 0.5 indicates an allosteric site, and a predicted value below 0.5 indicates a nonallosteric site.

TABLE 3 | Classifying power and ranking power of AutoGluon models on the test sets.

Indicator	Mean value	Top position	Mean value
Precision	0.850	Top 1	65.1%
Recall	0.616	Top 2	77.8%
F1 score	0.701	Top 3	82.7%

Based on AutoKeras and AutoGluon model performance using cutoff values ranging 4–8, the value of 6 leads to the balance between the precision and recall with the highest F1 score. When the cutoff is smaller than 6, the unsatisfactory performance might result from insufficient data for models to learn. When the cutoff is larger than 8, the performance starts to drop because of the unbalanced and low-quality data. Therefore, the cutoff value of 6 was selected to produce the final model.

In the final model, the mean values of accuracy, precision, recall, and F1 score for the AutoKeras model were calculated as 0.955, 0.853, 0.595, and 0.675, respectively. These values for the AutoGluon model are 0.976, 0.919, 0.656, and 0.754, respectively. The results show that the AutoGluon model has a better performance than the AutoKeras model and thus was selected for further test and final deployment.

3.3 Test Set Performance

The final AutoGluon model using the cutoff position as 6 was tested on the test set, where the model was used to evaluate all the detected pockets. The metric values shown are comparable to its performance using the validation set (**Table 3**), indicating the good prediction power of this model.

It is also expected that a powerful machine learning model is capable of ranking allosteric sites in the top positions. In the

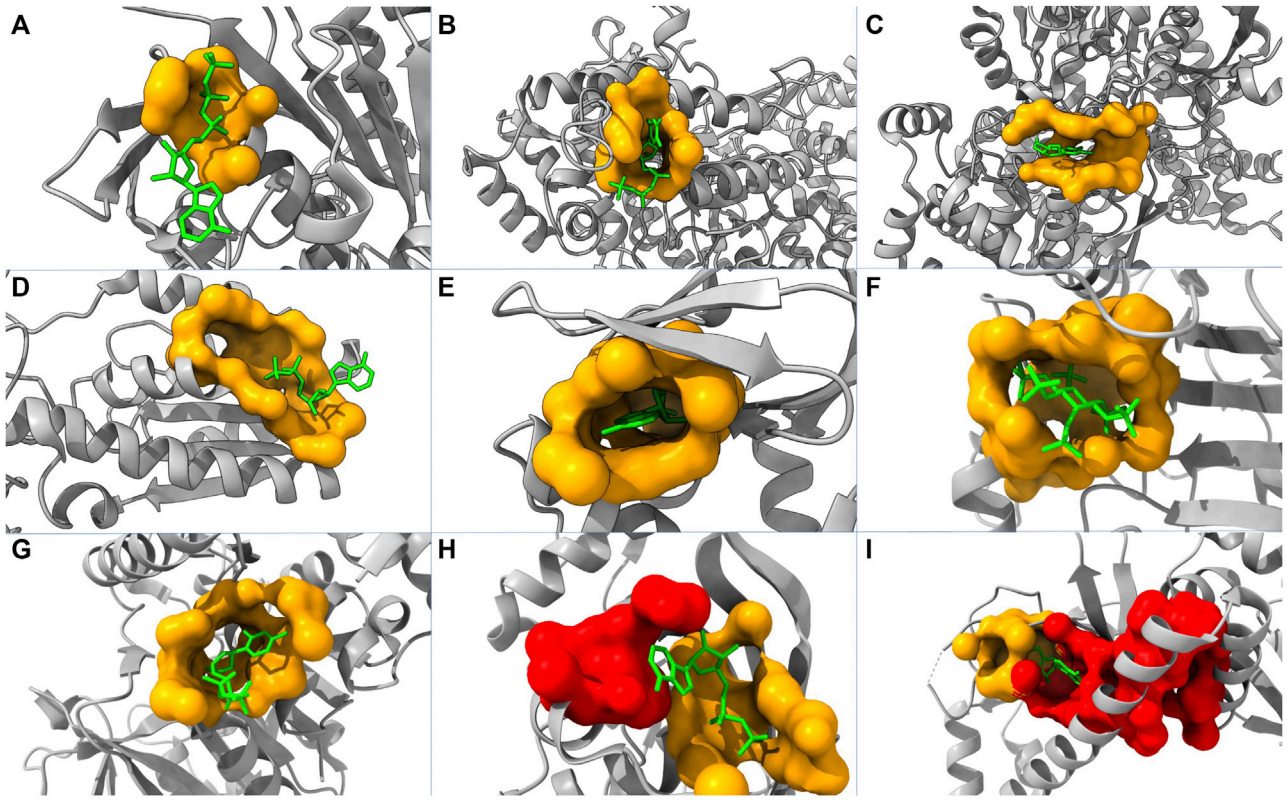


FIGURE 4 | Structures of nine proteins with modulators and predicted pockets. PDB IDs of these proteins are: **(A)** 2FPL, **(B)** 2R1R, **(C)** 3BCR, **(D)** 4PFK, **(E)** 1Q5O, **(F)** 3PEE, **(G)** 4HO6, **(H)** 1XMV, and **(I)** 2OZ6. The yellow pockets are labeled as allosteric, and the lime molecules are modulators. For **(A–G)**, the allosteric pockets are successfully predicted as top one by our model. For **(H,I)**, the red pockets are predicted as the first place, and the allosteric pockets are predicted as the second place.

current study, we evaluated the ranking power of our models by calculating the ranking probabilities of the allosteric sites at the top 1, 2, and 3 positions. The probabilities of allosteric sites, shown in **Table 3**, indicate that the final prediction model could rank the known allosteric sites among the top three positions for the majority of the test set. Taking the classifying power and ranking power together, our method has a great performance on allosteric site predictions.

3.4 Novel Protein Prediction

To further evaluate the performance of our model, we tested our model using 50 randomly picked proteins that are in the core set but not included in the core-diversity set in ASBench. Among these proteins, 22, 11, and 3 of their allosteric sites are ranked as first, second, and third, respectively. This leads to 72% of the additional test set with their true allosteric sites being ranked among the top three by our model. We also plot nine structures highlighting predicted allosteric sites and modulators (**Figure 4**). Our model successfully predicted allosteric sites as the top site for seven out of these proteins (**Figures 4A–G**), with the probabilities of 58.94%, 79.40%, 78.30%, 82.16%, 95.78%, 96.12%, and 85.11%, respectively. For protein in **Figure 4H**, the top pocket has a probability of 80.37%, and the real allosteric site is predicted at the second place with a probability of 77.20%. For protein in **Figure 4I**, the top pocket has a probability of 77.24%, and the real allosteric site is predicted at the second place with a probability of 51.09%.


```
{
    "1": {
        "prob": "95.33%",
        "residues": "chain A and resid 307 319 317
287 286 253 304 303 333 348 262 329 300 290 288 308
350 255 291 346 331 261"
    },
    "2": {
        "prob": "36.57%",
        "residues": "chain A and resid 328 326 353
355 321 327 318 352"
    },
    "3": {
        "prob": "22.74%",
        "residues": "chain A and resid 266 268 250
273 269 246 249 270 274 321 251 352"
    }
}
```


FIGURE 5 | Allosteric probability results of chain A of protein 5DKK returned by command line API of PASSer.

In some cases, the fallaciously predicted top one pockets are close to and even merge into the pocket labeled as allosteric (**Figures 4H,**

I). Consequently, it is not straightforward to determine whether the predicted top one pockets are false-positive. This complication of model interpretation could result from the data preprocessing (pocket detection and pocket labeling). In reality, two pockets might collectively act as one allosteric site in a biological process but being identified as two individual pockets in our model.

