UNIVERSITY OF CALIFORNIA, SAN DIEGO

Integrating conformational and protonation equilibria in biomolecular modeling

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Chemistry

by

Meekyum Olivia Kim

Committee in charge:

Professor J. Andrew McCammon, Chair
Professor Rommie Amaro
Professor Michael K. Gilson
Professor Susan Taylor
Professor Wei Wang

2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

The dissertation of Meekyum Olivia Kim is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2015

EPIGRAPH

Simplicity is the highest goal, achievable when you have overcome all difficulties. After one has played a vast quantity of notes and more notes, it is simplicity that emerges as the crowning reward of art.

Frédéric Chopin

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SCHEMES

LIST OF TABLES

LIST OF ABBREVIATIONS

ε – Dielectric constant
Aβ – Amyloid β
ABZ – Albendazole
APP – Amyloid precursor protein
AUC – Area under the curve
B3LYP – Becke exchange with Lee, Yang and Parr correlation (3-parameter) hybrid functional
BACE-1 – β-site APP cleaving enzyme 1
BZ – Benzimidazole
BPH – Bisphosphonate
CB[7] – Cucurbit[7]uril
CBZ – Carbendazim
CITPP – Citronellyl diphosphate
cMD – Conventional molecular dynamics
CpHMD – Constant pH molecular dynamics
EF – Enrichment factor
ESP – Electrostatic potential
E,E-FPP – *E,E*-farnesyl diphosphate
E,Z-DPP – *E,Z*-decaprenyl diphosphate
E,Z-DPPS – *E,Z*-decaprenyl diphosphate synthase
E,Z-FPP – *E,Z*-farnesyl diphosphate
E,Z-FPPS – *E,Z*-farnesyl diphosphate synthase
FBZ – Fuberidazole
FEP – Free energy perturbation
FN – False negative
FP – False positive
FPR – False positive rate
GAFF – General Amber force field
GB – Generalized Born
GB-OBC – Generalized Born model by Onufriev, Bashford, Case
GPP – Geranyl diphosphate
H^+ – Proton
HF – Hartree-Fock
HIV-1 – Human immunodeficiency virus type 1
HTS – High-throughput screening
IC_{50} – Half inhibitory concentration
IPP – Isopentenyl diphosphate
k_B – Boltzmann constant
MC – Monte Carlo
MD – Molecular dynamics
MK – Merz-Kollman
n – Hill coefficient
NMR – Nuclear magnetic resonance
NPT – Constant pressure and temperature
NVT – Constant volume and temperature
PC – Principal component
PCA – Principal component analysis
PDB – Protein Data Bank
pH – Negative logarithmic value of activity of hydrogen ion in an aqueous solution
pH-REMD – Constant pH replica exchange molecular dynamics
pK_a – Negative logarithmic value of the acid dissociation constant
POVME – Pocket Volume Measurer

PV – Pressure multiplied by volume
QM – Quantum mechanics
QM/MM – Quantum mechanics/molecular mechanics
RESP – Restrained electrostatic potential
RmlC – dTDP-6-deoxy-D-xylo-4-hexulose 3',5'-epimerase
RMSD – Root-mean-square deviation
RMSF – Root-mean-square fluctuation
ROC – Receiver operating characteristic
Rv3378c – Tuberculosynol/tuberculosinyl adenosine synthase
s – Fraction of deprotonated species
SE – Standard error
TBZ – Thiabendazole
TI – Thermodynamic integration
TN – True negative
TP – True positive
TPR – True positive rate
TRH - 2'-Deoxy-thymidine-β-L-rhamnose
UPP – Undecaprenyl diphosphate
UPPS – Undecaprenyl diphosphate synthase
UV – Ultraviolet
vdW – van der Waals
VS – Virtual screening
Z_{LR} – Charge of ligand-receptor complex
Z_{L} – Charge of ligand
Z_{P} – Charge of protein
Z_{PL} – Charge of protein-ligand complex
Z_{R} – Charge of receptor

ACKNOWLEDGEMENTS

I first thank my parents, who have showed infinite love and support, taught me patience and persistence, and given me confidence throughout my life. Your constant assurance that I can make things happen has helped me begin and complete my first adventurous journey to here. To my brother, who also sent me vast amount of love and support, I always appreciate and learn from your persistence. To Levi, thanks for being there for me with love, passion, and support, over 2,500 miles from New York, sharing my joy and holding me when I feel weak. Your infinite inspiration and creativity have further sophisticated my knowledge and personality to who I am now. This summer will mean so much to us, together with Riblet (with a coral bow tie).

To Andy, I owe a great amount of thanks to you for the numerous opportunities and guidance you provided, through which I learned so much during my graduate years. Besides science, owing to your generosity, I could indulge in fanciful cultural activities that otherwise I would rarely have had the chance to experience. I also thank the McCammon group members, whom I learned very much from their joy and curiosity in science by collaborating and communicating. Especially, I must thank Pat Blachly, who has been an amazing teacher for the last couple years in numerous aspects, including Sarah Palin (and for that reason, I thank Alysia as well). To Mehrnoosh Arrar, Joe Kaus, Nuo Wang, Leo Boechi, Steffen Lindert, Bill Sinko, César de Oliveira, Ferran Feixas, Yi Wang, and Sara Swift, thanks for making my graduate years so lively and productive with occasional fun hangouts. To Patti, thanks for making my life easier and particularly, for letting my days not bore by sharing your dinnerware collection. To Robert Konecny and Brian fox, thanks for always coming to check my computer whenever I said “my computer is not working.”

Finally, I would also like to thank my doctoral committee, Professors Rommie Amaro, Mike Gilson, Susan Taylor, and Wei Wang, for their guidance and support for me to successfully complete my degree.

Chapter 2 is a minimally modified reprint of the material as it appears in Meekyum Olivia Kim, Sara E. Nichols, Yi Wang, and J. Andrew McCammon, “Effects of histidine protonation and

rotameric states on virtual screening of *M. tuberculosis* RmlC," *Journal of Computer-Aided Molecular Design*, 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is a minimally modified reprint of the material as it appears in Meekyum Olivia Kim, Xinxin Feng, Ferran Feixas, Wei Zhu, Steffen Lindert, Shannon Bogue, William Sinko, César de Oliveira, Guodong Rao, Eric Oldfield, and J. Andrew McCammon, "A molecular dynamics investigation of *Mycobacterium tuberculosis* prenyl synthases: Conformational flexibility and implications for computer-aided drug discovery," *Chemical Biology & Drug Design*, 2014. The dissertation author was one of three primary investigators, who each contributed equally to performing research and writing this paper.

Chapter 4 is a minimally modified reprint of the material as it appears in Meekyum Olivia Kim, Patrick G. Blachly, Joseph W. Kaus, and J. Andrew McCammon, "Protocols Utilizing Constant pH Molecular Dynamics to Compute pH-Dependent Binding Free Energies," *Journal of Physical Chemistry B*, 2015. The dissertation author was one of two primary investigators, each of whom contributed equally to performing research and writing this paper.

Chapter 5 is a minimally modified reprint of the material as it appears in Meekyum Olivia Kim, Patrick G. Blachly, and J. Andrew McCammon, "Conformational dynamics and inhibitor binding of BACE-1: From the perspective of protonation equilibria," submitted to *PLoS Computational Biology*, 2015. The dissertation author was the primary investigator and author of this paper.

VITA

2011	B.S. in Chemistry Sookmyung Women's University, Seoul, South Korea
2013	M.S. in Chemistry University of California, San Diego, La Jolla, USA
2015	Ph.D. in Chemistry University of California, San Diego, La Jolla, USA

PUBLICATIONS

Kim M. O., Blachly P. G., McCammon J. A. Conformational dynamics and inhibitor binding of BACE-1: From the perspective of protonation equilibria, *PLoS Comp. Biol.* In press.

Kim M. O.*, Blachly P. G.*, Kaus J. W., McCammon J. A. Protocols utilizing constant pH molecular dynamics to compute pH-dependent binding free energies, *J. Phys. Chem. B.* **2015**, *119*, 861-872.
* These authors contributed equally to this work.

Kim M. O.*, Feng X., Feixas F.*, Zhu W., Lindert S., Bogue S., Sinko W., de Oliveira C., Rao G., Oldfield E.*, McCammon J. A. A molecular dynamics investigation of *Mycobacterium tuberculosis* prenyl synthases: Conformational flexibility and implications for computer-aided drug discovery. *Chem. Biol. Drug Des.* **2014**, doi: 10.1111/cbdd.12463.
* These authors contributed equally to this work.

Kim M. O., Nichols S. E., Wang Y., McCammon J. A. Effects of histidine protonation and rotameric states on virtual screening of *M. tuberculosis* RmlC, *J. Comput. Aided Mol. Des.* **2013**, *27*, 235-246.

ABSTRACT OF THE DISSERTATION

Integrating conformational and protonation equilibria in biomolecular modeling

by

Meekyum Olivia Kim

Doctor of Philosophy in Chemistry

University of California San Diego, 2015

Professor J. Andrew McCammon, Chair

Due to high sensitivity of biomolecular systems to the electrostatic environments, coupled treatment of conformational and protonation equilibria is required for an accurate characterization of true ensemble of a given system. The research presented in this dissertation examines the effects of conformational and protonation equilibria of varying extent on diverse aspects of computational biomolecular modeling, as introduced in Chapter 1. The effects of protonation and stereoisomerism of two histidines on virtual screening against the *M. tuberculosis* enzyme RmlC are presented in Chapter 2. In Chapter 3, conformational flexibility of three *M. tuberculosis* prenyl synthases is probed using molecular dynamics simulations, with implications for computer-aided drug discovery effort for the new generation antibacterial and antivirulence therapeutics. Chapters 4 and 5 consider the conformational and protonation equilibria simultaneously by utilizing constant pH molecular dynamics, in which fluctuations in both conformation and protonation state are possible. In Chapter 4, a

computational protocol utilizing constant pH molecular dynamics to compute pH-dependent binding free energies is presented. The methodology is further applied to protein-ligand complexes in Chapter 5, where the thermodynamic linkage between protonation equilibria, conformational dynamics, and inhibitor binding is illustrated.

Chapter 1

Integrating conformational and protonation equilibria in biomolecular modeling

Introduction

The ensemble of protein conformations determines the possible functions of a protein. The instantaneous conformation in this structural ensemble is sensitive to changes in its surrounding environment, such as pH, temperature, pressure, and presence of a ligand. Protonation equilibria are of particular interest because virtually all proteins depend on tight regulation of the intracellular pH to maintain stability and function.[1, 2] The surrounding electrostatic environment is reflected as the microscopic pK_a of the titratable residues in the protein, the pH at which individual acidic and basic residues are 50% protonated. Changes in conformation of the protein can alter the electrostatics, which may induce a shift in the pK_a of titratable groups.[3] Furthermore, the formation of a complex between the protein and small molecules, nucleic acids, or other proteins can also cause changes in the pK_a of titratable groups on either binding partner.[4] Therefore, coupling of conformational and protonation equilibria is essential to an accurate characterization of structural ensemble of a given biomolecular system.

Conventional biomolecular modeling or free energy computations involving the complex formation typically employ fixed protonation states for the titratable groups in both binding partners set *a priori*, which are identical for the free and bound states. Clearly, in cases where the ligand binding accompanies a net proton transfer to the system, this assumption ignores the possibility of protonation

states changing while the chemical environments change upon binding. Hence, when the true ensemble of conformations consists of varying electrostatic environments, the use of a single, static protonation state may hinder the accurate description of the ensemble and can lead to significant errors.[4] In this dissertation, the effects of conformational and protonation equilibria are examined in diverse aspects of computational biomolecular modeling, with varying degrees of each taken into account. The following sections briefly describe the motivations and methodologies behind each chapter in addressing the influence of conformational and protonation ensembles.

Protonation and stereoisomerism of active site histidines of RmlC

While multiple tautomers and protonated forms of ligand molecules are considered in a virtual screening (VS), titratable groups in protein receptors are assigned a single protonation state that is identical for both bound and unbound forms. Given a goal of VS to identify small molecules binding to a target protein with high affinities, such presumption precludes the possibility of finding diverse ligands that bind to the receptors with different protonation states. An example of where these effects may be important is the *M. tuberculosis* enzyme RmlC, as the active site of the enzyme has two histidines that engage in proton transfer during the catalysis. Depending on the net charge and the location of proton(s), a histidine can adopt three states: HIP (+1 charged, both δ- and ε-nitrogens protonated), HID (neutral, δ-nitrogen protonated), and HIE (neutral, ε-nitrogen protonated). In addition, three additional states are possible for a histidine with its imidazole ring flipped. In Chapter 2, we evaluate the performance of VS using 36 receptor models with different protonation and rotameric states of two active site histidines to investigate the sensitivity of VS to the protonation states of titratable residues located in the receptor active site.

Conformational flexibility of antibacterial and antivirulence targets

In addition to the assignment of a fixed protonation state, the protein receptors are treated rigidly in typical VS study by the use of static X-ray crystallographic structures. However, not only is the intrinsic conformational flexibility of the receptor ignored but also the structural rearrangement

induced by ligand binding in such rigid treatment of receptors. Consequently, an induced fit mechanism has been incorporated into the docking algorithms in order to account for receptor flexibility.[5, 6] While the docking protocols that consider induced fit show improvements in the accuracy of the VS results by accounting for binding-induced conformational changes,[7-10] the search algorithm is largely limited to side chain rotations, and therefore, global structural changes upon ligand binding or rare receptor conformations are not addressed.

Another approach to account for receptor flexibility in structure-based drug discovery is the relaxed complex scheme, which utilizes an ensemble of various conformations usually generated by molecular dynamics (MD) simulations.[11] In an MD simulation, the Hamiltonian is computed using the classical molecular mechanics force field, and the time evolution of the system configurations is obtained by solving Newton's equations of motions.[12, 13] The use of the structural ensemble generated by MD simulations within the relaxed complex scheme framework enables sampling of the global structural changes in the receptor conformation that occur on longer timescales, as well as incorporating infrequently observed, yet biologically important, conformations upon ligand binding into the VS protocol. The method has been successfully applied to various structure-based drug discovery studies, showing promising results in finding the inhibitors that bind to rarely sampled receptor conformers.[14, 15] In Chapter 3, we apply the relaxed complex scheme to the computer-aided drug discovery effort to develop the novel new generation antibacterial and antivirulence therapeutics targeting three *M. tuberculosis* prenyl synthases: *E,Z*-decaprenyl diphosphate synthase; *E,Z*-farnesyl diphosphate synthase; and tuberculosinol/tuberculosinyl adenosine synthase.

Computing pH-dependent binding free energies

In Chapters 2 and 3, the effects of protonation and conformational equilibria on computational biomolecular modeling are respectively analyzed. However, due to the tight coupling between the conformation and its surrounding electrostatic environments, as described above, the conformational and protonation spaces have to be addressed together for an accurate description of true ensemble for a given biomolecular system. This is particularly important to the biomolecular association, where

protein-ligand binding can accompany a net transfer of protons, and hence, the binding process is pH-dependent.

Constant pH molecular dynamics (CpHMD) is a technique that incorporates pH as an additional external thermodynamic variable to the conventional MD framework, enabling concurrent sampling of conformational and protonation spaces according to the semi-grand canonical ensemble.[16] In CpHMD simulations, conformational fluctuations propagate with MD simulations while intermittent Monte Carlo steps evaluate a change in the protonation state of a given titratable residue.[16, 17] The new protonation state is either accepted or rejected contingent on the application of the Metropolis criterion, and the MD simulation continues. Repeated application of these steps builds an ensemble of conformations and protonation states along the MD trajectory. In Chapter 4, we describe a computational protocol combining CpHMD simulations with the binding polynomial formalism[18] to compute pH-dependent binding free energies. In this approach, the CpHMD method provides a framework through which the pH dependence of binding processes can be examined with respect to the reference state, which is taken either from experiment or from thermodynamic integration computations. Tested to various host-guest complexes, the methodology highlights the errors associated with incorrect assignment of protonation states in free energy computations and represents a promising approach to properly address the changes in protonation states upon complex formation.

Proton-linked conformational dynamics and inhibitor binding of BACE-1

In Chapter 5, conformational dynamics and binding free energies of inhibitors of BACE-1, the β-secretase responsible for the amyloidogenesis in Alzheimer's disease, are described from the perspective of protonation equilibria. The catalytic mechanism of BACE-1 involves a water-mediated proton transfer from the aspartyl dyad in the active site to the substrate, as well as structural flexibility in the flap region. Therefore, the coupling of protonation and conformational equilibria is essential to a full *in silico* characterization of BACE-1. This work applies the CpHMD-based free energy method

described in Chapter 4 to quantify the binding affinity in protein-ligand systems and highlights the importance of correctly addressing the binding-induced changes in protonation states for cases where protein-ligand binding accompanies a net proton transfer. Also, the value of CpHMD technique as an all-purpose tool for obtaining pH-dependent dynamics and binding free energies of biological systems is illustrated, suggesting high utility of the CpHMD method in computer-aided drug design workflows.

References

1. Yang, A. S.; Honig, B. On the pH Dependence of Protein Stability. *J. Mol. Biol.* **1993,** *231*, 459-474.

2. Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Preiction of pH-Dependent Properties of Proteins. *J. Mol. Biol.* **1994,** *238*, 415-436.

3. García-Moreno, B. Adaptations of Proteins to Cellular and Subcellular pH. *J. Biol.* **2009,** *8*, 98-101.

4. Onufriev, A. V.; Alexov, E. Protonation and pK Changes in Protein-Ligand Binding. *Q. Rev. Biophys.* **2013,** *46*, 181-209.

5. Birch, L.; Murray, C. W.; Harshom, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of Molecular Docking to Induced Fit Effects in Influenza Virus Neuraminidase. *J. Comput. Aided Mol. Des.* **2002,** *16*, 855-869.

6. Sherman, W.; Day, T.; Jacobson, M. P.; Freisner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006,** *49*, 534-553.

7. Betts, M. J.; Sternberg, M. J. An Analysis of Conformational Changes on Protein-Protein Association: Implications for Predictive Docking. *Protein Eng.* **1999,** *12*, 271-283.

8. Murray, C. W.; Baxter, C. A.; Frenkel, A. D. The Sensitivity of the Results of Molecular Docking to Induced Fit Effects: Application to Thrombin Thermolysin and Neuraminidase. *J. Comput. Aided Mol. Des.* **1999,** *13*, 547-562.

9. Najmonovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Side-Chain Flexibility in Proteins Upon Ligand Binding. *Proteins* **2000,** *39*, 261-268.

10. McRobb, F. M.; Capuano, B.; Crosby, I. T.; Chalmers, D. K.; Yuriev, E. Homology Modeling and Docking Evaluation of Aminergic G Protein-Coupled Receptors. *J. Chem. Inf. Model.* **2010,** *50*, 626-637.

11. Amaro, A. E.; Baron, R.; McCammon, J. A. An Improved Relaxed Complex Scheme for Receptor Flexibility in Computer-Aided Drug Design. *J. Comput. Aided Mol. Des.* **2008,** *22*, (693-705).

12. Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002,** *9*, 646-652.

13. Monticelli, L.; Tieleman, D. P. Force Fields for Classical Molecular Dynamics. *Methods Mol. Biol.* **2013,** *924*, 197-213.

14. Amaro, A. E.; Schnaufer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A. Discovery of Drug-Like Inhibitors of an Essential RNA-Editing Ligase in Trypanosoma Brucei. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *105*, 17278-17283.

15. Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.* **2004,** *47*, 1879-1881.

16. Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH Molecular Dynamics Using Stochastic Titration. *J. Chem. Phys.* **2002,** *117*, 4184-4200.

17. Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH Molecular Dyamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004,** *25*, 2038-2048.

18. Wyman, J. Heme Proteins. *Adv. Protein. Chem.* **1948,** *4*, 407-531.

Chapter 2

Effects of histidine protonation and rotameric states on virtual screening of *Mycobacterium tuberculosis* RmlC

Abstract

While it is well established that protonation and tautomeric states of ligands can significantly affect the results of virtual screening, such effects of ionizable residues of protein receptors are less well understood. In this study, we focus on histidine protonation and rotameric states and their impact on virtual screening of *Mycobacterium tuberculosis* enzyme RmlC. Depending on the net charge and the location of proton(s), a histidine can adopt three states: HIP (+1 charged, both δ- and ε-nitrogens protonated), HID (neutral, δ-nitrogen protonated), and HIE (neutral, ε-nitrogen protonated). Due to common ambiguities in X-ray crystal structures, a histidine may also be resolved as three additional states with its imidazole ring flipped. Here, we systematically investigate the predictive power of 36 receptor models with different protonation and rotameric states of two histidines in the RmlC active site by using results from a previous high-throughput screening. By measuring enrichment factors and area under the receiver operating characteristic curves, we show that virtual screening results vary depending on hydrogen bond networks provided by the histidines, even in the cases where the ligand does not obviously interact with the side chain. Our results also suggest that, even with the help of widely used pK_a prediction software, assigning histidine protonation and rotameric states for virtual screening can

still be challenging and requires further examination and systematic characterization of the receptor-ligand complex.

Introduction

The effect of ligand protonation and tautomeric states on virtual screening (VS) has been the subject of extensive research.[1-4] It is well known that different protonated forms or tautomers of a ligand may have significantly different rankings in VS.[1, 2] Unlike ligand molecules, for which multiple tautomers and protonated forms can be included in a VS study, the ionizable residues of protein receptors are assigned a single state prior to the screening. For instance, in the standard protonation model, all Asp, Glu, and His residues are deprotonated while all Arg and Lys residues are protonated at pH 7.4, the typical extracellular pH for the biological systems. Various algorithms, such as PROPKA,[5-8] H++,[9-11] and MCCE,[12-14] have been developed to improve the quality of the proton assignment. However, few studies have investigated the effect of such assignment of the titratable residues of protein receptors on VS results.[15] In this study, we focus on the impact of histidine protonation and rotameric states on VS by systematically analyzing a screen using results from a previous high-throughput screening (HTS) of the enzyme RmlC (dTDP-6-deoxy-D-xylo-4-hexulose 3',5'-epimerase) of *Mycobacterium tuberculosis*.[16]

Histidines participate in a large number of important biochemical reactions. As catalytic residues in the enzymatic active site,[17] proton shufflers in proton transfer reactions,[18-20] coordinating ligands in metalloproteins and hemoglobin,[21, 22] histidines play essential roles for proper functions of a cell. The side chain of a histidine has a pK_a around 6.0, which is close to the physiological pH.[23] Depending on the pH of its environment, a histidine readily switches between the doubly protonated cationic form and the neutral state (Figure 2-1): At low pH, both δ-nitrogen and ε-nitrogen of the imidazole ring are protonated and the amino acid has a +1 charge (HIP). At high pH, the histidine is neutral with either δ-nitrogen (HID) or ε-nitrogen (HIE) protonated. Apart from the above three states, positions of carbon and nitrogen atoms in the imidazole ring may be switched due to common ambiguities in X-ray crystallographic structures.[24] As a result, a histidine can adopt three additional

rotameric states, namely, flipped HIP, flipped HID, or flipped HIE (see Figure 2-1).[25] In this work, we set out to evaluate the impact of all six protonation and rotameric states of a histidine on the virtual screening results.

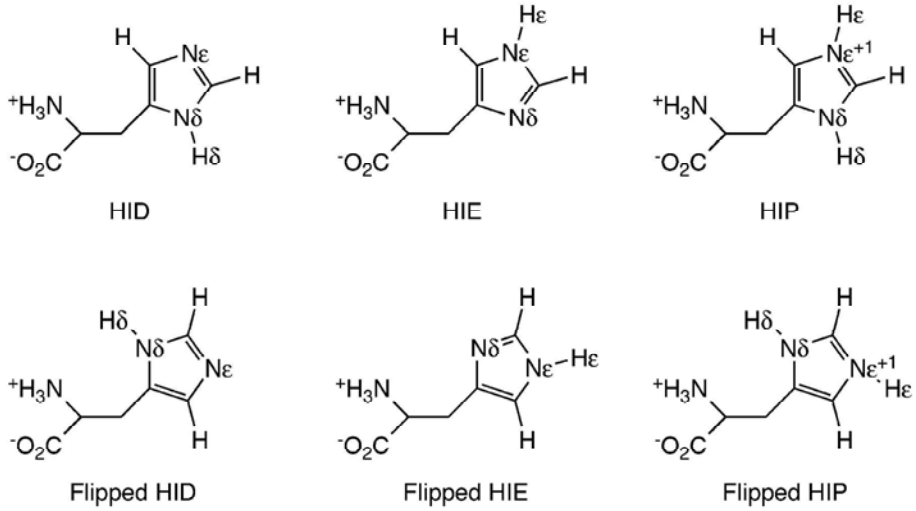


Figure 2-1. Six possible protonation and rotameric states of a histidine. Formal charges on nitrogen in HIP states are marked.

M. tuberculosis is the primary causative pathogen of the lethal, contagious disease tuberculosis. It has a three-layered cell wall composed of peptidoglycan, arabionogalactan, and mycolic acids.[26] This highly impermeable cellular envelope provides natural resistance against a large variety of antibiotics, which renders the inhibition of the cell wall biosynthesis a promising target for drug discovery against tuberculosis.[16, 26] The enzyme RmlC participates in the synthesis of an indispensible linker molecule 2'-deoxy-thymidine-β-$_L$-rhamnose (TRH), connecting the peptidoglycan and arabinogalactan layer in the bacterial cell wall.[6, 16] Based on the crystal structure of the *M. tuberculosis* RmlC in complex with TRH (Figure 2-2A), it has been suggested that the enzyme uses a histidine (His62) as a key catalytic site that pairs with Tyr132 in an acid-base couple for proton transfer.[27] Apart from His62, the active site contains another histidine, *i.e.*, His119, involved in the interaction with TRH.

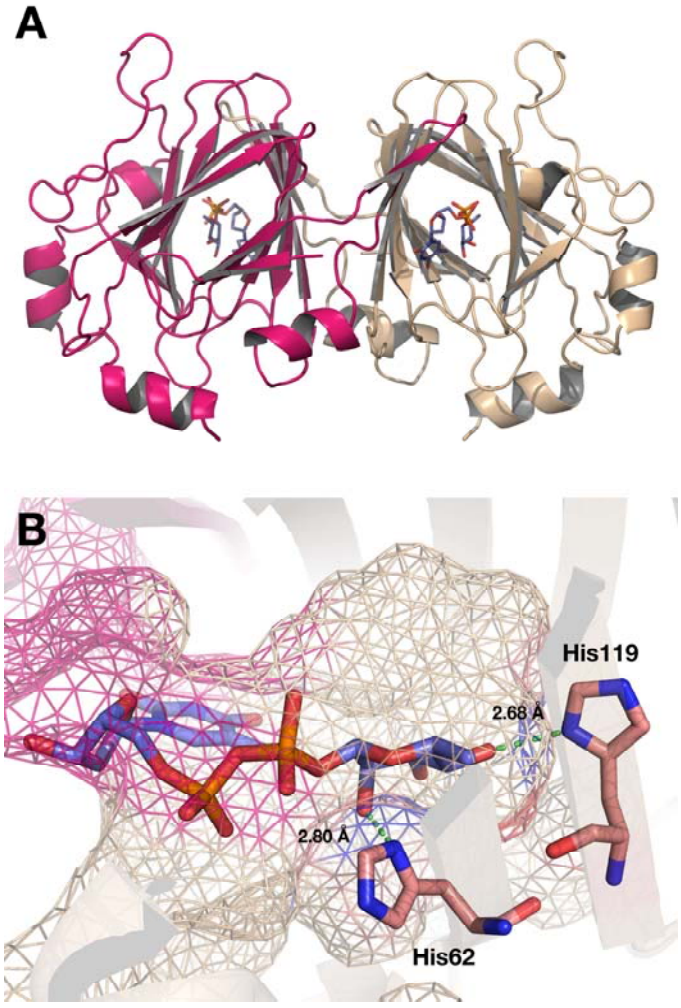


Figure 2-2. (A) RmlC homodimer in complex with co-crystalized TRH (PDB ID 2IXC). The two monomers are colored in pink and beige, respectively. (B) Close view of the co-crystal ligand TRH, with His62 and His119 highlighted. The binding surface of receptor is represented as wire frame. Hydrogen bonds are shown with dashed green lines.

As a part of a drug discovery campaign against tuberculosis, 201,368 compounds were screened in a previous HTS against RmlC, revealing a series of hits with the best half inhibitory concentration (IC_{50}) of 0.12 μM at pH 7.4.[16] Based on these results, we constructed a library of 2,010 compounds, including 2,000 decoys and ten actives. The library was screened against 36 receptor models with different protonation and rotameric states of His62 and His119 of RmlC. Through enrichment factors (EF), receiver operating characteristic (ROC) curves, and area under the curve (AUC) metrics, we systematically evaluated the relative VS performance of various protonated receptor

models. In the remainder of the text, we will discuss these analyses in detail and examine pK_a predictions for the two histidines made by commonly used software packages.

Methods

Crystal structure and initial preparation

The X-ray crystallographic structure of RmlC in complex with the product analog 2'-deoxy-thymidine-β-L-rhamnose (TRH) was obtained from Protein Data Bank (PDB ID 2IXC).[27] One dimer containing chains A and B, each in complex with a TRH molecule, was submitted to the Protein Preparation Wizard of Schrödinger Suite 2011.[28] Missing hydrogens in the crystal structure were added while water and TRH molecules were removed, followed by a brief optimization of hydrogen positions at pH 7.0. Receptor models with 36 different protonation and rotameric states of His62 and His119 in chain A were then generated and refined with the OPLS2005 force field. Two other titratable residues in the active site, Lys72 and Asp83, were kept charged. Subsequent virtual screening (VS) was performed on the active site of chain A. See Figure 2-S1 for a schematic description of the hydrogen and nitrogen of His62 and His119 acting as a hydrogen bond donor or acceptor in each receptor model.

Receptor grid generation

A set of 36 receptor models based on different protonation states of His62 and His119 were generated using Glide 5.7 in Schrödinger Suite 2011.[29] The grid center was set to where the center of mass of the co-crystallized TRH molecule in chain A had been before removal. The sizes of the inner and outer grid boxes were set to 10 Å and 20 Å in each direction, respectively. The models were assigned unique numbers from 1 to 36 as listed in Table 2-1.

Ligand preparation

A library containing 2,000 inactive and ten active compounds was generated from the previous HTS result of the NIH Molecular Libraries Small Molecule Repository (BioFocus DPI).[16] First, the entire library of 201,386 compounds used in the HTS was obtained from PubChem, with BioAssay IDs 1532, 1533, 1695, and 1696 (including primary screening results and dose-response assays).[16] Ten verified actives were selected from BioAssay 1696 and 2,000 inactive compounds were

randomly selected from the remaining compounds. The final library of 2,010 ligands was then subjected to LigPrep of Schrödinger Suite 2011[30] with OPLS2005 force field. The ligands were ionized within the pH range of 7.0 ± 2.0 using Epik[31, 32] and tautomers and stereoisomers were generated for the inactives, resulting in a library of total 3,934 compounds. The Canvas tool of Schrödinger Suite 2012[33-35] was used to compare the similarity of the actives and decoys. Tanimoto coefficients of the compounds in the library to each of the actives were calculated based on the molecular binary fingerprints, as described in Supporting Information 2. The co-crystal ligand TRH was also prepared in the same way and docked to all models for initial assessment of pose prediction.

Docking

After experimenting with both the Glide SP and XP docking modes[29, 36-38] we found that XP outperformed SP in ranking the actives over decoys (data not shown). The different performance of SP and XP mainly stems from differences in their scoring functions. The hydrophobic enclosure term in the XP algorithm may be particularly suitable for our study, given the strong hydrophobic interactions between many active compounds and the binding site.[38] Hence, in the remainder of the study, we used the Glide XP precision to perform docking on 36 receptor models described above.

Predictive performance analysis

We analyzed the predictive performance of VS using 36 receptor models described above by calculating enrichment factors (EF), receiver operating characteristic (ROC) curves, and areas under the curve (AUC). The statistical significance of the AUC values of different receptor models was evaluated with a *p*-test with 95% confidence limit.

The EF is a widely used metric to evaluate the efficiency of VS.[39] The value of $EF^{x\%}$ indicates how much more likely an active compound is ranked in the top x% of a VS result compared to a random selection, *i.e.*, how many times the database is enriched. Specifically, EF is calculated as Eq. 1:

$$EF^{x\%} = \frac{N^{x\%}_{experimental}}{N_{active} \times x\%} \quad (1)$$

where $N_{experimental}^{x\%}$ is the number of experimentally verified actives in the top x% of the database and N_{active} is total number of actives in the database.[39] In this study, $EF^{1\%}$ and $EF^{10\%}$ were calculated from the top 1% and 10% of the VS result, respectively.

To investigate the docking performance in a threshold-independent manner, the AUC value was calculated from the ROC curve. The ROC curve allows a straightforward visualization of the performance of VS in ranking the actives higher over decoys.[40] In our study, we have a list of experimentally verified actives, or positives, and decoys, or negatives. These positives and negatives are further categorized into true or false according to their rank above or below, a certain threshold of the VS result, respectively, *e.g.*, the actives ranked above a chosen threshold becomes true positive (TP). To generate the ROC curve, the true positive rate (TPR) and false positive rate (FPR) are calculated as Eqs. 2 and 3:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{3}$$

In the ROC curve, the TPR is plotted as a function of the FPR. The AUC is then calculated to compare the performance of different receptor models quantitatively.[23] An AUC of 0.5 corresponds to a random selection of the ligand by a receptor.

To evaluate the statistical significance of the AUC values of different receptor models, we performed the two-sided *p*-test at the 95% level. A two-sided *p*-value of less than 0.05 (corresponding to 5%) rejects the null hypothesis that the AUC values of a pair of receptors are statistically identical and accepts the alternative hypothesis that their difference is statistically meaningful. Hence the pair of receptors with statistically different AUC values is differentiated by their abilities to rank the actives and decoys. The two-sided *p*-values were calculated following Craig *et al*. and references therein,[41, 42] which is described below briefly. As in Eq. 4, the AUC is first calculated as the mean TPR of decoys:

$$\text{AUC} = \sum_{i}^{\text{decoys}} \Delta\text{FPR}\ \text{TPR}_i = \frac{1}{N_{\text{decoys}}} \sum_{i}^{\text{decoys}} \text{TPR}_i = \left\langle \text{TPR} \right\rangle_{\text{decoys}} = 1 - \left\langle \text{FPR} \right\rangle_{\text{actives}} \tag{4}$$

where TPR_i is the true positive rate at decoy i and ΔFPR is the constant increment in the false positive rate. The difference between the AUC values of the pair of receptor models A and B becomes as Eq. 5:

$$\Delta AUC = AUC_A - AUC_B = \langle TPR \rangle_{decoys,A} - \langle TPR \rangle_{decoys,B} = \langle TPR_A - TPR_B \rangle_{decoys} \quad (5)$$

where the last step arose from docking of the same library into all receptor models, which statistically indicates the pairing of samples. As a result,

$$\Delta AUC = \frac{1}{N_{decoys}} \sum_{i}^{decoys} \left(TPR_{i,A} - TPR_{i,B} \right) = \langle TPR_A - TPR_B \rangle_{decoys} \quad (6)$$
$$= \frac{1}{N_{actives}} \sum_{i}^{actives} \left(FPR_{i,B} - FPR_{i,A} \right) = \langle FPR_B - FPR_A \rangle_{actives}$$

Then the variances for the actives and decoys are given by:

$$Var_{\Delta,a} = \frac{1}{N_{actives}} \sum_{i}^{actives} \left\{ \left(FPR_{i,A} - FPR_{i,B} \right) - \langle FPR_A - FPR_B \rangle_{actives} \right\}^2 \quad (7)$$

$$Var_{\Delta,d} = \frac{1}{N_{decoys}} \sum_{i}^{decoys} \left\{ \left(TPR_{i,A} - TPR_{i,B} \right) - \langle TPR_A - TPR_B \rangle_{decoys} \right\}^2 \quad (8)$$

with the standard error in ΔAUC given as:

$$SE_\Delta = \sqrt{\frac{Var_{\Delta,a}}{N_{actives}} + \frac{Var_{\Delta,d}}{N_{decoys}}} \quad (9)$$

Finally, the two-sided *p*-value for ΔAUC between the two receptors is obtained as a Gaussian distribution with a standard deviation equal to SE_Δ:

$$p = \mathrm{erfc}\left(\frac{|\Delta AUC|}{\sqrt{2} SE_\Delta} \right) \quad (10)$$

where erfc is the complementary error function. All analyses of the receptor predictive power were done with Matlab R2011a (Version 7.12).[43]

pK_a Prediction analysis

PROPKA,[5-8] Maestro,[28, 44] H++,[9-11] and MCCE[12-14] were employed to predict the protonation states of His62 and His119 in the active site of RmlC. Along with the dimer of chains A and B from the X-ray crystal structure containing the co-crystallized TRH molecule, we also generated an additional structure by removing the TRH to examine the effect of the ligand on pK_a prediction. Therefore, the dimers with and without the ligand were subjected to pK_a calculation using PROPKA, Maestro, H++, and MCCE.

PROPKA[5-8] predicts the pK_a through an empirical method by calculating a pK_a shift, ΔpK_a, arising from perturbation of electrostatic energy of an ionizable residue between its charged and neutral states. Thus the pK_a is predicted by:

(11) $$pK_a = pK_{Model} + \Delta pK_a$$

with additional terms and parameters describing the Coulombic interaction, desolvation, unfavorable electrostatic reorganization energies, and hydrogen bond networks. The model pK_a used for histidine in PROPKA is 6.50.

The Protein Preparation Wizard of Maestro[28, 44] has been updated to employ PROPKA by default, from its previous version using Epik.[31, 32] The pK_a calculation using Epik relies on the well-established Hammett and Taft (HT)[45] linear free energy approach and the quality of hydrogen bond networks. In this study, we compared the pK_a prediction of Maestro both with and without PROPKA.

H++ is a single-structure continuum electrostatics methodology that predicts the pK_a values of the titratable residues based on either Generalized Born or Poisson-Boltzmann method using the AMBER 10 force field.[9-11] Multi-conformation continuum electrostatics (MCCE) calculates the pK_a values of the ionizable groups in protein and ligands by generating various conformations by side chain rotations throughout a titration.[12-14] The changes in the conformation create a position-dependent dielectric response and the degrees of freedom of the conformers are added to calculate the Boltzmann distribution of the ionization states and atomic positions. The pairwise electrostatic interactions between different conformers are calculated by the DelPhi Poisson-Boltzmann solver.

Results and discussion

In order to evaluate the effect of histidine protonation and rotameric states on the predictive performance of receptors, we performed virtual screening (VS) for the *M. tuberculosis* enzyme RmlC based on the results of a previous high-throughput screening (HTS) study. Below, we will first examine the typical interactions of the co-crystal ligand TRH to probe the ligand pose dependence on histidine protonation. We further contextualize analysis of enrichment performance and predictive power of various receptor models, by discussing the interactions with the receptor to show the effect of different histidine protonation states on VS. Finally, we compare the predicted pK_a values calculated by several common pK_a calculation packages to the receptor protonation states with the best predictive power.

Docking of TRH

Docking the co-crystal ligand TRH back into 36 receptor models was carried out to show the pose, or ligand orientation relative to the receptor, dependence on histidine protonation and rotameric states. Hydrogen bonding patterns for the crystal coordinates of His62 and His119, shown in Figure 2-2B, imply the potential significance of hydrogen bonds in docking of TRH. Docking of this ligand allowed preliminarily examination of the dependence of pose on possible hydrogen bond networks with the receptor.

Varying histidine protonation states has a clear effect on pose prediction for the determined co-crystal ligand. RMSD of docking pose of self-docked TRH into the crystal coordinates for different protonation and rotameric states of His62 and His119 varied from 2.91 Å to 5.44 Å. The protonation state of both histidines with the best average RMSD is HIE, which agrees with the most probable protonation states of the crystal coordinates in the presence of TRH. Also, in all cases, the docking algorithm predicts the position of the pyrophosphate of the ligand correctly, but the large deviation from the crystal coordinates mainly stems from the flipping of the thymidine and rhamnose moieties around the pyrophosphate, resulting in different hydrogen bonding patterns between TRH and two histidines. This indicates the importance of hydrogen bond networks with His62 and His119 in the pose prediction of the co-crystal ligand TRH. Therefore, after examining the pose dependence upon hydrogen bonds

provided by two histidines, we expanded our study to look systematically at ranking compounds in VS and how it is affected by the protonation and rotameric states of histidines.

Virtual screening

Molecular docking was carried out to examine compound ranking dependence on histidine protonation and rotomeric states. The ligand set included ten actives and 2,000 inactives selected at random from a HTS. We note that Tanimoto scores indicate that most of our decoys have a low similarity to the actives. Such a decoy set presents a smaller challenge to the docking algorithm and the predictive performance of VS itself may be affected when decoys with greater similarity to the actives are used. However, this study aimed to examine not the predictive performance of the docking algorithm *per se*, but how histidine protonation states affect the relative performance in VS.

Docked active ligands and the product analog were examined to characterize important interactions in the RmlC binding site. In all receptor models, hydrophobic pi-pi stacking interactions contribute significantly to docking scores of the active compounds within the RmlC active site. The initial hit compound from HTS, SID7975595, is ranked high in most receptor models, between 8th and 51st rank in 26 out of 36 receptors. Although there is only limited structural similarity between SID7975595 and the co-crystal ligand TRH, the tricyclic ring of SID7975595 readily replaces the TRH thymidine moiety, while the benzimidazolone ring replaces the rhamnose moiety, providing structural basis of the inhibition. As shown in Figure 2-3, the hydrophobic interaction between the actives and receptor often involves Tyr132 and Tyr138 from chain A and Phe26 from chain B (note that a part of chain B intrudes in the active site of chain A). Through interacting with the essential binding site residues and preventing water molecules from accessing Phe26 and Tyr132, the actives provide abundant hydrophobic contacts to achieve the high binding affinity. As discussed in Sivendran *et al.*,[16] substitution of the ethyl group attached to the nitrogen on the tricyclic ring of SID7975595 by an allyl group (*e.g.*, the active compound 77074) further enhances the binding affinity by forming an even tighter hydrophobic seal. In comparison, substitution of this group by a smaller methyl group or a hydrogen atom results in a lower binding affinity.[16] In addition to the hydrophobic contacts described

above, some of the actives also form hydrogen bonds with Ser51, Arg59, and Arg170. The interactions of the docked actives can be found in Figure 2-S3.

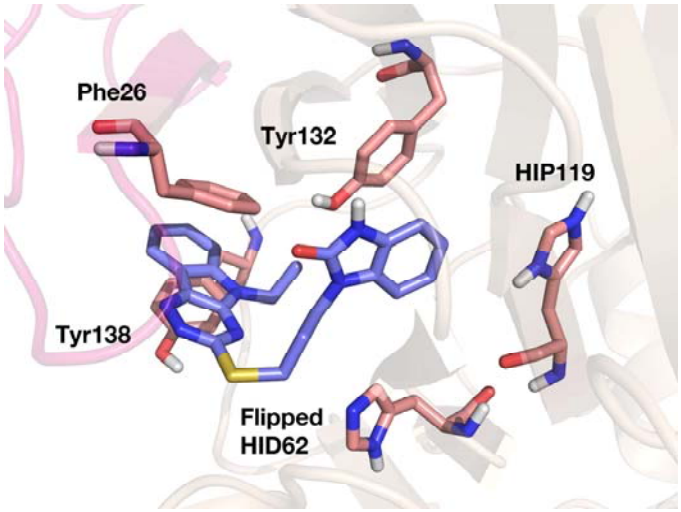


Figure 2-3. Predicted interaction of the initial hit compound SID7975595 with flipped HID62 and HIP119 in receptor model 23. Generally, the actives do not have strong interactions with His62 or His119, yet varying histidine protonation states have a profound effect on the ranked results. Favorable interactions are observed with other binding site residues, such as Tyr132 and Tyr138 as depicted here.

Interestingly, the actives generally do not achieve polar interactions with His62 and His119. As shown in Figure 2-3, the carbonyl oxygen and two benzimidazolone nitrogens of SID7975595 face away from His62 and His119. The direction of aromatic hydrogens of the actives is often unable to participate in hydrogen bond networks with the two histidines. Nevertheless, different protonation and rotameric states of these histidines do affect the VS results through their interactions with the decoys.

Assessment of differences in ranking

It is not uncommon that only the top 1% of compounds screened can be tested experimentally in a VS study, due to the limited resources. Therefore, the enrichment factor $(EF)^{1\%}$ metric, which reflects the database enrichment performance in the top 1% (*i.e.*, 20 docked compounds) of a library, becomes particularly relevant in assessing the predictive power of VS. The $EF^{1\%}$ ranges from 0 to 80 for 36 receptor models (Table 2-1), indicating that the VS results are sensitive to the protonation and rotameric states of His62 and His119 of RmlC. Nevertheless, 28 out of 36 receptors rank more than eight actives within the top 10% in the VS, as reflected by the $EF^{10\%}$ (Table 2-1), suggesting that most

receptors are able to distinguish the actives and decoys when a larger portion (10%) of the database is considered. The EF results also suggest that the receptor models with HIP62 or HIP119 tend to have poor enrichment performance, likely due to the extensive hydrogen bond networks with the decoys, as discussed later.

The area under the receiver operating characteristic curve (AUC) for each receptor model was evaluated to report the enrichment performance of models upon different protonation and rotameric states of His62 and His119. As shown in Figure 2-4A and Table 2-1, the AUC values of all receptor models range from 0.868 to 0.996, indicating an overall good predictive performance (an AUC of 0.5 corresponds to no differentiation between the actives and decoys). In general, the AUC result is complementary to the EF assessment for receptor predictive performance. Summarizing Table 2-1, Figure 2-4C shows how the range of the receptor performance depends on the two histidine protonation and rotameric states. Considering the 25% to 75% range of the AUCs (Figure 2-4C, indicated by the thicker lines), the His62 models show a larger variation across His119 states. The His119 models, on the other hand, have a more consistent performance regardless of the protonation states of His62, with the exception of HIP state. This indicates that different protonation states of His62 have a smaller influence than those of His119 on the receptor performance in our screening.

Table 2-1. AUC values of 36 receptor models with different protonation and rotameric states.

Receptor Model	His62	His119	AUC	$EF^{1\%}$	$EF^{10\%}$
1	HIE	HIE	0.942	60	8
2	HIE	Flipped HIE	0.992	60	10
3	HIE	HID	0.868	30	5
4	HIE	Flipped HID	0.961	50	9
5	HIE	HIP	0.875	30	6
6	HIE	Flipped HIP	0.996	80	10
7	Flipped HIE	HIE	0.963	70	9
8	Flipped HIE	Flipped HIE	0.945	30	7
9	Flipped HIE	HID	0.991	60	10
10	Flipped HIE	Flipped HID	0.938	50	6
11	Flipped HIE	HIP	0.918	0	8
12	Flipped HIE	Flipped HIP	0.963	60	8
13	HID	HIE	0.989	30	10
14	HID	Flipped HIE	0.991	40	10
15	HID	HID	0.989	30	10
16	HID	Flipped HID	0.990	20	10
17	HID	HIP	0.916	0	8
18	HID	Flipped HIP	0.991	20	10
19	Flipped HID	HIE	0.992	80	10
20	Flipped HID	Flipped HIE	0.957	40	8
21	Flipped HID	HID	0.969	50	9
22	Flipped HID	Flipped HID	0.987	50	10
23	Flipped HID	HIP	0.981	40	10
24	Flipped HID	Flipped HIP	0.988	60	10
25	HIP	HIE	0.971	40	8
26	HIP	Flipped HIE	0.991	30	10
27	HIP	HID	0.982	0	10
28	HIP	Flipped HID	0.933	40	9
29	HIP	HIP	0.869	0	7
30	HIP	Flipped HIP	0.936	30	9
31	Flipped HIP	HIE	0.945	30	9
32	Flipped HIP	Flipped HIE	0.933	10	6
33	Flipped HIP	HID	0.964	40	8
34	Flipped HIP	Flipped HID	0.917	0	6
35	Flipped HIP	HIP	0.950	0	8
36	Flipped HIP	Flipped HIP	0.969	20	9

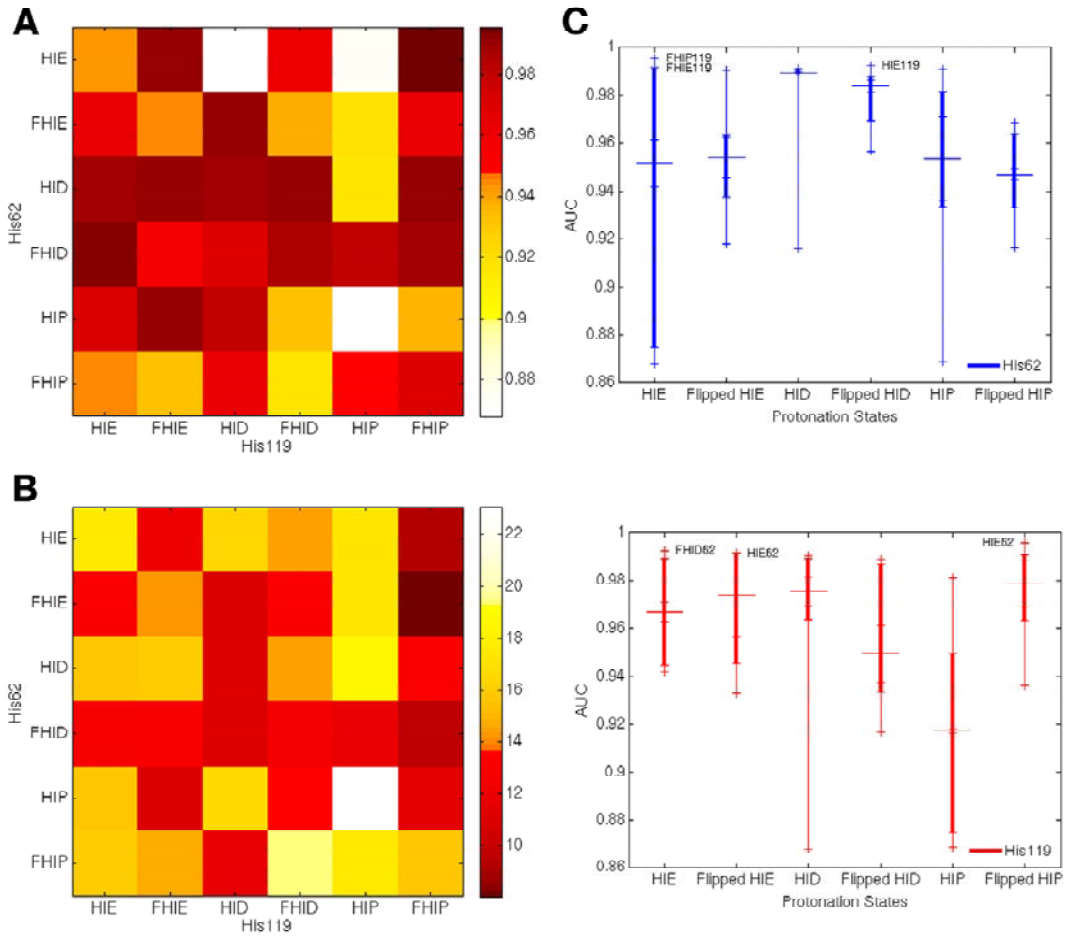


Figure 2-4. (A) AUC values of 36 receptor models. Protonation and rotameric states are marked for each histidine. Flipped states are marked with the letter F. Darker color indicates higher AUC and better predictive performance of the corresponding receptor model. (B) Average hydrogen bond percentage of the top 1 % compounds in 36 VS runs. Protonation and rotameric states are marked for each histidine. Lighter color indicates higher hydrogen bond percentage, with % unit for the colorbar. The R^2 for the correlation between the AUCs and average hydrogen bond percentage for each VS run is 0.42 (see Figure 2-S4 for the scatter plot). (C) Receptor performance dependence on His62 (top) and His119 (bottom). The median of the AUC values of each protonation state is shown with large horizontal line. The small ticks in each histidine model mark six different protonation states of the other histidine. The thicker vertical lines represent 25–75 % range of the AUCs. The best receptor models are shown explicitly with the models' protonation states.

A stronger dependence of enrichment on the protonation states of His119 is observed in the HIE62 and HIP62 models. With HIE62 state, the models with flipped HIP119 (model 6) and flipped HIE119 (model 2) yield the highest receptor performance. Models 3 and 5 with HID119 and HIP119, respectively, lead to the worst enrichment. In examining why HIE62 state has the largest variation in AUCs, one finds that His62 has either pi-pi stacking or no interactions with ligands, and makes only a few hydrogen bonds with high-ranking decoys. Therefore, the receptor performance depends on the interaction of His119 with the decoys. This is also seen when examining the broad performance range of AUCs of the HIP62 models. The hydrogen bond networks with the decoys will be discussed later in the following section.

In order to evaluate the statistical significance of difference of the AUC values between a pair of receptor models, we performed a two-sided *p*-test at the 95% level on the null hypothesis that the pair has statistically comparable AUC values, against the alternative hypothesis that their difference in the AUC values and predictive power is statistically meaningful. On average, the receptors have more than 16 *p*-values less than 0.05, demonstrating the sensitivity of VS on histidine protonation and rotameric states. As one might expect, the receptors with the most significant differences correspond to the models with the highest (model 6) or lowest AUC values (models 3, 29, and 5). Model 6 is statistically better at ranking the actives over decoys than 26 other receptors in the ensemble. Models 3, 29, and 5 are distinguishably worse at ranking the actives than 29, 25, and 31 other receptors, respectively.

Quantitative analysis of the hydrogen bonding interactions was carried out for the top 1% (20 docked compounds) of each VS result to account for the abundant hydrogen bonding interactions with the binding site residues often observed with the decoys. The results indicate an inverse correlation between the hydrogen bonding contribution and receptor performance. Figure 2-4B shows the average hydrogen bond percentage of each receptor model for the top 1% docked compounds. Hydrogen bond percentage is defined as the portion of the Glide XP hydrogen bonding term in the total docking score. Comparison of Figure 2-4A and 4B reveals the inverse relationship between the hydrogen bond percentage and AUC with a R^2 of 0.42 ($y = -56.18x + 67.95$, the correlation is plotted in Figure 2-S4). The inverse relationship is commonly observed with models with HIP119, flipped HIP119, or HID62,

where the high hydrogen bond percentage resulted in poor enrichment. For example, receptor model 29 with HIP62 and HIP119, where both histidines presenting active site facing hydrogen bond donors, has one of the worst AUCs due to the high percentage of hydrogen bonds in the top hits.