3.5 Web Server

The model has been integrated into the Protein Allosteric Site Server. The server can be either accessed at https://passer.smu.edu or through the command line. Here is an example using the command line to test the chain A of protein 5DKK using the AutoML model.

```
# !/bin/bash
curl -X POST \
-d pdb=5dkk -d chain=A -d model=autoML \
https://passer.smu.edu/api
```

This returns the top 3 pocket probabilities with residues in the json format, as shown in **Figure 5**, which can be easily parsed for further usage. Therefore, this provides a chance for large-scale searching applications for allosteric drug discovery.

4 CONCLUSION

Several machine learning-based methods have been developed for allosteric site prediction over the past few years. In this study, we applied an emerging ML technique, automated machine learning, to further improve the performance of protein allosteric site prediction models. The AutoML framework is capable of automating the machine learning model pipeline. The developed allosteric site prediction model, PASSer2.0, performs well under multiple indicators and is shown to have a good ranking power with a high percentage of ranking allosteric sites at top positions.

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

SX organized the data, performed the analysis, and wrote the first draft of the manuscript. HT performed the validation and reviewed and edited the manuscript. PT supervised the project and reviewed and edited the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

Research reported in this article was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R15GM122013.

ACKNOWLEDGMENTS

Computational time was generously provided by the Southern Methodist University's Center for Research Computing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.879251/full#supplementary-material

REFERENCES

Amor, B. R., Schaub, M. T., Yaliraki, S. N., and Barahona, M. (2016). Prediction of Allosteric Sites and Mediating Interactions through Bond-To-Bond Propensities. *Nat. Commun.* 7, 12477. doi:10.1038/ncomms12477

An, J., Totrov, M., and Abagyan, R. (2005). Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* 4, 752–761. doi:10.1074/mcp.m400159-mcp200

Bian, Y., Jing, Y., Wang, L., Ma, S., Jun, J. J., and Xie, X.-Q. (2019). Prediction of Orthosteric and Allosteric Regulations on Cannabinoid Receptors Using Supervised Machine Learning Classifiers. *Mol. Pharm.* 16, 2605–2615. doi:10.1021/acs.molpharmaceut.9b00182

Chen, A. S.-Y., Westwood, N. J., Brear, P., Rogers, G. W., Mavridis, L., and Mitchell, J. B. O. (2016). A Random Forest Model for Predicting Allosteric and Functional Sites on Proteins. *Mol. Inf.* 35, 125–135. doi:10.1002/minf.201500108

Chen, L., Lu, Y., Wu, C.-T., Clarke, R., Yu, G., Van Eyk, J. E., et al. (2021a). Data-driven Detection of Subtype-specific Differentially Expressed Genes. *Sci. Rep.* 11, 1–12. doi:10.1038/s41598-020-79704-1

Chen, T., and Guestrin, C. (2016). "Xgboost: A Scalable Tree Boosting System," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., et al. (2021b). Ilearnplus: a Comprehensive and Automated Machine-Learning Platform for Nucleic Acid and Protein Sequence Analysis, Prediction and Visualization. *Nucleic acids Res.* 49, e60. doi:10.1093/nar/gkab122

Elshawi, R., Maher, M., and Sakr, S. (2019). Automated Machine Learning: State-Of-The-Art and Open Challenges. *arXiv Prepr. arXiv:1906.02287*.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., et al. (2020). Autogluon-tabular: Robust and Accurate Automl for Structured Data. *arXiv Prepr. arXiv:2003.06505*.

Goncearenco, A., Mitternacht, S., Yong, T., Eisenhaber, B., Eisenhaber, F., and Berezovsky, I. N. (2013). Spacer: Server for Predicting Allosteric Communication and Effects of Regulation. *Nucleic acids Res.* 41, W266–W272. doi:10.1093/nar/gkt460

Greener, J. G., and Sternberg, M. J. (2015). Allopred: Prediction of Allosteric Pockets on Proteins Using Normal Mode Perturbation Analysis. *BMC Bioinforma.* 16, 335–337. doi:10.1186/s12859-015-0771-1

Gunasekaran, K., Ma, B., and Nussinov, R. (2004). Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins* 57, 433–443. doi:10.1002/prot.20232

Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., et al. (2013). Allosite: a Method for Predicting Allosteric Sites. *Bioinformatics* 29, 2357–2359. doi:10.1093/bioinformatics/btt399

Huang, W., Wang, G., Shen, Q., Liu, X., Lu, S., Geng, L., et al. (2015). ASBench: Benchmarking Sets for Allosteric Discovery: Fig. 1. *Bioinformatics* 31, 2598–2600. doi:10.1093/bioinformatics/btv169

Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., et al. (2011). Asd: a Comprehensive Database of Allosteric Proteins and Modulators. *Nucleic acids Res.* 39, D663–D669. doi:10.1093/nar/gkq1022

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. Berlin, Germany: Springer Nature.

Jin, H., Song, Q., and Hu, X. (2019). "Auto-keras: An Efficient Neural Architecture Search System," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 1946–1956. doi:10.1145/3292500.3330648

Karaglani, M., Gourlia, K., Tsamardinos, I., and Chatzaki, E. (2020). Accurate Blood-Based Diagnostic Biosignatures for Alzheimer's Disease via Automated Machine Learning. *J. Clin. Med.* 9, 3016. doi:10.3390/jcm9093016

Kipf, T. N., and Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks. *arXiv Prepr. arXiv:1609.02907*.

Laine, E., Goncalves, C., Karst, J. C., Lesnard, A., Rault, S., Tang, W.-J., et al. (2010). 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

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an Open Source Platform for Ligand Pocket Detection. *BMC Bioinforma.* 10, 168. doi:10.1186/1471-2105-10-168

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* 18, 559–563.

Liaw, A., and Wiener, M. (2002). Classification and Regression by Randomforest. *R. news* 2, 18–22.

Liu, X., Lu, S., Song, K., Shen, Q., Ni, D., Li, Q., et al. (2020). Unraveling Allosteric Landscapes of Allosterome with Asd. *Nucleic Acids Res.* 48, D394–D401. doi:10.1093/nar/gkz958

Lu, S., Huang, W., and Zhang, J. (2014). Recent Computational Advances in the Identification of Allosteric Sites in Proteins. *Drug Discov. today* 19, 1595–1600. doi:10.1016/j.drudis.2014.07.012

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

Nussinov, R., Tsai, C.-J., and Csermely, P. (2011). Allo-network Drugs: Harnessing Allostery in Cellular Networks. *Trends Pharmacol. Sci.* 32, 686–693. doi:10.1016/j.tips.2011.08.004

Panagopoulou, M., Karaglani, M., Manolopoulos, V. G., Iliopoulos, I., Tsamardinos, I., and Chatzaki, E. (2021). Deciphering the Methylation Landscape in Breast Cancer: Diagnostic and Prognostic Biosignatures through Automated Machine Learning. *Cancers* 13, 1677. doi:10.3390/cancers13071677

Panjkovich, A., and Daura, X. (2012). Exploiting Protein Flexibility to Predict the Location of Allosteric Sites. *BMC Bioinforma.* 13, 273. doi:10.1186/1471-2105-13-273

Panjkovich, 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

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

Srinivasan, B., Forouhar, F., Shukla, A., Sampangi, C., Kulkarni, S., Abashidze, M., et al. (2014). Allosteric Regulation and Substrate Activation in Cytosolic Nucleotidase II fromLegionella Pneumophila. *Febs J.* 281, 1613–1628. doi:10.1111/febs.12727

Suykens, J. A. K., and Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* 9, 293–300. doi:10.1023/a:1018628609742

Tian, H., Jiang, X., Trozzi, F., Xiao, S., Larson, E. C., and Tao, P. (2021b). Explore Protein Conformational Space with Variational Autoencoder. *Front. Mol. Biosci.* 8, 781635. doi:10.3389/fmolb.2021.781635

Tian, H., Jiang, X., and Tao, P. (2021a). Passer: Prediction of Allosteric Sites Server. *Mach. Learn. Sci. Technol.* 2, 035015. doi:10.1088/2632-2153/abe6d6

Tian, H., Jiang, X., Xiao, S., La Force, H., Larson, E. C., and Tao, P. (2022). Last: Latent Space Assisted Adaptive Sampling for Protein Trajectories. *arXiv Prepr. arXiv:2204.13040*.

Tian, H., Trozzi, F., Zoltowski, B. D., and Tao, P. (2020). Deciphering the Allosteric Process of the Phaeodactylum Tricornutum Aureochrome 1a Lov Domain. *J. Phys. Chem. B* 124, 8960–8972. doi:10.1021/acs.jpcb.0c05842

Wagner, J. R., Lee, C. T., Durrant, J. D., Malmstrom, R. D., Feher, V. A., and Amaro, R. E. (2016). Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem. Rev.* 116, 6370–6390. doi:10.1021/acs.chemrev.5b00631

Waring, J., Lindvall, C., and Umeton, R. (2020). Automated Machine Learning: Review of the State-Of-The-Art and Opportunities for Healthcare. *Artif. Intell. Med.* 104, 101822. doi:10.1016/j.artmed.2020.101822

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., et al. (2018). Taking Human Out of Learning Applications: A Survey on Automated Machine Learning. *arXiv Prepr. arXiv:1810.13306*.

Zha, J., Li, M., Kong, R., Lu, S., and Zhang, J. (2022). Explaining and Predicting Allostery with Allosteric Database and Modern Analytical Techniques. *J. Mol. Biol.* 2022, 167481. doi:10.1016/j.jmb.2022.167481

Zhang, Q., Heldermon, C. D., and Toler-Franklin, C. (2020). "Multiscale Detection of Cancerous Tissue in High Resolution Slide Scans," in International Symposium on Visual Computing (Berlin, Germany: Springer), 139–153. doi:10.1007/978-3-030-64559-5_11

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.

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fmolb.2022.953984

OPEN ACCESS

EDITED BY
J. Andrew McCammon,
University of California, San Diego,
United States

REVIEWED BY
Mithun Radhakrishna,
Indian Institute of Technology
Gandhinagar, India
Qin Xu,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Vineetha Menon,
vineetha.menon@uah.edu
Jerome Baudry,
jerome.baudry@uah.edu

SPECIALTY SECTION
This article was submitted to Molecular Recognition,
a section of the journal
Frontiers in Molecular Biosciences

RECEIVED 26 May 2022
ACCEPTED 16 December 2022
PUBLISHED 12 January 2023

CITATION
Gupta S, Baudry J and Menon V (2023), Big Data analytics for improved prediction of ligand binding and conformational selection.
Front. Mol. Biosci. 9:953984.
doi: 10.3389/fmolb.2022.953984

Big Data analytics for improved prediction of ligand binding and conformational selection

Shivangi Gupta[1], Jerome Baudry[2]* and Vineetha Menon[1]*

[1]Department of Computer Science, The University of Alabama in Huntsville, Huntsville, AL, United States, [2]Department of Biological Sciences, The University of Alabama in Huntsville, Huntsville, AL, United States

This research introduces new machine learning and deep learning approaches, collectively referred to as Big Data analytics techniques that are unique to address the protein conformational selection mechanism for protein:ligands complexes. The novel Big Data analytics techniques presented in this work enables efficient data processing of a large number of protein:ligand complexes, and provides better identification of specific protein properties that are responsible for a high probability of correct prediction of protein:ligand binding. The GPCR proteins ADORA2A (Adenosine A2a Receptor), ADRB2 (Adrenoceptor Beta 2), OPRD1 (Opioid receptor Delta 1) and OPRK1 (Opioid Receptor Kappa 1) are examined in this study using Big Data analytics techniques, which can efficiently process a huge ensemble of protein conformations, and significantly enhance the prediction of binding protein conformation (i.e., the protein conformations that will be selected by the ligands for binding) about 10–38 times better than its random selection counterpart for protein conformation selection. In addition to providing a Big Data approach to the conformational selection mechanism, this also opens the door to the systematic identification of such "binding conformations" for proteins. The physico-chemical features that are useful in predicting the "binding conformations" are largely, but not entirely, shared among the test proteins, indicating that the biophysical properties that drive the conformation selection mechanism may, to an extent, be protein-specific for the protein properties used in this work.

KEYWORDS

protein conformation selection, Big Data, deep learning, machine learning, feature selection, drug discovery

1 Introduction

The prediction of which small molecules, e.g., substrates or modulators, are more likely than other small molecules to bind to a specific protein, is one the most formidable challenges of contemporary biology, chemical biology and pharmacology. Only a small fraction of the large number of small organic molecules present in living organisms will, in most cases, bind to a specific protein. There is a considerable amount of work that aims at improving the biophysical approaches to predicting such protein:ligand interactions.

As exemplified in the current special issue of Frontiers, the dynamics of the protein target is increasingly taken into account in such predictive approaches. Indeed, a protein cycles through multiple conformations, a few of which will be bound by its ligands, as conceptualized in the "conformational selection" mechanism of ligand binding. Virtual docking (Amaro et al., 2018) that aims at predicting if a given small chemical binds to a given protein, usually considers only one protein conformation in an "induced fit" mechanism. Advances beyond a simple induced fit mechanisms have been proposed, such as submitting the protein:ligands complexes to molecular dynamics simulations after docking (Seelinger and de Groot, 2010), which identifies binding modes of known ligands close to that of their experimental co-crystallized structures, or generating an ensemble of holo structures from experimental structures deposited in the PDB for a given protein target (Aggarwal et al., 2021). This present work, continuing in that direction, aims at using the information contained in molecular dynamics simulations of a single protein target structure prior to any docking.

In principle the "binding" protein conformations will correspond to the free energy minima of the (protein + ligand) complex free energy hypersurface. In our research, we are looking into whether we can identify these rare apo-conformations that possess this capacity to bind their ligands, while the vast majority of the other apo-protein conformations do not. This paper describes our Big Data analytics work toward such characterization of what properties of an apo-protein conformation more likely lead to conformational selection.

The data we used here has been obtained using supermassive "ensemble docking" from proteins' molecular dynamics simulations, and is described in (Evangelista et al., 2016). The data corresponds to about 1.5 millions of protein conformation and protein:ligand complex structures and their associated docking scores.

Big Data analytics provides an efficient approach to analyzing such a large amount of data, and also addresses the class imbalance problem (Abd Elrahman and Abraham, 2013), which is a result of imbalanced groups or sub-categories present in the data, where the majority class or larger group of data consists of non-binding protein conformations and it overshadows the minority class or smaller data group, which comprises the data-of-interest i.e., the binding protein conformations. In our prior work (Akondi et al., 2019; Gupta et al., 2022; Sripriya Akondi et al., 2022), a novel two-stage sampling-based classifier framework was proposed with the primary goal of addressing the class imbalance problem and maximizing the detection of potential binding protein conformations as conventional machine learning (ML) algorithms are ill-equipped to deal with the issue of class imbalance during the data-learning phase. This paper extends on our previous work by presenting additional improvements to our two-stage sampling-based classification approach (Gupta et al., 2022) using deep learning techniques and four different feature selection methods in conjunction with an Enrichment ratio framework.