Notably, the hydrogen bonding potential of His119 often determines the receptor performance. For example, the model with HID62 and HIP119 was an outlier among the HID62 models in Figure 2-4C, with noticeably low enrichment compared to the overall good performance of the other five HID62 models. The HID62 models have a high median AUC of 0.989, despite the frequent hydrogen bonding to the decoys from HID62. This is due to His119 states achieving few hydrogen bonding interactions with the decoys. Only with HIP119 state does the HID62 model make hydrogen bonds with a number of decoys, resulting in the relatively low AUC. This observation agrees with the stronger dependence of the receptor performance on the protonation states of His119, as discussed above. Figure 2-S4 describes the AUC distribution and hydrogen bond percentage along with the direction of hydrogen bond donor or acceptor from two histidines facing the receptor.

Above analyses highlight the hindering effect of hydrogen bonding to the decoys on the predictive power of VS, due to the various coordinates of two histidines with different protonation and rotameric states. The scatter of the observed correlation with the R^2 of 0.42 is likely attributed to several causes, including the chemical nature of the decoy dataset, as well as the slight differences in geometry of each receptor upon minimization in the initial preparation of the protein. By clearly showing the sensitivity of virtual screening results on different protonation and rotameric states of histidines in the active site, we emphasize that care should be taken when preparing the atomic coordinates of a receptor for VS, particularly considering the general properties of the ligands being screened. This includes taking into account the hydrogen bonding to the co-crystal ligand and its effect on protein preparation, as well as a comprehensive analysis of proximal hydrogen bond networks. This is usually achieved by examining the results from widely used pK_a prediction software packages, and to this point, we have compared results from different packages relative to our VS results and discuss them further.

Docking of the decoys

Various factors lead to differences in ranking across the receptors, particularly with respect to the decoys. Generally, the decoys that ranked higher than the actives were of high molecular weight and had more potential to have hydrogen bonds with the receptor. In this section, we further analyze the frequent interaction patterns observed between the decoys and receptor, with a focus on the receptor models with poor enrichment.

Decoys tend to have larger molecular weight and more ring structures than the actives (Table 2-2). This results in the decoys ranking higher, due to hydrophobic interactions in the absence of hydrogen bonds to the receptor. Figure 2-5A shows the hydrophobic interactions achieved through the large inactive compound 16952387 in receptor model 19. This compound is often ranked within the top five in many VS runs for its substantial pi-pi stacking interactions with Phe26, Tyr132, and Tyr138. This trend is frequently observed in virtual screening where larger molecules rank better as a result of extensive interactions with the receptor.[46]

Table 2-2. Comparison of molecular weight, number of hydrogen bond donor, and number of hydrogen bond acceptor for the actives and decoys.

	Actives	**Decoys**
Average Molecular Weight (g/mol)	417.69	353.85
Stdev. of Molecular Weight (g/mol)	27.07	80.85
Average Number of Hydrogen Bond Donor	1.1	1.04
Average Number of Hydrogen Bond Acceptor	5.8	5.89

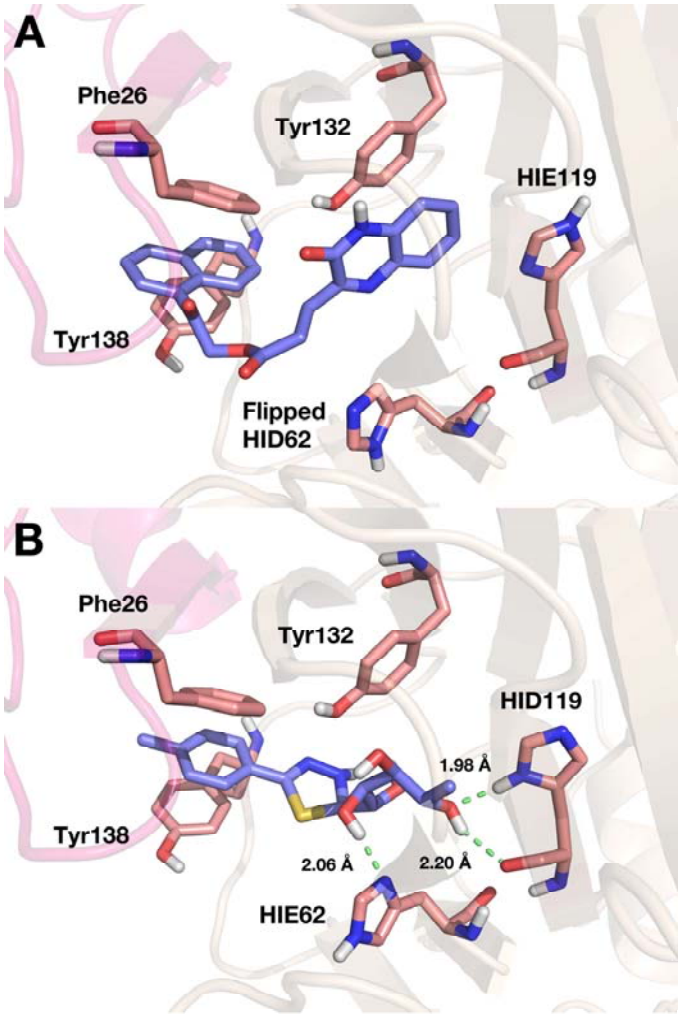


Figure 2-5. (A) Interaction of the inactive compound 16952387 with flipped HID62 and HIE119 in receptor model 19. The compound has no interaction with either histidine. Pi–pi stacking interactions with Phe26 from chain B, Tyr132, and Tyr138 contribute to its high rank, along with hydrogen bonds with Arg23, Arg59, Arg170, and Ser51 (not shown). (B) Interaction of the inactive compound 17388064 with HIE62 and HID119 in receptor model 3. Both histidines provide hydrogen bonds to the compound.

The enrichment performance is particularly low for the receptors providing abundant hydrogen bond networks to the decoys. Interactions through His62 and His119 were not widely observed for the actives, and therefore compounds with larger enthalpic contributions erroneously rank more favorably. An example shown in Figure 2-5B depicts the interaction of inactive compound 17388064 in receptor model 3 (AUC 0.868), ranked as first. In this receptor, which is the worst at ranking compounds based on AUC, compound 17388064 forms two hydrogen bonds with two histidines, one between its

hydroxyl hydrogen and the δ-nitrogen of HIE62 and the other between its hydroxyl oxygen and the hydrogen on δ-nitrogen of HID119. This compound has five hydrogen bond donors and nine acceptors, a large number compared to the respective averages of those of the decoys and actives (Table 2-2). Therefore, with a high hydrogen bonding contribution of 34.7 ± 6.6% to the total score, this decoy compound is frequently observed to form at least one hydrogen bond with either of the two histidines, thereby achieving high ranks in multiple VS runs.

Two other receptor models, model 29 with HIP62 and HIP119 and model 5 with HIE62 and HIP119, show similar interaction patterns to decoys as model 3. These three models have the lowest AUC values, with an average of o.870 among them. As discussed above, their AUC values differ significantly from other receptors, reflecting the subtle relationship between hydrogen bonds achieved through His62 and His119 and poor enrichment. An additional figure describing the hydrogen bond networks between the decoys and receptors is provided in Supporting Information 5.

pK_a Prediction for His62 and His119

Our results clearly demonstrate the sensitivity of virtual screening on histidine protonation and rotameric states. In many computational biophysical studies, the protonation states of the titratable residues are determined using various pK_a prediction programs. To assess the performance of these programs to identify the receptor model with the best predictive power in docking, we compared the pK_a prediction results of His62 and His119 from PROPKA, Maestro, H++, and MCCE, as shown in Table 2-3 with the calculated pK_a values.

Table 2-3. Comparison of the predicted pK_a values and protonation states of His62 and His119 of RmlC by commonly used software.

		RmlC-TRH complex		RmlC without TRH	
		His62	**His119**	**His62**	**His119**
PROPKA 3.1	pK_a	4.24	5.8	5.16	6.12
	Protonation	Neutral	Neutral	Neutral	Neutral
Maestro with PROPKA	pK_a	7.09	5.6	4.99	6.12
	Protonation	HIP	HIE	HIE	HIE
Maestro with Epik	Protonation	HIP	HIP	HIP	HIP
H++	pK_a	3	<0.0	3	<0.0
	Protonation	HIE	HIE	HIE	HIE
MCCE	pK_a	<0.0	7.201	2.771	1.347
	Protonation	Neutral	HIP	Neutral	HIP

First, PROPKA 3.1 predicts that both His62 and His119 are neutral regardless of the presence of TRH during preparation. The program, however, cannot assign rotameric states of histidines. Therefore, a state of HID, flipped HID, HIE, or flipped HIE must be determined manually. Similar to PROPKA, the program H++, which uses a single-structure continuum electrostatics, also finds both histidines to be neutral, although the predicted pK_a values are different from those from PROPKA. The program MCCE, which is based on multi-conformation continuum electrostatics, predicts His62 to be neutral while His119 to be protonated.

Next, we used the Protein Preparation Wizard in Maestro to calculate pK_a of His62 and His119 with and without TRH. Note that Maestro is able to vary rotameric states, whereas PROPKA cannot. A recent update enables Maestro to employ PROPKA in its pK_a prediction instead of Epik. With Epik, Maestro predicts both His62 and His119 in doubly protonated states, regardless of the presence of TRH. Interestingly, the receptor model that corresponds to this di-doubly protonated state has the worst predictive power with an AUC of 0.869. When PROPKA is used, HIP62 and HIE119 are predicted for the protein-TRH complex and HIE62 and HIE119 for the apo protein. These two predictions by PROPKA in Maestro correspond to the models of moderate enrichment performance, with AUCs of 0.971 for model 25 (HIP62 and HIE119) and 0.942 for model 1 (HIE62 and HIE119), respectively.

Given that the above predictions made by different software vary significantly from each other, caution should be taken when using these results as a guideline to prepare a protein for virtual screening. Without intimate knowledge of the true protonation state of the receptor as well as the ligands being screened, it is difficult to address this problem. Therefore, we suggest that a small-scale analysis, like one performed in this study, and comparison with experimental data, if available, could provide a more accurate description of protonation and rotameric states of the titratable residues in protein receptors for future larger-scale screenings. Alternatively, a model that includes explicit incorporation of alternative side chain protonation and rotameric states during docking, potentially with information stored in the grid as exists for rotatable hydroxyls and thiols in Glide, may be worth pursuing. Examination of the results with respect to the protonation states and rankings based on interactions with histidines should be carefully examined before proceeding to experimental testing.

Additionally, receptor flexibility will likely affect the protonation states of the ionizable residues. While this was not explicitly studied here, aside from minimization of each receptor after assigning protonation states, protein flexibility is clearly important for drug design and development.[47, 48] Coupling of conformational and protonation space becomes quickly intractable with physical methods such as those described here, but enhanced sampling methods show promise in tackling such difficulties.[49] This includes constant pH molecular dynamics simulations, for which the pH is an external thermodynamic variable, used for blind prediction of pK_a values of the titratable residues.[50-52] Effectively applying results from these simulations to molecular design is an ongoing area of interest. Equilibrium ensembles from such simulations can be used in conjunction with docking as an application of relaxed complex scheme, where virtual screening is conducted with an ensemble of structures with different protonation states, to improve the enrichment results.[53] Taking receptor flexibility into account in the target preparation will lead to broader sampling of conformational and protonation space, thus enhancing the performance of VS.

Conclusions

Protein-ligand recognition is of central importance in structure-based drug discovery. Correctly accounting for the chemical environment surrounding the ligand is imperative for characterizing and predicting the molecular interactions. The possible effect of the various protonation and rotameric states of the ionizable residues of the receptor on virtual screening (VS) is critical yet often overlooked. In this study, we thoroughly examined the influence of the protonation and rotameric states of histidine on the predictive power of the docking protocol for drug discovery.

A histidine can adopt three different forms depending on the net charge and the location of proton(s). Due to common ambiguities in X-ray crystal structures, three additional states may be generated through flipping of the imidazole ring. In this work, we performed a VS study on the *M. tuberculosis* enzyme RmlC to investigate the effect of six protonation and rotameric states of histidines. We systematically examined the contribution of hydrogen bonding interactions provided by two histidines in the active site, His62 and His119. The predictive performance of receptors was assessed by quantitatively analyzing enrichment factors and area under the receiver operating characteristic curve. We showed that the hydrogen bond networks involving His62 and His119 are important in the interaction between the co-crystal ligand TRH and active site, validating the significance of accurate description of protonation and rotameric states of the two histidines in VS. We compared the typical patterns of interactions achieved with the active site residues observed for the active compounds and decoys; whereas the actives often involve only hydrophobic interactions, the high-ranking decoys are erroneously enriched by additional hydrogen bonds provided by His62 and His119. Our analyses reveal the sensitivity of virtual screening on protonation, ionization, and rotameric states of active site histidines. We recommend a priori analysis of receptor-ligand hydrogen bonding interactions, in addition to the usage of protonation assignment software packages, to prepare a receptor for virtual screening. Systematically assessing binding site protonation state effects before conducting a large virtual high-throughput screening, beyond empirical state prediction, may therefore result in enrichment gains.

Acknowledgements

This work has been supported in part by the National Science Foundation, the National Institutes of Health, Howard Hughes Medical Institute, Center for Theoretical Biological Physics, the National Biomedical Computation Resource, and the NSF supercomputer centers.

Chapter 2 is a minimally modified reprint of the material as it appears in Meekyum Olivia Kim, Sara E. Nichols, Yi Wang, and J. Andrew McCammon, "Effects of histidine protonation and rotameric states on virtual screening of *M. tuberculosis* RmlC," *Journal of Computer-Aided Molecular Design*, 2013. The dissertation author was the primary investigator and author of this paper.

Supporting Information

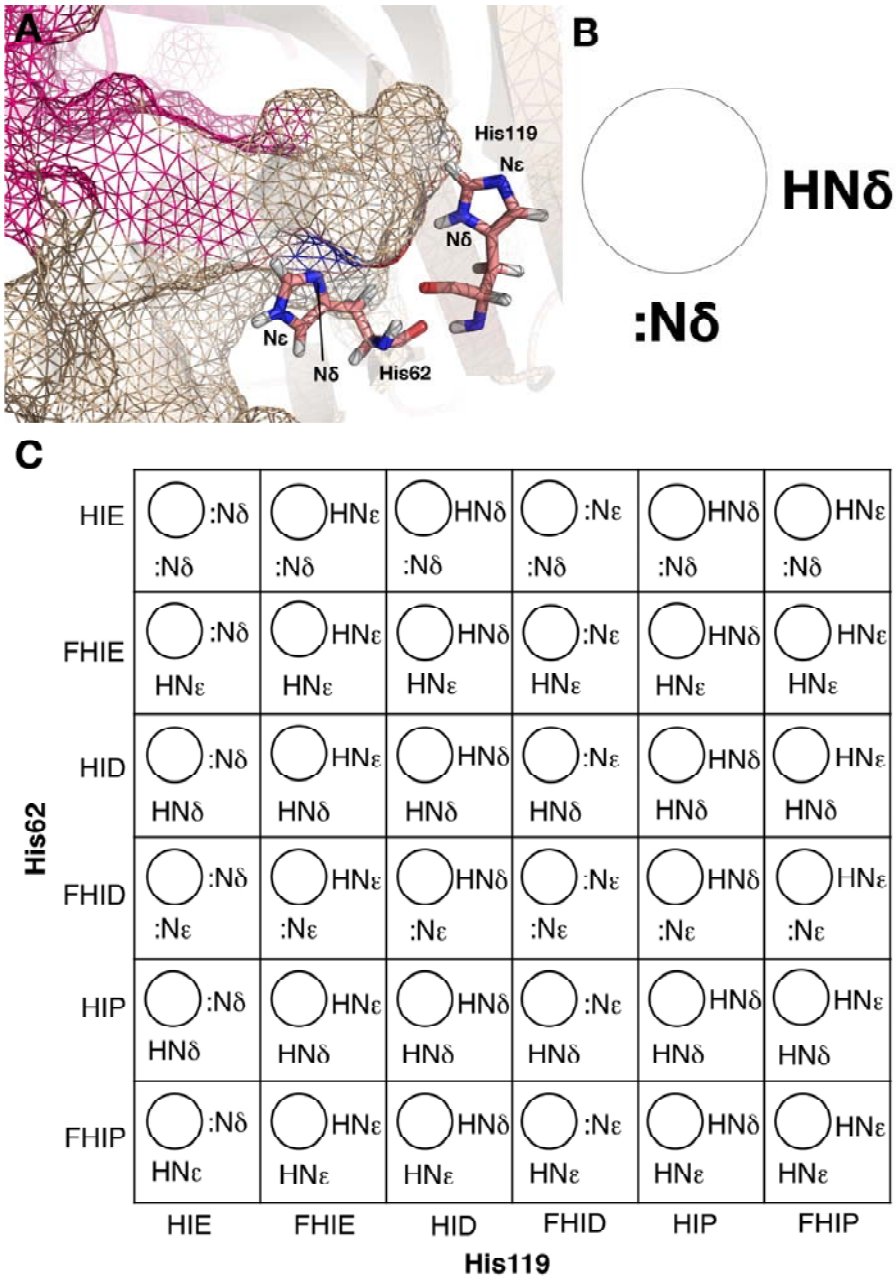


Figure 2-S1. (A) Location of His62 and His119 with respect to the active site of RmlC. Atom types of nitrogen of histidine are shown as either Nδ or Nε and the surface is represented as wireframes. (B) Schematic description of hydrogen and nitrogen of two histidines in (A). The active site is shown as a circle and His62 is located below the circle and His119 right next to it. Colon (:) represents a lone pair of electrons on nitrogen. (C) Schematic description of hydrogen and nitrogen of two histidines in 36 receptor models following (B).

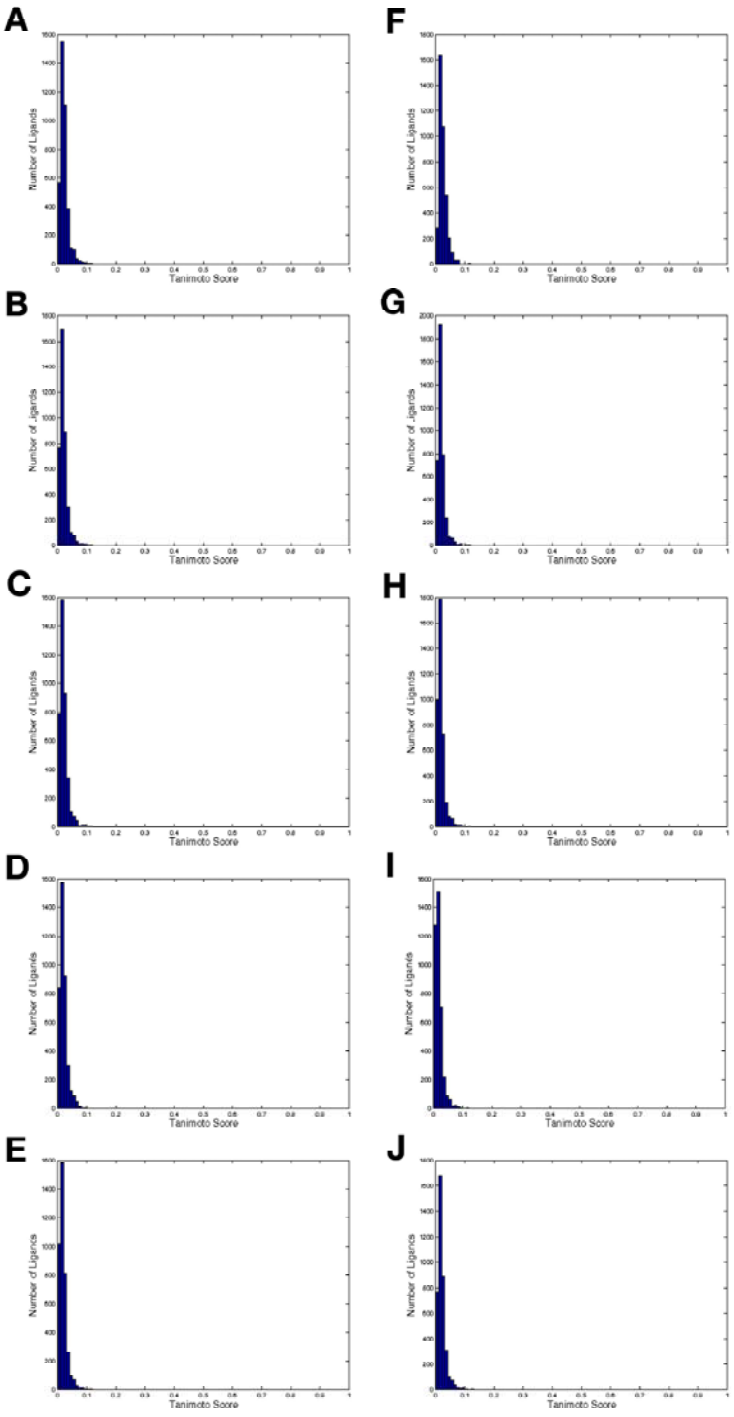


Figure 2-S2. Distribution of Tanimoto scores of the library of total 3,934 compounds against each of ten active compounds as a reference. The active compound used to calculate Tanimoto score is (A) 77070; (B) 77071; (C) 77072; (D) 77073; (E) 77074; (F) 78531; (G) 78532; (H) 78533; (I) SID7972845; and (J) SID7975595.

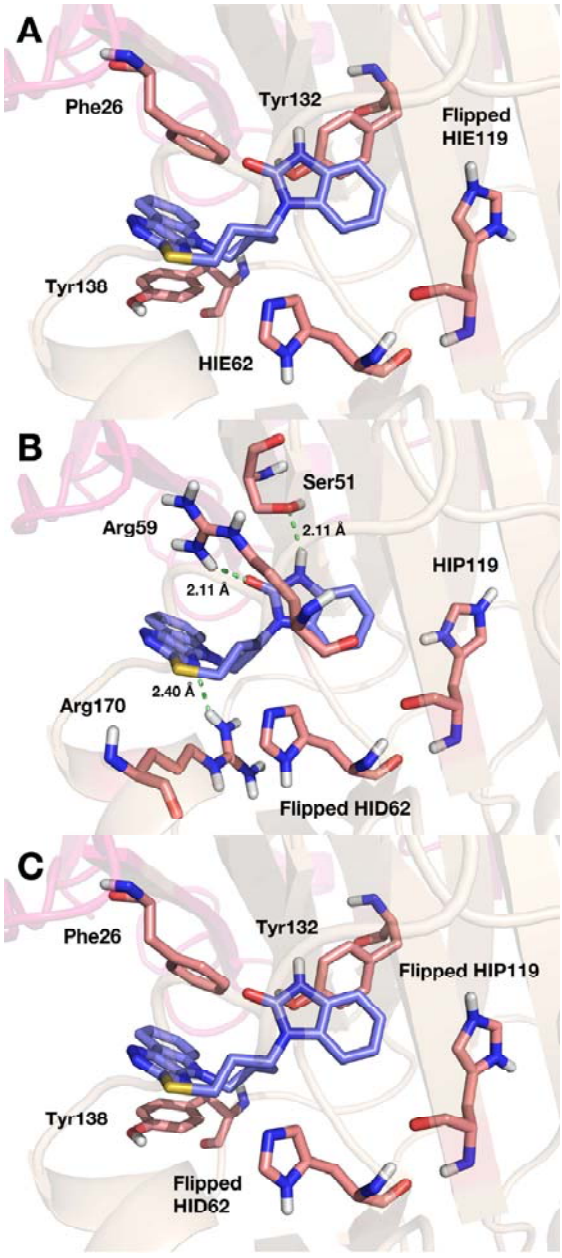


Figure 2-S3. (A) Predicted interaction of the active compound SID7975595 with HIE62 and flipped HIE119 in receptor model 2 (AUC 0.992). The compound mainly has hydrophobic interactions. (B) Predicted hydrogen bonding networks between the active compound SID7975595 and Arg59, Arg170, and Ser51 in the active site in receptor model 23 with flipped HID62 and HIP119 (AUC 0.981). (C) Interaction of the active compound 77074 with flipped HID62 and flipped HIP119 in receptor model 24 (AUC 0.988). The compound 77074, where the ethyl group attached to the nitrogen on the tricyclic ring of the compound SID7975595 has been substituted by an allyl group, has the lower IC_{50}. Hydrophobic pi-pi stacking comprises the main interaction between the compound and the active site residues.

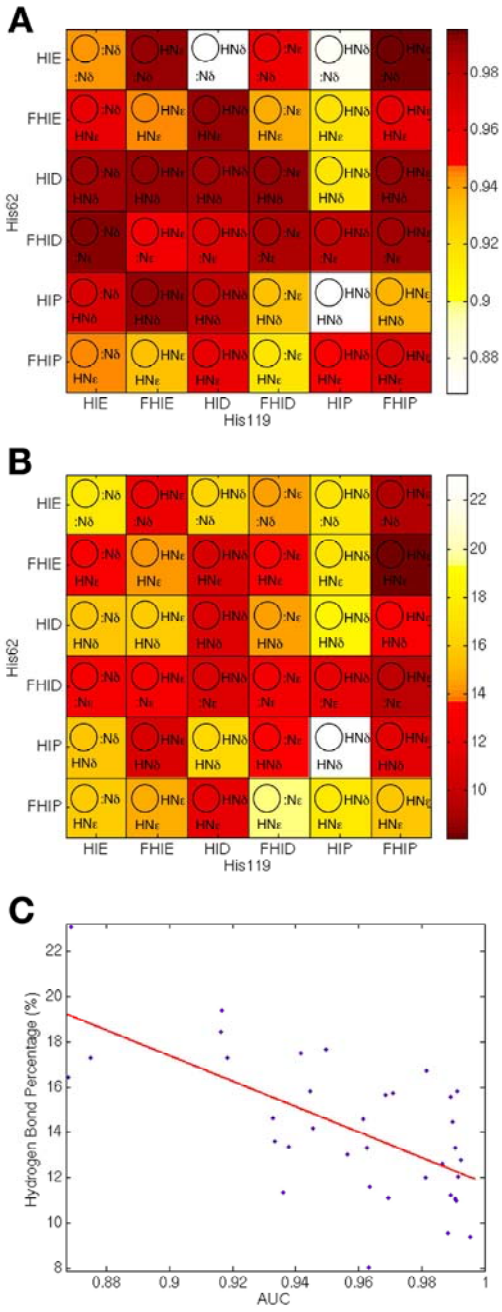


Figure 2-S4. (A) AUC values of 36 receptor models (Figure 2-4A) shown with hydrogen bond donor or acceptor of two histidines for each model as described in Figure 2-S1. (B) Average hydrogen bond percentage of the top 1% compounds in 36 VS runs (Figure 2-4B). Hydrogen bond donor or acceptor from two hisidines is shown together in the same way as in (A). (C) Scatter plot showing the correlation between the AUCs and average hydrogen bond percentage for the top 1% compounds of each VS run. The correlation is observed as $y = -56.18x + 67.95$ with the R^2 of 0.42.

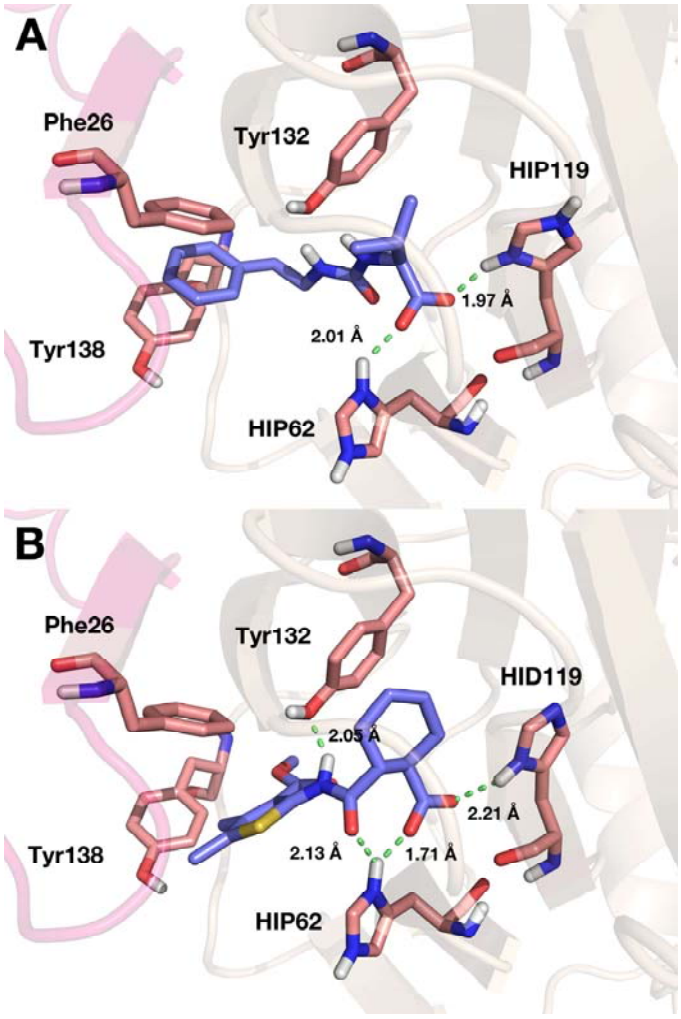


Figure 2-S5. (A) Interaction of the inactive compound 14741063 with HIP62 and HIP119 in receptor model 29 (docking score -8.513, AUC 0.869). Both hydrogens on δ-nitrogens of HIP62 and HIP119 accept lone electron pairs from carboxylic acid group of the compound. Phe26 and Tyr132 interact with the compound via pi-pi stacking interaction. (B) Interaction of the inactive compound 14736762 with HIP62 and HID119 in receptor model 27 (AUC 0.982). Hydrogen on δ-nitrogen of HIP62 makes two hydrogen bonds with the amide oxygen and carboxylic oxygen of the compound while that of HIP119 also forms a hydrogen bond with carboxylic oxygen. Additional hydrogen bonding networks include the one between the hydroxyl oxygen of Tyr132 and the amide nitrogen of the compound.

References

1. Knox, A. J. S.; Meegan, M. J.; Carta, G.; Lloyd, D. G. Considerations in Compound Database Preparation - "Hidden" Impact on Virtual Screening Results. *J. Chem. Inf. Model.* **2005,** *45*, 1908-1919.

2. Ten Brink, T.; Exner, T. E. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *J. Chem. Inf. Model.* **2009,** *49*, 1535-1546.

3. Martin, Y. C. Let's Not Forget Tautomers. *J. Comput. Aided Mol. Des.* **2009,** *23*, 693-704.

4. Kalliokoski, T.; Salo, H. S.; Lahtela-Kakkonen, M.; Poso, A. The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2009,** *49*, 2742-2748.

5. Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein pKa Values. *Proteins* **2005,** *61*, 704-721.

6. Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein-Ligand Complexes. *Proteins* **2008,** *73*, 765-783.

7. Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011,** *7*, 525-537.

8. Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011,** *7*, 2284-2295.

9. Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating pKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005,** *33*, 368-371.

10. Myers, J. B.; Grothaus, G.; Narayana, S.; Onufriev, A. V. A Simple Clustering Algorithm Can Be Accurate Enough for Use in Calculations of Pks in Macromolecules. *Proteins* **2006,** *63*, 928-938.

11. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling Simulation. *Nucleic Acids Res.* **2012,** *40*, 537-541.

12. Alexov, E.; Gunner, M. R. Incorporating Protein Conformational Flexibility into pH Titration Calculations: Results on T4 Lysozyme. *Biophys. J.* **1997,** *74*, 2075-2093.

13. Georgescu, R. E.; Alexov, E.; Gunner, M. R. Combining Conformational Flexibility and Continuum Electrostatics for Calculating pKa's in Proteins. *Biophys. J.* **2002,** *83*, 1731-1748.

14. Song, Y.; Gunner, M. R. Mcce2: Improving Protein pKa Calculations with Extensive Side Chain Rotamer Sampling. *J. Comput. Chem.* **2009,** *30*, 2231-2247.

15. Park, M. S.; Gao, C.; Stern, H. A. Estimating Binding Affinities by Docking/Scoring Methods Using Variable Protonation States. *Proteins* **2011,** *79*, 304-314.

16. Sivendran, S.; Jones, V.; Sun, D.; Wang, Y.; Grzegorzewicz, A. E.; Scherman, M. S.; Napper, A. D.; McCammon, J. A.; Lee, R. E.; Diamond, S. L., et al. Identification of Triazinoindol-Benzimidazolones as Nanomolar Inhibitors of Mycobacterium tuberculosis Enzyme Tdp-6-Deoxy-D-Xylo-4-Hexopyronosid-4-Ulose 3,5-Epimerase (RmlC). *Bioorg. Med. Chem.* **2010,** *18*, 896-908.

17. Cleland, W. W. Low-Barrier Hydrogen Bonds and Enzymatic Catalysis. *Arch. Biochem. Biophys.* **2000,** *382*, 1-5.

18. Tu, C.; Silverman, D. N.; Forsman, C.; Jonsson, B. H.; Blindskog, S. Role of Histidine 64 in the Catalytic Mechanism of Human Carbonic Anhydrase Ii Studied with a Site-Specific Mutant. *Biochemistry* **1989,** *28*, 7913-7918.

19. Ren, X.; Tu, C.; Laipis, P. J.; Silverman, D. N. Proton Transfer by Histidine 67 in Site-Directed Mutants of Human Carbonic Anhydrase III. *Biochemistry* **1995,** *34*, 8492-8498.

20. Eigen, M. Proton Transfer, Acid-Base Catalysis, and Enzymatic Hydrolysis. Part I: Elementary Processes. *Angew. Chem. Int. Ed. Engl.* **1964,** *3*, 1-19.

21. Stockel, J.; Safar, J.; Wallace, A. C.; Cohen, F. E.; Prusiner, S. B. Prion Protein Selevtively Binds Copper (Ii) Ions. *Biochemistry* **1998,** *37*, 7185-7193.

22. Olson, J. S.; Mathews, A. J.; Rohlfs, R. J.; Springer, B. A.; Egeberg, K. D.; Sligar, S. G.; Tame, J.; Renuad, J. P.; Nagai, K. The Role of Distal Histidine in Myoglobin and Haemoglobin. *Nature* **1998,** *336*, 265-266.

23. Li, S.; Hong, M. Protonation, Tautomerization, and Rotameric Structure of Histidine: A Comprehensive Study by Magic-Angle-Spinning Solid-State NMR. *J. Am. Chem. Soc.* **2011,** *133*, 1534-1544.

24. Gluster, J.; Lewis, M.; Rossi, M. *Crystal Structure Analysis for Chemists and Biologists*. VCH Publishers: New York, 1994.

25. Li, X.; Jacobson, M. P.; Zhu, K.; Zhao, S.; Friesner, R. A. Assignment of Polar States for Protein Amino Acid Residues Using an Interaction Cluster Decomposition Algorithm and Its Application to High Resolution Protein Structure Modeling. *Proteins* **2006,** *66*, 824-837.

26. Ma, Y.; Stern, R. J.; Scherman, M. S.; Vissa, V. D.; Yan, W.; Jones, V. C.; Zhang, F.; Franzblau, S. G.; Lewis, W. H.; McNeil, M. R. Drug Targeting Mycobacterium tuberculosis Cell Wall Synthesis: Genetics of dTDP-Rhamnose Synthetic Enzymes and Development of a Microtiter Plate-Based Screen for Inhibitors of Conversion of dTDP-Glucose to dTDP-Rhamnose. *Antimicrob. Agents Chemother.* **2001,** *45*, 1407-1416.

27. Dong, C.; Major, L. L.; Srikannathasan, V.; Errey, J. C.; Giraud, M.; Lam, J. S.; Graninger, M.; Messner, P.; McNeil, M. R.; Field, R. A., et al. RmlC, a C3' and C5' Carbohydrate Epimerase, Appears to Operate Via an Intermediate with an Unusual Twist Boat Conformation. *J. Mol. Biol.* **2007,** *365*, 146-159.

28. *Schrödinger Suite 2011 Protein Preparation Wizard; Epik Version 2.2; Impact Version 5.7; Prime Version 3.0*, Schrödinger, LLC: New York, NY, 2011.

29. *Glide*, Version 5.7; Schrödinger, LLC: New York, NY, 2011.

30. *Ligprep*, Version 2.5; Schrödinger, LLC: New York, NY, 2011.

31. Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; R., T. M.; Uchiyama, M. Epik: A Software Program for pKa Prediction and Protonation State Generation for Druglike Molecules. *J. Comput. Aided Mol. Des.* **2007,** *21*, 681-691.

32. *Epik*, Version 2.2; Schrödinger, LLC: New York, NY, 2011.

33. *Maestro, Version 9.3*, Schrödinger, LLC: New York, NY, 2012.

34. Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and Comparison of 2d Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.* **2010,** *29*, 157-170.

35. Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-Scale Systematic Analysis of 2d Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model.* **2010,** *50*, 771-784.

36. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M., et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004,** *47*, 1739-1749.

37. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004,** *47*, 1750-1759.

38. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006,** *49*, 6177-6196.

39. Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005,** *45*, 1369-1375.

40. Fawcett, T. An Introduction to Roc Analysis. *Pattern Recognit. Lett.* **2006,** *27*, 861-874.

41. Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichment. *J. Chem. Inf. Model.* **2010,** *50*, 511-524.

42. Nicholls, A. How Do We Know and When Do We Know It? *J. Comput. Aided Mol. Des.* **2008,** *22*, 239-255.

43. *Matlab*, Version 7.12.0.635; The MathWorks, Inc.: Natick, MA, 2011.

44. *Maestro*, Version 9.2; Schrödinger, LLC: New York, NY, 2011.

45. Perrin, D. D.; Dempsey, B.; Sergeant, E. P. pKa Prediction for Organic Acids and Bases. *Chapman and Hall, London* **1981**.

46. Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D. Consideration of Molecular Weight Dring Compound Selection in Virtual Target-Based Database Screening. *J. Chem. Inf. Comput. Sci.* **2003,** *43*, 267-272.

47. Carlson, H. A.; McCammon, J. A. Accommodating Protein Flexibility in Computational Drug Design. *Mol. Pharmacol.* **2000,** *57*, 213-218.

48. Carlson, H. A. Protein Flexibility and Drug Design: How to Hit a Moving Target. *Curr. Opin. Chem. Biol.* **2002,** *6*, 447-452.

49. Sinko, W.; Lindert, S.; McCammon, J. A. Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem. Biol. Drug Des.* **2013,** *81*, 41-49.

50. Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH Molecular Dynamics Using Stochastic Titration. *J. Chem. Phys.* **2002,** *117*, 4184-4200.

51. Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004,** *25*, 2038-2048.

52. Williams, S. L.; Blachly, P. G.; McCammon, J. A. Measuring the Successes and Deficiencies of Constant pH Molecular Dynamics: A Blind Prediction Test. *Proteins* **2011,** *79*, 2281-2288.

53. Amaro, R. E.; Baron, R.; McCammon, J. A. An Improved Relaxed Complex Scheme for Receptor Flexibility in Computer-Aided Drug Design. *J. Comput. Aided Mol. Des.* **2008,** *22*, 693-705.

Chapter 3

A molecular dynamics investigation of *Mycobacterium tuberculosis* prenyl synthases: Conformational dynamics and implications for computer-aided drug discovery

Abstract

With the increasing rise in antibiotic resistance, there is considerable interest in discovering new drugs active against new targets. Here, we investigated the structural dynamics of three isoprenoid synthases from *Mycobacterium tuberculosis* by performing molecular dynamics (MD) simulations with a view to discovering new drug leads. Two of the enzymes, *E,Z*-farnesyl diphosphate synthase (*E,Z*-FPPS, Rv1086c) and *E,Z*-decaprenyl diphosphate synthase (*E,Z*-DPPS, Rv2361c), are involved in bacterial cell wall biosynthesis while the third, tuberculosinol/tuberculosinyl adenosine synthase (Rv3378c), is involved in virulence factor formation. The MD results for these three enzymes were then compared with previous results on undecaprenyl diphosphate synthase (UPPS) from *E. coli* by means of active site volume fluctuation and principal component analyses. In addition, an analysis of the binding mechanisms of prenyl diphosphates to *E,Z*-FPPS, *E,Z*-DPPS, and UPPS utilizing the new MD results is reported. We also screened libraries of inhibitors against *E,Z*-DPPS, finding ~ 1 μM inhibitors, and used the receiver operating characteristic-area under the curve (ROC-AUC) method to test the predictive

power of X-ray and MD-derived *E,Z*-DPPS receptors. We found one compound with potent *M. tuberculosis* cell growth inhibition activity as an IC_{50} ~ 0.5- to 20-μM inhibitor of *E,Z*-DPPS depending on substrate, a ~ 660-nM inhibitor of Rv3378c, as well as a 4.8-μM inhibitor of *E,Z*-FPPS, opening up the possibility of multi-target inhibition involving both cell wall biosynthesis and virulence factor formation.

Introduction

Complications associated with the growing drug-resistance of pathogenic bacteria to antibiotics have triggered the research for new tools and strategies to treat bacterial infections.[1, 2] Antibiotics such as penicillin, methicillin, and vancomycin all target bacterial cell wall biosynthesis, but all have either lost or are losing their efficacy.[3, 4] Drug-resistant forms of pathogenic *Mycobacterium tuberculosis* have reactivated the research on treatment of contagious disease tuberculosis, leading to a quest for new drugs that are active ideally against new targets.[5]

One attractive target for drug discovery against *M. tuberculosis* is to inhibit isoprenoid biosynthetic pathway in the bacterial cell wall biosynthesis, in particular, *E,Z*-decaprenyl diphosphate synthase (*E,Z*-DPPS; Rv2361c), an essential gene for the organism. *E,Z*-DPPS converts *E,Z*-farnesyl diphosphate (*E,Z*-FPP) to ω,*E*,poly*Z*-DPP by adding seven isopentenyl diphosphate (IPP) to *E,Z*-FPP, a reaction that is very similar to the one catalyzed by undecaprenyl diphosphate synthase (UPPS) in non-mycobacterial systems (Figure 3-1).[6-11] *E,Z*-FPP in *M. tuberculosis* is produced by *E,Z*-FPPS (Rv1086c) from geranyl diphosphate (GPP); however, *E,Z*-FPPS is not crucial for bacterial cell wall growth as its role can be replaced by DPPS in its absence.[12, 13] An implication of this observation is, therefore, that it would be desirable to inhibit DPPS activity with both GPP and *E,Z*-FPP as substrates, as we describe here.

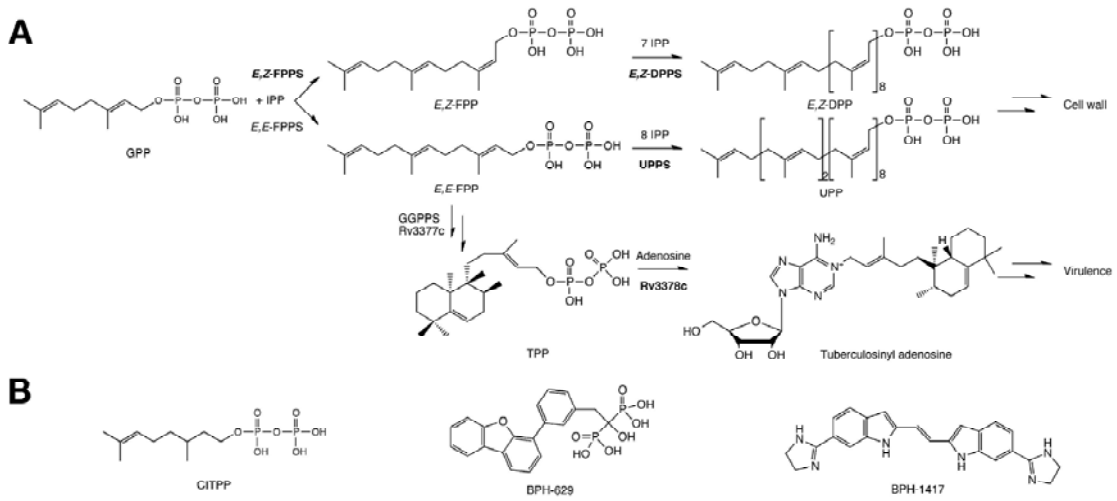


Figure 3-1. (A) Reactions catalyzed by *E,Z*-FPPS, *E,Z*-DPPS, UPPS, and Rv3378c. (B) Chemical structures of several inhibitors discussed in the text.

Structural insights into *E,Z*-FPPS, *E,Z*-DPPS, and UPPS reveal a highly similar structural motif, ζ-fold, as also found in another *M. tuberculosis* enzyme, tuberculosinol/tuberculosinyl adenosine synthase (Rv3378c).[14, 15] Rv3378c is an essential enzyme in persistent, non-replicative *M. tuberculosis* that resides within macrophages. It was originally thought to be involved in formation of the putative tuberculosinol and iso-tuberculosinol virulence factors,[16] then in formation of edaxadiene[17] and edaxadiene B,[18] and most recently, tuberculosinyl adenosine.[15] Hence, inhibition of Rv3378c is likely to represent a novel antivirulence approach to therapy.[17, 19] Despite its unique role in *M. tuberculosis*, the high structural similarity of Rv3378c to isoprenoid synthases brings special attention to this enzyme as an attractive target for antivirulence therapeutics. This also opens up a possibility of targeting multiple enzymes with a single potent inhibitor.

The first characterized isoprenoid diphosphate synthase with ζ-fold is undecaprenyl diphosphate synthase (UPPS) (Figure 3-2C).[20] The structure of UPPS is well known and promising inhibitors have been developed.[6, 9, 10, 21] Previously, we found from molecular dynamics (MD) studies that UPPS has considerable conformational flexibility and that there are up to four inhibitor binding sites (sites 1-4, Figure 3-2C).[10, 22] However, in *E,Z*-FPPS and *E,Z*-DPPS, there are far fewer available structures, and to date, inhibitors bind only in or close to the native binding site, *i.e.*, site 1 (Figure 3-2A

and 2B). Likewise, the tuberculosinyl diphosphate substrate for Rv3378c as well as an Rv3378c bisphosphonate inhibitor bind in or near the active site, *i.e.*, site 1 in UPPS (Figure 3-2D).[14]

In this work, we perform a series of long MD simulations on various *E,Z*-FPPS, *E,Z*-DPPS, and Rv3378c structures in order to explore conformational flexibility of these enzymes and to identify conformations that may be relevant for structure-based drug design. We compare the results with those obtained previously for UPPS by means of active site volume calculations and principal component analysis.[10, 23] Motivated by the presence of distinctive conformational transitions in *E,Z*-FPPS and *E,Z*-DPPS, the structures derived from the MD simulations were used to probe possible chain elongation mechanisms for prenyl diphosphates. In addition, as more emphasis has been focused on *E,Z*-DPPS as the promising drug target, we experimentally screened against *E,Z*-DPPS various libraries of small molecules using both native (*E,Z*-FPP) and non-native (GPP and *E,E*-FPP) substrates for *E,Z*-DPPS inhibition activity. Also, utilizing MD-derived structure in conjunction with computational docking, we investigated enrichment of known inhibitors of *E,Z*-DPPS in virtual screening. Finally, we explored the idea of multi-target inhibition to find inhibitors active against all three proteins of interest, *E,Z*-FPPS, *E,Z*-DPPS, and Rv3378c, as a route to new drug leads targeting both cell wall biosynthesis and virulence. Our results highlight the importance of addressing structural dynamics in both mechanistic understanding and inhibitor design for the promising drug targets.

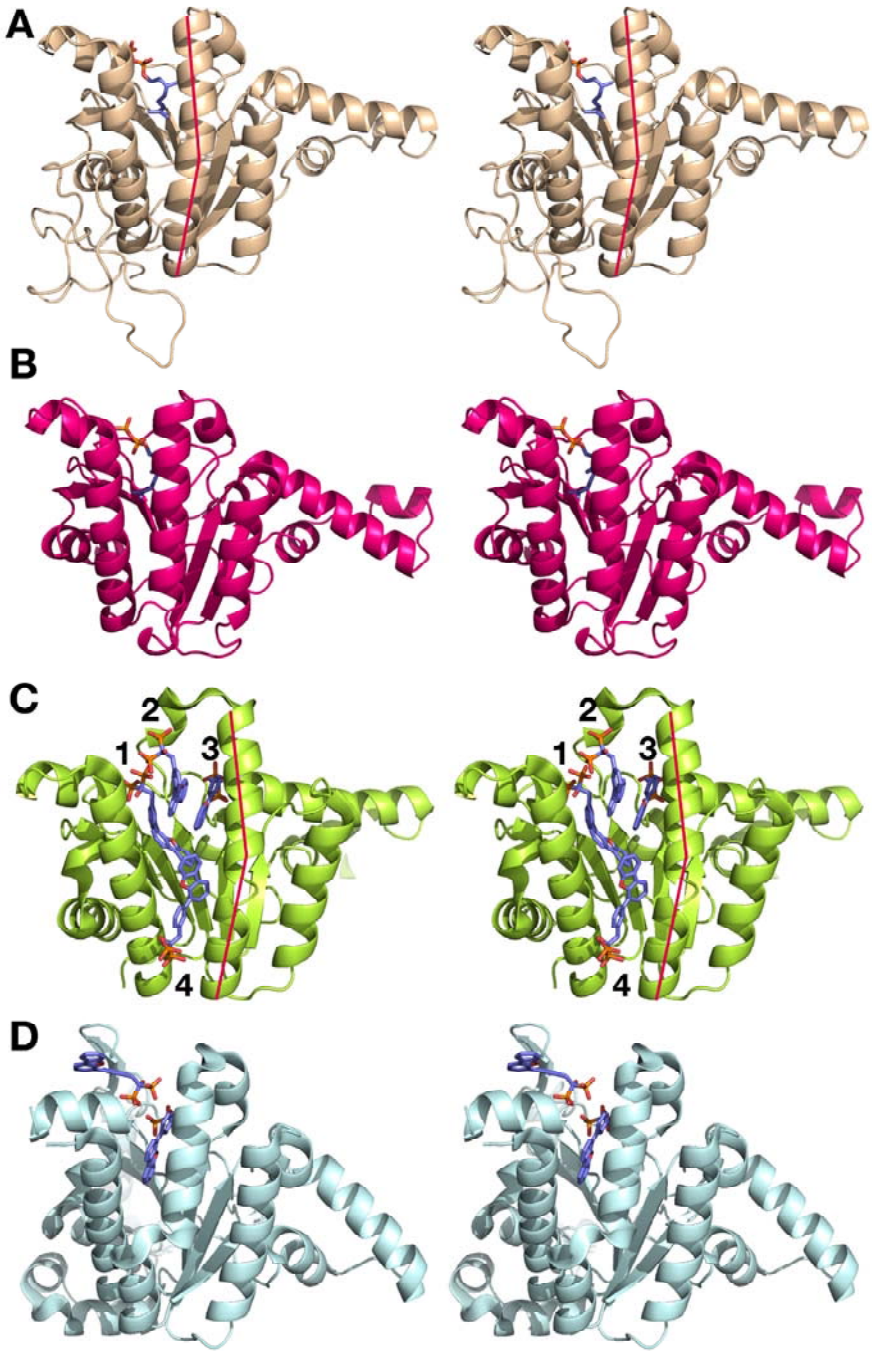


Figure 3-2. X-ray crystallographic structures of interest, shown as stereo views. (A) *M. tuberculosis E,Z*-FPPS (Rv1086c; PDB ID 2VG0) + CITPP. (B) *E,Z*-DPPS (Rv2361c; PDB ID 2VG3) + CITPP. (C) *E. coli* UPPS (PDB ID 2E98) + BPH-629. The numbers 1, 2, 3, and 4 denote the four ligand binding sties found in UPPS. (D) Rv3378c (PDB ID 3WQM) + BPH-629. The red lines indicate where the α3 helix can bend in *E,Z*-DPPS and UPPS: this helix is "broken" in Rv3378c.

Methods

Molecular dynamics simulations

The following *E,Z*-FPPS crystal structures were used: apo *E,Z*-FPPS (PDB ID 2VFW), *E,Z*-FPPS in complex with *E,E*-FPP (PDB ID 2VG1), and *E,Z*-FPPS in complex with citronellyl diphosphate (CITPP; PDB ID 2VG0).[21] *E,E*-FPP is a structural analog of the product *E,Z*-FPP but has two *trans* double bonds. For Rv3378c, two dimeric systems based on two different crystal structures were prepared for the MD simulations: Apo state (PDB ID 3WQL) and the inhibitor BPH-629 bound system (PDB ID 3WQM).[14] For each system, tleap program in AMBER 11 was used to neutralize the systems by adding Na^+ counterions and solvating using a TIP3P water box.[24-26] Minimization using the *sander* module of AMBER 11 was carried out in two stages: 1,000 steps of minimization of the solvent and ions with the protein and ligand restrained with a force constant of 500 kcal/mol $Å^2$, followed by a 2,500 step minimization of the entire system.[25, 27] An initial 20 ps MD simulation with a restraint of 10 kcal/mol $Å^2$ on the protein and ligand was then performed in order to heat the system to 300 K. Subsequently, 500 ns MD simulations were carried out on each system under the NPT ensemble at 300 K using AMBER 11 suite of programs with the ff99SBildn force field.[25, 27, 28] Periodic boundary conditions were used, along with a non-bonded interaction cutoff of 10 Å for Particle Mesh Ewald (PME) long-range electrostatic interaction calculations. Bonds involving hydrogen atoms were constrained using the SHAKE algorithm, allowing for a time step of 2 fs.[29]

For *E,Z*-DPPS, we used the following structures: apo *E,Z*-DPPS (PDB ID 2VG4), *E,Z*-DPPS in complex with IPP bound to monomer B (PDB ID 2VG2), and *E,Z*-DPPS in complex with CITPP bound to both monomers (PDB ID 2VG3).[21] Glycerol, phosphate, chloride, and sulfate ions used in crystallization were removed from the crystal structures while keeping the magnesium ions, which are essential for catalysis.[30] The protonation states of ionizable amino-acid residues were determined by using PROPKA and H++.[31-38] Geometries of the ligands were optimized using the B3LYP functional and a 6-31G(d) basis set in Gaussian 03 suite of programs[39] and their charges were assigned using the antechamber and RESP modules in AmberTools 11.[25, 40] All other force field parameters were taken

from the General AMBER force field (GAFF).[41] Proteins were solvated with TIP3P water molecules[26] with a buffer region of 10 Å in all directions and neutralized with counterions using the tleap program.[24] Each DPPS system was equilibrated using the *sander* with the MPI module of AMBER 11 and the ff99SBildn force field.[25, 27, 28] Water molecules were minimized with periodic boundary conditions in a constant volume with the protein and ligands fixed with a force constant of 2.0 kcal/mol $Å^2$, followed by a 150 ps MD simulation in the NPT ensemble. The entire system was minimized and heated from 0 K to 300 K over 500 ps, followed by two 20 ps MD simulations in the NVT and NPT ensembles, respectively. 500 ns MD simulations were performed on each DPPS system in the NVT ensemble with a Langevin thermostat using the *pmemd* module of Amber 11 with the ff99SBildn force field using a graphics card.[25, 27, 28] The PME summation method was used to describe the long-range electrostatic interactions, and short-range non-bonded interactions were truncated at 8 Å in the periodic boundary conditions.