2 Materials and methods

2.1 Dataset description

As described in our previous work (Gupta et al., 2022), Molecular Dynamics (MD) simulations of four proteins, namely, ADORA2A (Adenosine A2a Receptor), ADRB2 (Adrenoceptor Beta 2), OPRD1 (Delta Opioid Receptor) and OPRK1 (Opioid Receptor Kappa 1) were used to study the efficacy of our proposed method. The conformations of these four proteins have been well-studied, and the protein conformations that: a) will bind to ligands (binding conformations) and b) will not bind to ligands (non-binding conformations), are known and have been previously documented and published (Evangelista et al., 2016).

ADORA2A: This dataset has 50 attributes and consists of 2,998 protein conformations among which 851 protein conformations are "binding" and 2,147 protein conformations that are "non-binding". Here the imbalance ratio is 3:1 i.e., for every datasample belonging to minority class (binding conformations) there are three data samples belonging to the majority class (non-binding conformations).

ADRB2: This dataset has 51 attributes and consists of 2,565 protein conformations among which 156 are binding and 2,411 protein conformations are non-binding. Here the imbalance ratio is 16:1 i.e., for every datasample belonging to minority class (binding conformations) there are 16 data samples belonging to the majority class (non-binding conformations).

OPRD1: This dataset has 51 attributes and consists of 3,004 protein conformations among which 72 protein conformations are binding and 2,932 protein conformations are non-binding. Here the imbalance ratio is 41:1 i.e., for every datasample belonging to minority class (binding conformations) there are 41 data samples belonging to the majority class (non-binding conformations).

OPRK1: This dataset has 50 attributes and consists of 2,998 protein conformations among which 138 protein conformations are binding and 2,862 protein conformations are non-binding. Here the imbalance ratio is 20:1 i.e., for every data sample belonging to minority class (binding conformations) there are 20 data samples belonging to the majority class (non-binding conformations).

Tables describing the protein attributes/features/descriptors for ADORA2A, ADRB2, OPRD1, and OPRK1 datasets can be found in our previous work (Gupta et al., 2022). ADRB2 and OPRD1 have one additional feature - pro_pl_seq (Sequence based pI) in comparison to ADORA2A and OPRK1. The molecular descriptors were calculated using the protein

descriptors from the program MOE (Akondi et al., 2019; Chemical Computing Group, 2019; Gupta et al., 2022; Sripriya Akondi et al., 2022).

2.1.1 Analysis of variance

Analysis of variance (ANOVA) is a statistical analysis method used here to calculate the linear relationship between the various protein features and to select the important protein features that correspond to the highest F-values (Johnson and Synovec, 2002). The top "x" features with the greatest F-values were selected in this case, where the x features to be retained is determined experimentally by the user. Thus, ANOVA technique allows for selection of the primary physio-chemical protein properties that essay a critical role in protein:ligand interaction and conformation selection.

2.1.2 Mutual information

Mutual Information (MI) (Macedo et al., 2019) is a measure of the amount of information that can be inferred about a variable **U** through the use of the other given random variable **V**. The mutual information **I** (**U**; **V**) for random variables **U** and **V** can be defined as follows (Guyon and Elisseeff, 2003; Gupta et al., 2022):

$$\mathbf{I}(\mathbf{U};\mathbf{V}) = -\sum_{v\in V}\sum_{u\in U} p(u,v)log\frac{p(u,v)}{p(u)\,p(v)} \tag{1}$$

where

- p(u,v) is the joint probability density function.
- p(u) is the probability density function

In Eq. 1, if the MI value **I** is 1, then **U** and **V** are dependent on each other, i.e., protein features share similar information. If the MI value **I** is 0, then **U** and **V** are independent of each other i.e., no common (in other words unique) information between the features. The MI in physio-chemical properties are calculated as follows:

- First, calculate the MI value for all properties to determine how dependent the physio-chemical features vectors are and understand the common information contained in all the protein features.
- Then, sort the protein features according to their highest MI values. The top "x" protein features with the greatest MI values are retained, where x is user defined.

2.1.3 Recurrence quantification analysis

Recurrence Quantification Analysis (RQA) is a non-linear data-analysis method that is used to study the dynamical systems (Eckmann et al., 1987). The first step in the recurrence analysis is to quantify the repeating patterns of a dynamic system. One of the variables generated by the quantification of the recurrences is Entropy (ENT), which is the probability distribution p(j) of the diagonal line on the RQA plot and is defined as:

$$ENTR = -\sum_{j=j_{min}}^{M} p(j)\,ln(p(j)) \tag{2}$$

where M is the number of points on the state space trajectory and j is the length of the diagonal line in the RQA plot. We investigate the RQA-based entropy measure's link to the probability of detecting potential binding conformations in terms of time-space evolution of protein conformations.

2.1.4 Spearman correlation coefficient

Spearman correlation coefficient is a statistical measure (Hauke and Kossowski, 2011) of the strength and direction of the monotonic relationship between each protein feature and target variable. The correlation coefficient for each feature is obtained by applying the formula as defined below:

$$\rho = \frac{\Sigma_i\,(u_i-\bar{u})\,(v_i-\bar{v})}{\sqrt{\Sigma_i\,(u_i-\bar{u})^2\,(v_i-\bar{v})^2}} \tag{3}$$

where u is the feature vector and $\bar{u}$ is its corresponding mean. Similarly, v is the target vector and $\bar{v}$ is the mean of the target vector. The Spearman correlation coefficients for protein features are computed, sorted and ranked based on the absolute value of the correlation coefficient. A subset of the protein features were then selected based on the "x" highest rankings, where x is user-defined. Therefore, the Spearman correlation coefficient allows us to select protein features that are strongly correlated with each other.

2.1.5 Extreme gradient boosting

Extreme Gradient Boosting (XGBoost) is a tree ensemble boosting approach that merges a number of weak classifiers into a single strong classifier (Chen and Guestrin, 2016). Starting with a base learner, the strong learner is trained iteratively for best classification or prediction performance. Given a dataset X with m samples and n protein descriptors, let $(x_1, y_1), ..., (x_k, y_k)$ be a set of inputs x_i and corresponding outputs y_i (Babajide Mustapha and Saeed, 2016). The XGBoost algorithm uses "K" additive functions, each representing a classification and regression tree (CART) to predict the output label $\hat{y}_i$ as defined by:

$$\hat{y}_i = \sum_{k=1}^{K} t_k(x_i), \quad t_k \in T \tag{4}$$

where t_k corresponds to a distinct tree structure with leaf score "w" and T is the space of all classification and regression trees. The goal is to minimize the following regularized objective function (Babajide Mustapha and Saeed, 2016):

$$Obj(\Theta) = \sum_i^m l(y_i,\ \hat{y}_i) + \sum_k^K \Psi(t_k) \tag{5}$$

where l is the loss function that is used to measure the difference between the predicted value $\hat{y}_i$ and the actual value y_i and Ψ is the regularization term that is used to avoid overfitting and is defined as:

$$\Psi(t_k) = \gamma D + \frac{1}{2}\lambda\|w\|^2 \quad (6)$$

where D is the number of leaves, w is the weight of each leaf, γ and λ are constants to control the degree of regularization.

2.1.6 K-Means clustering

K-Means clustering is an unsupervised machine learning algorithm (Oyelade et al., 2010) that is used to understand the data patterns in the input data by grouping the instances in the dataset that are similar into different clusters. K-Means clustering is often used to produce compact clusters with minimum intra-cluster distances and maximum inter-cluster distances (Oyelade et al., 2010). This goal is achieved by splitting the data into a number of clusters "k" that the user specifies (Wilkin and Huang, 2007). Here we employ the K-Means clustering algorithm to under sample the data points from the majority class samples i.e., non-binding protein conformations as demonstrated in our prior work (Akondi et al., 2019).

2.1.7 Generative adversarial networks

Generative adversarial networks (GAN) is an unsupervised learning method that involves learning regularities or patterns in the input data to produce new examples that mimic the original dataset. The GAN technique uses two artificial models, the discriminator and generator, which compete for data learning (Jo and Kim, 2022). The discriminator focuses on discriminating or distinguishing between the original and synthetic data, whereas the generator tries to create synthetic data that is comparable to the real data. The loss function of GAN (Jo and Kim, 2022) is defined as:

$$\min_D \max_F V(F,D) = \mathbb{Q}_{u\sim p_u}[\log\ (F(u)] - \mathbb{Q}_{v\sim p_v}[\log(1-(F(D(v)))] \quad (7)$$

where.

- p_u is the data-generating distribution
- p_v is the noise distribution
- u is the real input data
- v is the noise input to the generator neural network
- F(u) is the output probability of the generator
- D(v) is the sample generated by the generator neural network

Here the GAN is used to oversample or replicate the minority class in the dataset to alleviate the class imbalance problem and in turn maximize the prediction of the potential binding protein conformations.

2.1.8 Convolutional neural networks

Convolutional neural network (CNN) is a supervised deep learning technique (Hossain and Sajib, 2019) that has emerged as the most widely used artificial neural network in many computer vision applications, including texture recognition (Cimpoi et al., 2016), remote sensing scene classification (Hu et al., 2015; Penatti et al., 2015) and structure-based protein analysis (Torng and Altman, 2017). Architectural design of a CNN consists of several convolutional, pooling and dropout layers followed by one or more fully-connected layers (FC) (Sultana et al., 2018). Figure 1 describes the architecture of the CNN used in our work.

The architectural design of the CNN in our work consists of a convolutional layer followed by dropout to reduce overfitting, a max pooling layer, a fully connected layer and an output layer. Rectified linear unit (ReLU) is used as the activation function for the convolution layer and fully connected layer. Binary cross-entropy L is used as the loss function for the CNN.

2.1.9 Recurrent neural networks

Recurrent neural network (RNN) is a class of neural networks which is used to detect patterns in a sequence of data (Ho and Wookey, 2020). In our work, the RNN architecture consists of two long-short term memory (LSTM) (Schmidt, 2019) layers with dropout followed by a dense layer with dropout and an output layer. The LSTM unit introduces a gate mechanism to select whether to retain or discard specific information in the existing memory. If the LSTM unit recognizes a pivotal protein descriptor from an input sequence early on, then it captures any potential long-distance dependencies between the protein descriptor and target value. Figure 2 describes the architecture of the RNN used in our work. Rectified linear unit (ReLU) is used as the activation function for the LSTM unit and dense layer, sigmoid function is used as the activation function in the output layer and binary cross-entropy as the loss function.

2.1.10 Evaluation metrics

The confusion matrix and its derived evaluation parameters such as classification accuracy, sensitivity, specificity, etc., are some of the most commonly used ML evaluation metrics to validate a classification or prediction performance of ML algorithms. In this case of binary classification between binding and non-binding protein conformations, the confusion matrix has four categories of classification results as follows:

- True Positive (TP): When the classifier accurately predicts "binding," indicating that the ligand and target protein did bind (Right predictions of class 1)
- True Negative (TN): When the classifier accurately predicts "non-binding," indicating that the ligand and target protein did not bind (Right predictions of class 0)
- False Negative (FN): When the classifier inaccurately predicted "non-binding," but the ligand and target protein did bind (Wrong predictions of class 0)
- False Positive (FP): When the classifier inaccurately predicted "binding," but the ligand and target protein did not bind (Wrong predictions of class 1)

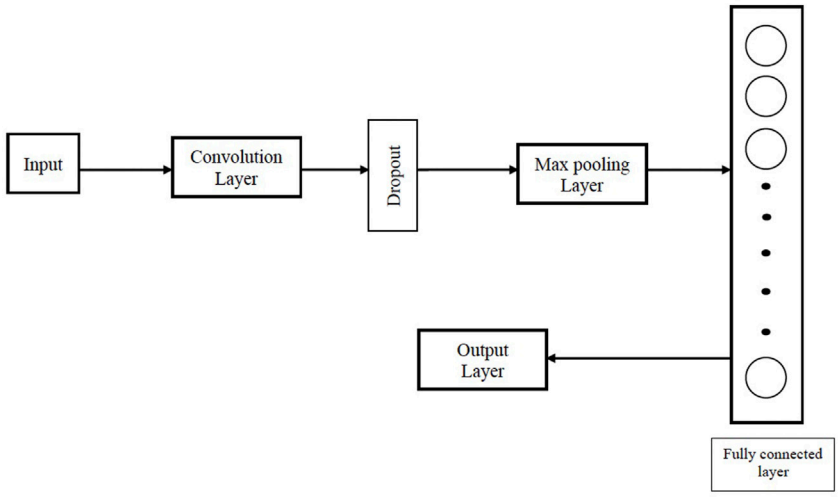


FIGURE 1
Architecture of the CNN used in our proposed Big Data analytics based AI/ML protein conformation selection/prediction framework.

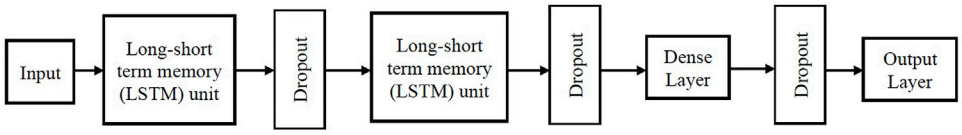


FIGURE 2
Architecture of the RNN used in our proposed Big Data analytics based AI/ML protein conformation selection/prediction framework.

Here class 0 refers to the non-binding protein conformations (majority class) and class 1 denotes the binding protein conformations (minority class).

Accuracy of an AI/ML framework is calculated as the sum of correctly predicted binding and non-binding protein conformations divided by the total number of conformations in the data set. It is defined as:

$$Accuracy = \frac{TP + FN}{TP + FP + FN + TN} \tag{8}$$

Sensitivity is the ability of the AI/ML framework to correctly predict binding protein conformations. It is calculated as the number of correctly predicted binding protein conformations divided by the total number of binding protein conformations in the data set as defined below:

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

Eqs 8, 9 are used for performance evaluation of the proposed AI/ML protein conformation selection/prediction framework.

2.1.11 Enrichment ratio framework

The enrichment was calculated using the TP and FN predictions from the Big Data analytics based AI/ML protein conformation selection/prediction framework, described in Section 2.2. The base enrichment ratio is calculated to measure the effectiveness of general predictive performance in the absence of the ML protein conformation selection framework as in our prior work (Gupta et al., 2022). For accurate base enrichment ratio we performed subset data selection on previously calculated and published anticipated protein:ligand interactions energies in (Evangelista et al., 2016). The assumption is that the computed protein:ligand interaction energies are quantitatively valid, i.e., a “preferred” binding conformation would be the one in which the protein binds the ligand stronger (i.e., with lower interaction energies) than other alternative conformations. Thus, the base enrichment was calculated from (Evangelista et al., 2016) by dividing the number of binding conformations by the total number of conformations. Eq. 10 calculates the base enrichment detected during the test phase if the ML algorithm is not implemented. We select different subsets of the TP and FN