Volume calculations

Active site volumes were calculated by using the POVME program with frames extracted every 25 ps from the simulations.[42] Points describing the binding pocket were manually defined along the hydrophobic cavity of monomer B of the apo *E,Z*-DPPS structure by locating a sphere with a 1 Å diameter at each point, removing any points where van der Waals clashes occurred with the protein. All points defined for monomer B of apo *E,Z*-DPPS were used for trajectories of *E,Z*-FPPS by aligning them to monomer B of apo *E,Z*-DPPS. Monomer A was also aligned to monomer B so that the results were comparable with the same points defining the active pocket. All ligands simulated in the MD were removed prior to volume calculation. The same procedure was repeated for Rv3378c and the UPPS X-ray crystallographic structures.

Principal component analysis

To compare the results obtained here with previous PCA results on UPPS obtained from 21 X-ray crystallographic structures,[10] the X-ray crystal structures of *E,Z*-FPPS and *E,Z*-DPPS that were used in the MD simulations were projected onto the UPPS PC space. Subsequently, the trajectories of the

apo states of *E,Z*-FPPS and *E,Z*-DPPS were projected onto the UPPS PC space. The PCA was carried out using the Bio3D package.[43]

Ligand docking

We docked a series of prenyl diphosphates with various chain lengths to *E,Z*-FPPS, *E,Z*-DPPS, and UPPS using the X-ray crystallographic structures 2VFW, 2VG4, and 2E98, respectively, in addition to the most open conformations taken from the MD simulations. The receptors were prepared by setting the receptor docking grid center to C1 of the co-crystallized ligand (CITPP) in the 2VG3 structure. Self-docking of the ligand CITPP was carried out as a preliminary test of the ability of the receptor grid center to recover the crystallographic pose of the ligand. The size of the inner and outer grid boxes were set to 14 Å and 20 Å in each direction, respectively. The prenyl diphosphates were then generated with various lengths (from C_{10} to C_{60}) using Maestro in Schrödinger Suite 2012.[44] The molecules were further prepared by using LigPrep with the OPLS2005 force field using all possible tautomers and stereoisomers generated in the pH range 5.2 ± 9.2, using Epik.[3, 45-47] Docking was carried out using the Glide XP precision in Schrödinger's Glide program.[48-51]

In vitro screening for *E,Z*-DPPS, *E,Z*-FPPS, and Rv3378c Inhibitors.

We screened an in-house library of 19 compounds against *E,Z*-DPPS using *E,Z*-FPP as substrate, 43 compounds using GPP as substrate, and 53 compounds using *E,E*-FPP as substrate. The structures and IC_{50} values for the active compounds (IC_{50} values in the range 31 nM to 880 μM) are shown in Figures 3-S1-S3. *E,Z*-DPPS was expressed and purified as described previously,[14] as were the *E,Z*-DPPS inhibition assays.[23] Briefly, the condensation of IPP and GPP, *E,E*-FPP, or *E,Z*-FPP catalyzed by *E,Z*-DPPS was monitored by using a continuous spectrophotometric assay for diphosphate release[52] in 96 well plates with 200 μL reaction mixtures containing 400 μM 2-amino-6-mercapto-7-methylpurine (MESG), 25 μM GPP, *E,Z*-FPP, or *E,E*-FPP, 200 μM IPP, 25 mM Tris-HCl (pH 7.5), 0.01% Triton X-100, and 1 mM $MgCl_2$. The IC_{50} values were obtained by fitting the inhibition data to a rectangular hyperbolic dose-response function in OriginPro 8.5 (OriginLab, Northampton, MA). The IC_{50} values for the most active hits were verified by radiometric assay as follows. A mixture of 15 μM substrate, 100 nM *E,Z*-DPPS, and inhibitors in the assay buffer (25 mM Tris-HCl, 1 mM $MgCl_2$, and

0.01% Triton X-100) was incubated for 10 min at 25 °C. 1.8 μL of 25 μM IPP (1% 1-^{3}H IPP, 15 μCi/mL, American Radiolabeled Chemicals, Inc.) was then added. The reaction was incubated at 37 °C for 10 min before quenching with 500 μL saturated NaCl solution. The saline solution was extracted with 500 μL butanol by vortexing, and 300 μL of the organic layer was transferred into scintillation vial for radiation readout. IC_{50} values were fitted in Origin 9.0 by analyzing the dose-response curves. For *E,Z*-FPPS, protein expression, purification and inhibition were all carried out as reported previously.[21]

For Rv3378c, protein expression and purification were all carried out as reported previously.[53] For inhibition assay of Rv3378c, a mixture of 100 μM TPP, 100 μM adenosine, 75 μg/mL Rv3378c, and inhibitors in the assay buffer (25 mM Tris-HCl, 1 mM $MgCl_2$, and 0.01% Triton X-100) was incubated for 2 hours at 37 °C. Tuberculosinyl adenosine formation was determined by LC/MS carried out using an Agilent LC/MSD Trap XCT Plus instrument. Compounds were separated on a Gemini 3 mm C18 110 Å (100*2 mm) column from Phenomenex using a 0–100% methanol (in water with 0.1% formic acid) gradient, and monitored by using positive-ion mode ESI at m/e=540.4.

Virtual screening

To incorporate receptor flexibility into computer-aided drug discovery as an application of the relaxed complex scheme, we carried out a virtual screening (VS) of the known actives (Figures 3-S1-S3) against an ensemble of 30 different *E,Z*-DPPS conformations.[54] The receptor conformers were selected by clustering the apo *E,Z*-DPPS trajectory based on the active site volumes. Any potential ligand binding sites in *E,Z*-DPPS were first explored by using the computational solvent mapping program FTMap, which suggests druggable hot spots in a protein by docking a number of small organic fragments into the protein.[55] Upon evaluating the results of FTMap (see Results and Discussion), we set the receptor grid center as the native binding pocket of the enzyme and used the same protocol as described above for ligand docking for receptor preparation.

The compound libraries consisted of the 43 (GPP substrate), 19 (*E,Z*-FPP), or 53 (*E,E*-FPP substrate) experimentally tested active inhibitors with IC_{50}s between 0.03 μM and 20 μM (GPP substrate), 0.6 μM and 90 mM (*E,Z*-FPPS substrate) or 0.65 μM and 880 mM (*E,E*-FPP substrate),

together with 1,000 decoys of average molecular weight 400 Da, from the Glide Decoy Set. The ligands were prepared using LigPrep in Schrödinger Suite 2012 with the OPLS2005 force field, and tautomers and stereoisomers were generated within the pH range of 5.2 ± 9.2 using Epik.[3, 45-47] The VS was carried out using the standard precision (SP) in Schrödinger's Glide program.[48-51] The VS results were analyzed by using the receiver operating characteristic (ROC) and area under the curve (AUC) metrics. ROC is a widely used method to evaluate the performance of VS in distinguishing known actives from decoy compounds.[56] The AUC then quantitatively compares the performance of different receptors; values of 0.5 correspond to a random selection of a compound in the library.

Results and discussion

M. tuberculosis E,Z-farnesyl diphosphate synthase (*E,Z*-FPPS), *E,Z*-decaprenyl diphosphate synthase (*E,Z*-DPPS), and tuberculosinol/tuberculosinyl adenosine synthase (Rv3378c) all crystallize as dimers as their functional units and adopt the same ζ-fold as seen in undecaprenyl diphosphate synthase (UPPS) (Figure 3-2). In the following sections, we begin by comparing conformational dynamics of *E,Z*-FPPS, *E,Z*-DPPS, and Rv3378c observed in the molecular dynamics (MD) simulations. Then, by performing principle component analysis, we compare the structural flexibility of *E,Z*-FPPS and *E,Z*-DPPS to UPPS in terms of the presence and type of bound ligand and size of the active site. We then propose a basis for chain length regulation by these enzymes and enhanced enrichment of active compounds of *E,Z*-DPPS in virtual screening. Finally, we report the experimental inhibition results of a lead compound that is active against *E,Z*-FPPS, *E,Z*-DPPS, and Rv3378c at nM and μM levels, suggesting a multi-target inhibition as a promising strategy for new generation drug development.

Structural flexibility of the enzyme active sites

To explore the conformational flexibility of the active sites of *E,Z*-FPPS, *E,Z*-DPPS, and Rv3378c, fluctuations in the volume of the binding pocket during the MD simulations were analyzed for each monomer of the various systems. Figure 3-3A-C shows the binding pocket volume fluctuations during the MD simulations for each monomer of the *E,Z*-DPPS systems investigated: apo (Figure 3-3A), isopentenyl diphosphate (IPP)-bound (Figure 3-3B), and citronellyl diophsphate (CITPP)-bound (Figure

3-3C). There are several clear transitions between the open and closed states in the active site, indicating large structural flexibility of the binding pocket of *E,Z*-DPPS. This flexibility is most pronounced in the apo and IPP-bound *E,Z*-DPPS simulations. An opening of the active site of the apo monomer B is particularly noticeable, as shown in Figure 3-4, in which the pocket opens up from 455 $Å^3$ in the X-ray crystal structure (Figure 3-4A) to 882 $Å^3$ in the early stage of the simulation (Figure 3-4B). Additional transitions between open and closed states of the active site between 200 $Å^3$ to 700 $Å^3$ range are subsequently observed. The large range in the volume fluctuation clearly indicates the flexible nature of the active site, with the largest volume likely being related to the need to accommodate the large product *E,Z*-DPP (C_{50}). Interestingly, the presence of CITPP (C_{10}, Figure 3-1B), a structural analog of GPP, restricts the active site volume to ~ 400 $Å^3$ during the entire simulation (Figure 3-3C). This is presumably due to strong hydrogen bond networks and salt bridges between the diphosphate moiety of CITPP and Arg residues, together with hydrophobic interactions in the active site.

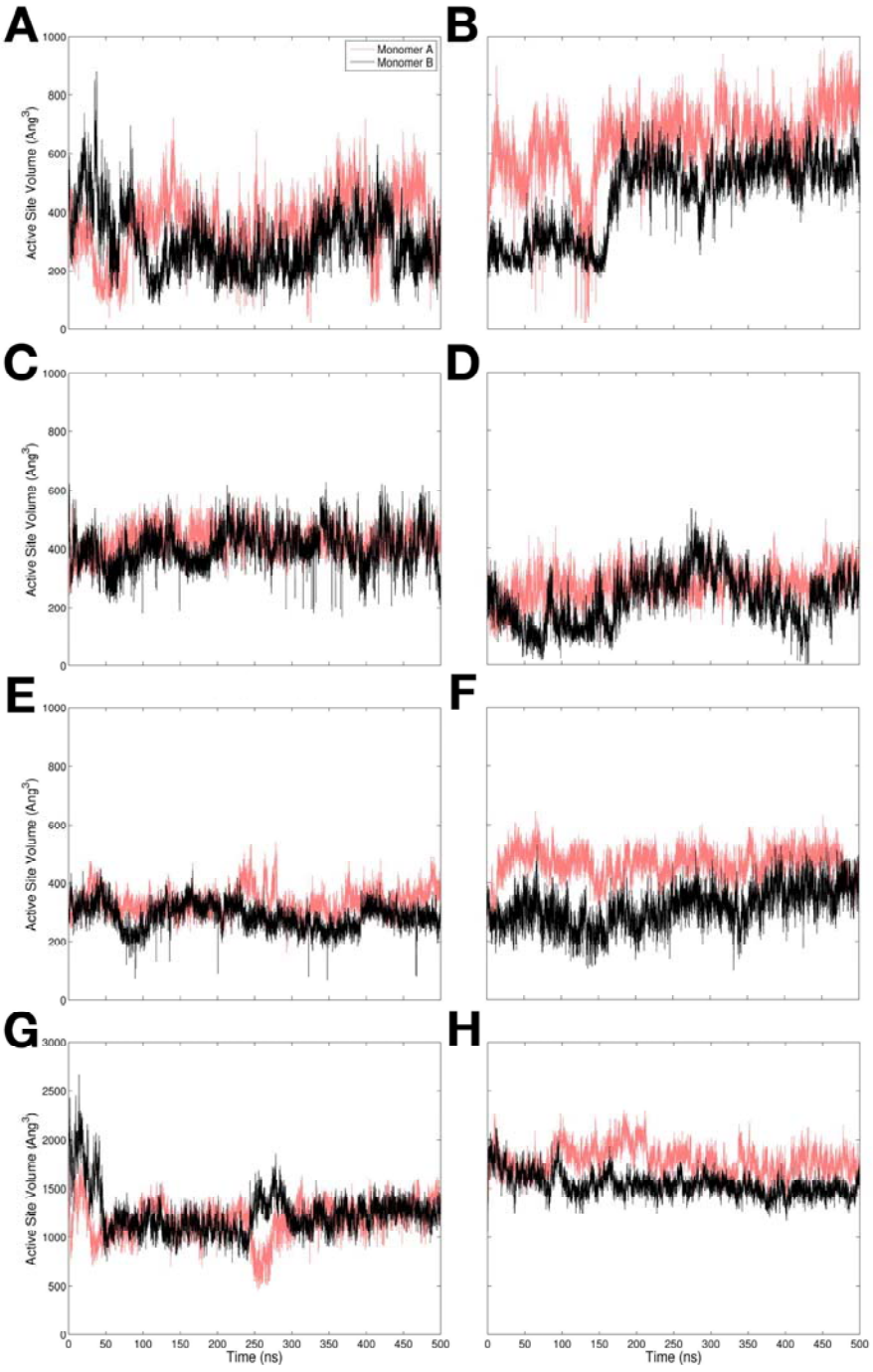


Figure 3-3. Volume fluctuations of the active site in each monomer of the various *E,Z*-FPPS, *E,Z*-DPPS, and Rv3378c systems. (A) Apo *E,Z*-DPPS. (B) *E,Z*-DPPS + IPP. (C) *E,Z*-DPPS + CITPP. (D) Apo *E,Z*-FPPS. (E) *E,Z*-FPPS + CITPP. (F) *E,Z*-FPPS + *E,E*-FPP. (G) Apo Rv3378c. (H) Rv3378c + BPH-629. Red and black lines are for monomers A and B, respectively.

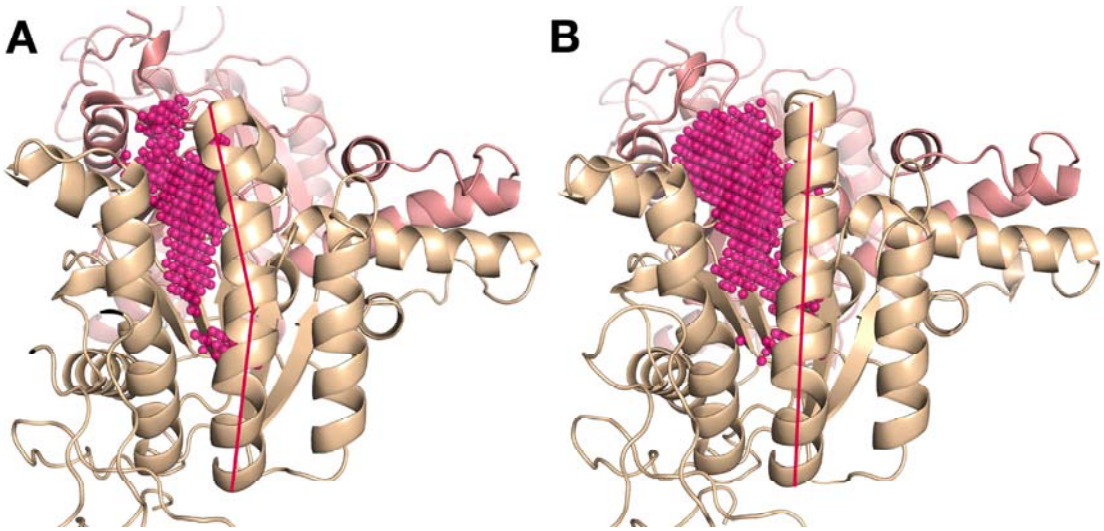


Figure 3-4. Comparison of the active site volumes. (A) Apo *E,Z*-DPPS in the X-ray crystal structure with V=455 Å^3. (B) The most open state in the MD simulation with V=882 Å^3.

Figure 3-4 clearly shows the large volume difference observed between the closed X-ray structure and most open MD structure. This large transition in the pocket volume arises from a pronounced bend in helix α3 between residues Ser126 and Phe136, which is very similar to what was previously seen in UPPS.[57] The X-ray crystallographic structures of the apo and holo *E,Z*-DPPS have this helix α3 in bent conformation and thus correspond to small pocket volumes. However, as discussed below in more detail, this pocket is too small to accommodate the larger *E,Z*-DPP product, which can, however, be docked to the more open structure (Figure 3-4B), and during this catalytic process, the bending motion of the α3 helix is likely to be important. Interestingly, we observe a closing of the pocket during an MD simulation in which we docked CITPP to the most open structure of apo DPPS (Figure 3-5). In this simulation, the active site has an initial volume of 756 Å^3 (Figure 3-5B), but it rapidly decreases during the simulation as the pocket closes (Figure 3-5B and 5C), with the closed structure having a volume of ~ 230 Å^3. Thus, the substrate-like ligand induces formation of a closed state with helix α3 in the bent conformation. Similar effects are also seen with the substrate-like ligands in UPPS and are important in the context of virtual screening and enzymatic catalysis, as described in more detail below.

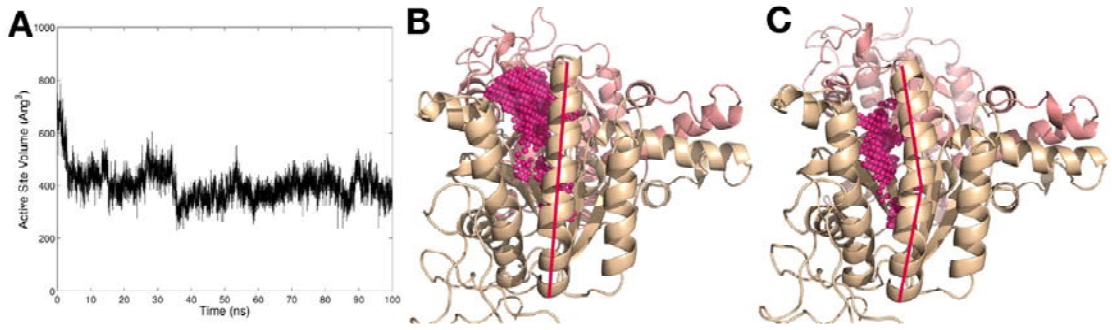


Figure 3-5. (A) Rapid closure of the expanded active site of monomer B in *E,Z*-DPPS upon incorporation of CITPP. (B) Binding pocket of the apo starting structure (V=756 Å^3). (C) Pocket in the closed state of the active site after MD in the presence of CITPP (V=229 Å^3). The helix bend illustrated helps close the pocket.

E,Z-FPPS, producing the C_{15} isoprenoid *E,Z*-FPP, has not unexpectedly a smaller active site than DPPS, which synthesizes the C_{50} compound, *E,Z*-DPP. As a result, the active site of *E,Z*-FPPS during the MD simulations remains in a relatively closed state, fluctuating only up to a volume of ~ 480 Å^3, even in the absence of ligands (Figure 3-3D). In the presence of the small, substrate-like inhibitor CITPP in both monomers, the active site volume remains constant around 300 Å^3 (Figure 3-3E). CITPP is a known inhibitor of *E,Z*-FPPS,[30] and as with CTIPP binding to *E,Z*-DPPS, it stabilizes the closed conformation through polar and hydrophobic interactions. With the non-native substrate *E,E*-FPP, the active site volume is larger as shown in Figure 3-3F, but again does not approach the much larger volumes seen in apo *E,Z*-DPPS.

Rv3378c shows quite different dynamics from *E,Z*-DPPS or *E,Z*-FPPS, as shown in Figure 3-3G and 3H. The enzyme has an inherently bigger active site with a volume larger than 1,000 Å^3. The enzyme in complex with a bisphosphonate inhibitor BPH-629A stays in a widely open state with a volume as large as ~ 2,300 Å^3. The large active site of Rv3378c can be traced to its different enzymatic function from those of *E,Z*-DPPS or *E,Z*-FPPS. Based on the X-ray crystallographic structure that contains tuberculosinyl diphosphate (TPP) substrate, the reaction catalyzed by Rv3378c may take place around the top region of the active site, rather than along the helical channel as isoprenoid synthases. In addition, it is possible that Rv3378c adopts the expanded conformation due to its nucleophilic reaction involving adenosine in addition to TPP, as suggested recently.[15]

Our results highlight different conformational flexibility of *E,Z*-DPPS, *E,Z*-FPPS, and Rv3378c that all adopt the same ζ-fold. The enzyme synthesizing longer product, *i.e.*, DPPS, has more pronounced plasticity in its active site while the ones producing a shorter chain remain more rigid. In the next section, we compare the conformational states of these enzymes to that of UPPS, which synthesizes even longer C_{55} product, and further relate the conformational dynamics to the enzymatic functions.

Pocket volume fluctuation and principal component analyses

As the first characterized enzyme with the ζ-fold and due to its essential role in the bacterial cell wall biosynthesis across different species, UPPS has been extensively studied in terms of its structural dynamics and inhibitor design. In our previous work, we showed that UPPS has a very dynamic nature around its active site and exists in distinct conformational states depending on the presence and type of the bound ligand in the active site.[10, 22] Inspired by the high structural similarity of *E,Z*-FPPS and *E,Z*-DPPS to UPPS, we compare the conformations of *E,Z*-FPPS, *E,Z*-DPPS, and UPPS using a criterion of the active site volume and principal component analyses (PCA).

Table 3-1 shows the pocket volumes for *E,Z*-FPPS, *E,Z*-DPPS, UPPS, and Rv3378c, using both X-ray crystallographic and in some cases, MD-derived structures. There is a broad range in volumes for each system: for *E,Z*-FPPS, the volume ranges between 240 $Å^3$ to 537 $Å^3$; for *E,Z*-DPPS, 227 $Å^3$ to 882 $Å^3$; for UPPS, 155 $Å^3$ to 1,440 $Å^3$; and for Rv3378c, 131 $Å^3$ to 1,436 $Å^3$. In Table 3-1, UPPS structures are clustered into three different conformational states based on its pocket volume: closed stated, semi-open, or ajar, state, and open state. Contingent on the clustering of UPPS structures, the small volumes of *E,Z*-FPPS and *E,Z*-DPPS active sites resemble the closed state of UPPS with the volume of 270-315 $Å^3$ (Table 3-1) where substrate-like ligands are bound. The open structures of *E,Z*-FPPS and *E,Z*-DPPS from the MD simulations have active site volumes of 537 $Å^3$ and 882 $Å^3$, respectively, and thus are more similar to the ajar or open state of UPPS. However, the active sites of *E,Z*-FPPS and *E,Z*-DPPS are clearly smaller on average than that of UPPS, likely due to the fact that UPPS have multiple bound ligands in many cases in its four binding sites while *E,Z*-FPPS and *E,Z*-DPPS do not.

Table 3-1. Computed active site volumes and PC values and corresponding conformational states for *E,Z*-FPPS, *E,Z*-DPPS, UPPS, and Rv3378c.

System	PDB ID	Pocket Volume (Å^3)	Ligand Volume (Å^3)	PC1	PC2	Conformation and average pocket volume (Å^3)
E,Z-FPPS	2VFW	240	0	-10.77	-2.42	Closed 269 ± 50
	2VG0	241	272	-11.14	-1.48	
	2VG1	327	344	-10.26	-2.42	
	2VFW Open state	537	0	-	-	Open (MD) 537
E,Z-DPPS	2VG2	227	189	-9.53	2.03	Closed 381 ± 133
	2VG4	455	0	-9.24	1.77	
	2VG3	461	272	-8.94	1.63	
	2VG4 Open state	882	0	-4.62	1.29	Open (MD) 882
UPPS	*Sa*UPPS	155	344	-10.87	4.2	Closed (substrate-bound) 272 ± 68
	1X07	270	189	-10.81	-1.16	
	1X09	307	189	-12.58	1.28	
	1X06	312	353	-9.75	-1.41	
	1X08	315	353	-10.58	-1.13	
	3SGV	456	351	0.99	-0.55	Ajar (apo/non-bisphosphonate-bound) 726 ± 167
	4H2M	581	397	1.71	1.69	
	4H2J	672	430	1.97	2.08	
	3TH8	765	338	0.61	4.01	
	4H38	804	387	1.87	1.70	
	4H3C	870	342	2.60	1.60	
	3SH0	932	329	4.57	2.29	
	2E9A	948	357	2.34	-3.90	Open (bisphosphonate-bound) 982 ± 31
	2E99	987	366	4.04	-4.80	
	2E9D	1010	357	3.22	-3.15	
	4H3A	1048	309	2.52	1.49	Ajar (apo/non-bisphosphonate-bound) 1064
	4H2O	1079	414	1.73	2.50	
	2E98	1082	355	5.45	-6.09	Open (bisphosphonate-bound) 1082 ± 22
	3QAS	1169	0	1.00	1.72	Ajar (apo/non-bisphosphonate-bound) 1205 ± 32
	3SGX	1213	393	4.19	6.11	
	3SGT	1232	327	2.03	-0.13	
	2E9C	1440	445	2.90	-4.18	Open (bisphosphonate-bound) 1440
Rv3378c	4CMX	131	246			Closed 131
	3WQK	714	0			Open 1046 ± 256
	3WQM	808	355			
	4KT8	905	412			
	3WQN	1051	412			
	4CMV	1171	152			
	3WQL	1240	0			
	4CMW	1436	152			

PCA is an attractive methodology for structure comparison by reducing the dimensionality of variables observed. In previous work on UPPS, we found three conformational clusters of UPPS from PCA using available X-ray crystallographic structures, which correspond to our clustering based on the pocket volumes: closed state containing substrate-like ligands; ajar, or semi-open, state for either apo or non-bisphosphonate bound structures; and open state with bound bisphosphonate ligands, often with multiple bound (circled as A, B, and C respectively, in Figure 3-6A).[10] Projection of X-ray structures of *E,Z*-FPPS and *E,Z*-DPPS onto the PC space of UPPS shows that *E,Z*-FPPS and *E,Z*-DPPS are close to the closed state of UPPS while the ajar and open states of UPPS are solely populated by UPPS (Figure 3-6A). On the other hand, projection of MD trajectories of *E,Z*-FPPS and *E,Z*-DPPS indicates that a large conformational space is sampled in the simulations (Figure 3-6B). However, the spaces sampled do not quite overlap with the ajar or open state of UPPS (Figure 3-6B). This is presumably due to the fact that the product of UPPS, *i.e.*, UPP (C_{55}), is slightly larger than those of *E,Z*-FPPS and *E,Z*-DPPS.

Our PCA results well characterize the existence of distinctive conformational states for different enzymes that synthesize the products with unique length. This indicates a high correlation between the structural dynamics around the active site and enzymatic functions. In the next section, we further investigate the chain elongation mechanisms of *E,Z*-FPPS and *E,Z*-DPPS predicted by means of computational docking of prenyl molecules with different chain lengths.

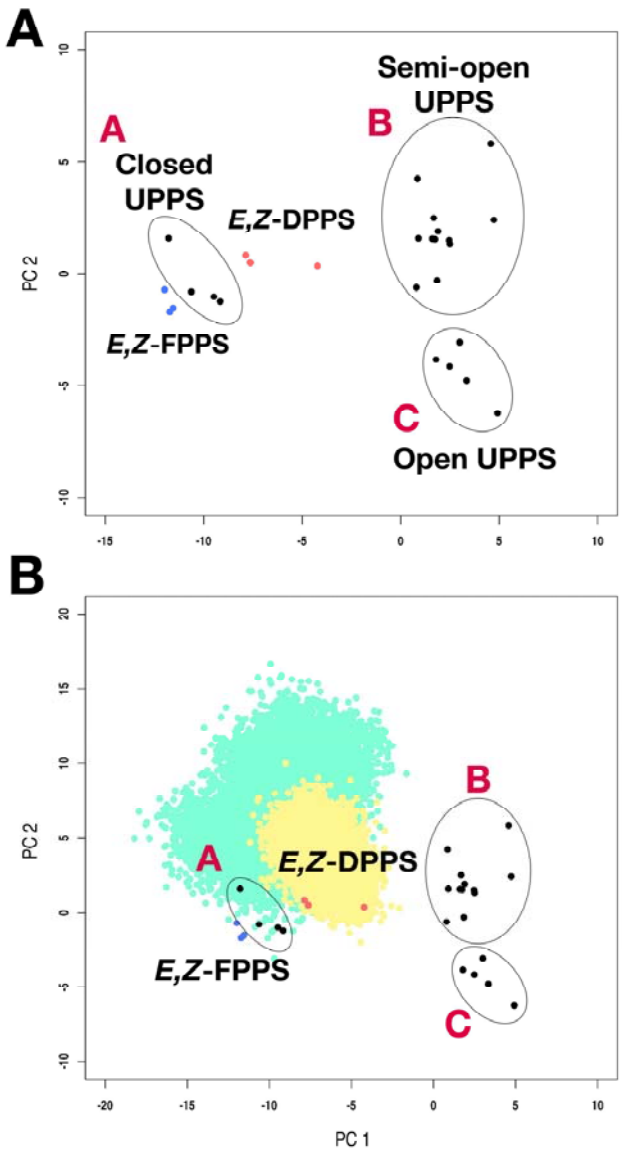


Figure 3-6. Separation of conformational states for different enzymes revealed by principal component analysis. (A) Projection of X-ray crystallographic structures of *E,Z*-FPPS (blue) and *E,Z*-DPPS (red) onto UPPS PC space. Distinctive conformational states corresponding to UPPS structures are marked. (B) Projection of trajectories of *E,Z*-FPPS (green) and *E,Z*-DPPS (yellow) onto UPPS PC space.

Chain elongation mechanisms of *E,Z*-FPPS, *E,Z*-DPPS, and UPPS

The length of the product synthesized have been used as an indicative of protein function in *E,E*-prenyl transferases.[58] In principle, one hypothesizes that the larger the active site, the longer will be the prenyl hydrocarbon chain produced by the enzyme. However, the active site volume calculations and PCA results described above indicate that the X-ray crystallographic structures of *E,Z*-FPPS and *E,Z*-DPPS correspond to the closed state of UPPS, in which case, certainly with *E,Z*-DPPS, it would potentially not be possible to accommodate the C_{50} product, as we indeed found experimentally (data not shown). Motivated by the structural flexibility of these enzymes observed in the MD simulations, we thus next probed the chain elongation mechanisms by docking a set of prenyl diphosphates with various chain lengths, *i.e.*, C_{10} to C_{55}, using the MD-derived structures of *E,Z*-FPPS, *E,Z*-DPPS, and UPPS. This approach is similar in sprit to that used by Wallrapp *et al*.,[58] but here we incorporated the MD structures to account for the protein flexibility.

Docking poses for the corresponding products to each enzyme are shown in Figure 3-7A-C, and docking score as a function of the prenyl chain length is plotted in Figure 3-7D. While *E,Z*-FPPS was able to accommodate the prenyl diphosphate up to C_{25} in its active site, the docking scores indicate the strongest binding with C_{10}, with C_{15} (*i.e.*, product of *E,Z*-FPPS) and C_{20} both being ~ 2-3 kcal/mol higher. Any prenyl molecules longer than C_{25} failed to dock into even the most open *E,Z*-FPPS structure due to the small size of the binding pocket, as indicated by docking scores of 0 kcal/mol (Figure 3-7D).

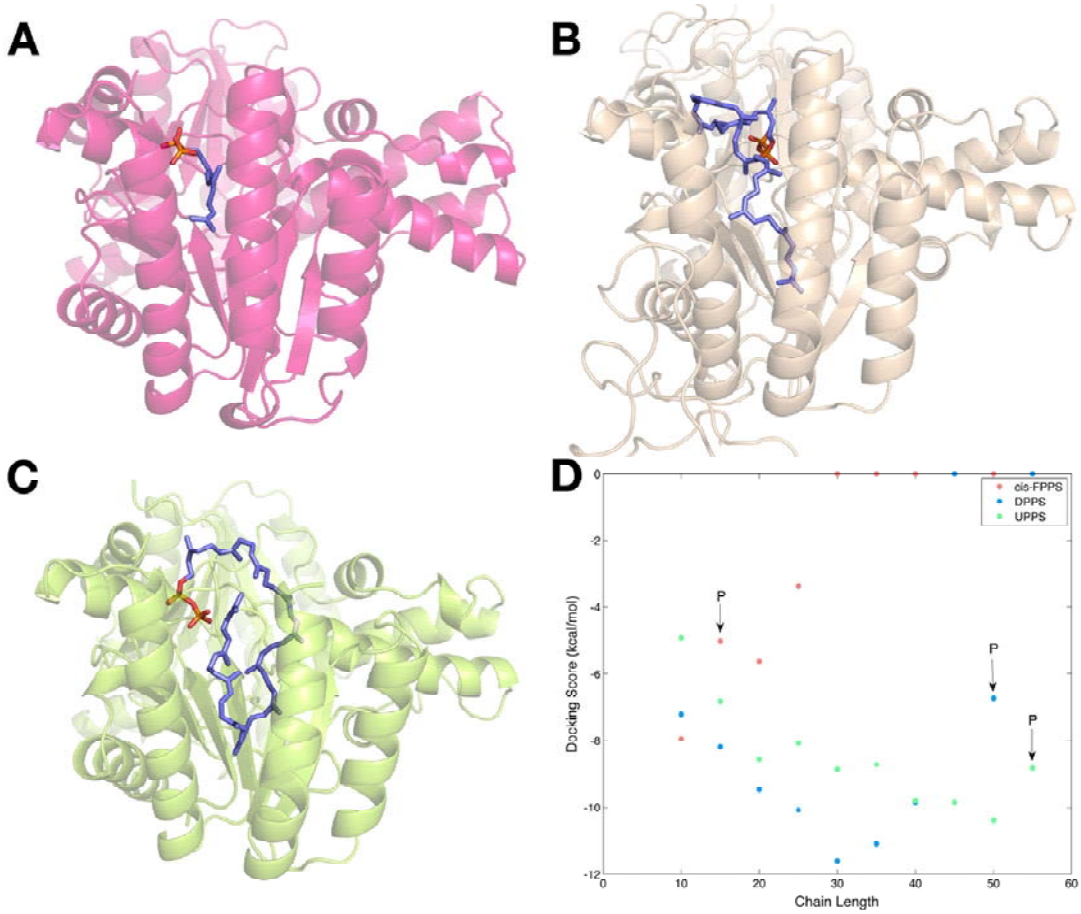


Figure 3-7. Docking poses of the products of the enzymes synthesizing prenyl molecules with various chain lengths. (A) *E,Z*-FPPS + FPP. (B) *E,Z*-DPPS + DPP. (C) UPPS + UPP. (D) Docking scores for the prenyl molecules with various chain lengths upon docking into *E,Z*-FPPS, *E,Z*-DPPS, and UPPS. The "P" signs indicate the products of three enzymes.

For *E,Z*-DPPS and UPPS, we see an expected U-shaped curve in which the best docking scores are found around chain lengths slightly shorter than the enzymes' own products; the energy minimum for the docking score is found at C_{30} for *E,Z*-DPPS and C_{50} for UPPS, which is not the normal product chain length (C_{50} and C_{55}, respectively). The score becomes then slightly worse with an increase in chain length, and both enzymes failed to dock any prenyl molecules longer than their products (Figure 3-7D). Interestingly, *E,Z*-DPP (C_{50}), the enzyme's final product, could not be docked to the X-ray crystallographic structure of *E,Z*-DPPS (data not shown) while it was able to fit into the most open MD-derived structure (Figure 3-7B). The docking score for *E,Z*-DPPS with the C_{50} product is about 4

kcal/mol higher than with C_{30}, which showed the strongest binding. Likewise, with UPPS, the energy minimum locates at C_{50} with the C_{55} product being ~ 1.5 kcal/mol higher. However, the expanded structures of both *E,Z*-DPPS and UPPS could not dock the prenyl diphosphates longer than their own products. Clearly, these results indicate that once the chain becomes sufficiently long, its binding energy increases, which results in product being released from the enzyme. In the mean time, the flexible nature of the active site, potentially induced by the bending motion of helix α3, likely enables the expansion of the binding pockets of *E,Z*-DPPS and UPPS, which is essential for synthesis of long prenyl diphosphates. Also, our method used here combining two computational techniques, *i.e.*, molecular docking and MD simulations, suggests future application to predict unknown enzymatic mechanisms for chain elongation.

Inhibition of *E,Z*-DPPS and receptor flexibility

While *E,Z*-FPPS and *E,Z*-DPPS work sequentially to produce *E,Z*-DPP in *M. tuberculosis* cell wall biosynthesis, *E,Z*-DPPS is able to substitute the role of *E,Z*-FPPS in the absence of the latter. While DPPS normally converts *E,Z*-FPP to *E,Z*-DPP (Figure 3-1), it can also utilize GPP, the substrate of *E,Z*-FPPS, as substrate. Ideally, then, an inhibitor should target *E,Z*-DPPS with either GPP or *E,Z*-FPP as substrate. Therefore, we screened our in-house library of putative prenyl synthase inhibitors against *E,Z*-DPPS using GPP, *E,Z*-FPP, and *E,E*-FPP as substrate; *E,E*-FPP, as with GPP, can be converted to the long chain prenyl diphosphates. This resulted in 43 inhibitors with IC_{50} values in the 30 nM to 20 μM range when using GPP as substrate (Figure 3-S1), 19 inhibitors with IC_{50} in the 600 nM to ~ 100 μM range when using *E,Z*-FPP as substrate (Figure 3-S2), and 53 inhibitors with IC_{50} as low as 650 nM using *E,E*-FPP as substrate (Figure 3-S3).

As well established, accounting for receptor flexibility in computational docking can significantly improve the results in structure-based drug discovery.[54] In UPPS, it was found that the rarely sampled structures in the MD simulations show the binding modes that are not observed with the static X-ray crystallographic structures.[22] Therefore, as we found a high degree of conformational plasticity of DPPS and several potent inhibitors of the enzyme with various substrates, we carried out a virtual screening (VS) with an ensemble of the receptors with varying active site volumes in order to

address the receptor flexibility. The receptor structures were taken from the apo *E,Z*-DPPS MD trajectory as it showed the largest fluctuation in the active site volume, in addition to the X-ray crystallographic structure.

As a preliminary examination of the ability to recover the crystallographic binding pose of the ligand, we first carried out self-docking of the co-crystallized ligand CITPP into *E,Z*-DPPS receptor. The docking pose shown in Figure 3-S4 confirmed the validity of the receptor grid center we used. We also investigated the possibility of the existence of other binding sites in *E,Z*-DPPS using FTMap.[55] The results of FTMap analysis shown in Figure 3-S5 suggest that DPPS does not have other binding sites. Therefore, we proceeded to perform the VS with the grid center located at the native binding site.

We evaluated the receptor performance in each VS using the receiver operating characteristic (ROC) curve and area under the curve (AUC) metrics. Figure 3-8A-C shows the AUC results of the MD-derived receptor structures using the inhibitors tested using GPP, *E,Z*-FPP, and *E,E*-FPP as substrate, respectively (individual ROC curves are shown in Figures 3-S7-S9). The populations of the sampled *E,Z*-DPPS structures as functions of the active site volume are shown in red and indicate that the most probable receptor volume is ~ 250 $Å^3$. When GPP is used as substrate, the AUC results show that the top scoring receptors all have volumes smaller than ~ 300 $Å^3$ (Figure 3-8A, in blue). This cutoff is more pronounced when *E,Z*-FPP is used as substrate (Figure 3-8B, in blue), where there is a clear step function behavior: The receptors with volume less than ~ 300 $Å^3$ does not have predictive power in selecting the actives from decoys while 90% of the receptors with volume bigger than 300 $Å^3$ have AUC greater than 0.5, with the best MD-derived receptor conformers f having AUC near 0.7 – a considerable enrichment in the actives. The presence of enrichment observed for the receptors with V < 300 $Å^3$ with GPP or *E,E*-FPP as substrate is due to the highly ranked small inhibitors used, while the inhibitors tested with *E,Z*-FPP are generally larger (Figures 3-S7-S9). Therefore, we speculate that the initial step in GPP chain elongation may be targeted by the small bisphosphonate inhibitors that can bind to a smaller receptor pocket, while with *E,Z*-FPP as substrate, the normal chain elongation step is being targeted. In all cases, however, neither the most populated conformations nor the X-ray

crystallographic receptors provide better enrichments in the actives than random selection of the compound in the library.

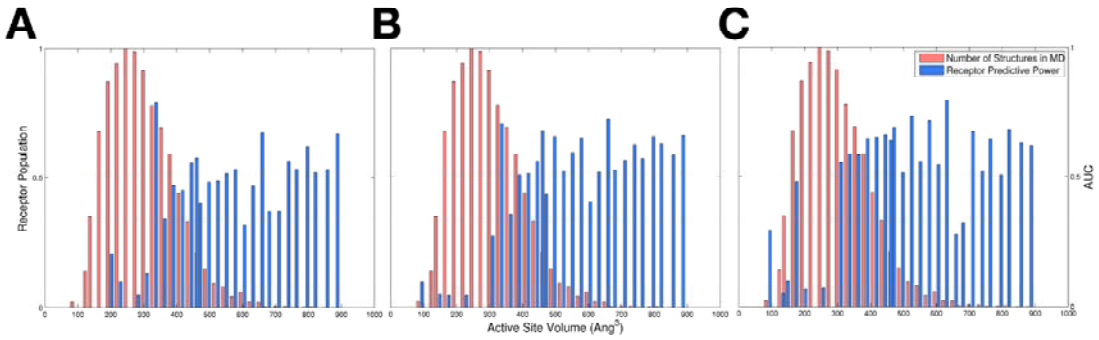


Figure 3-8. Active site volume of *E,Z*-DPPS and predictive performance in screening the compound library tested with (A) GPP substrate; (B) *E,Z*-FPP substrate; and (C) *E,E*-FPP substrate.

Detailed examination of docking poses of the actives in several high performance receptors provides a significant role of receptor flexibility. For example, when *E,Z*-FPP, the native substrate, was used, the receptors 10, 22, and 30 show the AUC values of 0.710, 0.726, and 0.662 (Figure 3-S7), with corresponding active site volumes of 331 $Å^3$, 654 $Å^3$, and 882 $Å^3$, respectively. This gives first insight that the binding pocket size may not be a critical factor. However, high enrichment observed in the small receptors is mainly due to the small inhibitors ranked high on the list, as mentioned above, in addition to location of the docked inhibitors far from the active site, *i.e.*, either limited to the top, solvent-exposed region of the active site or near the interface of two monomers, leading to the erroneously high enrichment. On the other hand, the large receptors are able to accommodate both large and small inhibitors in the binding pocket, enabling more realistic prediction of the binding modes of the inhibitors.

Overall, our results clearly show that there is remarkable improvement in VS upon using an ensemble of MD-derived receptor conformers, most of which, interestingly, are rarely sampled in the MD simulations. The enrichment of VS is noticeably enhanced upon utilizing the receptors with the expanded binding pocket compared to the poor performance when the most populated or X-ray crystallographic structures were used, emphasizing the significance of accounting for the receptor

flexibility in the structure-based drug discovery. Approaches using different metrics such as RMSD clustering to generate the receptor ensemble can possibly yield good results as well in application of the relaxed complex scheme.

Toward multi-target inhibition of *E,Z*-FPPS, *E,Z*-DPPS, and Rv3378c

In previous work, we reported the X-ray crystallographic structure of Rv3378c in the presence of its substrate, TPP, as well as in the presence of a bisphosphonate inhibitor, BPH-629 (Figure 3-1B).[14] Bisphosphonates are often found to be good inhibitors of prenyl synthases as they contain a diphosphate group isostere, a bisphosphonate. For example, bisphosphonates are known potent inhibitors of UPPS and *E,E*-FPPS, where inhibitors of the latter being used clinically to treat bone resorption diseases. However, there has been little progress in developing bisphosphonates that act against bacteria since they do not penetrate the bacterial cell wall. On the other hand, there are many other known UPPS inhibitors, and many of the *E,Z*-DPPS inhibitors shown in Figure 3-S1 also inhibit UPPS.[10] Therefore, we tested several of these inhibitors for activity against *E,Z*-FPPS (Figure 3-S9) and three for activity against Rv3378c (Figure 3-S10) using TPP and adenosine as substrates.

The most potent inhibitor for Rv3378c was BPH-629, which had an IC_{50} of 210 nM. This compound was also active against *E,Z*-DPPS with an IC_{50} of 610 nM when *E,Z*-FPP was used as substrate. However, BPH-629 was inactive against the *M. tuberculosis* model, *M. smegmatis*. In contrast, we found that the bisamidine BPH-1417 (Figure 3-1B) had an activity against DPPS with an IC_{50} of 0.5 to 20 μM, depending on substrate used, and Rv3378c with an IC_{50} of 660 nM. In addition, it also inhibited *E,Z*-FPPS with an IC_{50} of 4.8 mM (Figure 3-S9). The binding modes of BPH-1417 predicted by docking are shown in Figure 3-9; in each case, the native binding site, *i.e.*, site 1, is likely to be occupied.

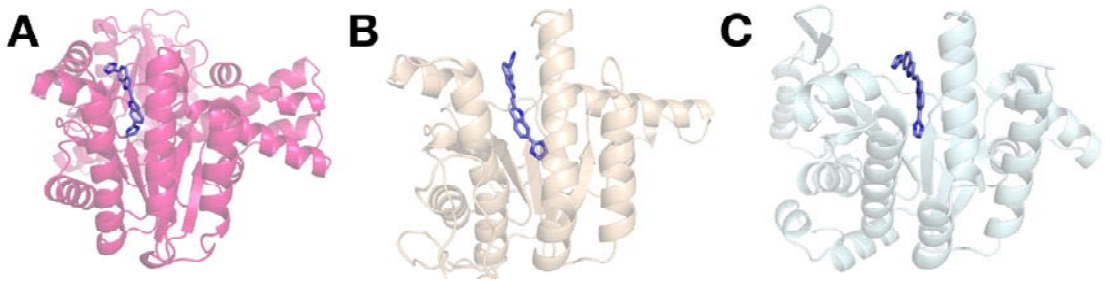


Figure 3-9. Predicted docking poses for BPH-1417. (A) *E,Z*-FPPS. (B) *E,Z*-DPPS. (C) Rv3378c.

While we previously showed that a bisamidine BPH-1358 (Figure 3-S10) was a potent UPPS inhibitor that also had *in vivo* activity in *S. aureus* mouse model of infection,[10] the compound was inactive against Rv3378c. However, the bisamidine BPH-1417 has potent *in vitro* as well as *in vivo* activity against *S. aureus*, and has also been reported to have an MIC in the range of 0.3 mg/mL to 1.3 mg/mL against *M. tuberculosis*.[59] It is also possible that DNA is a major target in these organisms, as with other bisamidines. Clearly, the ability of BPH-1417 to potentially target DNA, cell wall biosynthesis catalyzed by *E,Z*-FPPS and *E,Z*-DPPS, and virulence factor formation by Rv3378c is of particular interest in the context of multi-target inhibitor development for drug-resistant infections.[60]

Conclusions

In this work, we examined the structural plasticity of three prenyl synthases from *M. tuberculosis* by performing molecular dynamics (MD) simulations: *E,Z*-farnesyl diphosphate synthase (*E,Z*-FPPS), *E,Z*-decaprenyl diphosphate synthase (*E,Z*-DPPS), and tuberculosinol/tuberculosinyl adenosine synthase (Rv3378c), all of which contain a characteristic ζ-fold. In particular, *E,Z*-DPPS and Rv3378c are of special interest as new drug targets in treatment of tuberculosis. Through binding pocket volume calculations, we showed that the apo state of *E,Z*-DPPS exhibited the large structural flexibility around the active site, showing several transitions between open and closed states characterized by bending motion of helix α3. *E,Z*-FPPS sampled a limited conformational space and Rv3378c showed distinctive dynamics in the active site due to its different enzymatic function despite the highly similar structural motif. The structures were clustered based on the computed active site

volume and further analyzed on the basis of three distinctive conformational states of undecaprenyl diphosphate synthase (UPPS) by means of principal component analysis. The X-ray crystallographic structures of *E,Z*-FPPS and *E,Z*-DPPS resembled most closely the closed structures of UPPS where substrate-like ligands were bound. We speculate this to be due to many UPPS structures that have multiple bound ligands as in its semi-open or open states, while *E,Z*-FPPS and *E,Z*-DPPS typically have a single ligand bound.

We further incorporated the structural flexibility to probe the chain elongation mechanisms of *E,Z*-FPPS, *E,Z*-DPPS, and UPPS by docking prenyl phosphates with various chain lengths to the MD-derived structures. The prenyl molecules with chain lengths shorter than products showed the most favorable binding. Use of the most open MD conformer suggested that the conformational transition to the expanded state is essential for the chain length regulation in catalysis.

To address to the increasing attention to *E,Z*-DPPS as a promising drug target, we reported the experimental analysis of inhibition assays against *E,Z*-DPPS using both native and non-native substrates, finding several potent leads in μM to nM range in each case. Motivated by the large structural plasticity of DPPS, we combined the MD results with our drug discovery protocol to address the receptor flexibility in computer-aided drug discovery. We showed that use of an ensemble of receptor conformers with various binding pocket volumes in virtual screening largely improved the receptor predictive performance in distinguishing the known actives from decoy compounds. Particularly, the MD-derived structures with the pocket volume greater than 300 $Å^3$ exhibited the highly enhanced enrichment. We also tested several of the *E,Z*-DPPS inhibitors for activity against *E,Z*-FPPS and Rv3378c and found several bisphosphonates and bisamidines were active against several targets. In particular, the compound BPH-1417 was a 660-nM inhibitor of Rv3378c, a 4.8-μM inhibitor of *E,Z*-FPPS, and a 1.3-μM inhibitor of *E,Z*-DPPS, opening up the possibility of multi-target inhibition in which both cell wall biosynthesis and virulence factor formation may be inhibited. Our study shows the conformational flexibility of three essential *M. tuberculosis* enzymes and connects the functional mechanisms to structural dynamics. Also, the results reported here highlight the significance of

accounting for the receptor flexibility in the structure-based drug design, and further propose a possibility for multi-target inhibition as new generation drug development.

Acknowledgements

This work was supported by the National Science Foundation, the United States Public Health Service (NIH grants CA158191, GM031749 and GM065307); the Howard Hughes Medical Institute; the National Biomedical Computation Resource; the UCSD Center for Theoretical Biological Physics; a Harriet A. Harlin Professorship (E. O.); the University of Illinois Foundation/Oldfield Research Fund; and a Molecular Biophysics Training Grant from the NIH. F. F. acknowledges financial support of the Beatriu de Pinós program from AGAUR for a postdoctoral grant (2010 BP_A 00339). X. F. was supported by an American Heart Association, Midwest Affiliate, Pre-Doctoral Fellowship (13PRE14510056).

Chapter 3 is a minimally modified reprint of the material as it appears in Meekyum Olivia Kim, Xinxin Feng, Ferran Feixas, Wei Zhu, Steffen Lindert, Shannon Bogue, William Sinko, César de Oliveira, Guodong Rao, Eric Oldfield, and J. Andrew McCammon, "A molecular dynamics investigation of *Mycobacterium tuberculosis* prenyl synthases: Conformational flexibility and implications for computer-aided drug discovery," *Chemical Biology & Drug Design*, 2014. The dissertation author was one of three primary investigators, who each contributed equally to performing research and writing this paper.

Supporting Information

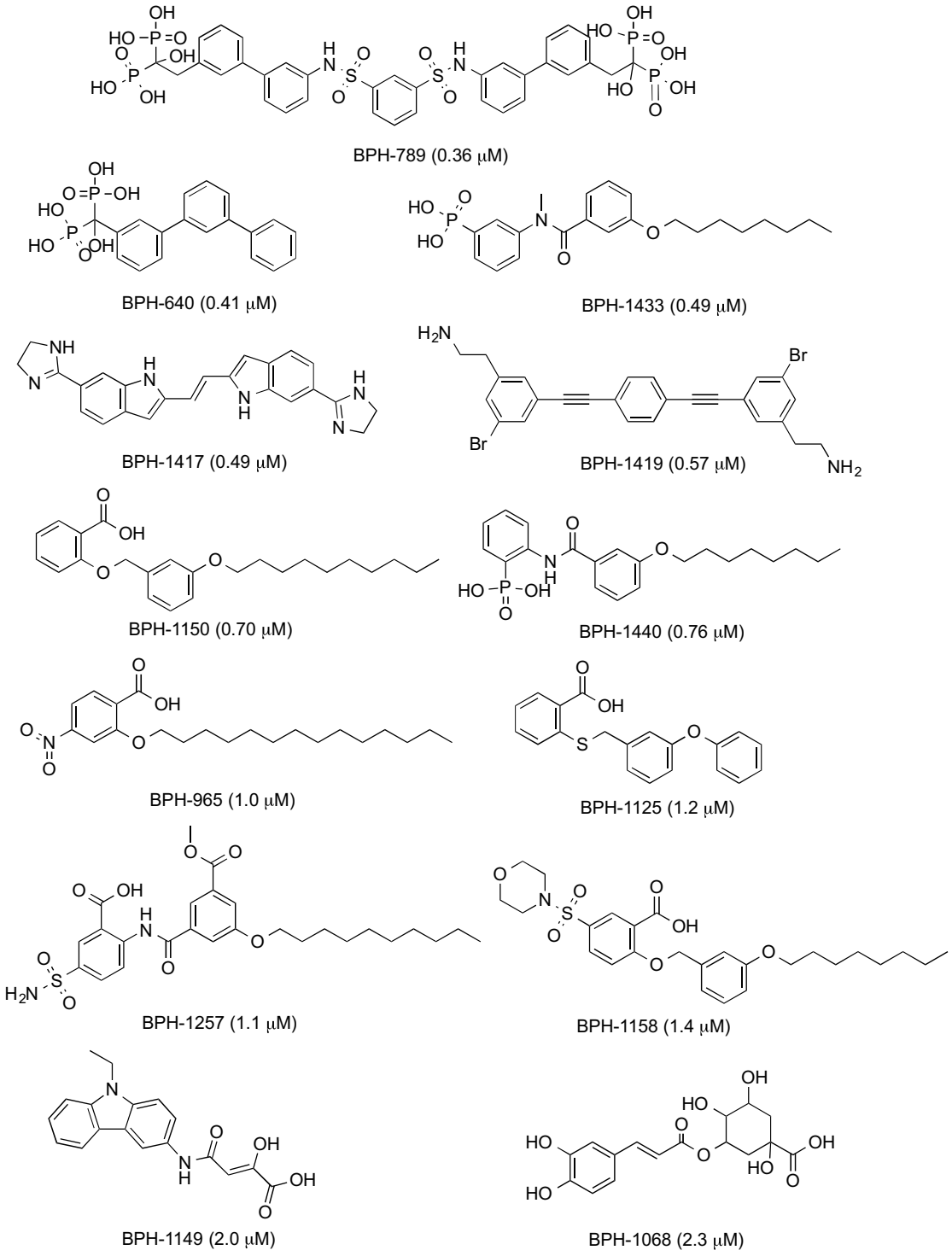


Figure 3-S1. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a GPP based assay.