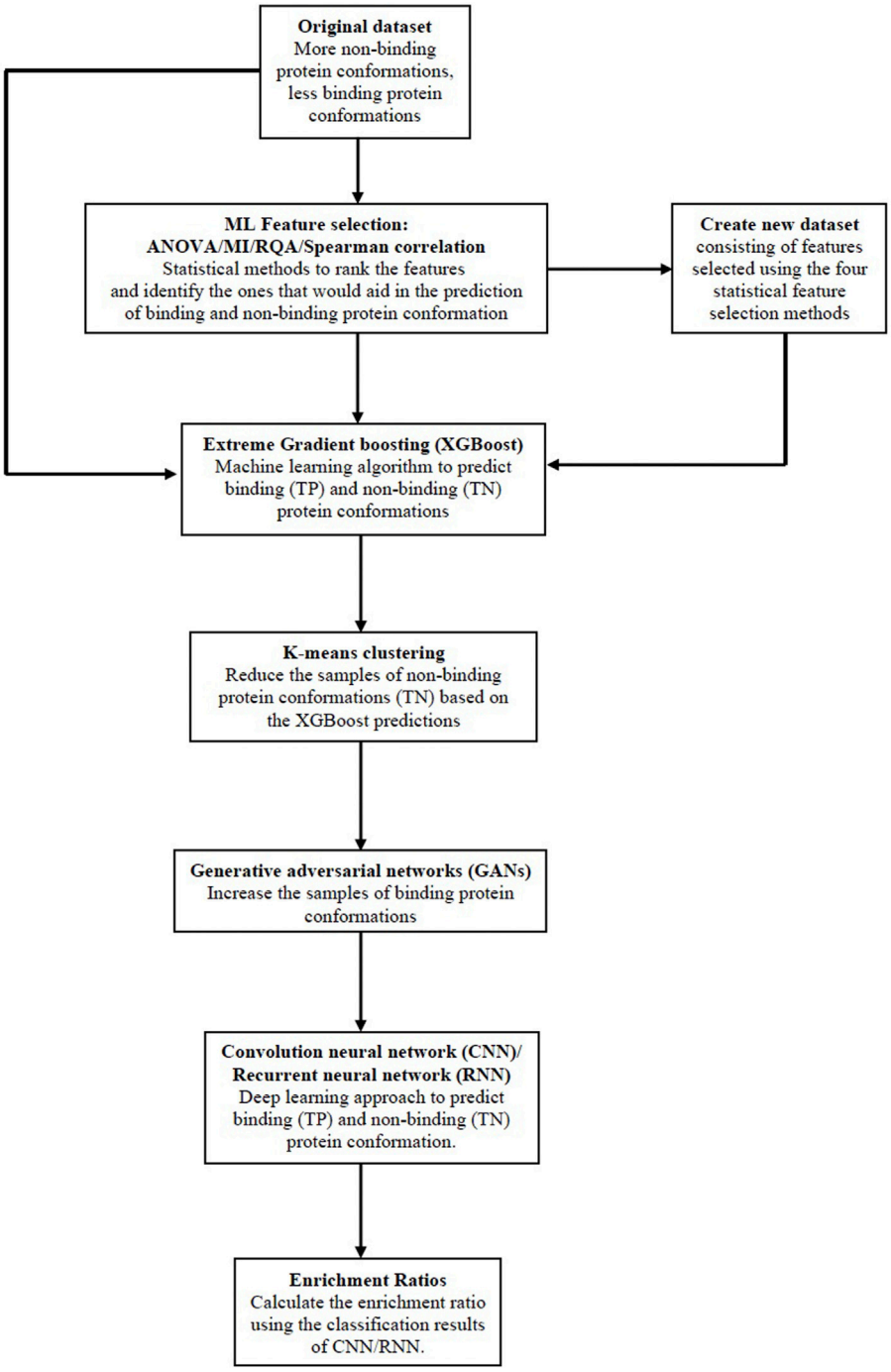


FIGURE 3
The proposed Big Data analytics based AI/ML protein conformation selection/prediction framework.

values in order to calculate the ML prediction framework enrichment ratios in Eq. 10. The values returned by both Eqs 10, 11 were then used to calculate the final enrichment ratio returned by each of the four filters (A,B,C,D) defined in Eq. 12.

$$\text{Base enrichment ratio} = \frac{number\ of\ binding\ conformations}{total\ number\ of\ conformations\ (binding\ and\ non-binding)} \quad (10)$$

$$\text{ML enrichment ratio} = \frac{number\ of\ binding\ conformations\ (TP)\ identified}{number\ of\ total\ conformations\ (TP\ and\ FN)\ identified} \quad (11)$$

$$\text{Final enrichment ratio} = \frac{ML\ enrichment\ ratio}{Base\ enrichment\ ratio} \quad (12)$$

The final enrichment ratios for proteins ADORA2A, ADRB2, OPRD1, and OPRK1 were calculated using four different filters (A,B,C,D) and have been described and published in our previous work (Gupta et al., 2022). The proposed enrichment ratio framework used is depicted in Supplementary Figure SI-1 (Gupta et al., 2022).

2.2 The proposed Big Data analytics based AI/ML protein conformation selection/ prediction framework

In this work, we combine the feature selection techniques discussed in Section 2.1 with the improved two-stage sampling based classification approach (Gupta et al., 2022) using deep learning techniques. The steps given below describe the new improved methodology and is illustrated in Figure 3:

- The first step in the methodology is to input the dataset and then apply the ML feature selection methods: i) Analysis of variance (ANOVA), ii) Mutual Information (MI), iii) Recurrence Quantification Analysis (RQA), and iv) Spearman correlation to select the important protein features from each of the methods respectively.
- We then obtain a feature ranking score for all features based on the common consensus of all the feature selection methods. Only the subset of protein features that are selected by all four feature selection methods are chosen to create a new dataset.
- Both the original dataset and a new dataset that is more biased towards samples in class 0 are sent as inputs to the XGBoost classifier. Samples of class 0 (TN) and class 1 (TP) are recorded as classification results 1.
- In order to create a new training dataset, the GAN algorithm was applied to both the original dataset and the new modified dataset.
- K-Means clustering (Akondi et al., 2019) is used on the XGBoost classifier's classification results, class 0 samples are undersampled, and the intended class 1 samples are oversampled. This step increases the detection rate of class 1 samples or binding protein conformations to address the class imbalance issue. The new training dataset has the same size as the initial training dataset in order to maintain consistency.
- Supervised classification using deep learning methodologies: CNN and RNN are applied to the newly created training dataset. Both classifiers are used to identify the binding and non-binding conformations in the new training dataset. The results of both classifiers are recorded.
- As a final step, the TP (binding conformations), and FN (binding conformations but are incorrectly predicted as non-binding conformations) by the AI/ML protein conformation prediction framework (CNN and RNN) are employed in the Enrichment ratio framework to calculate the Enrichment ratios. The outcomes of the framework for enrichment ratios are recorded.

3 Results

The overview of enrichment ratios for ADORA2A that were determined using the predicted binding conformations from the AI/ML framework is shown in Table 1. As indicated in Supplementary Table SI-1 through Supplementary Table SI-5, the AI/ML framework was evaluated on the remaining 70% of the dataset after being trained on 30% of it. It can be observed that the data selection filter A of the Enrichment ratio framework gave the maximum enrichment ratio of 7.1 using XGBoost + GANs–RNN framework predictions.

The list of protein descriptors that the four ML feature selection techniques determined to be significant is shown in Table 2 and it can be observed that 11 of the 50 features were chosen. Table 3 gives the overview of the enrichment ratios that were calculated using the features listed in Table 2. It can be observed that data selection filter A of Enrichment ratio framework gave the maximum enrichment ratio of 10.2 using XGBoost + GANs–CNN framework predictions.

The three common protein descriptors for the proteins ADORA2A, OPRK1, and OPRD1 that were determined to be significant by the four ML feature selection methods are listed in Supplementary Table SI-6. A summary of the enrichment ratios that were estimated using the characteristics indicated in Supplementary Table SI-6 is provided in Supplementary Table SI-7. It can be observed that employing data selection filter A, the XGBoost + GANs–CNN framework predictions provided the maximum enrichment ratio of 8.2.

The overview of enrichment ratios for the ADRB2 binding conformations predicted by the AI/ML framework is shown in Table 4. On 30% of the dataset, the AI/ML framework was trained, and on the remaining 70%, it was tested. It can be observed that employing data selection filter C, the XGBoost + GANs–RNN framework predictions provided the maximum enrichment ratio of 13.8.

The list of protein descriptors that the three out of four ML feature selection techniques determined to be significant is shown in Table 5. It can be seen from the table that 8 of the 51 features were chosen. Table 6 provides an overview of the enrichment ratios that were computed using the features listed in Table 5. It can be observed that employing data selection filter D, the XGBoost + GANs–RNN framework predictions provided the maximum enrichment ratio of 24.2.

TABLE 1 Enrichment Ratios of ADORA2A on the original dataset with no feature selection with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	5.6	Filter A	0.5	4.1	Filter C	1.0
XGboost + GANs–RNN	7.1	Filter A	1.0	5.9	Filter C	0.5

TABLE 2 11 features out of 50 were selected having a feature score of four using the feature scoring table for ADORA2A.

pro_asa_vdw	pro_dipole_moment	pro_patch_ion_n	pro_patch_neg_n
pro_asa_hyd	pro_hyd_moment	pro_app_charge	pro_zquadrupole
pro_volume	pro_patch_ion	pro_helicity	

TABLE 3 Enrichment Ratios of ADORA2A on the dataset consisting of features as shown in Table 2 with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	10.2	Filter A	0.5	8.1	Filter C	10.0
XGboost + GANs–RNN	9.0	Filter B	0.5	6.5	Filter D	1.0

TABLE 4 Enrichment Ratios of ADRB2 on the original dataset with no feature selection with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	9.4	Filter C	10.0	6.7	Filter B	0.5
XGboost + GANs–RNN	13.8	Filter C	1.0	7.6	Filter A	1.0

TABLE 5 8 features out of 51 were selected having a feature score of three using the feature scoring table for ADRB2.

pro_dipole_moment	pro_patch_hyd_5	pro_patch_pos_2
pro_patch_hyd	pro_patch_neg	pro_patch_hyd_1
pro_patch_hyd_4	pro_patch_neg_1	

TABLE 6 Enrichment Ratios of ADRB2 on the dataset consisting of features as shown in Table 5 with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	18.1	Filter D	10.0	8.4	Filter A	0.5
XGboost + GANs–RNN	24.2	Filter D	10.0	13.5	Filter B	1.0

The summary of enrichment ratios for OPRD1 that were determined using the predicted binding conformations from the AI/ML framework is shown in Table 7. On 30% of the dataset, the AI/ML framework was trained, and on the remaining 70%, it was tested. It can be observed that utilizing data selection filter B, the XGBoost + GANs–RNN

TABLE 7 Enrichment Ratios of OPRD1 on the original dataset with no feature selection with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	12.5	Filter B	0.5	4.9	Filter C	10.0
XGboost + GANs–RNN	37.5	Filter B	0.5	27.6	Filter D	5.0

TABLE 8 12 features out of 51 were selected having a feature score of four using the feature scoring table for OPRD1.

pro_asa_vdw	pro_hyd_moment	pro_patch_hyd_5	pro_patch_pos
pro_asa_hyd	pro_patch_hyd	pro_patch_neg	pro_net_charge
pro_asa_hph	pro_patch_hyd_4	pro_patch_neg_5	pro_app_charge

TABLE 9 Enrichment Ratios of OPRD1 on the dataset consisting of features as shown in Table 8 with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	16.5	Filter B	1.0	11.2	Filter A	0.5
XGboost + GANs–RNN	37.5	Filter B	0.5	25.7	Filter D	0.5

TABLE 10 Enrichment Ratios of OPRK1 on the original dataset with no feature selection with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	11.0	Filter A	10.0	6.9	Filter B	0.5
XGboost + GANs–RNN	27.6	Filter A	1.0	21.0	Filter D	0.5

framework predictions produced an enrichment ratio of up to 37.

The list of protein descriptors that the four ML feature selection techniques determined to be significant is shown in Table 8. It can be seen that 12 of the 51 features were chosen, and Table 9 provides an overview of the enrichment ratios that were computed using the features listed in Table 8. It can be observed that employing data selection filter B, the XGBoost + GANs–RNN framework predictions provided the maximum enrichment ratio of 37.5.

Supplementary Table SI-8 gives the overview of enrichment ratios that were calculated using the features that were listed in Supplementary Table SI-6. It can be seen that both XGBoost + GANs–CNN and XGBoost + GANs–RNN framework predictions gave the same enrichment ratio of 37.5, using data selection filter B.

A summary of the enrichment ratios for OPRK1 that were determined using the predicted binding conformations from the AI/ML framework is shown in Table 10. On 30% of the dataset, the AI/ML framework was trained, and on the remaining 70%, it was tested. It can be seen that employing data selection filter A, the XGBoost + GANs–RNN framework predictions provided the maximum enrichment ratio of 27.6.

TABLE 11 5 features out of 50 were selected having a feature score of four using the feature scoring table for OPRK1.

pro_asa_vdw	pro_hyd_moment	pro_patch_neg_1
pro_asa_hyd	pro_patch_hyd_5	

The list of protein descriptors that the four ML feature selection techniques determined to be significant is shown in Table 11. It can be seen that 5 of the 50 features were chosen, and Table 12 provides an overview of the enrichment ratios that were computed using the features listed in Table 11. It can be seen that both XGBoost + GANs–CNN and XGBoost + GANs–RNN frameworks gave the same enrichment ratio of 27.6, using data selection filter A.

An overview of the enrichment ratios that were calculated using the descriptors listed in Supplementary Table SI-6 is provided in Supplementary Table SI-9. It can be seen that employing data

TABLE 12 Enrichment Ratios of OPRK1 on the dataset consisting of features as shown in Table 11 with training size of 30%.

Classifier	Maxima	Filter	% of data	Minima	Filter	% of data
XGboost + GANs–CNN	27.6	Filter A	1.0	20.2	Filter D	0.5
XGboost + GANs–RNN	27.6	Filter A	1.0	21.0	Filter D	0.5

selection filter A, the XGBoost + GANs–RNN framework predictions provided the maximum enrichment ratio of 30.1.

4 Discussion

The Big Data analytics research outcomes in this study suggest that four proteins ADORA2A, ADRB2, OPRK1, and OPRD1, and their binding conformations considered in this work do possess similar global properties that can be leveraged to predict whether they will be more likely to bind their ligands than other conformations. The enrichment factors obtained with the best approaches are about 10 to about 40 times better than what would be available with a random selection of protein conformations for docking. For three out of the four targets of interest here (i.e., ADORA2A, OPRK1, and OPRD1), the physico-chemical features that are most associated with a high propensity to be selected for binding by the ligands are the water accessible surface area (MOE descriptor *pro_asa_vdw*), the hydrophobic surface area (MOE descriptor *pro_asa_hyd*) and the hydrophobicity moment (MOE descriptor *pro_hyd_moment)*. That these properties, which are global and not limited to the binding sites, are common to the important descriptor of all proteins point to a dual role of exposure to solvent and hydrophobicity as globally driving the capacity of proteins to bind, or not, their ligands. Note that this work is not a structure-activity relationship studies, i.e., we do not at this point give a range of values for these proteins that would be associated with ligand binding and a range of values that would be associated with non-ligand binding.

The fourth protein target that was used here, ADRB2, can also be analyzed by deep learning approaches to identify the ligand binding conformations about 24 times better than a random selection of conformations. Yet, that one protein target yields different physico-chemical features than the other three proteins used here, although the general role of surface hydrophobicity and electrostatics (negatively-charged regions, precisely) is conserved. We do not yet know if this difference observed between ADRB2 and the other proteins is a result of different actual physicochemical mechanisms involved in ligand binding, or if this is an artifact of the data and of specific issues with class imbalance from the MD trajectories of this target.