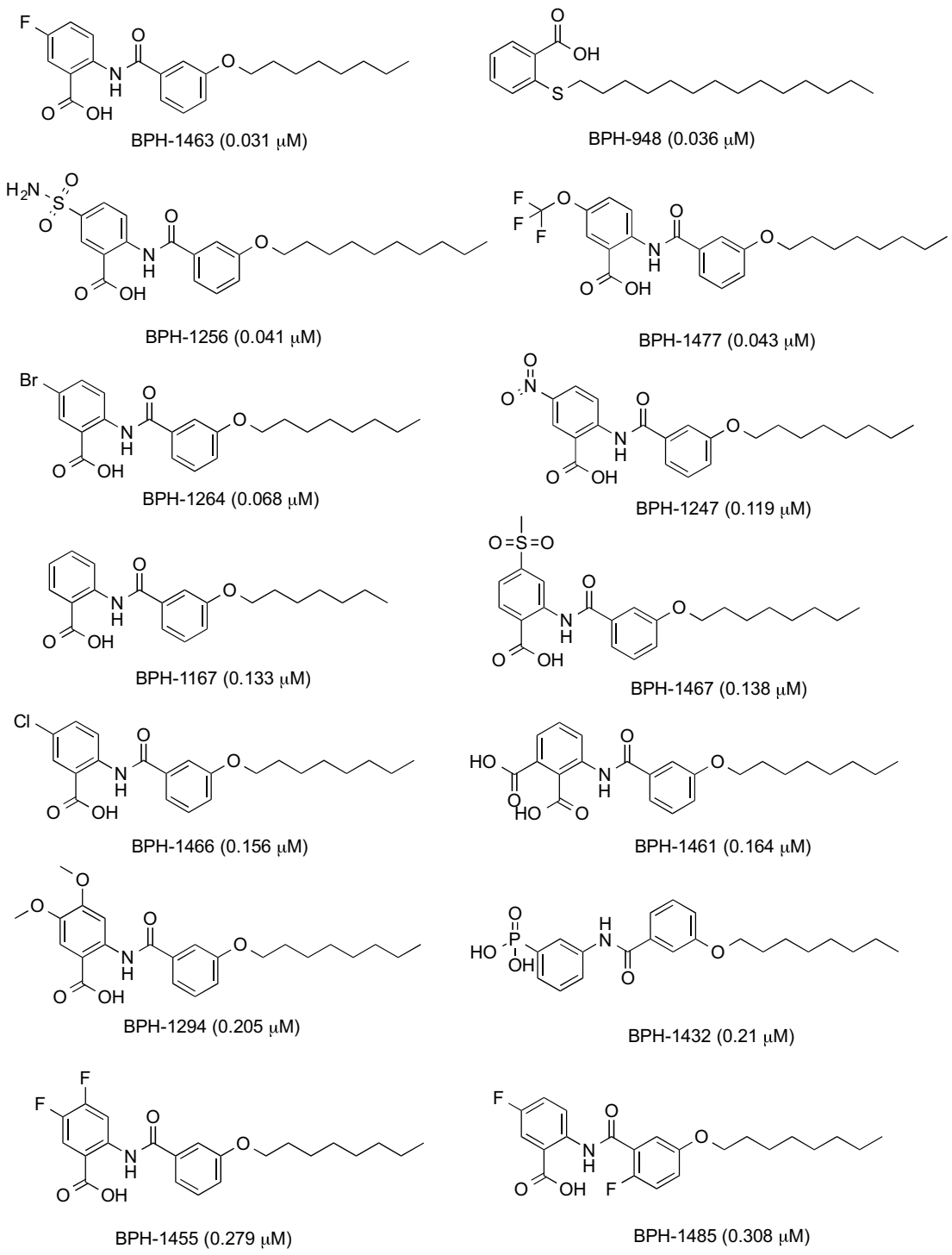


Figure 3-S1. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a GPP based assay, continued.

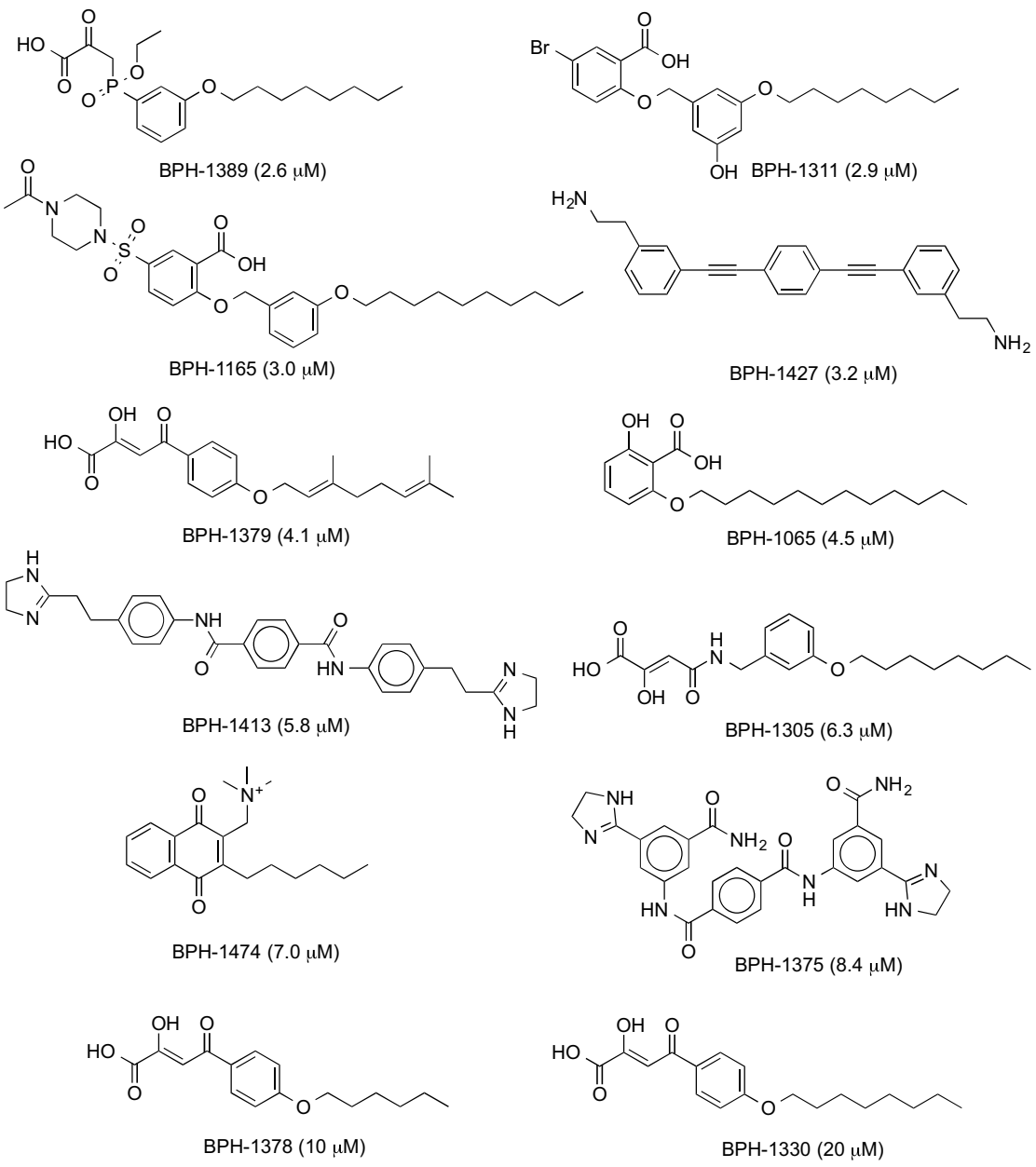


Figure 3-S1. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a GPP based assay, continued.

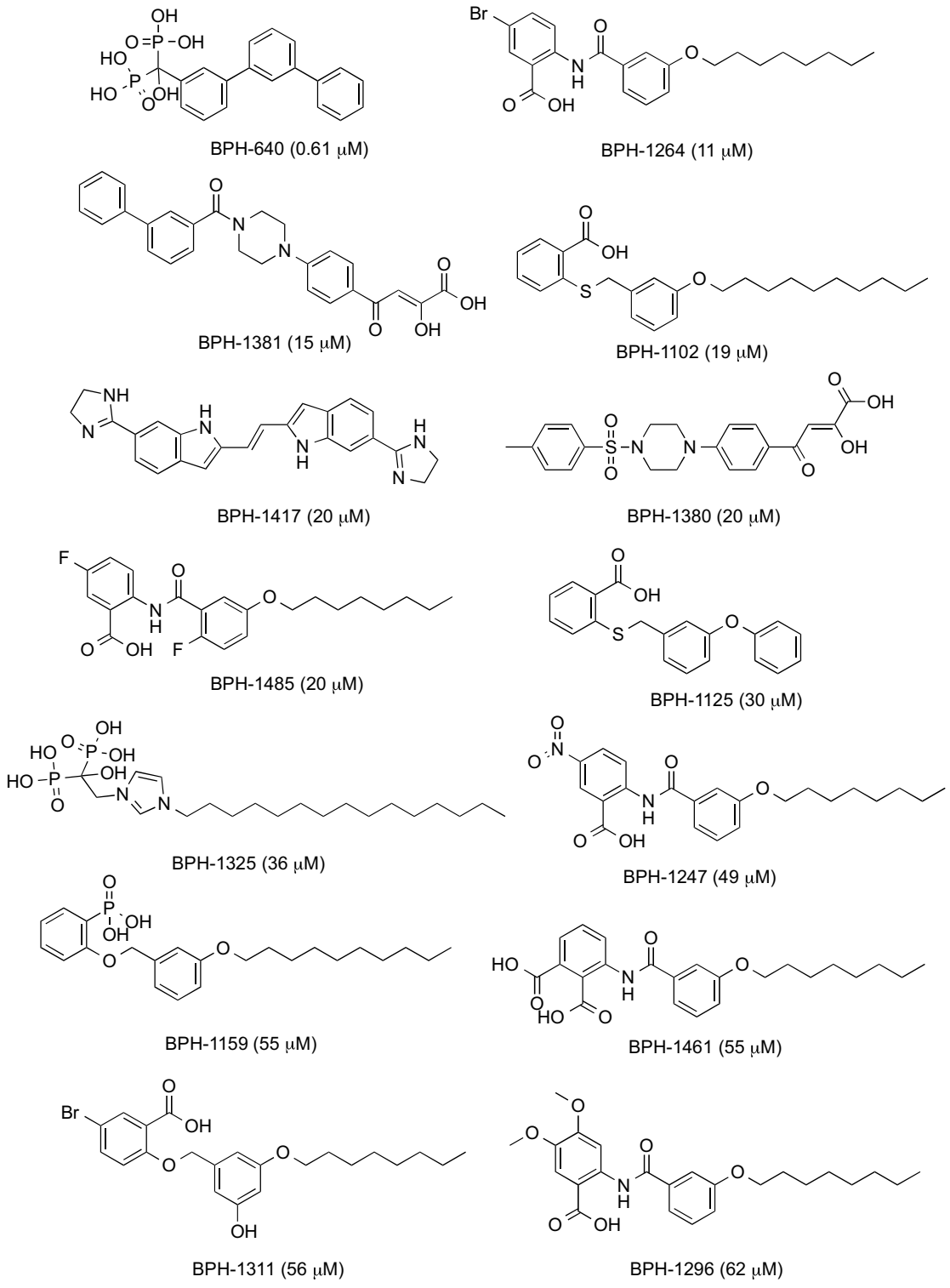


Figure 3-S2. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a *E,Z*-FPP based assay.

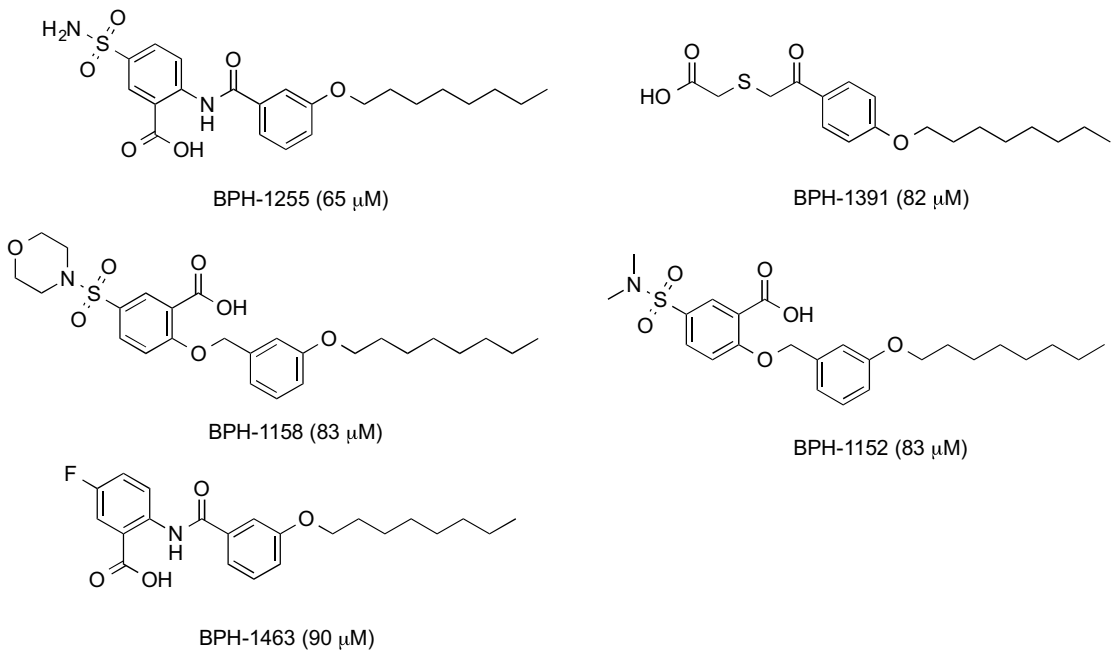


Figure 3-S2. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a *E,Z*-FPP based assay, continued.

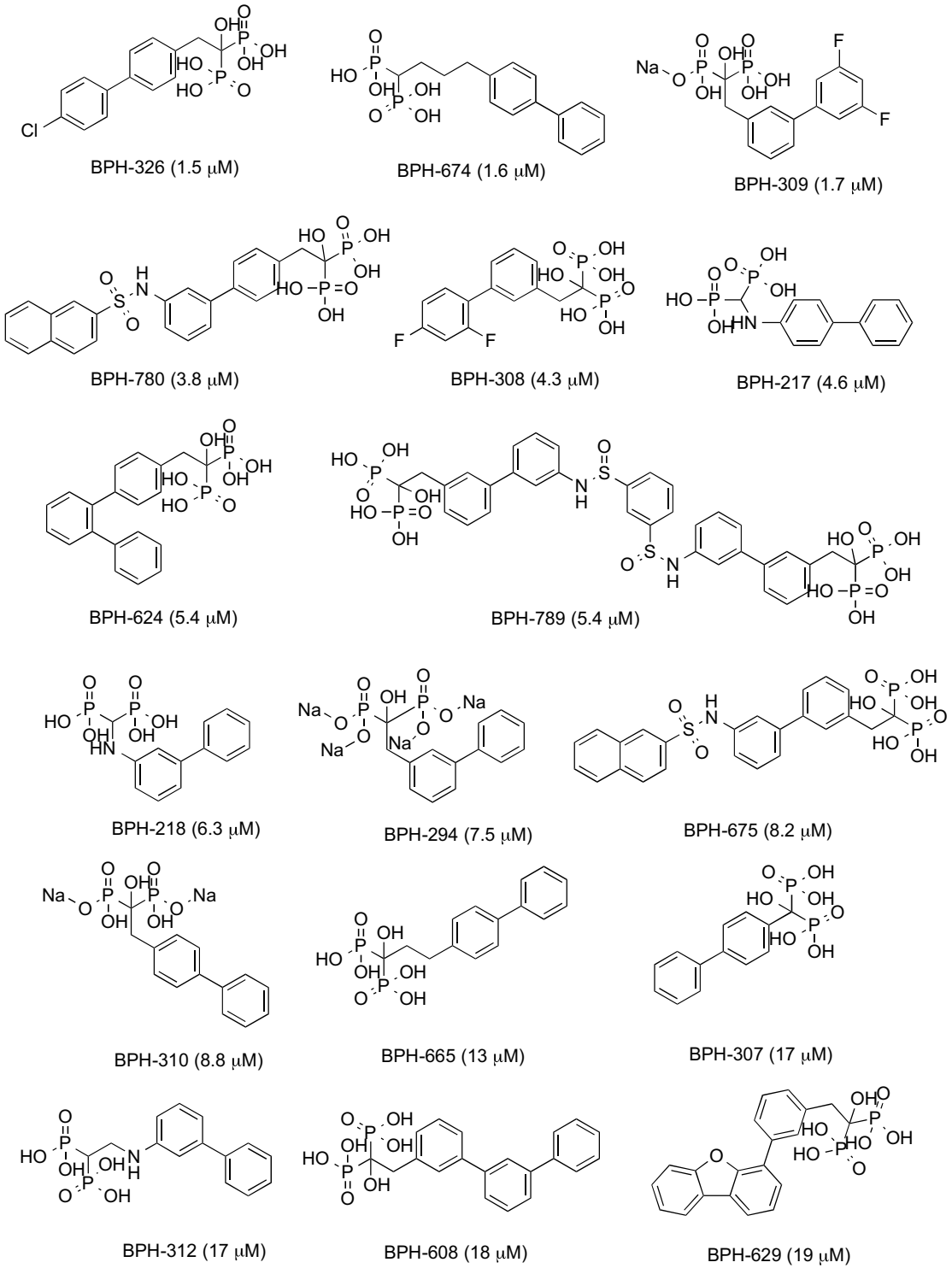


Figure 3-S3. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a *E,E*-FPP based assay.

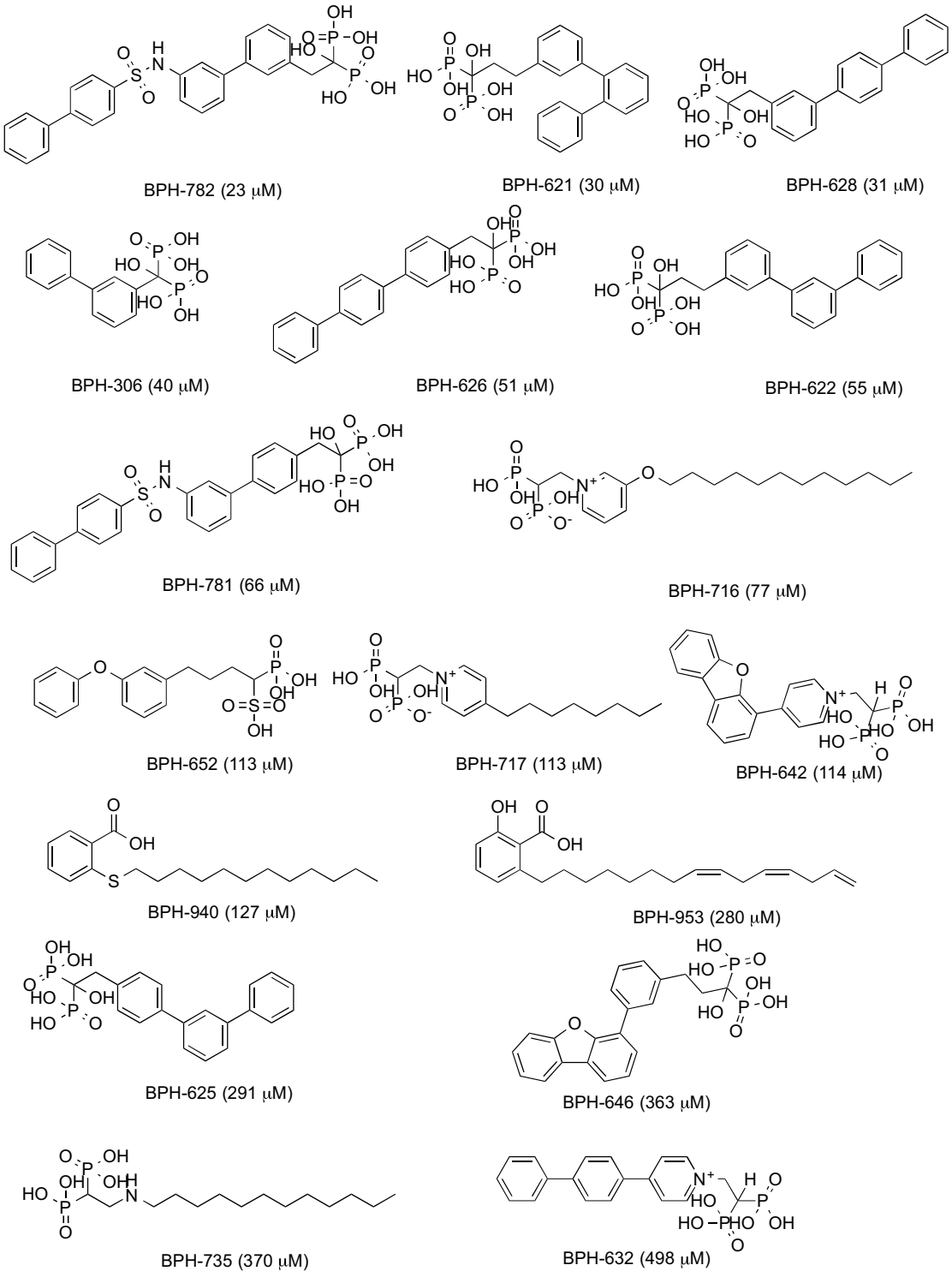
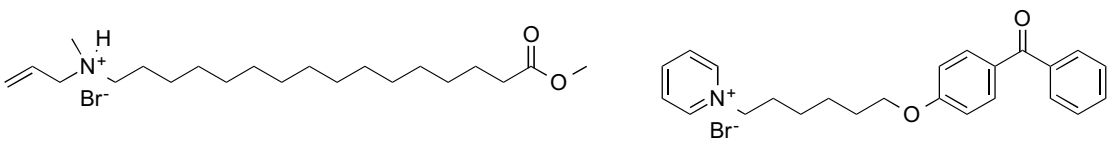


Figure 3-S3. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a *E,E*-FPP based assay, continued.



BPH-922 (620 μM)

BPH-921 (880 μM)

Figure 3-S3. Structures and IC_{50} values for *E,Z*-DPPS inhibitors in a *E,E*-FPP based assay, continued.

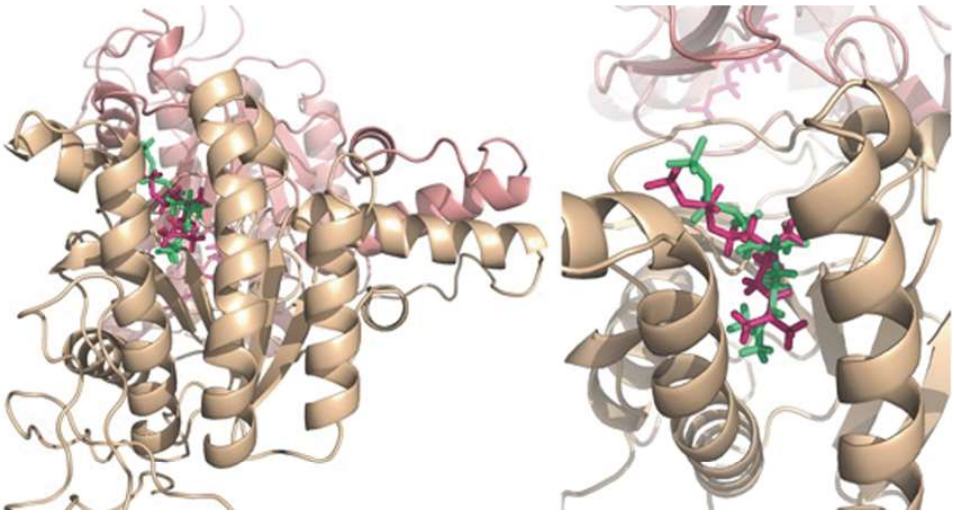


Figure 3-S4. Comparison of the docking pose of the co-crystallized ligand CITPP self-docked into the receptor grid center (green) with its crystallographic orientation in the 2VG3 crystal structure (pink).

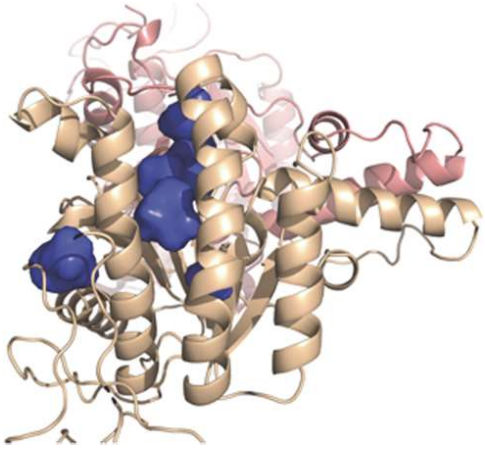


Figure 3-S5. Potential ligand binding sites of *E,Z*-DPPS predicted by FTMap.

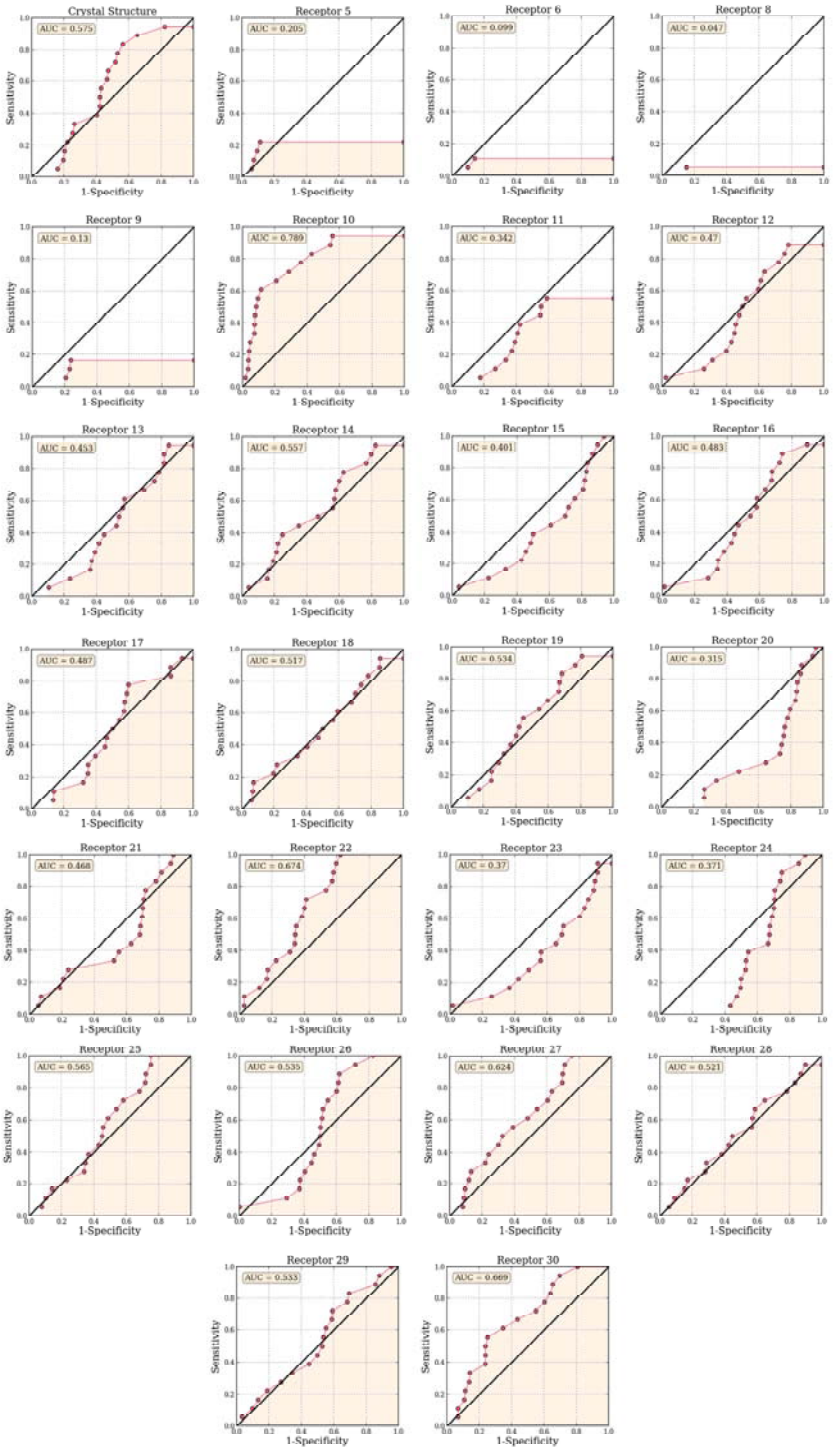


Figure 3-S6. ROC-AUC results for *E,Z*-DPPS inhibition using X-ray and MD structures with GPP as substrate.

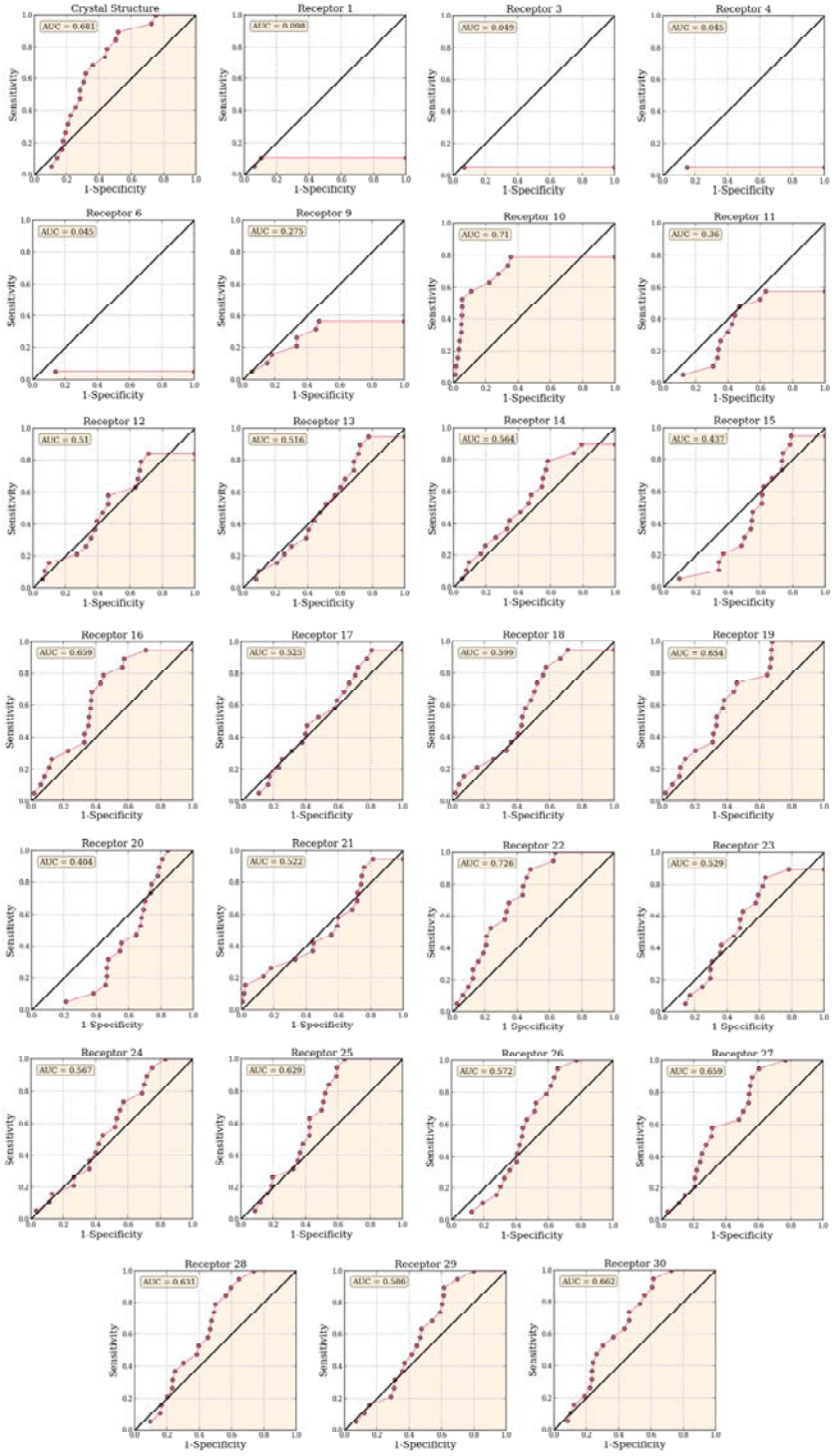


Figure 3-S7. ROC-AUC results for *E,Z*-DPPS inhibition using X-ray and MD structures with *E,Z*-FPP as substrate.

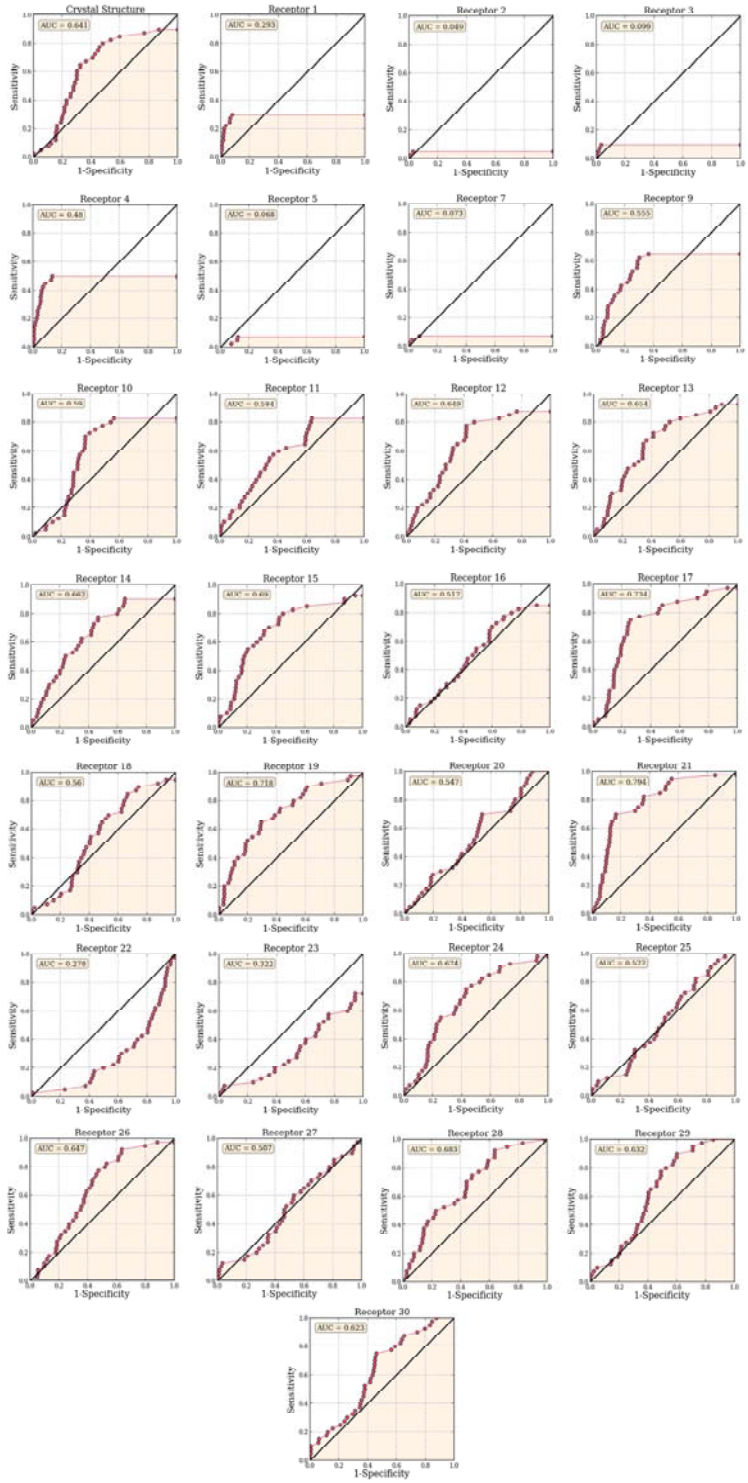


Figure 3-S8. ROC-AUC results for *E,Z*-DPPS inhibition using X-ray and MD structures with *E,E*-FPP as substrate.

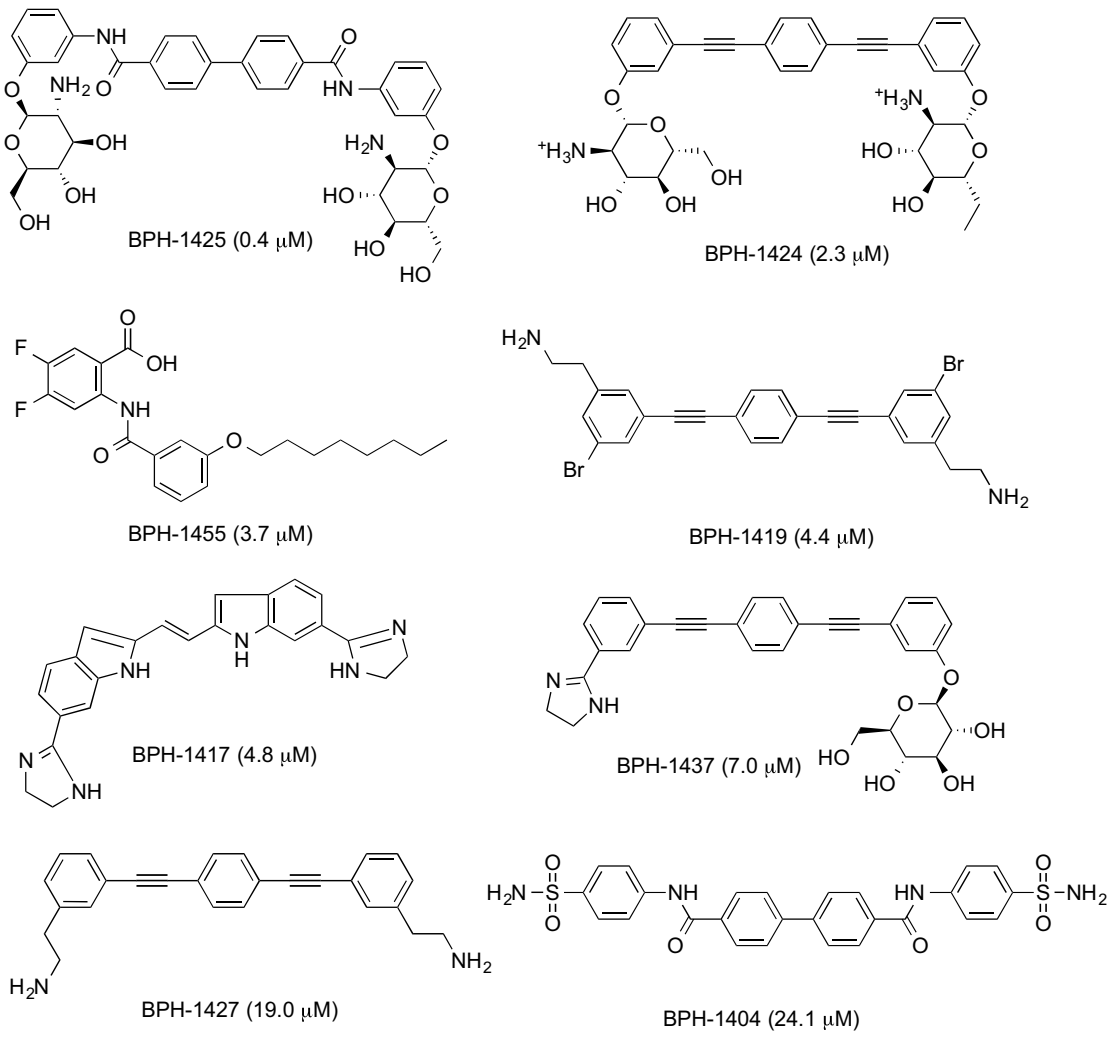


Figure 3-S9. Structures and IC_{50} values for *E,Z*-FPPS inhibitors.

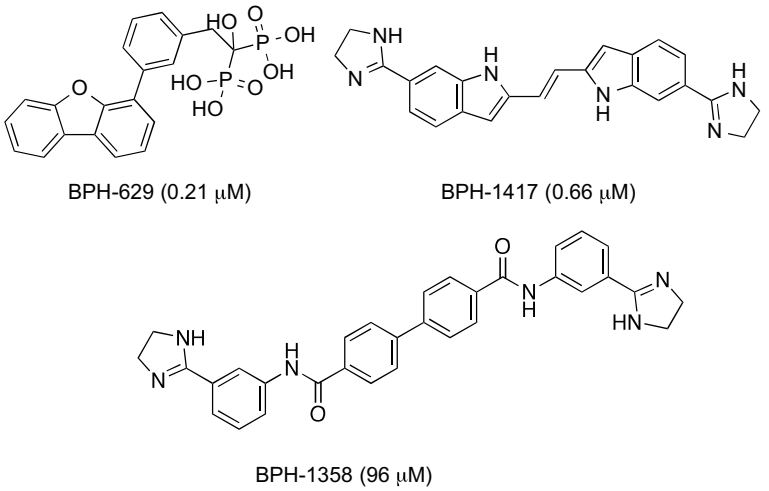


Figure 3-S10. Structures and IC_{50} values for Rv3378c inhibitors.

References

1. Chambers, H. F.; DeLeo, F. R. Waves of Resistance: Staphylococcus Aureus in the Antibiotic Era. *Nat. Rev. Microbiol.* **2009,** *7*, 629-641.

2. Fischbach, M. A.; Walsh, C. T. Antibiotics for Emerging Pathogens. *Science* **2009,** *325*, 1089-1093.

3. Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the Comprehensive, Rapid, and Accurate Prediction of the Favorable Tautomeric States of Drug-Like Molecules in Aqueous Solution. *J. Comput. Aided Mol. Des.* **2010,** *24*, 591-604.

4. Antimicrobial Resistance: Global Report on Surveillance.

5. Phillips, L. Infectious Disease: Tb's Revenge. *Nature* **2013,** *493*, 14.

6. Schulbach, M. C.; Brennan, P. J.; Crick, D. C. Identification of a Short (C15) Chain Z-Isoprenyl Diphosphate Synthase and a Homologous Long (C50) Chain Isoprenyl Diphosphate Synthase in Mycobacterium tuberculosis. *J. Biol. Chem.* **2000,** *275*, 22876-22881.

7. Takayama, K.; Goldman, D. S. Enzymatic Synthesis of Mannosyl-1-Phosphoryl-Decaprenol by a Cell-Free System of Mycobacterium tuberculosis. *J. Biol. Chem.* **1970,** *245*, 6251-6257.

8. Peukert, S.; Sun, Y.; Zhang, R.; Hurley, B.; Sabio, M.; Shen, X.; Gray, C.; Dzink-Fox, J.; Tao, J.; Cebula, R. Design and Structure-Activity Relationships of Potent and Selective Inhibitors of Undecaprenyl Pyrophosphate Synthase (UPPS): Tetramic, Tetronic Acids and Dihydropyridin-2-Ones. *Bioorg. Med. Chem. Lett.* **2008,** *18*, 1840-1844.

9. Guo, R.-T.; Cao, R.; Liang, P.-H.; Ko, T.-P.; Chang, T.-H.; Hudock, M. P.; Jeng, W.-Y.; Chen, C. K.-M.; Zhang, Y.; Song, Y. Bisphosphonates Target Multiple Sites in Both Cis-and Trans-Prenyltransferases. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *104*, 10022-10027.

10. Zhu, W.; Zhang, Y.; Sinko, W.; Hensler, M. E.; Olson, J.; Molohon, K. J.; Lindert, S.; Cao, R.; Li, K.; Wang, K., et al. Antibacterial Drug Leads Targeting Isoprenoid Biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **2013,** *110*, 123-128.

11. Jeong, Y. C.; Anwar, M.; Bikadi, Z.; Hazai, E.; Moloney, M. G. Natural Product Inspired Antibacterial Tetramic Acid Libraries with Dual Enzyme Inhibition. *Chem. Sci.* **2013,** *4*, 1008-1015.

12. Crick, D. C.; Schulbach, M. C.; Zink, E. E.; Macchia, M.; Barontini, S.; Besra, G. S.; Brennan, P. J. Polyprenyl Phosphate Biosynthesis in Mycobacterium tuberculosis and Mycobacterium Smegmatis. *J. Bacteriol.* **2000,** *182*, 5771-5778.

13. Sassetti, C. M.; Boyd, D. H.; Rubin, E. J. Genes Required for Mycobacterial Growth Defined by High Density Mutagenesis. *Mol. Microbiol.* **2003,** *48*, 77-84.

14. Chan, H.-C.; Feng, X.; Ko, T.-P.; Huang, C.-H.; Hu, Y.; Zheng, Y.; Bogue, S.; Nakano, C.; Hoshino, T.; Zhang, L., et al. Structure and Inhibition of Tuberculosinol Synthase and Decaprenyl Diphosphate Synthase from Mycobacterium tuberculosis. *J. Am. Chem. Soc.* **2014,** *136*, 2892-2896.

15. Layre, E.; Lee, H. J.; Young, D. C.; Martinot, A. J.; Buter, J.; Minnaard, A. J.; Annand, J. W.; Fortune, S. M.; Snider, B. B.; Matsunaga, I., et al. Molecular Profiling of Mycobacterium tuberculosis Identifies Tuberculosinyl Nucleoside Products of the Virulence-Associated Enzyme Rv3378c. *Proc. Natl. Acad. Sci. U.S.A.* **2014,** *111*, 2978-2983.

16. Hoshino, T.; Nakano, C.; Ootsuka, T.; Shinohara, Y.; Hara, T. Substrate Specificity of Rv3378c, an Enzyme from Mycobacterium tuberculosis, and the Inhibitory Activity of the Bicyclic Diterpenoids against Macrophage Phagocytosis. *Org. Biomol. Chem.* **2011,** *9*, 2156-2165.

17. Mann, F. M.; Xu, M. M.; Chen, X. M.; Fulton, D. B.; Russell, D. G.; Peters, R. J. Edaxadiene: A New Bioactive Diterpene from Mycobacterium tuberculosis. *J. Am. Chem. Soc.* **2009,** *131*, 17526-17527.

18. Prach, L.; Kirby, J.; Keasling, J. D.; Alber, T. Diterpene Production in Mycobacterium tuberculosis. *Febs Journal* **2010,** *277*, 3588-3595.

19. Pethe, K.; Swenson, D. L.; Alonso, S.; Anderson, J.; Wang, C.; Russell, D. G. Isolation of Mycobacterium tuberculosis Mutants Defective in the Arrest of Phagosome Maturation. *Proc. Natl. Acad. Sci. U.S.A.* **2004,** *101*, 13642-13647.

20. Shimizu, N.; Koyama, T.; Ogura, K. Molecular Cloning, Expression, and Purification of Undecaprenyl Diphosphate Synthase. No Sequence Similarity between E- and Z-Prenyl Diphosphate Synthases. *J. Biol. Chem.* **1998,** *273*, 19476-19418.

21. Wang, W.; Dong, C.; McNeil, M.; Kaur, D.; Mahapatra, S.; Crick, D. C.; Naismith, J. H. The Structural Basis of Chain Length Control in Rv1086. *J. Mol. Biol.* **2008,** *381*, 129-140.

22. Sinko, W.; de Oliveira, C.; Williams, S.; van Wynsberghe, A.; Durrant, J. D.; Cao, R.; Oldfield, E.; McCammon, J. A. Applying Molecular Dynamics Simulations to Identify Rarely Sampled Ligand-Bound Conformational States of Undecaprenyl Pyrophosphate Synthase, an Antibacterial Target. *Chem. Biol. Drug Des.* **2011,** *77*, 412-420.

23. Durrant, J. D.; Cao, R.; Gorfe, A. A.; Zhu, W.; Li, J.; Sankovsky, A.; Oldfield, E.; McCammon, J. A. Non-Bisphosphonate Inhibitors of Isoprenoid Biosynthesis Identified Via Computer-Aided Drug Design. *Chem. Biol. Drug Des.* **2011,** *78*, 323-332.

24. Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005,** *26*, 1668-1688.

25. Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M., et al. *Amber 11*, University of California, San Francisco: 2010.

26. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** *79*, 926-935.

27. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006,** *65*, 712-725.

28. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99sb Protein Force Field. *Proteins* **2010,** *78*, 1950-1958.

29. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.

30. Schulbach, M. C.; Mahapatra, S.; Macchia, M.; Barontini, S.; Papi, C.; Minutolo, F.; Bertini, S.; Brennan, P. J.; Crick, D. C. Purification, Enzymatic Characterization, and Inhibition of the Z-Farnesyl Diphosphate Synthase from Mycobacterium tuberculosis. *J. Biol. Chem.* **2001,** *276*, 11624-11630.

31. Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Interpretaion of Protein pKa Values. *Proteins* **2005,** *61*, 704-721.

32. Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein-Ligand Complexes. *Proteins* **2008,** *73*, 765-783.

33. Olsson, M. H. M.; Søndergard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA 3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011,** *7*, 525-537.

34. Søndergard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011,** *7*, 2284-2295.

35. Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating pKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005,** *33*, W368-371.

36. Myers, J.; Grothaus, G.; Narayanan, S.; Onufriev, A. A Simple Clustering Algorithm Can Be Accurate Enough for Use in Calculation of Pks in Macromolecules. *Proteins* **2006,** *63*, 928-938.

37. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulation. *Nucleic Acids Res.* **2012,** *40*, 537-541.

38. Schafmeister, C. E. A. F.; Ross, W. S.; Romanovski, V. *Leap*, University of california, San Francisco: 1995.

39. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T. K., K. N.; Burant, J. C.; Millam, J. M., et al. *Gaussian 03*, Gaussian, Inc.: Wallingford, CT, 2004.

40. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006,** *25*, 247-260.

41. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004,** *25*, 1157-1174.

42. Durrant, J. D.; de Oliveira, C. A. F.; McCammon, J. A. Povme: An Algorithm for Measuring Binding-Pocket Volumes. *J. Mol. Graph. Model.* **2011,** *29*, 773-776.

43. Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. Bio3d: An R Package for the Comparative Analysis of Protein Structures. *Bioinformatics* **2006,** *22*, 2695-2696.

44. *Maestro*, Version 9.3; Schrödinger, LLC: New York, NY, 2012.

45. *Ligprep*, Version 2.5; Schrödinger, LLC: New York, NY, 2012.

46. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulation. *Nucleic Acids Res.* **2012,** *40 (W1)*, W537-541.

47. Shelley, J. C.; Cholleti, A.; Frye, L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pKa Prediction and Protonation State Generation for Drug-Like Molecules. *J. Comput. Aided Mol. Des.* **2007,** *21*, 681-691.

48. *Glide, Version 5.8*, Schrödinger, LLC: New York, NY, 2012.

49. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M., et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004,** *47*, 1739-1749.

50. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004,** *47*, 1750-1759.

51. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006,** *49*, 6177-6196.

52. Webb, M. R. A Continuous Spectrophotometric Assay for Inorganic Phosphate and for Measuring Phosphate Release Kinetics in Biological Systems. *Proc. Natl. Acad. Sci. U.S.A.* **1992,** *89*, 4884-4887.

53. Chan, H. C.; Feng, X. X.; Ko, T. P.; Huang, C. H.; Hu, Y. M.; Zheng, Y. Y.; Bogue, S.; Nakano, C.; Hoshino, T.; Zhang, L. L., et al. Structure and Inhibition of Tuberculosinol Synthase and Decaprenyl Diphosphate Synthase from Mycobacterium tuberculosis. *Journal of the American Chemical Society* **2014,** *136*, (7), 2892-2896.

54. Amaro, A. E.; Baron, R.; McCammon, J. A. An Improved Relaxed Complex Scheme for Receptor Flexibility in Computer-Aided Drug Design. *J. Comput. Aided Mol. Des.* **2008,** *22*, (693-705).

55. Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-Based Identification of Druggable 'Hot Spots' of Proteins Using Fourier Domain Correlation Techniques. *Bioinformatics* **2009,** *25*, 621-627.

56. Fawcett, T. An Introduction to Roc Analysis. *Pattern Recogn. Lett.* **2006,** *27*, 861-874.

57. Teng, K. H.; Liang, P. H. Structures, Mechanisms and Inhibitors of Undecaprenyl Diphosphate Synthase: A Cis-Prenyltransferase for Bacterial Peptidoglycan Biosynthesis. *Bioorg. Chem.* **2012,** *43*, 51-57.

58. Wallrapp, F. H.; Pan, J.-J.; Ramamoorthy, G.; Almonacid, D. E.; Hillerich, B. S.; Seidel, R.; Patskovsky, Y.; Babbitt, P. C.; Almo, S. C.; Jacobson, M. P. Prediction of Function for the Polyprenyl Transferase Subgroup in the Isoprenoid Synthase Superfamily. *Proc. Natl. Acad. Sci. U.S.A.* **2013,** *110*, 1196-1202.

59. Panchal, R. G.; Lane, D.; Boshoff, H. I.; Butler, M. M.; Moir, D. T.; Bowlin, T. L.; Bavari, S. Bis-Imidazolinylindoles Are Active against Methicillin-Resistant Staphylococcus Aureus and Multidrug-Resistant Mycobacterium tuberculosis. *J. Antibiot.* **2013,** *66*, 47-49.

60. Morphy, J. R.; Harris, C. J. Designing Multi-Target Drugs. *Designing Multi-Target Drugs* **2012**, 1-365.

Chapter 4

Protocols utilizing constant pH molecular dynamics to compute pH-dependent binding free energies

Abstract

In protein-ligand binding, the electrostatic environments of the two binding partners may vary significantly in bound and unbound states, which may lead to protonation changes upon binding. In cases where ligand binding results in a net uptake or release of protons, the free energy of binding is pH-dependent. Nevertheless, conventional free energy calculations and molecular docking protocols typically do not rigorously account for changes in protonation that may occur upon ligand binding. To address these shortcomings, we present a simple methodology based on Wyman's binding polynomial formalism to account for the pH dependence of binding free energies and demonstrate its use on cucurbit[7]uril (CB[7]) host-guest systems. Using constant pH molecular dynamics and a reference binding free energy that is taken either from experiment or from thermodynamic integration computations, the pH-dependent binding free energy is determined. This computational protocol accurately captures the large pK_a shifts observed experimentally upon CB[7]:guest association and reproduces experimental binding free energies at different levels of pH. We show that incorrect assignment of fixed protonation states in free energy computations can give errors of > 2 kcal/mol in

these host-guest systems. Use of the methods presented here avoids such errors, thus suggesting their utility in computing proton-linked binding free energies for protein-ligand complexes.

Introduction

The changes in the electrostatic environment that accompany binding of small molecules, nucleic acids, or other proteins may thus induce changes in the protonation states of titratable groups in the protein.[1-8] Recently, Aguilar *et al*. conducted a computational survey of various protein-protein, protein-small molecule, and protein-nucleic acid complexes to ascertain the prevalence of protonation change in the protein receptor upon biomolecular association. Notably, in 60 percent of the protein-small molecule complexes considered, at least one titratable residue in the protein was found to assume different protonation states in its free and bound states.[9] Furthermore, protonation changes that accompany small molecule binding to proteins are not limited to the protein partner: an estimated 60-80 percent of orally-administered drugs are weak acids or bases, whose protonation states can also be tuned by the cellular pH and electrostatic environment of their protein binding partners.[10-13] In cases where protein-ligand binding accompanies a net transfer of protons to either binding partner, the binding process is pH-dependent, *i.e.* the observed binding free energy is a function of pH.

Conventionally, both computational docking and more rigorous free energy computations, such as the thermodynamic integration (TI) and free energy perturbation (FEP) methods, employ fixed protonation states that are identical for free and bound states in the computation of binding affinities. Clearly, in cases where ligand binding is linked to the (un)binding of protons, such approximations will lead to error. Improper assignment of protonation states in binding free energy computations may result in significant errors, making correct assignment of pK_a and protonation state essential to obtaining accurate free energies.

As stated, simulations of protein-ligand systems are typically preceded by the assignment of fixed protonation states to titratable groups on the two binding partners, often using programs such as H++[14-16] and PROPKA[17-20] to do so. Further, docking studies often employ empirical prediction algorithms, which sometimes use Hammett and Taft relations, to assign fixed protonation states to the

free ligands being docked.[21, 22] These approaches, however, fail to account for changes in protonation that may follow from the altered electrostatic environment surrounding the two binding partners upon complex formation. Several computational methods, however, have been developed that permit the protonation of titratable residues to respond to changes in the electrostatic environment.[2, 23-28] For instance, various flavors of constant pH molecular dynamics (CpHMD) methodologies have emerged to incorporate pH as an added external thermodynamic parameter to conventional molecular dynamics (MD) simulations, allowing fluctuations in the protonation of titratable residues to accompany conformational sampling.[29-34] To date, CpHMD simulations have been used to successfully predict pK_a values of titratable groups in proteins[29-37] and nucleic acids,[38-40] as well as to explain the mechanism behind the pH-dependent conformational changes critical to the function of proteins such as nitrophorin[41] and rhodopsin.[42]

The CpHMD method provides a framework through which the pH dependence of binding processes can be examined. To the best of our knowledge, there is currently no standard protocol available to rigorously account for proton-linked ligand binding. Multiple experimental and computational groups, however, have utilized the binding polynomial formalism devised by Wyman[43] to calculate the changes in binding free energy that accompany binding-induced protonation changes for both protein-protein[3, 4, 44] and protein-nucleic acid binding.[39, 45, 46] Motivated by Mason and Jensen's usage of this binding polynomial formalism to estimate the free energies of binding for protein-protein complexes using the PROPKA web server,[4] we adopt a similar approach in conjunction with the CpHMD method by Mongan *et al*.[33] to obtain pH-dependent free energy profiles *in silico* for the binding of small molecules to the cucurbit[7]uril (CB[7]) host.

CB[7] is a synthetic molecule with seven repeating glycoluril units bridged by methylene groups (Figure 4-1).[47, 48] This 7-fold symmetric host has gained much attention due to its ability to encapsulate drug-like small molecules with high affinity as a stable host-guest complex.[49-55] Benzimidazole (BZ) and a series of its derivatives (Figure 4-2) comprise a class of widely used fungicides and anthelmintic drugs[56-58] that have been shown to bind to the CB[7] host and undergo the pK_a shifts as large as 4 pK units upon complex formation (Table 4-1).[59] At neutral pH, these weakly

acidic guests are predominantly deprotonated when free in solution, but each binds a single proton upon encapsulation by CB[7]. Both the acid/base behaviors of BZ-derived guests and the small size and relative rigidity of CB[7] compared to a typical biomolecule make the CB[7]:BZ complexes ideal model systems to test theoretical methods for computing pH-dependent binding free energies.

In this work, we accurately reproduce the pK_a shifts of the various BZ derivatives upon binding to CB[7], using CpHMD simulations. Coupling these pK_a data with reference binding free energies taken either from experiment or from TI computations allows us to obtain a full description of CB[7]:guest binding free energies as functions of pH. Additionally, we show that improper assignment of guest protonation states in binding free energy computations can produce errors in excess of 2 kcal/mol at neutral pH, highlighting the importance of accurately accounting for the pH effects in free energy calculations or docking.

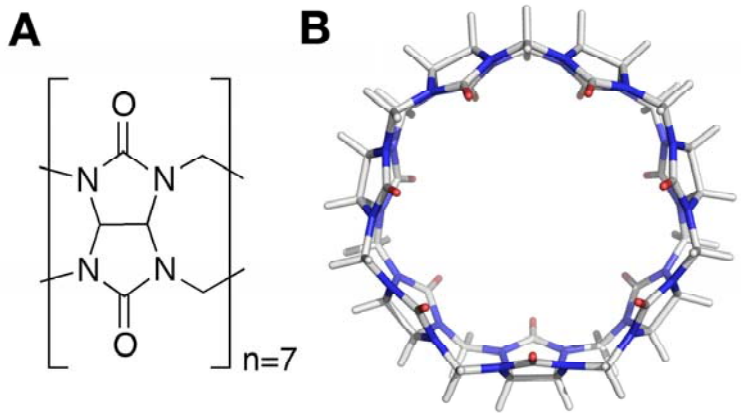


Figure 4-1. Structure of cucurbit[7]uril (CB[7]) host. (A) Glycoluril unit. (B) Top view of CB[7].

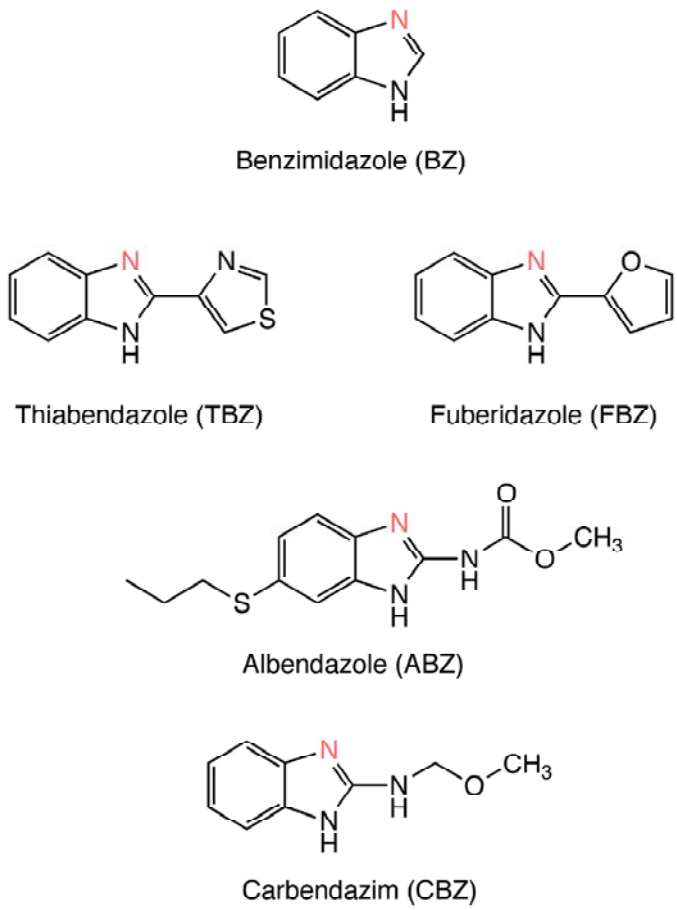


Figure 4-2. Chemical structures of benzimidazole and its derivatives.

Table 4-1. Experimental pK_a shifts of benzimidazole guests upon binding to CB[7].[59] pK_a^F denotes the pK_a of the free guest and pK_a^C represents the pK_a of the guest in complex with CB[7].

Guest	**pK_a^F**	**$pK_a^{C,exp}$**	**ΔpK**
BZ	5.5	9.0	3.5
TBZ	4.6	8.6	4.0
FBZ	4.8	8.6	3.8
ABZ	3.5	6.1	2.6
CBZ	4.5	7.0	2.5

Theory

Binding polynomial formalism for computing the pH dependence of binding free energies

Mason and Jensen recently examined the pH dependence of protein-protein binding[4] through an application of the binding polynomial formalism developed by Wyman[43] and used by Tanford to describe protein folding/unfolding.[60] Following the theoretical foundations of these groups, the binding of a titratable ligand (L) to a general macromolecular receptor (R) can be considered through a general equation for ligand association governed by the apparent equilibrium constant, K_{app}:

(1) $$R + \{L\} \overset{K_{app}}{\rightleftharpoons} \{RL\},$$

where the curly braces indicate that the ligand and complex (RL) ensembles may contain different protonated forms of the titratable ligand species. In the case of a ligand with a single titratable site binding to CB[7], which itself does not titrate in the biological range of pH levels, K_{app} can be written as

(2) $$K_{app} = \frac{[LR] + [HLR^+]}{[R]([L] + [HL^+])},$$

where the concentrations, rather than activities, of the given species are reported assuming ideal dilute solutions. Building from the thermodynamic cycle used to describe the proton-linked ligand binding to CB[7] (Scheme 4-1), K_{app} can be rewritten according to Eq. 3, in which the concentrations of all species are presented in binding polynomials with respect to the concentrations of the deprotonated complex and ligand species:

$$\begin{array}{ccc} R + L & \overset{K_b^\circ}{\rightleftharpoons} & R:L \\ K_a^F \downarrow\uparrow & & \uparrow\downarrow K_a^C \\ R + HL^+ & \overset{K_b^{+,\circ}}{\rightleftharpoons} & R:HL^+ \end{array}$$

Scheme 4-1. Thermodynamic cycle for complex formation between a receptor (R) and a titratable ligand (L).

(3) $$K_{app} = \frac{[LR]\left(1+\frac{[HLR^+]}{[LR]}\right)}{[R][L]\left(1+\frac{[HL^+]}{[L]}\right)} = K_b^\circ \frac{\left(1+\frac{[HLR^+]}{[LR]}\right)}{\left(1+\frac{[HL^+]}{[L]}\right)}.$$

Using the acid dissociation constants for the free ligand (K_a^F) and ligand-receptor complex (K_a^C), as illustrated by the vertical reactions in Scheme 4-1 (Eqs. 4 and 5):

(4) $$K_a^F = \frac{[HL^+]}{[L]a_{H^+}}$$

(5) $$K_a^C = \frac{[HLR^+]}{[LR]a_{H^+}},$$

where the proton activity is denoted by a_{H^+}, Eq. 3 can be rewritten in terms of the overall free energy of binding for the ligand L to the receptor R ($\Delta G°$):

(6) $$\Delta G^\circ(pH) = -k_B T \ln K_b^\circ - k_B T \ln\left(\frac{1+10^{pK_a^C - pH}}{1+10^{pK_a^F - pH}}\right) = \Delta G^\circ_{ref} - k_B T \ln\left(\frac{1+10^{pK_a^C - pH}}{1+10^{pK_a^F - pH}}\right),$$

where the proton activity and acid dissociation constants have been converted to their respective logarithmic constants, pH and pK_a. The pH dependence of the binding free energy can thus be obtained having only the pK_a values of the ligand molecule free in solution (pK_a^F) and in complex with the receptor (pK_a^C), as well as the free energy of binding for a reference reaction shown in Eq. 6 (the top reaction in Scheme 4-1), $\Delta G°_{ref}$, in which there is no net uptake or release of protons. This formalism for obtaining $\Delta G°$ as a function of pH can further be applied to cases where multiple ligand and receptor groups titrate in the pH range considered, assuming that proton binding occurs independently. In other words, Eq. 6 can only be applied when all titratable groups are uncoupled from each other.

As protein active sites often contain multiple titratable groups whose protonation states are coupled to perform a given function, it will sometimes be wrong to assume that all titratable groups remain uncoupled upon ligand binding. In such cases, Wyman[43, 61] derived a relation between K_{app} and pH such that

(7) $$\frac{\partial \ln K_{app}}{\partial \ln[H^+]} = \Delta\nu_{H^+} = Z_{LR} - (Z_L + Z_R),$$

where, using the notation used by Tanford,[60] $\Delta\nu_{H+}$ is the change in the number of bound protons in the receptor-ligand complex, relative to the number of protons bound to the ligand and receptor individually. Utilizing the unit charge of a proton, this relation is equivalent to the difference in total charge, Z, between reactants and product in Eq. 1. With $\Delta Z = Z_{LR} - (Z_L + Z_R)$, integration of Eq. 7 provides a thermodynamic relation that holds for proton-linked ligand binding in cases where titratable sites may interact (Eq. 8):

$$\Delta G^{\circ}(\mathrm{pH}) = \Delta G^{\circ}_{\mathrm{ref,pH}} - k_B T \ln(10) \int_{\mathrm{pH}_{\mathrm{ref}}}^{\mathrm{pH}} \left\{ Z_{LR}(\mathrm{pH}) - Z_L(\mathrm{pH}) \right\} d\mathrm{pH} \,, \tag{8}$$

where Z_R is omitted since the CB[7] receptor under consideration does not titrate in the pH range considered in this study. Since the integration is performed with respect to pH in the second term in Eq. 8, the reference binding free energy corresponds to the binding free energy at a specific pH.

Both Eqs. 6 and 8 thus provide frameworks for computing the pH-dependent binding free energy by adding a correction term to the reference free energy of binding. In the case of Eq. 6, the reference free energy, $\Delta G°_{ref}$, is obtained for receptor-ligand binding with protonation states fixed, such that no net change of protonation occurs. Analogously, the reference free energy in Eq. 8 is required to be the free energy of binding at a given value of pH. These two reference free energies are not necessarily equivalent; however, the reference reaction can be chosen such that they have the same value.

Constant pH molecular dynamics

Baptista and co-workers developed constant pH molecular dynamics (CpHMD) with stochastic titration to enable concurrent sampling of both conformational and protonation spaces according to the semi-grand canonical ensemble.[32] Here, we use the simplified CpHMD formulation implemented in the standard release of AMBER 12[62] that is similar to Baptista's formulation except that the simulation is performed in implicit solvent with generalized Born electrostatics.[33] In this method, an MD simulation is propagated from initial sets of coordinates and protonation states. After a chosen number of MD steps, the simulation is halted, at which point a Monte Carlo (MC) step evaluates whether a random titratable residue in the system should change protonation states. The acceptance of this new

protonation state is contingent on the application of the Metropolis criterion to the computed transition free energy, ΔG_{trans}, obtained using Eq. 9, where pH enters as an external thermodynamic parameter and k_BT is the Boltzmann constant multiplied by the temperature of the system.

(9) $$\Delta G_{trans} = k_B T(pH - pK_{a,ref})\ln 10 + \Delta G_{elec} - \Delta G_{elec,ref}$$

For the value of pH at which the simulation is conducted, the difference in electrostatic free energy that accompanies the change in protonation being considered, ΔG_{elec}, is computed with respect to the difference in electrostatic free energy that accompanies the analogous change in protonation for a model compound, $\Delta G_{elec,ref}$, which has a known pK_a value ($pK_{a,ref}$). In this manner, any non-classical contributions to the transition free energy cancel. For a given CB[7]:guest system, the model compound that enters Eq. 9 is the guest molecule free in solution, its $pK_{a,ref}$ is the experimentally obtained pK_a value of the free guest (pK_a^F, Table 1), and $\Delta G_{elec,ref}$ is defined to be the electrostatic free energy that equally populates the protonated and deprotonated forms of the free guest when the solution pH is equal to the experimental pK_a of the free guest. If the transition is accepted, then MD is continued with the new protonation state for the titratable residue; otherwise, MD continues without change in the protonation state. Repeated application of these steps builds an ensemble of protonation states along the MD trajectory.

Use of constant pH molecular dynamics in the binding polynomial scheme

The CpHMD method is applied to obtain values for pK_a^C in Eq. 6 and ΔZ in Eq. 8 to provide pH-dependent correction terms to the reference binding free energies. In the case of Eq. 6, values of pK_a^C are obtained from simulating the CB[7]:guest system at a range of pH values. Each CpHMD simulation obtains a fractional protonation for the titratable guest being considered. By tabulating the fraction of deprotonated guest species (s) computed at each value of pH, application of the Hill equation can be used to predict pK_a^C as the midpoint of the titration, as well as the Hill coefficient, n (Eq. 10):

(10) $$s = \frac{1}{1+10^{n(pK_a - pH)}}.$$

This method can reliably extract the pKa when the titratable residue exhibits typical titration behavior.[63] In all fits, the Hill coefficient obtained is approximately one, which is anticipated in the absence of cooperativity.

In the case of Eq. 8, the partial charges for the guest free in solution, Z_L, and the partial charges for the guest in complex with CB[7], Z_{LR}, can similarly be obtained from CpHMD simulations. BZ and its derivatives have charges of +1 when protonated and 0 when deprotonated. Consequently, Z_{LR} and Z_L are equivalent to the fraction of protonated species (1 – s) obtained from CpHMD simulations performed on the CB[7]:guest complex and the free guest, respectively.

Methods

Parameterization of CB[7] and benzimidazole ligands for molecular dynamics simulations

Partial charges for the CB[7] host have previously been derived[64] using the restrained electrostatic potential (RESP) procedure,[65-67] conventionally used to parameterize nonstandard residues for molecular simulations performed with AMBER force fields. Analogously, the geometries of benzimidazole (BZ), albendazole (ABZ), carbendazim (CBZ), fuberidizole (FBZ) and thiabendazole (TBZ, Figure 4-2), were optimized at the B3LYP/6-31G(d) level of theory[68-71] using the Gaussian 09 suite of programs.[72] Subsequently, the electrostatic potentials (ESP) associated with the optimized geometries of these guests were computed using MK radii[73] at the HF/6-31G(d) level of theory. The ESPs of the different guest molecules were submitted to the antechamber module[67] in the AmberTools 12 suite of programs,[62] which applies the RESP procedure to extract atomic point charges for use in molecular dynamics (MD) simulations. All other CB[7] and guest ligand force field terms, including Lennard-Jones parameters, were taken from the general AMBER force field (GAFF).[74]

Docking of guest molecules to CB[7]

To generate starting coordinates for MD simulations of different CB[7]:guest complexes, the various BZ-derivatives were docked rigidly into the CB[7] cavity using the extra precision mode (XP) in Schrödinger's Glide program.[75-77] Each CB[7]:guest docking experiment yielded a single pose for the CB[7]:guest complex, and all guest molecules bound CB[7] similarly. For illustrative purposes, the resulting CB[7]:FBZ complex obtained from Glide is shown in Figure 4-3. It is also worth noting that the docked poses of deprotonated and protonated guests in complex with CB[7] are similar. The hydrophobic core of the BZ guests is encapsulated by the CB[7] cavity, orienting the ligands similarly regardless of protonation, while additional furanyl, thiazole, amido, or thioether R-groups seen in the BZ derivatives protrude outside of the entrance to CB[7]. All poses show good agreement with experiments.[59]

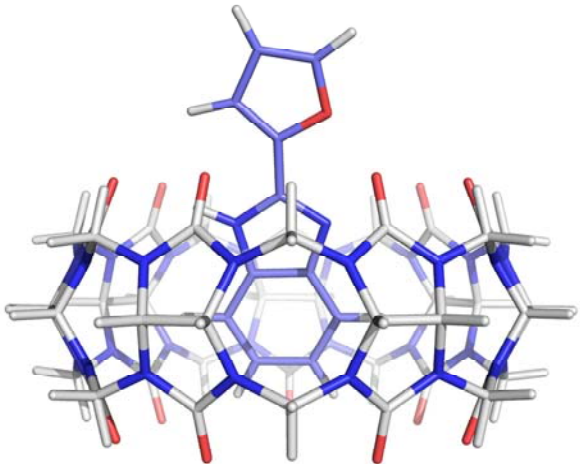


Figure 4-3. Structure of CB[7]:fuberidazole complex generated by docking.

Constant pH molecular dynamics simulation details

CpHMD simulations were performed using the AMBER 12 suite of programs for the range of pH values between 2 and 12 at increments of 0.5.[33, 62] All simulations employed the OBC generalized Born (GB) implicit solvent model (igb=5)[78] with a salt concentration of 0.1 M. Starting from the docked CB[7]:guest structures, all systems were minimized for 5,000 steps while applying positional restraints

to all heavy atoms with a force constant of 20 kcal/mol Å^2. Following minimization, the system was heated to 300 K over the course of 500 ps using a Langevin thermostat[79] while maintaining the positional restraints applied to all heavy atoms with a force constant of 5 kcal/mol Å^2. After heating, a 1 ns equilibration simulation was performed at 300K. Production simulations were then performed for 5 ns, with MC steps taken every 10 fs. In all equilibration and production steps, the bonds involving hydrogen were constrained using the SHAKE algorithm,[80] and a cutoff of 30 Å for the computation of nonbonded interactions was enforced.

Computing absolute binding free energy with thermodynamic integration

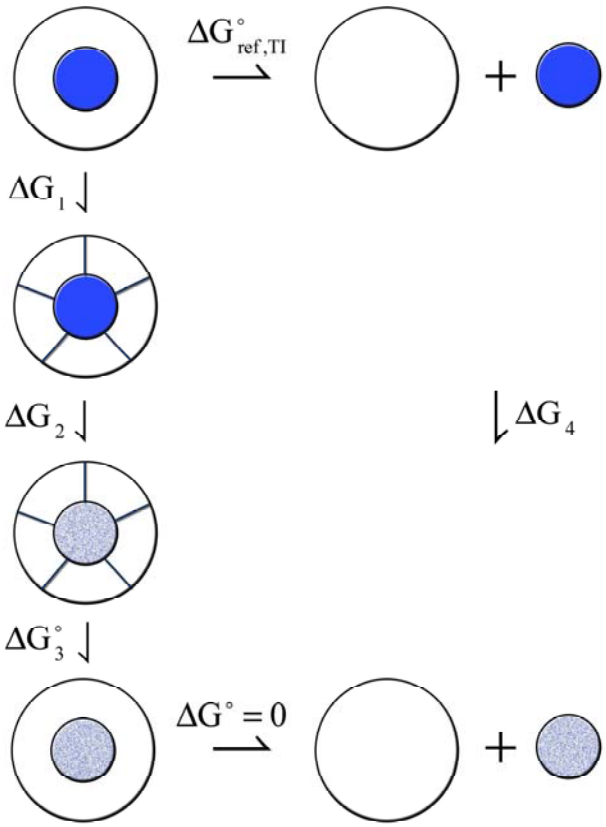
The calculation of the pH-dependent binding free energy requires a reference binding energy obtained either in the absence of protonation change (Eq. 6) or at a specified pH value (Eq. 8). TI computations were performed to obtain the absolute binding free energy between CB[7] and guest molecules that are deprotonated both free in solution and in complex using the double decoupling method.[81] In TI, the free energy change is evaluated as

(11) $$\Delta G_{0\to 1} = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle d\lambda ,$$

where U is the total potential energy of the system coupled to λ, which varies smoothly between the initial state of λ=0 and the final state of λ=1.[82] The reference binding free energy was obtained from the thermodynamic cycle shown in Scheme 2 and was calculated using

(12) $$\Delta G^{\circ}_{ref,TI} = -\Delta G_1 - \Delta G_2 - \Delta G^{\circ}_3 + \Delta G_4 ,$$

where ΔG_1 is the free energy for gradually turning on restraints (see below), ΔG_2 is the free energy for decoupling the guest while bound to the host in the presence of the restraints, $\Delta G_3°$ is the free energy for turning off the restraint and correcting for the standard state, and ΔG_4 is the solvation free energy for the decoupled guest (Scheme 4-2).



Scheme 4-2. Thermodynamic cycle for an absolute binding free energy calculation. The outer circle represents a CB[7] host and the inner blue circle shows a guest molecule in the reference deprotonated state. ΔG_1 is the free energy for gradually turning on the restraints; ΔG_2 is for decoupling the guest from the host in the presence of the restraints; $\Delta G_3°$ is the analytical correction for removing the restraints; and ΔG_4 is the solvation free energy for the guest.

The electrostatic and van der Waals (vdW) contributions to ΔG_2 and ΔG_4 are computed separately, the latter using the softcore potential algorithm.[83-85] To improve the convergence for these computations, the virtual bond algorithm developed by Karplus and co-workers was applied, where a set of restraints were used to fix the position and orientation of the guest relative to CB[7]. The free energy for turning on the restraint, ΔG_1, was computed using TI. The free energy for turning off the restraint, $\Delta G_3°$, was calculated using an analytical expression, which corrects for the presence of restraints and also accounts for the standard state:[86]

$$\Delta G^{\circ}_{3} = -k_B T \ln\left[\frac{8\pi^2 V^{\circ}}{r^2_{aA,^{\circ}} \sin\theta_{A,^{\circ}} \sin\theta_{B,^{\circ}}} \frac{(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{1/2}}{(2\pi k_B T)^3}\right] - k_B T \ln\frac{\sigma_{R\cdots L}}{\sigma_R \sigma_L} \cdot \quad (13)$$

Here V° is the standard state volume of 1661 Å^3 for ideal gas, r_{aA}, $\sin\theta_A$, $\sin\theta_B$ are the distance and angle values used for each restraint, having corresponding harmonic force constants (K's in Eq. 13), which are 5 kcal/mol Å^2 for the distance restraint and 20 kcal/mol rad^2 for the angle and dihedral restraints. The second term in Eq. 13 accounts for the symmetry in the system, where $\sigma_{R\cdots L}$, σ_R and σ_L are the symmetry numbers for the host-guest complex, CB[7] and the guest molecule, respectively. For our system, $\sigma_{R\cdots L}$ and σ_L are 1, and σ_R is 14.

The *pmemd* implementation of TI in AMBER 14 was used to calculate the reference binding free energy $\Delta G°_{ref,\ TI}$ for each guest to CB[7].[87, 88] The reference ionization state was chosen to be deprotonated because all experimental values of $\Delta G°_{ref}$ were measured with the guests deprotonated.[59] For the calculation of the electrostatic contribution to ΔG_2 and ΔG_4, 11 equally spaced λ values were used (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). For the calculation of the vdW contribution to ΔG_2 and ΔG_4, 21 λ values were used (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9, 0.925, 0.95, 0.975, 1.0). For the computation of the free energy for turning on the restraints, ΔG_1, 16 λ values were used (0.0, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). The unequal spacing of λ windows was needed to capture a smoother transition of $\partial U(\lambda)/\partial\lambda$ along the λ parameter and reduce errors in integration. Integration was performed numerically using the trapezoidal rule, and uncertainties in the free energies were propagated as standard deviations.

Each CB[7]:guest complex was solvated with TIP3P water[89] with a region of 12 Å in any direction using tleap program.[90] The system was minimized for 5,000 steps and heated to 300 K over 500 ps in the NVT ensemble using a Langevin thermostat,[79] followed by an equilibration for 500 ps in the NPT ensemble using a Berendsen barostat[91] with isotropic position scaling to bring the system to a stable density. All production simulations were performed in the NVT ensemble and were extended until the cumulative free energy computed for each individual transformation converges (changes in ΔG < 0.01 kcal/mol).

Results

Review of experimental results

Previously, Koner *et al*. observed enhancements in stabilities and solubilities of benzimidazole (BZ) derivatives upon encapsulation by the cucurbit[7]uril (CB[7]) host.[59] The authors obtained values of pK_a^F and pK_a^C by fitting the data from UV titrations and ^{1}H NMR spectroscopy.[92] Henceforth $pK_a^{F,exp}$ and $pK_a^{C,exp}$ will differentiate experimental pK_a values from their respective computed values, $pK_a^{F,calc}$ and $pK_a^{C,calc}$. The experimental data showed large shifts in pK_a ranging between 2.5 to 4 pK units upon complex formation with CB[7] (Table 5-1). Additionally, association constants of the complexes were obtained at basic pH where guests were presumably deprotonated in both bound and unbound states; association constants were also obtained for the binding of protonated guests through application of the thermodynamic cycle (see Scheme 2 in ref. 59). In all cases, measurements of the binding free energies for different CB[7]:guest complexes indicated that the protonated guests are favored in the CB[7] cavity (discussed later; see Table 5-3).

pK_a Shifts upon CB[7]:guest complex formation

To compute the pK_a values of various BZ derivatives in complex with CB[7], we perform constant pH molecular dynamics (CpHMD) simulations on five CB[7]:guest complexes. In Figure 4-4, representative titration curves are shown for benzimidazole (BZ) and albendazole (ABZ), both in complex with CB[7] and free in solution. Similar curves corresponding to the other guests can be found in Figure 4-S1. Titration of the free guest in solution offers preliminary examination of the ability of the CpHMD to reproduce the experimental pK_a; for example, in the case of BZ, the $pK_a^{F,calc}$ value matches $pK_a^{F,exp}$ value of 5.5, indicating proper calibration of the CpHMD method. From the titration curve of BZ free in solution (Figure 4-4A, green curve), it is apparent that free BZ is protonated at values of pH less than 4.5 and deprotonated at pH levels above 6.5. Between these pH levels, an ensemble of protonated and deprotonated states exists. Relative to the titration curve for free BZ, the titration curve for the CB[7]:BZ complex is shifted toward more basic values of pH (Figure 4-4A, purple curve). Indeed, the value of $pK_a^{C,calc}$ for BZ is found to be 8.7 – a shift of more than 3 pK units above its pK_a^F

(Table 4-2); consequently, complexed BZ is protonated at pH below 7.5, indicating the preferred protonation state of BZ at neutral (typical physiological) pH differs depending on its bound state. The observed preference for the protonated guest in the cavity of CB[7] is due to the additional hydrogen bond between the titratable proton on BZ and one of the carbonyl oxygens at the entrance to the CB[7] cavity (Figure 4-5). It is worth noting that the Hill equation provides a reasonable estimate of the pK_a values for BZ both free and in complex with CB[7], with fitting errors of ~ 0.01 pK units. Furthermore, the $pK_a^{C,calc}$ value for BZ underestimates its $pK_a^{C,exp}$ by only 0.3 pK units (Table 4-2).

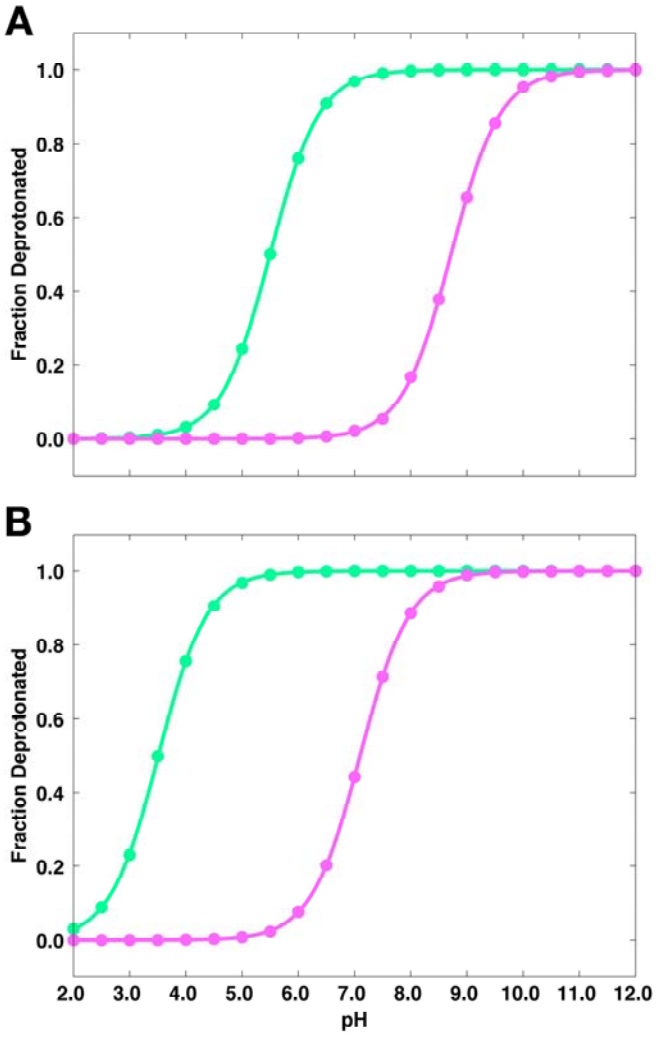


Figure 4-4. Titration curves from constant pH MD simulations of the guests free in solution (green) and in complex with CB[7] (purple). (A) Benzimidazole. (B) Albendazole.

Table 4-2. Comparison of pK_a^C values obtained from CpHMD simulations ($pK_a^{C,calc}$) with experimental data ($pK_a^{C,exp}$).[59] The corresponding Hill coefficients (n) are also shown.

Guest	$pK_a^{C, exp}$	$pK_a^{C,calc}$	n
BZ	9.0	8.71 ± 0.01	0.99
TBZ	8.6	8.19 ± 0.01	1.01
FBZ	8.6	8.61 ± 0.01	1.01
ABZ	6.1	7.10 ± 0.01	0.99
CBZ	7.0	7.40 ± 0.01	1.03

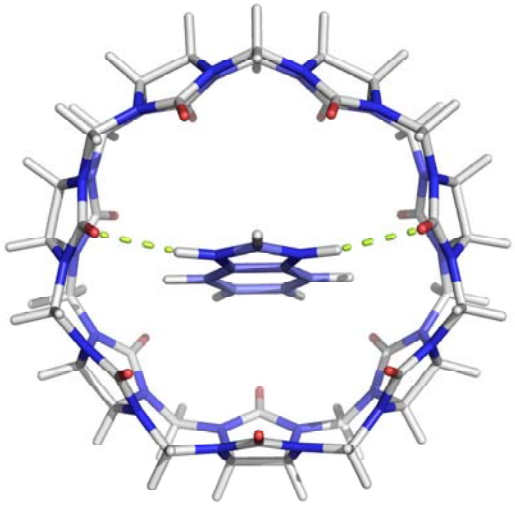


Figure 4-5. Hydrogen bonds formed between the protonated benzimidazole with the carbonyl oxygens of CB[7].

The chemical structure of ABZ differs from that of BZ by the presence amido and thioether R-groups attached to the BZ core. Additionally, the experimentally determined pK_a for the CB[7]:ABZ complex remains acidic ($pK_a^{C,exp}$=6.1). The titration curves obtained from CpHMD simulations of free and complexed ABZ are shown in Figure 4-4B. Qualitatively, the titration behavior of ABZ appears similar to that of BZ, as its $pK_a^{C,calc}$ of 7.1 is shifted toward a more basic value from its $pK_a^{F,exp}$ of 3.5. At neutral pH, these data suggest that ABZ is fully deprotonated when free in solution, whereas both its protonated and deprotonated forms are significantly populated when in complex with CB[7]. While the errors obtained for fitting the Hill equation to the titration data are minimal with errors observed for BZ

and all other guests of less than 0.01 pK units, the value of $pK_a^{C,calc}$ for the CB[7]:ABZ complex has the greatest deviation from the experiment ($\Delta pK=1.0$, Table 4-2).

The titration curves for the other guests follow a similar trend, where formation of the CB[7]:guest complex increases the pK_a of the guest (Figure 4-S1). These shifts are in line with the experimentally determined pK_a values (Table 4-2), with the largest deviation seen for ABZ as stated above. Overall, the CpHMD method provides accurate predictions of $pK_a^{C,calc}$ values, with a mean average error (MAE) of 0.42 pK units with respect to experiment ($pK_a^{C,exp}$, Table 4-2).

pH Dependence of the binding free energy

As discussed above, the pK_a values of the BZ-derived guests differ when bound to CB[7] and when free in solution (Figure 4-4, Table 4-2). Since there are no other titratable groups in the CB[7]:guest complexes in the pH ranges studied here, the binding of the guests to CB[7] can have a net uptake of protons, which makes their binding free energies depend on the solution pH. In this section, we compute binding free energies as functions of pH using Eqs. 6 and 8. Both of these equations can be used to obtain the pH-dependent binding free energy by adding a pH-dependent correction term to a reference binding free energy. In Eq. 6, this reference free energy corresponds to the free energy of binding in the absence of proton binding. In contrast, Eq. 8 requires that the reference free energy be obtained at a specific pH. The reference binding free energies in these two equations can be identical if obtained at a specific value of pH where the protonation states do not change. Since experimental association constants were obtained at pH levels where both the free and bound guests are deprotonated[59], we use reference binding free energies for the association of deprotonated guests with CB[7] in this work.

As a simple illustration of how pH-dependent binding free energies may be obtained, we first use the binding free energy measured experimentally for each of the different CB[7]:guest systems ($\Delta G°_{ref,exp}$) as the reference free energy term in Eqs. 6 and 8. We refer to this as a "hybrid" approach, as it obtains a pH-dependent binding free energy ($\Delta G°_{hybrid}$) from the experimental reference binding free energy ($\Delta G°_{ref,exp}$) and CpHMD-derived terms. These terms are either the pK_a for the complex, $pK_a^{C,calc}$,

used in Eq. 6 or charge differences between the binding partners in complex and free in solution, ΔZ in Eq. 8. While all results described here have been obtained using Eq. 6 with pH-dependent corrections requiring values of $pK_a^{C,calc}$, identical results have also been obtained using Eq. 8 (Figure 4-S3 and Table 4-S1).

Plots of binding free energies as functions of pH for CB[7] complexes with BZ, FBZ and ABZ are shown in Figure 4-6A-C. While these binding free energies are referenced to $\Delta G°_{ref,exp}$ (Figure 4-6, red line), the use of $pK_a^{C,calc}$ in Eq. 6 and ΔZ in Eq. 8 to generate the full curve as a function of pH can be assessed by how well the computed binding free energy at acidic pH ($\Delta G°^{+}_{hybrid}$ in Table 4-3) matches the analogous value derived from experiment ($\Delta G°^{+}_{exp}$, blue line in Figure 4-6). For all CB[7]:guest complexes, the values of $\Delta G°^{+}_{hybrid}$ deviate less than 1.35 kcal/mol from the respective experimental values (Table 4-3), with the greatest error observed for ABZ. These errors are entirely due to the errors in computing values of $pK_a^{C,calc}$, as the value for $\Delta G°^{+}_{exp}$ was derived using experimentally obtained values of pK_a^{C}, pK_a^{F}, and $\Delta G°_{ref}$.[59]

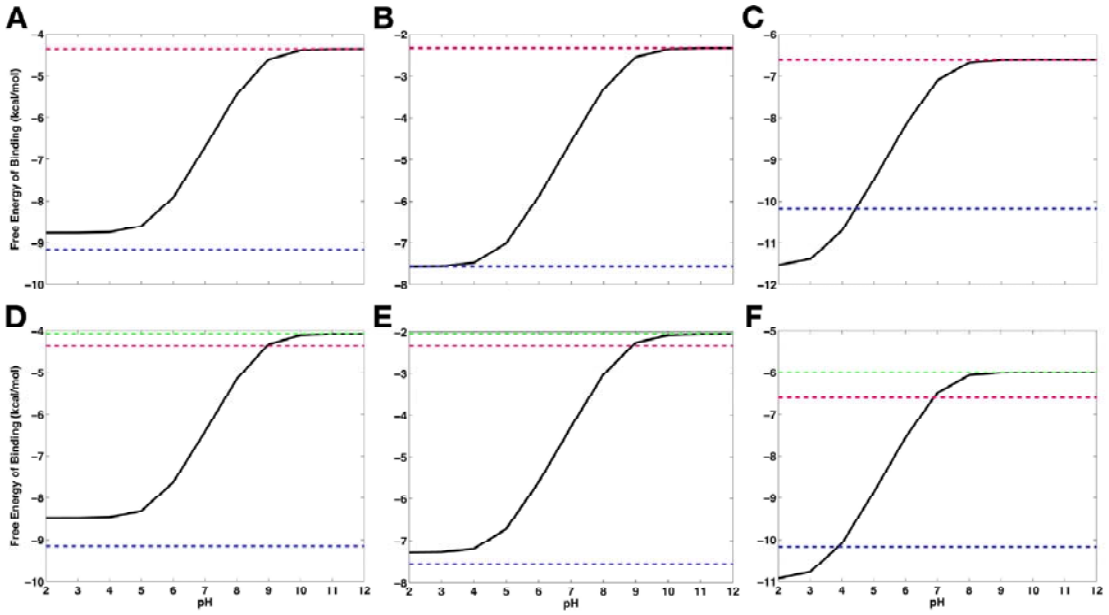


Figure 4-6. Binding free energies as functions of pH (black line). The top row is computed by the hybrid approach using the experimental reference binding energies ($\Delta G°_{ref,exp}$, red line) and the bottom row uses the full computational approach with the reference binding energies computed by thermodynamic integration ($\Delta G°_{ref,TI}$, green line). Experimentally derived binding free energies for the protonated guests are shown in blue. (A), (D) Benzimidazole. (B), (E) Fuberidazole. (C), (F) Albendazole.

Table 4-3. Binding free energies of the guests upon complex formation with CB[7], computed using the hybrid approach with Eq. 6. All energies are reported in kcal/mol. $\Delta G°_{ref,exp}$ is the experimental[59] binding free energy for the reference deprotonated guest; $\Delta G°^{+}_{exp}$ is the binding free energy for the protonated guest derived from the $\Delta G°_{ref,exp}$; and $\Delta G°^{+}_{hybrid}$ is the free energy obtained by using $pK_a^{C,calc}$ with $\Delta G°_{ref,exp}$ in Eq. 6.

Guest	**$\Delta G°_{ref,exp}$**	**$\Delta G°^{+}_{exp}$**	**$\Delta G°^{+}_{hybrid}$**	**$\Delta G°_{ref,exp}$ (pH 7)**	**$\Delta G°_{hybrid}$ (pH 7)**
BZ	-4.4	-9.2	-8.8	-7.1	-6.7
TBZ	-3.0	-8.6	-7.9	-5.2	-4.7
FBZ	-2.3	-7.6	-7.6	-4.5	-4.6
ABZ	-6.6	-10.2	-11.5	-6.7	-7.1
CBZ	-6.0	-9.5	-10.0	-6.4	-6.8

From Figure 4-6A-C, it is evident that all guests bind more favorably when protonated. Indeed, the binding free energies observed for deprotonated guests (at extremely basic pH) are 3.4 to 5.6 kcal/mol more positive (less favorable) than those obtained at acidic pH when the guests are protonated. This tendency is most pronounced in CB[7]:FBZ complex (Figure 4-6B), for which the binding free energy obtained when FBZ is predominantly protonated (-7.56 kcal/mol) is over 5 kcal/mol more favorable than its respective value when FBZ is deprotonated (-2.33 kcal/mol). This observation is consistent with experiment[59] and stems from the favorable hydrogen bond between the guest and CB[7] (Figure 4-5).

Taking a closer look at the pH-dependent binding free energies of the CB[7]:BZ complex, it is apparent that the binding free energy spans 4.8 kcal/mol between pH levels 4.5 to 10, a range that essentially encompasses the pH levels of most biological reactions (Figure 4-6A).[93] At physiological pH (~ 7), free BZ is predominantly deprotonated, whereas BZ in complex with CB[7] is protonated (Figure 4-4A). In conventional free energy computations, ligand protonation states are typically assigned as the preferred protonation state for the free ligand. Consistent with this convention, BZ would be considered deprotonated in free energy computations performed at pH 7. In making this assumption, the binding free energy deviates from the pH-dependent binding free energy obtained here by ~ 2.3 kcal/mol for the CB[7]:BZ complex. Similar deviations are noted for the binding of other guests as well, with the magnitudes ranging between 0.4 to 2.3 kcal/mol (Table 4-3, Figure 4-S2).

Full prediction of the pH-dependent free energy profile

To demonstrate the utility of our method when experimental binding free energies are unavailable, we perform thermodynamic integration (TI) computations based on the thermodynamic cycle shown in Scheme 4-2 to obtain $\Delta G°_{ref,TI}$, the reference binding free energies of the CB[7]:guest complexes with the guests deprotonated. The pH-dependent correction terms obtained either with Eq. 6 or 8 are then referenced to $\Delta G°_{ref,TI}$ to obtain a full computational prediction (CpHMD/TI) of the pH-dependent free energy profiles.

The free energy profiles of CB[7]:guest complexes using $\Delta G°_{ref,TI}$ are shown in Figure 4-6D-F, while the computed values of $\Delta G°_{ref,TI}$ are reported in Table 4-4 for comparison with experiment ($\Delta G°_{ref,exp}$, Table 4-4). All values of $\Delta G°_{ref,TI}$ agree well with experiment, showing absolute errors that are less than 1.3 kcal/mol. Further, the error with respect to experiment ($\Delta G°^{+}_{exp}$) in the predicted values for the binding free energy of protonated guests ($\Delta G°^{+}_{TI}$) are similarly low (< 1.4 kcal/mol). Errors in $\Delta G°^{+}_{TI}$ arise from both the computation of $pK_a^{C,calc}$ (or ΔZ, when using Eq. 8; Table 4-S2) using the CpHMD method and the binding free energy computation with TI. These errors are not always of the same sign; for example, the deviation from $\Delta G°^{+}_{exp}$ obtained for ABZ decreased from 1.4 kcal/mol when using $\Delta G°_{ref,exp}$ in the hybrid approach to 0.7 kcal/mol when using the $\Delta G°_{ref,TI}$ reference, indicating some cancellation of error in the full computational approach. In contrast, the deviation for TBZ increased from 0.7 kcal/mol when using the experimental $\Delta G°_{ref,exp}$ reference to 1.4 kcal/mol when using $\Delta G°_{ref,TI}$. Regardless, results obtained using the full computational approach with $\Delta G°_{ref,TI}$ show errors that are similar in magnitude to those observed using $\Delta G°_{ref,exp}$ in the hybrid experimental/computational approach. Furthermore, the errors associated with the full computational protocol can be lower than the errors that arise from performing binding free energy computations with fixed protonation states assigned to the unbound CB[7] and guest molecules (Table 4-4).

Table 4-4. Binding free energies of the guests, computed using full computational approach (CpHMD/TI) and compared to experiment.[59] All energies are reported in kcal/mol. $\Delta G°_{ref,exp}$ is the experimental[59] binding free energy for the reference deprotonated guest; $\Delta G°_{ref,TI}$ is the absolute binding free energy obtained from TI computations for the reference state; $\Delta G°^{+}_{exp}$ is the binding free energy for the protonated guest derived from the $\Delta G°_{ref,exp}$; and $\Delta G°^{+}_{TI}$ is the binding free energy obtained by using $pK_a^{C,calc}$ with $\Delta G°_{ref,TI}$ in Eq. 6.

Guest	**$\Delta G°_{ref,exp}$**	**$\Delta G°_{ref,TI}$**	**$\Delta G°^{+}_{exp}$**	**$\Delta G°^{+}_{TI}$**	**$\Delta G°_{ref,exp}$ (pH 7)**	**$\Delta G°_{ref,TI}$ (pH 7)**
BZ	-4.4	-4.1 ± 2.0	-9.2	-8.5	-7.1	-6.4
TBZ	-3.0	-2.3 ± 2.6	-8.6	-7.3	-5.2	-4.0
FBZ	-2.3	-2.1 ± 2.6	-7.6	-7.3	-4.5	-4.3
ABZ	-6.6	-6.0 ± 3.0	-10.2	-11.0	-6.7	-6.5
CBZ	-6.0	-4.8 ± 2.7	-9.5	-8.7	-6.4	-5.5

Discussion

Changes in the pK_a values and, consequently, the protonation states of ionizable species participating in biomolecular association processes are well documented. To address this phenomenon, we present a simple methodology for obtaining the pH dependence of binding free energies for a series of cucurbit[7]uril (CB[7]):guest complexes. Based on Wyman's binding polynomial formalism,[43] binding free energies are computed as pH-dependent corrections to a reference binding free energy. Combining this formalism with constant pH molecular dynamics (CpHMD) simulations and free energy computations yields a reasonable protocol for evaluating the pH-dependent binding free energies of biomolecular systems.

Focusing on the application of CpHMD to Eq. 6 in order to obtain pH-dependent relative binding free energies, we assess how well CpHMD simulations can capture the pK_a of BZ guests in complex with CB[7] (Table 4-2). With the exception of albendazole (ABZ), the values of $pK_a^{C,calc}$ obtained for the different CB[7]:guest complexes deviate from experiment by less than 0.41 pK units. The $pK_a^{C,calc}$ value obtained for the CB[7]:ABZ complex, however, exhibits an error of 1.0 pK unit. Since the CpHMD simulations conducted in this study are only 5 ns long, we extended the simulation of CB[7]:ABZ to 25 ns to ascertain whether the value of $pK_a^{C,calc}$ had converged; however, the resulting $pK_a^{C,calc}$ remains unchanged. Further, the process of fitting titration data obtained from CpHMD simulations of CB[7]:guest complexes to the Hill equation is achieved with very little statistical error (< 0.01 pK units, Table 4-2). Both of these findings indicate that the error in the $pK_a^{C,calc}$ values is not due to convergence problems in the CpHMD simulations. Instead, it is possible, though not explicitly demonstrated in this work, that inaccuracies in the computed pK_a^{C} values stem from problems with the force field due to the accuracies of similar magnitude to those seen in the previous CpHMD runs.[94, 95]

Since CpHMD simulations can reliably compute the values of $pK_a^{C,calc}$ (and, similarly, ΔZ in the case of Eq. 8) for the CB[7]:guest systems, we proceed to incorporate these $pK_a^{C,calc}$ values in Eq. 6 along with a reference experimental binding free energy ($\Delta G°_{ref,exp}$) to obtain binding free energies as functions of pH. This hybrid experimental/computational approach is followed and shown for

CB[7]:guest systems (Figure 4-6A-C). To evaluate the accuracy of this approach, we compare the computed value of $\Delta G^{\circ +}_{hybrid}$ to experiment ($\Delta G^{\circ +}_{exp}$) and observe good agreement, with errors of < 1.4 kcal/mol arising from the computation of $pK_a^{C,calc}$.

Having established that Eq. 6 can successfully recapitulate pH-dependent binding free energies with an experimentally determined reference binding free energy, we consider the use of thermodynamic integration (TI) computations to remove this dependence on experiment. TI computations effectively reproduce the reference binding free energies observed from experiment ($\Delta G^{\circ}_{ref,exp}$) with absolute errors less than 1.3 kcal/mol ($\Delta G^{\circ}_{ref,TI}$, Table 4-4). The resulting pH-dependent free energy profiles using $\Delta G^{\circ}_{ref,TI}$ are similar to those computed with $\Delta G^{\circ}_{ref,exp}$, as shown in Figure 4-6. Furthermore, the absolute errors in predicting the free energies of the protonated guests, $\Delta G^{\circ +}_{TI}$, using the $\Delta G^{\circ}_{ref,TI}$ reference are less than 1.3 kcal/mol. These errors arise from both $\Delta G^{\circ}_{ref,TI}$ and the use of CpHMD simulations to obtain $pK_a^{C,calc}$ values. In regard to the computation of $\Delta G^{\circ}_{ref,TI}$, we do observe large statistical uncertainties for all CB[7]:guest complexes considered, which stem largely from the van der Waals decoupling simulations (Table 4-S3); however, the free energies computed for every transformation in the thermodynamic cycle shown in Scheme 4-2 have all converged, with the cumulative computed $\Delta G < 0.01$ kcal/mol.

The use of our CpHMD/TI approach to provide a full computational prediction of pH-dependent binding free energies is particularly advantageous when experimental association constants are not available, as most experimental measurements face the limitations at extreme pH levels due to the highly possible destabilization or denaturation of the proteins under such severe conditions. Therefore, when combined with computational free energy calculations, our method is free from such concerns, eliminating the reliance on the availability of experimental data. While the CpHMD/TI computation of pH-dependent binding free energies is prone to greater error than the hybrid experimental/computational approach described previously, we find the absolute errors in the CpHMD-derived pK_a^C values and reference binding free energies obtained from TI computations are not necessarily of the same sign for the CB[7]:guest systems considered and may cancel out. Further, the

observed errors with respect to the $\Delta G^{\circ +}_{exp}$ are relatively low (< 1.4 kcal/mol). In contrast, the error in assigning incorrect protonation states in free energy computations without correcting for the pH dependence of the binding free energy can give errors in excess of 2 kcal/mol (Table 4-3). This observation underscores the importance of accounting for the linkage of proton binding or release to ligand binding in free energy computations and demonstrates the high utility of the CpHMD/TI approach.

Our results highlight the significant changes in pK_a and free energy of binding upon complex formation that accompanies a net proton uptake. Noting that the guests used here have a single titratable site, corresponding changes in free energy may sometimes be larger in protein-ligand binding where multiple titratable groups exist. Therefore, we believe that our method will have great utility in computer-aided drug discovery, where early stages of the structure-based drug design often focus on finding a high-affinity binder to a target protein. Extensions of our methodology to such more complex protein systems may require improvements to the computational protocols employed. The simple framework developed here allows for trivial incorporation of CpHMD methods that incorporate explicit solvent models[35] and/or enhanced sampling techniques, such as accelerated molecular dynamics[37] or replica exchange[36] to improve pK_a computations in systems where convergence is difficult.[95] Similarly, our protocol accommodates the use of alternative methods for obtaining the reference binding free energy required by Eqs. 6 and 8. Thus the computational methodology for performing CpHMD/TI computations can be chosen to best address the system under consideration.

While we have focused on the results obtained using Eq. 6, which assumes all titratable groups are decoupled, the CpHMD/TI method is also compatible with the expression for obtaining the pH-dependent binding free energy given in Eq. 8, and these two expressions yield identical results in the case of the CB[7]:guest systems considered here. We intend to build on the computational protocol developed here, with a natural extension being the application of the CpHMD/TI method to obtain pH-dependent binding free energies of protein-ligand complexes. As protein-ligand systems are more complicated than the CB[7]:guest systems considered in this work, we believe the use of Eq. 8 will have high utility to address potential interactions between titratable groups.

Given the magnitude of errors in computed binding free energies obtained with fixed protonation states in the CB[7]:guest systems, our computational protocol represents a promising approach to remove these errors, thus implicating its utility in drug discovery workflows.[96] Though not specifically addressed in this work, similar philosophies may also be applicable to the scoring functions in docking protocols.

Conclusions

In this work, we determined the pH-dependent changes in binding free energies for complex formation between cucurbit[7]uril (CB[7]) and a series of benzimidazole guests. Using constant pH molecular dynamics simulations combined with experimental data, we developed a hybrid protocol that could capture the significant changes in the CB[7]:guest binding free energies with high accuracy. Subsequently, we combined our method with thermodynamic integration (TI) to enable a full computational prediction of the pH-dependent free energy profiles. This protocol successfully accounted for the pH-dependent changes in the binding free energies during complex formation. Future work will include examination of pH-dependent binding free energies for protein-ligand complexes.

Acknowledgements

This work was supported by the NSF (MCB-1020765), NIH (NIH GM31749), Howard Hughes Medical Institute, Center for Theoretical Biological Physics (CTBP), National Biomedical Computation Resource (NBCR) and NSF supercomputer centers. P. G. B. and J. W. K. acknowledge support from the NIH Molecular Biophysical Training Grant (2T32GM008326-21).

Chapter 4 is a reprint of the material as it appears in Meekyum Olivia Kim, Patrick G. Blachly, Joseph W. Kaus, and J. Andrew McCammon, "Protocols Utilizing Constant pH Molecular Dynamics to Compute pH-Dependent Binding Free Energies," *Journal of Physical Chemistry B*, 2014. The dissertation author was one of two primary investigators, each of whom contributed equally to performing research and writing this paper.

Supporting Information

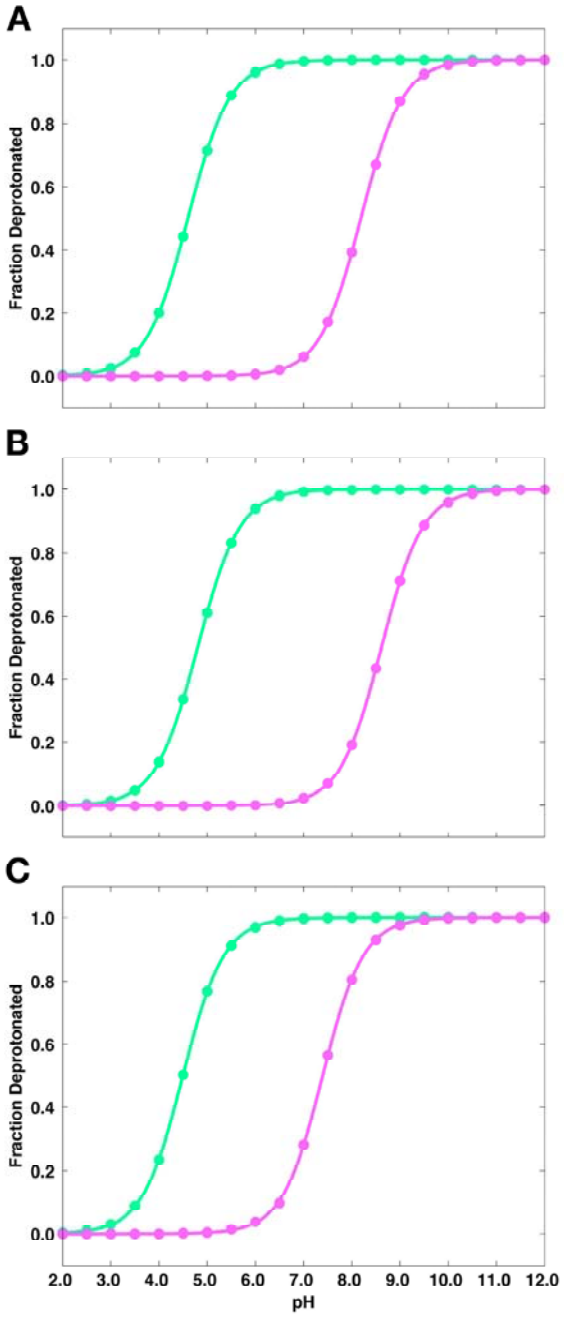


Figure 4-S1. Titration curves from constant pH MD simulations of the guests free in solution (green) and in complex with CB[7] (purple). (A) Thiabendazole. (B) Fuberidazole. (C) Carbendazim.

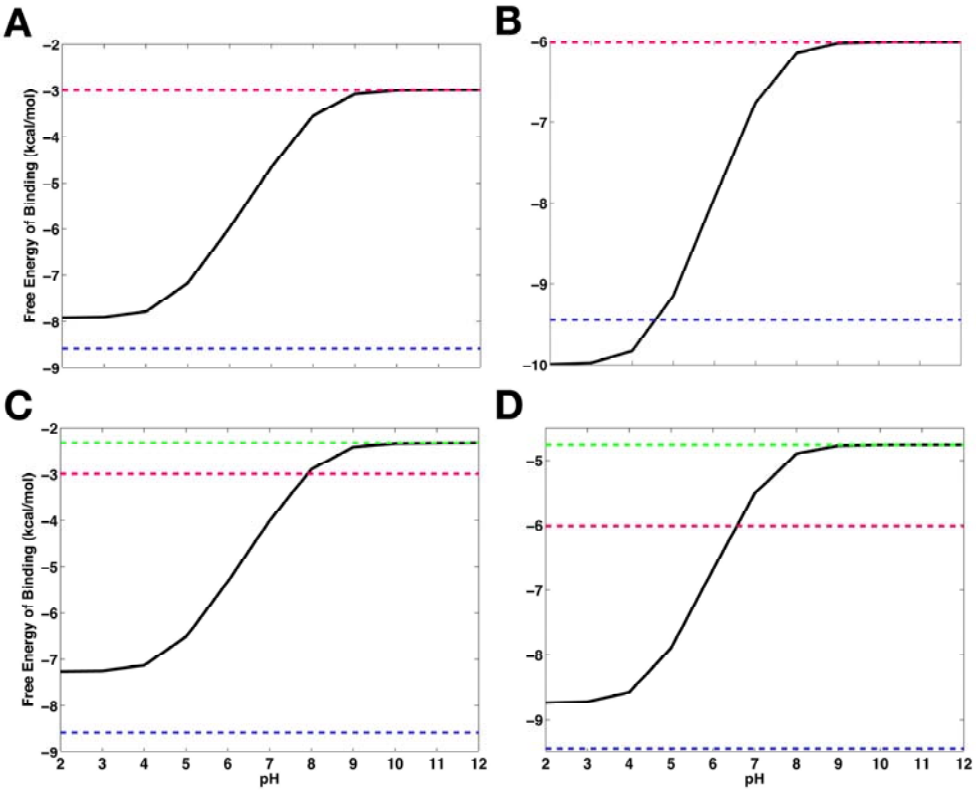


Figure 4-S2. Binding free energies as functions of pH (black line). The top row is computed by the hybrid approach using the experimental reference binding energies ($\Delta G°_{ref,exp}$, red line) and the bottom row uses the full computational approach with the reference binding energies computed by thermodynamic integration ($\Delta G°_{ref,TI}$, green line). Experimentally derived binding free energies for the protonated guests are shown in blue. (A), (C) Thiabendazole. (B), (D) Carbendazim.

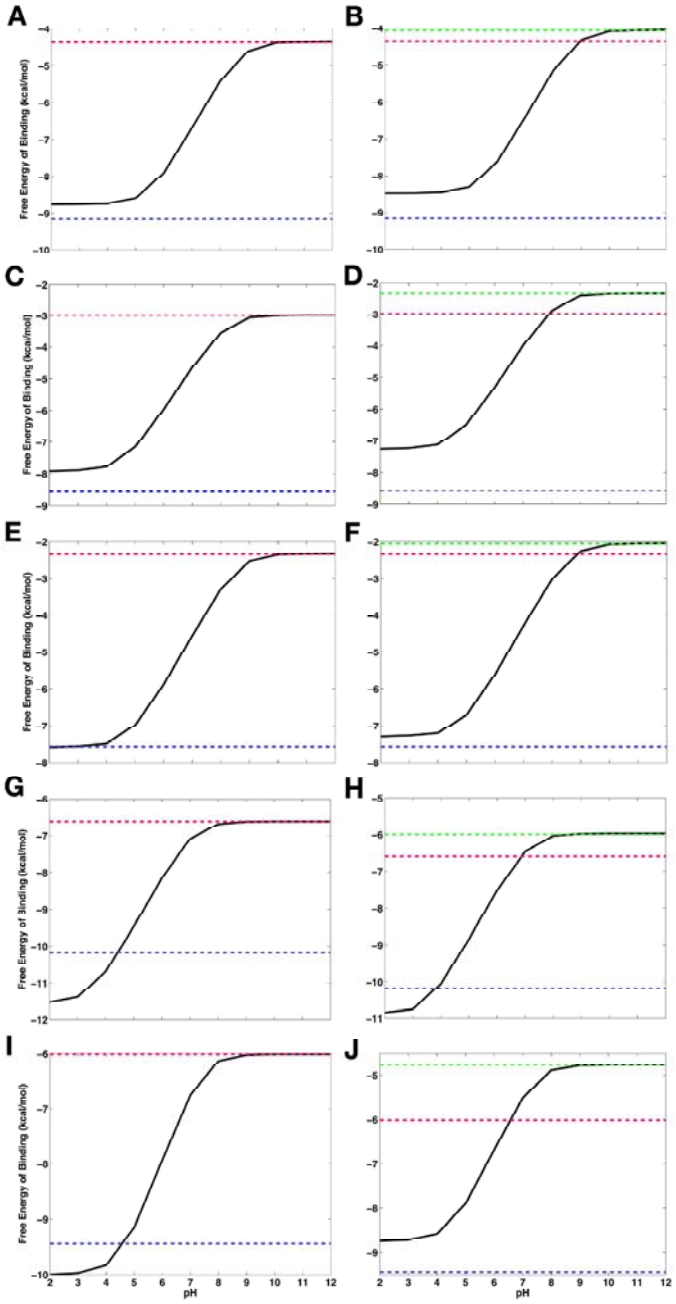


Figure 4-S3. Binding free energies as functions of pH (black line), computing using Eq. 8. The left row is computed by the hybrid approach using the experimental reference binding energies ($\Delta G^{\circ}_{ref,exp}$, red line) and the right row uses the full computational approach with the reference binding energies computed by thermodynamic integration ($\Delta G^{\circ}_{ref,TI}$, green line). Experimentally derived binding free energies for the protonated guests are shown in blue. (A), (B) Benzimidazole. (C), (D) Thiabendazole. (E), (F) Fuberidazole. (G), (H) Albendazole. (I), (J) Carbendazim.

Table 4-S1. Binding free energies of the guests upon complex formation with CB[7], computed using the hybrid approach with Eq. 8. All energies are reported in kcal/mol. $\Delta G°_{ref,exp}$ is the experimental[59] binding free energy for the reference deprotonated guest; $\Delta G°^{+}_{exp}$ is the binding free energy for the protonated guest derived from the $\Delta G°_{ref,exp}$; and $\Delta G°^{+}_{hybrid}$ is the free energy obtained by using $pK_a^{C,calc}$ with $\Delta G°_{ref,exp}$ in Eq. 8.

Guest	**$\Delta G°_{ref,exp}$**	**$\Delta G°^{+}_{exp}$**	**$\Delta G°^{+}_{hybrid}$**	**$\Delta G°_{ref,exp}$ (pH 7)**	**$\Delta G°_{hybrid}$ (pH 7)**
BZ	-4.4	-9.2	-8.8	-7.1	-6.7
TBZ	-3.0	-8.6	-7.9	-5.2	-4.6
FBZ	-2.3	-7.6	-7.6	-4.5	-4.6
ABZ	-6.6	-10.2	-11.4	-6.7	-7.1
CBZ	-6.0	-9.5	-10.0	-6.4	-6.8

Table 4-S2. Binding free energies of the guests, computed using full computational approach (CpHMD/TI) and compared to experiment.[59] All energies are reported in kcal/mol. $\Delta G°_{ref,exp}$ is the experimental[59] binding free energy for the reference deprotonated guest; $\Delta G°_{ref,TI}$ is the absolute binding free energy obtained from TI computations for the reference state; $\Delta G°^{+}_{exp}$ is the binding free energy for the protonated guest derived from the $\Delta G°_{ref,exp}$; and $\Delta G°^{+}_{TI}$ is the binding free energy obtained by using $pK_a^{C,calc}$ with $\Delta G°_{ref,TI}$ in Eq. 8.

Guest	**$\Delta G°_{ref,exp}$**	**$\Delta G°_{ref,TI}$**	**$\Delta G°^{+}_{exp}$**	**$\Delta G°^{+}_{TI}$**	**$\Delta G°_{ref,exp}$ (pH 7)**	**$\Delta G°_{ref,TI}$ (pH 7)**
BZ	-4.4	-4.1 ± 2.0	-9.2	-8.5	-7.1	-6.4
TBZ	-3.0	-2.3 ± 2.6	-8.6	-7.3	-5.2	-4.0
FBZ	-2.3	-2.1 ± 2.6	-7.6	-7.3	-4.5	-4.3
ABZ	-6.6	-6.0 ± 3.0	-10.2	-10.8	-6.7	-6.5
CBZ	-6.0	-4.8 ± 2.7	-9.5	-8.7	-6.4	-5.5

Table 4-S3. Free energy for each segment in the thermodynamic cycle (Scheme 4-2) for absolute binding free energy computations for the guests. All energies are reported in kcal/mol. ΔG_1 is the free energy for gradually turning on restraints on; ΔG_2 is the free energy for decoupling the guest while bound to the host in the presence of the restraints; $\Delta G_3°$ is the free energy for turning off the restraint and correcting for the standard state; and ΔG_4 is the solvation free energy for the decoupled guest. Statistical errors are reported as standard deviations. Following Scheme 4-2, $\Delta G°_{ref,TI}$ is computed as - ΔG_1 - ΔG_2 - $\Delta G_3°$ + ΔG_4.

System	Segment	ΔG_{elec}	ΔG_{vdw}	ΔG_1	$\Delta G_3°$	$\Delta G°_{ref,TI}$	$\Delta G°_{ref,exp}$
CB[7]:BZN	ΔG_2	6.14 ± 0.60	16.22 ± 1.10	6.28 ± 0.74	-14.28	-4.07 ± 2.05	-4.36
	ΔG_4	10.13 ± 0.85	0.16 ± 1.16				
CB[7]:TBN	ΔG_2	-36.45 ± 0.82	16.10 ± 1.51	4.48 ± 0.62	-14.36	-2.33 ± 2.58	-2.99
	ΔG_4	-33.68 ± 0.97	1.12 ± 1.55				
CB[7]:FBN	ΔG_2	-10.64 ± 0.76	15.40 ± 1.50	4.34 ± 0.59	-14.36	-2.05 ± 2.52	-2.33
	ΔG_4	-7.88 ± 0.92	0.59 ± 1.54				
CB[7]:ABN	ΔG_2	249.13 ± 0.94	17.65 ± 1.84	3.61 ± 0.57	-14.85	-6.00 ± 3.01	-6.61
	ΔG_4	248.75 ± 1.05	0.79 ± 1.83				
CB[7]:CBN	ΔG_2	257.78 ± 0.89	16.63 ± 1.54	4.42 ± 0.62	-14.69	-4.76 ± 2.69	-6.01
	ΔG_4	258.17 ± 1.11	1.21 ± 1.56				

References

1. Onufriev, A. V.; Alexov, E. Protonation and pK Changes in Protein-Ligand Binding. *Q. Rev. Biophys.* **2013,** *46*, 181-209.

2. Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and Rationalization of Protein pKa Values Using QM and QM/MM Methods. *J. Phys. Chem. A* **2005,** *109*, 6634-6643.

3. Jensen, J. H. Calculating pH and Salt Dependence of Protein-Protein Binding. *Curr. Pharm. Biotechnol.* **2008,** *9*, 96-102.

4. Mason, A. C.; Jensen, J. H. Protein-Protein Binding Is Often Associated with Changes in Protonation State. *Proteins* **2008,** *71*, 81-91.

5. Cassidy, C. S.; Lin, J.; Frey, P. A. A New Concept for the Mechanism of Action of Chymotrypsin: The Role of the Low-Barrier Hydrogen Bond. *Biochemistry* **1997,** *36*, 4576-4584.

6. Stranzl, G. R.; Gruber, K.; Steinkellner, G.; Zangger, K.; Schwab, H.; Kratky, C. Observation of a Short, Strong Hydrogen Bond in the Active Site of Hydroxynitrile Lyase from Hevea Brasiliensis Explains a Large pKa Shift of the Catalytic Base Induced by the Reaction Intermediate. *J. Biol. Chem.* **2004,** *279*, 2699-3707.

7. Brandsdal, B. O.; Smalås, A. O.; Åqvist, J. Free Energy Calculations Show That Acidic P1 Variants Undergo Large pKa Shifts Upon Binding to Trypsin. *Proteins* **2006,** *64*, 740-748.

8. Trylska, J.; Antosiewicz, J.; Geller, M.; Hodge, C. N.; Klabe, R. M.; Head, M. S.; Gilson, M. K. Thermodynamic Linkage between the Binding of Protons and Inhibitors to HIV-1 Protease. *Protein. Sci.* **1999,** *8*, 180-195.

9. Aguilar, B.; Anandakrishnan, R.; Ruscio, J. Z.; Onufriev, A. V. Statistics and Physical Origins of pK and Ionization State Changes Upon Protein-Ligand Binding. *Biophys. J.* **2010,** *98*, 872-880.

10. Manallack, D. T. The Acid-Base Profile of a Contemporary Set of Drugs: Implications for Drug Discovery. *SAR QSAR Environ. Res.* **2009,** *20*, 611-655.

11. Leeson, P. D.; St-Gallay, S. A.; Wenlock, M. C. Impact of Ion Class and Time on Oral Drug Molecular Properties. *Med. Chem. Commun.* **2011,** *2*, 91-105.

12. Meanwell, N. A. Improving Drug Candidates by Design: A Focus on Physicochemical Properties as a Means of Improving Compound Disposition and Safety. *Chem. Res. Toxicol.* **2011,** *24*, 1420-1456.

13. Manallack, D. T.; Prankerd, R. J.; Yuriev, E.; Oprea, T. I.; Chalmers, D. K. The Significance of Acid/Base Properties in Drug Discovery. *Chem. Soc. Rev.* **2013,** *42*, 485-496.

14. Gordon, J. C.; Myers, J. B.; Folta, T.; Soja, V.; Heath, L. S.; Onufriev, A. V. H++: A Server for Estimating pKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005,** *33*, W368-371.

15. Myers, J. B.; Grothaus, G.; Narayana, S.; Onufriev, A. V. A Simple Clustering Algorithm Can Be Accurate Enough for Use in Calculations of Pks in Macromolecules. *Proteins* **2006,** *63*, 928-938.

16. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulation. *Nucleic Acids Res.* **2012,** *40 (W1)*, W537-541.

17. Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Interpretation of Protein pKa Values. *Proteins* **2005,** *61*, 704-721.

18. Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein-Ligand Complexes. *Proteins* **2008,** *73*, 765-783.

19. Olsson, M. H. M.; Søndergard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011,** *7*, 525-537.

20. Søndergard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Realization of pKa Values. *J. Chem. Theory Comput.* **2011,** *7*, 2284-2295.

21. Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pKa Prediction for Organic Acids and Bases*. Chapman and Hall: London, 1981.

22. Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pKa Prediction and Protonation State Generateion for Drug-Like Molecules. *J. Comput. Aided Mol. Des.* **2007,** *21*, 681-691.

23. Alexov, E.; Gunner, M. R. Incorporating Protein Conformational Flexibility into pH Titration Calculations: Results on T4 Lysozyme. *Biophys. J.* **1997,** *74*, 2075-2093.

24. Georgescu, R. E.; Alexov, E.; Gunner, M. R. Combining Conformational Flexibility and Continuum Electrostatics for Calculating pKa's in Proteins. *Biophys. J.* **2002,** *83*, 1731-1748.

25. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Seminanalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990,** *112*, 6127-6129.

26. Warshel, A.; Russel, S. Calculations of Electrostatic Interactions in Biological Systems and in Solutions. *Q. Rev. Biophys.* **1984,** *17*, 283-422.

27. Sham, Y. Y.; Chu, Z. T.; Warshel, A. Consistent Calculations of pKa's of Ionizable Residues in Proteins: Semi-Microscopic and Microscopic Approaches. *J. Phys. Chem. B* **1997,** *101*, 4458-4472.

28. Li, H.; Robertson, A. D.; Jensen, J. H. The Determinants of Carboxyl pKa Values in Turkey Ovomucoid Third Domain. *Proteins* **2004,** *55*, 689-704.

29. Baptista, A. M.; Martel, P. J.; Petersen, S. B. Simulation of Protein Conformational Freedom as a Function of pH: Constant-pH Molecular Dynamics Using Implicit Titration. *Proteins* **1997,** *27*, 523-544.

30. Borjesson, U.; Hunenberger, P. H. Explicit-Solvent Molecular Dynamics Simulation at Constant pH: Methodology and Application to Small Amines. *J. Chem. Phys.* **2001,** *114*, 9706-9719.

31. Lee, M. S.; Salsbury Jr., F. R.; Brooks III, C. L. Constant-pH Molecular Dynamics Using Continuous Titration Coordinates. *Proteins* **2004,** *56*, 738-752.

32. Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH Molecular Dynamics Using Stochastic Titration. *J. Chem. Phys.* **2002,** *117*, 4184-4200.

33. Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004,** *25*, 2038-2048.

34. Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. Simulating Proteins at Constant pH: An Approach Combining Molecular Dynamics and Monte Carlo Simulation. *Proteins* **2002,** *47*, 469-480.

35. Swails, J.; York, D. M.; Roitberg, A. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput.* **2014,** *10*, 1341-1352.

36. Meng, Y. L.; Roitberg, A. Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model. *J. Chem. Theory Comput.* **2010,** *6*, 1401-1412.

37. Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. Coupling Constant pH Molecular Dynamics with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2010,** *6*, 560-568.

38. Goh, G. B.; Knight, J. L.; Brooks III, C. L. Constant pH Molecular Dynamics Simulations of Nucleic Acids in Explicit Solvent. *J. Chem. Theory Comput.* **2012,** *8*, 36-46.

39. Goh, G. B.; Knight, J. L.; Brooks, I., C. L. pH-Dependent Dynamics of Complex RNA Macromolecules. *J. Chem. Theory Comput.* **2013,** *9*, 935-943.

40. Goh, G. B.; Knight, J. L.; Brooks III, C. L. Towards Accurate Prediction of Protonation Equilibrium of Nucleic Acids. *J. Phys. Chem. Lett.* **2013,** *4*, 760-766.

41. Di Russo, N. V.; Estrin, D. A.; Marti, M. A.; Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Comput. Biol.* **2012,** *8*, e1002761.

42. Laricheva, E. N.; Arora, K.; Knight, J. L.; Brooks, I., C. L. Deconstructing Activation Events in Rhodopsin. *J. Am. Chem. Soc.* **2013,** *135*, 10906-10909.

43. Wyman, J. Heme Proteins. *Adv. Protein Chem.* **1948,** *4*, 407-531.

44. Mitra, R. C.; Zhang, Z.; Alexov, E. In Silico Modeling of pH-Optimum of Protein-Protein Binding. *Proteins* **2011,** *79*, 925-936.

45. Misra, V. K.; Hecht, J. L.; Yang, A. S.; Honig, B. Electrostatic Contributions to the Binding Free Energy of the Λcl Repressor to DNA. *Biophys. J.* **1998,** *75*, 2262-2273.

46. Moody, E. M.; Lecompte, J. T. J.; Belivacqua, P. C. Linkage between Proton Binding and Folding in RNA: A Thermodynamic Framework and Its Experimental Application for Investigating pKa Shifting. *RNA* **2005,** *11*, 157-172.

47. Freeman, W. A.; Mock, W. L.; Shih, N. Y. Cucurbituril. *J. Am. Chem. Soc.* **1981,** *103*, 7367-7368.

48. Kim, J.; Jung, I. S.; Kim, S. Y.; Lee, E.; Kang, J. K.; Sakamoto, S.; Yamaguchi, K.; Kim, K. New Cucurbituril Homologues: Syntheses, Isolation, Characterization, and X-Ray Crystal Structures of Cucurbit[N]Uril (N=5, 7, and 8). *J. Am. Chem. Soc.* **2000,** *122*, 540-541.

49. Liu, S.; Ruspic, C.; Mukhopadhyay, P.; Chakrabarti, S.; Zavalij, P. Y.; Isaacs, L. The Cucurbit[N]Uril Family: Prime Components for Self-Sorting Systems. *J. Am. Chem. Soc.* **2005,** *127*, 15959-15967.

50. Rekharsky, M. V.; Mori, T.; Yang, C.; Ko, Y. H.; Selvapalam, N.; Kim, H.; Sobransingh, D.; Kaifer, A. E.; Liu, S.; Isaacs, L., et al. A Synthetic Host-Guest System Achieves Avidin-Biotin Affinity by Overcoming Enthalpy-Entropy Compensation. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *109*, 20006-20011.

51. Macartney, D. H. Encapsulation of Drug Molecules by Cucurbiturils: Effects on Their Chemical Properties in Aqueous Solution. *Isr. J. Chem.* **2011,** *51*, 600-615.

52. Ghosh, I.; Nau, W. M. The Strategic Use of Supramolecular pKa Shifts to Enhance the Bioavailability of Drugs. *Adv. Drug Deliv. Rev.* **2012,** *64*, 764-783.

53. Ma, D.; Hettiarachchi, G.; Nguyen, D.; Zhang, B.; Wittenberg, J. B.; Zavalij, P. Y.; Briken, V.; Isaacs, L. Acyclic Cucurbit[N]Uril Molecular Containers Enhance the Solubility and Bioactivity of Poorly Soluble Pharmaceuticals. *Nat. Chem.* **2012,** *4*, 503-510.

54. Hettiarachchi, G.; Nguyen, D.; Wu, J.; Lucas, D.; Ma, D.; Isaacs, L.; Briken, V. Toxicology and Drug Delivery by Cucurbit[N]Uril Type Molecular Containers. *PLoS One* **2010,** *5*, e10514.

55. Saleh, N.; Meetani, M.; Al-Kaabi, L.; Ghosh, I.; Nau, W. M. Effects of Cucurbiturils on Tropicamide and Potential Applications in Ocular Drug Delivery. *Supramol. Chem.* **2011,** *23*, 654-661.

56. Horton, J. Albendazole: A Review of Anthelmintic Efficacy and Safety in Humans. *Parasitology* **2000,** *121*, S113-132.

57. Gould, M.; Nelson, L.; Waterere, D.; Hynes, R. Biocontrol of Fusarium Sambucinum, Dry Rot of Potato, by Serratia Plymuthica 5-6. *Biocontrol. Sci. Technol.* **2008,** *18*, 1005-1016.

58. Gallidis, E.; Papadopoulos, E.; Ptochos, S.; Arsenos, G. The Use of Targeted Selective Treatments against Gastrointestinal Nematodes in Milking Sheep and Goats in Greece Based on Parasitological and Performance Criteria. *Vet. Parasitol.* **2009,** *164*, 53-58.

59. Koner, A. L.; Ghosh, I.; Saleh, N.; Nau, W. M. Supramolecular Encapsulation of Benzimidazole-Derived Drugs by Cucurbit[7]Uril. *Can. J. Chem.* **2011,** *89*, 139-147.

60. Tanford, C. Protein Denaturation. Part C. Theoretical Models for the Mechanism of Denaturation. *Adv. Protein Chem.* **1970,** *24*, 1-95.

61. Wyman, J. The Binding Potential, a Neglected Linkage Concept. *J. Mol. Biol.* **1965,** *11*, 631-644.

62. Case, D. A.; Darden, T. A.; Cheatham, I., T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Merz, K. M.; Roberts, B., et al. *Amber 12*, University of Calfornia, San Francisco: 2012.

63. Onufriev, A. V.; Case, D. A.; Ullmann, G. M. A Novel View of pH Titration in Biomolecules. *Biochemistry* **2001,** *40*, 3413-3419.

64. Rogers, K. E.; Ortiz-Sánchez, J. M.; Baron, R.; Fajer, M.; de Oliveira, C. A. F.; McCammon, J. A. On the Role of Dewetting Transitions in Host-Guest Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2013,** *9*, 46-53.

65. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Determining Atom-Centered Charges: The RESP Model. *J. Phys. Chem.* **1993,** *97*, 10269-10280.

66. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. Application of RESP Charges to Calculate Conformational Energies, Hydrogen Bond Energies and Free Energies of Solvation. *J. Am. Chem. Soc.* **1993,** *115*, 9620-9631.

67. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006,** *25*, 247-260.

68. Vosko, S. H.; Nusair, L. W. M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980,** *58*, 1200-1211.

69. Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988,** *37*, 785-789.

70. Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.

71. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Caculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994,** *98*, 11623-11627.

72. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A., et al. *Gaussian 09*, Revision D.01; Gaussian, Inc.: Wallingford, CT, 2009.

73. Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990,** *11*, 431-439.

74. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004,** *25*, 1157-1174.

75. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K., et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004,** *47*, 1739-1749.

76. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006,** *49*, 6177-6196.

77. *Glide, Version 5.8*, Schrödinger, LLC: New York, NY, 2012.

78. Onufriev, A. V.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004,** *55*, 383-394.

79. Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* **2004,** *120*, 6363-6374.

80. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.

81. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997,** *72*, 1047-1069.

82. Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935,** *3*, 300-313.

83. Shirts, M. R.; Pande, V. S. Solvation Free Energies of Amino Acid Side Chain Analogs for Common Molecular Mechanics Water Models. *J. Chem. Phys.* **2005,** *122*, 134508.

84. Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear Scaling Schemes for Lennard-Jones Interactions in Free Energy Calculations. *J. Chem. Phys.* **2007,** *127*, 214108.

85. Steinbrecher, T.; Joung, I.; Case, D. A. Soft-Core Potentials in Thermodynamic Integration: Comparing One- and Two-Step Transformations. *J. Comput. Chem.* **2011,** *32*, 3253-3263.

86. Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **2003,** *107*, 9535-9551.

87. Kaus, J. W.; Pierce, L. T.; Walker, R. C.; McCammon, J. A. Improving the Efficiency of Free Energy Calculations in the Amber Molecular Dynamics Package. *J. Chem. Theory Comput.* **2013,** *9*, 4131-4139.

88. Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, I., T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H., et al. *Amber 14*, University of California, San Francisco: 2014.

89. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** *79*, 926-935.

90. Schafmeister, C. E. A. F.; Ross, W. S.; Romanovski, V. *Leap*, University of california, San Francisco: 1995.

91. Zhang, Y.; Feller, S. E.; Brooks, B. R.; Pastor, R. W. Computer Simulation of Liquid/Liquid Interfaces. I. Theory and Application to Octane/Water. *J. Chem. Phys.* **1995,** *103*, 10252-10266.

92. Zhang, X.; Gramlich, G.; Wang, X.; Nau, W. M. A Joint Structural, Kinetic, and Thermodynamic Investigation of Substituent Effects on Host-Guest Complexation of Bicyclic Azoalkanes by B-Cyclodextrin. *J. Am. Chem. Soc.* **2002,** *124*, 254-263.

93. Casey, J. R.; Grinstein, S.; Orlowski, J. Sensors and Regulators of Intracellular pH. *Nat. Rev. Mol. Cell Biol.* **2010,** *11*, 50-61.

94. M., M.; Baptista, A. M. Is the Prediction of pKa Values by Constant-pH Molecular Dynamics Being Hindered by Inherited Problems? *Proteins* **2011,** *79*, 3437-3447.

95. Williams, S. L.; Blachly, P. G.; McCammon, J. A. Measuring the Successes and Deficiencies of Constant pH Molecular Dynamics: A Blind Prediction Study. *Proteins* **2011,** *79*, 3381-3388.

96. Jorgensen, W. L. Foundations of Biomolecular Modeling. *Cell* **2013,** *155*, 1199-1202.

Chapter 5

Conformational dynamics and binding free energies of inhibitors of BACE-1: From the perspective of protonation equilibria

Abstract

BACE-1 is the β-secretase responsible for the initial amyloidogenesis in Alzheimer's disease, catalyzing hydrolytic cleavage of substrate in a pH-sensitive manner. The catalytic mechanism of BACE-1 requires water-mediated proton transfer from aspartyl dyad to the substrate, as well as structural flexibility in the flap region. Thus, the coupling of protonation and conformational equilibria is essential to a full *in silico* characterization of BACE-1. In this work, we perform constant pH replica exchange molecular dynamics simulations on both apo BACE-1 and five BACE-1-inhibitor complexes to examine the effect of pH on dynamics and inhibitor binding properties of BACE-1. In our simulations, we find that solution pH controls the conformational flexibility of apo BACE-1, whereas bound inhibitors largely limit the motions of the holo enzyme at all levels of pH. The microscopic pK_a values of titratable residues in BACE-1 including its aspartyl dyad are computed and compared between apo and inhibitor-bound states. Changes in protonation between the apo and holo forms suggest a thermodynamic linkage between binding of inhibitors and protons localized at the dyad. Utilizing our recently developed computational protocol applying the binding polynomial formalism to the constant pH molecular dynamics (CpHMD) framework, we are able to obtain the pH-dependent binding free

energy profiles for various BACE-1-inhibitor complexes. Our results highlight the importance of correctly addressing the binding-induced protonation changes in protein-ligand systems where binding accompanies a net proton transfer. This work comprises the first application of our CpHMD-based free energy computational method to protein-ligand complexes and illustrates the value of CpHMD as an all-purpose tool for obtaining pH-dependent dynamics and binding free energies of biological systems.

Introduction

Alzheimer's disease is a neurodegenerative disorder characterized by loss of memory and failure in cognitive abilities, resulting from synaptic dysfunction and neuronal death in the brain.[1-5] Major damages found in the brains of Alzheimer's patients include cerebral and vascular deposits of insoluble amyloid plaques, consisting of aggregates of amyloid β-peptide (Aβ).[6-8] Aβ occurs in two different forms, $A\beta_{40}$ and $A\beta_{42}$, and the overproduction and oligomerization of $A\beta_{42}$ is associated with the early onset of Alzheimer's disease.[9-12] Aβ is produced by sequential proteolytic cleavage of the type 1 transmembrane protein amyloid precursor protein (APP) by β- and γ-secretases.[13, 14] While γ-secretase generates several Aβ peptides varying in the length of C-termini, β-secretase, or β-site APP cleaving enzyme 1 (BACE-1), cleavage precisely gives the fibrillogenic $A\beta_{42}$.[13-15] Therefore, as it catalyzes the initial site-specific hydrolysis step of Aβ production, BACE-1 is an attractive therapeutic target for the treatment of Alzheimer's disease.[1-3, 16, 17]

As an aspartyl protease, the catalytic mechanism of BACE-1 involves two highly conserved aspartyl residues, Asp32 and Asp228, which form a symmetric dyad at the base of the catalytic cleft of the enzyme (Figure 5-1).[16] Analogous aspartyl dyads are found in the aspartyl protease family including pepsin, cathepsin D, renin, and HIV-1 protease.[18-21] The dyad is central to the hydrolytic cleavage of the substrate through a nucleophilic attack of water bound to the dyad.[19-23] Due to the general acid-base catalytic nature of the mechanism, the enzymatic activity of BACE-1 is maximal at pH 4.5 and strongly depends on solution pH.[24, 25]

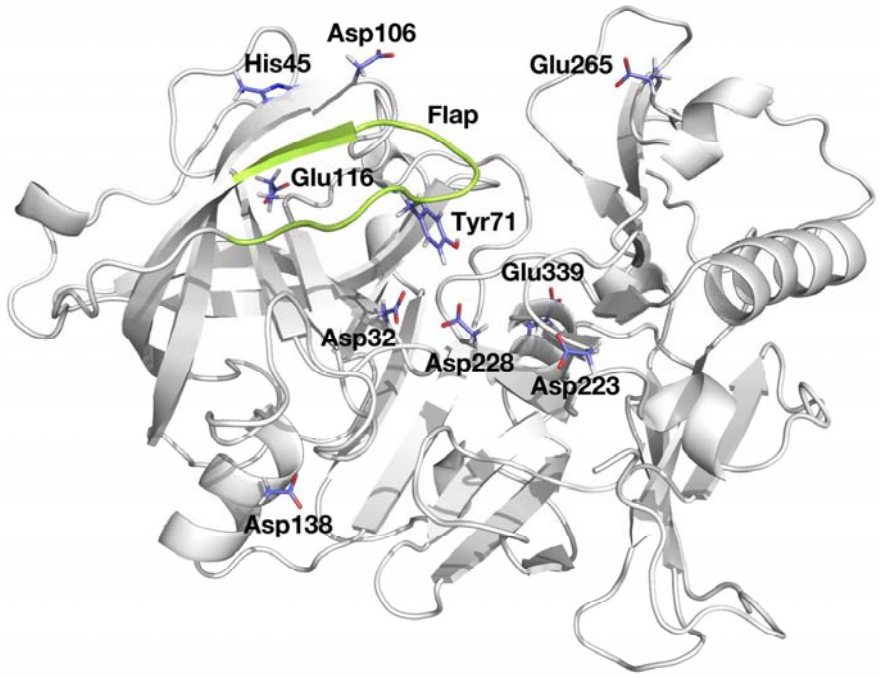


Figure 5-1. Structure of BACE-1, highlighted with titratable residues considered here and flap region (residues 67 to 77) in green.

The active site of BACE-1 is covered by an antiparallel β hairpin (henceforth referred to as the flap region; residues 67 to 77 shown in green in Figure 5-1) that is characteristic of aspartyl proteases.[16, 26-29] The X-ray crystal structures of other aspartyl proteases indicate that the flap is inherently flexible.[26-29] The flexibility of the flap region is likely utilized in catalysis, with transitions between open and closed conformations facilitating the entrance of substrates into the active site and release of hydrolytic products.[21, 29-31] The conserved Tyr71[20] located at the tip of the flap region is particularly essential for the conformational transitions of the flap. Observations from X-ray crystallographic structures and molecular dynamics (MD) simulations suggest that variation in hydrogen bond patterns between Tyr71 and surrounding residues such as Lys107, Lys75, Gly74, Glu77, and Trp76 enables the flexible motions of the flap.[21, 29, 31-33] In the presence of inhibitors, Tyr71 can directly interact with bound inhibitors and lock the flap in the closed state.[31, 33, 34]

Given that the enzymatic activity of BACE-1 depends on solution pH and that the structural flexibility is intrinsic to catalysis, a comprehensive understanding of the pH dependence of BACE-1 dynamics would greatly benefit drug design efforts. A detailed description of the protonation state of the aspartyl dyad is also important as all known bound inhibitors directly contact the dyad. Several computational efforts have attempted to determine the protonation state of the dyad, employing methods such as molecular mechanics (MM),[35] quantum mechanics (QM),[36] QM/MM,[37] molecular docking,[38] and continuum electrostatics calculations.[39] However, the conformational flexibility of BACE-1 was not rigorously addressed in these computations.

The importance of accounting for conformational flexibility in pK_a computations has been well established.[40-45] The instantaneous pK_a of a titratable group is determined by its electrostatic environment, which is affected by the given conformation of protein and protonation states of other titratable residues.[46] Changes to the conformation of the protein can alter the electrostatics, which may, in turn, induce a shift in the pK_a of titratable groups. The prevalence of such coupling of protonation and conformational equilibria has been observed in various systems both computationally and experimentally.[43, 47-58] Furthermore, complex formation between protein and small molecules can also induce changes in the pK_a values of titratable groups on either binding partner.[50, 59-66]

Consequently, several computational methods have been developed to explicitly account for conformational changes in pK_a computations.[48, 67-72] Among these, various flavors of constant pH molecular dynamics (CpHMD) methodologies have emerged to incorporate pH as an additional external thermodynamic variable to the conventional MD framework.[47, 73-77] CpHMD simulations have been successfully applied to predict pK_a values of titratable groups in proteins[47, 73-79] and nucleic acids,[80-82] as well as to explain the acid-base catalysis by RNase A[83] and to understand the mechanisms behind the pH-dependent conformational changes.[58, 84]

Conventional molecular simulations or free energy computations typically employ fixed protonation states that are identical for both free and bound states, set prior to the computations. This assumption ignores the possibility of protonation states changing upon binding and can lead to significant errors when protein-ligand binding is a pH-dependent process.[59] Furthermore, the pH-

dependent conformational dynamics cannot be appropriately addressed if the protonation states are fixed while conformational fluctuations propagate.

Recognizing the lack of a standard protocol to rigorously account for proton-linked ligand binding to protein, we recently developed a protocol utilizing CpHMD to compute pH-dependent binding free energies.[85] In our computational method, the binding polynomial formalism devised by Wyman[86] is applied with the CpHMD framework to obtain a pH-dependent correction to a reference free energy of binding obtained at a given level of pH ($\Delta G°_{ref,pH}$). The proton-linked binding free energy then can be expressed using the notation used by Tanford:[87]

$$\Delta G^{\circ}(\mathrm{pH}) = \Delta G^{\circ}_{\mathrm{ref,pH}} - k_B T \ln(10) \int_{\mathrm{pH_{ref}}}^{\mathrm{pH}} \left\{ Z_{PL}(\mathrm{pH}) - Z_P(\mathrm{pH}) - Z_L(\mathrm{pH}) \right\} d\mathrm{pH} \,, \quad (1)$$

where Z_{PL}, Z_P, and Z_L are the total charges of the protein-ligand complex, protein, and ligand, respectively, as obtained from the CpHMD simulations. In this work, Z_L is omitted from consideration since the inhibitors considered here do not titrate in the physiological pH range (Figure 5-2). The integral in the second term in Eq. 1 provides a thermodynamic relation that holds for pH-dependent ligand binding in cases where proton binding to different titratable sites may be cooperative (*i.e.*, no assumptions are made about sites titrating independently). When applied to binding of small molecules to the cucurbit[7]uril host, this CpHMD-based free energy method accurately obtained the pH-dependent binding free energy profiles.[85] Also, the method demonstrated that the traditional use of fixed protonation states for both free and bound states predicted based on pH 7 in free energy computations could give errors larger than 2 kcal/mol in the host-guest systems with a single titratable site.[85] Given the complexity of protein environment where multiple titratable groups exist, the corresponding error in free energy may be even larger in protein-ligand binding, highlighting the significance of accurate description of the binding-induced pK_a shifts in free energy computations.

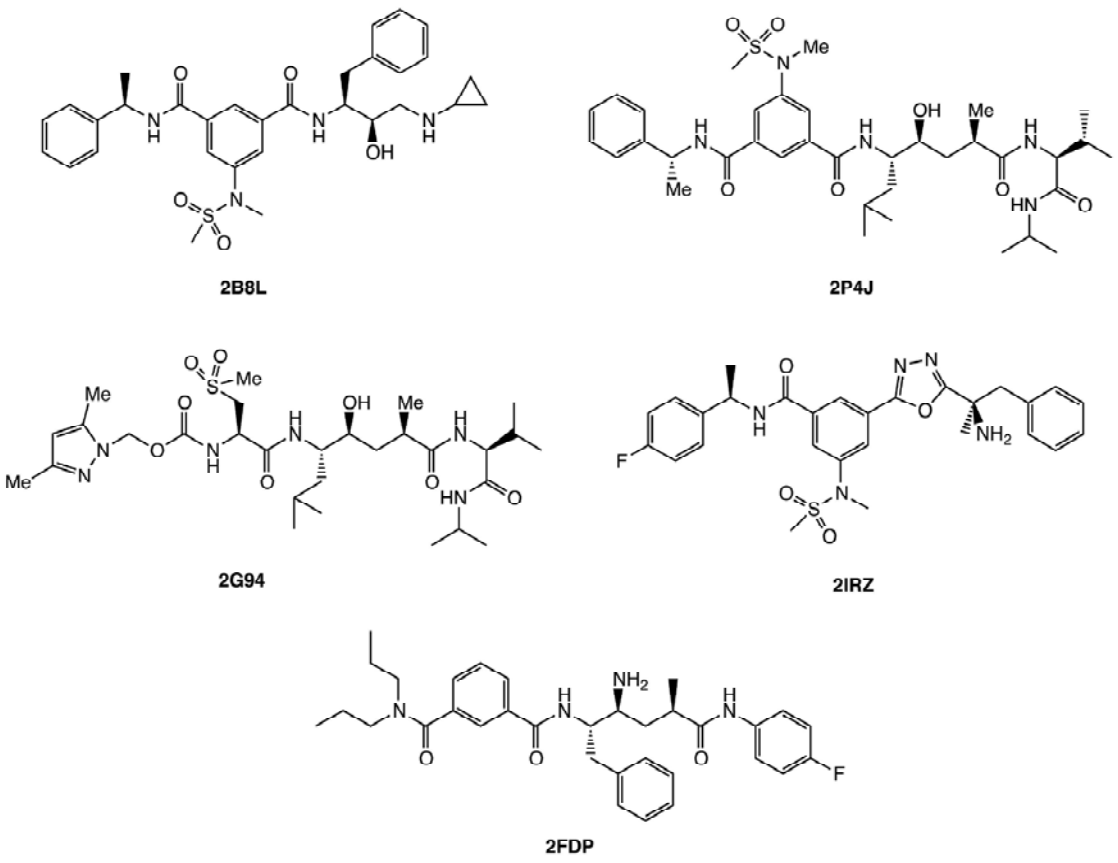


Figure 5-2. Chemical structures of BACE-1 inhibitors considered in this study. The PDB IDs are shown for complex structures of BACE-1 bound with respective inhibitors.

In this work, we have performed constant pH replica exchange molecular dynamics (pH-REMD) simulations to study the proton-linked conformational dynamics and binding free energies of inhibitors to BACE-1. The conformational changes of the flap region of BACE-1 in the absence and presence of inhibitors are analyzed with respect to solution pH, which is found to act as a conformational switch. The microscopic pK_a values of ten titratable residues including the aspartyl dyad in BACE-1-inhibitor complexes are obtained from the pH-REMD simulations and compared with those computed for free enzyme. The results show significant binding-induced shifts in the pK_a values. We further apply our CpHMD-based computational protocol to these results, computing the pH-dependent binding free energy profiles of various inhibitors to BACE-1. The results demonstrate that incorrect assignment of protonation state to the titratable groups can result in errors of over 8 kcal/mol in free

energy computations for the systems considered here, highlighting the significance of correctly addressing the binding-induced protonation changes. To the best of our knowledge, this work presents the first application of CpHMD simulations to quantify binding in protein-ligand systems and shows high utility for addressing pH effects in computer-aided drug discovery workflows.

Methods

Constant pH replica exchange molecular dynamics

Baptista *et al*. developed the constant pH molecular dynamics (CpHMD) method to enable concurrent sampling of discrete protonation states and conformational space according to the semi-grand canonical ensemble.[74] In this work, we apply the flavor of CpHMD coupled with replica exchange (pH-REMD)[88] implemented in the AMBER 14 suite of programs[89] with generalized Born (GB) electrostatics. In the CpHMD simulations, a conventional molecular dynamics (MD) simulation is periodically interrupted by a Monte Carlo (MC) step, in which a change in the protonation state of a random titratable residue is considered.[77] Acceptance of the new protonation state is contingent on the computed transition free energy, ΔG_{trans}:

$$\Delta G_{trans} = k_B T(pH - pK_{a,ref})\ln 10 + \Delta G_{elec} - \Delta G_{elec,ref}, \quad (2)$$

where pH enters as an external thermodynamic parameter and k_BT is the Boltzmann constant multiplied by the temperature of the system. For a given value of pH, the difference in electrostatic free energy that accompanies the change in protonation being considered, ΔG_{elec}, is computed with respect to the difference in electrostatic free energy that accompanies analogous change in protonation for a model compound, $\Delta G_{elec,ref}$, which has a known pK_a value ($pK_{a,ref}$). As all titratable residues in this study are protein residues, the model compounds referenced in Eq. 2 are individual amino acids in GB solvent. The respective $pK_{a,ref}$ values for CpHMD in AMBER 14 are 4.0 for Asp, 4.4 for Glu, 6.5 and 7.1 for His, 9.6 for Tyr, and 10.4 for Lys.[89] Computing ΔG_{trans} with respect to these model compounds enables cancellation of non-classical terms. The Metropolis criterion is then applied to ΔG_{trans} to determine

whether to accept the proposed change in protonation, and the MD simulation is resumed. Repeated application of these steps builds an ensemble of protonation states along the MD trajectory.

In pH-REMD, the exchange between adjacent replicas is achieved in the pH dimension at a fixed conformation, whose acceptance ratio is dependent on the MC exchange probability for replicas i and j:

$$P_{i\to j} = \min\left\{1, \exp\left[\ln 10(N_i - N_j)(pH_i - pH_j)\right]\right\}, \quad (3)$$

where N_i is the number of titratable protons in replica i and pH_i is the pH of replica i prior to the exchange attempt.[88, 90] By enhancing the sampling through an application of replica exchange scheme, the method has been shown to achieve faster convergence and better sampling in both conformational and protonation spaces compared to original CpHMD.[79, 88]

From the pH-REMD simulations, the pK_a of a given residue is computed as the midpoint of titration by fitting titration data to the Hill equation:

$$s = \frac{1}{1+10^{n(pK_a - pH)}}, \quad (4)$$

where s is the fraction of deprotonated species for a given residue and n is the Hill coefficient. The fraction of deprotonated species (s) for a titratable group is obtained at each value of pH from the pH-REMD simulations. In using Eq. 1 shown in Introduction, the fractions of protonated species (1-s) of the protein-inhibitor complex and protein can be translated into Z_{PL} and Z_P, respectively.

Computing the pH dependence of protein-ligand binding free energies

The binding polynomial formalism devised by Wyman[86] was used by Tanford to study protein denaturation[87] and by Mason and Jensen to examine the pH dependence of protein-protein binding.[91] Following their theoretical foundations, we recently applied it to binding of a general receptor to a ligand with a single titratable site, and the detailed derivation of the formalism can be found therein.[85] Here we briefly outline the theoretical basis of the method and show its usage for protein-ligand binding with multiple titratable sites.

First, consider complex formation between a protein (P) with a single titratable site and a ligand (L) that does not titrate in the biological range of pH levels. The association can be expressed as a general equation with the apparent equilibrium constant, K_{app}:

(5) $$\{P\} + L \overset{K_{app}}{\rightleftharpoons} \{PL\},$$

where the curly braces indicate that the ensembles of protein and protein-ligand complex (PL) may contain different protonated forms of the titratable species. K_{app} can be expressed in terms of binding polynomials through an application of the thermodynamic cycle for proton-linked ligand binding shown in Scheme 5-1:

$$\begin{array}{ccc} P + L & \overset{K_b^\circ}{\rightleftharpoons} & P:L \\ K_a^F \downarrow\uparrow & & \uparrow\downarrow K_a^C \\ HP^+ + L & \overset{K_b^{+,\circ}}{\rightleftharpoons} & HP^+:L \end{array}$$

Scheme 5-1. Thermodynamic cycle for complex formation between a protein (P) with a single titratable site and a ligand (L).

(6) $$K_{app} = \frac{[PL] + [HPL^+]}{([P] + [HP^+])[L]} = K_b^\circ \frac{\left(1 + \frac{[HPL^+]}{[PL]}\right)}{\left(1 + \frac{[HP^+]}{[P]}\right)},$$

where the concentrations of the given species, instead of activities, are shown assuming ideal dilute solutions and $K°_b$ is the equilibrium constant of binding for a reference reaction in which net proton transfer is ignored.

The overall free energy of binding ($\Delta G°$) can then be expressed by using logarithmic representations of the acid dissociation constants for the free protein (pK_a^F) and protein-ligand complex (pK_a^C):

(7) $$\Delta G^\circ(pH) = \Delta G^\circ_{ref} - k_B T \ln\left(\frac{1 + 10^{pK_a^C - pH}}{1 + 10^{pK_a^F - pH}}\right),$$

where $\Delta G°_{ref}$ is the free energy of binding for the reference reaction.

In cases where proton binding to different titratable sites may be cooperative, Wyman[92] derived a relation between K_{app} and pH such that

$$\frac{\partial \ln K_{app}}{\partial \ln[H^+]} = \Delta\nu_{H^+} = Z_{PL} - (Z_P + Z_L) \quad (8)$$

where, using the notation used by Tanford,[87] $\Delta\nu_{H+}$ is the change in the number of bound protons in the protein-ligand complex, relative to the number of protons bound to the protein and ligand individually. With $\Delta Z = Z_{PL} - (Z_P + Z_L)$, integration of Eq. 8 provides a thermodynamic relation for proton-linked ligand binding where titratable sites may interact:

$$\Delta G^\circ(pH) = \Delta G^\circ_{ref,pH} - k_B T \ln(10) \int_{pH_{ref}}^{pH} \left\{ Z_{PL}(pH) - Z_P(pH) - Z_L(pH) \right\} dpH \,, \quad (9)$$

where Z_{PL}, Z_P, and Z_L are the total charges for protein-ligand complex, protein, and ligand, respectively, obtained from the pH-REMD simulations. Eq. 9, or Eq. 1, provides framework for computing the pH-dependent binding free energy through a correction term to the reference free energy of binding. In this work, Z_L is omitted from consideration since the inhibitors considered here do not titrate in the physiological pH range.

Preparation of the BACE-1 systems for simulations

The X-ray crystallographic structures of BACE-1 in complex with inhibitors N-[(1S,1R)-bnezyl-3-(cyclopropylamino)-2-hydroxypropyl]-5-[methyl(methylsulfonyl)amino-N'-[(1R)-1-phenylethyl]isophthalamide (PDB ID 2B8L);[93] N-[(1S,2S,4R)-2-hydroxy-1-isobutyl-5-({(1S)-1-[(isopropylamino)carbonyl]-2-methylpropyl}amino-4-methyl-5-oxopentyl]-5-methyl(methylsulfonyl)amino]-N'-[(1R)-1-phynylethyl]isophthalamide (PDB ID 2P4J);[94] N~2~-[(2R,4S,5S)-5-{[N-{[3,5-dimethyl-1H-pyrazol-1-yl)methoxy]carbonyl}-3-(methylsulfonyl)-$_L$-alanyl]amino}-4-hydroxy-2,7-dimethyloctanoyl]-N-isobutyl-$_L$-valinamide (PDB ID 2G94);[95] 3,{5-[(1R)-1-amino-1-methyl-2-phenylethyl]-1,3,4-oxadiazol-2-yl}-N-[(1R)-1-(4-fluorophenyl)ethyl]-5-[methyl(methylsulfonyl)amino]benzamide (PDB ID 2IRZ);[96] and N1-((2S,3S,5R)-3-amino-6-(4-fluorophenylamino)-5-methyl-6-oxo-1-phenylhexan-2-yl)-N3,N3-dipropylisophthalamide (PDB ID 2FDP)[97] were used to study the pH-dependent inhibitor binding to BACE-1. The chemical structures of

the inhibitors are shown in Figure 5-2. A segment between residues Gly158 and Ser169 is not solved in 2B8L, 2IRZ, and 2FDP crystal structures and hence this loop was built by homology modeling using the Structure Prediction Wizard module of Schrödinger's Prime program.[98-100] The FASTA sequence of protein including the missing loop region for each X-ray structure was obtained from UniProt.[101] The mutation AWA that exist in 2B8L and 2IRZ structures made for crystallography were corrected to the original sequence of KWE.[93, 96] Using the homologs found by BLAST search algorithm,[102] a chimera model containing the missing loop region was built for each structure. The homology-modeled loop region was energy-refined for relaxation using the Refine Loops panel of the Prime program.[98] Apo structure of BACE-1 was generated by removing the bound inhibitor from the refined 2B8L structure to even out any effects that may arise from homology modeling.

The geometries of the inhibitors were optimized at the B3LYP/6-31G(d) level of theory[103-106] using the Gaussian 09 suite of programs.[107] The electronic potentials (ESP) for the optimized geometries of the inhibitors were computed using MK radii[108] at the HF/6-31G(d) level of theory. Subsequently, the atomic point charges were computed from the ESPs by applying the RESP procedure using the antechamber module[109] in the AmberTools 14 suite of programs.[89] All other force field terms including Lennard-Jones parameters for use in molecular dynamics (MD) simulations were taken from the general AMBER force field (GAFF).[110]

Conventional molecular dynamics simulations

Prior to the pH-REMD simulations, each system was subject to short conventional molecular dynamics (cMD) simulations for equilibration purposes. Protonation states of the titratable groups are assigned using the PROPKA web server.[71, 111-113] All protein force field parameters are taken from the AMBER ff14SB force field,[89] while the ligand parameters are taken from the AMBER GAFF force field.[110] Each system was solvated with TIP3P water[114] and counterions were added to neutralize the system by tleap program.[115] Water molecules were first minimized and simulated for 150 ps in the NPT ensemble with a harmonic restraint of 2.0 kcal/mol $Å^2$ on the protein and ligand heavy atoms to relax the water. The entire system was then minimized and heated to 300 K over 500 ps. Two equilibrations with respective duration of 200 ps were performed. First, the system was equilibrated at constant

volume and temperature (NVT) using a Langevin thermostat.[116] Following this, the second equilibration was carried out at constant pressure and temperature (NPT) using a Berendsen barostat (ntp=1)[117] with isotropic position scaling to bring the system to a stable density. A 100 ns cMD production was then performed in the NVT ensemble. The Particle Mesh Ewald summation method was used to compute long-range electrostatic interactions,[118, 119] and short-range non-bonded interactions are truncated at 8 Å in the periodic boundary conditions. All dynamics are conducted using the *pmemd.cuda* module of AMBER 14 suite of programs.[89, 116, 120] The RMSD of the apo structure indicated a convergence to the starting conformation after first 20 ns, ensuring the stability of the system (Figure 5-S1).

Constant pH replica exchange molecular dynamics simulations details

Preliminary investigation of the pK_a shift of the titratable groups of BACE-1 upon inhibitor binding was carried out using the PROPKA web server by submitting the apo and holo structures.[71, 111-113] The results indicated binding-induced pK_a shifts for a number of ionizable residues, with the most pronounced shift for the aspartyl dyad (data not shown). A total of ten ionizable residues within 12 Å of the active site were chosen for titration, Asp32, Asp106, Asp138, Asp223, Asp228, Glu116, Glu265, Glu339, His45, and Tyr71. As the titratable groups in the inhibitors considered in this study have pK_a's above 12, the titration was carried out on the chosen protein side chains only.

pH-REMD simulations were performed using the *pmemd.cuda.MPI* module of AMBER 14 suite of programs for the pH range between 1 and 12 stepped by 1 pH unit.[88, 89] All simulations employed the generalized Born (GB)-Neck 2 implicit solvent model (igb=8)[121] with a salt concentration of 0.1 M. To ensure equilibration in the implicit solvent, a 5,000 step minimization was carried out for each system starting from the conformation obtained from the cMD simulations, with positional restraints on all heavy atoms with a force constant of 20 kcal/mol $Å^2$. The system was then heated to 300 K over 500 ps using a Langevin thermostat while maintaining the positional restraints applied to all heavy atoms with a force constant of 5 kcal/mol $Å^2$, followed by a 1 ns equilibration. The Monte Carlo (MC) moves for titration were performed during the production stage only, where the MC steps taken every 10 fs and exchange between replicas at adjacent pH attempted every 100 steps, *i.e.*, 200 fs with a 2 fs time step. The production simulations were carried out for 60 ns and data from last 50 ns were used

for analyses. In the equilibration and production steps, the bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[122]

Simulation analyses

RMSF and clustering analyses, and reconstruction of the pH-based trajectories from the pH-REMD simulations were performed using cpptraj program in the AmberTools 14 suite of programs.[89] Clustering analyses for the pH-based trajectories used pairwise RMSDs computed for Cα atoms between frames to divide the trajectories into five clusters using the average-linkage algorithm.[123] Fitting of titration data to the Hill equation (Eq. 4) to obtain the pK_a values was carried out using Matlab.[124]

Results

pH-Dependent conformational dynamics of apo BACE-1

As conformational transitions of BACE-1 are suggested to play a role in catalysis,[29-31] we first examine the dynamics intrinsic to apo BACE-1 before exploring the effect of inhibitor binding in the following section. Conventional molecular dynamics (cMD) simulations of duration 100 ns are carried out prior to constant pH molecular dynamics (pH-REMD) simulations in order to provide equilibration phase for apo and three inhibitor-bound systems prepared by homology modeling (see Methods). Using the protonation states assigned using the PROPKA program,[71, 111-113] the cMD simulations provide a benchmark for comparing pH-REMD simulations.

In order to quantify the extent to which each residue fluctuates, we compute the root-mean-square fluctuation (RMSF) of each residue in apo BACE-1 from the cMD trajectory. As shown in Figure 5-3A, higher RMSF values are noted for the flap region (residues 67 to 77), consistent with the suggestion by others.[29, 31-34] Taking a closer look at the flap region, we measure the distance between the center of mass of the aspartyl dyad and Tyr71, which is located at the tip of the flap region. From the change of the dyad-flap distance plotted in Figure 5-3B, it is evident that the flap region undergoes transitions between open and closed conformations, within the distance range of 10 Å (closed) to 30 Å (open). In the closed conformation of the flap, we observe a water-mediated hydrogen bond network

that includes the dyad, Ser35, Tyr71, Arg128, Thr231, and Thr329 (Figure 5-3C), agreeing with the findings from previous studies.[23, 31, 34] On the other hand, when the flap is open, the dyad forms contacts mediated by water with Ser35, Thr231, and Arg235, while Tyr71 is entirely exposed to solvent (Figure 5-3C).

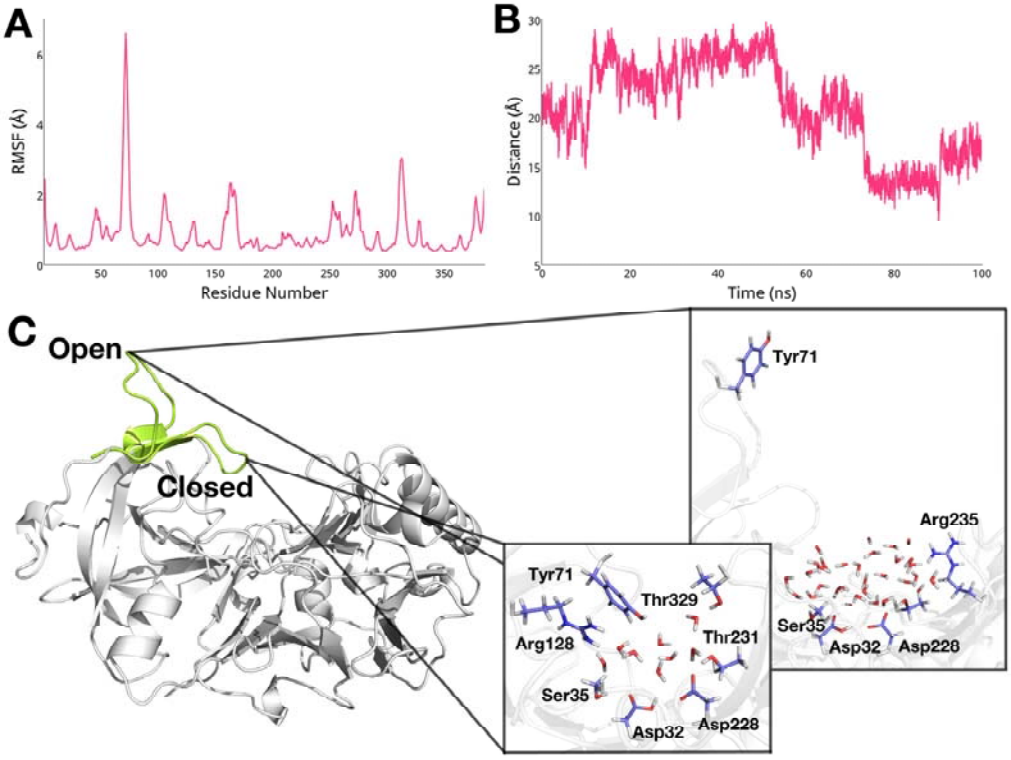


Figure 5-3. (A) RMSF of apo BACE-1 from the cMD simulations. (B) Change in distance between the dyad and Tyr71 in the flap region. (C) Open and closed conformations of the flap and interactions between the dyad and surrounding residues in each conformation.

While the cMD simulation is performed with fixed protonation states, we choose to examine the effect of protonation equilibria on the conformational flexibility of apo BACE-1 as the enzymatic activity of BACE-1 is shown to be pH-dependent.[24, 25] We focus this investigation on a comparison of the dynamics of the flap region at acidic (pH 1 to 3) and basic (pH 9 to 11) pH levels. The conformational space of the flap region sampled at these differing levels of pH is quantified by measuring the distances between the center of mass of the dyad and Tyr71 (Figure 5-4A). Since the pH-based trajectories reconstructed from the pH-REMD simulations are not time-dependent, distributions

of the measured distances are presented here. The distance between the dyad and Tyr71 exhibits a bimodal distribution at acidic pH, with the flap sampling both open and closed states. When the flap is closed, an average distance between the dyad and flap is about 8 Å, while open conformations are also populated, having an average dyad-flap distance of ~ 17 Å. In order to visualize structural characteristics typical at varying solvent environments, we carry out a clustering analysis on the conformations sampled at acidic and basic pH. We find three dominant conformers that encompass 86% of total conformations sampled at acidic pH. The flap regions from representative structures of these three clusters are shown in red in Figure 5-4B, further illustrating the flap region sampling both open and closed conformations at acidic pH.

At basic pH, the flap exhibits noticeably different dynamics compared to acidic pH. The flap region remains over 10 Å away from the dyad and most frequently found in the open conformation with a distance of about 17 Å from the dyad (Figure 5-4A). From the clustering analysis, a single conformer is found to represent 82% of total conformations sampled at basic pH. In this typical conformer, the flap is in widely open state with a distance of 17.6 Å from the dyad and completely exposed to the solvent area, as shown in blue in Figure 5-4B.

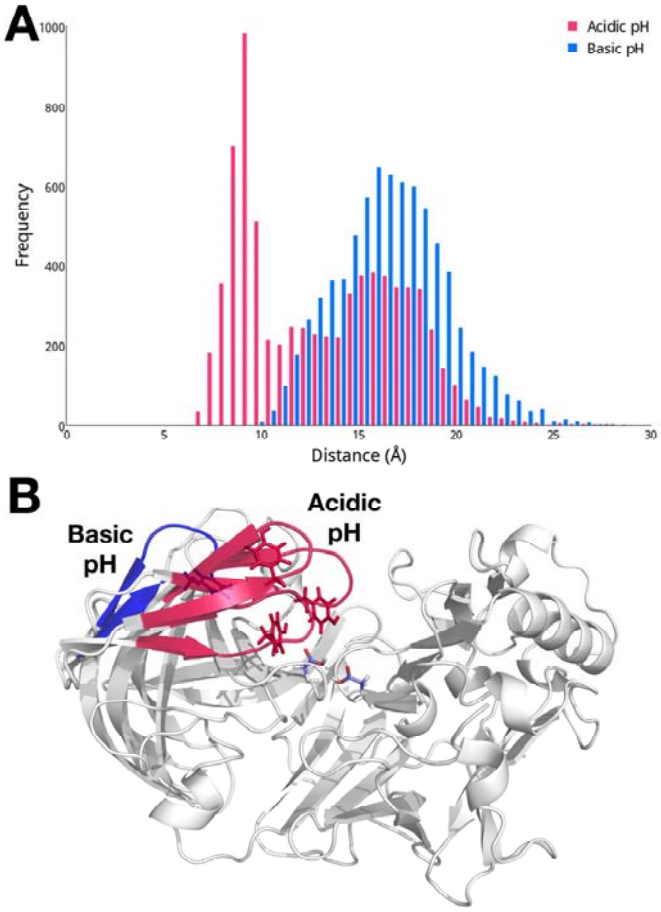


Figure 5-4. (A) Distributions of distances between the dyad and Tyr71 in apo BACE-1 at acidic (pH 1 to 3; red) and basic (pH 9 to 11; blue) pH. (B) Conformations of the flap region in the cluster representative structures of apo BACE-1 at acidic and basic pH. The distances between the dyad and flap of the representative structures are 9.17 Å, 11.0 Å, and 15.7 Å at acidic pH; and 17.6 Å at basic pH.

Conformational dynamics of BACE-1 in complex with inhibitors

We continue to probe the changes in dynamics of BACE-1 that accompany inhibitor binding. While the shifts between open and closed conformations of the flap region are observed in the cMD simulations of apo BACE-1, the flap remains in a closed state in the inhibitor-bound cMD simulations. The distances between the center of mass of the dyad and Tyr71 observed in the cMD simulations of BACE-1 in complex with the inhibitors 2B8L and 2FDP, respectively, are shown in Figure 5-5A. In comparison to the distribution of distances observed in the cMD simulation of apo BACE-1 (Figure 5-3B), we observe significantly less flexibility in the flap region in the 2B8L and 2FDP systems.

Fluctuation of the dyad-flap distance in the 2B8L system during the early stage of the simulation is likely due to structural instability arising from the homology modeling. However, after 40 ns, the flap region in the 2B8L complex achieves a stable state and remains closed at a distance of about 15 Å from the dyad. It is worth noting that the measured dyad-flap distances in the inhibitor-bound systems are inherently longer than that in apo enzyme due to the presence of the inhibitors in the active site. Similarly, the flap region in the 2FDP system maintains an average dyad-flap distance of ~ 11 Å. The bound inhibitors have hydrophobic interactions with Tyr71 and Phe108 while forming hydrogen bonds with the polar residues including the dyad, Tyr71, Thr231, and Ser325. These hydrogen bond networks effectively lock the flap in the closed state, as shown in Figure 5-5B for the 2B8L system. Similar trends in the dyad-flap distance are observed in other inhibitor-bound systems (Figure 5-S2).

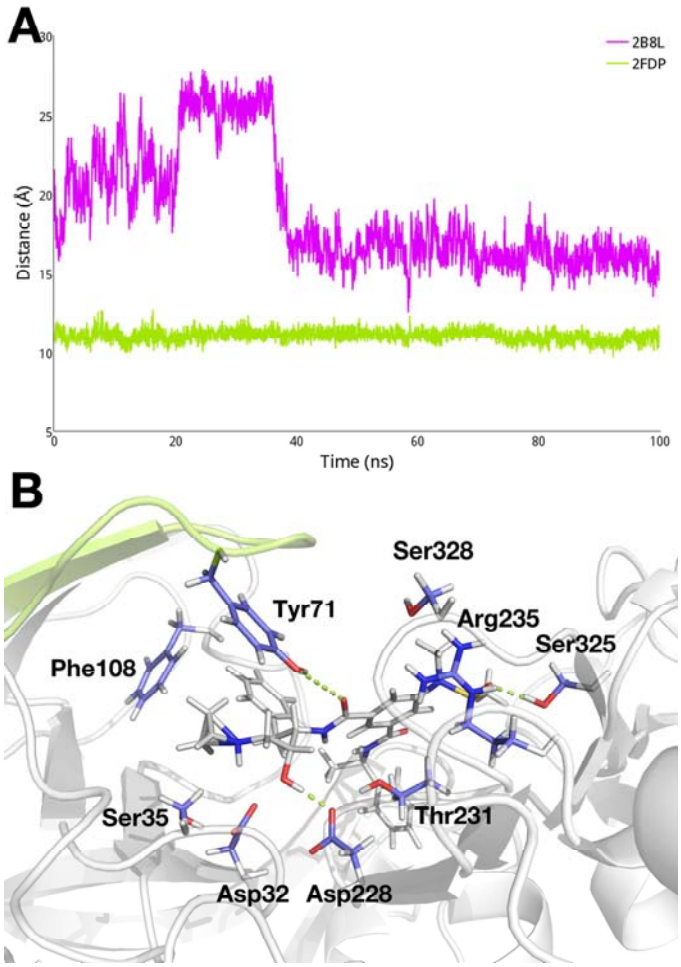


Figure 5-5. (A) Change in distance between the dyad and flap region in the 2B8L (purple) and 2FDP (green) systems in the respective cMD simulations. (B) Interactions between the bound inhibitor in the 2B8L complex and surrounding residues.

As we observe the pH-controlled dynamics of the flap region in apo BACE-1, we also carry out similar analyses on the holo systems to examine the effect of pH on conformations of the flap in the presence of inhibitors. Unlike the apo system, a clear distinction in the dynamics of BACE-1 contingent on pH levels is not found in the inhibitor-bound systems. The distances between the center of mass of the dyad and Tyr71 indicate that a closed state of the flap is dominant in the 2B8L system at both pH, with average distances of 15.1 Å and 17.5 Å, respectively (Figure 5-6A). From the clustering analysis, the flap is in closed state in the typical structure representing 67 to 77 % of the total conformations sampled at both pH conditions, as shown in Figure 5-6C, forming similar contacts with the bound

inhibitor and dyad as observed in the cMD simulations (Figure 5-5B). Other cluster representative structures also exhibit similarly closed conformations the dyad-flap distances varying in the range of 15 to 19 Å. In the 2FDP system, the flap region exhibits essentially identical dynamics regardless of solvent pH. The flap is found primarily in a closed conformation where the dyad-flap distance is 15.2 Å on average (Figure 5-6B). Consistent with these findings, the cluster representative structures of the 2FDP system at both acidic and basic pH have the flap in closed state, similar to the 2B8L complex (Figure 5-6C). Similar trends in the distributions of the dyad-flap distance in varying pH conditions are observed in other inhibitor-bound systems (Figure 5-S3).

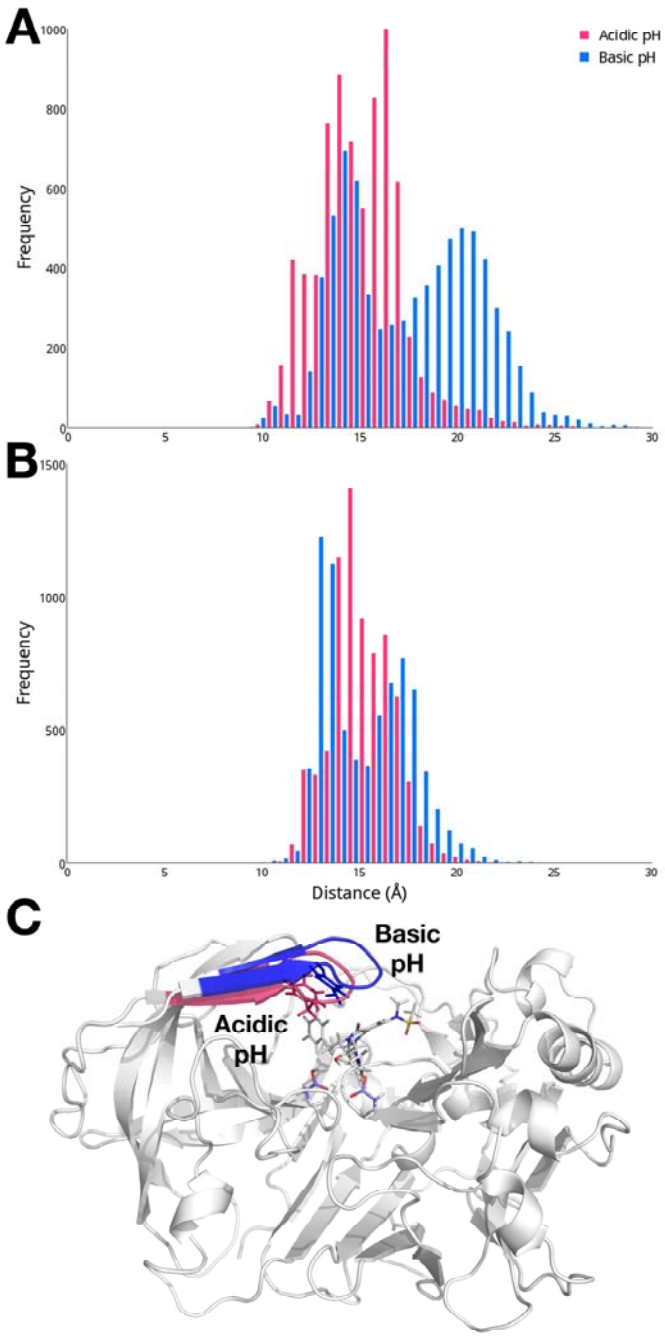


Figure 5-6. Distribution of distances between the dyad and flap at acidic (pH 1 to 3; red) and basic (pH 9 to 11; blue) pH. (A) 2B8L. (B) 2FDP. (C) Conformations of the flap in the cluster representative structures of the 2B8L system at acidic and basic pH.

Computation of pK_a values of apo BACE-1

We further investigate the acidic properties of the aspartyl dyad and surrounding titratable sites in both apo and holo BACE-1. The pK_a values of the different titratable residues considered are obtained by fitting the Hill equation (Eq. 4; see Methods) to titration data obtained from the pH-REMD simulations conducted at different levels of pH (Table 5-1). We first analyze the computed pK_a values of the titratable groups in apo BACE-1 and compare the values obtained for inhibitor-bound systems in the following section.

Table 5-1. pK_a values obtained from pH-REMD simulations. The corresponding Hill coefficients are shown in parentheses.

	Apo	**2B8L**	**2P4J**	**2G94**	**2IRZ**	**2FDP**
Asp32	5.0 ± 0.2 (0.51)	9.7 ± 0.2 (1.72)	8.7 ± 0.6 (0.34)	9.54 ± 0.05 (2.19)	8.4 ± 0.7 (0.40)	8.9 ± 0.7 (0.33)
Asp106	3.99 ± 0.01 (1.05)	3.94 ± 0.01 (0.89)	4.00 ± 0.01 (0.97)	3.91 ± 0.01 (0.97)	3.85 ± 0.01 (0.96)	3.85 ± 0.01 (0.99)
Asp138	6.26 ± 0.01 (1.16)	6.10 ± 0.02 (0.93)	5.59 ± 0.04 (0.86)	5.72 ± 0.01 (0.92)	6.16 ± 0.01 (1.04)	6.40 ± 0.01 (1.11)
Asp223	5.20 ± 0.03 (0.91)	5.99 ± 0.03 (1.07)	5.03 ± 0.02 (0.92)	5.34 ± 0.01 (0.95)	4.53 ± 0.03 (1.02)	5.46 ± 0.02 (1.06)
Asp228	5.9 ± 0.5 (0.41)	8.4 ± 0.1 (1.26)	8.3 ± 0.1 (1.75)	8.1 ± 0.2 (0.90)	8.3 ± 0.1 (0.79)	9.7 ± 0.6 (0.30)
Glu116	7.23 ± 0.02 (1.02)	7.9 ± 0.1 (0.77)	7.51 ± 0.04 (0.86)	7.08 ± 0.01 (1.26)	7.42 ± 0.04 (0.77)	8.27 ± 0.09 (1.34)
Glu265	4.31 ± 0.01 (0.96)	4.39 ± 0.01 (0.95)	4.28 ± 0.01 (0.95)	4.32 ± 0.01 (0.97)	4.32 ± 0.01 (0.98)	4.33 ± 0.01 (0.98)
Glu339	6.6 ± 0.1 (0.73)	7.01 ± 0.07 (0.88)	7.46 ± 0.02 (1.01)	6.16 ± 0.03 (1.03)	7.35 ± 0.02 (1.66)	6.88 ± 0.02 (1.00)
His45	5.6 ± 0.1 (1.15)	4.3 ± 0.1 (0.91)	6.33 ± 0.02 (1.06)	6.26 ± 0.02 (1.16)	5.99 ± 0.08 (0.87)	5.95 ± 0.02 (1.17)
Tyr71	9.68 ± 0.01 (1.02)	9.4 ± 0.3 (1.05)	9.62 ± 0.02 (1.13)	9.71 ± 0.03 (1.99)	9.4 ± 0.03 (0.86)	9.92 ± 0.09 (0.85)

The predicted pK_a values of the aspartyl dyad, Asp32 and Asp228, are 5.0 ± 0.2 and 5.9 ± 0.5, respectively. Titration curves for the dyad are shown in Figure 5-7, plotted as fraction of deprotonated species of each residue as a function of pH. From the titration curves, both aspartates are completely protonated at acidic pH levels (pH < 3), while fully deprotonated at basic pH (pH > 8). Between pH 4 and 8, Asp32 and Asp228 exist in an ensemble of protonated and deprotonated forms. To illustrate this, Asp32 and Asp228 are approximately 20% and 40% deprotonated, respectively, at pH 4.5, the pH at which BACE-1 is most active. These observed shifts from the typical pK_a of Asp residues (4.0)[89, 125] may aid in the proton transfer step required in BACE-1 catalysis. The Hill coefficients of the dyad deviate from one, suggesting that titration of Asp32 and Asp228 is coupled. The coupled titration observed for these residues also contributes to greater noise in their respective titration curves, leading to larger errors in in their pK_a (Table 5-1).

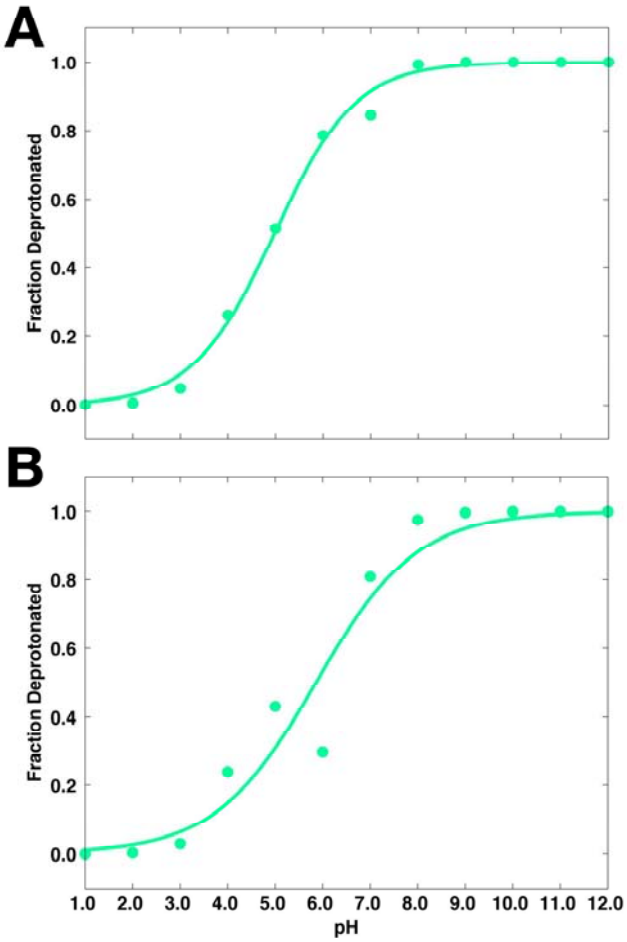


Figure 5-7. Titration curves from the pH-REMD simulations of apo BACE-1. (A) Asp32. (B) Asp228.

The pK_a values of the remaining titratable residues are also reported in Table 5-1. All ionizable residues besides the dyad appear to titrate independently of each other, as suggested by their Hill coefficients that are approximately one. Furthermore, the statistical errors from fitting procedure to obtain the pK_a's of these residues are minimal. Similar to those of the dyad, the computed pK_a values of Asp138, Asp223, Glu116, and Glu339 are shifted higher than the canonical pK_a values for Asp (4.0) and Glu (4.4),[89, 125] whereas Asp106 exhibits titration behavior in line with model Asp. As these residues are distant from the active site, these deviations in the computed pK_a values arise mainly from the microenvironments surrounding them. For instance, Glu116 and Glu339 are buried in the protein interior (Figure 5-1), where the microscopic dielectric constants can be different from that of the bulk

solvent.[42, 43, 126, 127] In this case, the neutral, protonated forms of the glutamates are favored at the solution pH where they would normally be charged if they were exposed to the bulk solvent. Titration curves for the remaining titratable residues are provided in Figure 5-S4.

Binding-induced pK_a shifts of titratable residues in BACE-1

Having observed the pK_a values of various titratable groups in apo BACE-1, we shift our attention to the inhibitor-bound systems. First considering the dyad, the pK_a values of Asp32 and Asp228 are shifted toward more basic values of pH when various inhibitors are present, with computed pK_a values greater than 8.1. Representative titration curves for the dyad in 2B8L and 2FDP systems (purple curves) are compared with those in apo BACE-1 (green curves) in Figure 5-8. Similar curves corresponding to other inhibitor-bound systems are shown in Figures 5-S5-S7.

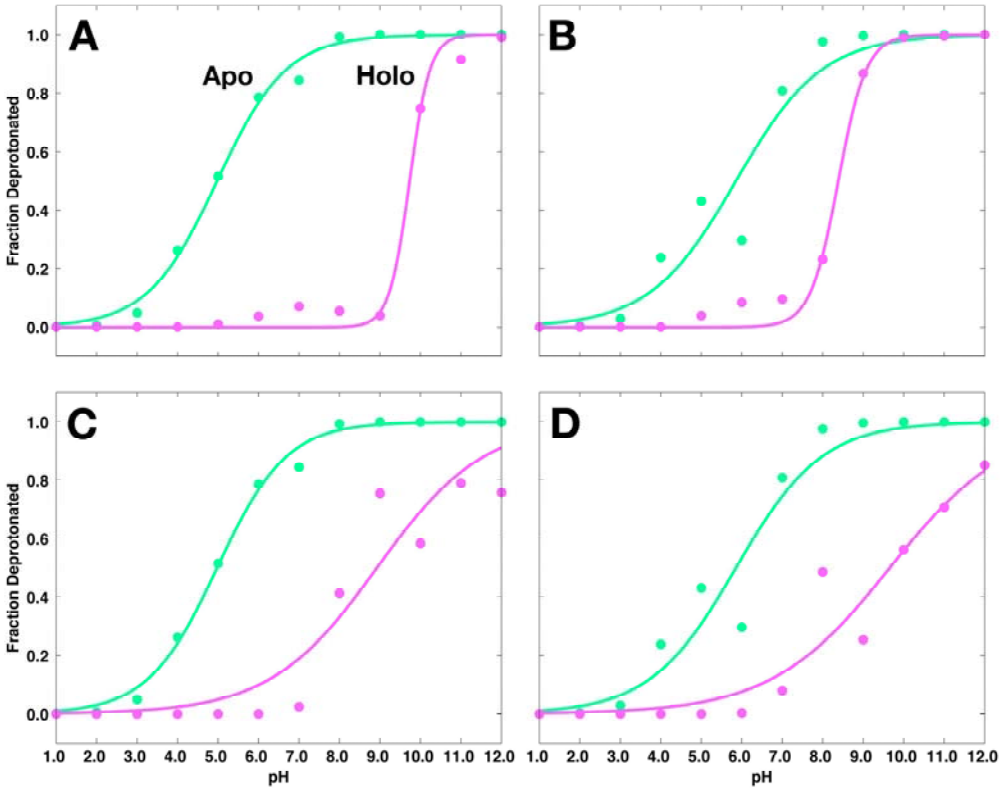


Figure 5-8. Titration curves for the aspartyl dyad in the 2B8L and 2FDP systems, shown in purple. (A) Asp32, (B) Asp228 in the 2B8L system. (C) Asp32, (D) Asp228 in the 2FDP system. Respective titration curves of the dyad in apo BACE-1 are shown in green.

In the 2B8L system, Asp32 and Asp228 have pK_a values of 9.7 ± 0.2 and 8.4 ± 0.1, respectively (shifted + 4.7 and + 2.5 pK units relative to their values in apo BACE-1). From the titration curves in Figure 5-8A and 8B, it is apparent that both Asp32 are Asp228 are significantly protonated under pH 8 – 9 in the 2B8L system. Above this pH, the dyad exists in an ensemble of different protonation states between pH 8 and 10. Similar binding-induced pK_a shifts are observed for the dyad in 2FDP system. In this case, the computed pK_a's of Asp32 and Asp228 are 8.9 ± 0.7 and 9.7 ± 0.6, respectively. As shown in Figure 5-8C and 8D, in the 2FDP complex, both aspartates are predominately protonated at the pH levels below 7, and exist as an ensemble of different protonated forms at basic pH. In both cases, as in the apo enzyme, we observe the coupled titration behavior of the dyad and relatively large errors during the fitting of titration data to the Hill equation.

The remaining titratable groups examined, all of which are distant from the binding site, do not undergo pK_a shifts larger than 1.3 pK units upon inhibitor binding (Figures 5-S5-S9). The narrow pK_a shifts of these titratable groups upon complex formation suggest that binding of inhibitors to BACE-1 is thermodynamically linked to a proton transfer that is primarily localized at the catalytic dyad.

pH Dependence of the binding free energies of inhibitors

The large shifts in pK_a's of the aspartyl dyad upon binding of inhibitors to BACE-1 indicate that proton binding is linked to complex formation in BACE-1 systems. Utilizing our recently developed computational protocol,[85] we present the pH-dependent changes in free energies of binding of inhibitors to BACE-1. Application of the binding polynomial formalism to the results obtained from the pH-REMD simulations provides pH-dependent corrections to the reference binding free energies obtained for a given pH, $\Delta G°_{ref,pH}$ (Eq. 1). The reference free energies of binding of inhibitors for BACE-1-inhibitor systems are obtained from experimental association constants measured at pH 4.5.[95, 97, 128, 129]

Binding free energy profiles as functions of pH of the 2B8L and 2FDP complexes are shown in Figure 5-9. Significant changes in binding free energies from the reference free energies at pH 4.5 are observed as solution pH increases. Considering the 2FDP complex, we see that binding is most favorable at acidic pH, where the maximum affinity is -11.2 kcal/mol. As pH increases, binding

becomes less favorable. This is most pronounced in the pH range of 5 to 10, where the aspartyl dyad that interacts directly with the bound inhibitor begins to populate deprotonated states, leading to an ensemble of protonated and deprotonated species of the dyad at the pH levels between 5 and 10 (Figure 5-8B). As the deprotonated forms of the dyad develop, hydrogen bonds made between the diprotonated dyad and bound inhibitors at low pH are lost, and consequently, the binding affinity becomes weaker in this pH range and is least favorable with an affinity of 2.2 kcal/mol at pH 12. Similar binding free energy profiles as functions of pH for other inhibitor-bound systems can be found in Figure 5-S10.

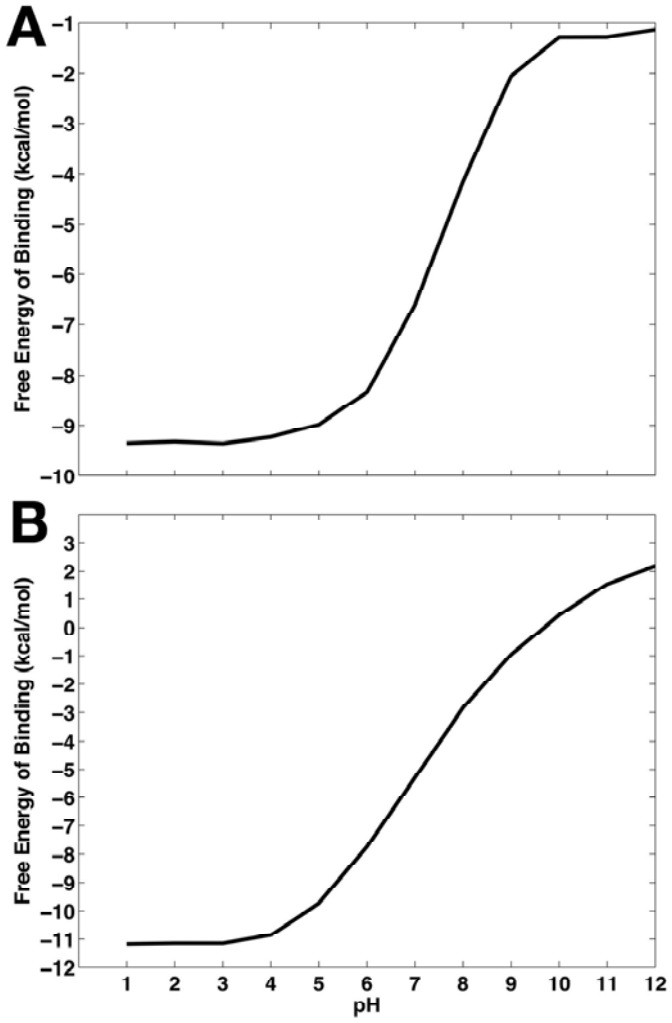


Figure 5-9. Binding free energy profiles of inhibitors as functions of pH. (A) 2B8L. (B) 2FDP.

In Table 5-2, the binding free energies are compared in the pH ranges that are most relevant to biological conditions. The binding free energy for the 2FDP system changes by 5.1 kcal/mol between pH 4.5 and 7. Comparison to the binding free energy at pH 10 leads to more dramatic changes; for instance, the binding affinity differs by 10.8 kcal/mol between pH 4.5 and pH 10 for the 2FDP complex. As mentioned above, such large changes are due to the shifts in protonation state of the dyad at these pH levels from the diprotonated state observed at pH 4.5, and further highlight the significance of correctly accounting for the protonation states of the titratable groups which accompany a net proton transfer upon inhibitor binding.

Table 5-2. Binding free energies of the inhibitors upon complex formation with BACE-1. All energies are reported in kcal/mol.

	$\Delta G°_{ref,pH}$ (pH 4.5)	$\Delta G°$ (pH 7)	$\Delta\Delta G°_{pH7-ref}$	$\Delta G°$ (pH 10)	$\Delta\Delta G°_{pH10-ref}$
2B8L	-9.1[128]	-6.6	2.5	-1.3	7.8
2P4J	-12.3[128]	-8.7	3.6	-4.0	8.3
2G94	-13.1[95]	-9.4	3.7	-4.9	8.2
2IRZ	-10.1[129]	-6.1	4.0	-2.6	7.5
2FDP	-10.4[97]	-5.3	5.1	0.4	10.8

Discussion

The dependence of BACE-1 enzymatic activity on solution pH and the need for conformational change to accompany the catalysis are both well established.[24, 25, 30, 31, 34] As a promising therapeutic target for the treatment of Alzheimer's disease, understanding the detailed mechanism underlying the pH dependence of BACE-1 dynamics and enzymatic activity is imperative for structure-based drug design. In this work, we have performed constant pH replica exchange molecular dynamics (pH-REMD) simulations to examine the proton-linked conformational dynamics and inhibitor binding properties of BACE-1.

Significant flexibility of the flap region (residues 67 to 77), resulting in transitions between open and closed conformations, is noted for apo BACE-1 during conventional molecular dynamics (cMD) simulations (Figure 5-3). As we further probe the effect of pH on conformational flexibility of apo BACE-1, distinctive conformations that are characteristic of different pH environments are

captured from the pH-REMD simulations. At acidic pH, both open and closed conformations of apo BACE-1 are significantly populated, whereas a single conformer with the flap closed predominates in basic conditions (Figure 5-4).

While we observe the flexible nature of apo BACE-1, the presence of inhibitors at the active site of BACE-1 greatly reduces conformational mobility of the enzyme (Figure 5-5). The bound inhibitors form various hydrophobic and polar interactions with the surrounding residues, holding the flap in a tightly closed state. Similarly closed conformations of the flap are observed regardless of the varying pH conditions, indicating that the structural flexibility of BACE-1 is largely limited by the presence of bound inhibitors.

We determine the microscopic protonation states of the dyad residues and surrounding titratable residues to further probe the mechanism underlying the pH dependence of the catalytic activity of BACE-1. First, the pK_a values of ten titratable residues in various BACE-1 systems are obtained from the pH-REMD simulations (Table 5-1). The computed pK_a values of Asp32 and Asp228 in apo BACE-1 are 5.0 ± 0.2 and 5.9 ± 0.5, respectively, shifted from the typical pK_a of Asp (Figure 5-7). At acidic pH, protonated states are dominant for both aspartates of the dyad. This allows the dyad to form hydrogen bonds with the flap residues, *i.e.*, Tyr71 and Thr72 and allows for closed conformations to be sampled at low pH. On the other hand, open conformations of the flap are also populated at acidic pH; these open conformations likely aid in substrate binding and product release in the course of catalysis.[29-31] When we solvate the open and closed conformers from the pH-REMD simulations, which are performed in implicit solvent, similar water occupancies to those in the cMD simulations are observed (Figure 5-3C). When the flap is open, the active site becomes largely accessible to water, which is needed to act as nucleophile for the hydrolytic catalysis by BACE-1. Also, water molecules entering the active site help to compensate for the breaking of hydrogen bonds between the dyad and flap by forming alternative hydrogen bonds and mediating the hydrogen bond networks with surrounding charged residues such as Ser35 and Arg235. Therefore, conformational transitions between open and closed states of the flap at acidic pH allow for channeling of solvent, substrates, and hydrolytic products to and from the active site in catalysis. With the pK_a values near the pH of optimal

enzymatic activity, *i.e.*, pH 4.5, the dyad is also able to easily gain and release proton(s) during the catalytic cycle. Hence the pK_a values of the dyad of BACE-1 shifted from the typical value may be an evolutionary result to achieve the maximal activity at pH 4.5.

As pH increases, deprotonated species of both the dyad and Tyr71 start to emerge. Consequently, the hydrogen bond networks observed in acidic conditions no longer persist and are instead replaced by water molecules, leading to the primarily open conformation of the flap at basic pH. While this open state of the flap is stabilized energetically through water-dyad interaction, it is likely that the persistence of the open conformation at basic pH disables the enzyme's ability to corral the substrate into the binding site for catalysis. Therefore, observation of the invariably open state of the flap at basic pH is consistent with the suggested role of flexibility of flap in the catalysis at acidic pH.[29-31] Our results thus indicate that the conformational dynamics intrinsic to the enzymatic catalysis of BACE-1 are modulated by solution pH, further suggesting the enzyme's structural adaptation during the evolution for its maximal activity.

The Hill coefficients for fitting the titration data to the Hill equation suggest the titration for the aspartyl dyad is cooperative in both free and complexed BACE-1, while independent, uncoupled titrations are observed for other residues considered. Such coupled titration behavior makes computing the microscopic pK_a values difficult. Hence, the statistical errors associated with computing the pK_a's of the dyad are higher than those of other titratable residues.

At pH 4.5, the aspartyl dyad exists in an ensemble of protonated and deprotonated species in apo BACE-1 (Figure 5-7). Upon binding of inhibitors, however, significant shifts in the pK_a values are observed for the dyad, with both Asp32 and Asp228 having pK_a values between 8 and 10 (Table 5-1). Inhibitor binding effectively alters the protonation state of the dyad at pH 4.5 to its diprotonated form in all cases studied here. The protonated forms are preferred for both aspartates in the presence of inhibitor. This diprotonated state likely compensates for the unfavorable energetics associated with desolvation upon inhibitor binding, allowing for hydrogen bonds between the dyad and the bound inhibitor. The minimal pK_a shifts observed for the remaining titratable residues of BACE-1 upon

complex formation imply a thermodynamic linkage between inhibitor binding and proton transfer primarily localized at the dyad.

Among several computational efforts to determine protonation state of the aspartyl dyad in BACE-1, a recent work by Domínguez *et al*. also examined the pK_a's of buried titratable groups in the 2B8L and 2IRZ systems,[39] whose computed values are compared with ours in Table 5-3. At first glance, the predicted pK_a values differ the most between the two works for Asp138 and Asp228. In Domínguez *et al.*, the pK_a of Asp228 does not deviate much from $pK_{a,ref}$ of Asp (4.0) in both systems compared, while Asp32 was predicted to undergo more significant pK_a shift relative to its dyad partner. On the other hand, pK_a's of both aspartates in the dyad shifted to more basic values of pH in our calculations.

Table 5-3. Comparison of computed pK_a values of the titrable groups in 2B8L and 2IRZ.

	2B8L		**2IRZ**	
	This work	**Domínguez *et al.* [39]**	**This work**	**Domínguez *et al.* [39]**
Asp32	9.7 ± 0.2	8.35	8.4 ± 0.7	6.88
Asp138	6.10 ± 0.02	3.23	6.16 ± 0.01	3.38
Asp228	8.4 ± 0.1	4.47	8.3 ± 0.1	4.01
Glu116	7.9 ± 0.1	5.95	7.42 ± 0.04	6.09
Glu339	7.01 ± 0.07	6.10	7.35 ± 0.02	6.26

When comparing the results in Table 5-3, several distinctions in the methods used in two studies for pK_a prediction should be noted. First, the GB-OBC implicit solvent model[130] was used in Domínguez *et al*.[131] while the GB-Neck 2 model, in which improved results have been obtained with the added parameters to the GB-OBC,[121] was employed in the pH-REMD method used here. Also, the internal dielectric constant of 10[131] was used in their work while the GB-Neck 2 implementation in Amber 14 employs 1.[121] In addition to the difference in the force fields utilized, perhaps more importantly, the conformational changes upon inhibitor binding were not rigorously accounted for in their study. Although the detailed comparison of the algorithms used for pK_a calculation is beyond the scope of this work, addressing the conformational aspect is particularly important for studying BACE-1 due to its flexible dynamic nature. This is especially crucial when computing the pK_a values of the titratable groups in the absence of bound inhibitors, as the conformational fluctuations in the flap also imply the change in solvent accessibility. Hence, the dielectric response of the aspartyl dyad can be

different in the presence and absence of bound inhibitors, which in turn can affect the computed pK_a values. Consequently, it is evident that conformational transitions accompanying binding of inhibitors should be accounted for in calculation of pK_a values of BACE-1. As all previous attempts to compute the pK_a's of the titratable groups in BACE-1 have been largely limited to the use of static X-ray crystallographic structures, our results obtained from concurrent sampling of protonation and conformational spaces by pH-REMD provide a new insight into the microscopic pK_a values of BACE-1.

Application of our recently developed constant pH molecular dynamics (CpHMD)-based computational protocol,[85] which applies the binding polynomial formalism to address the pH dependence of binding free energies, enables us to obtain proton-linked binding free energy profiles of various inhibitors. As shown in Figures 5-9 and 5-S10, all inhibitors bind most strongly at acidic pH. The changes in binding free energies are most pronounced in the pH range of 4 to 10, which essentially encompasses most biological reactions. The deviations in binding free energies within this pH range from the reference binding free energies at pH 4.5 arise from the shift in populations of major protonated species of the titratable residues, primarily those of the dyad. Between pH 5 and 10, the dyad starts to populate the deprotonated species (Figure 5-8), and as the deprotonated forms of the dyad develop, hydrogen bonds made between the diprotonated state of the dyad and bound inhibitors at low pH break. Subsequently, the binding free energies of the inhibitors become very unfavorable as pH increases. Such observations are impossible with cMD simulations where the protonation states are fixed and fractional protonation is not allowed. This highlights the benefit of using CpHMD method in order to address cases in which changes in protonation states are critical.

Furthermore, our results emphasize the importance of correctly addressing the binding-induced changes in protonation states in protein-ligand systems where binding accompanies a net proton transfer. In conventional molecular modeling or free energy computations, the protonation states of the titratable groups, which are set ahead of time, are fixed and assumed to be identical for both free and bound states. Consistent with this convention, consider a hypothetical scenario in which both Asp32 and Asp228 are assumed to be completely protonated in both apo and holo states. In the case of 2B8L

system, such protonation state of the dyad will result in binding free energy of the inhibitor of -9.3 kcal/mol (Figure 5-9A). On the other hand, when both aspartates are considered fully deprotonated, the binding free energy of the inhibitor is -1.3 kcal/mol. In these two extreme scenarios where the identical, discrete protonation state of the dyad is assumed for both free and bound states, the binding free energies deviate from the true free energy in which the protonation states are considered separately for apo and holo states and fractional protonation is allowed. The errors are as large as 8 kcal/mol for the 2B8L system and similar deviations are noted for other inhibitor-bound systems considered here, ranging between 8 and 12.6 kcal/mol. Such errors are nontrivial and the magnitude is in great excess of typical errors from free energy computations.[132, 133]

In addition, we note the lack of binding free energies of inhibitors to BACE-1 that are experimentally measured at pH levels other than pH 4.5. For BACE-1, the inhibition assays are traditionally carried out at pH 4.5 where the catalytic activity of BACE-1 is maximal. However, from a free energy computational standpoint, it would be greatly beneficial if binding free energies were measured at other pH levels to incorporate the effect of pH into free energy computations. Availability of experimental reference binding energies at various pH will be of great importance to pushing the free energy computation field forward.

The results presented here demonstrate the dynamics of BACE-1 controlled by solvent pH. The flexible motions of the flap region at low pH, assisted by the diprotonated state of the aspartyl dyad, enable the enzyme's optimal catalytic mechanism at acidic environment, implying a linkage between the protonation equilibria, conformational dynamics, and catalytic activity of BACE-1. In addition, we show the thermodynamic relation between binding of inhibitors and protons at the active site of BACE-1. Our results highlight the importance of accurately accounting for the protonation states of the titratable groups in protein-ligand systems where ligand binding is pH-dependent. Furthermore, we show that the CpHMD method can be used as an all-purpose tool to assess the pH-dependent dynamics and to quantify the binding free energies for protein-ligand systems where the protonation equilibria play an important role. To the best of our knowledge, this work presents the first application of our CpHMD-based free energy method to protein-ligand systems. In using the method, absolute

binding free energies obtained by computational free energy calculations such as thermodynamic integration can be used in cases where experimental association constants are not available. Our results highlight high utility of CpHMD method to address the effect of pH on conformational dynamics and inhibitor binding in computer-aided drug discovery workflows.

Acknowledgements

This work was supported by the NSF (MCB-1020765), NIH (NIH GM31749), Howard Hughes Medical Institute, National Biomedical Computation Resource (NBCR), and NSF supercomputer centers.

Chapter 5 is a minimally modified reprint of the material as it appears in Meekyum Olivia Kim, Patrick G. Blachly, and J. Andrew McCammon, "Conformational dynamics and inhibitor binding of BACE-1: From the perspective of protonation equilibria," submitted to *PLoS Computational Biology*, 2015. The dissertation author was the primary investigator and author of this paper.

Supporting Information

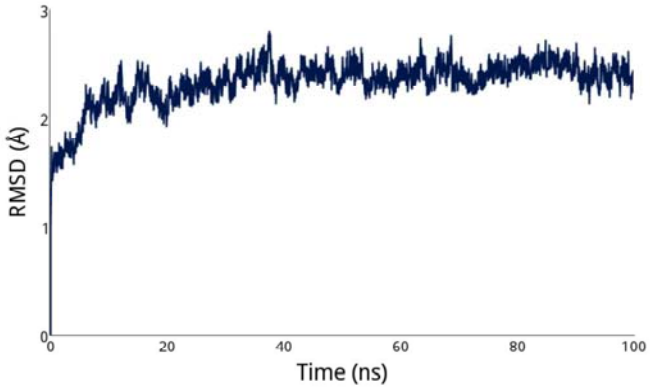


Figure 5-S1. RMSD of apo BACE-1 in the cMD simulation.

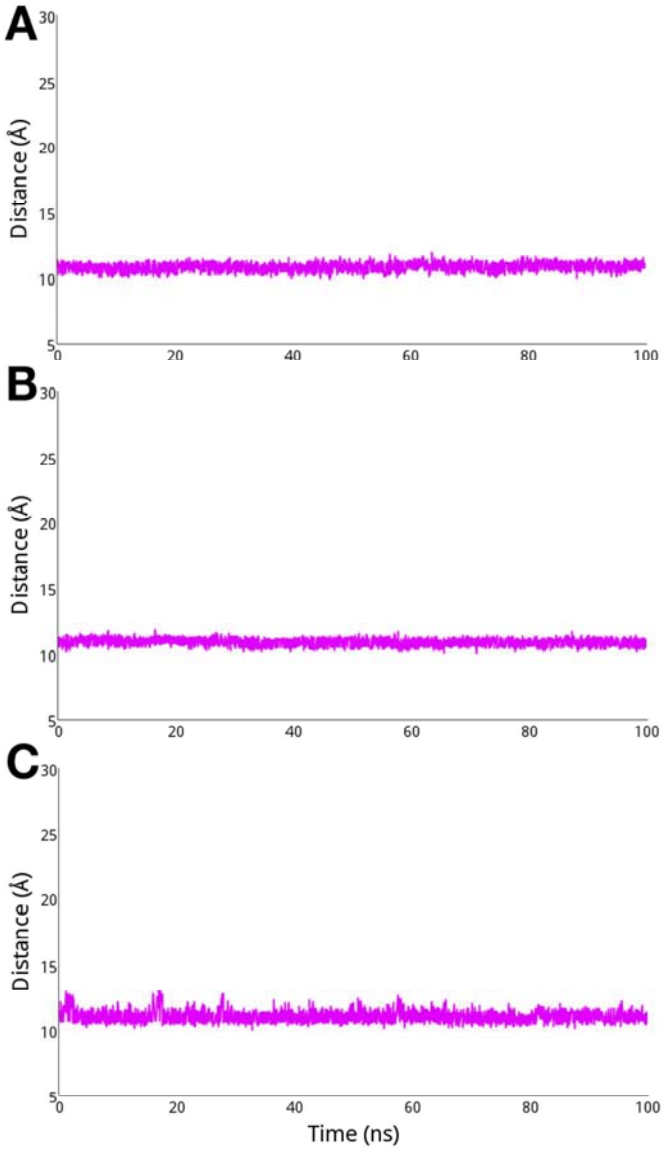


Figure 5-S2. Change in distance between the dyad and flap region observed in the cMD simulations. (A) 2P4J. (B) 2G94. (C) 2IRZ.

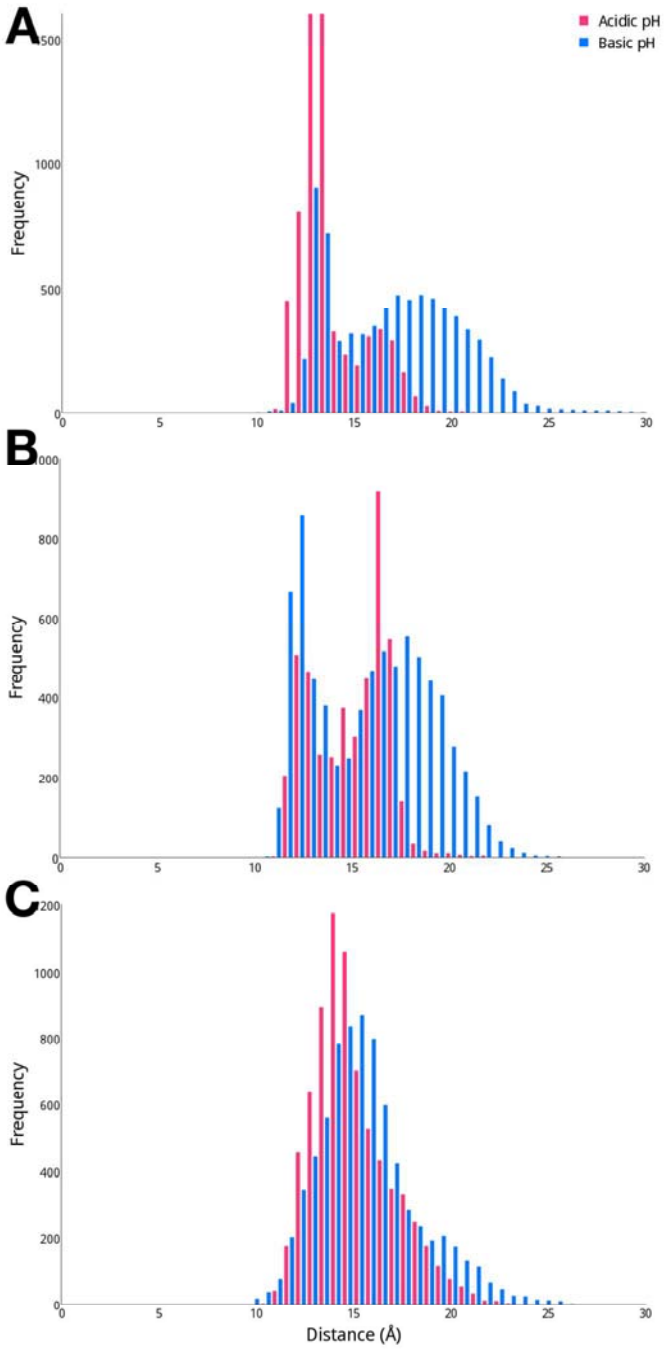


Figure 5-S3. Distribution of distances between the dyad and flap at acidic (pH 1 to 3; red) and basic (pH 9 to 11; blue) pH. (A) 2P4J. (B) 2G94. (C) 2IRZ.

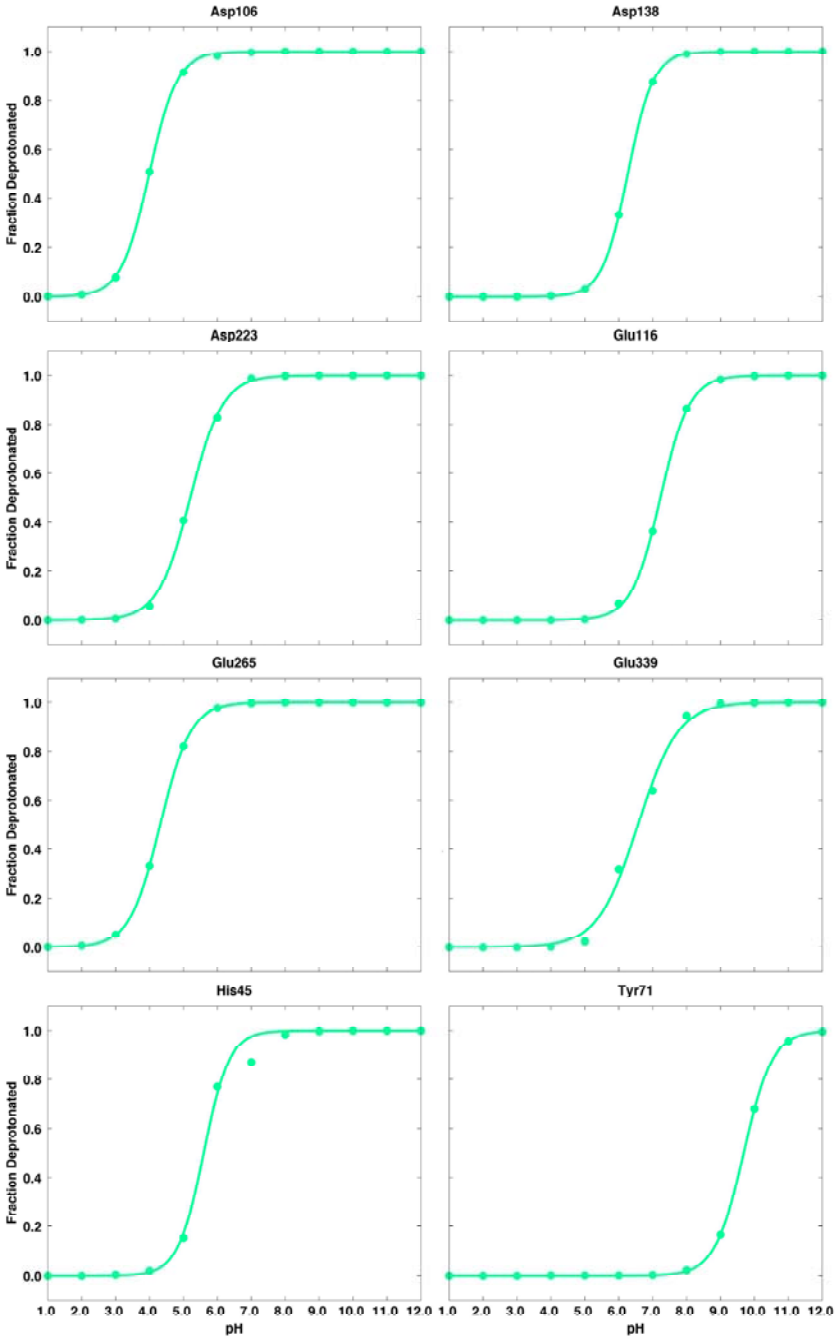


Figure 5-S4. Titration curves from the pH-REMD simulations of apo BACE-1.

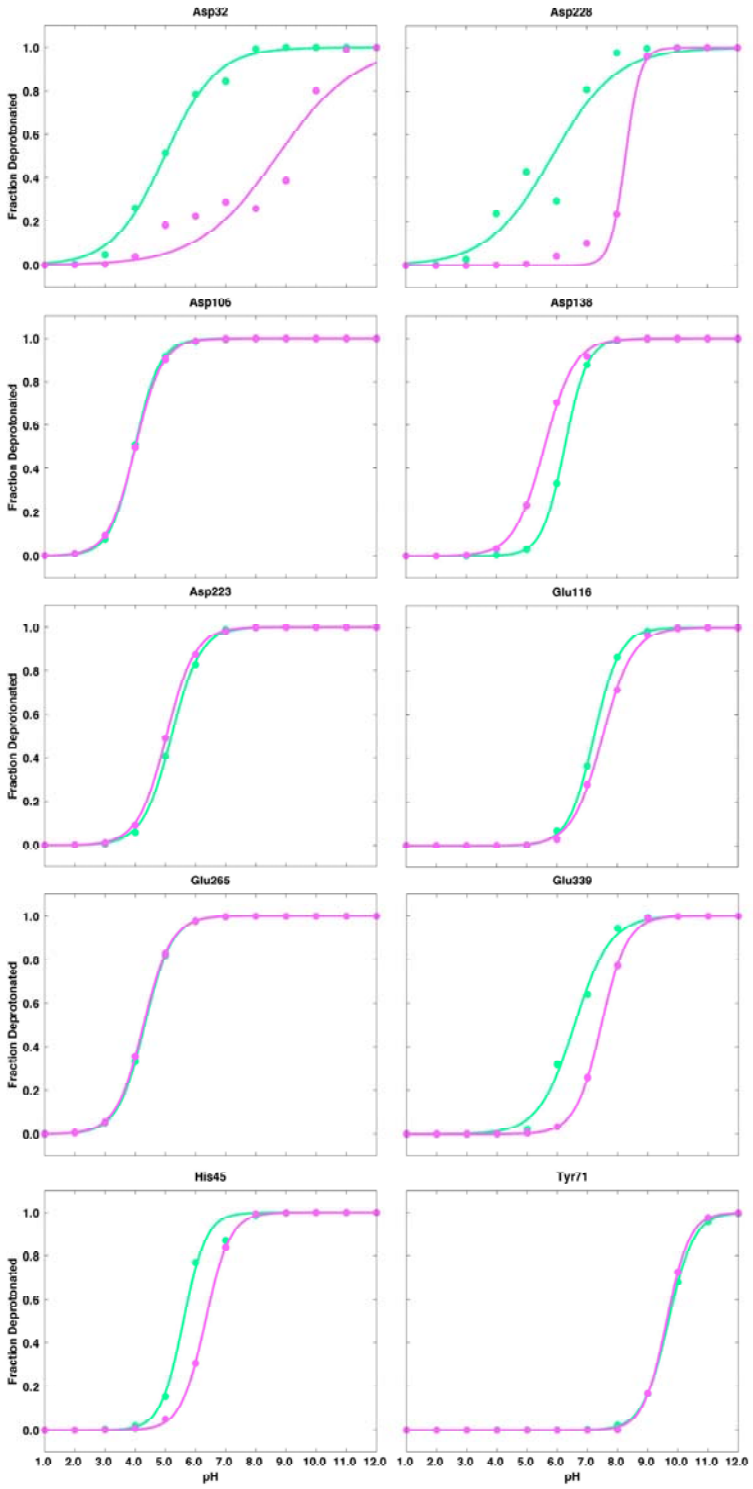


Figure 5-S5. Titration curves from the pH-REMD simulations of the 2P4J system, shown in purple. Respective titration curves of the dyad in apo BACE-1 are shown in green.

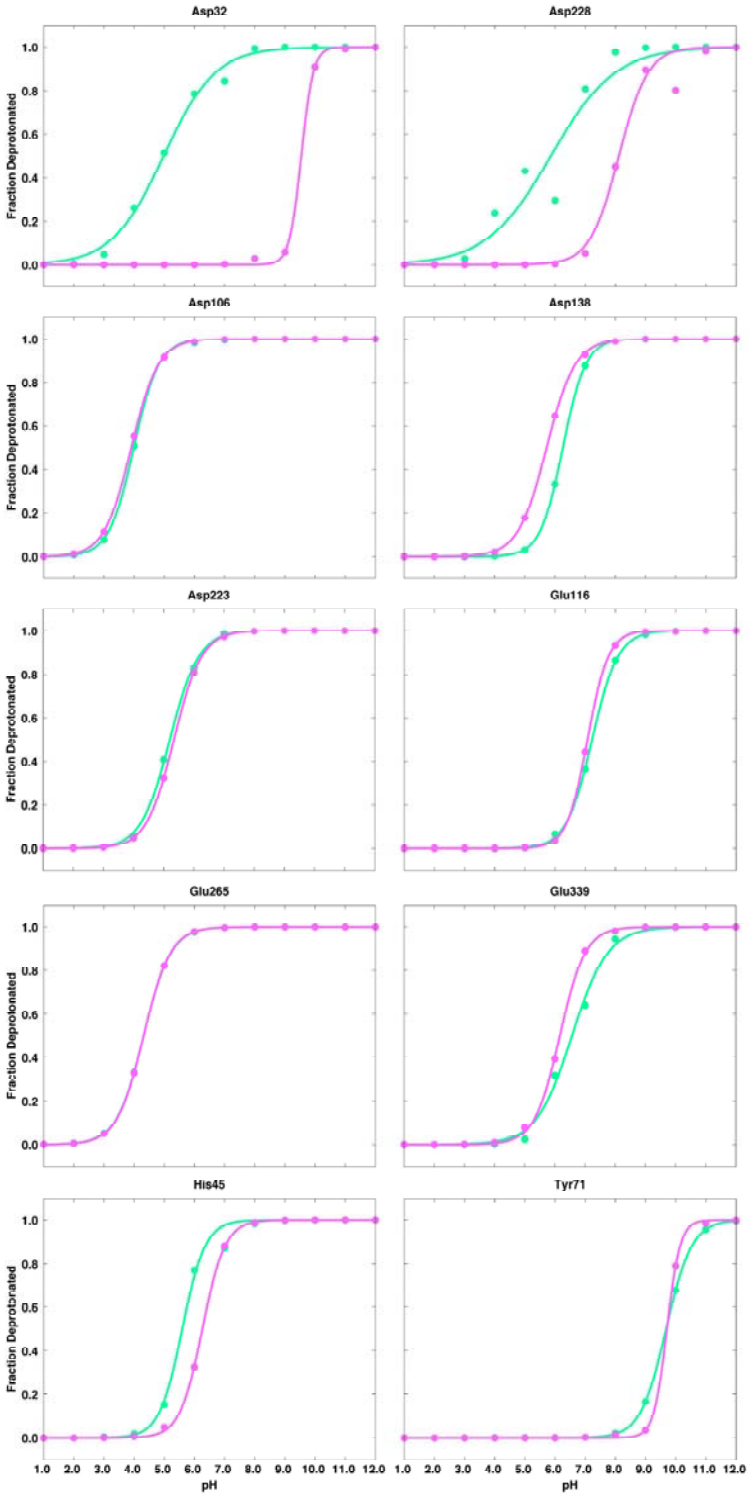


Figure 5-S6. Titration curves from the pH-REMD simulations of the 2G94 system, shown in purple. Respective titration curves of the dyad in apo BACE-1 are shown in green.

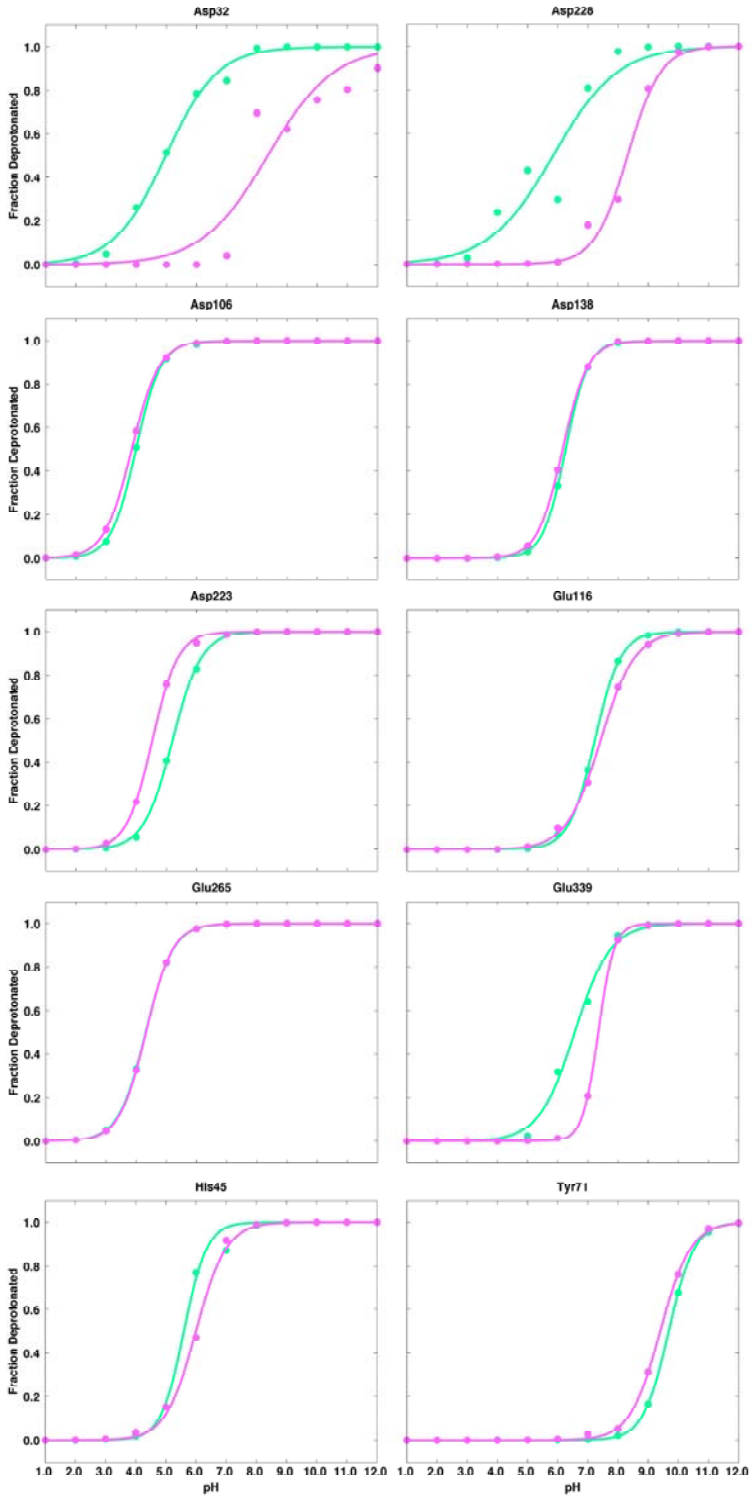


Figure 5-S7. Titration curves from the pH-REMD simulations of the 2IRZ system, shown in purple. Respective titration curves of the dyad in apo BACE-1 are shown in green.

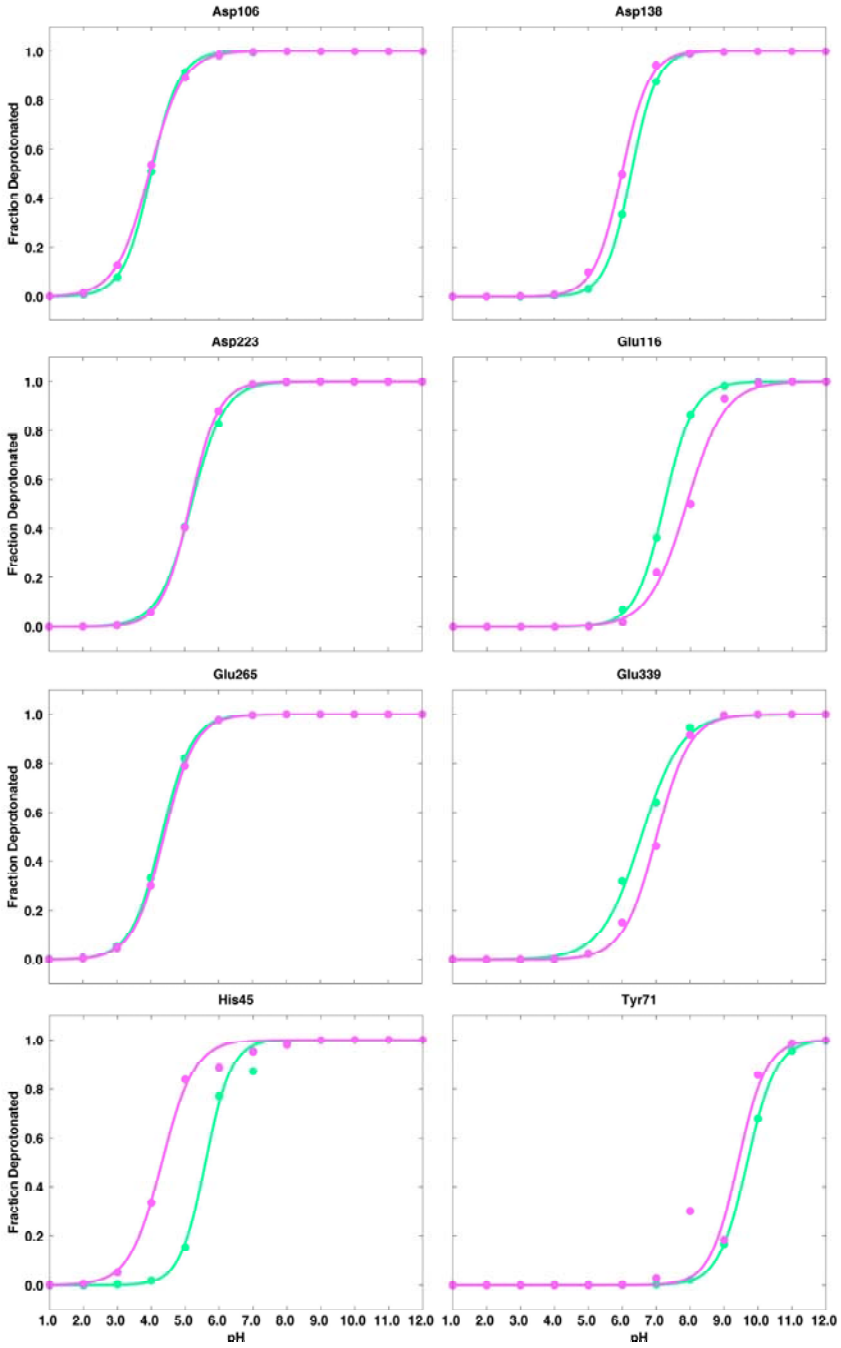


Figure 5-S8. Titration curves from the pH-REMD simulations of the 2B8L system, shown in purple. Respective titration curves of the dyad in apo BACE-1 are shown in green.

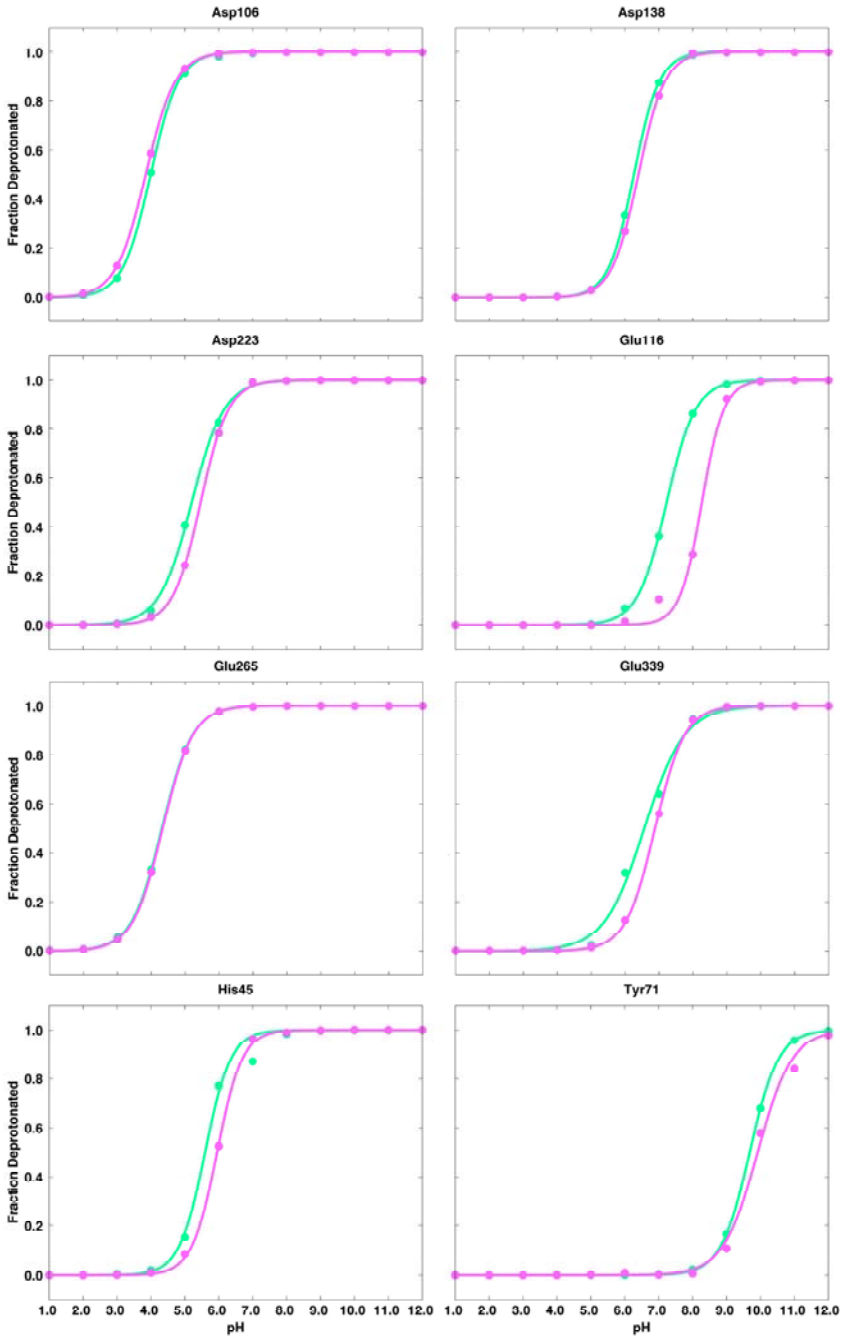


Figure 5-S9. Titration curves from the pH-REMD simulations of the 2FDP system, shown in purple. Respective titration curves of the dyad in apo BACE-1 are shown in green.

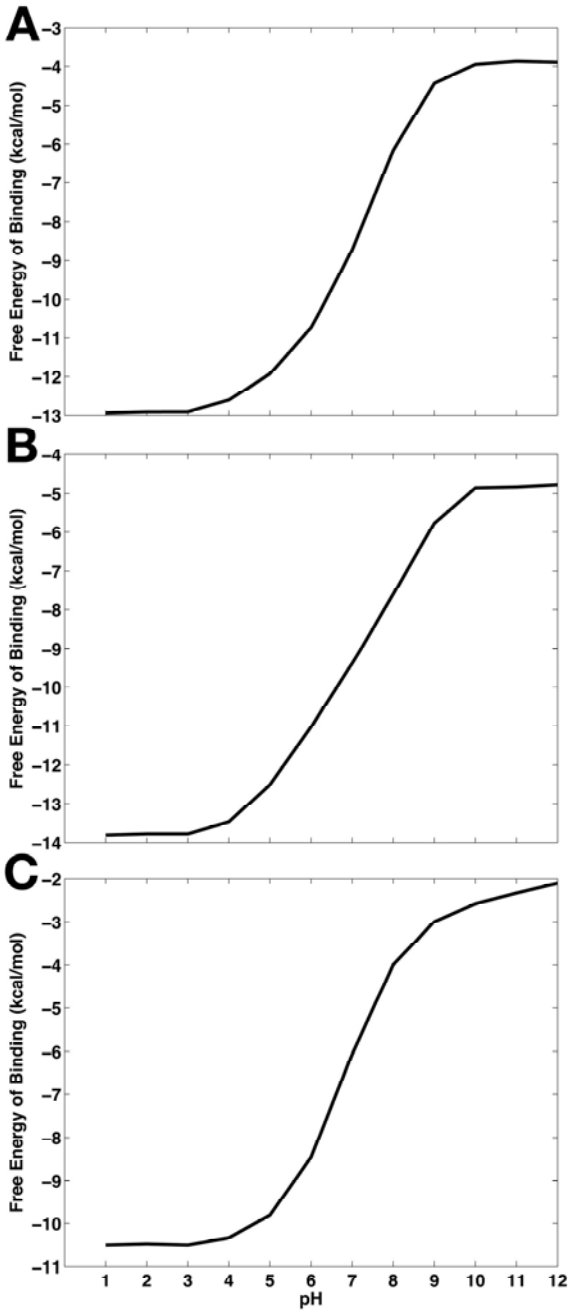


Figure 5-S10. Binding free energy profiles of inhibitors as functions of pH. (A) 2P4J. (B) 2G94. (C) 2IRZ.

References

1. Mattson, M. P. Pathways Towards and Away from Alzheimer's Disease. *Nature* **2004,** *430*, 631-639.

2. Haass, C.; Selkoe, D. J. Soluble Protein Oligomers in Neurodegeneration: Lessons from the Alzheimer's Amyloid
B-Peptide. *Nat. Rev. Mol. Cell Biol.* **2007,** *8*, 101-112.

3. Vassar, R.; Kovacs, D. M.; Yan, R.; Wong, P. C. The B-Secretase Enzyme BACE in Health and Alzheimer's Disease: Regulation, Cell Biology, Function, and Therapeutic Potential. *J. Neurosci.* **2009,** *41*, 12787-12794.

4. Ghosh, A. K.; Osswald, H. L. BACE1 (B-Secretase) Inhibitors for the Treatment of Alzheimer's Disease. *Chem. Soc. Rev.* **2014,** *43*, 6765-6813.

5. Cleary, J. P.; Walsh, D. M.; Hofmeister, J. J.; Shankar, G. M.; Kuskowski, M. A.; Selkoe, D. J.; Ashe, K. H. Natural Oligomers of the Amyloid-Beta Protein Specifically Disrupt Cognitive Function. *Nat. Neurosci.* **2005,** *8*, 79-84.

6. Selkoe, D. J. The Molecular Pathology of Alzheimer's Disease. *Neuron* **1991,** *6*, 487-498.

7. Selkoe, D. J. Translating Cell Biology into Therapeutic Advances in Alzheimer's Disease. *Nature* **1999,** *399*, A23-A31.

8. Selkoe, D. J. The Origins of Alzheimer's Disease. *JAMA* **2000,** *283*, 1615-1617.

9. Glenner, G. G.; Wong, C. W. Alzheimer's Disease: Initial Report of He Purification and Characterization of a Novel Cerebrovascular Amyloid Protein. *Biochem. Biophys. Res. Commun.* **1984,** *120*, 885-890.

10. Younkin, S. G. The Role of Aβ42 in Alzheimer's Disease. *J. Physiol. Paris* **1998,** *92*, 289-292.

11. Hardy, J. A.; Higgins, G. A. Alzheimer's Disease: The Amyloid Cascade Hypothesis. *Science* **1992,** *256*, 184-185.

12. Hardy, J. A.; Selkoe, D. J. The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics. *Science* **2002,** *297*, 353-356.

13. Vassar, R.; Bennett, B. D.; Babu-Khan, S.; Kahn, S.; Mendiaz, E. A.; Denis, P.; Teplow, D. B.; Ross, S.; Amarante, P.; Leoloff, R., et al. B-Secretase Cleavage of Alzheimer's Amyloid Precursor Protein by the Transmembrane Aspartic Protease BACE. *Science* **1999,** *286*, 735-741.

14. Haass, C. Take Five-BACE and the Γ-Secretase Quartet Conduct Alzheimer's Amyloid B-Peptide Generation. *EMBO J.* **2004,** *23*, 483-488.

15. Cole, S. L.; Vassar, R. The Alzheimer's Disease B-Secretase Enzyme, BACE1. *Mol. Neurodegener.* **2007,** *2*, 22.

16. Hong, L.; Koelsch, G.; Lin, X.; Wu, S.; Terzyan, S.; Ghosh, A. K.; Zhang, X. C.; Tang, J. Structure of the Protease Domain of Memapsin 2(B-Secretase) Complexed with Inhibitor. *Science* **2000,** *290*, 150-153.

17. Ghosh, A. K.; Shin, D.; Downs, D.; Koelsch, G.; Lin, X.; Ermolieff, J.; Tang, J. Design of Potent Inhibitors for Human Bran Memapsin 2 (B-Secretase). *J. Am. Chem. Soc.* **2000,** *122*, 3522-3523.

18. Tang, J.; James, M. N. G.; Hsu, I. N.; Jenkins, J. A.; Blundell, T. L. Structural Evidence for Gene Duplication in the Evolution of the Acid Proeases. *Nature* **1978,** *271*, (618-621).

19. Suguna, K.; Padlan, E. A.; Smith, C. W.; Carlson, W. D.; Davies, D. R. Binding of a Reduced Peptide Inhibitor to the Aspartic Proeinase from Rhizopus Chinensis: Implications for a Mechanism of Action. *Proc. Natl. Acad. Sci. U.S.A.* **1987,** *84*, 7009-7013.

20. Dunn, B. M. Structure and Mechanism of the Pepsin-Like Family of Aspartic Peptidases. *Chem. Rev.* **2002,** *102*, 4413-4458.

21. Cascella, M.; Micheletti, C.; Rothlisberger, U.; Carloni, P. Evolutionarily Conserved Functional Mechanics across Pepsin-Like and Retroviral Aspartic Proteases. *J. Am. Chem. Soc.* **2005,** *127*, 3734-3742.

22. Hunt, C. E.; Turner, A. J. Cell Biology, Regulation and Inhibition of B-Secretase (BACE-1). *FEBS J.* **2009,** *276*, 1845-1859.

23. Prasad, B. V.; Suguna, K. Role of Water Molecules in the Structure and Function of Aspartic Proteinases. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002,** *58*, 250-259.

24. Lin, X.; Koelsch, G.; Wu, S.; Downs, D.; Dashti, A.; Tang, J. Human Aspartic Protease Memapsin 2 Cleaves the B-Secretase Site of B-Amyloid Precursor Protein. *Proc. Natl. Acad. Sci. U.S.A.* **2000,** *97*, 1456-1460.

25. Stachel, S. J.; Coburn, C. A.; Rush, D.; Jones, K. L. G.; Zhu, H.; Rajapakse, H. A.; Graham, S. L.; Simon, A. J.; Holloway, M. K.; Allison, T. J., et al. Discovery of Aminoheterocycles as a Novel B-Secretase Inhibitor Class: pH Dependence on Binding Activity Part 1. *Bioorg. Med. Chem. Lett.* **2009,** *19*, 2977-2980.

26. Blundell, T. L.; Jenkins, J. A.; Sewell, B. T.; Pearl, L. H.; Cooper, J. B.; Tickle, I. J.; Veerapandian, B.; Wood, S. P. X-Ray Analyses of Aspartic Proteinases: The Three-Dimensional Structure at 2.1 Å Resolution of Endothiapepsin. *J. Mol. Biol.* **1990,** *20*, 919-941.

27. Sielecki, A. R.; Fedorov, A. A.; Boodhoo, A.; Andreeva, N. S.; James, M. N. G. Molecular and Crystal Structures of Monoclinic Porcine Pepsin Refined at 1.8 Å Resolution. *J. Mol. Biol.* **1990,** *214*, 143-170.

28. Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Hosur, M. V.; Sowder, R. C., II.; Cachau, R. E.; Collins, J.; Silva, A. M.; Erickson, J. W. Crystal Structures of Native and Inhibited Forms of Human Cathepsin D: Implications for Lysosomal Targeting and Drug Design. *Proc. Natl. Acad. Sci. U.S.A.* **1993,** *90*, 6796-6800.

29. Hong, L.; Tang, J. Flap Position of Free Memapsin 2 (B-Secretase), a Model for Flap Opening in Aspartic Protease Catalysis. *Biochemistry* **2004,** *43*, 4689-4695.

30. Tang, J.; Koelsch, G. A Possible Function of a Flap of Aspartic Proteases: The Capture of Substrate Side Chains Determines the Specificity of Cleavage Positions. *Peptide Lett.* **1995,** *2*, 257-266.

31. Andreeva, N. S.; Rumsh, L. D. Analysis of Crystal Structures of Aspartic Proteinases: On the Role of Amino Acid Residues Adjacent to the Catalytic Site of Pepsin-Like Enzymes. *Protein Sci.* **2001,** *10*, 2439-2450.

32. Patel, S.; Vuillard, L.; Cleasby, A.; Murray, C. W.; Yon, J. Apo and Inhibitor Complex Structures of BACE (B-Secretase). *J. Mol. Biol.* **2004,** *343*, 407-416.

33. Spronk, S. A.; Carlson, H. A. The Role of Tyrosine 71 in Modulating the Flap Conformations of BACE1. *Proteins* **2011,** *79*, 2247-2259.

34. Gorfe, A. A.; Caflisch, A. Functional Plasticity in the Substrate Binding Site of B-Secretase. *Structure* **2005,** *13*, 1487-1498.

35. Park, H.; Lee, S. Determination of the Active Site Protonation State of B-Secretase from Molecular Dynamics Simulation and Docking Experiment: Implications for Structure-Based Inhibitor Design. *J. Am. Chem. Soc.* **2003,** *125*, 16416-16422.

36. Rajamani, R.; Reynolds, C. H. Modeling the Protonation States of the Catalytic Aspartates in B-Secretase. *J. Med. Chem.* **2004,** *47*, 5159-5166.

37. Yu, N.; Hayik, S. A.; Wang, B.; Liao, N.; Reynolds, C. H.; Merz, K. M., Jr. Assigning the Protonation States of Key Aspartates in B-Secretase Using QM/MM X-Ray Structure Refinement. *J. Chem. Theory Comput.* **2006,** *2*, 1057-1069.

38. Barman, A.; Prabhakar, R. Protonation States of the Catalytic Dyad of B-Secretace (BACE1) in the Presence of Chemically Diverse Inhibitors: A Molecular Docking Study. *J. Chem. Inf. Model.* **2012,** *52*, 1275-1287.

39. Domínguez, J. L.; Christopeit, T.; Vilaverde, M. C.; Gossas, T.; Otero, J. M.; Nyström, S.; Baraznenok, V.; Lindström, E.; Danielson, U. H.; Sussman, F. Effect of the Protonation State of the Titratable Residues on the Inhibitor Affinity to BACE-1. *Biochemistry* **2010,** *49*, 7255-7263.

40. Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. The Determinants of pKas in Proteins. *Biochemistry* **1996,** *35*, 7819-7833.

41. Onufriev, A. V.; Case, D. A.; Ullmann, G. M. A Novel View of pH Titration in Biomolecules. *Biochemistry* **2001,** *40*, 3413-3419.

42. Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; Velu, P. D.; García-Moreno, B. Charges in the Hydrophobic Interior of Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2010,** *107*, 16096-16100.

43. Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; García-Moreno, B. Large Shifts in pKa Values of Lysine Residues Buried inside a Protein. *Proc. Natl. Acad. Sci. U.S.A.* **2011,** *108*, 5260-5265.

44. Alexov, E.; Mehler, E. L.; Baker, N.; Baptista, A. M.; Huang, Y.; Milletti, F.; Nielsen, J. E.; Farrell, D.; Carstensen, T.; Olsson, M. H. M., et al. Progress in the Prediction of pKa Values in Proteins. *Proteins* **2011,** *79*, 3260-3275.

45. Nielsen, J. E.; Gunner, M. R.; García-Moreno, B. The pKa Cooperative: A Collaborative Effort to Advance Structure-Based Calculations of pKa Values and Electrostatic Effects in Proteins. *Proteins* **2011,** *79*, 3249-3259.

46. García-Moreno, B. Adaptations of Proteins to Cellular and Subcellular pH. *J. Biol.* **2009,** *8*, 98-101.

47. Baptista, A. M.; Martel, P. G.; Petersen, S. B. Simulation of Protein Conformational Freedom as a Function of pH: Constant-pH Molecular Dynamics Using Implicit Titration. *Proteins* **1997,** *27*, 523-544.

48. Alexov, E.; Gunner, M. R. Incorporating Protein Conformational Flexibility into pH Titration Calculations: Results on T4 Lysozyme. *Biophys. J.* **1997,** *74*, 2075-2093.

49. Cassidy, C. S.; Lin, J.; Frey, P. A. A New Concpet for the Mechanism of Action of Chymotrypsin: The Role of the Low-Barrier Hydrogen Bond. *Biochemistry* **1997,** *36*, 4576-4584.

50. Alexov, E. Calculating Proton Uptake/Release and Binding Free Energy Taking into Account Ionization and Conformation Changes Induced by Protein-Inhibitor Association Application to Plasmepsin, Cathepsin D and Endothiapepsin-Pepstatin Complexes. *Proteins* **2004,** *56*, 572-584.

51. Luo, R.; Head, M. S.; Moult, J.; Gilson, M. K. pKa Shifts in Small Molecules and HIV Protease: Electrostatics and Conformation. *J. Am. Chem. Soc.* **1998,** *120*, 6138-6146.

52. Trylska, J.; Antosiewicz, J.; Geller, M.; Hodge, C. N.; Klabe, R. M.; Head, M. S.; Gilson, M. K. Thermodynamic Linkage between the Binding of Protons and Inhibitors to HIV-1 Protease. *Prot. Sci.* **1999,** *8*, 180-195.

53. Sakurai, K.; Goto, Y. Principal Component Analysis of the pH-Dependent Conformational Transitions of Bovine Beta-Lactoglobulin Monitored by Heteronuclear NMR. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *104*, 15346-15351.

54. Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; García-Moreno, B. High Apparent Dielectric Constant inside a Protein Reflects Structural Reorganization Coupled to the Ionization of an Internal Asp. *Biophys. J.* **2007,** *92*, 2041-2053.

55. Søndergard, C. R.; McIntosh, L. P.; Pollastri, G.; Nielsen, J. E. Determination of Electrostatic Interaction Energies and Protonation State Populations En Enzyme Active Sites. *J. Mol. Biol.* **2008,** *376*, 269-287.

56. Hass, M. A.; Jensen, M. R.; Led, J. J. Probing Electric Fields in Proteins in Solution by NMR Spectroscopy. *Proteins* **2008,** *72*, 333-343.

57. Karp, D. A.; Stahley, M. R.; García-Moreno, B. Conformational Consequences of Ionization of Lys, Asp, and Glu Buried at Position 66 in Staphylococcal Nuclease. *Biochemistry* **2010,** *49*, 4138-4146.

58. Di Russo, N. V.; Estrin, D. A.; Marti, M. A.; Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Comput. Biol.* **2012,** *8*, e1002761.

59. Onufriev, A. V.; Alexov, E. Protonation and pK Changes in Protein-Ligand Binding. *Q. Rev. Biophys.* **2013,** *46*, 181-209.

60. Aguilar, B.; Anandakrishnan, R.; Ruscio, J. Z.; Onufriev, A. V. Statistics and Physical Origins of pK and Ionization State Changes Upon Protein-Ligand Binding. *Biophys. J.* **2010,** *98*, 872-880.

61. Stranzl, G. R.; Gruber, K.; Steinkellner, G.; Zangger, K.; Schwab, H.; Kratky, C. Observation of a Short, Strong Hydrogen Bond in the Active Site of Hydroxynitrile Lyase from Hevea Brasiliensis Explain a Large pKa Shift of the Catalytic Base Induced by the Reaction Intermediate. *J. Biol. Chem.* **2004,** *279*, 2699-3707.

62. Szakács, Z.; Béni, S.; Varga, Z.; Örfi, L.; Kéri, G.; Noszál, B. Acid-Base Profiling of Imatinib (Gleevec) and Its Fragments. *J. Med. Chem.* **2005,** *48*, 249-255.

63. Brandsdal, B. O.; Smalås, A. O.; Åqvist, J. Free Energy Calculations Show That Acidic P1 Variants Undergo Large pKa Shifts Upon Biniding to Trypsin. *Proteins* **2006,** *64*, 740-748.

64. Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Protonation Changes Upon Ligand Binding to Trypsin and Thrombin: Structural Interpretation Based on pKa Calculations and Itc Experiments. *J. Mol. Biol.* **2007,** *375*, 1347-1356.

65. Ragona, L.; Fogolari, F.; Catalano, M.; Ugolini, R.; Zetta, L.; Molinari, H. Ef Loop Conformational Change Triggers Ligand Binding in B-Lactoglobulins. *J. Biol. Chem.* **2003,** *278*, 38840-38846.

66. Wang, L.; Witham, S.; Zhang, Z.; Li, L.; Hodsdon, M.; Alexov, E. In Silico Investigation of pH-Depencdence of Prolactin and Human Growth Hormone Binding to Humen Prolactin Receptor. *Commun. Comput. Phys.* **2013,** *13*, 207-222.

67. Warshel, A.; Russel, S. Calculations of Electrostatic Interactions in Biological Systems and in Solutions. *Q. Rev. Biophys.* **1984,** *17*, 283-422.

68. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990,** *112*, 6127-6129.

69. Sham, Y. Y.; Chu, Z. T.; Warshel, A. Consistent Calculations of pKa's of Ionizable Residues in Proteins: Semi-Microscopis and Microscopic Approaches. *J. Phys. Chem. B* **1997,** *101*, 4458-4472.

70. Georgescu, R. E.; Alexov, E.; Gunner, M. R. Combining Conformational Flexibility and Continuum Electrostatics for Calculating pKa's in Proteins. *Biophys. J.* **2002,** *83*, 1731-1748.

71. Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Interpretaion of Protein pKa Values. *Proteins* **2005,** *61*, 704-721.

72. Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and Rationalization of Protein pKa Values Using QM and QM/MM Methods. *J. Phys. Chem. A* **2005,** *109*, 6634-6643.

73. Borjesson, U.; Hunenberger, P. H. Explicit-Solvent Molecular Dynamics Simulation at Constant pH: Methodology and Application to Small Amines. *J. Chem. Phys.* **2001,** *114*, 9706-9719.

74. Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH Molecular Dynamics Using Stochastic Titration. *J. Chem. Phys.* **2002,** *117*, 4184-4200.

75. Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. Simulating Proteins at Constant pH: An Approach Combining Molecular Dynamics and Monte Carlo Simulation. *Proteins* **2002,** *47*, 469-480.

76. Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. Constant-pH Molecular Dynamics Using Continuous Titration Coordinates. *Proteins* **2004,** *56*, 738-752.

77. Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004,** *25*, 2038-2048.

78. Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. Coupling Constant pH Molecular Dynamics with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2010,** *6*, 560-568.

79. Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput.* **2014,** *10*, 1341-1352.

80. Goh, G. B.; Knight, J. L.; Brooks, C. L., III. Constant pH Molecular Dynamics Simulations of Nucleic Acids in Explicit Solvent. *J. Chem. Theory Comput.* **2012,** *8*, 36-46.

81. Goh, G. B.; Knight, J. L.; Brooks, C. L., III. pH-Dependent Dynamics of Complex RNA Macromolecules. *J. Chem. Theory Comput.* **2013,** *9*, 935-943.

82. Goh, G. B.; Knight, J. L.; Brooks, C. L., III. Towards Accurate Prediction of Protonation Equilibrium of Nucleic Acids. *J. Phys. Chem. Lett.* **2013,** *4*, 760-766.

83. Dissanayake, T.; Swails, J. M.; Harris, M. E.; Roitberg, A. E.; York, D. M. Interpretation of pH-Activity Profiles for Acid-Base Catalysis from Molecular Simulations. *Biochemistry* **2015**.

84. Laricheva, E. N.; Arora, K.; Knight, J. L.; Brooks, C. L., III. Deconstructing Activation Events in Rhodopsin. *J. Am. Chem. Soc.* **2013,** *135*, 10906-10909.

85. Kim, M. O.; Blachly, P. G.; Kaus, J. W.; McCammon, J. A. Protocols Utilizing Constant pH Molecular Dynamics to Compute pH-Dependent Binding Free Energies. *J. Phys. Chem. B* **2015,** *119*, 861-872.

86. Wyman, J. Heme Proteins. *Adv. Protein Chem.* **1948,** *4*, 407-531.

87. Tanford, C. Protein Denaturation. Part C. Theoretical Models for the Mechanism of Denaturation. *Adv. Protein Chem.* **1970,** *24*, 1-95.

88. Swails, J. M.; Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J. Chem. Theory Comput.* **2012,** *8*, 4393-4404.

89. Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E. I.; Darden, T. A.; Duke, R. E.; Gohlke, H., et al. *Amber 14*, University of California, San Francisco: 2014.

90. Itoh, S. G.; Damjanović, A.; Brooks, B. R. pH Replica-Exchange Method Based on Discrete Protonation States. *Proteins* **2011,** *79*, 3420-3436.

91. Mason, A. C.; Jensen, J. H. Protein-Protein Binding Is Often Associated with Changes in Protonation State. *Proteins* **2008,** *71*, 81-91.

92. Wyman, J. The Binding Potential, a Neglected Linkage Concept. *J. Mol. Biol.* **1965,** *11*, 631-644.

93. Stachel, S. J.; Coburn, C. A.; Steele, T. G.; Crouthamel, M.-C.; Pietrak, B. L.; Lai, M.-T.; Holloway, M. K.; Munshi, S. K.; Graham, S. L.; Vacca, J. P. Conformationally Biased P3 Amide Replacements of B-Secretase Inhibitors. *Bioorg. Med. Chem. Lett.* **2006,** *16*, 641-644.

94. Ghosh, A. K.; Bilcer, G.; Harwood, C.; Kawahama, R.; Shin, D.; Hussain, K. A.; Hong, L.; Loy, J. A.; Nguyen, C.; Koelsch, G., et al. Structure-Based Design: Potent Inhibitors of Human Brain Memapsin 2 (B-Secretase). *J. Med. Chem.* **2001,** *44*, 2865-2868.

95. Ghosh, A. K.; Kumaragurubaran, N.; Hong, L.; Lei, H.; Hussain, K. A.; Liu, C. F.; Devasamudram, T.; Weerasena, V.; Turner, R.; Koelsch, G., et al. Design, Synthesis and X-Ray Structure of Protein-Ligand Complexes: Important Insight into Selectivity of Memapsin 2 (B-Secretase) Inhibitors. *J. Am. Chem. Soc.* **2006,** *128*, 5310-5311.

96. Rajapakse, H. A.; Nantermet, P. G.; Selnick, H. G.; Munshi, S. K.; McGaughey, G. B.; Lindsley, S. R.; Young, M. B.; Lai, M.-T.; Espeseth, A. S.; Shi, X.-P., et al. Discovery of Oxadiazoyl Tertiary Carbinamine Inhibitors of B-Secretase (BACE-1). *J. Med. Chem.* **2006,** *49*, 7270-7273.

97. Yang, W.; Lu, W.; Lu, Y.; Zhong, M.; Sun, J.; Thomas, A. E.; Wilkinson, J. M.; Fucini, R. V.; Lam, M.; Randal, M., et al. Aminoethylenes: A Tetrahedral Intermediate Isostere Yielding Potent Inhibitors of the Aspartyl Protease BACE-1. *J. Med. Chem.* **2006,** *49*, 839-842.

98. *Prime*, Version 3.6; Schrödinger, LLC: New York, NY, 2014.

99. Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of Crystal Packing Forces in Determining Protein Sidechain Conformations. *J. Mol. Biol.* **2002,** *320*, 597-608.

100. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins* **2004,** *55*, 351-367.

101. Consortium, T. U. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015,** *43*, D204-D212.

102. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Logal Alignment Search Tool. *J. Mol. Biol.* **1990,** *215*, 403-410.

103. Vosko, S. H.; Nusair, L. W. M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980,** *58*, 1200-1211.

104. Lee, C.; Yang, W.; Parr, R. G. Development of the Cole-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988,** *37*, 785-789.

105. Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.

106. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994,** *98*, 11623-11627.

107. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A., et al. *Gaussian 09*, Revision D.01; Gaussian, Inc.: Wallingford, CT, 2009.

108. Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990,** *11*, 431-439.

109. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006,** *25*, 247-260.

110. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004,** *25*, 1157-1174.

111. Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein-Ligand Complexes. *Proteins* **2008,** *73*, 765-783.

112. Olsson, M. H. M.; Søndergard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Inthernal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011,** *7*, 525-537.

113. Søndergard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Realization of pKa Values. *J. Chem. Theory Comput.* **2011,** *7*, 2284-2295.

114. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** *79*, 926-935.

115. Schafmeister, C. E. A. F.; Ross, W. S.; Romanovski, V. *Leap*, University of california, San Francisco: 1995.

116. Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular-Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* **2004,** *120*, 6363-6374.

117. Zhang, Y.; Feller, S. E.; Brooks, B. R.; Pastor, R. W. Computer Simulation of Liquid/Liquid Interfaces. I. Theory and Application to Octane/Water. *J. Chem. Phys.* **1995,** *103*, 10252-10266.

118. Darden, T. A.; York, D. M.; Pedersen, L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993,** *98*, 10089-10092.

119. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T. A.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995,** *103*, 8577-8593.

120. Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations Wifh Amber on Gpus. 1. Generalized Born. *J. Chem. Theory Comput.* **2012,** *8*, 1542-1555.

121. Nguyen, H.; Roe, D. R.; Simmerling, C. L. Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J. Chem. Theory Comput.* **2013,** *9*, 2020-2034.

122. Ryckaret, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.

123. Shao, J. Y.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. I. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007,** *3*, 2312-2334.

124. *Matlab*, Version 8.1; The MathWorks, Inc.: Natick, MA, 2013.

125. Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. Electrostatic Calculations of Side-Chain pKa Values in Myoglobin and Comparison with NMR Data for Histidines. *Biochemistry* **1993,** *32*, 8045-8056.

126. Kukic, P.; Farrell, D.; McIntosh, L. P.; García-Moreno, B.; Jensen, K. S.; Toleikis, Z.; Teilum, K.; Nielsen, J. E. Protein Dielectric Constants Determined from NMR Chemical Shift Perturbations. *J. Am. Chem. Soc.* **2013,** *135*, 16968-16976.

127. Nielsen, J. E.; McCammon, J. A. Calculating pKa Values in Enzyme Active Sites. *Protein Sci.* **2003,** *12*, 1894-1901.

128. Ghosh, A. K.; Kumaragurubaran, N.; Hong, L.; Kulkarni, S. S.; Xu, X.; Chang, W.; Weerasena, V.; Turner, R.; Koelsch, G.; Bilcer, G., et al. Design, Synthesis, and X-Ray Structure of Potent Memapsin 2 (B-Secretase) Inhibitors with Isophthalamide Derivatives as the P2-P3-Ligands. *J. Med. Chem.* **2007,** *50*, 2399-2407.

129. McGaughey, G. B.; Colussi, D.; Graham, S. L.; Lai, M.-T.; Munshi, S. K.; Nantermet, P. G.; Pietrak, B. L.; Rajapakse, H. A.; Seinick, H. G.; Stauffer, S. R., et al. B-Secretase (BACE-1) Inhibitors: Accounting for 10s Loop Flexibility Using Rigid Active Sites. *Bioorg. Med. Chem. Lett.* **2007,** *17*, 1117-1121.

130. Onufriev, A. V.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000,** *104*, 3712-3720.

131. Spassov, V. Z.; Yan, L. A Fast and Accurate Computational Approach to Protein Ionization. *Prot. Sci.* **2009,** *17*, 1955-1970.

132. Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The Sampl2 Bind Prediction Challenge: Introduction and Overview. *J. Comput. Aided Mol. Des.* **2010,** *24*, 259-279.

133. Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014,** *10*, 2632-2647.