Nonetheless, the fact that the apo-proteins' global physico chemicals properties may—to an extent—predict the ligand-binding character of conformations is remarkable. Naturally, this does not mean that only global protein properties are "holding" the keys to the conformational selection mechanisms. This work will have to be continued and repeated with features that are specific to the binding sites' conformations rather than describing the global protein structure.

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

Conceptualization, SG, JB, and VM; methodology, SG; software, SG; validation, SG, JB, and VM; formal analysis, SG, JB, and VM; investigation, SG; resources, SG and JB; data curation, SG and JB; writing—original draft preparation, SG, JB, and VM; writing—review and editing, JB and VM; visualization, SG; supervision, JB and VM; project administration, JB and VM; funding acquisition, VM. All authors have read and agreed to the published version of the manuscript.

Funding

Funding to VM and JB was provided by The University of Alabama.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their

affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.953984/full#supplementary-material

References

Abd Elrahman, S. M., and Abraham, A. (2013). A review of class imbalance problem. *J. Netw. Innov. Comput.* 1, 332–340. doi:10.20943/01201706.4351

Aggarwal, R., Gupta, A., and Priyakumar, U. D. (2021). APObind: A dataset of ligand unbound protein conformations for machine learning applications in de novo drug design. Available at: https://arxiv.org/abs/2108.09926 (Accessed 08 25, 2021).

Akondi, V. S., Menon, V., Baudry, J., and Whittle, J. (2019). "Novel K-means clustering-based undersampling and feature selection for drug discovery applications," in Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019, 2771–2778.

Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö., McCammon, J. A., Miao, Y., et al. (2018). Ensemble docking in drug discovery. *Biophys. J.* 114, 2271–2278. doi:10.1016/j.bpj.2018.02.038

Babajide Mustapha, I., and Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. *Molecules* 21, 983. doi:10.3390/molecules21080983

Chemical Computing Group (2019). Molecular operating environment (MOE). Available at: https://www.chemcomp.com/Products.htm.

Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. arXiv: 1603.02754 2016.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv, abs/1412.3555.

Cimpoi, M., Maji, S., Kokkinos, I., and Vedaldi, A. (2016). Deep filter banks for texture recognition, description, and segmentation. *IJCV* 118 (1), 65–94. doi:10.1007/s11263-015-0872-3

Eckmann, J.-P., Kamphorst, S. O., and Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhys. Lett.* 4, 973–977. doi:10.1209/0295-5075/4/9/004

Evangelista, W., Weir, R. L., Ellingson, S. R., Harris, J. B., Kapoor, K., Smith, J. C., et al. (2016). Ensemble-based docking: From hit discovery to metabolism and toxicity predictions. *Bioorg. Med. Chem.* 24, 4928–4935. doi:10.1016/j.bmc.2016.07.064

Gupta, S., Baudry, J., and Menon, V. (2022). Using big data analytics to "back engineer" protein conformational selection mechanisms. *Molecules* 27 (8), 2509. doi:10.3390/molecules27082509

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Hauke, J., and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaest. Geogr.* 30, 87–93. doi:10.2478/v10117-011-0021-1

Ho, Y., and Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* 8, 4806–4813. doi:10.1109/ACCESS.2019.2962617

Hossain, M. A., and Sajib, M. (2019). Classification of image using convolutional neural network (CNN). *Glob. J. Comput. Sci. Technol.* 19, 13–18. doi:10.34257/GJCSTDVOL19IS2PG13

Hu, F., Xia, G.-S., Hu, J., and Zhang, L. (2015). Transferring deep convolutional neural networks for the classification of high- resolution remote sensing imagery. *Remote Sens.* 7 (11), 14680–14707. doi:10.3390/rs71114680

Jo, W., and Kim, D. (2022). Obgan: Minority oversampling near borderline with generative adversarial networks. *Expert Syst. Appl.* 197, 116694. doi:10.1016/j.eswa.2022.116694

Johnson, K. J., and Synovec, E. R. (2002). Pattern recognition of jet fuels: Comprehensive GCGC with ANOVA-based feature selection and principal component analysis. *Chemom. Intell. Lab. Syst.* 60, 225–237. doi:10.1016/s0169-7439(01)00198-8

Macedo, F., Oliveira, M. R., Pacheco, A., and Valadas, R. (2019). Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing* 325, 67–89. doi:10.1016/j.neucom.2018.09.077

Oyelade, O. J., Oladipupo, O. O., and Obagbuwa, I. C. (2010) Application of k means clustering algorithm for prediction of students academic performance. arXiv preprint arXiv:1002.2425, 2010.

Penatti, O., Nogueira, K., and Santos, J. (2015). "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, June 7 2015 to June 12 2015.

Schmidt, R. (2019). Recurrent neural networks (RNNs): A gentle introduction and overview.

Seelinger, D., and de Groot, B. L. (2010). Conformational transitions upon ligand binding: Holo-structure prediction from apo conformations. *PLOS Comput. Biol.* 6 (1), e1000634. doi:10.1371/journal.pcbi.1000634

Sripriya Akondi, V., Menon, V., Baudry, J., and Whittle, J. (2022). Novel big data-driven machine learning models for drug discovery application. *Molecules* 27, 594. doi:10.3390/molecules27030594

Sultana, F., Sufian, A., and Dutta, P. (2018). "Advancements in image classification using convolutional neural network," in 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 22-23 Nov. 2018, 122–129. doi:10.1109/ICRCICN.2018.8718718

Torng, W., and Altman, R. B. (2017). 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinforma.* 18, 302. doi:10.1186/s12859-017-1702-0

Wilkin, G. A., and Huang, X. (2007). "K-Means clustering algorithms: Implementation and comparison," in Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007), Iowa, USA, August 13-15, 2007 (IEEE), 133–136.

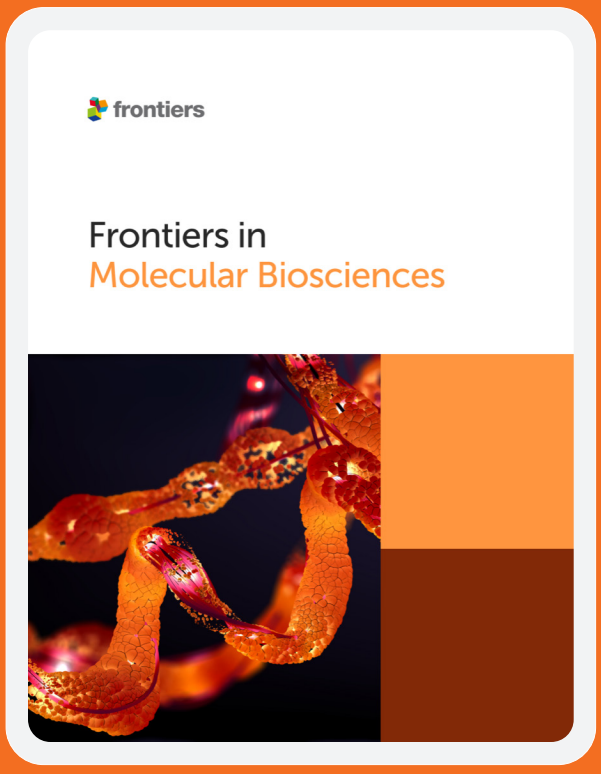
Frontiers in
Molecular Biosciences

Explores biological processes in living organisms on a molecular scale

Focuses on the molecular mechanisms underpinning and regulating biological processes in organisms across all branches of life.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact