UNIVERSITY OF CALIFORNIA,
IRVINE

Computational Characterization of Hv1 Proton Channel Inhibitors and Investigation of the Tools in Molecular Simulations

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Victoria T. Lim

Dissertation Committee:
Professor David L. Mobley, Co-Chair
Professor Douglas J. Tobias, Co-Chair
Professor Rachel W. Martin

2020

DEDICATION

I dedicate this dissertation to all the teachers—most importantly my parents, Tran and Lieh Lim—who have guided and inspired me throughout my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I thank my advisors David Mobley and Doug Tobias for their support and guidance over the last five years. Doug's scientific vision and astute research feedback has led me to become a more thoughtful scientist with particular appreciation for simple but valuable "back of the envelope" calculations. David's genuine care about the work that he does and the people that he mentors is remarkable in its rarity. From working with David, I've been able to pursue and develop my scientific curiosities, improve my technical writing, and seek answers to the greater questions of life.

I am also grateful to many mentors who shared their expertise and advice on molecular simulations and the complex life of research. Specifically, I acknowledge Alfredo Freites, Christopher Bayly (at OpenEye Scientific Software), Andrew Geragotelis, Nathan Lim, Caitlin Bannan, and Eric Wong. These people were integral to my success.

On a more personal note, I appreciate the research discussions and the camaraderie of the Mobley group. Beyond the science itself, my experiences in graduate school were immeasurably enriched by long chats with Caitlin, lunches with Hannah Baumann and Léa El Khoury, and impromptu breaks with Sukanya Sasmal. I enjoyed the company of the Mobsters as a whole on numerous dinner events, outdoor trips, museum ventures, and long days in the lab.

While it is less common for PhD students to partake in research outside of their lab, I had the valuable experience to spend a summer in Stockholm, Sweden, with Erik Lindahl's group. There, I dove into a new project outside my main area of expertise, and successfully developed a workflow to automate cryo-EM model fitting with molecular dynamics. Efforts to publish this work are currently underway. On this, I value the time and guidance of Erik Lindahl, Christian Blau, Reba Howard, Urška Rovšnik, Yuxuan Zhang, and Joe Jordan.

Through various initiatives and programs at UC Irvine, I've had many opportunities for personal and professional development. I credit Bri McWhorter and Sandra Tsing Loh for advancing my scientific communication skills. I applied these skills in my positions as Communications Consultant for the UCI Graduate Resource Center and as the lead managing editor (2019–2020) for the Loh Down on Science. Both of these were incredibly rewarding experiences. I also particularly appreciate the Invention Transfer Group Fellowship with UCI Applied Innovation, where I learned to evaluate scientific data from the standpoint of intellectual property protection, marketing, and licensing. Stephen Ro and Lisa Stutts were amazing mentors who always encouraged me to push forward, often realizing my potential even before I did.

Finally, I humbly acknowledge financial support from the National Science Foundation for both the Graduate Research Fellowship (DGE-1321846) and the Graduate Research Opportunities Worldwide fellowship. I further acknowledge funding from the National Institutes of Health (GM098973) and the Swedish Research Council. My funding sources allowed me the flexibility and opportunities to pursue a variety of projects, contributing to a graduate research experience far more fulfilling than I could have ever planned.

Chapter 5 of this dissertation is reprinted with permission from the American Chemical Society regarding the published work: V. T. Lim, C. I. Bayly, L. Fusti-Molnar, and D. L. Mobley, "Assessing the Conformational Equilibrium of Carboxylic Acid via Quantum Mechanical and Molecular Dynamics Studies on Acetic Acid," J. Chem. Inf. Model., 59(5): 1957-1964, 2019. I thank Chris Bayly, Laszlo Fusti-Molnar, and David Mobley for their collaborative efforts.

VITA

Victoria T. Lim

EDUCATION

Doctor of Philosophy in Chemistry — **2020**
University of California, Irvine — *Irvine, CA*

Bachelor of Science in Chemistry;
Bachelor of Science in Mathematics — **2015**
Belmont University — *Nashville, TN*

PROFESSIONAL EXPERIENCE

NSF Graduate Research Fellow — **2015–2020**
Mobley and Tobias Labs; University of California, Irvine — *Irvine, CA*

Technology Transfer Fellow — **2018–2020**
Beall Applied Innovation; University of California, Irvine — *Irvine, CA*

Visiting Scientist — **Summer 2019**
Lindahl Lab; Stockholm University — *Stockholm, Sweden*

Computational Chemistry Intern — **Summer 2015**
Advised by Alejandro Crespo; Merck & Co. — *Rahway, NJ*

Research Assistant – Synthetic Organic Chemistry — **2013–2015**
Advised by Jeffrey Johnston; Vanderbilt University — *Nashville, TN*

REU Scholar – Computational Quantum Chemistry — **Summer 2014**
Advised by Wesley Allen, University of Georgia — *Athens, GA*

Product Validation Intern — **Summer 2013**
Croda, Inc. — *Edison, NJ*

COMMUNICATION EXPERIENCE

Lead Managing Editor — **2019–2020**
The Loh Down on Science — *Irvine, CA*

Writer — **2018–2019**
The Loh Down on Science — *Irvine, CA*

Communications Consultant — **2017–2019**
Graduate Resource Center; University of California, Irvine — *Irvine, CA*

TEACHING AND MENTORING

Research Mentor **2017–2019**
Jordan Ehrman; Mobley Lab; UCI *Irvine, CA*

Research Mentor **2016–2017**
Daisy Kyu; Mobley Lab; UCI *Irvine, CA*

Research Mentor **2016–2017**
Nam Thi; Mobley Lab; UCI *Irvine, CA*

Graduate Teaching Assistant **Spring 2016**
Physical Sciences Online Coursework Development; UCI *Irvine, CA*

Graduate Teaching Assistant **Winter 2016**
CHEM 2LB – Honors General Chemistry Lab; UCI *Irvine, CA*

Graduate Teaching Assistant **Fall 2015**
CHEM 1LD – General Chemistry Lab; UCI *Irvine, CA*

HONORS AND ACCOMPLISHMENTS

Graduate Research Fellowship **2016–2020**
National Science Foundation

U.S. Patent and Trademark Office Registration **2020**
Registration Number: 78,984

Penny J. Gilmer Travel Grant **2018, 2020**
OpenEye Scientific Software, Annual CUP Meeting

GROW International Research Fellowship **2019**
National Science Foundation

Biophysical Society Travel Award **2019**
2019 Biophysical Society Annual Meeting

2nd Place Poster Presentation **2018**
Fourth Annual Ewha-UCI Joint Symposium

Conference Involvement Award **2018**
2018 Computational Chemistry Gordon Research Conference

UCI School of Physical Sciences Travel Award **2018**
2018 Computational Chemistry Gordon Research Conference

2nd Place Oral Presentation **2017**
Graduate Women in Science 27th Student Research Conference

1st Place Oral Presentation **2017**
UCI Graduate Research Symposium; Chemistry/Physics Division

UCI School of Physical Sciences Travel Award **2017**
253rd ACS National Meeting

Summer Institute Scholarship **2017**
San Diego Supercomputer Center

Outstanding Teaching Assistant Contributions **2016**
UCI Department of Chemistry

PUBLICATIONS

1. **Lim, V. T.**; Mobley, D. L. Benchmark assessment of molecular geometries and energies from small molecule force fields. *In preparation.*

2. **Lim, V. T.**; Freites, J. A.; Tobias, D. Potential of mean force calculations for the membrane permeation of Hv1 channel blockers. *In preparation.*

3. Rovsnik, U.; **Lim, V. T.**; Blau, C.; Howard, R. J.; Lindahl, E. Cryo-EM Structure Determination and Model Fitting of the Proton-Gated Ligand-Gated Ion Channel glic at Multiple pH States. *In preparation.*

4. **Lim, V. T.**; Geragotelis, A. D.; Lim, N. M.; Freites, J. A.; Tombola, F.; Mobley, D.; Tobias, D. Insights on Small Molecule Binding to the Hv1 Proton Channel from Free Energy Calculations with Molecular Dynamics Simulations. *Under review.*

5. **Lim, V. T.**; Bayly, C. I.; Fusti-Molnar, L.; Mobley, D. L. Assessing the Conformational Equilibrium of Carboxylic Acid via Quantum Mechanical and Molecular Dynamics Studies on Acetic Acid. *Journal of Chemical Information and Modeling* **2019**, 59, 1957–1964, PMID: 30742770.

6. Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; **Lim, V. T.**; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *Journal of Chemical Theory and Computation* **2018**, 14, 6076–6092, PMID: 30351006

7. Crespo, A.; Rodriguez-Granillo, A.; **Lim, V. T.** Quantum-Mechanics Methodologies in Drug Discovery: Applications of Docking and Scoring in Lead Optimization. *Current Topics in Medicinal Chemistry* **2017**, 17, 2663–2680.

8. **Lim, V. T.**; Tsukanov, S. V.; Stephens, A. B.; Johnston, J. N. Enantioselective Synthesis of α-Bromonitroalkanes for Umpolung Amide Synthesis: Preparation of tert-Butyl ((1R)-1-(4-(benzyloxy)phenyl)-2-bromo-2-nitroethyl)carbamate. *Organic Syntheses* **2017**, 88–99.

ABSTRACT OF THE DISSERTATION

Computational Characterization of Hv1 Proton Channel Inhibitors and Investigation of the Tools in Molecular Simulations

By

Victoria T. Lim

Doctor of Philosophy in Chemistry

University of California, Irvine, 2020

Professor David L. Mobley, Co-Chair

Professor Douglas J. Tobias, Co-Chair

Computer simulation methods have the power to greatly optimize drug design, reducing the time and cost that it takes to bring a new drug from lab to market. Two factors of concern in rational drug design are the binding affinity of some small molecule to a biomolecular target and the molecule's ability to permeate the cellular membrane to reach said target. In the first half of this work, I describe my efforts analyzing the protein-ligand binding affinity and membrane permeability of inhibitors of the Hv1 proton channel. Hv1 is a voltage-gated proton channel with clinical relevance, being highly expressed in highly metastatic cancer cell lines. The inhibition of Hv1 has been found to slow the growth of cancer cells. My research on Hv1 channel inhibitors lays the groundwork for future studies on adapting these inhibitors for more effective channel blocking.

In the second half of this work, I examine the tools that are foundational to molecular simulations—force fields—as well as explore the conventional wisdom behind proton assignment in carboxylic acids. Biomolecular simulation represents a valuable computational microscope for drug discovery efforts; however, it is only through correct setup protocols and well-calibrated tools that we can obtain the most accurate and meaningful results.

Chapter 1

Introduction

Proteins are one of the fundamental macromolecules of life. Extraordinarily diverse in structure and function, these tiny molecular machines are critical to nearly every cellular process in our bodies. Given their integral involvement at the most basic level in our ability to function, the modulation of proteins by small molecules may modify a disease process, such as aspirin which inhibits the enzyme (cyclooxygenase) that produces pain-signaling prostaglandin molecules, [304] or loratadine (Claritin) which treats allergies by blocking the histamine receptor. [17] This principle of protein modulation for therapeutic intervention is often difficult to achieve in practice, however, as drug discovery and development have been reported to exceed $2 billion per new drug over an average time span of at least ten years. [70, 13] The majority of costs may be attributed to failures that happen along the drug discovery pipeline. [69] More efficient methods for drug design and evaluation, such as computer-aided drug design techniques, [173, 282, 148, 220, 187, 55] can lead to identification of more promising candidates as well as quicker decisions to terminate drugs that will eventually fail. Specifically, we can use computational techniques to validate experimental results and, from there, apply atomistic insights to guide future experiments.[127]

In the first part of this dissertation, I focus on the Hv1 protein, a proton channel whose primary function is to release proton buildup from within the cell. Hv1 is coded by the HVCN1 gene [243, 262] which is evolutionarily conservatively in a variety of species (e.g., humans, [172] dinoflagellates, [281] and insects [39]) and cell types (e.g., lung epithelia,[84] brain microglia [334], lymphocytes [265, 33]). In humans, Hv1 contributes to the respiratory burst of white blood cells, [244, 62, 339] the capacitation of sperm, [288, 181, 339, 240, 179] and limiting the release of neutrophil granules. [219] Inhibitors of Hv1 may provide treatment for several diseases, as Hv1 has been found to enhance brain damage in ischemic stroke [334] and is also overexpressed in highly metastatic human breast tumor tissues and cell lines. [323] Effective inhibition of Hv1 can be achieved by ligands that bind the extracellular side of the channel, including Zn^{2+}, [262, 243, 42] small peptide toxins such as Hanatoxin, C6, and AGAP/W38F, [4, 339, 290] or by ligands that bind the cytoplasmic side, such as guanidine derivatives. [128, 248] My research involves examining the channel binding and membrane permeability properties of the Hv1-blocking guanidine derivatives using the computational technique of molecular dynamics.

Classical molecular dynamics (MD) implements basic principles of classical mechanics to analyze the physical movements and interactions of a molecular system over time. MD can be applied to study a myriad of aspects related to drug discovery including pharmacophore modeling, drug binding, and ADME (absorption, distribution, metabolism, and excretion) properties. [61, 279, 28] In atomistic molecular dynamics, the molecular system (e.g., a protein-ligand complex in water) is represented as a collection of atoms that interact with each other through bonded and non-bonded forces described by a potential energy function. MD simulations propagate the system forward in time by solving Newton's equations of motion. With the assumption of ergodicity, that is, assuming that the full relevant phase space is sampled over long time scales, thermodynamic quantities of interest can be obtained from time-averaged quantities of the pertinent variables of coordinates and velocities. [168]

Part of my work on Hv1 involved investigating the characteristics and energetics of a small molecule inhibitor binding to the channel. Specifically, I computed relative binding free energies of a prototypical guanidine derivative, 2-guanidinobenzimidazole (2GBI), to Hv1 and a series of mutants. In the context of drug discovery, relative binding free energies may be applied to a congeneric series of ligands to evaluate particular chemical changes of the ligand that may strengthen protein-ligand binding, for example. In application, this requires a reliable structure of the protein model. Since the discovery of the proton channel gene (HVCN1) in 2006, [262, 243] a crystal structure of human Hv1 has been elusive in large part due to the overall difficulty of crystallizing membrane proteins. [35, 229, 210] Lack of a human protein model notwithstanding, the Tobias group at UC Irvine developed an atomistic homology model of human Hv1 in an open channel state using a closed state chimeric structure of mouse Hv1. [94, 289] While this model has been supported by experimental data on gating charge, salt-bridge interactions, and proximity of specific residues, we reasoned that further investigation of the open state model with respect to small molecule inhibitor binding would contribute to its validation as well as pave the way to designing more effective channel blockers. With reference data from electrophysiological measurements following site-directed mutagenesis of Hv1, [128] I computed relative binding free energies of 2GBI to the wild type and six mutants of Hv1.

While straightforward, "vanilla" molecular dynamics simulations can be used to generate equilibrated structures which can themselves be studied, more complicated processes require the use of extended techniques such as enhanced sampling methods [336] or specialized free energy techniques. [156] The computation of binding free energies falls into the latter category as direct simulation of binding and unbinding processes would involve timescales many times the residence time of the complex (e.g., microseconds to milliseconds [291, 25, 293]) which quickly becomes computationally intractable. [226, 227] In lieu of simulating direct binding and unbinding events, alchemical free energy perturbation represents one of the primary approaches used today for computing binding free energies via simulations, being

both computationally feasible and theoretically rigorous. [206, 46]

Alchemical free energy perturbation (FEP) entails a non-physical transformation between an initial reference state (e.g., bound ligand in wild type protein) and a final target state (e.g., bound ligand in mutant protein). FEP is a work-based method [156, 333] based on equilibrium sampling in which the transformation occurs instantaneously. [238, 5] From some reference state A to some target state B, the difference in free energy between the two states can be calculated from the work involved in perturbing state A to state B according to Jarzynski's equality [140, 139] as:

$$\exp(-\beta\,\Delta G) = \Big\langle \exp(-\beta\,W_{A\to B}) \Big\rangle\,. \tag{1.1}$$

In the limit that the transformation is performed instantaneously as in FEP, Jarzynski's equality reduces to the Zwanzig equation [342, 44]:

$$\exp(-\beta\,\Delta G) = \Big\langle \exp(-\beta\,\Delta U \Big\rangle_{A}\,. \tag{1.2}$$

In other words, the free energy difference between A and B is computed from an ensemble average relying on the Boltzmann weighted potential energy difference between those states as in the equation:

$$\Delta G_{A,B} = -k_{\mathrm{B}}T \ln \left\langle e^{-\beta(U_B(\boldsymbol{x})-U_A(\boldsymbol{x}))} \right\rangle_{A}\,. \tag{1.3}$$

Given that free energy is a state function and is path-independent, the same free energy difference would be obtained, in theory, if the reference and target states were reversed:

$$\Delta G_{B,A} = +k_{\mathrm{B}}T \ln \left\langle e^{-\beta(U_A(\boldsymbol{x})-U_B(\boldsymbol{x}))} \right\rangle_{B}\,. \tag{1.4}$$

In practice, the perturbation introduced in the reference state to yield the target state will

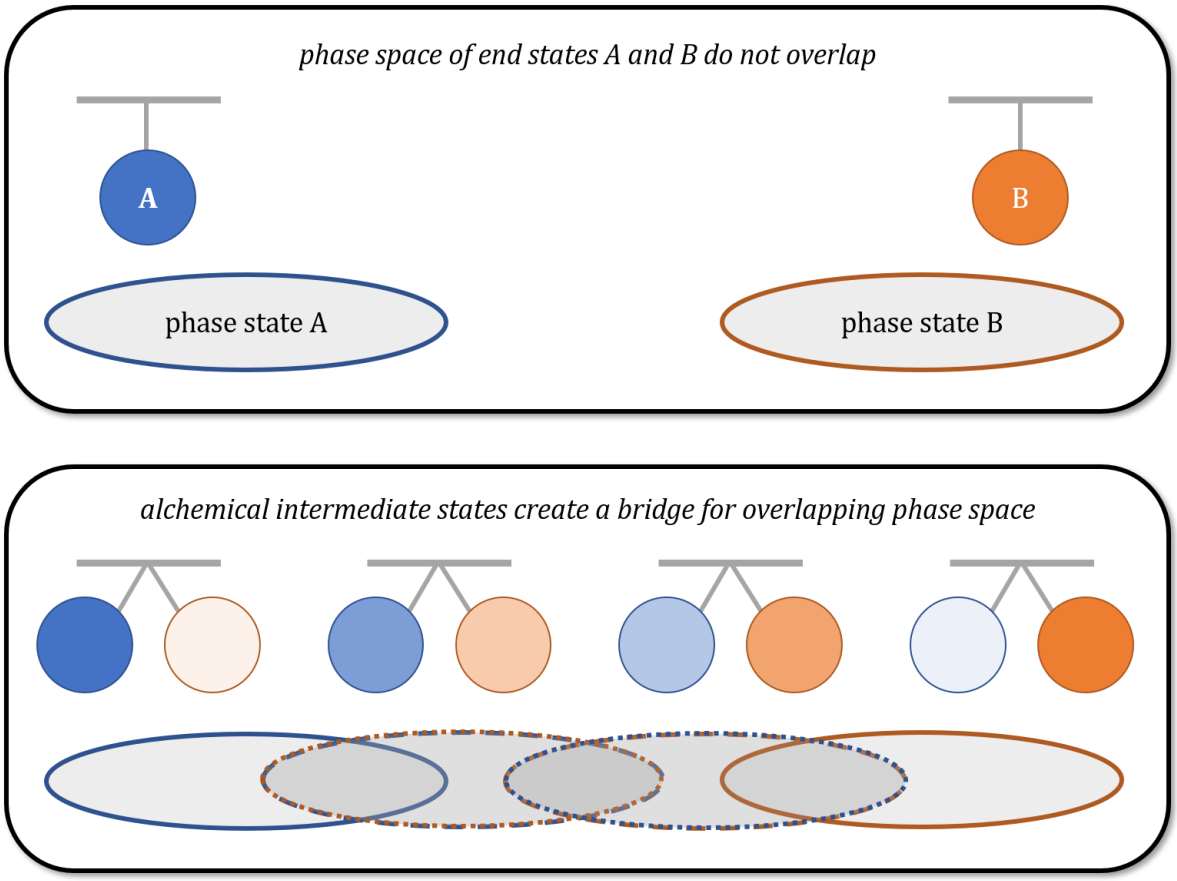


Figure 1.1: Schematic of two states, A and B, of which to compute the free energy difference. Individual simulation of A and B alone would sample separate areas of phase space, leading to a poor estimate of the work involved in perturbing between the two states. The addition of intermediate alchemical states can provide a way to bridge the phase space between these two states.

in many cases be significant enough that the important low-energy configurations are not the same for states A and B. [274, 185] Directly computing equation 1.3 would not yield reliable estimates of the free energy difference as there would be little to no overlap in the phase space that they sample. [44, 156] However, we can effectively build a bridge in configurational space by introducing states, or "windows," which are intermediate to A and B (Figure 1.1). These states do not need to be physically meaningful, in accordance with the path-independent formalism of free energy perturbation. The full transformation between the initial/reference and final/target states then entails a series of reversible steps comprising several intermediate alchemical structures.

Alchemical transformations can be achieved using a dual topology paradigm, in which both the reference and target states are present in the simulation for all windows. [89, 233] In the case of computational mutagenesis, such as the mutation from an aspartate to a glutamate residue in a protein, a hybrid residue is constructed to contain both aspartate (Asp) and glutamate (Glu). All atoms in both residues are present at the same time in the system. In the alchemical states between the Asp system and the Glu system, the non-bonded interactions of the disappearing atoms (Asp) are gradually decreased to zero. Concurrently, the non-bonded interactions of the appearing atoms (Glu) are slowly increased from zero to full. In this way, all intermediate windows have both end states present, but to varying degrees based on where the window is throughout the transformation. In this example, the initial window would have aspartate fully interacting with the system while glutamate is a "ghost" residue, and the final window would have glutamate fully interacting with the rest of the system while aspartate is present only as a non-interacting ghost residue.

Using the alchemical free energy perturbation technique with atomistic molecular dynamics, I quantitatively and qualitatively characterized the protein-ligand binding of 2GBI to the Hv1 proton channel. This research is presented in **Chapter 2**.

In addition to probing the channel binding of 2GBI to Hv1, I explored differences in the thermodynamic properties underlying membrane permeability between 2GBI and another Hv1 inhibitor, 5-chloro-2-guanidinobenzimidazole (ClGBI). Generally speaking, many orally administered drugs must pass through several cellular membranes to reach their targets, [167] and, accordingly, membrane permeability is a major determinant of drug bioavailability and *in vivo* efficacy. [6, 12, 302] The guanidine derivatives 2GBI and ClGBI are both intracellular open channel blockers of Hv1 which share the same molecular identity except for the replacement of one hydrogen atom by a chlorine atom. Unlike 2GBI, however, ClGBI has been found to be able to access the protein binding site when present in the extracellular medium. [295] It is hypothesized that ClGBI is able to permeate the cellular membrane to

act as a channel blocker to Hv1. I investigated this hypothesis by computing the potentials of mean force characterizing the lipid membrane permeation of 2GBI and ClGBI based on the principles of the inhomogeneous solubility-diffusion model.

The passive transport of solutes through an inhomogeneous environment such as that of a lipid membrane can be described with the inhomogeneous solubility-diffusion model. [190, 66] Derived from the steady-state flux which defines the number of molecules passing through the membrane per unit area per unit time, this model relates the experimentally accessible permeability coefficient P_m to the position-dependent potential of mean force $W(z)$ and the local diffusion coefficient $D(z)$ as:

$$\frac{1}{P_m} = \int_{-L/2}^{L/2} \frac{\exp W(z)/k_B T}{D(z)} \, \mathrm{d}z \,, \tag{1.5}$$

In this equation, k_B is the Boltzmann constant, T is the temperature, and z is the relative position of the solute along the transmembrane axis. In general terms, the membrane permeability of some solute can be modeled by a thermodynamic-kinetic relationship which depends exponentially on the thermodynamic component $W(z)$ and linearly on the kinetic component $D(z)$. My focus was on calculating the potentials of mean force for 2GBI and ClGBI to evaluate the thermodynamic contributions that differentiate their relative membrane permeabilities.

Similarly with obtaining binding free energies using MD simulations, computing a potential of mean force for membrane permeability requires specialized sampling techniques, as brute force sampling only works well for small, rigid neutral solutes, for instance oxygen or water. [308] We can apply potential of mean force (PMF) methods to determine free energy differences, as was the goal of alchemical free energy perturbation introduced earlier. In contrast, the PMF allows one to assess the free energy change along some collective variable, or physically meaningful reaction coordinate, such as the measure of a dihedral angle or the distance

between two atoms. The collective variable must be defined prior to performing the PMF simulation(s) and must be chosen carefully to adequately and efficiently construct the free energy profile for the process in question. [83, 40, 72] In the case of membrane permeation, the typical one-dimensional collective variable employed is the relative position of the center of mass of the solute along the membrane normal. Higher dimensional collective variables may also be employed, such as to account for the orientational degrees of freedom of the solute in the membrane, [287, 52] internal conformational changes of the solute, [138, 34] or the coordination numbers between the solute and the surrounding lipid or water molecules. [272, 96] This increased detail in the reaction coordinate may allow a more thorough exploration of the primary factors underlying membrane permeation but often comes as a trade-off due to the increased computational expense.

One method to compute the potential of mean force using molecular dynamics simulations is known as adaptive biasing force. A general challenge of MD simulations is to achieve efficient sampling of the relevant areas of phase space, given that biomolecular processes often involve considerable free energy barriers on a rugged energy landscape. Adaptive biasing force (ABF) seeks to transform this complex potential surface to one that is effectively flat in the long-time limit, allowing uniform sampling throughout the space. [58, 122] In contrast to many other enhanced sampling methods which use a biasing potential (e.g., harmonic potentials in umbrella sampling or Gaussian potentials in metadynamics), ABF applies a biasing *force* on the specified collective variable. These two quantities are directly related as the force is the negative gradient of the potential. However, there is an important nuance between them in terms of applying a bias to the system: the potential is a global property whereas the force (i.e., the gradient) is a local property. That is to say, appropriately biasing the potential requires knowledge of the potential beyond the currently occupied state, whereas a biasing force can be applied from just the knowledge of the local state. Therefore, it is in principle more efficient to enhance simulation sampling using adaptively biased forces compared to adaptively biased potentials. [50]

To achieve an essentially flat potential surface in the long-time limit, ABF applies a force, changing over time and value of the reaction coordinate, which effectively cancels out the estimate of the average force. As a short mathematical digression, the standard dynamics become altered with the introduction of the mean force $A'(z)$, which is the ensemble average of the instantaneous force F_ξ:

$$A'(z) = -\frac{\int F_\xi(\mathbf{x}) \exp\left[-\beta \mathrm{V}(\mathbf{x})\right] \delta_{\xi(\mathbf{x})-\mathbf{z}}(\mathrm{d}\mathbf{x})}{\int \exp\left[-\beta \mathrm{V}(\mathbf{x})\right] \delta_{\xi(\mathbf{x})-\mathbf{z}}(\mathrm{d}\mathbf{x})} \,. \tag{1.6}$$

In this equation, $\mathbf{x}$ is the position, ξ is the position-dependent expression for the collective variable, z is the value for the collective variable ξ, and V is the (unbiased) potential of the system. The classic expression for the instantaneous force is the following:

$$F_\xi = -\nabla V \cdot \frac{\partial \mathbf{x}}{\partial \xi} + \beta^{-1} \frac{\partial \ln|\mathrm{J}|}{\partial \xi} \,, \tag{1.7}$$

where the second term involves the Jacobian matrix arising from the transformation from Cartesian coordinates to generalized coordinates involving ξ. In other words, the mean force along the collective variable is canceled out by an equal and opposite force; this force is introduced into the system to facilitate traversal of the energy landscape fraught with barriers and valleys. Technically, it would be difficult to employ equation 1.7 in its given form for a complex coordinate definition of the collective variable as the equation involves cumbersome second derivatives, but mathematical approaches can be implemented to circumvent this issue as discussed by the original authors of ABF.[58] In the end, the potential of mean force can, in the simple case of a one-dimensional collective variable, be obtained from integration of the estimated mean force. The ABF method has been widely used to study a variety of biomolecular processes including protein-ligand dissociation [217], peptide folding, [294] membrane permeability. [171]

With the adaptive biasing force algorithm, I simulated the dynamics of 2GBI and ClGBI,

two inhibitors of Hv1, permeating a lipid bilayer. I computed the potentials of mean force to examine the thermodynamic effects of their passive membrane permeation. This research is presented in **Chapter 3**.

Thus far I have discussed my work on the application of molecular simulations to gain understanding and insight on the therapeutically relevant inhibition of Hv1. In the last two chapters of this dissertation, I focus on evaluating the tools and setup methodology that contribute to the models upon which the simulations are conducted.

An important tool of classical molecular dynamics is the force field. The time evolution of molecular systems in MD simulations relies on evaluating the forces acting on each atom at each time step. These forces are derived from potentials which describe the energy of various components of the system, typically including bonds, valence angles, dihedral angles, improper dihedral angles, electrostatic interactions, and van der Waals interactions. In the simple case of a bond, its functional form is represented by a harmonic potential:

$$U_{\text{bond}} = k(r_{ij} - r_0)^2 \,, \tag{1.8}$$

which involves two preset parameters, the spring constant k and the equilibrium distance r_0. The force field of molecular modeling refers to the complete set of functional forms and parameters from which the time-dependent forces are calculated.

A force field contains hundreds to thousands of parameters depending on the range of chemistry covered. Considering bond terms again, the values of the two parameters will differ depending on the bond type; for example, the carbon-carbon bond in ethyne will necessarily have a higher spring constant and lower equilibrium distance than the carbon-carbon bond in ethane. These parameters, along with the parameters corresponding to the other force field terms, can be obtained from reference data from experiments or quantum mechanics. The process of, and inputs for, parameterization vary based on force field. Ultimately, molecular

mechanics force fields are inherently approximations. Given the wide applicability of MD simulations, having accurate force fields–that is, suitable functional forms with appropriate parameters based on the chemistry modeled–is crucial to achieve meaningful results.

Force fields are developed to model particular biomolecules, including proteins, [134, 183, 109] nucleic acids, [85, 159] and lipids. [153, 68] There are also several force fields to model a general class of small molecules, for example, to represent drugs in binding sites. Small molecule force fields must be able to handle arguably greater complexity than force fields for biopolymers which have narrower ranges of available chemistries (e.g., a finite library of amino acids). Despite this increased complexity, there has traditionally not been much attention or concerted effort towards the development of small molecule force fields, which are sometimes born from the generalization of biomolecular force fields to cover more chemical space. [202, 319, 318] As a result, different small molecule force fields will likely differ in how well they represent chemistry.

Understanding the strengths and weaknesses of various small molecule force fields is important to catalyze the development of new force fields, as the Open Force Field Initiative aims to do. [202] In **Chapter 4**, I present my work on benchmarking a set of small molecule force fields with respect to reference quantum mechanical data. I analyzed force field-optimized geometries and conformer energies from six small molecule force fields on a set of over 20,000 molecular structures. The insights gained from this research may be useful to future force field development efforts.

Shifting gears from assessing the critical tools of molecular simulations, I examine a default setup protocol regarding accurate hydrogen placement in carboxylic acid systems. This work was motivated by a discussion with Christopher Bayly from OpenEye Scientific Software, who strongly believed that carboxylic acids would generally only be found in their *syn* conformation and that the *anti* conformation would be sparsely, if at all, populated. Classical molecular modeling packages typically protonate neutral carboxylic acids into a

syn conformation. The directionality of a hydrogen bond determines the interactions that it will form, such as hydrogen bonding between a drug and its receptor. These interactions may contribute 1-5 kcal/mol of energetic stability [271, 98] and may also confer specificity of ligand binding. [41, 315] In **Chapter 5**, I question the prevailing wisdom of *syn*- and *anti*-conformation proton assignment, quantifying the considerable energetic barrier for interconversion between the two states as well as approximating the energy distributions of the populations of both states.

To you, dear reader, this conclusory paragraph is only the end of the beginning. For me, it is the end of the end. In this introductory chapter, I have summarized the broader concepts of what motivates this work, to optimize the process of drug design using computer simulation methods. I discussed a few of those methods, namely classical molecular dynamics, alchemical free energy perturbation, and adaptive biasing force. I introduced two examples of my efforts in computer-aided drug design in **Chapters 2 and 3**, where I investigated the channel binding and membrane permeability of Hv1 channel blockers. Finally, while biomolecular simulation represents a valuable computational microscope to probe targets and processes with potential therapeutic applications, a microscope with a flawed lens will distort the observations. In **Chapters 4 and 5** of this work, I examined various parts of the microscope itself, specifically force fields and the default hydrogen placement of carboxylic acids. As I conclude this section, I hearken back to the final words of the personal statement I wrote upon starting my PhD, hopeful that this work "advances knowledge and leaves a positive impact in the scientific community and the world around us."

Chapter 2

Insights on small molecule binding to the Hv1 proton channel from free energy calculations with molecular dynamics simulations

Hv1 is a voltage-gated proton channel whose main function is to facilitate extrusion of protons from the cell. The development of effective channel blockers for Hv1 can lead to new therapeutics for the treatment of maladies related to Hv1 dysfunction. Although the mechanism of proton permeation in Hv1 remains to be elucidated, a series of small molecules have been discovered to inhibit Hv1. Here, we computed relative binding free energies of a prototypical Hv1 blocker on a model of human Hv1 in an open state. We used alchemical free energy perturbation techniques based on atomistic molecular dynamics simulations. The results support our proposed open state model and shed light on the preferred tautomeric state of the channel blocker. This work lays the groundwork for future studies on adapting the blocker molecule for more effective inhibition of Hv1.

2.1 Introduction

The flux of ions across the cell membrane is regulated by a variety of ion channels. Hv1 is a voltage-dependent ion channel whose main function is to conduct protons through the cell membrane. Like voltage-dependent metal ion channels, the activation of Hv1 is regulated by voltage-sensing domains (VSDs), which are transmembrane modular units that detect changes in the membrane potential. However, in contrast to voltage-dependent metal ion channels, Hv1 lacks a separate pore domain as the VSD is sufficient for proton permeation in addition to mediating channel gating.[262, 243, 242] The action of Hv1 contributes to physiological processes such as the production of reactive oxygen species, the bioluminescence of dinoflagellates, and the maturation of human sperm.[77, 258, 180] Hv1 was found to be specifically and highly expressed in highly metastatic human breast tumor tissues, and inhibition of Hv1 was shown to reduce cancer metastasis and tumor.[322, 323] More generally, the development of inhibitors to block the Hv1 channel may also lead to therapeutic benefit for other Hv1-related maladies, including allergies[270, 341] or exacerbated brain damage in ischemic stroke.[334]

The structure of human Hv1 has not yet been experimentally determined. Human Hv1 is a homodimer, but each monomer has its own pore and can function independently.[295, 155] We recently developed an atomistic model of the human Hv1 VSD open state using molecular dynamics (MD) simulations on the microsecond timescale in a hydrated lipid bilayer under an applied membrane potential.[94] As discussed in the work by Geragotelis et al., there exist other proposed structures for the human Hv1 open state;[332, 161, 38, 97, 245] however, these were either modeled from crystallographic structures of voltage-gated ion channels which do not contain a native permeation pathway through the VSD,[332, 161, 38, 245] or they have alternative VSD conformations from artificial restraints during MD simulations.[97] Our structure was modeled starting from the crystal structure of a chimeric construct based on mouse Hv1 in a putative closed state[289] from which we generated closed

and open state configurations exclusively of the applied membrane potential. The elicited conformational changes were consistent with the mechanism of activation proposed for metal ion channel VSDs.[145, 129, 241] The resulting structural models are supported by available experimental data.[94] Further investigation of our open state model with respect to small molecule inhibitor binding would contribute to its validation and refinement and pave the way to designing more effective channel blockers.

Effective inhibition of Hv1 can be achieved by ligands that bind the extracellular side of the channel, including Zn^{2+};[42, 243, 262] small peptide toxins such as Hanatoxin, C6, and AGAP/W38F;[4, 339, 290] or by ligands that bind the cytoplasmic side, such as guanidine derivatives.[128, 248] We focus on a representative guanidine derivative, 2-guanidinobenzimidazole (2GBI), to understand how it binds to the channel. Compounds targeting the VSD often have multiple targets.[108, 36] However, the open state VSD in Hv1 is more hydrated compared to non-conducting VSDs and offers enough space for 2GBI to bind at a location that, in other VSDs, is occupied by S4 arginines. We propose that this feature could be exploited for the development of Hv1-specific drugs. 2GBI and related compounds can produce maximal inhibition of human Hv1 of $> 90\%$ and can block Hv1 both at the plasma membrane and in intracellular compartments. The binding of 2GBI to Hv1 is described as a "foot in the door" mechanism of block, by which binding of the ligand prevents channel closure and slows down channel deactivation.[129] There is currently no known experimental structure of 2GBI in complex with Hv1, and the details of 2GBI binding are not completely known; however, mutagenesis experiments[128] suggest several key residues that play a role, including the selectivity filter D112[212] as well as F150 and R211 lining the central constriction region.[332]

The structure of 2GBI (Figure 2.1, top) was rationalized by the fact that Hv1 is inhibited by the guanidinium ion[295] which is structurally similar to the voltage-sensing arginine residues in the S4 helix.[100] 2GBI is a more effective inhibitor for Hv1 than guanidinium with the

addition of its benzimidazole moiety. Despite that fact, replacement of one nitrogen atom on the five-membered ring by an oxygen atom resulted in less effective inhibition than even guanidinium.[129] Details of the binding site aside, this raises the question of the structure and charge distribution of 2GBI in relation to its tautomeric forms.[124]

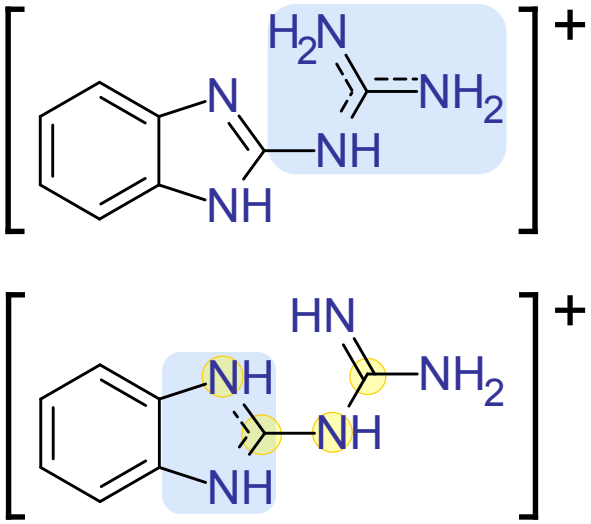


Figure 2.1: Two tautomers of 2GBI are investigated in Hv1 binding: "gbi1" (top) and "gbi2" (bottom). Blue regions have higher positive charge density. Yellow circles mark atoms of the dihedral angle selected for the dihedral scan in this work.

In this work, we studied the binding of 2GBI to our open state human Hv1 model[94] using alchemical free energy perturbation with atomistic molecular dynamics simulations. We considered a series of six protein mutations previously shown to alter 2GBI apparent binding affinity, and we computed relative binding free energies of these mutations compared to wild type Hv1 for two tautomeric forms of the inhibitor. This work reveals structural characteristics of 2GBI binding, emphasizes the prevalence of gbi2 in binding Hv1, and may inform future efforts to optimize small molecule blockers of Hv1.

2.2 Methods

2.2.1 Ligand parameterization with CGenFF

Our study examines two tautomers of the positively charged 2GBI, depicted in Figure 2.1. Both of these structures contain a guanidine-like moiety but have different centers of excess charge, located in either the guanidine region ("gbi1") or in the central imidazole moiety ("gbi2"). The force field for each tautomer was developed using the CHARMM General Force Field (CGenFF).[305, 306, 94] We generated reference structures from gas phase quantum mechanical (QM) geometry optimizations using the Psi4 software package.[230] We employed the MP2/6-31G* method[120, 326, 87, 88, 246] to stay consistent with CGenFF development.[305, 306] The generated gbi2 force field was modified to achieve better agreement regarding the geometry-optimized structures between the force field and QM results. The force field modification details and final parameters for both ligands can be found in the supporting information ("Ligand parameterization for 2GBI").

2.2.2 Docking calculations and pose refinement with molecular dynamics simulations

We docked each tautomer of 2GBI into Hv1 with AutoDock Vina version 1.1.2.[297] Twenty protein configurations were selected from a ~10 microsecond simulation,[94] and we removed the membrane and all water molecules during docking. While the presence or absence of water molecules may influence results from docking,[330, 255] specific water molecules involved in binding 2GBI are as of yet unknown. Thus, our aim in docking was just to generate an initial set of binding poses to be further refined in MD simulations, which allow waters to rearrange as they prefer. The ligand dihedral angle, defined by the four atoms marked in Figure 2.1 (bottom), was held fixed in a planar conformation to prevent high-energy initial

structures resulting from strained non-planar conformations (see supporting information, Figures A.2 and A.4).

The resulting poses were filtered by a "reverse clustering" technique. We used the Clustering plugin[103] in VMD[135] to cluster all poses by root mean square deviation (RMSD) with a 3.0 Å cutoff. We retained poses that were distinct from each other by more than this threshold. This process yielded a diverse set of starting poses ensuring good coverage of the multiple possible binding modes in the binding site, allowing minor differences between similar poses to be explored by our subsequent MD simulations. Each binding pose was placed back into its hydrated protein-membrane configuration, and we removed water molecules overlapping within 2 Å of the ligand, otherwise retaining original hydration of the pore.

We then ran short (5 nanosecond [ns]) MD simulations for all distinct poses of each 2GBI tautomer using the NAMD software package, version 2.11.[235] Prior to dynamics, 5000 minimization steps were performed using the conjugate gradient algorithm. Backbone alpha carbon atoms were fixed during minimization using harmonic position restraints with a force constant of 1 kcal/mol/Å^2. The restraints were gradually turned off over 400 picoseconds (ps). Dynamics were then run with a 2 femtosecond (fs) time step in the isothermal-isobaric ensemble at a temperature of 300 K and a pressure of 1 bar. The Langevin thermostat was applied with a damping constant of 5 ps^{-1}, and the Langevin piston method was applied with an oscillation period of 200 fs and a damping time of 100 fs. Periodic boundary conditions were applied in all dimensions, with the z-axis normal to the membrane bilayer. The SHAKE algorithm was used to restrain all bonds involving hydrogen atoms to their equilibrium values. Long-range electrostatic interactions were applied with the Particle Mesh Ewald (PME) algorithm; short-range Lennard-Jones and Coulombic interactions were calculated using a cutoff of 12 Å and a switching function applied beyond 10 Å. Bonded interactions and short-range forces were calculated every 2 fs, and long-range forces were calculated every 4 fs. An electric field was applied using a field vector of 0.14 kcal/(mol Å e) in the z direction,

corresponding to a depolarizing membrane potential of +150 mV.

The final poses selected for free energy calculations were chosen based on the stability of the ligand in the binding position as well as evaluation of the hypothesized placement of 2GBI from experimental double-mutant cycle analyses.[128] Our evaluation criteria for pose analysis is outlined in the supporting information ("Pose selection and refinement with MD"). The poses were equilibrated for at least 20 ns before initiating free energy calculations. We verified that the pore retained hydration in each of the docked configurations during the equilibration simulations (see supporting information: Figure A.6).

Configurations from our MD simulations for pose refinement and subsequent free energy calculations were all rendered using VMD version 1.9.3.[135]

2.2.3 Alchemical free energy calculations

We applied the thermodynamic cycle approach as depicted in Figure 2.2 to calculate the change in the binding free energy of 2GBI from wild type Hv1 to the following protein mutations: D112E, V178A, S181A, V109A, R208K, and R211S. The first and last single letter codes denote the starting and final amino acids, respectively. The relative binding free energy in this cycle is calculated as the difference in the free energies of mutation from the holo state to the apo state, i.e., $\Delta\Delta G = (\Delta G_4 - \Delta G_3)$. The computed values were evaluated alongside the relative binding energies from experimental mutagenesis studies,[128] which correspond to processes ΔG_1 and ΔG_2 of the thermodynamic cycle; i.e., $\Delta\Delta G = (\Delta G_2 - \Delta G_1)$.

Processes ΔG_3 and ΔG_4 of the thermodynamic cycle represent the alchemical transformations conducted computationally. The initial and final states for each of ΔG_3 and ΔG_4 are connected through a series of non-physical intermediate states ("λ windows") that comprise the alchemical transformation. This transformation is controlled by a parameter λ, starting

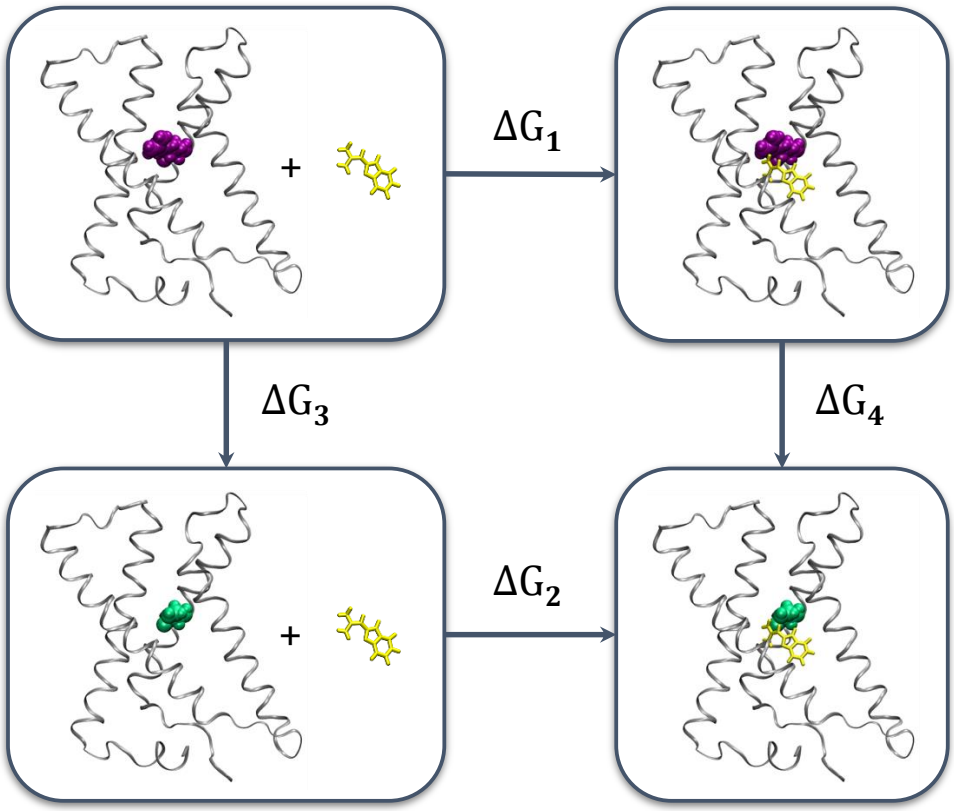


Figure 2.2: Thermodynamic cycle used to compute relative binding free energies. The initial side chain from wild type Hv1 is shown in purple vdW spheres, and the final side chain after alchemical mutation of Hv1 is shown in green vdW spheres. The 2GBI ligand is drawn in yellow licorice. The membrane and water molecules are not shown for clarity. The relative binding free energy from experiment is calculated from $\Delta\Delta G = (\Delta G_2 - \Delta G_1)$, which is thermodynamically equivalent to the *alchemically* computed relative binding free energy of $\Delta\Delta G = (\Delta G_4 - \Delta G_3)$.

from a value of zero at the initial state (i.e., wild-type protein residue) and reaching a value of one at the final state (i.e., mutant protein residue).

The alchemical transformation is subdivided into 40 equivalently spaced λ windows, such that the first window ranges from $\lambda = 0$ to $\lambda = 0.025$ and the final window ranges from $\lambda = 0.975$ to $\lambda = 1$. With simple linear modification of interactions, "end point catastrophes" may result when $(\lambda \to 0)$ or $(\lambda \to 1)$, during which incoming atoms may appear where other particles already exist.[27] We avoid this by applying a soft-core potential on perturbed atoms to gradually scale their short-range nonbonded interactions with the rest of the system.[337]

Each alchemical transformation was conducted in both the forward and reverse directions (e.g., aspartate mutated to glutamate as well as glutamate mutated to aspartate) to facilitate convergence and sampling of overlapping phase space of the end states. While the free energy difference between two states can be determined by the Zwanzig relationship,[342] this method is often slow to converge and results in poor phase space overlap.[274, 185] The Bennett acceptance ratio (BAR)[23] minimizes the statistical variance between two ensembles with overlapping configurational space and yields the optimal averaging of the forward and reverse simulations. We used BAR to combine the data from both directions to estimate the final free energies for each of ΔG_3 and ΔG_4. The reported uncertainties of the relative binding free energies are the root sum square values of the ΔG_3 and ΔG_4 standard errors from BAR.

A dual-topology paradigm was applied throughout the alchemical transformation in which both the initial and final states were simultaneously present but non-interacting. Only nonbonded interactions for perturbed (incoming or outgoing) atoms contributed to the cumulative free energy. Van der Waals interactions were scaled as a function of λ across the full range from zero to one. Electrostatic interactions of the annihilating atoms were linearly decoupled over the first half of the transformation ($\lambda = 0$ to $\lambda = 0.5$), after which electrostatic interactions were linearly coupled for the incoming atoms.

We computed the relative binding free energies using alchemical free energy perturbation with atomistic molecular dynamics simulations in NAMD.[235] Each window started from the same initial hybrid structure and comprised 1000 minimization steps, 1 ns of equilibration, and 4 ns of production. The MD simulation settings were maintained as described earlier. For each protein mutation, we simulated a total of 400 ns, consisting of 5 ns in each of the 40 λ windows for the forward and reverse directions. The data from each λ window was subsampled to extract uncorrelated, effectively independent samples using the `pyMBAR` Python package.[273] For all tautomeric states and mutations represented in this study,

including mutation S211R (see Results and Discussion), the total simulation time is 5.6 microseconds.

We modified our standard setup for free energy calculations for mutations R208K and R211S. In these two cases, we additionally applied flat-bottom distance restraints to prevent hypermobility of the arginine and lysine residues. During normal MD simulations, these positively charged side chains are involved in salt bridge interactions with nearby acidic residues. Residue 208 is in proximity to D119, D123, and E192, and residue 211 is in proximity to D112 and D185. However, these salt bridges are not able to be maintained during the alchemical transformations because of the decoupled nonbonded interactions in the dual-topology hybrid molecule. In other words, considering the case of R208K when arginine is mutated to lysine, arginine is fully interacting and present at the start, when $\lambda = 0$. In the final λ stage when $\lambda = 1$, arginine is now non-interacting with the system and therefore does not form its usual contacts. The non-interacting, flexible side chains end up sampling an artificially broadened area of phase space. This improperly skews the phase space overlap between the forward and reverse calculations and adversely affects the free energy estimate (see Results and Discussion). To avoid this issue and improve convergence of the calculations, we added additional distance restraints on the flexible mutating residues for R208K and R211S. The flat-bottom restraints were applied from the terminal nitrogen atom in arginine or the terminal carbon of lysine to surrounding protein residues that were not involved in the alchemical transformation. The reference distances were chosen to be large enough such that the restraint energies were zero when applied to the initial or final structure yet restrictive enough to prevent excessive side chain mobility during the free energy calculations (see supporting information: "Flat-bottom restraints for mutations involving Arg or Lys residues"). The force constants on the restraint walls were set to 10 kcal/mol/Å.

Because some of our free energy calculations involved changes in formal charge, it was necessary to take particular care in these cases. Specifically, following the free energy simulations,

we accounted for the energetic contribution arising from the change in net charge in the R211S mutation. To do this, we applied analytical corrections for the electrostatic finite-size effects as described by Rocklin et al.[257] That being said, these corrections were found to have almost no impact on the final relative binding free energies as we show in the Results and Discussion.

2.3 Results and Discussion

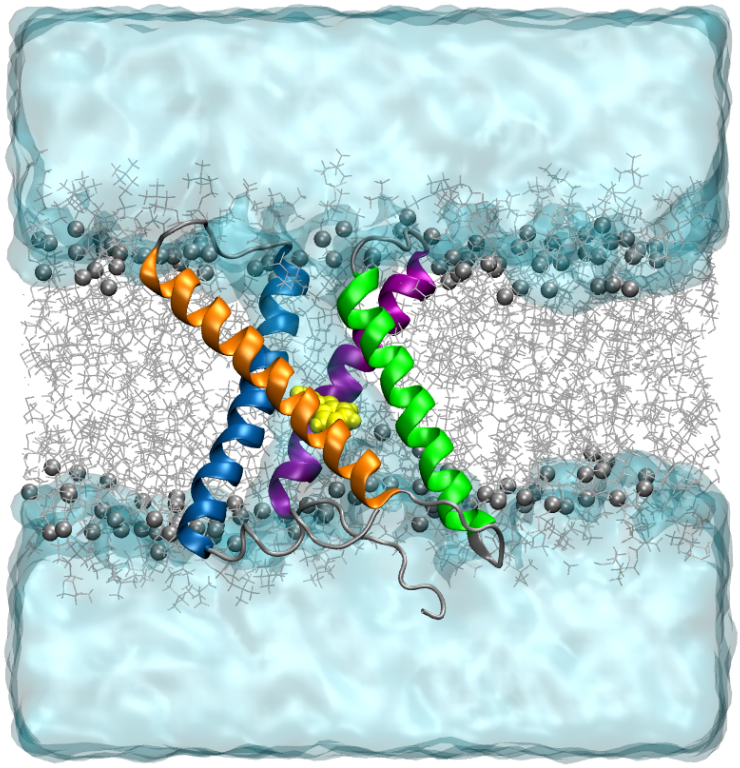


Figure 2.3: Configuration snapshot of 2GBI (gbi1) bound to Hv1, rendered in VMD.[135] The four alpha helical segments of Hv1 are colored as follows: S1 in blue, S2 in orange, S3 in green, and S4 in purple. 2GBI is shown in yellow with the VDW representation. Waters are drawn in slabs of light blue, and lipids are depicted as gray sticks with their carbonyl carbons shown as vdW spheres. The extracellular region is on top.

We employed docking calculations, atomistic molecular dynamics simulations, and alchemical free energy calculations to examine the binding of 2GBI to human Hv1 (Figure 2.3). We computed relative binding free energies for a series of six protein mutations compared to wild type Hv1. Each set of binding free energy calculations was conducted for two tautomers of 2GBI, gbi1 and gbi2, which vary in their centers of excess charge (Figure 2.1). Our results show general qualitative agreement with experimental mutagenesis data and suggests that one tautomer may be more relevant than the other in the protein-bound configuration. We also discuss some limitations when carrying out alchemical transformations of flexible protein side chains.

2.3.1 Positioning of 2GBI tautomers within the Hv1 open state

From the docking calculations, we employed a "reverse clustering" approach to identify 2GBI binding poses (see Methods). That is to say, because the binding mode is unknown and given that docking does not predict definitive binding modes, we aimed to sample a variety of potential poses of 2GBI within the expected binding region in order to find the most reasonable binding mode. The resulting poses from docking calculations were assessed using the predicted interactions between 2GBI and Hv1 from experimental mutagenesis data.[128] Hong et al. proposed a model of binding involving residues D112, F150, S181, and R211. For both gbi1 and gbi2, we selected binding poses where 2GBI was in close enough proximity to interact with these residues. Based on these poses, we note likely contacts with two additional residues, D185 and F182.

Following refinement in MD simulations, we selected a binding pose (Figure 2.4) with the following features: interactions of the guanidine region with the charged residues D112 and R211; hydrogen bond donation from one of the imidazole NH groups to acidic residue D185; and proximity of the hydrophobic residues F150, I154, and F182 to the benzimidazole moiety.

While D112 and R211 are believed to interact in the open state conformation of Hv1,[24] the presence of 2GBI as a channel blocker in the proposed binding site necessarily interrupts this salt bridge formation. We chose the same binding pose for gbi1 and gbi2.

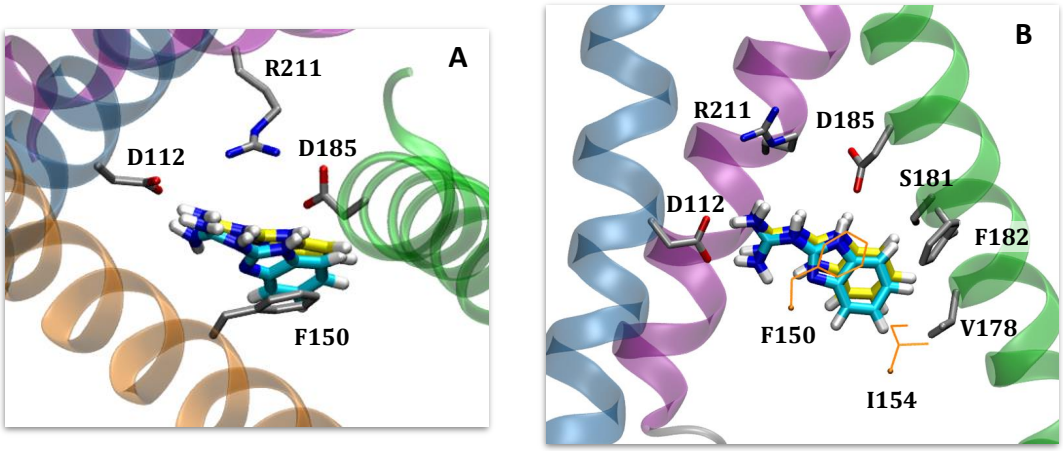


Figure 2.4: 2GBI docked into the open state model of human Hv1, rendered in VMD.[135] The binding pose is consistent with experimental mutagenesis data. Segments are colored in the same manner as in Figure 2.3. The ligand is colored cyan for gbi1 and yellow for gbi2. (**A**) Top-down view from the extracellular end of the channel of bound 2GBI. Loops are hidden for viewing clarity. (**B**) Lateral view of Hv1 (extracellular side on top) with S2 hidden for clarity. The two residues F150 and I154 are part of S2.

Our pose is distinct from that proposed in recent work by Gianti et al.[97] who performed induced fit docking of 2GBI (our gbi1 structure). They obtained a binding pose in which 2GBI is roughly in the same region of Hv1 but angled differently such that the benzo moiety of 2GBI is closer to S1 than to S3. The location of the guanidine group in our pose pointing to R211 is more consistent with experimental work.[128] Differences in binding configuration may be in part due to differences in our open state models.[94]

Chamberlin et al. also described a binding site for 2GBI, using the ligand model we refer to as gbi1, docked into a model of *Ciona intestinalis* Hv1, which itself was based on the crystal structure for the open state chimeric channel Kv1.2-2.1.[38, 182] Their ligand position is similar to our results in that the 2GBI benzo moiety is adjacent to F150; however, their binding pose has the guanidine moiety of 2GBI pointing towards the intracellular end of

the channel in contact with E153, D174, and E171. While the salt bridge patterns are similar between both protein models, the Chamberlin et al. protein structure has S2 in a relatively higher position (towards extracellular side) than the other segments. Protein models notwithstanding, this intracellular-pointing ligand orientation would preclude any interactions with S181 or R211 as proposed by Hong et al.[128] (see supporting information: Figure A.7). Neither of Gianti's or Chamberlin's studies considered alternate tautomers or binding free energy calculations.

2.3.2 Calculation of relative binding free energies

We calculated the relative binding free energies for two tautomers of 2GBI binding to wild type and mutant Hv1 using our open state model. The relative binding free energies were computed using molecular dynamics simulations with alchemical free energy perturbation. We considered six Hv1 mutations, previously reported by Hong et al. using electrophysiological and site-directed mutagenesis experiments.[128] Compared to 2GBI binding to wild type Hv1, these mutations comprise the effects of unfavorable binding (D112E, S181A), favorable binding (V178A, V109A), and relatively no change in the binding free energy (within ~0.1 kcal/mol: R208K, R211S). These values are plotted in Figure 2.5 and listed in Table 2.1. Overall, with the exception of R211S, the calculated values are in good qualitative agreement with the experimental mutagenesis data which contributes to the validity of our open state Hv1 model. Between the two 2GBI tautomers, the results for gbi2 are more consistent with experiment. The mean absolute error (MAE) for gbi2 is 2.1 kcal/mol, compared to gbi1 with an MAE of 3.2 kcal/mol, which suggests that gbi2 may be the more prevalent species. Without consideration of R211S (discussed later), the MAE improves to 1.1 kcal/mol for gbi2 and 2.5 kcal/mol for gbi1.

Protein mutations which involve charged side chains (D112E, R208K, and R211S) are more

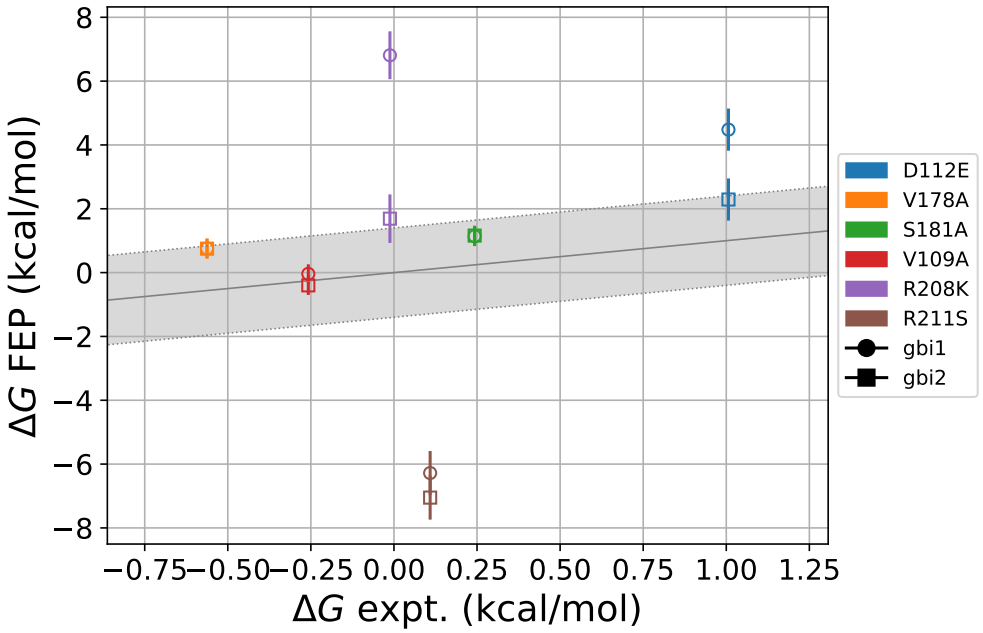


Figure 2.5: Relative binding free energies for the two 2GBI tautomers and six protein mutations considered. The error bars represent the root sum square values of standard errors from the Bennett acceptance ratio for the apo and holo states. The shaded region represents ±1.4 kcal/mol from the experimental value, representing a 10-fold change in the equilibrium constant. Alternate tautomers use different symbols. The markers for V178A (orange) and S181A (green) overlap for gbi1 and gbi2. Computed relative binding free energies are in qualitative agreement with experimental data, especially for gbi2.

challenging for which to compute binding free energies compared to mutations having neutral residue end states. D112E involves charge movement by the addition of a methylene group, R208K changes the charge density of the cationic region,[10] and R211S changes the charge of the system from neutral to -1 *e*. There are greater disparities in the relative binding free energies of gbi1 and gbi2 for these three mutations, with gbi2 values generally closer to experiment.

Compared to WT, the reported experimental value for the D112E mutation leads to a change in binding free energy of +1.01 ± 0.05 kcal/mol. In our simulations, this mutation induced a shift in the gbi1 ligand orientation after ~5 ns as gbi1 rotates to prevent crowding of the glutamate residue and to maximize interactions of its guanidine moiety (Figure 2.6). Lengthening of the side chain of residue 112 appears to encroach on the binding site and affect the native hydrogen bonding interactions between the positively-charged 2GBI and the nearby acidic side chains of residues 112 and 185 (see supporting information: Figure A.8). This crowding may explain the unfavorable binding represented by the relative binding free energy.

Of the mutations considered here, the V178A mutation is most favorable for binding 2GBI, with an experimental relative binding free energy of -0.56 ± 0.05 kcal/mol compared to wild type. In contrast, our predicted value was between 0.7 and 0.8 kcal/mol for both tautomers,

mutation	expt	gbi1	gbi2
D112E	1.01 ± 0.05	4.9 ± 0.7	2.3 ± 0.7
V178A	-0.56 ± 0.05	0.8 ± 0.3	0.8 ± 0.3
S181A	0.24 ± 0.05	1.1 ± 0.3	1.1 ± 0.3
V109A	-0.26 ± 0.06	0.0 ± 0.3	-0.4 ± 0.3
R208K	-0.01 ± 0.05	6.8 ± 0.8	1.7 ± 0.8
R211S	0.11 ± 0.05	-6.3 ± 0.7	-7.1 ± 0.7

Table 2.1: Binding free energies of 2GBI (gbi1 and gbi2) to Hv1, for mutant Hv1 compared to wild type Hv1. The uncertainties represent the root sum square values of standard errors from the Bennett acceptance ratio for the apo and holo states. The data presented here is also plotted in Figure 2.5.

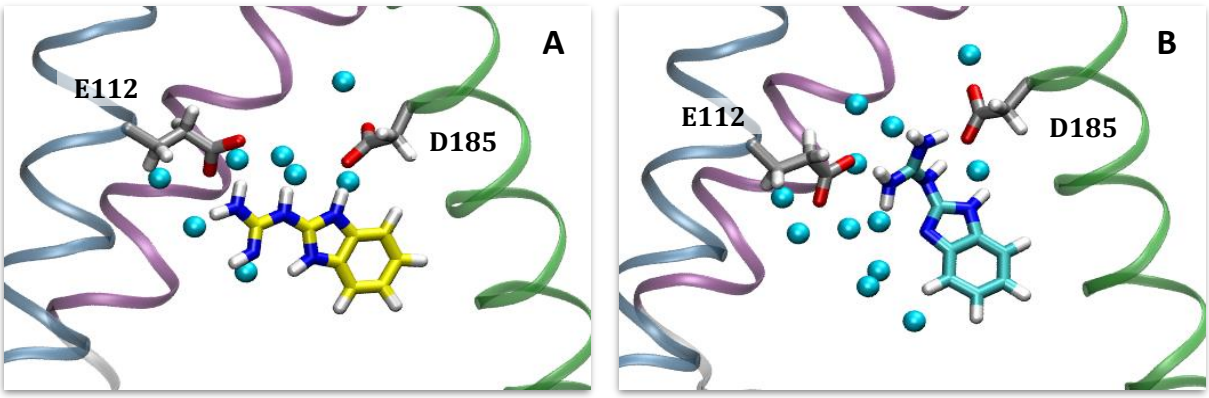


Figure 2.6: 2GBI bound to the human Hv1 D112E mutant in the (**A**) gbi2 and (**B**) gbi1 systems, rendered in VMD.[135] The ligand and surrounding residues are displayed in licorice representation, and the oxygen atoms of waters within 3 Å of the ligand are shown as light blue spheres. Concerning hydrogen bonding interactions between 2GBI and the neighboring acidic residues E112 and D185, the tautomer gbi1 does not contain a protonated imidazole region as in gbi2 (see Figure 2.1). It seems for this reason that gbi2 maintains its original binding pose after mutation, while gbi1 reorients in order to maximize the interactions of its more positively charged guanidine moiety with both E112 and D185.

indicating binding becomes less favorable. We wondered if the experimental value reflected additional protein rearrangements not captured in the timescale of our MD simulations. To explore this further, we focused on a nearby aromatic residue, F182, in proximity to both V178 and the benzo moiety of 2GBI. This residue has been found to play a role in 2GBI binding via π-stacking interactions.[338] We hypothesized that F182 may reorient itself, occupying some of the space from the mutation of valine to alanine, in order to better interact with the bound ligand. We evaluated the energetic contribution of this putative conformational change by computing the potential of mean force for rotating the F182 χ_1 dihedral angle. Our results did not support this theory but rather suggested an energetic cost of around 8 kcal/mol for the F182 aromatic ring being in a parallel configuration with the 2GBI aromatic moiety (see supporting information: "Potential of mean force calculation for F182 rotation after V178A mutation").

2.3.3 Hysteresis in free energy calculations involving "floppy" residues

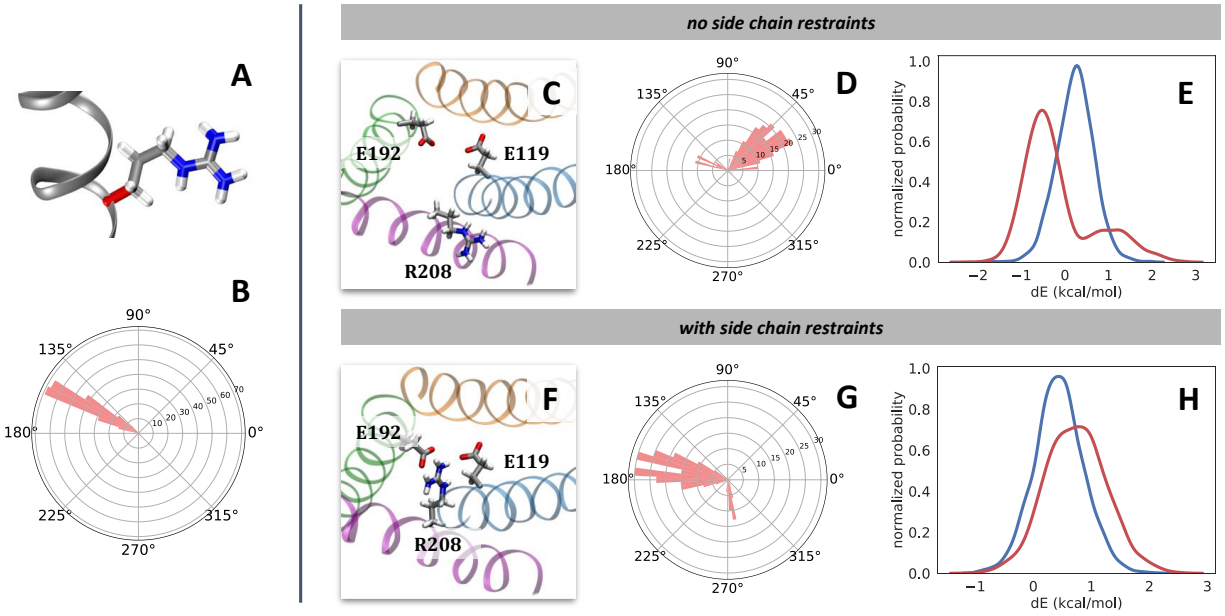


Figure 2.7: Correlation of the mutating arginine χ_1 dihedral angle to the work overlap in the R208K transformation. Without side chain restraints, χ_1 takes unexpected values, causing the side chain to point outside the Hv1 channel and leading to poor overlap of forward and reverse work values. (**A**) Arginine χ_1 dihedral angle drawn in red. (**B**) Histogram in polar coordinates of dihedral angle values belonging to R208 χ_1 during a normal, fully-interacting MD simulation. (**C**) View of the channel from the extracellular side. The incoming K208 residue is not shown for clarity. (**D**) Polar histogram of dihedral angle values belonging to R208 χ_1 during alchemical free energy calculation of R208K. The angle values are plotted from the $\lambda = [0.400, 0.425]$ window. (**E**) Overlap of forward (blue) and reverse (red) work values for the $\lambda = [0.400, 0.425]$ window in the R208K mutation of the apo protein. Both R208 and K208 partially interact with the system as per the alchemical potential. (**F, G, H**) Analogous to C, D, E, respectively, but with the addition of flat-bottom distance restraints.

One way that we evaluate convergence in our free energy calculations is by analyzing the histogram overlap of work values from the forward and reverse processes.[238] We observe that mutations involving flexible and charged side chains, such as arginine and lysine, are more prone to have poor overlap, signifying inadequate sampling of overlapping regions of phase space during the alchemical transformations. This presumably arises from the scaling of non-bonded interactions. Towards the start of the transformation (λ close to 0), the incoming

atoms are mostly decoupled from, or not interacting with, the rest of the system. Similarly, at states where λ is close to 1 near the end of the alchemical transformation, the outgoing atoms are decoupled from the rest of the system. The decoupled nonbonded interactions may "free" the flexible, charged side chains to adopt various nonphysical conformations.

We illustrate this hysteresis in context of the R208K Hv1 mutation, focusing on the arginine residue. In wild type Hv1, the measure of the R208 χ_1 dihedral angle (Figure 2.7A) averaged around 154° in equilibrium simulations (histogrammed in Figure 2.7B), and R208 is in close proximity to the acidic side chains E119 and E192. However, during free energy calculations with the scaled nonbonded interactions, the side chain pivots such that R208 is artifactually outside the channel (Figure 2.7C), and the χ_1 dihedral angle values are inconsistent with the range expected from wild type Hv1 simulations (Figure 2.7D). The lack of fully coupled van der Waals and electrostatic interactions of the floppy, charged residue with the rest of the system interrupts native ionic interactions and leads to unnatural wandering of the side chain in simulations of these artificial intermediate states. The sampled unphysical configurations were directly correlated with poor overlap between the forward and reverse transformations, evidenced here by bimodal histograms of the work values from the forward and reverse transformations (Figure 2.7E). Similar observations were made for the lysine χ_1 dihedral angle.

We reduced this hysteresis by imposing a series of flat-bottom distance restraints with nearby interacting acidic residues (see Methods and supporting information: "Flat-bottom restraints for mutations involving Arg or Lys residues"). This yielded fewer abnormal side chain configurations (Figure 2.7F-G) and better overlap of the forward and reverse work (Figure 2.7H).

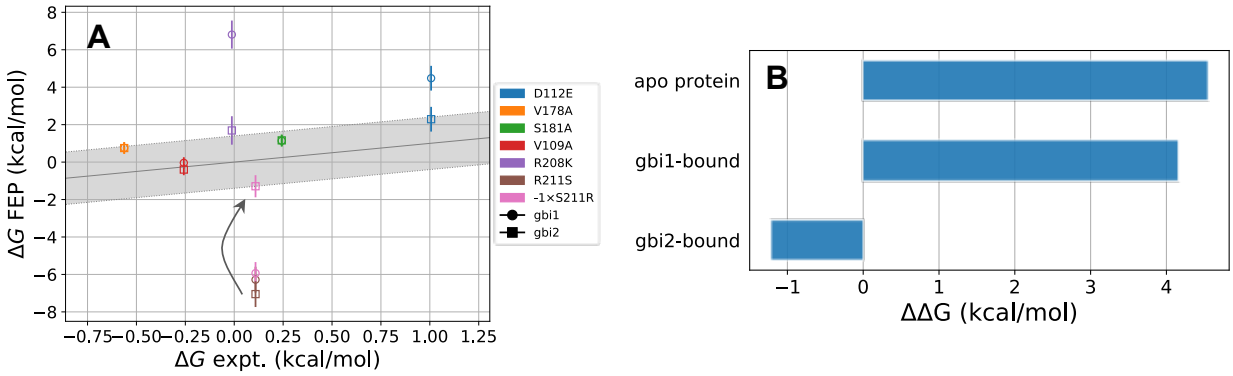


Figure 2.8: Calculated relative binding free energies for the mutation of residue 211 between the arginine and serine end states reveal discrepancies based on the starting structure. (**A**) Computed relative binding free energies, analogous to Figure 2.5, with the addition of the S211R mutation. The pink markers show the negative of the S211R values to compare on the same scale as R211S. The R211S relative binding free energies ideally should be the same as those for $-1\times$ S211R. The error bars represent the root sum square values of standard errors from the Bennett acceptance ratio for the apo and holo states. (**B**) Bar plot showing the change in the computed ΔG_3 (apo protein) or ΔG_4 (holo protein) of individual processes of the thermodynamic cycle (Figure 2.2) for the R211S/S211R mutation. The apo protein value represents $\Delta G_{3,\,R211S} - (-1 \times \Delta G_{3,\,S211R})$, and the analogous expression is applied to the ΔG_4 values of gbi1 and gbi2. This value should be zero since free energy is a state function which is independent of the path taken.

2.3.4 Further exploration on R211S

The computed binding free energies of R211S are least consistent with experimental results. Our results were predicted to be overly favorable by -6.3 kcal/mol for gbi1 and -7.1 kcal/mol for gbi2 (the experimental reference value is 0.1 kcal/mol). We considered whether a potential contribution may have been the change in total charge of the system. The mutation from a positively charged arginine to a neutral serine changes the system's net charge from neutral to -1 *e*. Our simulations employed the particle-mesh Ewald (PME) method for treatment of long-range electrostatics in the periodic system. However, PME enforces neutrality by introducing a uniform neutralizing background charge. Therefore, with the non-neutral end state, the calculated free energy not only has the expected contribution of turning off the arginine charge but also undesired contribution for turning on the neutralizing background charge. We corrected for the electrostatic finite-size effects by applying an analytical correction based on the Poisson-Boltzmann continuum electrostatics model in CHARMM (version c40b1).[257, 30] The corrections to the binding free energy were not trivial when considered individually (apo = -3.1042 kcal/mol, gbi1 = -3.1030 kcal/mol, gbi2 = -3.1019 kcal/mol). However, they become negligible when considering relative binding free energies. As a result, we believe that the change in net charge is not the primary reason for poor agreement of the experimental and computed R211S relative binding free energies.

We investigated the possibility of unconverged conformational sampling as a cause for the R211S outlying behavior. We performed the same transformation in the opposite direction, i.e., preparing the system with S211 and perturbing to R211. The S211 structures for the apo, gbi1-bound, and gbi2-bound states came from the end of the respective R211S transformations. Each configuration was equilibrated for 5 ns before the S211R free energy calculations. We maintained the same λ schedule and simulation parameters as described earlier.

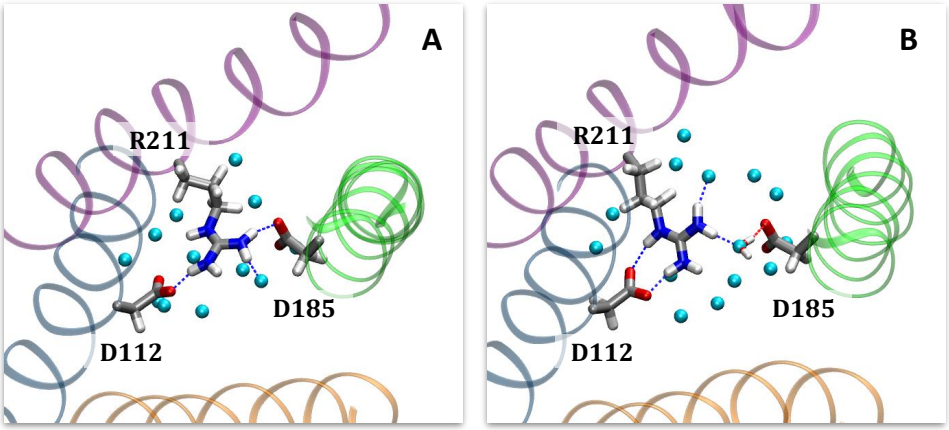


Figure 2.9: View from the extracellular side of the apo channel showing representative interaction networks around R211 (**A**) before and (**B**) after simulating a null transformation of R211 to S and back to R. Oxygen atoms of waters within 2 Å of the R211, D112, and D185 are colored as light blue spheres. The interaction differences between (**A**) and (**B**) reveal potential sampling barriers in what ought to be equivalent systems. Figure rendered using VMD.[135]

We consider the negative value of S211R in order to adequately compare the free energy change in the same direction, and we subsequently refer to this mutation as $-1 \times$ S211R. The free energies for R211S and $-1 \times$ S211R should be the same ideally, as free energy is a state function which is independent of the path or direction taken between two states. The computed $-1 \times$ S211R binding free energies for gbi1 and gbi2 are -5.9 kcal/mol and -1.3 kcal/mol respectively (Figure 2.8A, brown vs. pink). In this plot, gbi1 seems to be more consistent between R211S and $-1 \times$ S211R. However, the plotted values are relative binding free energies where the apo value is subtracted from each of the gbi1 and gbi2 values (see the thermodynamic cycle in Figure 2.2). Here, the ΔG_3 of apo and the ΔG_4 of gbi1 increase by about 4 kcal/mol between R211S and $-1 \times$ S211R, but the ΔG_4 for gbi2 only changes by about 1 kcal/mol between R211S and $-1 \times$ S211R (Figure 2.8B).

We posit that this discrepancy may arise from differences in R211 interactions with other protein residues as well as differences in hydration patterns in the local region. In simulations of wild type Hv1, R211 is a hydrogen bond donor to two aspartate residues, D112 and D185 (Figure 2.9A). These interactions are disrupted when the arginine is mutated to serine,

and there is a slight influx of waters to the mutation site to stabilize D185. The S211R mutation reintroduces arginine, but R211 does not reform native contacts with D185, now surrounded by the water molecules (Figure 2.9B). This may occur if the system is slower to dispel waters from the pore than it is to allow waters inside, especially near the buried region of the mutated residue (Figure 2.10). Hence the underlying end states of R211 are not the same for the R211S and S211R mutations, potentially due to insufficient rearrangement of the surrounding solvent network. In other words, non-overlapping ensembles are being sampled in the R211S and S211R transformations, precluding reasonable comparison of these calculations with experimental mutagenesis results. Thus, our results for this transformation are reported here, but the apparent agreement between $\Delta\Delta G_{calc}$ and $\Delta\Delta G_{expt}$ for gbi2 and the $-1 \times$ S211R mutation (Figure 2.8A) cannot be used meaningfully to assess the validity of our structural model due to the difficulties with this particular calculation.

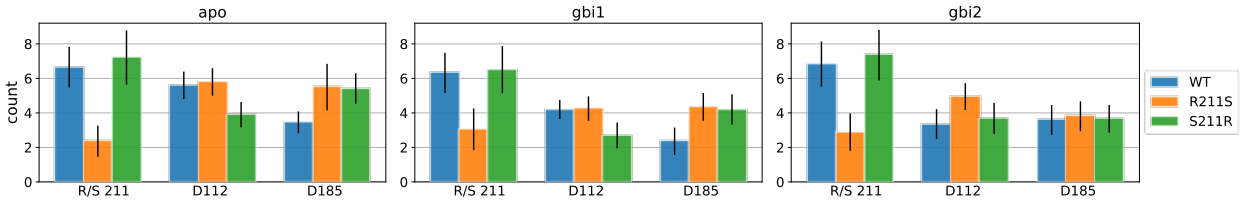


Figure 2.10: Average number of water molecules within 3 Å of the mutating residue (Arg or Ser 211), D112, and D185. The error bars represent standard deviations. Hydration patterns are not always consistent in the vicinity of R211 during the R211S and S211R mutations. For example, in the apo state, the number of water molecules near D185 increases (left subplot, right group, blue → orange bars). When mutating the system with S211 back to R211, the number of water molecules decrease near D112 instead of near D185 (left subplot, middle group, orange → green bars). It might be expected that the blue bars (wild type protein with R211) and the green bars (twice mutated protein also with R211) should be the same in all three systems. Between the two R211 states of before and after mutation, both the apo and gbi1 systems differ by 1-2 water molecules in the specified vicinity of D112 and D185, whereas gbi2 is more consistent in hydration around the two acidic residues.

2.4 Conclusions

In this study, we employed free energy calculations with atomistic molecular dynamics simulations to characterize the binding of 2GBI to Hv1 as a foundational step to optimize small molecule blockers of Hv1. Our molecular insights of the binding mechanism of 2GBI to Hv1 agree with double-mutant cycle experiments, helping to validate our structural model for the channel and the bound state.

The mutations that were most computationally challenging were those involving charged side chains: D112E, R208K, and R211S. The R211S calculation might be further examined using enhanced sampling methods such as REST2,[141, 142, 320] with the inclusion of protein side chains and water molecules in the vicinity of R211.

Results from alchemical free energy calculations show that both the gbi1 and gbi2 tautomers are comparable in most relative binding free energies calculated. The gbi1 structure is limited in its extent of forming protein-ligand contacts without the protonated imidazole region. We reason that gbi2, the tautomer with excess charge in the imidazole region, is likely the primary state to bind Hv1. We hope to investigate the extent to which free energy calculations may be used to design new compounds based on the 2GBI scaffold, particularly with an imidazole-based pharmacophore with substitutions at the *2*-position that stabilize ring protonation, for more effective binding to Hv1 for potential therapeutic benefit.

2.5 Acknowledgments

The authors appreciate helpful discussions from Christopher Bayly on the tautomeric state of 2GBI and from Wei Jiang on water molecules in the binding site. The authors are grateful for computing support from Nathan Crawford and the UCI GreenPlanet cluster, supported

in part by NSF Grant CHE-0840513. VTL and NML acknowledge funding from the National Science Foundation Graduate Research Fellowship Program. The authors appreciate financial support from the National Institutes of Health (grant GM108889 to DLM, and GM098973 to FT).

2.6 Author contributions statement

VTL conducted the research, prepared the figures, and wrote the manuscript. ADG and NML contributed to setup of the system and molecular simulations. JAF contributed to analysis of the results. FT provided insight on results and research directions. DLM and DJT jointly supervised the work. All authors reviewed the manuscript.

Chapter 3

Potential of mean force calculations for the membrane permeation of Hv1 channel blockers

The voltage-gated proton channel Hv1 mediates efflux of protons from the cell. Hv1 integrally contributes to various physiological processes including pH homeostasis and the respiratory burst of phagocytes. Inhibition of Hv1 may provide therapeutic avenues for the treatment of inflammatory diseases, breast cancer, and ischemic brain damage. In this work, we investigate two prototypical Hv1 blockers, 2-guanidinobenzimidazole (2GBI) and 5-chloro-2-guanidinobenzimidazole (ClGBI), from an experimentally screened class of guanidine derivatives. Of these two compounds, 2GBI demonstrates inhibitory effect only on intracellular application, whereas ClGBI is thought to permeate the cellular membrane to access the intracellular side of the channel. We compute the potential of mean force for each compound to partition into the membrane using atomistic molecular dynamics simulations with the adaptive biasing force method. Our results rationalize the putative distinction between these two blockers with respect to their abilities to permeate the cellular membrane.

3.1 Introduction

Hv1 is a voltage-gated proton channel whose primary function is the release of excess protons from the cell. Human Hv1 is a homodimer, and each monomer is composed of four transmembrane alpha helices forming the voltage-sensing domain (VSD) which also serves as the proton conduction pathway.[262, 243, 242] Hv1 integrally contributes to various physiological processes including pH homeostasis and the respiratory burst of phagocytes.[77] Inhibition of Hv1 may provide therapeutic avenues for the treatment of inflammatory diseases,[270] breast cancer,[322, 323] and ischemic brain damage during stroke.[334]

Guanidine derivatives have been found to block Hv1 in the VSD in its open state.[309, 129] Two molecules in this class are 2-guanidinobenzimidazole (2GBI) and 5-chloro-2-guanidinobenzimidazole (ClGBI). In order to be useful as therapeutics, these small molecule inhibitors should be able to permeate the cellular membrane as research suggests that binding occurs from the intracellular side of the channel.[129, 128] ClGBI differs from 2GBI by the replacement of a hydrogen atom with a chlorine atom on the benzene moiety. This structural modification may affect the ligand's pharmacokinetic properties. 2GBI demonstrates inhibitory effect only on intracellular application, whereas ClGBI can inhibit Hv1 when present in the extracellular medium. It is hypothesized that ClGBI permeates the cellular membrane in order to access the intracellular side of the channel.[129, 128]

Membrane permeability rates have not been experimentally determined for 2GBI or ClGBI, but the permeation process can be examined using molecular simulations. These simulations typically have an advantage over experimental techniques in that they can provide spatially resolved information with atomistic resolution. Molecular dynamics (MD) simulations enable investigation of transport processes at the molecular level to describe and explain important steps including membrane partitioning and molecular flip-flop across the hydrophobic membrane core.[67]

The inhomogeneous solubility-diffusion model[66, 190] describes the permeation process derived from the steady-state flux of solute molecules through the membrane.[228] The permeability P_m of some solute to cross a membrane of thickness L is expressed as:

$$\frac{1}{P_m} = \int_{-L/2}^{L/2} \frac{\exp W(z)/k_B T}{D(z)} \, \mathrm{d}z, \tag{3.1}$$

where $W(z)$ is the potential of mean force (PMF), $D(z)$ is the local diffusion constant, k_B is the Boltzmann constant, T is the temperature, and z is a collective variable which describes the position of the solute along the transmembrane axis. Our focus in this work is on calculating the potentials of mean force for various Hv1 channel blockers, given the exponential dependence of the permeability coefficient on the PMFs, to gain understanding of their relative permeabilities from a thermodynamic standpoint.

We aimed to investigate the propensity of ClGBI to permeate the lipid bilayer, shedding light on why it may uniquely be able to block Hv1 from the extracellular compartment. We employed atomistic molecular dynamics simulations to compute the potentials of mean force for 2GBI and ClGBI to partition into the membrane using the adaptive biasing force (ABF) enhanced sampling method,[58, 122] which applies a continually updated biasing force to effectively yield a flat free energy surface in the long time limit. ABF has been demonstrated to reproduce experimental membrane permeability trends.[171, 298] Our results rationalize the putative distinction between these two blockers on their abilities to permeate the cellular membrane to inhibit Hv1.

3.2 Methods

In this work, we compute and examine the one-dimensional potentials of mean force for 2-guanidinobenzimidazole (2GBI) and 5-chloro-2-guanidinobenzimidazole (ClGBI) to perme-

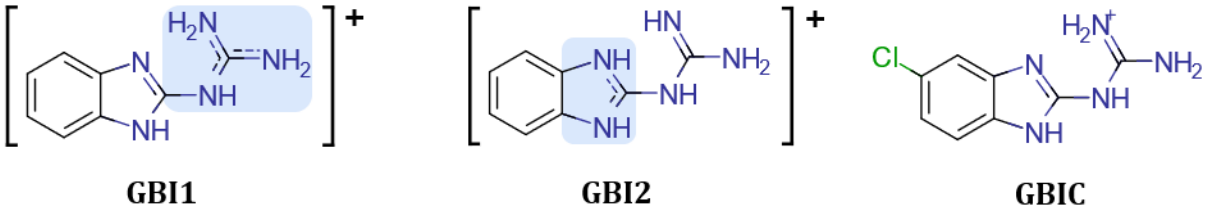


Figure 3.1: Chemical structures of molecules in this study. GBI1 and GBI2 refer to distinct tautomeric states of the positively charged molecule, 2-guanidinobenzimidazole. The shaded blue regions have higher positive charge density. GBIC refers to 5-chloro-2-guanidinobenzimidazole.

ate through a 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) bilayer. Chemical structures of the molecules that we study in this work are depicted in Figure 3.1. For 2GBI, we consider two tautomers, which we refer to as GBI1 and GBI2. Both GBI1 and GBI2 are positively charged but differ in the placement of one hydrogen atom, leading to different regions of partial charge concentration (denoted by the shaded blue regions in Figure 3.1). Both of these tautomers were previously investigated in the computation of relative binding free energies to Hv1,[177] and GBI2 was predicted to be to be the primary tautomeric state that binds to Hv1.

We obtained an initial configuration of a 72-lipid POPC bilayer from the end of a 35 nanosecond (ns) simulation at 303 K and 1 atm. This configuration was generated by the group of Jeff Klauda and is available at `https://terpconnect.umd.edu/~jbklauda/memb.html`. The membrane spanned the xy plane with the z-axis normal to the membrane plane. Using the CHARMM36 force field[153] for the lipid molecules with the modified TIP3P water mode,[143, 189] we further equilibrated the membrane for 849.6 ns using Desmond 2.4.[29] Short-range real-space interactions were cut off at 12 Å by using a force-based switching function between 8 Å and 12 Å. An r-RESPA algorithm[299] was employed to integrate the equations of motion with a time step of 4 fs for the long-range non-bonded forces, 2 fs for short-range non-bonded, and 1 fs for bonded forces. The particle mesh Ewald method[57, 78]

was used to treat long-range electrostatics. All bond lengths involving hydrogen atoms were constrained using SHAKE.[259] The simulations were performed at constant temperature (303 K) and pressure (1 atm), using Nose-Hoover chains[192] and the Martyna-Tobias-Klein barostat.[193]

Given that permeation events may affect membrane structure and to diminish the possibility of boundary effects, we extended the size of the membrane by duplicating it to form a 2x2 patch to yield a final system size of 288 POPC molecules (144 per leaflet) and 16,444 water molecules. The cell dimensions were 100 $\times$ 100 $\times$ 90 $Å^3$. Atomistic molecular dynamics simulations were then performed using the NAMD 2.13 software package.[235] The system was equilibrated for an additional 20 ns. The temperature was maintained at 295 K using the stochastic velocity rescaling (canonical sampling through velocity rescaling)[31] thermostat with a rescale period of 1 picosecond (ps) and 20 time steps between rescaling. The pressure was maintained at 1 atm using the Langevin piston method[82] with an oscillation period of 200 fs and a damping time scale of 100 fs. The O-H bond lengths in water were constrained using the SETTLE algorithm,[200] and all other covalent bonds involving hydrogen atoms were constrained using the RATTLE algorithm.[7] Long-range electrostatic interactions were evaluated using the particle mesh Ewald method[57, 78] with a grid spacing of 1.0 Å. Short-range van der Waals and electrostatic interactions were calculated using a cutoff of 9 Å with a switching function applied beyond 8 Å. Multiple time step integration was applied with timesteps of 6 fs for long-range electrostatic forces, 2 fs for short-range nonbonded forces, and 2 fs for bonded forces using the r-RESPA method.[299]

We used the same configuration of the equilibrated bilayer to prepare the three separate systems for GBI1, GBI2, and GBIC. We used the force fields for GBI1 and GBI2 from our previous relative binding free energies study.[177] The CGenFF parameters for GBIN, GBIC, and GBIN were obtained from the ParamChem server, `https://cgenff.umaryland.edu/`.[305, 306] We compared the energy minimized structure from the force field with that

of a gas phase MP2/6-31G* geometry optimization to ensure that we had comparable minimum energy structures. We also assessed the torsional energy profiles and refined dihedral parameters to yield more consistent results with quantum mechanical data (refined parameters in supplementary information). Each solute was added into the hydrated region, and we equilibrated the system for an additional 10 ns, maintaining the same MD simulation parameters.

We then performed MD simulations with the adaptive biasing force (ABF) method[58, 122] in NAMD. Using the `colvars`[83] module in NAMD, we defined a one-dimensional collective variable (denoted as z) to describe the relative position of the solute along the transmembrane axis, specified as the center of mass position of the solute relative to the center of mass position of the carbonyl carbon atoms of the POPC molecules. The collective variable was stratified into eleven overlapping windows whose distance boundaries are listed in Table 3.1 and depicted at the bottom of Figure 3.2. The window boundaries were maintained using upper and lower wall force constants of 10 kcal/mol/Å^2. The forces applied by ABF were collected with a bin width of 0.1 Å. A total of 1000 samples were collected per bin prior to application of ABF to avoid nonequilibrium effects in system dynamics. Each window was sampled separately for each permeant, at a minimum of 40 ns per window per permeant. The windows comprising the collective variable in bulk water (01, 11) were initiated from equilibrated MD simulations, and each subsequent window approaching the center of the bilayer was initiated from the previous overlapping window. Window 06 was sampled in duplicate, bidirectionally initiated from window 05 and window 07. Convergence was reached more quickly for windows in which the solute was only in bulk water ($|z|$ = 30 to 40 Å). Convergence took longer in both inhomogeneous environments (e.g., membrane headgroups) as well as in the hydrophobic interior of the membrane. The minimum, total amount of simulation time for each solute was at least one microsecond. For all molecules in aggregate, we simulated a total of 6.7 microseconds.

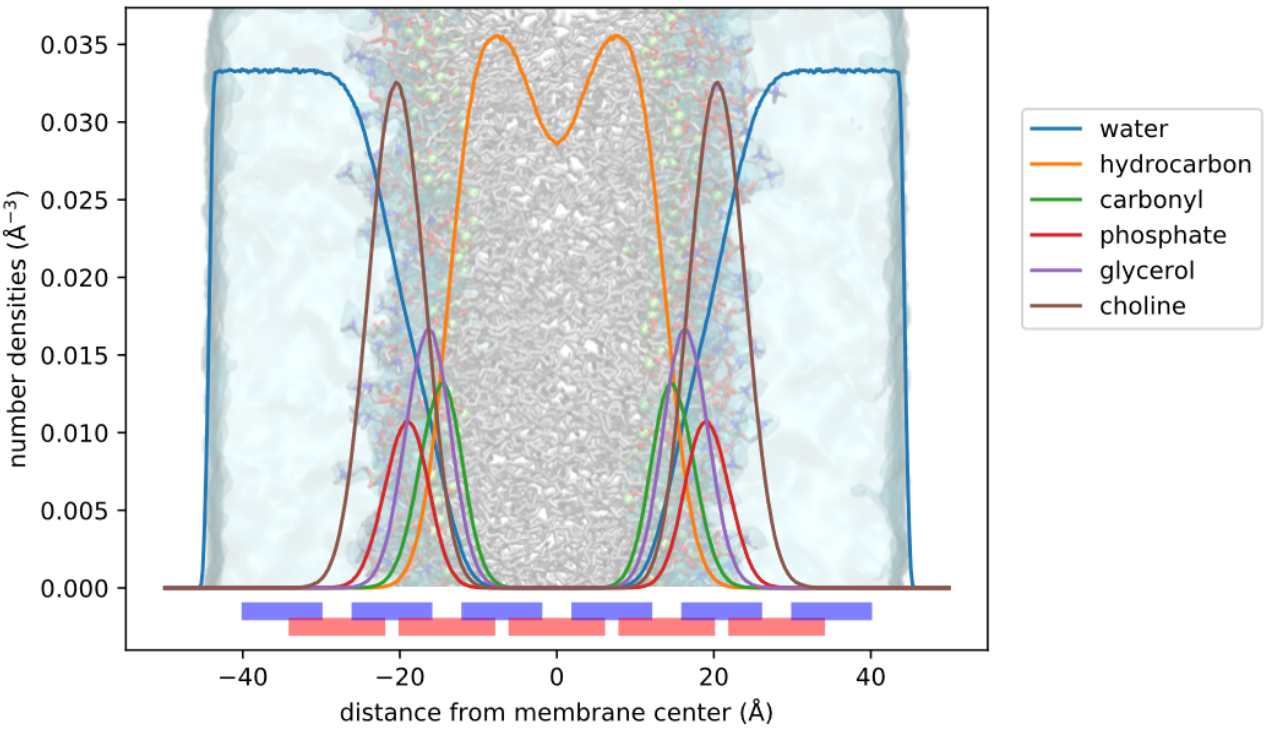


Figure 3.2: Number density profile of the hydrated POPC bilayer used in this work. The one-dimensional collective variable for adaptive biasing force simulations is defined as the distance between the z coordinate of the center of mass of the solute to the z coordinate of the center of mass of the carbonyl carbons of the POPC lipids. Waters are drawn in slabs of light blue, and lipids are in licorice representation with their carbonyl carbons shown as green vdW spheres. The collective variable is stratified into eleven overlapping windows whose distance boundaries are denoted in the red and blue rectangles and listed in Table 3.1.

Window	CV: lower bound (Å)	CV: upper bound (Å)
01	30	44
02	22	24
03	16	26
04	08	20
05	02	12
06	-06	06
07	-12	-02
08	-20	-08
09	-26	-16
10	-34	-22
11	-40	-30

Table 3.1: Separation of the collective variable (CV) into windows. The collective variable is defined by the difference of the z coordinate of the permeant center of mass and the z coordinate of the carbonyl carbon atoms of the POPC bilayer.

One of the key events of membrane permeation is solute "flip-flop" across the hydrophobic membrane core.[67] This flip-flop may occur over long timescales,[14] preventing adequate sampling of the interior hydrophobic region of the membrane (also see example in supplementary information). For this reason, we employed a bidirectional approach following the protocol of Li et al.[175] to compute membrane permeation, with the solute approaching the membrane center from both the upper and lower leaflets of the bilayer. For each leaflet of the membrane, we combined the gradients obtained from ABF simulations over successive windows by averaging the overlapping gradients between neighboring windows. The combined gradient data for each leaflet was integrated to yield the PMF corresponding to each half of the membrane. The PMFs were shifted such that the average value in bulk water ($|z|$ = 35 to 40 Å) is equal to zero kcal/mol. The upper and lower leaflet PMFs were then combined by the following equation:

$$W(z) = -k_B T \ln\{e^{-W_{\text{upper}}(z)/k_B T} + e^{-W_{\text{lower}}(z)/k_B T}\} \,. \tag{3.2}$$

Noting the symmetry of the bilayer, we symmetrized the resulting PMF by averaging about

the bilayer center. Error bars were calculated as the difference between the unsymmetrized and symmetrized PMFs.

Finally, from the potential of mean force $W(z)$, we computed the free energy difference for the small molecule to partition into the membrane from bulk water using the equation[63, 188, 311]:

$$\Delta G(\text{wat} \rightarrow \text{mem}) = -k_B T \ln \left\{ \frac{1}{z_2 - z_1} \int_{z_1}^{z_2} e^{-\frac{W(z)-W(z_1)}{k_B T}} \, \mathrm{d}z \right\}, \tag{3.3}$$

where z_1 and z_2 represent points in bulk water on opposite sides of the membrane. In general terms, the transfer free energy of a solute from one solvent to another is proportional to the log of the partition coefficient ($\Delta G(z) = -k_B T \ln K(z)$). The partition coefficient, $K(z)$, is a form of equilibrium constant describing the ratio of the concentrations of the solute in each solvent[310, 133, 16]—the two solvents here are the aqueous phase regions on opposite sides of the membrane. We don't simulate these concentrations directly but determine this equilibrium constant from integrating the exponential of the Boltzmann-weighted free energy along this coordinate (the potential of mean force). Error bars were estimated by propagating[169] the uncertainty values of the PMFs.

3.3 Results and Discussion

In this work, we used atomistic molecular dynamics simulations with adaptive biasing force to compute the free energy profiles for various Hv1 inhibitors to permeate a lipid bilayer (Figure 3.2). We rationalize the apparent increased ability of ClGBI to permeate the membrane, in contrast to 2GBI, primarily based on its energetically favorable state in the lipid headgroups.

The potentials of mean force (PMFs) for membrane permeation of GBI1, GBI2, and GBIC

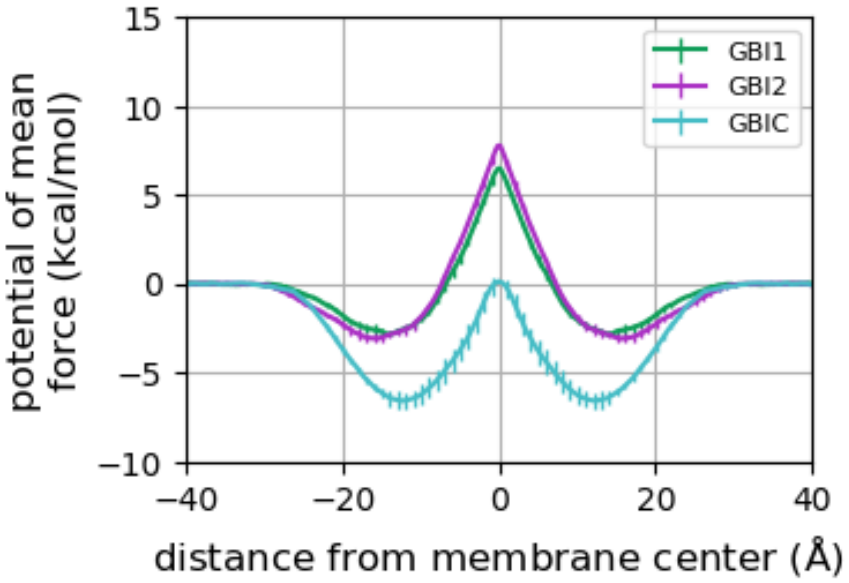


Figure 3.3: Potentials of mean force of some Hv1 channel blockers to permeate the membrane. Error bars signify deviation from symmetry.

are presented in Figure 3.3. Comparing the bilayer center ($z = 0$ Å) to bulk water ($z = 30$ Å), the two tautomers of 2GBI have highly similar PMFs, with GBI2 having a slightly higher barrier than GBI1. The profile for GBIC, in contrast, reveals more favorable energetics to membrane permeation. Comparing the barrier height energy differences of the PMFs (from the minimum to the maximum of the potentials), the results are (9.3 ± 0.4 kcal/mol) GBI1, (10.9 ± 0.5 kcal/mol) GBI2, and (6.8 ± 0.7 kcal/mol) GBIC. The PMFs show that the most energetically favorable position for the solute in the bilayer is generally in the range of $|z|$ = 15-20 Å. This corresponds to the regions of POPC headgroups, including the negatively charged phosphate groups (Figure 3.2).

We also computed the water-membrane partitioning free energies by integrating the Boltzmann-weighted PMFs (equation 3.3). These values, shown in Table 3.2, suggest a greater likelihood for GBIC to permeate the membrane compared to GBI1 and GBI2. Molecular dynamics simulations based on the inhomogeneous solubility-diffusion model have previously been found to overestimate permeability by several orders of magnitude relative to experi-

Permeant	$\Delta G(wat \rightarrow mem)$ (kcal/mol)
GBI1	-1.79 ± 0.02
GBI2	-2.13 ± 0.02
GBIC	-5.49 ± 0.04

Table 3.2: Partitioning free energies estimated from the potentials of mean force. Error bars obtained by propagation of uncertainties.

mental values, so we do not propose our results to be quantitative estimates of membrane permeation. Complicating factors include membrane composition, subdiffusive behavior, insufficient sampling, and force field quality.[171, 53, 254, 186, 308, 298, 43] We take these results more qualitatively of ClGBI's apparent increased propensity to permeate the membrane bilayer compared to 2GBI.

Due to interactions with the membrane, a translocating molecule may influence the structure of its local environment in the lipid bilayer, potentially leading to membrane deformation.[92, 14, 272, 51, 308, 266] To obtain a general sense of how the solute affects the bilayer from permeation, we calculated the two-dimensional electrostatic potential[3] of the POPC bilayer averaged over simulation snapshots in which the molecule was located at the center of the membrane ($|z| \leq 1$ Å). The potential is shifted relative to the bulk water. In these electrostatic potential plots, shown in Figure 3.4, the computed electrostatic potential spans a range of 800 mV. Panels A-C show a considerable amount of membrane distortion effects to a similar degree among GBI1, GBI2, and GBIC. Panel D shows a reference electrostatic potential plot of the membrane with the solute in bulk water without influence on the membrane structure. Panel E depicts a representative snapshot from MD simulations in which the solute remains coordinated with the membrane and solvated by water molecules as it permeates to the hydrophobic bilayer interior. These molecules are positively charged and are able to interact with the negatively charged phosphate groups of the POPC lipid molecules. These electrostatic interactions appear to be the driving force of membrane deformation, although it has been previously noted that even a permeating water molecule can cause

distortion of the bilayer.[51]

To gain a more specific understanding of solute interactions as it translocates across the bilayer, we computed hydrogen bond counts averaged as a function of distance from the membrane center (Figure 3.5). Error bars denote standard deviations. These results reveal that, at the bilayer center, GBI2, and to a lesser extent GBI1, forms a greater number of hydrogen bonding interactions with water compared to GBIC (compare the volumetric maps representing the solvation shell structures in the bottom of Figure 3.5). Concurrent with inducing membrane deformation, a charged solute in the bilayer will also tend to be coordinated by water molecules.[63, 312, 73, 86, 266] GBI2's higher average hydrogen bond count towards the center of the bilayer compared to GBIC may be correlated with the somewhat sharper membrane deformation of GBI2 as evidenced in the electrostatic potential plots.

Also notable in Figure 3.5 is that GBIC has the highest average number of hydrogen bonding interactions with POPC phosphate groups within 20 Å of the bilayer center (see also solvation shell structure of GBIC). Both GBI1 and GBIC have a more concentrated charge region in the peripheral guanidine moiety (Figure 3.1), which likely interacts more with the negatively charged phosphate groups. However, GBI1 does not have the same level of interactions to the phosphate groups as GBIC. We theorized that GBIC's increased interactions with phosphate may be influenced by the dispersion interactions of its chlorine atom with the POPC hydrocarbon tails, leading to a stronger directional preference of the solute relative to the bilayer. We next turn our focus to directional preferences of each solute in the bilayer.

We evaluated the effect of position, normal to the bilayer, on each solute's orientational preferences. We defined a molecule vector whose origin is one of the two distal carbon atoms on the benzene moiety and whose endpoint is the nitrogen atom that bridges the imidazole and guanidine moieties. We computed the cosine similarity between this vector and the axis normal to the membrane to assess the orientation of the molecule as a function of position

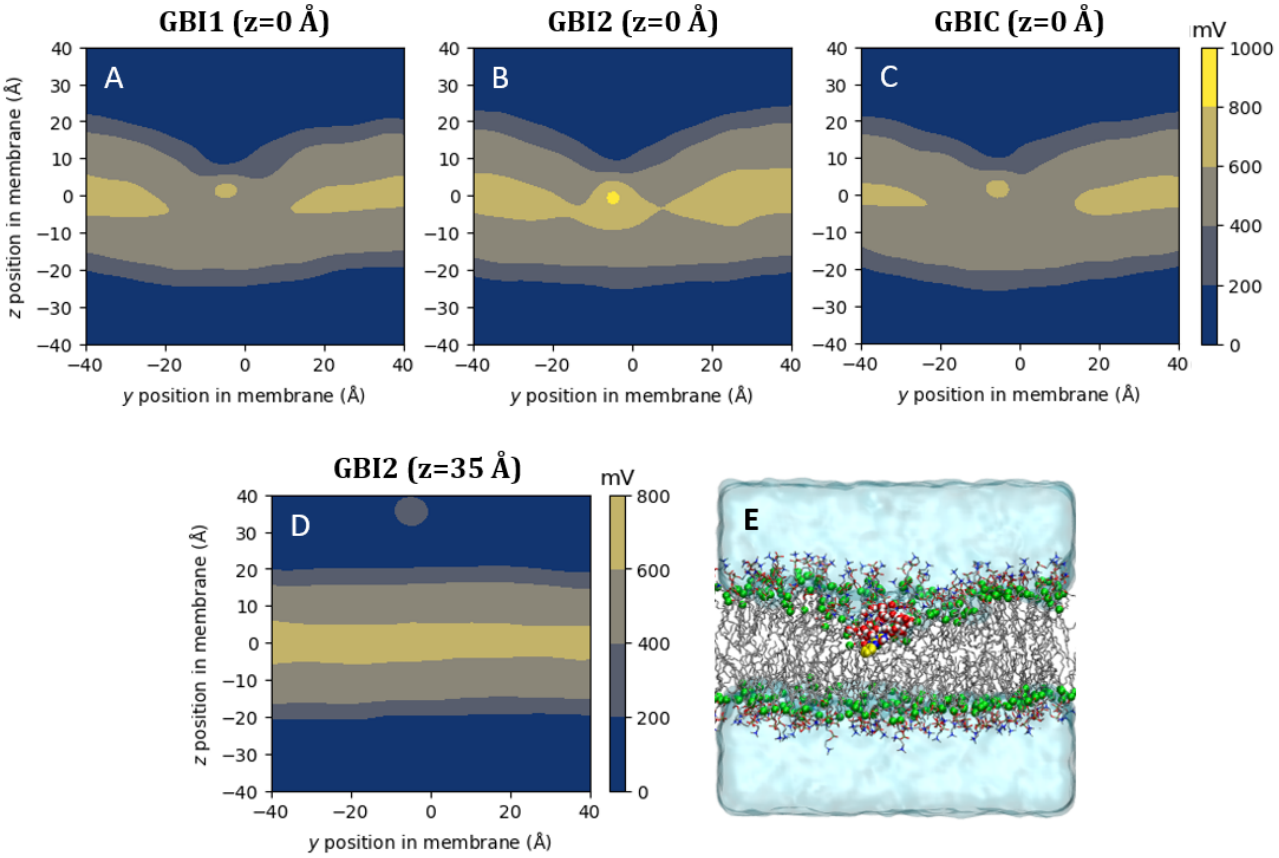


Figure 3.4: Contour plots representing the averaged electrostatic potentials of the membrane responding to molecular permeation for the solute at position $|z| \leq 1.0$ Å. Panels **A**, **B**, and **C** show the system with GBI1, GBI2, and GBIC, respectively. Panel **D** shows a representative plot of the average electrostatic potential when the solute is in bulk water (here, GBI2 at $z = 35$ Å). Membrane permeation leads to considerable distortion for all molecules. Panel **E** shows a typical configuration snapshot of the membrane system corresponding to GBI2 located at $z = 0$ Å. Waters are drawn in slabs of light blue, and lipids are in licorice representation with their carbonyl carbons shown as green vdW spheres. GBI2 is shown in vdW representation, with carbon atoms in yellow. Water molecules within 8 Å of GBI2 are shown in red (oxygen) and white (hydrogen) vdW spheres.

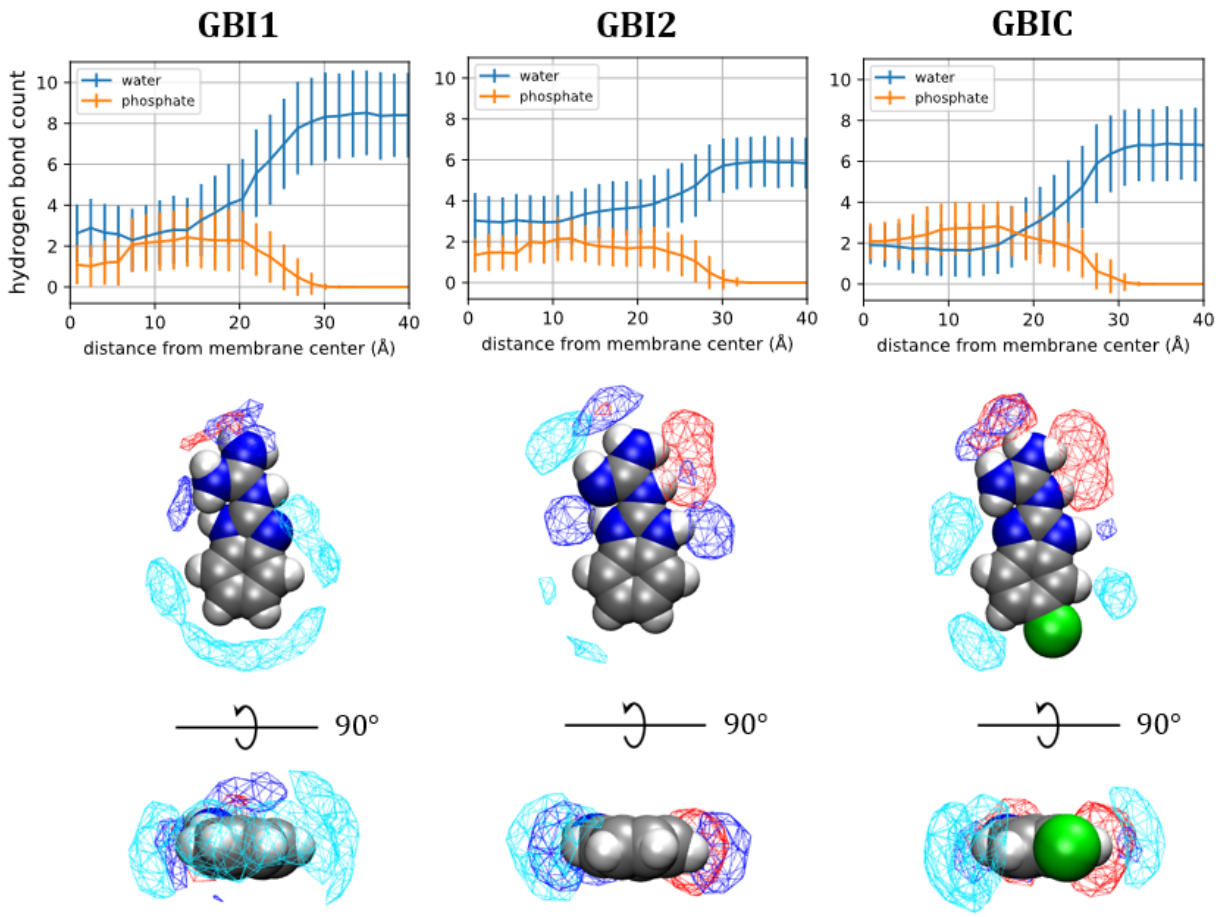


Figure 3.5: **Top**: Average number of hydrogen bonds formed by the solutes with water molecules (blue) and phosphate groups (orange) as a function of the solute center of mass position. **Bottom**: Evaluation of the local environment of the solute being in the interior of the membrane ($|z| \leq 1.0$ Å). We show the solvation shell structures of GBI1, GBI2, and GBIC for the surrounding water molecules, POPC phosphate groups, and POPC hydrocarbon chains. The solvation shells were computed using the oxygen atom of water molecules, oxygen atoms of the phosphate groups, and carbon atoms of the hydrocarbon chains. The surfaces represent equal number density for each group: water (blue, 0.013 Å^{-3}), phosphate (red, 0.008 Å^{-3}), and hydrocarbon (cyan, 0.0023 Å^{-3}).

along the membrane normal. The cosine similarity measures were evaluated in each of the simulation windows and histogrammed in Figure 3.6. Each row of the figure represents a predefined range of the collective variable sampled by the solute. Symmetry of the bilayer leads to antisymmetric orientation results for the solute, so we combined the upper and lower leaflet data by taking the negative of the cosine similarity measures of the lower leaflet.

We interpret the orientation plots from the direction of the molecule being in bulk water approaching the center of the lipid bilayer. The top row of Figure 3.6 shows uniform distributions from -1 (guanidine pointing towards bilayer center) to +1 (guanidine pointing away from bilayer center), indicating that molecules in bulk water have no orientational preference as they are in an isotropic medium. As the solute approaches the bilayer (row 2 of Figure 3.6), there are marginally higher counts towards the -1 direction. This is in the region in which the solute begins to approach the lipid headgroups (Figure 3.2) and will likely orient such that the guanidine interacts with the headgroups upon initially partitioning. This seems to be emphasized in the $|z|$ = 16-26 Å region. At some point, the molecule reorients, flipping around such that as it permeates to the interior of the membrane beyond the headgroups, its positively charged regions can interact with the negatively charged phosphates, while the more hydrophobic benzene moiety is solvated by the lipid hydrocarbon chains (rows 4-6 of Figure 3.6). The orientational preference is present but least strong for GBI2, possibly due to greater solvation by surrounding water molecules towards the membrane center (Figure 3.5). On the other hand, GBI1 or GBIC with a more positively charged guanidine tail (as opposed to a more positively charged imidazole center) may be more likely to exhibit orientational partiality such that the guanidine region points out away from the bilayer center and the benzene moiety is solvated by the hydrophobic lipid tails. Between GBI1 and GBIC which are nearly identical save for the chlorine atom of GBIC, GBIC has a much stronger orientation bias with its guanidine moiety pointing towards bulk water away from the hydrophobic interior.

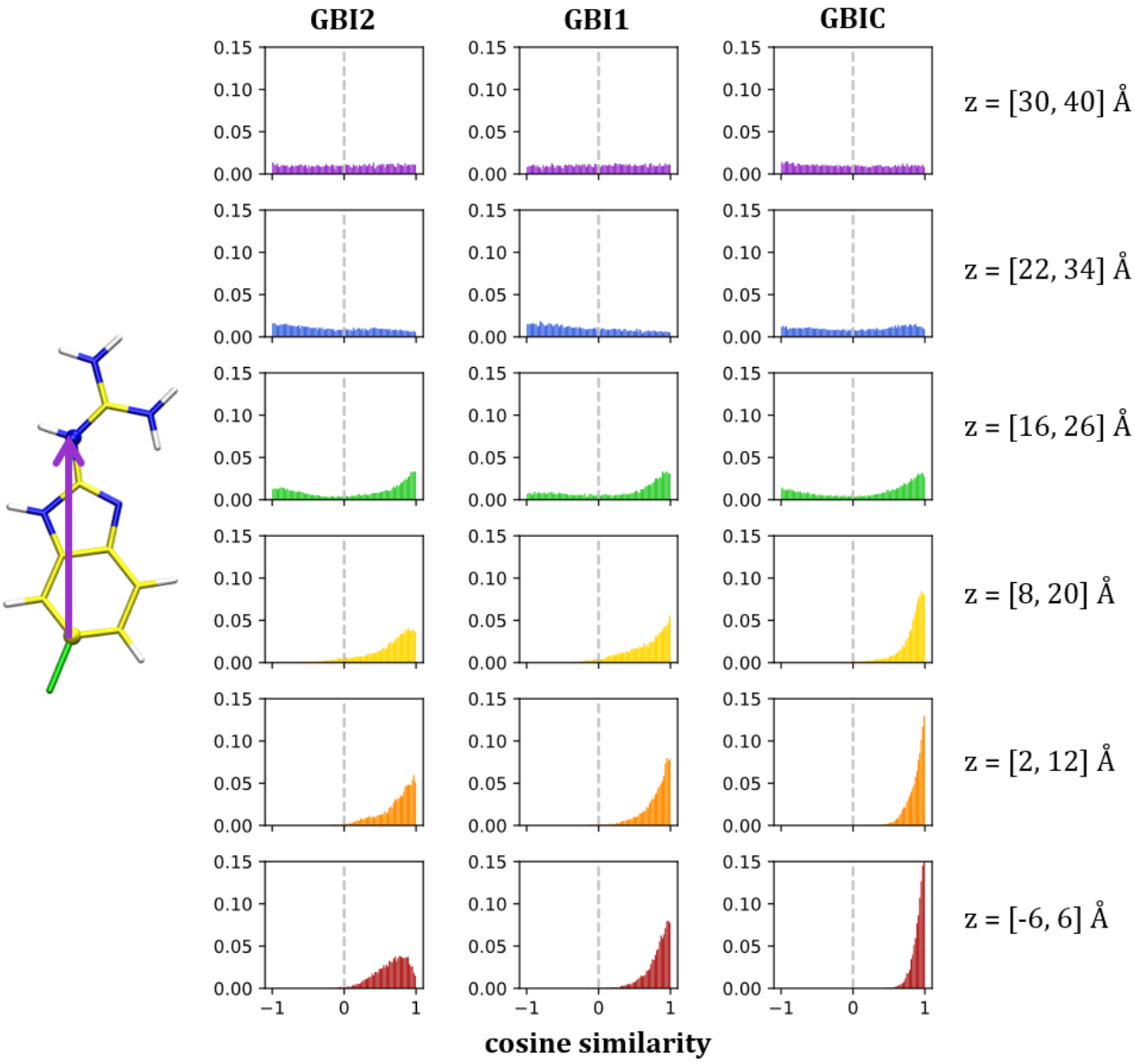


Figure 3.6: Histograms of molecule orientation distinguished by location along the lipid bilayer. The cosine similarity score represents the directionality of the solute, where a value of 1 is defined such that the guanidine moiety is pointing towards $+z$, and a value of -1 means that the guanidine moiety is pointing towards $-z$. A cosine similarity of 0 represents the solute being at 90° relative to the z axis. In this figure, each column shows the histograms for a separate solute, and each row represents a particular range along the collective variable. Each histogram is normalized such that the total area is unity. The symmetry of the upper and lower leaflets of the bilayer results in antisymmetric cosine similarities between the two leaflets (e.g., range [8, 20] Å would be antisymmetric to range [-20, -8] Å). For each permeant, we take the negative values of the cosine similarity scores of the lower leaflet and histogram them with the cosine similarity scores of the corresponding ranges of the upper leaflet.

To explain this stronger orientation preference of GBIC, we take a look at the specific environment around GBIC as distinguished from GBI1. It appears that the molecular size and composition of GBIC is such that the solute can both interact with the negatively charged POPC phosphate groups with its guanidine moiety while also forming dispersion interactions with the hydrocarbon chains with its chlorobenzene moiety (Figure 3.7). Between two atoms, dispersion forces generally increase as the number of electrons increase, and a chlorine atom would therefore have significantly more dispersion contributions compared to a hydrogen atom. This atomic substitution seems to explain the strong orientational preference of GBIC in the membrane headgroups, as it can be energetically stabilized by dispersion interactions of the chlorobenzene moiety with the hydrophobic lipid tails in addition to the Coulombic interactions of its guanidine moiety with the lipid phosphate groups.

Finally, it is possible that 2GBI and ClGBI do not remain in their protonated forms from bulk water to the varied environments along the transmembrane direction. To estimate whether the molecules will remain protonated as they permeate the bilayer, we computed the pK_a shift profile using the method of Li et al.[175] for 2GBI to analyze the effect of the inhomogeneous environment on the molecule's pK_a (see details in supplementary information). The pK_a of neutral 2GBI has been experimentally determined to be 7.09.[2] The pK_a of positively charged 2GBI, whether GBI1 or GBI2, is necessarily greater than 7.09. Based on the pK_a shift profile, the pK_a is expected to change by up to +2 pK units. This means that the effect of the membrane will lead to a change in pK_a of positively charged 2GBI to be at least 9. This means that GBI1/GBI2 is predicted to exhibit more basic character as it permeates the membrane, particularly in the glycerol-phosphate region, and therefore we do not expect deprotonation during membrane permeation. As a point of reference, even a positively charged arginine residue is predicted to remain 70-99% protonated while partitioning into a dipalmitoylphosphatidylcholine (DPPC) lipid bilayer at neutral pH.[175]

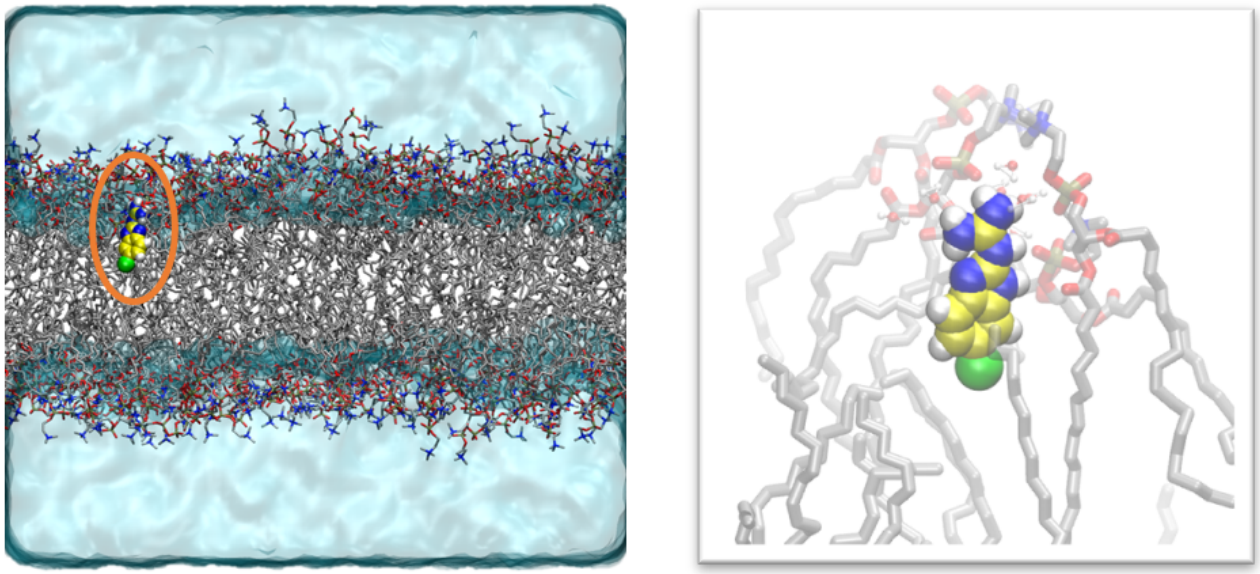


Figure 3.7: In addition to electrostatic interactions of the charged guanidine region of the solute with the negative phosphate moieties of POPC, GBIC may be further stabilized by dispersion interactions of its chlorine atom with the hydrocarbon tails of the lipid bilayer. The left panel shows a representative snapshot of GBIC in the membrane in the region z = 8-20 Å. GBIC (circled in orange for ease of identification) is drawn as vdW spheres, with carbon atoms in yellow, nitrogen atoms in blue, hydrogen atoms in white, and the chlorine atom in green. The right panel shows a close up view of the local environment around GBIC. Water molecules within 3 Å of the solute are depicted in CPK representation, with farther water and lipid molecules hidden for viewing clarity.

3.4 Conclusions

In this work, we compared the potentials of mean force to estimate the relative energetic contributions for two Hv1 channel blockers to permeate the membrane. Experimental evidence suggests that ClGBI can permeate the membrane and block the channel when present in the extracellular medium.[128] We draw general qualitative conclusions from our results to show that there is increased energetic favorability for ClGBI to permeate compared to 2GBI, and rationalize that a major driving factor may be considerable stabilization in the membrane headgroups by both electrostatic and dispersion interactions. A survey of existing literature of both experimental and computational results demonstrates that halogen atoms tend to increase the membrane permeability between similar chemical compounds.[252, 317, 93, 95, 102] We do not purport our results to be quantitatively accurate given the limitations of predicting permeability rates from molecular simulations,[171, 53, 254, 186, 308, 298, 43]. Further exploration of permeability from a different angle that takes into account kinetic properties to compute transition rates, such as in milestoning[327, 314] or the weighted ensemble[343, 47] methods, may yield further insight into the properties of cell-absorption by these Hv1 blockers.

3.5 Acknowledgments

The authors thank David Mobley, Michael von Domaros, and Caitlin Bannan for helpful discussions. VTL appreciates editorial input from Justin Smith and Matthew Agee. We acknowledge financial support from NIH grant GM098973.

3.6 Supplementary Information

Results and analysis of the pKa shift profile, as well as a discussion on long timescales for solute flip flop, are presented in the supplementary information of Appendix B.

Chapter 4

Benchmark assessment of molecular geometries and energies from small molecule force fields

Force fields are used in a wide variety of contexts for classical molecular simulation, including studies on protein-ligand binding, membrane permeation, and thermophysical property prediction. The quality of these studies relies on the quality of the force fields used to represent the systems. Focusing on small molecules of fewer than 50 heavy atoms, our aim in this work is to compare six force fields: GAFF, GAFF2, MMFF94, MMFF94S, SMIRNOFF99Frosst, and the Open Force Field version 1.0 (Parsley) force field. On a dataset comprising over 20,000 molecular structures, we analyzed their force field-optimized geometries and conformer energies compared to reference quantum mechanical (QM) data. We show that most of these force fields are comparable in accuracy at reproducing gas-phase QM geometries and energetics, but that GAFF/GAFF2/Parsley do slightly better in reproducing QM energies and that MMFF94/MMFF94S perform slightly better in geometries. Parsley shows considerable improvement over its predecessor SMIRNOFF99Frosst, and we identify particular

outlying chemical groups for further force field improvement.

4.1 Introduction

The study of chemical and biological systems relies on an accurate assessment of the energetics and geometries of the systems. Many computational methods serve to help investigate these systems, ranging from more accurate, higher cost quantum mechanical techniques to more approximate methods which compromise accuracy in favor of increased efficiency. Classical mechanics-based calculations fall into the latter group, and have an advantage over more theoretically rigorous calculations of being able to model larger systems over longer time timescales.[101, 49, 163, 164]

The modeling and simulation of molecular systems in classical mechanical calculations typically requires the use of a force field, a set of energy functions and associated parameters comprising the potential energy function. This potential energy function defines interactions between components in the system based on the coordinates of its particles.[253, 239]

Force fields have a long history of development. Strategies for force field development vary in terms of the chemical space covered, the types of data used for training, and the approach to optimize parameters given a set of input data.[216, 208] The training data used to develop a force field typically includes input data from both experimental and reference QM calculations. This finite amount of input data is carefully chosen to be representative of the systems for which the force field is designed. The limit of accuracy for some force field is measured by its ability to reproduce experimental observables, such as hydration free energies. When such experimental evidence is unavailable, the force field can be assessed with respect to quantum mechanical data, such its ability to reproduce QM geometries and relative energies. Given the complexity of force field development, including multidimensional input data, various

functional forms, and approaches to chemical perception[202], force fields vary in how accurately they can compute properties of interest. Indeed, many examples serve to highlight the limitations of force fields.[32, 91, 130, 157, 199]

Our focus in this work is on force fields for small molecules, which are instrumental for evaluating binding free energies and modeling ligand binding poses. Relatively few literature studies evaluate force field accuracy on general small drug-like molecules, in contrast to force fields for proteins,[109, 191, 123, 48, 178, 132, 275, 256, 194] nucleic acids,[251, 247, 81] carbohydrates,[195, 234, 11, 286] and other specific chemical systems.[321, 218, 198, 236, 79, 277, 237, 232, 20, 211, 158, 110, 278] On small molecules, this work comprises predictions of solvation free energies,[307, 205] strain energies,[269] experimental osmotic coefficients,[340] partition coefficients,[146, 80] conformer energies,[250, 149, 147, 328, 116] conformer geometries,[328, 131] and robustness of parameterization[131]. Most of these studies assess four or fewer force fields on limited molecule sets. We aim to present a broader assessment of general small molecule force fields on a large, diverse library of drug-like compounds to provide a general evaluation of how accurately these force fields perform. We use QM data as a valuable source of information for force field assessment to explore chemical space relatively quickly and easily.

In this work, we benchmarked small molecule force fields with respect to quantum mechanical results. We assessed six force fields: the General Amber Force Field, first and second generations (GAFF[319] and GAFF2[37]); the Merck Molecular Force Field, initial and "static" versions (MMFF94[111, 112, 113, 117, 114] and MMFF4S[115, 116]); and the SMIRKS-based force fields from the Open Force Field Initiative (OpenFF 1.0 "Parsley"[316] and its predecessor SMIRNOFF99Frosst[203]). On a dataset of over 25,000 small molecules, we evaluated optimized geometries and energies from molecular mechanics (MM) energy minimizations using force fields and compared them with reference to quantum mechanical data. We also identified particular chemical groups that represent systematic outliers in the force field

geometries and energies. This work provides a general understanding of the strengths of different small molecule force fields and identifies areas of improvement for future force field development.

4.2 Methods

4.2.1 We acquired reference geometries and energies of molecules from QCArchive and grouped them by connectivity

We obtained the molecule set in this work from QCArchive[280] from the dataset labeled `OpenFF Full Optimization Benchmark 1`, which was created for the purpose of benchmarking OpenFF 1.0.[221] The benchmark set was chosen to include a broad range of drug-like compounds.[201] This QCArchive dataset contains QM geometry-optimized structures and energies from the B3LYP-D3BJ/DZVP method.[22, 170, 313, 285, 99] This method and basis set were chosen by the Open Force Field initiative to provide reasonably accurate conformational energies and geometries at moderate computational cost.[250, 149] The database was accessed on November 11, 2019, at which point it contained 26,313 completed geometry optimization results (i.e., the total number of conformers for all molecules), 420 structures with error, and 3 calculations incomplete. We took only the completed results.

In our dataset, we organized molecular structures such that conformers of the same molecule were grouped together if they have the same absolute (non-isomeric) graph. Importantly, we do not use the SMILES string listed in the QCArchive DataFrame to represent the molecule itself, because the identity of the molecule may change during QM geometry optimization due to changes in bonding/changes in tautomer; see, for example, Figure 4.1. Molecules with different tautomerization states, which have different chemical connectivity, are treated as distinct molecules in our study. While two molecular structures may start QM optimizations

from the same molecule (same connectivity), we only use their final geometries to identify and distinguish molecules based on their connectivity. We grouped together all structures in the dataset whose final geometries yielded the same canonical isomeric SMILES string, as evaluated by OEMolToSmiles from the OpenEye OEChem Python toolkit. The structures were then organized into conformer sets as perceived by OEChem's OEAbsCanonicalConfTest. This dataset organization procedure takes into account any molecular identity changes during QM optimization, such as if two same molecules no longer had the same tautomerization state after QM optimization or if two different molecules ended up in the same tautomerization state. We ensured that what we identified as a molecule, and all of its given conformers, contained the same chemical connectivity.

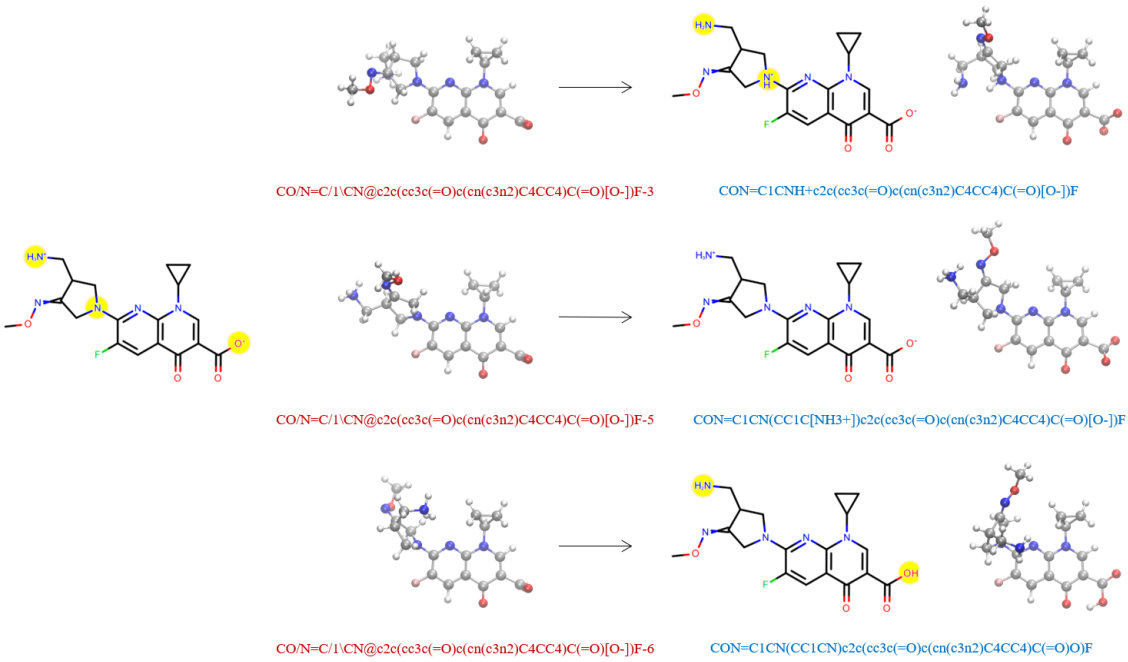


Figure 4.1: Molecules may change connectivity during QM optimization leading to different tautomers. On the left hand side, we show the Lewis structure and three associated conformers of an example molecule from the QCArchive `OpenFF Full Optimization Benchmark 1` dataset. Yellow circles highlight the regions of potential tautomerization changes. The QCArchive SMILES labels are colored in red. The right hand side shows the structures after QM optimization. The canonical isomeric SMILES labels representing the optimized molecules are colored in blue. Only the middle structure retains the original tautomeric identity. In our dataset, the geometries on the right hand side would be analyzed as distinct molecules.

The resulting QM geometries were used as input structures for gas phase energy minimizations using the following small molecule force fields: GAFF,[319] GAFF2,[37] MMFF94,[111, 112, 113, 117] MMFF4S,[115, 116] Parsley[316], and SMIRNOFF99Frosst[203]. Parsley refers to the Open Force Field (OpenFF) Parsley Force Field (v1.0.0), specifically

`openff_unconstrained-1.0.0-RC2.offxml`. SMIRNOFF99Frosst is a SMIRKS Native Open Force Field (SMIRNOFF) and descends from AMBER's parm99 force field as well as Merck-Frosst's parm@frosst. The SMIRNOFF99Frosst version used here is `SMIRNOFF99Frosst-1.1.0.offxml`. SMIRNOFF99Frosst is the predecessor of Parsley.

4.2.2 We assigned FF parameters then energy minimized all molecules

Each structure was assigned AM1 Mulliken-type partial charges with bond-charge corrections (AM1-BCC charges)[136, 137] from the electrostatically least-interacting functional group technique. The partial charges were generated using the openmoltools wrapper[21] to OpenEye's oequacpac charging engine[222] calling OEAM1BCCELF10Charges().

To assign force field parameters to each molecule, we used antechamber and tleap[37] via openmoltools[21] for the GAFF(2) force fields, and we used the Open Force Field toolkit for Parsley and SMIRNOFF99Frosst, in all cases using pre-assigned charges applied as described above. Parameter assignment as well as energy minimization for the MMFF94(S) force fields were handled using OpenEye oeszybki.[222] All other energy minimizations were completed in OpenMM[74] using the L-BFGS algorithm with an energy tolerance of 5.0e-9 kJ/mol and 1500 maximum number of iterations.

We removed any molecular structure that was not successfully parameterized with all force fields. This set consisted of 721 structures that were unable to be parameterized by GAFF or GAFF2, and an additional 522 structures that raised an error during OpenMM setup

through the Open Force Field toolkit. Our pruned set contained 25,070 structures.

4.2.3 We evaluated relative energies and geometric agreement with optimized QM geometries

We compared the energy-minimized geometries and energies for each force field with respect to the QM reference data by computing the following metrics: relative energy differences (ddE), root-mean-square deviation of atomic positions (RMSD), and Torsion Fingerprint Deviation (TFD)[268, 76]. The relative energy difference (ddE) between the FF and QM energy for the ith conformer of some molecule was computed using the following expression:

$$ddE_i = dE_{FF,i} - dE_{QM,i} \tag{4.1}$$

$$= [\text{FF energy}(i) - \text{FF energy}(0)] - [\text{QM energy}(i) - \text{QM energy}(0)] \ , \tag{4.2}$$

where the 0th conformer is defined as the conformer with the lowest QM energy for the given molecule.

Molecular structures may change conformation after energy minimizations which may lead to lower agreement between FF and QM energies for a particular conformer. To address this potential issue, we performed a conformer matching process for each FF structure which considered the final optimized geometries and energy differences. We ensured that every MM conformer was within 1.0 Å of a QM reference structure. The QM reference conformer was removed from analysis if there were no FF conformers that matched it within 1.0 Å RMSD. Furthermore, if a molecule ends up with two of the "same" FF-minimized conformers compared to a QM reference structure, we only keep the FF conformer with the lowest RMSD score while any redundant conformers are removed from analysis. For this reason, the number

of total molecular structures for each force field will likely differ after conformer matching as the intricate conformational energy landscapes are represented differently by various QM methods and force fields. Then, the mean signed deviation (MSD) was computed over the N conformers of each molecule with equation 4.3, iterating over the relative energy dE of each conformer i. The reference conformer with $dE = 0$ was removed from the MSD calculation. The molecule MSDs were then aggregated to compare among all force fields.

$$MSD = \frac{1}{N} \sum_{i=1}^{N} dE_{FF,i} - dE_{QM,i} \tag{4.3}$$

To compare FF geometries with QM geometries, we used RMSD and TFD scores. The RMSD values, calculated with OpenEye OERMSD, took into account hydrogen atoms, symmetry-related transformations, and overlaid structures to yield the lowest possible RMSD. TFDs were computed using the RDKit Python library. We evaluated each of these three metrics individually and looked for potential correlations between energies and structures in terms of agreement with reference QM data.

The complete Python code used for the setup, FF minimizations, and analysis of this work is available on Github at `https://github.com/MobleyLab/benchmarkff`. The Github repository also contains SDF files of the starting molecular dataset (QM energies and geometries; `full_opt_benchmark1.sdf`) and the structures removed due to parameterization or setup errors (in the `benchmarkff/molecules/issues` directory).

4.3 Results and Discussion

Here, we present and discuss our results comparing several general small molecule force fields against reference quantum mechanical data. We are interested in two major categories of comparison: Energetic agreement and geometric agreement. Particularly, an ideal force field will yield the same energy minima or optimized geometries as the QM energy landscape, with no additional minima, and the relative energies of those minima will agree between QM and MM. Thus, to assess performance in these two categories, we computed relative conformer energies and compared these between MM and QM, as well as assessing geometric agreement of MM optimized geometries with those from QM. We also identified specific parameters for improvement for future versions of the Open Force Field small molecule force field.

Our study relies on the assumption that force field accuracy can be evaluated using gas phase energies and geometries. One of the greater goals of force field science, such as that of the Open Force Field Initiative, is building force fields that will work well in the condensed phase (e.g., small molecules in solution or binding to biomolecules). That being said, we make our assumption based on two key observations. First, force fields—especially those in the AMBER family—are usually fitted to reproduce gas phase conformational energies and geometries.[319] This means that we are testing these force fields on properties they are fitted to reproduce. Second, bonded parameters are not expected to change significantly on transfer to the condensed phase. Rather, non-bonded interactions are particularly important in condensed phase simulations. Of the non-bonded interactions, electrostatics models are often polarized beyond what would be expected in the gas phase in order to reproduce condensed-phase properties, and Lennard-Jones parameters can be tuned to reproduce condensed phase properties (as has been a particular focus of the OPLS force fields[144, 71]). Even when these are done, force fields retain bonded terms parameterized to reproduce QM geometries and energetics, further emphasizing the importance of testing in such a context. We therefore believe our assumption is reasonable and that this work warrants investigation.

We start our force field benchmark analysis by comparing FF energies to QM energies. Here, since our choice of reference energy for MM is arbitrary, we choose to compare relative conformer energies. For any given molecule, an ideal force field would have relative energies for different conformers in MM that agree with those for the same conformers in QM. For the differences in relative conformer energies that we computed—that is, the difference between the MM relative conformer energies and the QM relative conformer energies—a FF with greater agreement to QM should have more values around or at 0 kcal/mol, and a FF with lower agreement with QM would exhibit a more spread out distribution away from 0 kcal/mol.

The computed energy differences for over 20,000 molecular structures with the six force fields were generally within ±50 kcal/mol (Table 4.1). However, GAFF had outlying energies which were several orders of magnitude beyond this range (row 1 of Table 4.1). These energies were traced back to six molecules (62 conformers thereof) shown in Figure 4.2. These molecules all contain a polar hydrogen atom which, after geometry optimization, overlaps with its parent atom. The spurious overlap of these hydrogen atoms, and associated energy extremes, results from a missing van der Waals parameter in GAFF. In GAFF2 (and SMIRNOFF99Frosst and subsequent OpenFF force fields[316, 203, 202]) hydroxyl hydrogens no longer have zero Lennard-Jones parameters, which seems to eliminate the problem for these molecules. Similar collapse of hydroxyl groups in close proximity has been observed previously in force fields with zero LJ parameters for hydroxyl hydrogens.[202]

After excluding the 62 GAFF outliers, the ddE energies are histogrammed in Figure 4.3. The relative conformer energy differences exhibit very similar distributions for all force fields. All distributions appear asymmetric, with a skew towards more negative ddE values than positive ones. Force fields of the same family tend to be more consistent with each other (GAFF and GAFF2, MMFF94 and MMFF94S). Concentration around ddE = 0 kcal/mol is greatest in the GAFF(2) family. From these results, the qualitative ordering of force

Force field	min ddE	max ddE
GAFF	-5549.7	35002325.4
GAFF[a]	-14.8	44.1
GAFF2	-15.6	43.7
MMFF94	-29.8	52.1
MMFF94S	-25.1	49.5
Parsley	-19.3	38.4
SMIRNOFF99Frosst	-18.8	42.8

[a] With outliers removed.

Table 4.1: Minimum and maximum ddE values as computed in equation 4.2 for all structures of each force field. Energy units are in kcal/mol.

fields from lowest to highest agreement with QM energies goes as SMIRNOFF99Frosst $<$ MMFF94 $\sim$ MMFF94S $<$ GAFF $\sim$ GAFF2 $\sim$ Parsley. In other words, GAFF, GAFF2, and Parsley seem to exhibit roughly similar performance by this metric, with the other force fields performing somewhat worse.

Given that two conformers starting from the same geometry may optimize to two distinct conformers after FF minimization, we took another approach to analyzing energy distributions, only considering the FF conformers that correspond to a QM counterpart. A FF conformer is deemed to have a "match" with a QM conformer if their RMSD is less than or equal to 1 Å (see more details in Methods). Out of 24,535 structures (excluding those for which *no* force field had a match), the number of matched conformers for each force field are: 22995 (GAFF), 22965 (GAFF2), 22822 (MMFF94), 22826 (MMFF94S), 21038 (Parsley), and 19714 (SMIRNOFF99Frosst). The mean signed deviation of the matched conformer energies are shown as violin plots in Figure 4.4. The violin plots are scaled such that each violin has the same area. This figure shows that the mean signed deviation of relative conformer energies is also fairly consistent between different force fields as seen in Figure 4.3. Upon closer inspection, the violins for GAFF, GAFF2, and Parsley are slightly wider around 0 kcal/mol (and those narrower elsewhere), signifying marginally higher agreement with QM energies. Note that this conformer filtering step was only used for analyzing the energies in the violin plots, and other results throughout this work do not rely on matched conformers.

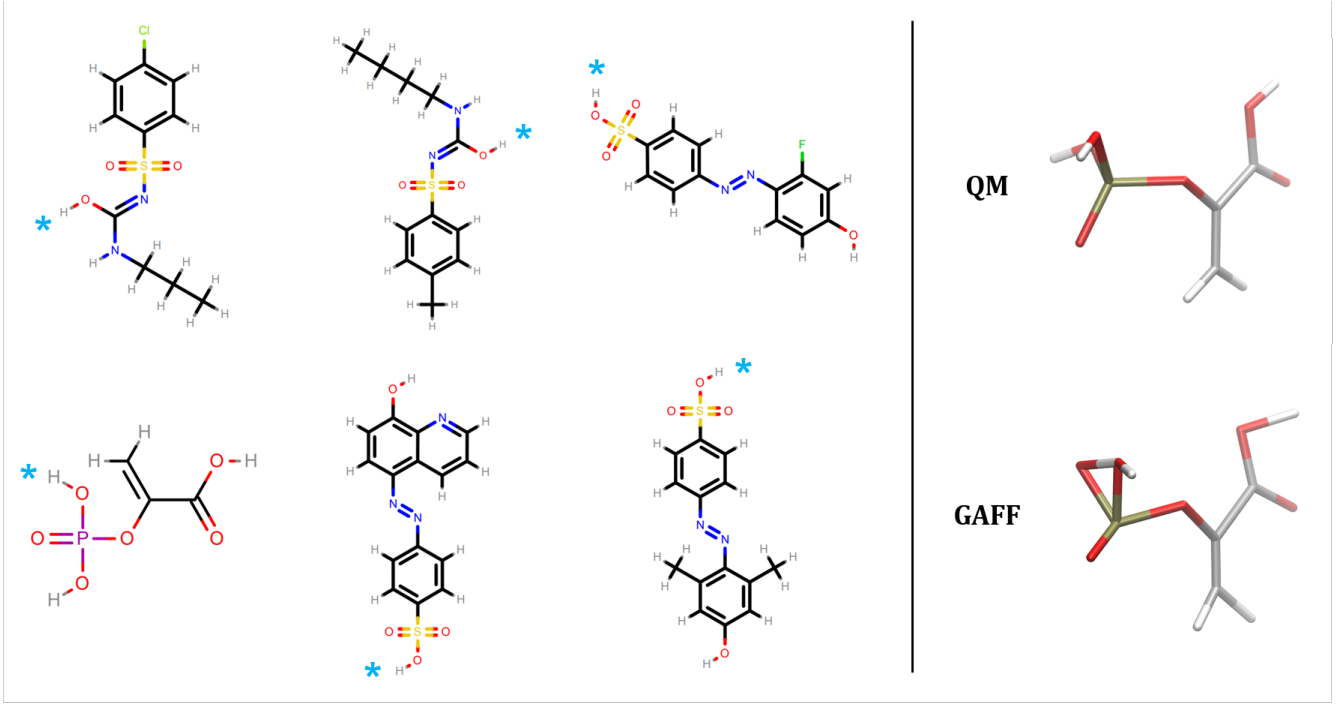


Figure 4.2: Molecules with extreme relative conformer energies for GAFF. The right hand side depicts the QM and FF geometries for phosphoenolpyruvic acid. The GAFF structure shows a representative overlap of a polar hydrogen atom with its connected parent atom due to a missing van der Waals parameter. On the left hand side, the overlapping hydrogen for the six molecules are denoted by cyan asterisks.

We next examine agreement between FF-optimized geometries and those from QM, as calculated by each molecule's root-mean-square deviation of atomic positions (RMSD) and Torsion Fingerprint Deviation (TFD) scores with reference to the parent QM-optimized geometries. While RMSD is the more common metric, it may depend on the molecule size, complicating interpretation of geometric agreement.[249, 261] In contrast, TFD was designed to be more independent of molecule size in order to compare molecular conformations more meaningfully.[268] This can help offset issues with RMSD where larger, more flexible molecules can contribute the most to RMSD. The TFD score between two molecular structures is evaluated by computing, normalizing, and Gaussian weighting the (pseudo-)torsion deviation for each bond and ring system. While TFD is normalized from 0 to 1, RMSD is unbounded. Both RMSD and TFD are similar in that a higher value signifies lower agree-

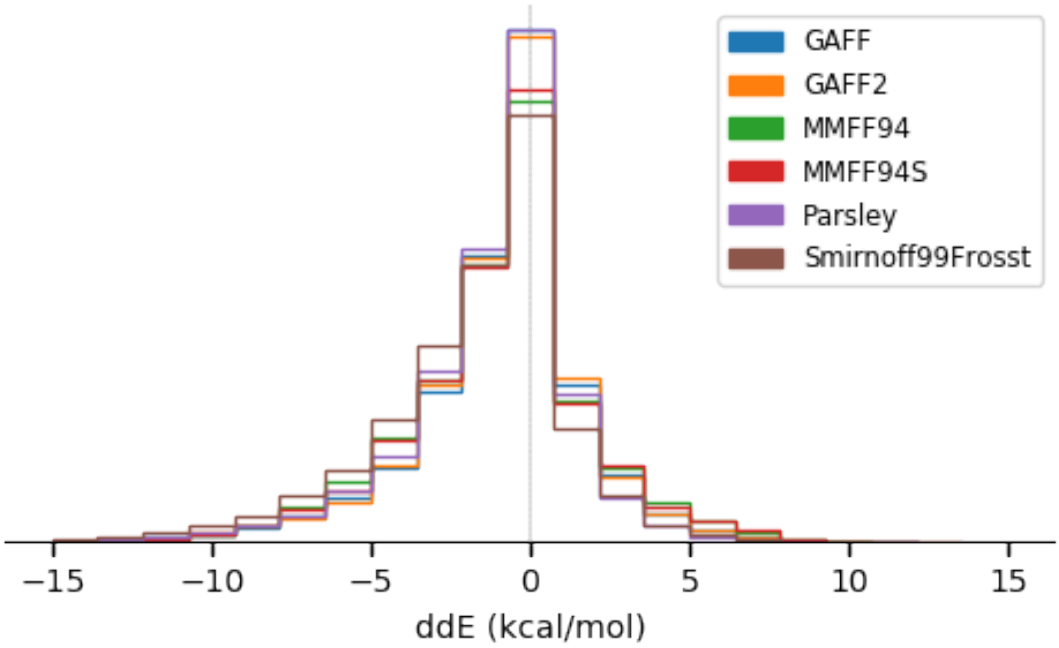


Figure 4.3: Histograms of the relative conformer energy differences as computed in equation 4.2 for each force field relative to QM. Each molecular structure, including different conformers of the same molecule, is counted separately. A force field having higher agreement with QM would have a higher bin centered at ddE = 0 kcal/mol. Parsley (purple) overlaps with GAFF (blue) in the central bin.

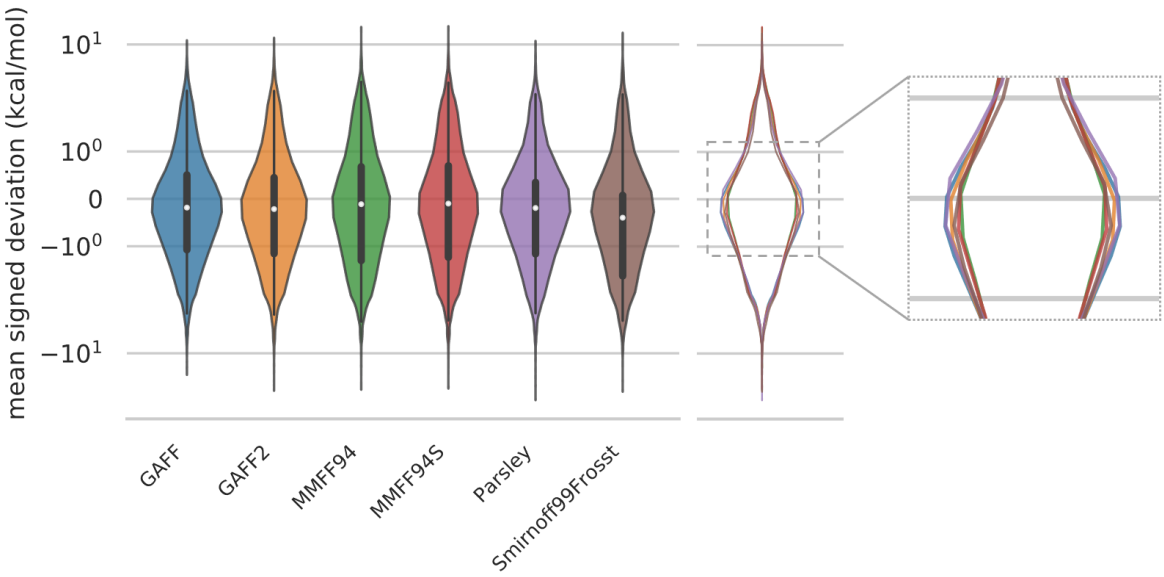


Figure 4.4: Violin plots of the mean signed deviation of conformer energies as computed in equation 4.3. The energy MSDs only take into account structures matched within 1 Å of the QM reference structure, so there are minor differences in the amount of data used to plot each violin (see text). The vertical axis is shown on a logarithmic scale. An overlay of the violin plots on the right panel better shows the subtle distinctions between the force fields in the most populated region, near zero error.

ment between the geometries of two molecules. A FF which yields optimized geometries closer to those of QM would have generally smaller RMSD/TFD values.

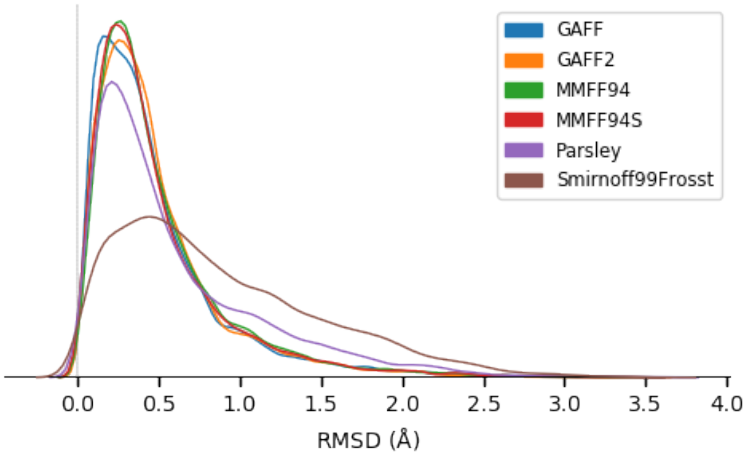
We calculated RMSD and TFD scores for the 24,958 structures with all force fields with respect to QM geometries. We plotted this data in histograms, applying a kernel density estimate (KDE), in Figure 4.5. While the smoothened distributions allow for easier visual distinction among the force fields, RMSD and TFD scores can never be lower than zero, so the KDE results in slight plotting artifacts. The raw histograms without smoothing can be found in the supplementary information.

In terms of geometry agreement, we observed similar results between the RMSD and TFD plots. For example, we note a higher deviation of geometries in Open Force Field's SMIRNOFF99Frosst and Parsley relative to other force fields, as seen in the elongated distributions towards higher RMSD/TFD scores. This effect is more pronounced for SMIRNOFF99Frosst. The larger

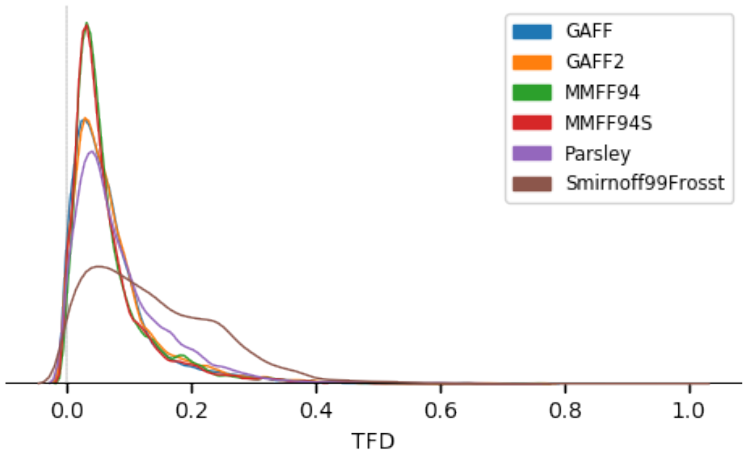
subset of molecules in Parsley and SMIRNOFF99Frosst, relative to the other force fields in this study, which have poorer agreement with QM geometries also explains why these two force fields have a lower number of conformer-matched structures as described earlier regarding Figure 4.4. Otherwise, force fields of the same family tend to have more consistent results with each other using the TFD measure. In the TFD distributions, the MMFF94(S) force fields seem to best reproduce QM reference structures for the molecules in our dataset, followed by GAFF(2), then the force fields from the Open Force Field Initiative.

We then sought to determine if there was a dependence between the relative energies and geometries. Scatter plots of ddE versus RMSD/TFD are shown for the Parsley force field in Figure 4.6. Each structure in our dataset is plotted as a single point. The ddE values are plotted on a logarithmic scale. We include in the supplementary information analogous plots for the other force fields, as well as all plots with ddE represented on a linear scale. Given tens of thousands of points on each plot leading to many overlapping points, we applied a color gradient from red to blue to represent regions from low to high density, respectively. Similar to the data represented as one-dimensional histograms (Figures 4.3 and 4.5), a higher density of points at the origin indicates results in better agreement with the reference QM data. There seems to be no general correlation between the energies and geometries for Parsley (Figure 4.6) or for other force fields (see supplementary information). However, using this visualization we identified particular chemical moieties that represent outlying energies or geometries (vide infra).

In this diverse set of molecules, we point out three particular moieties, those containing an N-N single bond (4221 structures), those containing an azetidine ring (639 structures), and a highly substituted octahydrotetracene (50 structures). These subsets are highlighted for Parsley in Figure 4.7 (see supplementary information for other force field results). Molecules containing an N-N single bond have a wide spread of energies with several ddE outliers between -10 to -20 kcal/mol. Structures with azetidine revealed both energy and geometry



(a)



(b)

Figure 4.5: Kernel density estimates of the RMSD and TFD values between force field structures as compared to QM structures. Values closer to zero indicate higher geometric similarity for both RMSD and TFD. The minimum value possible for either measure is zero; however, the smoothing generates curves that extend below zero. We include the smoothing to more easily distinguish between different force fields, but we also include the non-smoothened histograms in the supplementary information.

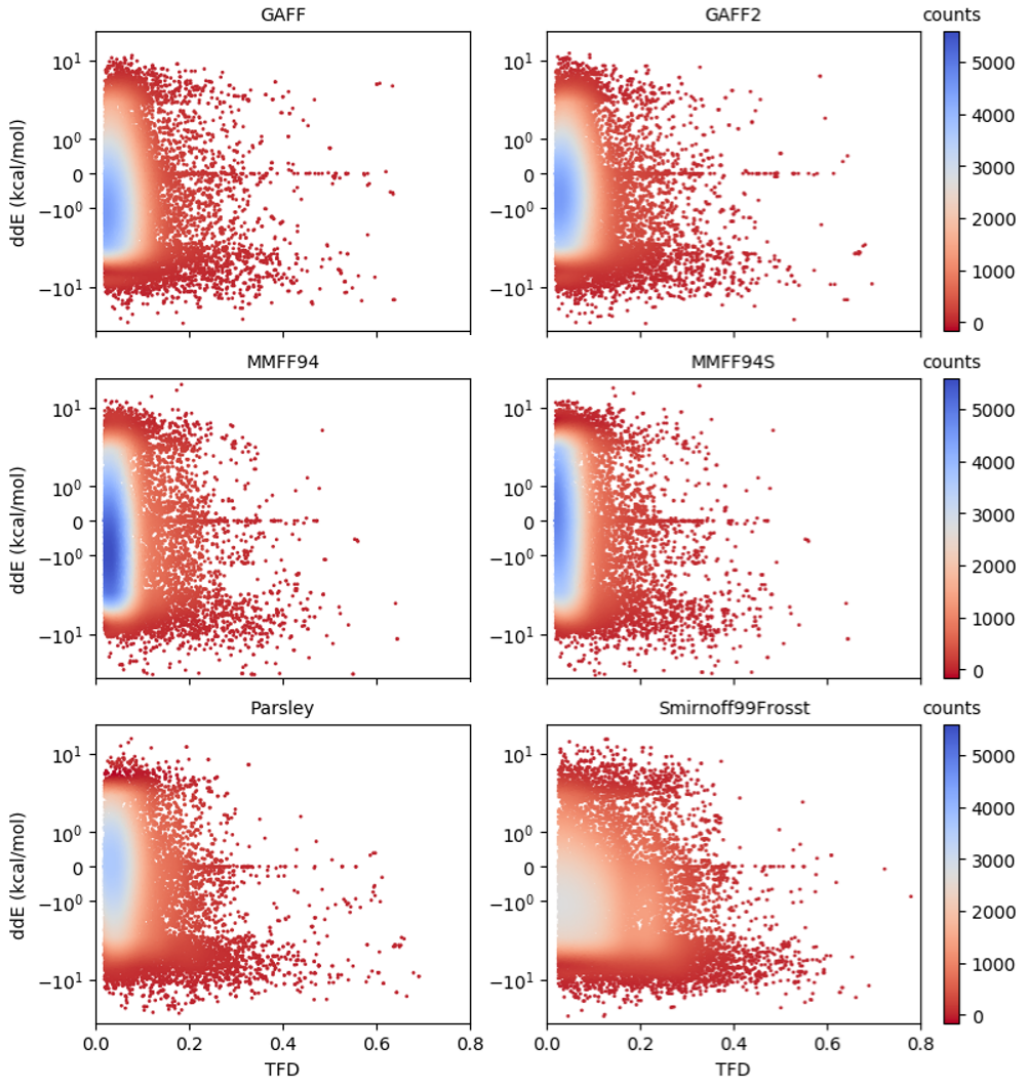


Figure 4.6: Scatter plots of relative conformer energies versus TFD scores. The points are colored by the interpolated density of points in a certain area. Blue indicates region of high density, that is, high compactness of points in that area. A force field having better agreement in both relative energies and geometries with the QM reference would have more points around the origin (ddE = 0, TFD = 0), though it is presumably possible for a force field to improve along one axis without improving along the other. The vertical axis is represented on logarithmic scale; the same plots with linear scaling can be found in the supplementary information.

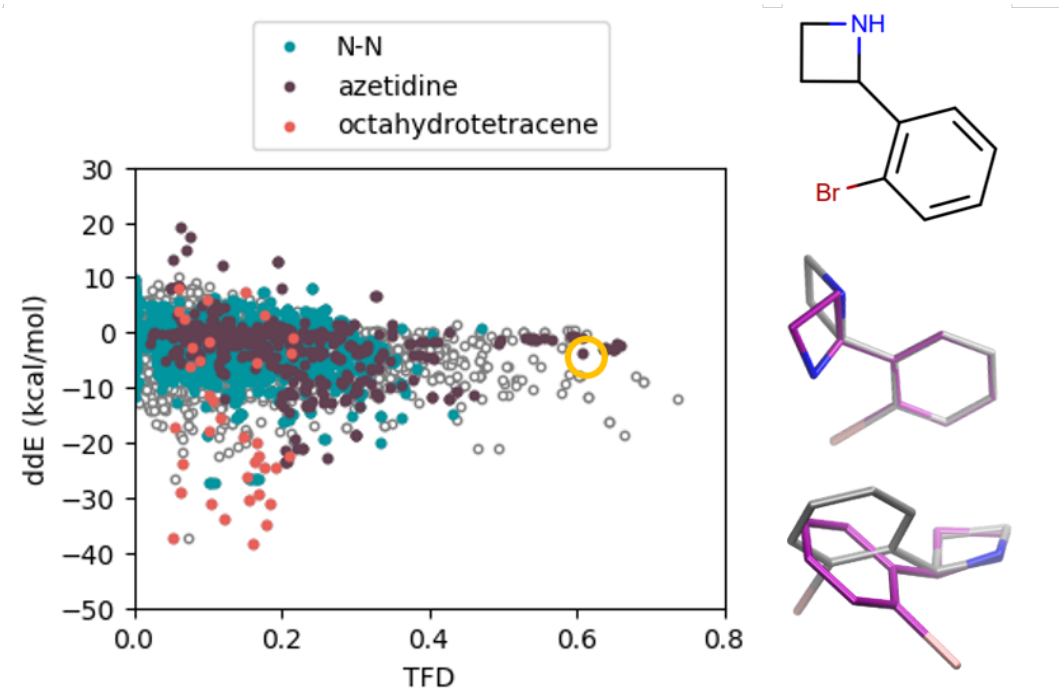


Figure 4.7: Scatter plot of the Parsley force field of relative conformer energies versus TFD scores. Colors highlight particular chemical groups that appear to be systematic outliers in energies or geometries. On the right hand side, we show a figure with high TFD and low ddE as circled in the scatter plot. The QM structure is in purple, and the force field structure is colored in silver.

outliers in the Amber and OpenFF force field families. Lastly, the substituted octahydrotetracene scaffold was found to be challenging to all force fields in reproducing QM energies (an example is presented in the supplementary information). These moieties represent systematic outliers that can be used in future studies investigating particular shortcomings of force fields or improving future versions of force fields.

4.4 Conclusions

In this work, we presented a large-scale analysis of six small molecule force fields in terms of their conformational energies and geometries compared to reputable QM data. The force fields, GAFF, GAFF2, MMFF94, MMFF94S, SMIRNOFF99Frosst, and Parsley, perform

similarly in terms of reproducing QM conformer energies, with slightly higher accuracy in GAFF, GAFF2, and Parsley. The MMFF94 and MMFF94S tend to capture QM geometries better than the other force fields, especially when using a more size-independent geometry measure. Finally, we identified particular chemical moieties that were systematic outliers in terms of relative energies or geometries. These N-N, azetidine, and octahydrotetracene-like compounds represent potential areas for improvement in future force field development.

We share our Python code comprising the setup, minimization, and analysis of this research on Github, available at: `https://github.com/MobleyLab/benchmarkff`.

Beyond these specific conclusions, we believe the general strategies employed here for benchmarking force field performance will be useful far more broadly than this specific study. Particularly, comparing performance by both geometric and energetic measures is particularly important, as the analysis we have done demonstrates. Additionally, the availability of a large amount of public data in QCArchive facilitates straightforward large scale benchmarking in a way it has not been done previously.

4.5 Acknowledgments

The authors thank the Open Force Field Initiative researchers for helpful planning and feedback, especially Caitlin Bannan, Jordan Ehrman, Jessica Maat, Lee-Ping Wang, and Hyesu Jang. We value discussions and insights from Christopher I. Bayly (OpenEye) who helped motivate this work. We also thank Jeffrey R. Wagner and Daniel G. A. Smith for code review and improvements. We appreciate funding from NIH grants GM098973 and (to DLM) 1R01GM108889-01 and 1R01GM124270-01A1.

4.6 Supplementary Information

The supplementary information in Appendix C contains (1) histograms without kernel density smoothing for RMSD and TFD relative to QM reference data for all force fields investigated in this work, (2) plots similar to those in Figure 4.6 with linear scaling of the vertical axis, and (3) plots in the same manner of Figure 4.7 for all force fields in this work. We also present an example of one of the octahydrotetracene-based structures having high deviation in ddE.

Chapter 5

Assessing the Conformational Equilibrium of Carboxylic Acid via QM and MD Studies on Acetic Acid

Accurate hydrogen placement in molecular modeling is crucial for studying the interactions and dynamics of biomolecular systems. The carboxyl functional group is a prototypical example of a functional group that requires protonation during structure preparation. To our knowledge, when in their neutral form, carboxylic acids are typically protonated in the *syn* conformation by default in classical molecular modeling packages, with no consideration of alternative conformations, though we are not aware of any careful examination of this topic. Here, we investigate the general belief that carboxylic acids should always be protonated in the *syn* conformation. We calculate and compare the relative energetic stabilities of *syn* and *anti* acetic acid using *ab initio* quantum mechanical calculations and atomistic molecular dynamics simulations. We focus on the carboxyl torsional potential and configurations of microhydrated acetic acid from molecular dynamics simulations, probing the effects of solvent, force field (GAFF vs. GAFF2), and partial charge assignment of acetic acid. We show

that while the *syn* conformation is the preferred state, the *anti* state may in some cases also be present under normal NPT conditions in solution.

5.1 Introduction

The carboxyl functional group, –COOH, is widespread in nature and highly biochemically relevant. It is present in amino acids that compose proteins, fatty acids of cell membranes, and naturally occurring organic compounds (*e.g.*, niacin, citric acid, biotin). This group is very common in medicinal compounds, found in over 450 marketed drugs including nonsteroidal anti-inflammatory drugs (*e.g.*, aspirin, ibuprofen), antibiotics (*e.g.*, penicillin) and cholesterol-lowering statins (*e.g.,* atorvastatin (Lipitor)).[15, 184] The presence of the hydrophilic carboxyl moiety on organic compounds can confer high solubility in water,[60, 196, 65] which can be important to consider when designing new chemical reactions or developing new medicinal compounds. This group can also have important implications for pharmaceutical drugs; for example, drugs with a carboxyl functional group can be more metabolically unstable[165] or have more difficulty diffusively crossing membranes.[165, 15] Given the carboxyl group's ubiquitous presence in nature and its importance as a functional group, understanding its conformational preferences in various settings is fundamental for the design, modification, and property prediction of new and existing molecules.

The preferred orientation of hydroxyl in the carboxyl functional group in solution is a matter of some debate, even for acetic acid, an archetypal carboxylic acid. Given a typical pK_a of less than 5, the carboxyl group will usually be in the unprotonated, anionic form at neutral pH when exposed to the environment. However, the pK_a may be significantly shifted as part of a ligand in a protein binding pocket or on protein side chains involved in reaction mechanisms. The two equilibrium conformations of the protonated carboxyl group are denoted *syn* (Figure 5.1(a)), where the O=C–O–H dihedral angle is defined here to be 0°, and *anti* (Figure

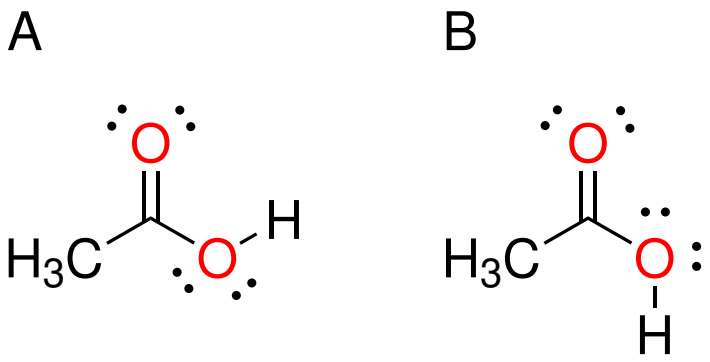


Figure 5.1: Lewis structure of acetic acid in (a) *syn* and (b) *anti* conformation.

5.1(b)), where the O=C–O–H dihedral angle is defined here as 180°. It is widely believed that the preferred conformation of carboxyl is the *syn* arrangement, from which there is a large energetic penalty to reach the *anti* arrangement. The reasoning behind this idea lies in the perceived extra stability of intramolecular hydrogen bonding that occurs in the *syn* structure. This belief is supported by a number of experimental and theoretical studies done in gas phase, and there is no doubt that this is the preferred conformation in the gas phase.[213, 215, 263, 329, 64]

The orientational preference of COOH is considerably more complex outside of gas phase. While some workers remain convinced that *syn* will be more stable, a variety of evidence indicates that this may not always be the case. A recent review article[214] discusses the competition between intramolecular and intermolecular hydrogen bonds in solution, stating that an intramolecular hydrogen bond may be disrupted in protic solution, such as water, when the increase in internal energy is offset by two or more solute-solvent intermolecular hydrogen bonds. Another study found that the carboxyl group has no strong preference, kinetically and thermodynamically, for the *syn* (or *anti*) conformation in proton transfer catalysis.[209] The *anti* state may also be important to consider when calculating solvation free energies.[154] In addition, the *anti* conformation is not insignificantly represented in structures from the Cambridge Structural Database,[59] supported by related crystallographic and theoretical charge density studies.[225, 197] The carboxyl group may also be

strongly influenced by its surroundings, such as that within a protein binding site, to prefer either the *syn* or *anti* state. In general, the local environment plays a large role in the conformational state of the carboxyl group, and the preferred orientation is not always obvious.

Past work investigating acetic acid in solvent predominantly considers the *syn* state, such as in studies characterizing hydrogen-bonding interactions of acetic acid microhydrates using DFT-B3LYP calculations,[90] or assessing the dimer form in various stages of hydration theoretically[231, 45] and experimentally.[224] One recent work examining solvent stabilization using DFT-ωB97X-D calculations[160] indicates that water may modulate the conformational preferences of acetic acid; however, to our knowledge there has not been a systematic investigation of the preferred conformational state of the carboxyl group in solution. We believe that this collection of evidence on the orientational preference of COOH in solvent lacks a clear, definitive answer on whether both conformational states of the carboxyl group may reasonably be populated in normal aqueous solution when this group is in its neutral protonation state.

In this work, we aim to understand the relative conformational stability and energetic barrier for carboxyl functional group interconversion in both gas and aqueous phases. We present our investigation on monomeric acetic acid using both *ab initio* quantum mechanical (QM) calculations and atomistic molecular dynamics (MD) simulations.

5.2 Methods

Past gas phase QM studies clearly indicate a preference for the *syn* structure of the COOH group such as in acetic acid. However, classical all-atom MD simulations show that both are equally favorable in solution, at least with the energy model ("force field") employed.[154]

This could be a real effect of water on the conformational preferences, or a limitation of the force field employed. Therefore, we need to examine a more intermediate region between gas phase QM and solution phase atomistic MD to settle the issue more definitively. Specifically, we look at QM in implicit solvent as well as QM data of snapshots pulled from MD simulations; on the MD end, we consider the effects of force field as well as solvation state.

We present an overview of our approach then discuss the methods in further detail. A torsion drive was conducted on acetic acid over the aforementioned dihedral angle. We conduct restrained geometry optimizations using two different QM methods, each with and without the presence of implicit solvent. Then, we carry out a set of geometry optimizations on pentahydrated acetic acid with varied water configurations obtained from MD simulations. We compared these energies for both *syn* and *anti* structures.

On the MD side, we compute a series of free energy landscapes, also known as potentials of mean force (PMFs), from driving the relevant torsion in acetic acid. We evaluated the sensitivity of these one-dimensional free energy surfaces to the force field (GAFF or GAFF2), partial charge assignment, and solvation state. We consider the force field because this factor is likely to vary among users running MD simulations. The partial charge set assigned to a solute depends on the initial conformation and is typically fixed throughout MD simulations, so we investigate potential implications of choosing one set or another. Finally, we compare the results of gaseous and aqueous phases to shed light on how reasonably the *syn* and *anti* states may be occupied in either scenario.

5.2.1 *Ab initio* torsion drive of acetic acid

Acetic acid configurations of the carboxyl O=C–O–H dihedral angle were generated and used as input for both QM torsion drives and MD umbrella sampling simulations. The dihedral angle was rotated using VMD[135] in 15° increments from 0° to 360°, yielding 24

total conformations.

The QM torsion drives were run using Turbomole version 7.1[300] with two different levels of theory: HF/6-31G* and TPSSh-D3BJ/def2-TZVP. The former method, using Hartree-Fock reference[119] with the Pople 6-31G* basis set,[121, 118] was chosen for consistency with the methods often employed in parameterization of force fields used for molecular simulation.[54] This low level method also provides historical perspective contributing to the strong bias favoring the *syn* conformation of the carboxyl group. Taking a more rigorous approach, we employed the TPSSh hybrid functional[284] with Grimme's D3 dispersion correction[105] and Becke-Johnson damping,[106] in combination with the Karlsruhe triple-zeta basis set def2-TZVP.[325] We chose the TPSSh functional because prior work indicates it is a suitable approach for treating the molecular dipole moments and polarizabilities of these hydrogen-bonding systems.[166, 126] We also run calculations in implicit solvent with each of the aforementioned methods using the conductor-like screening model (COSMO) with outlying charge corrections.[152, 151, 75, 276]

5.2.2 *Ab initio* geometry optimizations from molecular dynamics configurations

We sample various configurations of water molecules around acetic acid by running separate MD simulations of the *syn* and *anti* conformations in a box of TIP3P water molecules. The structures were solvated using Antechamber[318] within a cubic box with TIP3P waters[143] such that the minimum distance between the solute and the edge of the periodic box was 12 Å. Dynamics were run using GROMACS version 5.0.4 with the leap-frog stochastic dynamics integrator and a 2 fs time step. We use a Langevin thermostat for the temperature at 298.15 K with a frictional constant of 2.0 ps^{-1}. The pressure was maintained at 1 atm using the Parrinello-Rahman pressure coupling scheme with a time constant of 10 ps^{-1} and

an isothermal compressibility of 4.5×10^{-5} bar^{-1}. All bonds involving hydrogen atoms were constrained using the LINCS algorithm.[125] The systems were simulated with 2500 steps of steepest descent minimization, 50 ps constant volume and temperature (NVT) equilibration, 5 ns of constant pressure and temperature (NPT) equilibration, and then 5 ns of NPT production. Trajectory snapshots were extracted of the most similar configurations for acetic acid and its five closest waters using a root-mean-square deviation clustering of geometries with a 2 Å cutoff. This yielded 14 snapshots for the penta-hydrated *syn* conformation and 17 snapshots for the penta-hydrated *anti* conformation. Each snapshot was MM-optimized via OpenEye's OEChem Python Toolkit[222] using the MMFF94S force field,[112, 113, 114, 115, 116, 117] then subsequently QM-optimized using Turbomole version 7.1[300] with COSMO-TPSSh-D3BJ/def2-TZVP.[152, 151, 75, 276, 284, 105, 106, 325]

5.2.3 MD simulations with umbrella sampling along carboxyl dihedral angle

We used umbrella sampling[296] molecular dynamics to compute a potential of mean force (PMF) to analyze the free energy landscape projected onto this one-dimensional coordinate. We compared MD results with the GAFF[319] and GAFF2[37] classical all-atom force fields, with partial charges assigned by the AM1-BCC[136, 137] approach. We consider effects of the solute partial charges in the MD simulations by carrying out MD simulations with AM1-BCC charges assigned from the *syn* configuration as well as charges assigned from the *anti* configuration. Energetics were examined in gas phase, then in solvent using explicit TIP3P water molecules.[143]

These simulations were run using GROMACS version 5.0.4.[1] Each acetic acid configuration generated in VMD was set with partial charges from the AM1-BCC charge model[136, 137] on the *syn* (0°) conformation as implemented in OpenEye's Python toolkits.[222] The partial

charges of the solute depend on initial configuration, so we also consider the *anti* (180°) conformation for computing partial charges. The O=C–O–H dihedral angle was restrained in both gas phase and aqueous MD simulations, using a harmonic force constant of 300 kJ/mol/(rad^2) (approximately 0.022 kcal/mol/(deg^2)).

For the gas phase simulations, the reference temperature of 298.15 K was maintained using Langevin dynamics with a frictional constant of 1.0 ps^{-1}. Maintaining the GROMACS parameters described earlier, the systems underwent steepest descent minimization over 2500 steps, NVT equilibration for 50 ps, and NVT production for 1 ns.

For the explicit solvent simulations, the solvation parameters and other MD simulation settings were maintained as described earlier in the section, "*Ab initio* geometry optimizations from molecular dynamics configurations." These systems were simulated with 2500 steps of steepest descent minimization, 50 ps NVT equilibration, 50 ps NPT equilibration, and 5 ns NPT production. The configurations with dihedral angle around 270° seemed not converged, so six conformations were extended 5 ns for a total of 10 ns each: 65°, 90°, 105°, 255°, 270°, 285°. However, there was little to no change in the resulting PMFs.

Analysis of all umbrella sampling simulations was completed with the MBAR algorithm[273] to produce the potentials of mean force (PMFs) for rotation of the carboxyl dihedral angle.

5.3 Results and Discussion

Results from both QM and MD approaches support former work and the general understanding that *syn* is favored in gas phase. They also indicate that the *anti* conformation may also be populated to a significant extent in water. We address our QM results first and then discuss MD results.

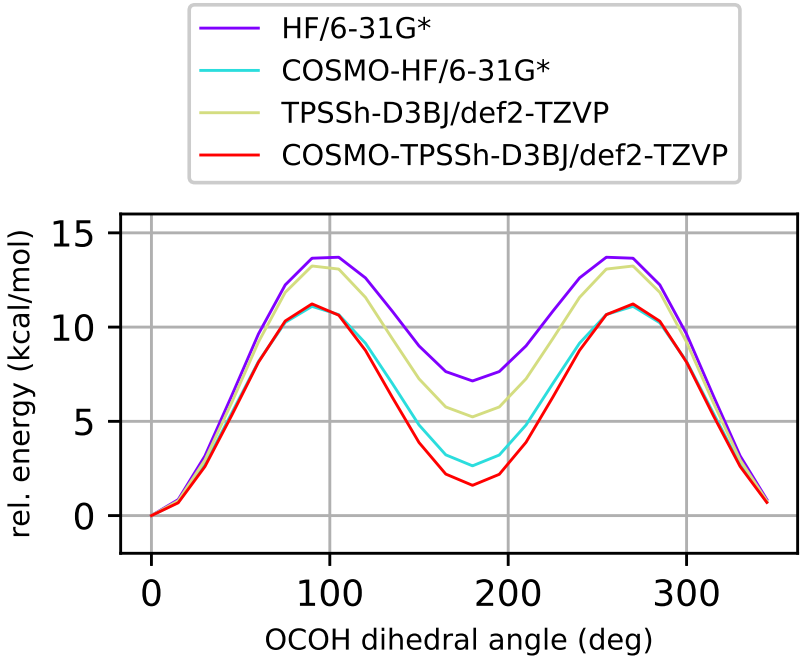


Figure 5.2: QM torsion drive of acetic acid carboxyl dihedral angle for HF and TPSSh methods. In each case, implicit solvation with COSMO reduces the energy barrier and the relative minima energy to 5-7 kcal/mol and 2-3 kcal/mol respectively.

5.3.1 *Ab initio* torsion drive of acetic acid

Our QM calculations in gas phase and implicit solvent show that *syn* is highly favored in the gas phase but the difference becomes less significant in solvent. From the torsion drive obtained via *ab initio* QM calculations, the *syn*-*anti* energy difference is 7.14 kcal/mol with the basic HF/6-31G* method and decreases to 5.24 kcal/mol with the higher level of theory using the TPSSh functional (Figure 5.2). With COSMO, a similar trend is seen in which the higher level of theory yields a smaller energy difference between the *syn* and *anti* structures. With either level of theory, adding implicit solvent significantly lowers the relative energy difference between *syn* and *anti* from 5-7 kcal/mol to 2-3 kcal/mol. A 5-7 kcal/mol difference is large enough that such configurations would occur only extremely rarely, whereas 2-3 kcal/mol is enough that such conformations will occur sporadically in solution (3-7% of the time) and could potentially easily be stabilized by interactions with a

nearby receptor or other biomolecule with a strain energy no larger than that reported in many binding interactions,[204, 260] making it potentially relevant functionally.

We now turn our focus to the energy barrier from the *syn* state to the *anti* state. This feature is not particularly critical in molecular simulation, as in most cases systems will be at equilibrium given sufficient relaxation time and sampling. That being said, the energy barrier has implications for interconversion between the two states. One conformation may be more structurally relevant than the other in certain scenarios, and a modeler may wish to achieve an accurate representation of the populations of both conformations. The barrier associated with the rotation of the carboxyl dihedral angle determines how easy it is to interconvert between and sample different conformations. From our QM results, we see a large energetic cost or barrier of 13-14 kcal/mol separating the *syn* form from the *anti* form in gas phase. Solvation with COSMO reduces this barrier height to around 11 kcal/mol.

Overall, the relative energy difference between the *syn* and *anti* conformations of acetic acid appears not very large, especially in the aqueous conditions relevant to biochemistry. The relative energy comparisons from the QM torsion drives are summarized in the top four lines of Table 5.1. Note that, from our QM results, these are relative energies rather than relative free energies; with MD in the following section; we obtain relative free energies.

5.3.2 *Ab initio* geometry optimizations from molecular dynamics configurations

To rule out the possibility that stabilization of the *anti* form in the torsion drive is due to implicit solvent model alone, and to determine whether explicit water might provide additional stabilization, we examined acetic acid with explicit water molecules. We first examined trihydrated *syn* and *anti* acetic acid (details in supporting information). However, recent work on the microhydration of acetic acid suggests that the particular arrange-

ment of water molecules may be important when comparing energetic stabilities of acetic acid conformations.[160] Given that we are interested in solution-phase behavior, the actual solution-phase geometry of water molecules around acetic acid then becomes very important. In order to reduce any artificial effects of water placement, we sample various conformations of water molecules around acetic acid by running molecular dynamics simulations for each of the *syn* and *anti* forms. Both simulations were run using the *syn* charges for context as these are predominantly used in present-day molecular simulations. Configurations of acetic acid with its five nearest waters were clustered by root-mean-square deviation of geometries. The most common arrangements were extracted for QM optimization in implicit solvent using the method COSMO-TPSSh-D3BJ/def2-TZVP.

The violin plots in Figure 5.3 display the distributions for the relative energies of the *syn* (left side) and *anti* (right side) pentahydrated configurations of acetic acid. Here, we see that the distribution for the *syn* configurations skews toward lower energies compared to the *anti* configurations. However, the energy values of the extrema are quite similar, and the population of the *anti* form at low energies is nonnegligible.

5.3.3 MD simulations with umbrella sampling along carboxyl dihedral angle

Our above QM calculations study only conformational energies, not free energies, so we computed the one-dimensional free energy landscape (the potential of mean force, or PMF) of rotating the acetic acid dihedral angle with classical molecular dynamics. The MD results in gas phase and in explicit solvent are in qualitative agreement with our QM data and indicate that water substantially increases the stability of the *anti* conformation. We considered various force fields, partial charge sets, and solvation states for a total of eight PMFs. Atomic partial charges are held fixed within our simulations, as is typical in MD, but these charges are

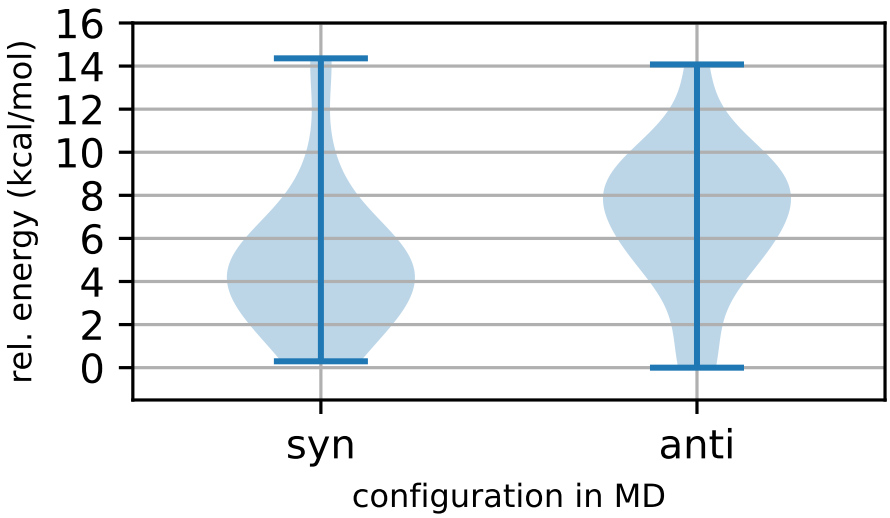


Figure 5.3: Violin plots for relative energy distributions of pentahydrated *syn* and *anti* conformations of acetic acid. The data represent COSMO-TPSSh-D3BJ/def2-TZVP energies of configurations taken from MD simulations of the *syn* form (14 snapshots) and the *anti* form (17 snapshots).

sensitive to the molecular conformation when assigning charges, so we assigned charges using both conformations. Hereafter we use the notation `SC` for acetic acid partial charges obtained from the *syn* conformation and `AC` for charges obtained from the *anti* conformation. Error bars on the PMFs are obtained from the MBAR estimator.[273] We present a comprehensive comparison in Figure 5.4 and in Table 5.1 and discuss each of these three factors (force field, charge set, and solvation state) separately.

Considering the GAFF and GAFF2 force fields, the PMFs are in good agreement with each other in both gas and aqueous phases as well as with either `SC` or `AC` (Figures 5.4, 5.5). We observe consistent relative free energies between the *syn* and *anti* minima. In gas phase, for the `SC` solute, the *syn* structure is favored in free energy by 6.2±0.2 kcal/mol with GAFF and 5.9±0.2 kcal/mol with GAFF2 (Figure 5.5). In aqueous phase, the *anti* structure is favored in free energy by -0.7±0.1 kcal/mol with GAFF and -1.4±0.1 kcal/mol with GAFF2 (Figure 5.4, teal vs. brown). These qualitative conclusions are the same when considering the `AC` solute. Thus, GAFF and GAFF2 give very similar results for the conformational equilibrium

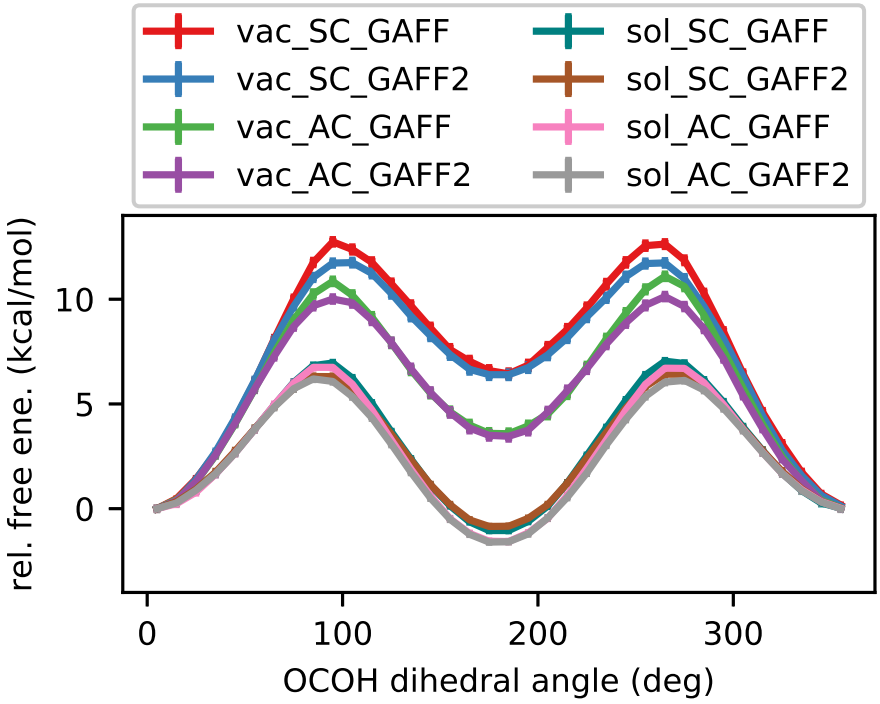


Figure 5.4: PMFs of rotating the acetic acid carboxyl dihedral angle. We consider variations on the force field (GAFF, GAFF2), solute AM1-BCC partial charges (starting from *syn* or *anti*), and solvation state (gas phase, explicit TIP3P waters).

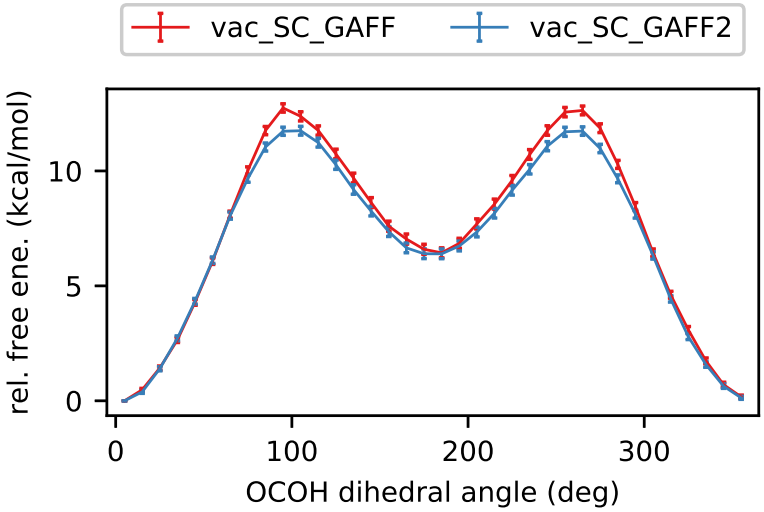


Figure 5.5: Comparison of GAFF and GAFF2 force fields in PMFs of rotating the acetic acid carboxyl dihedral angle. Both are in strong agreement with each other. The PMFs displayed in this figure came from gas phase simulations with *syn* charges. Similar conclusions were drawn for PMFs from aqueous simulations and from using *anti* charges (Figure 5.4).

Method	Solvation	minima[a]	barrier
HF/6-31G*	gas	7.1	13.7
HF/6-31G*	COSMO	2.8	11.2
TPSSh/def2-TZVP[b]	gas	5.2	13.2
TPSSh/def2-TZVP	COSMO	1.6	11.2
vac_SC_GAFF	gas	6.2±0.2	12.7±0.3
vac_SC_GAFF2	gas	5.9±0.2	11.7±0.3
vac_AC_GAFF	gas	3.4±0.2	11.0±0.3
vac_AC_GAFF2	gas	3.3±0.2	10.1±0.3
sol_SC_GAFF	TIP3P	-0.8±0.1	7.0±0.2
sol_SC_GAFF2	TIP3P	-0.7±0.1	6.4±0.2
sol_AC_GAFF	TIP3P	-1.3±0.1	6.7±0.2
sol_AC_GAFF2	TIP3P	-1.4±0.1	6.1±0.2

[a] All relative energy differences are taken with respect to acetic acid's *syn* conformation.
[b] Dispersion corrections added with all TPSSh calculations in this work. See details in text.

Table 5.1: Summary of relative energy differences between *syn* and *anti* conformations of acetic acid as well as free energy barriers of interconversion. The first four lines are results from QM torsion drives, and the last eight from umbrella sampling are from atomistic molecular dynamics simulations. Energies are listed in units of kcal/mol.

of acetic acid which holds true regardless of the partial charge set. Overall these results, at least within the classical framework, indicate that explicit solvent provides approximately 5-8 kcal/mol of stabilization of the *anti* conformation relative to the *syn* conformation. This trend is in the same direction as that provided by COSMO implicit solvent, but provides further stabilization.

We also compare the two force fields in terms of the conformational transition barriers. We note that the GAFF barrier height is higher than the GAFF2 barrier in each pairwise combination of the two force fields with various solvent and charge models. The barrier height differences are 0.9±0.4 kcal/mol in gas phase (compare barrier heights in Figure 5.4 for red vs. blue and for green vs. purple). The rotational barriers differ by 0.6±0.2 kcal/mol in aqueous phase (compare barrier heights in Figure 5.4 for teal vs. brown and for pink vs. gray). For both gaseous and aqueous states, the effects of the partial charges on the PMFs are stronger than those of the force field. For example, in Figure 5.4, the red and green curves are more distinct from each other, while the red and blue curves are more similar.

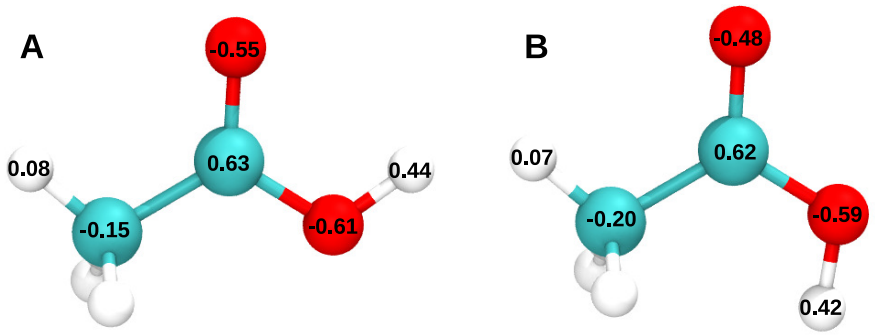


Figure 5.6: AM1-BCC charges generated for (a) *syn* and (b) *anti* configurations of acetic acid.

Since the partial charges of the solute may affect the PMFs more so than the force field, as shown here, one should carefully consider other likely conformations when assigning partial charges. Next we further investigate the solute partial charge sets.

There is a pronounced difference in the PMFs depending on the conformation used to charge acetic acid (Figure 5.6). Charges are typically fixed throughout a molecular dynamics simulation, meaning that initial charge assignment is important for capturing correct energetics throughout a simulation. The free energy difference between the *syn* and *anti* structures is notably larger in gas phase than in water. When we use the *syn* form to obtain AM1-BCC charges (`SC`), the gas phase PMFs are higher in energy for both the barrier height and the two minima (Figure 5.7(a)) compared to using the *anti* form to obtain AM1-BCC charges (`AC`). Qualitatively, the `SC` set is slightly stronger in magnitude than the `AC` set, meaning a slightly stronger polarization along the bonds of the carboxyl group; this is consistent with the intramolecular hydrogen bonding aspect of the *syn* conformation. The stronger `SC` partial charges contribute to increased stabilization of the lower-energy *syn* structure in gas phase, which results in a greater free energy difference and barrier height compared to `AC`. On the other hand, in water, (Figure 5.7(b)), *syn* and *anti* are closer in relative free energy for `SC` than for `AC`. In this setting, *syn* is higher in energy than *anti*. Once again, the stronger `SC` partial charges contribute to increased stabilization of *syn*, in this case via more stabilizing

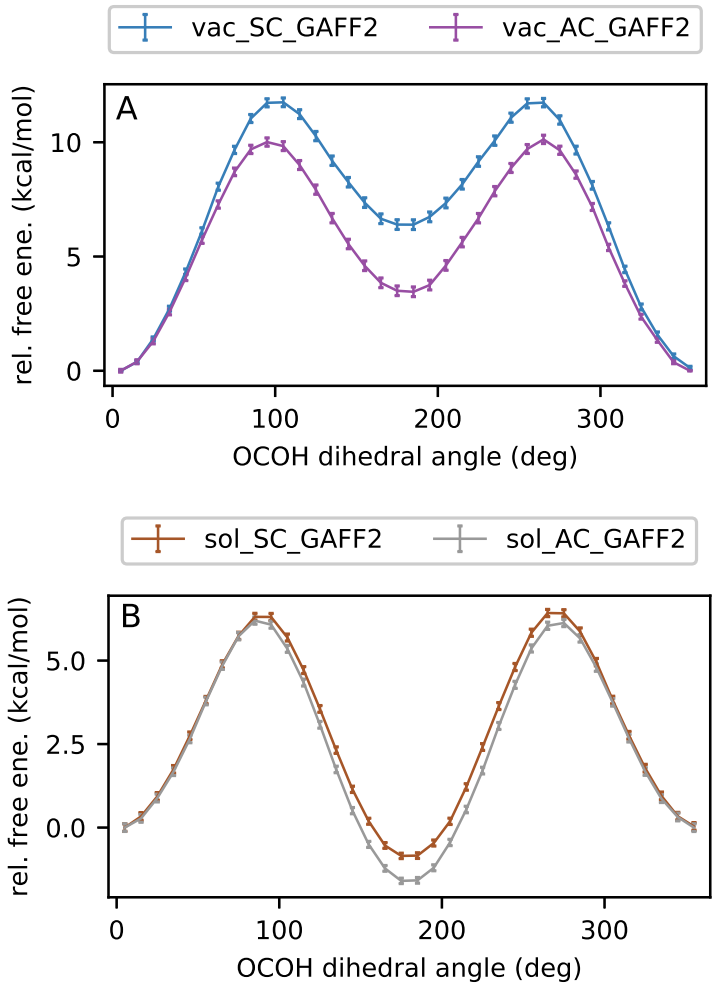


Figure 5.7: Comparison of *syn* and *anti* solute charges in PMFs of rotating the acetic acid carboxyl dihedral angle. In each situation with *anti* charges (A) and *syn* charges (B), the `AC` set more strongly stabilizes the *anti* conformation than the `SC` set.

interactions with the solvent. Here, the two minima are closer in free energy. Therefore we see again that the relative free energies at the minima are governed more strongly by solute charges than by force field.

We take a final look at the MD PMFs in the lens of gaseous versus aqueous phases. These results are in harmony with earlier work on ibuprofen (a carboxylic acid) which found that the *syn* conformation was favorable in vacuum but the *anti* conformation was slightly preferred in water.[154] The major takeaway from the aqueous phase PMFs is that the *anti* conformation of acetic acid is the lower free energy state in solution due to an increased ability to form stabilizing interactions with the solvent. This conclusion qualitatively parallels the result obtained with COSMO-QM calculations on microhydrated acetic acid which showed that the *anti* conformation is lower in energy than the *syn* conformation by about 1.6 kcal/mol, at least for certain arrangements of water molecules.

Overall, the MD results are qualitatively consistent with QM calculations in determination of relative energy differences of the minima and energy barriers for conformational interconversion. The `SC` charge set seems better than the `AC` set in reproducing the relative energy differences obtained with QM DFT in gas phase and in implicit solvent, consistent with our previous practice of considering this conformation more important when assigning charges.

To summarize our PMF results, we considered the effects of force field, charge set, and solvation state on the relative minima free energies as well as on the transition barriers between the two minima. The force fields GAFF and GAFF2 yielded generally similar results to each other. The PMFs in both gas phase and aqueous phase revealed strong dependence on solute charges, especially at the minima. More specifically, the set of partial charges assigned to acetic acid is sensitive to the orientation of O–H in the carboxyl group, leading to variations of up to several kcal/mol in the free energy difference between the *syn* and *anti* structures. Lastly, the dihedral rotation free energy barriers between the *syn*-*anti* conformations are more dependent on the charge set than the force field in gas phase

simulations, while they are more influenced by the force field in aqueous phase simulations. All eight PMFs, obtained from permutation of the force field, solute charges, and solvation state are summarized in Figure 5.4 and Table 5.1.

5.4 Conclusions

Our results call into question the conventional wisdom that carboxylic acids will almost always be in the "more stable" *syn* conformation in biomolecular systems. Typically, the increased stability of the *syn* form is understood to be from the stabilizing intramolecular interaction between the hydrogen atom in the hydroxyl group and the carbonyl oxygen. This idea is in tune with gas phase results we present in this work. However, in aqueous phase, we conclude that the *anti* state may nearly be as populated as the *syn* state due to stabilizing interactions from the solvent. Thus, for MD studies that involve a carboxylic acid or other functional group with possible intramolecular hydrogen bonds, it may be necessary to ensure sufficient sampling of all potentially relevant conformations in solution. This can be challenging given the particularly large barrier associated with rotation of the carboxylic acid torsion.

Our findings also have implications for partial charge calculations for parameter assignment for MD simulations. Carboxylic acids are a case in which neither partial charge set adequately represents the electrostatics of the solute as it samples various conformations. When generating an empirical force field, such as for a small molecule ligand, charges are typically computed for a particular given conformation. These fixed charges are then used for scenarios involving conformational change. In this work, we observe that different solute charges may lead to deviations in relative free energies to as large as 3 kcal/mol. Interconversion is not expected to be frequent, given that the torsional barrier is at least 6 kcal/mol. For that reason, one may wish to treat *syn* and *anti* conformation charges individually, though

this could present difficulties in cases that interconversion is needed for convergence (*e.g.,* a carboxylic acid in a binding site where one conformation forms better contacts than the other). As an alternative approach, the use of polarizable charges may provide a more holistic picture of the carboxyl group's variable nature.

The carboxyl conformational equilibrium has implications for several other types of studies. Hydration free energy calculations may lead to results which depend substantially on the starting conformation. For example, kinetic trapping into one particular conformation can lead to computed hydration free energies which are sensitive to starting conformation and vary by more than 2 kcal/mol because of large torsional barriers.[154] This work also informs efforts to accurately calculate pK_{a} values for ionizable side chains in proteins, *i.e.,* aspartate and glutamate.[303, 174, 18, 26, 150, 335, 292, 283, 324, 9, 8] An accurate insight into the preferred aqueous phase structure of the carboxyl group is important for catalysis, with impacts in atmospheric science and industrial processes.[162] Further impact may be in crystal engineering and drug co-crystallization, in which the carboxyl group is often used to promote aqueous solubility.[65] Theoretical studies on proton transfer such as on solvated acetic acid[107] or on green fluorescent protein[264, 176] typically employ the *syn* conformation due to its expected energetic preference; however, it is worth investigating possible adaptations of carboxyl groups to their local environments. Being aware of the carboxyl moiety's nuanced conformational preferences in different environments may thus lead to better insight for calculated properties, reactivity, and molecular design.

5.5 Acknowledgments

The authors thank Prof. Filipp Furche and Matthew Agee for helpful discussions on QM methods and for support in using the Turbomole software package, respectively. VTL acknowledges funding the National Science Foundation Graduate Research Fellowship Pro-

gram. DLM appreciates financial support from the National Institutes of Health (1R01-GM108889-01) and the National Science Foundation (CHE 1352608), and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513.

5.6 Supplementary Information

Appendix D contains the raw data and Python scripts for all figures, configuration input files for all calculations, initial and final coordinates of all structures, Python script for computing free energies from PMFs, computations and discussion on trihydrated acetic acid, comparison of acetic acid geometries to existing literature, and a discussion on choice of QM method with additional calculations on water hexamers.

Bibliography

[1] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, Sept. 2015.

[2] C. Acerete, J. Catalan, F. Fabero, M. Sanchez-Cabezudo, R. M. Claramunt, and J. Elguero. Structure and Basicity of 2-Guanidiobenzimidazoles. *HETEROCYCLES*, 26(6):1581, 1987.

[3] A. Aksimentiev and K. Schulten. Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophysical Journal*, 88(6):3745–3761, June 2005.

[4] A. A. Alabi, M. I. Bahamonde, H. J. Jung, J. I. Kim, and K. J. Swartz. Portability of paddle motif function and pharmacology in voltage sensors. *Nature*, 450(7168):370–375, Nov. 2007.

[5] M. Aldeghi, B. L. de Groot, and V. Gapsys. Accurate Calculation of Free Energy Changes upon Amino Acid Mutation. In T. Sikosek, editor, *Computational Methods in Protein Evolution*, Methods in Molecular Biology, pages 19–47. Springer, New York, NY, 2019.

[6] G. L. Amidon, H. Lennernäs, V. P. Shah, and J. R. Crison. A Theoretical Basis for a Biopharmaceutic Drug Classification: The Correlation of in Vitro Drug Product Dissolution and in Vivo Bioavailability. *Pharm Res*, 12(3):413–420, Mar. 1995.

[7] H. C. Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52(1):24–34, Oct. 1983.

[8] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. Prediction of Ph-dependent Properties of Proteins. *Journal of Molecular Biology*, 238(3):415–436, May 1994.

[9] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. The Determinants of pKas in Proteins. *Biochemistry*, 35(24):7819–7833, Jan. 1996.

[10] C. T. Armstrong, P. E. Mason, J. L. R. Anderson, and C. E. Dempsey. Arginine side chain interactions and the role of arginine as a gating charge carrier in voltage sensitive ion channels. *Sci. Rep.*, 6(1):1–10, Feb. 2016.

[11] J. L. Asensio, M. Martin-Pastor, and J. Jimenez-Barbero. The use of CVFF and CFF91 force fields in conformational analysis of carbohydrate molecules. Comparison with AMBER molecular mechanics and dynamics calculations for methyl α-lactoside. *International Journal of Biological Macromolecules*, 17(3):137–148, Jan. 1995.

[12] B. J. Aungst. Novel Formulation Strategies for Improving Oral Bioavailability of Drugs with Poor Membrane Permeation or Presystemic Metabolism. *Journal of Pharmaceutical Sciences*, 82(10):979–987, Oct. 1993.

[13] J. Avorn. The $2.6 Billion Pill — Methodologic and Policy Considerations. *N. Engl. J. Med.*, 372(20):1877–1879, May 2015. _eprint: https://doi.org/10.1056/NEJMp1500848.

[14] E. Awoonor-Williams and C. N. Rowley. Molecular simulation of nonfacilitated membrane permeation. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858(7, Part B):1672–1687, July 2016.

[15] C. Ballatore, D. M. Huryn, and A. B. Smith. Carboxylic Acid (Bio)Isosteres in Drug Design. *ChemMedChem*, 8(3):385–395, Mar. 2013.

[16] C. C. Bannan, G. Calabró, D. Y. Kyu, and D. L. Mobley. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *J. Chem. Theory Comput.*, 12(8):4015–4024, Aug. 2016.

[17] H. A. Barenholtz and D. C. McLeod. Loratadine: A Nonsedating Antihistamine with Once-Daily Dosing:. *DICP*, June 1989.

[18] D. Bashford and M. Karplus. pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry*, 29(44):10219–10225, Nov. 1990.

[19] D. M. Bates and G. S. Tschumper. CCSD(T) Complete Basis Set Limit Relative Energies for Low-Lying Water Hexamer Structures. *J. Phys. Chem. A*, 113(15):3555–3559, Apr. 2009.

[20] J. Baucom, T. Transue, M. Fuentes-Cabrera, J. M. Krahn, T. A. Darden, and C. Sagui. Molecular dynamics simulations of the d(CCAACGTTGG)2 decamer in crystal environment: Comparison of atomic point-charge, extra-point, and polarizable force fields. *J. Chem. Phys.*, 121(14):6998–7008, Oct. 2004.

[21] K. Beauchamp, A. Rustenburg, A. Rizzi, J. Behr, G. Matos, L. Wang, R. McGibbon, D. Mobley, and J. Chodera. Openmoltools.

[22] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, Apr. 1993.

[23] C. H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268, Oct. 1976.

[24] T. K. Berger and E. Y. Isacoff. The Pore of the Voltage-Gated Proton Channel. *Neuron*, 72(6):991–1000, Dec. 2011.

[25] M. Bernetti, M. Masetti, W. Rocchia, and A. Cavalli. Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem.*, 70(1):143–171, 2019. _eprint: https://doi.org/10.1146/annurev-physchem-042018-052340.

[26] P. Beroza and D. A. Case. Including Side Chain Flexibility in Continuum Electrostatic Calculations of Protein Titration. *J. Phys. Chem.*, 100(51):20156–20163, Jan. 1996.

[27] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539, June 1994.

[28] D. W. Borhani and D. E. Shaw. The future of molecular dynamics simulations in drug discovery. *J Comput Aided Mol Des*, 26(1):15–26, Jan. 2012.

[29] K. J. Bowers, D. E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, pages 43–43, Nov. 2006.

[30] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[31] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, Jan. 2007.

[32] F. Cailliez and P. Pernot. Statistical approaches to forcefield calibration and prediction uncertainty in molecular simulation. *J. Chem. Phys.*, 134(5):054124, Feb. 2011.

[33] M. Capasso, M. K. Bhamrah, T. Henley, R. S. Boyd, C. Langlais, K. Cain, D. Dinsdale, K. Pulford, M. Khan, B. Musset, V. V. Cherny, D. Morgan, R. D. Gascoyne, E. Vigorito, T. E. DeCoursey, I. C. M. MacLennan, and M. J. S. Dyer. HVCN1 modulates BCR signal strength via regulation of BCR-dependent generation of reactive oxygen species. *Nat. Immunol.*, 11(3):265–272, Mar. 2010.

[34] A. E. Cardenas and R. Elber. Computational study of peptide permeation through membrane: Searching for hidden slow variables. *Mol. Phys.*, 111(22-23):3565–3578, Dec. 2013. _eprint: https://doi.org/10.1080/00268976.2013.842010.

[35] E. P. Carpenter, K. Beis, A. D. Cameron, and S. Iwata. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, 18(5):581–586, Oct. 2008.

[36] W. A. Carterall. Molecular Mechanisms of Gating and Drug Block of Sodium Channels. In *Sodium Channels and Neuronal Hyperexcitability*, pages 206–225. John Wiley & Sons, Ltd, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470846682.ch14.

[37] D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman, AMBER 2016, University of California, San Francisco, 2016.

[38] A. Chamberlin, F. Qiu, S. Rebolledo, Y. Wang, S. Y. Noskov, and H. P. Larsson. Hydrophobic plug functions as a gate in voltage-gated proton channels. *Proc Natl Acad Sci USA*, 111(2):E273, Jan. 2014.

[39] G. Chaves, C. Derst, A. Franzen, Y. Mashimo, R. Machida, and B. Musset. Identification of an HV1 voltage-gated proton channel in insects. *FEBS J.*, 283(8):1453–1464, 2016. _eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1111/febs.13680.

[40] C. Chen. Fast exploration of an optimal path on the multidimensional free energy surface. *PLOS ONE*, 12(5):e0177740, May 2017.

[41] D. Chen, N. Oezguen, P. Urvil, C. Ferguson, S. M. Dann, and T. C. Savidge. Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Sci. Adv.*, 2(3):e1501240, Mar. 2016.

[42] V. V. Cherny and T. E. DeCoursey. Ph-Dependent Inhibition of Voltage-Gated H+ Currents in Rat Alveolar Epithelial Cells by Zn2+ and Other Divalent Cations. *J Gen Physiol*, 114(6):819–838, Dec. 1999.

[43] C. Chipot and J. Comer. Subdiffusion in Membrane Permeation of Small Molecules. *Sci. Rep.*, 6(1):1–14, Nov. 2016.

[44] C. Chipot and A. Pohorille. Calculating Free Energy Differences Using Perturbation Theory. In C. Chipot and A. Pohorille, editors, *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer Series in CHEMICAL PHYSICS, pages 33–75. Springer, Berlin, Heidelberg, 2007.

[45] J. Chocholoušová, J. Vacek, and P. Hobza. Acetic Acid Dimer in the Gas Phase, Nonpolar Solvent, Microhydrated Environment, and Dilute and Concentrated Acetic Acid: Ab Initio Quantum Chemical and Molecular Dynamics Simulations. *J. Phys. Chem. A*, 107(17):3086–3092, May 2003.

[46] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande. Alchemical free energy methods for drug discovery: Progress and challenges. *Current Opinion in Structural Biology*, 21(2):150–160, Apr. 2011.

[47] L. T. Chong, A. S. Saglam, and D. M. Zuckerman. Path-sampling strategies for simulating rare events in biomolecular systems. *Current Opinion in Structural Biology*, 43:88–94, Apr. 2017.

[48] E. A. Cino, W.-Y. Choy, and M. Karttunen. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 8(8):2725–2740, Aug. 2012.

[49] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne, and W. L. Jorgensen. Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.*, 12(5):2312–2323, May 2016.

[50] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, and C. Chipot. The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *J. Phys. Chem. B*, 119(3):1129–1151, Jan. 2015.

[51] J. Comer, K. Schulten, and C. Chipot. Calculation of Lipid-Bilayer Permeabilities Using an Average Force. *J. Chem. Theory Comput.*, 10(2):554–564, Feb. 2014.

[52] J. Comer, K. Schulten, and C. Chipot. Diffusive Models of Membrane Permeation with Explicit Orientational Freedom. *J. Chem. Theory Comput.*, 10(7):2710–2718, July 2014.

[53] J. Comer, K. Schulten, and C. Chipot. Permeability of a Fluid Lipid Bilayer to Short-Chain Alcohols from First Principles. *J. Chem. Theory Comput.*, 13(6):2523–2532, June 2017.

[54] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, May 1995.

[55] A. Crespo, A. Rodriguez-Granillo, and V. T. Lim. Quantum-Mechanics Methodologies in Drug Discovery: Applications of Docking and Scoring in Lead Optimization. *CTMC*, 17(23), Aug. 2017.

[56] E. E. Dahlke, R. M. Olson, H. R. Leverentz, and D. G. Truhlar. Assessment of the Accuracy of Density Functionals for Prediction of Relative Energies and Geometries of Low-Lying Isomers of Water Hexamers. *J. Phys. Chem. A*, 112(17):3976–3984, May 2008.

[57] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, June 1993.

[58] E. Darve, D. Rodríguez-Gómez, and A. Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.*, 128(14):144120, Apr. 2008.

[59] L. D'Ascenzo and P. Auffinger. A comprehensive classification and nomenclature of carboxyl–carboxyl(ate) supramolecular motifs and related catemers: Implications for biomolecular systems. *Acta Cryst B*, 71(2):164–175, Apr. 2015.

[60] S. E. David, P. Timmins, and B. R. Conway. Impact of the counterion on the solubility and physicochemical properties of salts of carboxylic acid drugs. *Drug Dev. Ind. Pharm.*, 38(1):93–103, Jan. 2012.

[61] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.*, 59(9):4035–4061, May 2016.

[62] T. E. DeCoursey. Voltage-Gated Proton Channels Find Their Dream Job Managing the Respiratory Burst in Phagocytes. *Physiology*, 25(1):27–40, Feb. 2010.

[63] K. R. DeMarco, S. Bekker, C. E. Clancy, S. Y. Noskov, and I. Vorobyov. Digging into Lipid Membrane Permeation for Cardiac Ion Channel Blocker d-Sotalol with All-Atom Simulations. *Front. Pharmacol.*, 9, 2018.

[64] J. L. Derissen. A reinvestigation of the molecular structure of acetic acid monomer and dimer by gas electron diffraction. *Journal of Molecular Structure*, 7(1):67–80, Jan. 1971.

[65] G. R. Desiraju. Crystal Engineering: From Molecule to Crystal. *J. Am. Chem. Soc.*, 135(27):9952–9967, July 2013.

[66] J. M. Diamond and Y. Katz. Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water. *J. Memb. Biol.*, 17(1):121–154, Dec. 1974.

[67] C. J. Dickson, V. Hornak, R. A. Pearlstein, and J. S. Duca. Structure–Kinetic Relationships of Passive Membrane Permeation from Multiscale Modeling. *J. Am. Chem. Soc.*, 139(1):442–452, Jan. 2017.

[68] C. J. Dickson, B. D. Madej, Å. A. Skjevik, R. M. Betz, K. Teigen, I. R. Gould, and R. C. Walker. Lipid14: The Amber Lipid Force Field. *J. Chem. Theory Comput.*, 10(2):865–879, Feb. 2014.

[69] J. A. DiMasi. The Value of Improving the Productivity of the Drug Development Process. *Pharmacoeconomics*, 20(3):1–10, Dec. 2002.

[70] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski. The price of innovation: New estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185, Mar. 2003.

[71] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, and W. L. Jorgensen. LigParGen web server: An automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res*, 45(W1):W331–W336, July 2017.

[72] J. Domański, G. Hedger, R. B. Best, P. J. Stansfeld, and M. S. P. Sansom. Convergence and Sampling in Determining Free Energy Landscapes for Membrane Protein Association. *J. Phys. Chem. B*, 121(15):3364–3375, Apr. 2017.

[73] S. Dorairaj and T. W. Allen. On the thermodynamic stability of a charged arginine side chain in a transmembrane helix. *PNAS*, 104(12):4943–4948, Mar. 2007.

[74] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, July 2017.

[75] F. Eckert and A. Klamt. Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE J.*, 48(2):369–385, Feb. 2002.

[76] J. N. Ehrman, C. C. Bannan, V. T. Lim, N. Thi, D. Y. Kyu, and D. L. Mobley. Improving Force Fields by Identifying and Characterizing Small Molecules with Parameter Inconsistencies, Sept. 2019.

[77] A. El Chemaly, Y. Okochi, M. Sasaki, S. Arnaudeau, Y. Okamura, and N. Demaurex. VSOP/Hv1 proton channels sustain calcium entry, neutrophil migration, and superoxide production by limiting cell depolarization and acidification. *J Exp Med*, 207(1):129–139, Jan. 2010.

[78] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577–8593, Nov. 1995.

[79] J. P. Ewen, C. Gattinoni, F. M. Thakkar, N. Morgan, H. A. Spikes, and D. Dini. A Comparison of Classical Force-Fields for Molecular Dynamics Simulations of Lubricants. *Materials*, 9(8):651, Aug. 2016.

[80] S. Fan, B. I. Iorga, and O. Beckstein. Prediction of octanol-water partition coefficients for the SAMPL6-$$\log P$$logP molecules using molecular dynamics simulations with OPLS-AA, AMBER and CHARMM force fields. *J Comput Aided Mol Des*, 34(5):543–560, May 2020.

[81] M. Feig and B. M. Pettitt. Structural Equilibrium of DNA Represented with Different Force Fields. *Biophysical Journal*, 75(1):134–149, July 1998.

[82] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.*, 103(11):4613–4621, Sept. 1995.

[83] G. Fiorin, M. L. Klein, and J. Hénin. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.*, 111(22-23):3345–3362, Dec. 2013.

[84] H. Fischer. Function of proton channels in lung epithelia. *Wiley Interdiscip. Rev. Membr. Transp. Signal.*, 1(3):247–258, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wmts.17.

[85] N. Foloppe and A. D. M. Jr. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.*, 21(2):86–104, 2000. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291096-987X%2820000130%2921%3A2%3C86%3A%3AAID-JCC2%3E3.0.CO%3B2-G.

[86] J. A. Freites, D. J. Tobias, G. von Heijne, and S. H. White. Interface connections of a transmembrane voltage sensor. *PNAS*, 102(42):15059–15064, Oct. 2005.

[87] M. J. Frisch, M. Head-Gordon, and J. A. Pople. A direct MP2 gradient method. *Chemical Physics Letters*, 166(3):275–280, Feb. 1990.

[88] M. J. Frisch, M. Head-Gordon, and J. A. Pople. Semi-direct algorithms for the MP2 energy and gradient. *Chemical Physics Letters*, 166(3):281–289, Feb. 1990.

[89] J. Gao, K. Kuczera, B. Tidor, and M. Karplus. Hidden thermodynamics of mutant proteins: A molecular dynamics analysis. *Science*, 244(4908):1069–1072, June 1989.

[90] Q. Gao and K. T. Leung. Hydrogen-bonding interactions in acetic acid monohydrates and dihydrates by density-functional theory calculations. *The Journal of Chemical Physics*, 123(7):074325, Aug. 2005.

[91] M. T. Geballe and J. P. Guthrie. The SAMPL3 blind prediction challenge: Transfer energy overview. *J Comput Aided Mol Des*, 26(5):489–496, May 2012.

[92] R. B. Gennis. Interactions of Small Molecules with Membranes: Partitioning, Permeability, and Electrical Effects. In R. B. Gennis, editor, *Biomembranes: Molecular Structure and Function*, Springer Advanced Texts in Chemistry, pages 235–269. Springer, New York, NY, 1989.

[93] C. L. Gentry, R. D. Egleton, T. Gillespie, T. J. Abbruscato, H. B. Bechowski, V. J. Hruby, and T. P. Davis. The effect of halogenation on blood–brain barrier permeability of a novel peptide drug. *Peptides*, 20(10):1229–1238, Oct. 1999.

[94] A. D. Geragotelis, M. L. Wood, H. Göddeke, L. Hong, E. K. Wong, J. A. Freites, F. Tombola, and D. J. Tobias. Voltage-dependent structural models of the human Hv1 proton channel from long-timescale molecular dynamics simulations. *Under review*, 2019.

[95] G. Gerebtzoff, X. Li-Blatter, H. Fischer, A. Frentzel, and A. Seelig. Halogenation of Drugs Enhances Membrane Binding and Permeation. *ChemBioChem*, 5(5):676–684, May 2004.

[96] Z. Ghaemi, M. Minozzi, P. Carloni, and A. Laio. A Novel Approach to the Investigation of Passive Molecular Permeation through Lipid Bilayers from Atomistic Simulations. *J. Phys. Chem. B*, 116(29):8714–8721, July 2012.

[97] E. Gianti, L. Delemotte, M. L. Klein, and V. Carnevale. On the role of water density fluctuations in the inhibition of a proton channel. *PNAS*, 113(52):E8359–E8368, Dec. 2016.

[98] P. Gilli, V. Ferretti, G. Gilli, and P. A. Borea. Enthalpy-entropy compensation in drug-receptor binding. *J. Phys. Chem.*, 98(5):1515–1518, Feb. 1994.

[99] N. Godbout, D. R. Salahub, J. Andzelm, and E. Wimmer. Optimization of Gaussian-type basis sets for local spin density functional calculations. Part I. Boron through neon, optimization technique and validation. *Can. J. Chem.*, 70(2):560–571, Feb. 1992.

[100] C. Gonzalez, S. Rebolledo, M. E. Perez, and H. P. Larsson. Molecular mechanism of voltage sensing in voltage-gated proton channels. *J Gen Physiol*, 141(3):275–285, Mar. 2013.

[101] M. A. González. Force fields and molecular dynamics simulations. *JDN*, 12:169–200, 2011.

[102] V. Govindaraj, H. Ungati, S. R. Jakka, S. Bose, and G. Mugesh. Directing Traffic: Halogen-Bond-Mediated Membrane Transport. *Chemistry – A European Journal*, 25(48):11180–11192, Aug. 2019.

[103] L. Gracia. VMD clustering plugin. `https://github.com/luisico/clustering`, 2018.

[104] S. Grimme. Improved second-order Møller–Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.*, 118(20):9095–9102, May 2003.

[105] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, Apr. 2010.

[106] S. Grimme, S. Ehrlich, and L. Goerigk. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32(7):1456–1465, May 2011.

[107] W. Gu, T. Frigato, T. P. Straatsma, and V. Helms. Dynamic Protonation Equilibrium of Solvated Acetic Acid. *Angewandte Chemie International Edition*, 46(16):2939–2943, Apr. 2007.

[108] J. Guo, S. Durdagi, M. Changalov, L. L. Perissinotti, J. M. Hargreaves, T. G. Back, S. Y. Noskov, and H. J. Duff. Structure Driven Design of Novel Human Ether-A-Go-Go-Related-Gene Channel (hERG1) Activators. *PLOS ONE*, 9(9):e105553, Sept. 2014.

[109] O. Guvench and A. D. MacKerell. Comparison of Protein Force Fields for Molecular Dynamics Simulations. In A. Kukol, editor, *Molecular Modeling of Proteins*, Methods Molecular Biology™, pages 63–88. Humana Press, Totowa, NJ, 2008.

[110] A. T. Hagler, S. Lifson, and P. Dauber. Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 2. A benchmark for the objective comparison of alternative force fields. *J. Am. Chem. Soc.*, 101(18):5122–5130, Aug. 1979.

[111] T. A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.*, 17(5-6):490–519, Apr. 1996.

[112] T. A. Halgren. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.*, 17(5-6):520–552, Apr. 1996.

[113] T. A. Halgren. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem.*, 17(5-6):553–586, Apr. 1996.

[114] T. A. Halgren. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.*, 17(5-6):616–641, Apr. 1996.

[115] T. A. Halgren. MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.*, 20(7):720–729, May 1999.

[116] T. A. Halgren. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J. Comput. Chem.*, 20(7):730–748, 1999. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291096-987X%28199905%2920%3A7%3C730%3A%3AAID-JCC8%3E3.0.CO%3B2-T.

[117] T. A. Halgren and R. B. Nachbar. Merck molecular force field. IV. conformational energies and geometries for MMFF94. *J. Comput. Chem.*, 17(5-6):587–615, Apr. 1996.

[118] P. C. Hariharan and J. A. Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoret. Chim. Acta*, 28(3):213–222, Sept. 1973.

[119] M. Häser and R. Ahlrichs. Improvements on the direct SCF method. *J. Comput. Chem.*, 10(1):104–111, Jan. 1989.

[120] M. Head-Gordon, J. A. Pople, and M. J. Frisch. MP2 energy evaluation by direct methods. *Chemical Physics Letters*, 153(6):503–506, Dec. 1988.

[121] W. J. Hehre, R. Ditchfield, and J. A. Pople. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics*, 56(5):2257–2261, Mar. 1972.

[122] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein. Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *J. Chem. Theory Comput.*, 6(1):35–47, Jan. 2010.

[123] J. Henriques, C. Cragnell, and M. Skepö. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J. Chem. Theory Comput.*, 11(7):3420–3431, July 2015.

[124] R. M. Hernandez-Garcia, N. Barba-Behrens, R. Salcedo, and G. Höjer. Theoretical study of 2-guanidinobenzimidazole. HF, MP2 and DFT calculations. *Journal of Molecular Structure: THEOCHEM*, 637(1):55–72, Oct. 2003.

[125] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[126] A. L. Hickey and C. N. Rowley. Benchmarking Quantum Chemical Methods for the Calculation of Molecular Dipole Moments and Polarizabilities. *J. Phys. Chem. A*, 118(20):3678–3687, May 2014.

[127] S. A. Hollingsworth and R. O. Dror. Molecular Dynamics Simulation for All. *Neuron*, 99(6):1129–1143, Sept. 2018.

[128] L. Hong, I. H. Kim, and F. Tombola. Molecular determinants of Hv1 proton channel inhibition by guanidine derivatives. *Proc Natl Acad Sci USA*, 111(27):9971, July 2014.

[129] L. Hong, M. M. Pathak, I. H. Kim, D. Ta, and F. Tombola. Voltage-Sensing Domain of Voltage-Gated Proton Channel Hv1 Shares Mechanism of Block with Pore Domains. *Neuron*, 77(2):274–287, Jan. 2013.

[130] C. W. Hopkins and A. E. Roitberg. Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem. *J. Chem. Inf. Model.*, 54(7):1978–1986, July 2014.

[131] V. Hornak. Comparison of robustness and quality of three small molecule force-fields. Proceedings of CUP XI, Santa Fe, NM, USA. 2010. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.700.6112&rep=rep1&type=pdf`.

[132] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinforma.*, 65(3):712–725, 2006. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21123.

[133] S. Hossain, A. Kabedev, A. Parrow, C. A. S. Bergström, and P. Larsson. Molecular simulation as a computational pharmaceutics tool to predict drug solubility, solubilization processes and partitioning. *European Journal of Pharmaceutics and Biopharmaceutics*, 137:46–55, Apr. 2019.

[134] J. Huang and A. D. MacKerell. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.*, 34(25):2135–2145, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.23354.

[135] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, Feb. 1996.

[136] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.*, 21(2):132–146, Jan. 2000.

[137] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.*, 23(16):1623–1641, Dec. 2002.

[138] J. P. M. Jämbeck and A. P. Lyubartsev. Exploring the Free Energy Landscape of Solutes Embedded in Lipid Bilayers. *J. Phys. Chem. Lett.*, 4(11):1781–1787, June 2013.

[139] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E*, 56(5):5018–5035, Nov. 1997.

[140] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.*, 78(14):2690–2693, Apr. 1997.

[141] W. Jiang, J. Thirman, S. Jo, and B. Roux. Reduced Free Energy Perturbation/Hamiltonian Replica Exchange Molecular Dynamics Method with Unbiased Alchemical Thermodynamic Axis. *J. Phys. Chem. B*, 122(41):9435–9442, Oct. 2018.

[142] S. Jo and W. Jiang. A generic implementation of replica exchange with solute tempering (REST2) algorithm in NAMD for complex biophysical simulations. *Computer Physics Communications*, 197:304–311, Dec. 2015.

[143] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, July 1983.

[144] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, Nov. 1996.

[145] J. Kalia and K. J. Swartz. Common Principles of Voltage-Dependent Gating for Hv and Kv Channels. *Neuron*, 77(2):214–216, Jan. 2013.

[146] G. Kamath, I. Kurnikov, B. Fain, I. Leontyev, A. Illarionov, O. Butin, M. Olevanov, and L. Pereyaslavets. Prediction of cyclohexane-water distribution coefficient for SAMPL5 drug-like compounds with the QMPFF3 and ARROW polarizable force fields. *J Comput Aided Mol Des*, 30(11):977–988, Nov. 2016.

[147] I. Y. Kanal, J. A. Keith, and G. R. Hutchison. A sobering assessment of small-molecule force field methods for low energy conformer predictions. *Int. J. Quantum Chem.*, 118(5):e25512, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.25512.

[148] I. M. Kapetanovic. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chemico-Biological Interactions*, 171(2):165–176, Jan. 2008.

[149] M. K. Kesharwani, A. Karton, and J. M. L. Martin. Benchmark ab Initio Conformational Energies for the Proteinogenic Amino Acids through Explicitly Correlated Methods. Assessment of Density Functional Methods. *J. Chem. Theory Comput.*, 12(1):444–454, Jan. 2016.

[150] K. P. Kilambi and J. J. Gray. Rapid Calculation of Protein pKa Values Using Rosetta. *Biophysical Journal*, 103(3):587–595, Aug. 2012.

[151] A. Klamt, V. Jonas, T. Bürger, and J. C. W. Lohrenz. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A*, 102(26):5074–5085, June 1998.

[152] A. Klamt and G. Schüürmann. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, (5):799–805, Jan. 1993.

[153] J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell, and R. W. Pastor. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J. Phys. Chem. B*, 114(23):7830–7843, June 2010.

[154] P. V. Klimovich and D. L. Mobley. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J Comput Aided Mol Des*, 24(4):307–316, Apr. 2010.

[155] H. P. Koch, T. Kurokawa, Y. Okochi, M. Sasaki, Y. Okamura, and H. P. Larsson. Multimeric nature of voltage-gated proton channels. *Proc Natl Acad Sci USA*, 105(26):9111, July 2008.

[156] D. A. Kofke. Free energy methods in molecular simulation. *Fluid Phase Equilibria*, 228-229:41–48, Feb. 2005.

[157] A. Köster, T. Spura, G. Rutkai, J. Kessler, H. Wiebeler, J. Vrabec, and T. D. Kühne. Assessing the accuracy of improved force-matched water models derived from Ab initio molecular dynamics simulations. *J. Comput. Chem.*, 37(19):1828–1838, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.24398.

[158] M. L. C. E. Kouwijzer, B. P. V. Eijck, S. J. Kroes, and J. Kroon. Comparison of two force fields by molecular dynamics simulations of glucose crystals: Effect of using ewald sums. *J. Comput. Chem.*, 14(11):1281–1289, 1993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540141104.

[159] M. Krepl, M. Zgarbová, P. Stadlbauer, M. Otyepka, P. Banáš, J. Koča, T. E. Cheatham, P. Jurečka, and J. Šponer. Reference Simulations of Noncanonical Nucleic Acids with Different χ Variants of the AMBER Force Field: Quadruplex DNA, Quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput.*, 8(7):2506–2520, July 2012.

[160] P. Krishnakumar and D. K. Maity. Microhydration of Neutral and Charged Acetic Acid. *J. Phys. Chem. A*, 121(2):493–504, Jan. 2017.

[161] K. Kulleperuma, S. M. Smith, D. Morgan, B. Musset, J. Holyoake, N. Chakrabarti, V. V. Cherny, T. E. DeCoursey, and R. Pomès. Construction and validation of a homology model of the human voltage-gated proton channel hHV1. *Journal of General Physiology*, 141(4):445–465, Mar. 2013.

[162] M. Kumar, D. H. Busch, B. Subramaniam, and W. H. Thompson. Organic Acids Tunably Catalyze Carbonic Acid Decomposition. *J. Phys. Chem. A*, 118(27):5020–5028, July 2014.

[163] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande. To milliseconds and beyond: Challenges in the simulation of protein folding. *Current Opinion in Structural Biology*, 23(1):58–65, Feb. 2013.

[164] O. F. Lange, D. van der Spoel, and B. L. de Groot. Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data. *Biophysical Journal*, 99(2):647–655, July 2010.

[165] T. Lassila, J. Hokkanen, S.-M. Aatsinki, S. Mattila, M. Turpeinen, and A. Tolonen. Toxicity of Carboxylic Acid-Containing Drugs: The Role of Acyl Migration and CoA Conjugation Investigated. *Chem. Res. Toxicol.*, 28(12):2292–2303, Dec. 2015.

[166] A. D. Laurent and D. Jacquemin. TD-DFT benchmarks: A review. *Int. J. Quantum Chem.*, 113(17):2019–2039, 2013.

[167] J. Le. Drug Absorption - Clinical Pharmacology. https://www.merckmanuals.com/professional/clinical-pharmacology/pharmacokinetics/drug-absorption, 2019.

[168] A. R. Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[169] E. O. Lebigot. Uncertainties: a python package for calculations with uncertainties. *URL http://pythonhosted. org/uncertainties*, 2010.

[170] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, Jan. 1988.

[171] C. T. Lee, J. Comer, C. Herndon, N. Leung, A. Pavlova, R. V. Swift, C. Tung, C. N. Rowley, R. E. Amaro, C. Chipot, Y. Wang, and J. C. Gumbart. Simulation-Based Approaches for Determining Membrane Permeability of Small Compounds. *J. Chem. Inf. Model.*, 56(4):721–733, Apr. 2016.

[172] S.-Y. Lee, J. A. Letts, and R. MacKinnon. Functional Reconstitution of Purified Human Hv1 H+ Channels. *Journal of Molecular Biology*, 387(5):1055–1060, Apr. 2009.

[173] S. P. Leelananda and S. Lindert. Computational methods in drug discovery. *Beilstein J. Org. Chem.*, 12(1):2694–2718, Dec. 2016.

[174] H. Li, A. D. Robertson, and J. H. Jensen. Very fast empirical prediction and rationalization of protein pKa values. *Proteins Struct. Funct. Bioinforma.*, 61(4):704–721, 2005.

[175] L. Li, I. Vorobyov, and T. W. Allen. Potential of Mean Force and pKa Profile Calculation for a Lipid Membrane-Exposed Arginine Side Chain. *J. Phys. Chem. B*, 112(32):9574–9587, Aug. 2008.

[176] M. A. Lill and V. Helms. Proton shuttle in green fluorescent protein studied by dynamic simulations. *PNAS*, 99(5):2778–2781, May 2002.

[177] V. T. Lim, A. D. Geragotelis, N. M. Lim, J. A. Freites, F. Tombola, D. M. Mobley, and D. J. Tobias. Insights on small molecule binding to the hv1 proton channel from free energy calculations with molecular dynamics simulations. *Under review*, 2020.

[178] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic Validation of Protein Force Fields against Experimental Data. *PLOS ONE*, 7(2):e32131, Feb. 2012.

[179] P. V. Lishko, I. L. Botchkina, A. Fedorenko, and Y. Kirichok. Acid Extrusion from Human Spermatozoa Is Mediated by Flagellar Voltage-Gated Proton Channel. *Cell*, 140(3):327–337, Feb. 2010.

[180] P. V. Lishko and Y. Kirichok. The role of Hv1 and CatSper channels in sperm activation. *J. Physiol.*, 588(23):4667–4672, 2010.

[181] P. V. Lishko, Y. Kirichok, D. Ren, B. Navarro, J.-J. Chung, and D. E. Clapham. The Control of Male Fertility by Spermatozoan Ion Channels. *Annu. Rev. Physiol.*, 74(1):453–475, Mar. 2012.

[182] S. B. Long, X. Tao, E. B. Campbell, and R. MacKinnon. Atomic structure of a voltage-dependent K+ channel in a lipid membrane-like environment. *Nature*, 450(7168):376–382, Nov. 2007.

[183] P. E. M. Lopes, O. Guvench, and A. D. MacKerell. Current Status of Protein Force Fields for Molecular Dynamics Simulations. In A. Kukol, editor, *Molecular Modeling of Proteins*, Methods in Molecular Biology, pages 47–71. Springer, New York, NY, 2015.

[184] Y. Lou and J. Zhu. Carboxylic Acid Nonsteroidal Anti-Inflammatory Drugs (NSAIDs). In C. Lamberth and J. Dinges, editors, *Bioactive Carboxylic Compound Classes*, pages 221–236. Wiley-VCH Verlag GmbH & Co. KGaA, 2016.

[185] N. Lu, J. K. Singh, and D. A. Kofke. Appropriate methods to combine forward and reverse free-energy perturbation averages. *J. Chem. Phys.*, 118(7):2977–2984, Jan. 2003.

[186] H. V. Ly and M. L. Longo. The Influence of Short-Chain Alcohols on Interfacial Tension, Mechanical Properties, Area/Molecule, and Permeability of Fluid Lipid Bilayers. *Biophysical Journal*, 87(2):1013–1033, Aug. 2004.

[187] S. J. Y. Macalino, V. Gosu, S. Hong, and S. Choi. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.*, 38(9):1686–1701, Sept. 2015.

[188] J. L. MacCallum and D. P. Tieleman. Computer Simulation of the Distribution of Hexane in a Lipid Bilayer: Spatially Resolved Free Energy, Entropy, and Enthalpy Profiles. *J. Am. Chem. Soc.*, 128(1):125–130, Jan. 2006.

[189] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, 102(18):3586–3616, Apr. 1998.

[190] S.-J. Marrink and H. J. C. Berendsen. Simulation of water transport through a lipid membrane. *J. Phys. Chem.*, 98(15):4155–4168, Apr. 1994.

[191] F. Martín-García, E. Papaleo, P. Gomez-Puertas, W. Boomsma, and K. Lindorff-Larsen. Comparing Molecular Dynamics Force Fields in the Essential Subspace. *PLOS ONE*, 10(3):e0121114, Mar. 2015.

[192] G. J. Martyna, M. L. Klein, and M. Tuckerman. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*, 97(4):2635–2643, Aug. 1992.

[193] G. J. Martyna, D. J. Tobias, and M. L. Klein. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.*, 101(5):4177–4189, Sept. 1994.

[194] D. Matthes and B. L. de Groot. Secondary Structure Propensities in Peptide Folding Simulations: A Systematic Comparison of Molecular Mechanics Interaction Schemes. *Biophysical Journal*, 97(2):599–608, July 2009.

[195] J. F. Matthews, G. T. Beckham, M. Bergenstråhle-Wohlert, J. W. Brady, M. E. Himmel, and M. F. Crowley. Comparison of Cellulose Iβ Simulations with Three Carbohydrate Force Fields. *J. Chem. Theory Comput.*, 8(2):735–748, Feb. 2012.

[196] D. P. McNamara, S. L. Childs, J. Giordano, A. Iarriccio, J. Cassidy, M. S. Shet, R. Mannion, E. O'Donnell, and A. Park. Use of a Glutaric Acid Cocrystal to Improve Oral Bioavailability of a Low Solubility API. *Pharm Res*, 23(8):1888–1897, Aug. 2006.

[197] M. G. Medvedev, I. S. Bushmarinov, and K. A. Lyssenko. Z-effect reversal in carboxylic acid associates. *Chem. Commun.*, 52(39):6593–6596, 2016.

[198] M. Mijaković, K. D. Polok, B. Kežić, F. Sokolić, A. Perera, and L. Zoranić. A comparison of force fields for ethanol–water mixtures. *Mol. Simul.*, 41(9):699–712, June 2015. _eprint: https://doi.org/10.1080/08927022.2014.923567.

[199] S. K. Mishra, G. Calabró, H. H. Loeffler, J. Michel, and J. Koča. Evaluation of Selected Classical Force Fields for Alchemical Binding Free Energy Calculations of Protein-Carbohydrate Complexes. *J. Chem. Theory Comput.*, 11(7):3333–3345, July 2015.

[200] S. Miyamoto and P. A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13(8):952–962, 1992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540130805.

[201] D. Mobley. Constructing benchmark/test sets for OpenFF quantum chemistry benchmarks. 2019. `https://github.com/openforcefield/release-1-benchmarking/tree/master/QM_molecule_selection`.

[202] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson, and P. K. Eastman. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.*, 14(11):6076–6092, Nov. 2018.

[203] D. L. Mobley, C. C. Bannan, J. R. Wagner, A. Rizzi, N. M. Lim, and M. Henry. openforcefield/smirnoff99frosst: Version 1.1.0, July 2019.

[204] D. L. Mobley and K. A. Dill. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure*, 17(4):489–498, Apr. 2009.

[205] D. L. Mobley, É. Dumont, J. D. Chodera, and K. A. Dill. Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent. *J. Phys. Chem. B*, 111(9):2242–2254, Mar. 2007.

[206] D. L. Mobley and P. V. Klimovich. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.*, 137(23):230901, Dec. 2012.

[207] C. Møller and M. S. Plesset. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.*, 46(7):618–622, Oct. 1934.

[208] L. Monticelli and D. P. Tieleman. Force Fields for Classical Molecular Dynamics. In L. Monticelli and E. Salonen, editors, *Biomolecular Simulations: Methods and Protocols*, Methods in Molecular Biology, pages 197–213. Humana Press, Totowa, NJ, 2013.

[209] T. A. Montzka, S. Swaminathan, and R. A. Firestone. Reversal of Syn-Anti Preference for Carboxylic Acids along the Reaction Coordinate for Proton Transfer. Implications for Intramolecular Catalysis. *J. Phys. Chem.*, 98(50):13171–13176, Dec. 1994.

[210] I. Moraes, G. Evans, J. Sanchez-Weatherby, S. Newstead, and P. D. S. Stewart. Membrane protein structure determination — The next generation. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1838(1, Part A):78–87, Jan. 2014.

[211] Mu, D. S. Kosov, and G. Stock. Conformational Dynamics of Trialanine in Water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS Force Fields to NMR and Infrared Experiments. *J. Phys. Chem. B*, 107(21):5064–5073, May 2003.

[212] B. Musset, S. M. E. Smith, S. Rajan, D. Morgan, V. V. Cherny, and T. E. DeCoursey. Aspartate 112 is the selectivity filter of the human voltage-gated proton channel. *Nature*, 480(7376):273–277, Dec. 2011.

[213] P. I. Nagy. The syn–anti equilibrium for the COOH group reinvestigated. Theoretical conformation analysis for acetic acid in the gas phase and in solution. *Computational and Theoretical Chemistry*, 1022:59–69, Oct. 2013.

[214] P. I. Nagy. Competing Intramolecular vs. Intermolecular Hydrogen Bonds in Solution. *Int. J. Mol. Sci.*, 15(11):19562–19633, Oct. 2014.

[215] P. I. Nagy, D. A. Smith, G. Alagona, and C. Ghio. Ab initio studies of free and monohydrated carboxylic acids in the gas phase. *J. Phys. Chem.*, 98(2):486–493, Jan. 1994.

[216] P. S. Nerenberg and T. Head-Gordon. New developments in force fields for biomolecular simulations. *Current Opinion in Structural Biology*, 49:129–138, Apr. 2018.

[217] Y. Niu, D. Shi, L. Li, J. Guo, H. Liu, and X. Yao. Revealing inhibition difference between PFI-2 enantiomers against SETD7 by molecular dynamics simulations, binding free energy calculations and unbinding pathway analysis. *Sci. Rep.*, 7(1):1–11, Apr. 2017.

[218] G. Odegard, T. Clancy, and T. Gates. Prediction of Mechanical Properties of Polymers with Various Force Fields. In *46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*. American Institute of Aeronautics and Astronautics, 2005. _eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2005-1850.

[219] Y. Okochi, Y. Aratani, H. A. Adissu, N. Miyawaki, M. Sasaki, K. Suzuki, and Y. Okamura. The voltage-gated proton channel Hv1/VSOP inhibits neutrophil granule release. *J. Leukoc. Biol.*, 99(1):7–19, 2016. _eprint: https://jlb.onlinelibrary.wiley.com/doi/pdf/10.1189/jlb.3HI0814-393R.

[220] F. Ooms. Molecular Modeling and Computer Aided Drug Design. Examples of their Applications in Medicinal Chemistry. *CMC*, 7(2):141–158, Feb. 2000.

[221] Open Force Field Initiative. "The Open Force Field 1.0 small molecule force field, our first optimized force field (codename 'Parsley')". `https://openforcefield.org/news/introducing-openforcefield-1.0/`.

[222] OpenEye Python Toolkits. OpenEye Scientific Software Inc.: Santa Fe, NM, USA.

[223] R. J. Ouellette and J. D. Rawn. 23 - Amines and Amides. In R. J. Ouellette and J. D. Rawn, editors, *Organic Chemistry Study Guide*, pages 465–494. Elsevier, Boston, Jan. 2015.

[224] B. Ouyang and B. J. Howard. The monohydrate and dihydrate of acetic acid : A high-resolution microwave spectroscopic study. *Phys. Chem. Chem. Phys.*, 11(2):366–373, 2009.

[225] R. Pal, M. B. M. Reddy, B. Dinesh, M. A. Venkatesha, S. Grabowsky, C. Jelsch, and T. N. Guru Row. Syn vs Anti Carboxylic Acids in Hybrid Peptides: Experimental and Theoretical Charge Density and Chemical Bonding Analysis. *J. Phys. Chem. A*, 122(14):3665–3679, Apr. 2018.

[226] A. C. Pan, D. W. Borhani, R. O. Dror, and D. E. Shaw. Molecular determinants of drug–receptor binding kinetics. *Drug Discovery Today*, 18(13):667–673, July 2013.

[227] A. C. Pan, H. Xu, T. Palpant, and D. E. Shaw. Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 13(7):3372–3377, July 2017.

[228] G. Parisio, M. Stocchero, and A. Ferrarini. Passive Membrane Permeability: Beyond the Standard Solubility-Diffusion Model. *J. Chem. Theory Comput.*, 9(12):5236–5246, Dec. 2013.

[229] J. L. Parker and S. Newstead. Membrane Protein Crystallisation: Current Trends and Future Perspectives. In I. Moraes, editor, *The Next Generation in Membrane Protein Structure Determination*, Advances in Experimental Medicine and Biology, pages 61–72. Springer International Publishing, Cham, 2016.

[230] R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.*, 13(7):3185–3197, July 2017.

[231] H. Pašalić, D. Tunega, A. J. A. Aquino, G. Haberhauer, M. H. Gerzabek, and H. Lischka. The stability of the acetic acid dimer in microhydrated environments and in aqueous solution. *Phys. Chem. Chem. Phys.*, 14(12):4162–4170, Feb. 2012.

[232] M. Patra and M. Karttunen. Systematic comparison of force fields for microscopic simulations of NaCl in aqueous solutions: Diffusion, free energy of hydration, and structural properties. *J. Comput. Chem.*, 25(5):678–689, 2004. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.10417.

[233] D. A. Pearlman. A Comparison of Alternative Approaches to Free Energy Calculations. *J. Phys. Chem.*, 98(5):1487–1493, Feb. 1994.

[234] S. Pérez, A. Imberty, S. B. Engelsen, J. Gruza, K. Mazeau, J. Jimenez-Barbero, A. Poveda, J.-F. Espinosa, B. P. van Eyck, G. Johnson, A. D. French, M. L. C. E. Kouwijzer, P. D. J. Grootenuis, A. Bernardi, L. Raimondi, H. Senderowitz, V. Durier,

G. Vergoten, and K. Rasmussen. A comparison and chemometric analysis of several molecular mechanics force fields and parameter sets applied to carbohydrates. *Carbohydrate Research*, 314(3):141–155, Dec. 1998.

[235] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16):1781–1802, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20289.

[236] V. V. Pisarev and S. A. Zakharov. Comparison of forcefields for molecular dynamics simulations of hydrocarbon phase diagrams. *J. Phys.: Conf. Ser.*, 946:012100, Jan. 2018.

[237] K. Pluhackova, S. A. Kirsch, J. Han, L. Sun, Z. Jiang, T. Unruh, and R. A. Böckmann. A Critical Comparison of Biomembrane Force Fields: Structure and Dynamics of Model DMPC, POPC, and POPE Bilayers. *J. Phys. Chem. B*, 120(16):3888–3903, Apr. 2016.

[238] A. Pohorille, C. Jarzynski, and C. Chipot. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B*, 114(32):10235–10253, Aug. 2010.

[239] J. W. Ponder and D. A. Case. Force Fields for Protein Simulations. In *Advances in Protein Chemistry*, volume 66 of *Protein Simulations*, pages 27–85. Academic Press, Jan. 2003.

[240] L. C. Puga Molina, G. M. Luque, P. A. Balestrini, C. I. Marín-Briggiler, A. Romarowski, and M. G. Buffone. Molecular Basis of Human Sperm Capacitation. *Front. Cell Dev. Biol.*, 6, 2018.

[241] F. Qiu, S. Rebolledo, C. Gonzalez, and H. P. Larsson. Subunit Interactions during Cooperative Opening of Voltage-Gated Proton Channels. *Neuron*, 77(2):288–298, Jan. 2013.

[242] I. S. Ramsey, Y. Mokrab, I. Carvacho, Z. A. Sands, M. S. P. Sansom, and D. E. Clapham. An aqueous H + permeation pathway in the voltage-gated proton channel Hv1. *Nat. Struct. Mol. Biol.*, 17(7):869–875, July 2010.

[243] I. S. Ramsey, M. M. Moran, J. A. Chong, and D. E. Clapham. A voltage-gated proton-selective channel lacking the pore domain. *Nature*, 440(7088):1213–1216, Apr. 2006.

[244] I. S. Ramsey, E. Ruchti, J. S. Kaczmarek, and D. E. Clapham. Hv1 proton channels are required for high-level NADPH oxidase-dependent superoxide production during the phagocyte respiratory burst. *PNAS*, 106(18):7642–7647, May 2009.

[245] A. L. Randolph, Y. Mokrab, A. L. Bennett, M. S. Sansom, and I. S. Ramsey. Proton currents constrain structural models of voltage sensor activation. *eLife*, 5:e18017, Aug. 2016.

[246] V. A. Rassolov, J. A. Pople, M. A. Ratner, and T. L. Windus. 6-31G* basis set for atoms K through Zn. *J. Chem. Phys.*, 109(4):1223–1229, July 1998.

[247] S. Y. Reddy, F. Leclerc, and M. Karplus. DNA Polymorphism: A Comparison of Force Fields for Nucleic Acids. *Biophysical Journal*, 84(3):1421–1449, Mar. 2003.

[248] A. Rennhack, E. Grahn, U. B. Kaupp, and T. K. Berger. Photocontrol of the Hv1 Proton Channel. *ACS Chem. Biol.*, 12(12):2952–2957, Dec. 2017.

[249] B. A. Reva, A. V. Finkelstein, and J. Skolnick. What is the probability of a chance prediction of a protein structure with an rmsd of 6 å? *Folding and Design*, 3(2):141–147, Apr. 1998.

[250] J. Řezáč, D. Bím, O. Gutten, and L. Rulíšek. Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set. *J. Chem. Theory Comput.*, 14(3):1254–1266, Mar. 2018.

[251] C. G. Ricci, A. S. C. de Andrade, M. Mottin, and P. A. Netz. Molecular Dynamics of DNA: Comparison of Force Fields and Terminal Nucleotide Definitions. *J. Phys. Chem. B*, 114(30):9882–9893, Aug. 2010.

[252] K. E. Riley and P. Hobza. The relative roles of electrostatics and dispersion in the stabilization of halogen bonds. *Phys. Chem. Chem. Phys.*, 15(41):17742–17751, 2013.

[253] S. Riniker. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.*, 58(3):565–578, Mar. 2018.

[254] T. Rivel, C. Ramseyer, and S. Yesylevskyy. The asymmetry of plasma membranes and their cholesterol content influence the uptake of cisplatin. *Sci. Rep.*, 9(1):1–14, Apr. 2019.

[255] B. C. Roberts and R. L. Mancera. Ligand-Protein Docking with Water Molecules. *J. Chem. Inf. Model.*, 48(2):397–408, Feb. 2008.

[256] P. Robustelli, S. Piana, and D. E. Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *PNAS*, 115(21):E4758–E4766, May 2018.

[257] G. J. Rocklin, D. L. Mobley, K. A. Dill, and P. H. Hünenberger. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *J. Chem. Phys.*, 139(18):184103, Nov. 2013.

[258] J. D. Rodriguez, S. Haq, T. Bachvaroff, K. F. Nowak, S. J. Nowak, D. Morgan, V. V. Cherny, M. M. Sapp, S. Bernstein, A. Bolt, T. E. DeCoursey, A. R. Place, and S. M. E. Smith. Identification of a vacuolar proton channel that triggers the bioluminescent flash in dinoflagellates. *PLOS ONE*, 12(2):e0171594, Feb. 2017.

[259] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, Mar. 1977.

[260] M. A. Sahai and P. C. Biggin. Quantifying Water-Mediated Protein–Ligand Interactions in a Glutamate Receptor: A DFT Study. *J. Phys. Chem. B*, 115(21):7085–7096, June 2011.

[261] K. Sargsyan, C. Grauffel, and C. Lim. How Molecular Size Impacts RMSD Applications in Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 13(4):1518–1524, Apr. 2017.

[262] M. Sasaki, M. Takagi, and Y. Okamura. A Voltage Sensor-Domain Protein Is a Voltage-Gated Proton Channel. *Science*, 312(5773):589, Apr. 2006.

[263] H. Sato and F. Hirata. The syn-/anti-conformational equilibrium of acetic acid in water studied by the RISM-SCF/MCSCF method1. *Journal of Molecular Structure: THEOCHEM*, 461–462:113–120, Apr. 1999.

[264] C. Scharnagl, R. Raupp-Kossmann, and S. F. Fischer. Molecular Basis for pH Sensitivity and Proton Transfer in Green Fluorescent Protein: Protonation and Conformational Substates from Electrostatic Calculations. *Biophysical Journal*, 77(4):1839–1857, Oct. 1999.

[265] T. Schilling, A. Gratopp, T. E. DeCoursey, and C. Eder. Voltage-activated proton currents in human lymphocytes. *J. Physiol.*, 545(1):93–105, 2002. _eprint: https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.2002.028878.

[266] E. V. Schow, J. A. Freites, P. Cheng, A. Bernsel, G. von Heijne, S. H. White, and D. J. Tobias. Arginine in Membranes: The Connection Between Molecular Dynamics Simulations and Translocon-Mediated Insertion Experiments. *J Membrane Biol*, 239(1):35–48, Jan. 2011.

[267] Schrödinger. The PyMOL Molecular Graphics System, Version 1.7.

[268] T. Schulz-Gasch, C. Schärfer, W. Guba, and M. Rarey. TFD: Torsion Fingerprints As a New Measure To Compare Small Molecule Conformations. *J. Chem. Inf. Model.*, 52(6):1499–1512, June 2012.

[269] B. D. Sellers, N. C. James, and A. Gobbi. A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.*, 57(6):1265–1275, June 2017.

[270] T. Seredenina, N. Demaurex, and K.-H. Krause. Voltage-Gated Proton Channels as Novel Drug Targets: From NADPH Oxidase Regulation to Sperm Biology. *Antioxidants & Redox Signaling*, 23(5):490–513, Feb. 2014.

[271] S.-Y. Sheu, D.-Y. Yang, H. L. Selzle, and E. W. Schlag. Energetics of hydrogen bonds in peptides. *PNAS*, 100(22):12683–12687, Oct. 2003.

[272] W. Shinoda. Permeability across lipid membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858(10):2254–2265, Oct. 2016.

[273] M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, Sept. 2008.

[274] M. R. Shirts and V. S. Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.*, 122(14):144107, Apr. 2005.

[275] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.*, 119(11):5740–5761, Aug. 2003.

[276] S. Sinnecker, A. Rajendran, A. Klamt, M. Diedenhofen, and F. Neese. Calculation of Solvent Shifts on Electronic g-Tensors with the Conductor-Like Screening Model (COSMO) and Its Self-Consistent Generalization to Real Solvents (Direct COSMO-RS). *J. Phys. Chem. A*, 110(6):2235–2245, Feb. 2006.

[277] M. S. Skaf. Molecular dynamics simulations of dielectric properties of dimethyl sulfoxide: Comparison between available potentials. *J. Chem. Phys.*, 107(19):7996–8003, Nov. 1997.

[278] A. A. Skelton, P. Fenter, J. D. Kubicki, D. J. Wesolowski, and P. T. Cummings. Simulations of the Quartz($10\bar{1}1$)/Water Interface: A Comparison of Classical Force Fields, Ab Initio Molecular Dynamics, and X-ray Reflectivity Experiments. *J. Phys. Chem. C*, 115(5):2076–2088, Feb. 2011.

[279] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe. Computational Methods in Drug Discovery. *Pharmacol Rev*, 66(1):334–395, Jan. 2014.

[280] D. Smith, D. Altarawy, L. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, and T. Crawford. The MolSSI QCArchive Project: An Open-Source Platform to Compute, Organize, and Share Quantum Chemistry Data. Mar. 2020.

[281] S. M. E. Smith, D. Morgan, B. Musset, V. V. Cherny, A. R. Place, J. W. Hastings, and T. E. DeCoursey. Voltage-gated proton channel in a dinoflagellate. *PNAS*, 108(44):18162–18167, Nov. 2011.

[282] C. M. Song, S. J. Lim, and J. C. Tong. Recent advances in computer-aided drug design. *Brief Bioinform*, 10(5):579–591, Sept. 2009.

[283] Y. Song, J. Mao, and M. R. Gunner. MCCE2: Improving protein pKa calculations with extensive side chain rotamer sampling. *J. Comput. Chem.*, 30(14):2231–2247, 2009.

[284] V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *The Journal of Chemical Physics*, 119(23):12129–12137, Dec. 2003.

[285] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.*, 98(45):11623–11627, Nov. 1994.

[286] C. A. Stortz, G. P. Johnson, A. D. French, and G. I. Csonka. Comparison of different force fields for the study of disaccharides. *Carbohydrate Research*, 344(16):2217–2228, Nov. 2009.

[287] R. Sun, Y. Han, J. M. J. Swanson, J. S. Tan, J. P. Rose, and G. A. Voth. Molecular transport through membranes: Accurate permeability coefficients from multidimensional potentials of mean force and local diffusion constants. *J. Chem. Phys.*, 149(7):072310, May 2018.

[288] X.-h. Sun, Y.-y. Zhu, L. Wang, H.-l. Liu, Y. Ling, Z.-l. Li, and L.-b. Sun. The Catsper channel and its roles in male fertility: A systematic review. *Reproductive Biology and Endocrinology*, 15(1):65, Aug. 2017.

[289] K. Takeshita, S. Sakata, E. Yamashita, Y. Fujiwara, A. Kawanabe, T. Kurokawa, Y. Okochi, M. Matsuda, H. Narita, Y. Okamura, and A. Nakagawa. X-ray crystal structure of voltage-gated proton channel. *Nat. Struct. Mol. Biol.*, 21(4):352–357, Apr. 2014.

[290] D. Tang, Y. Yang, Z. Xiao, J. Xu, Q. Yang, H. Dai, S. Liang, C. Tang, H. Dong, and Z. Liu. Scorpion toxin inhibits the voltage-gated proton channel using a Zn2+-like long-range conformational coupling mechanism. *Br. J. Pharmacol.*, 177(10):2351–2364, 2020. _eprint: https://bpspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/bph.14984.

[291] Z. Tang, S.-H. Chen, and C.-e. A. Chang. Transient States and Barriers from Molecular Simulations and the Milestoning Theory: Kinetics in Ligand–Protein Recognition and Compound Design. *J. Chem. Theory Comput.*, 16(3):1882–1895, Mar. 2020.

[292] R. L. Thurlkill, G. R. Grimsley, J. M. Scholtz, and C. N. Pace. pK values of the ionizable groups of proteins. *Protein Sci.*, 15(5):1214–1218, 2006.

[293] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *PNAS*, 112(5):E386–E391, Feb. 2015.

[294] D. S. Tomar, V. Weber, B. M. Pettitt, and D. Asthagiri. Importance of Hydrophilic Hydration and Intramolecular Interactions in the Thermodynamics of Helix–Coil Transition and Helix–Helix Assembly in a Deca-Alanine Peptide. *J. Phys. Chem. B*, 120(1):69–76, Jan. 2016.

[295] F. Tombola, M. H. Ulbrich, and E. Y. Isacoff. The Voltage-Gated Proton Channel Hv1 Has Two Pores, Each Controlled by One Voltage Sensor. *Neuron*, 58(4):546–556, May 2008.

[296] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, Feb. 1977.

[297] O. Trott and A. J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2):455–461, 2010.

[298] C. H. Tse, J. Comer, Y. Wang, and C. Chipot. Link between Membrane Composition and Permeability to Drugs. *J. Chem. Theory Comput.*, 14(6):2895–2909, June 2018.

[299] M. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(3):1990–2001, Aug. 1992.

[300] TURBOMOLE V7.1 2016, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from `http://www.turbomole.com`.

[301] J. M. Turney, A. C. Simmonett, R. M. Parrish, E. G. Hohenstein, F. A. Evangelista, J. T. Fermann, B. J. Mintz, L. A. Burns, J. J. Wilke, M. L. Abrams, N. J. Russ, M. L. Leininger, C. L. Janssen, E. T. Seidl, W. D. Allen, H. F. Schaefer, R. A. King, E. F. Valeev, C. D. Sherrill, and T. D. Crawford. Psi4: An open-source ab initio electronic structure program. *WIREs Comput Mol Sci*, 2(4):556–565, July 2012.

[302] H. van de Waterbeemd, H. Lennernas, and P. Artursson. *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability*. WILEY-VCH Verlag GmbH, 2009. Accepted: 2019-12-04T06:48:48Z.

[303] H. W. T. van Vlijmen, M. Schaefer, and M. Karplus. Improving the accuracy of protein pKa calculations: Conformational averaging versus the average structure. *Proteins Struct. Funct. Bioinforma.*, 33(2):145–158, 1998.

[304] J. R. Vane and R. M. Botting. The mechanism of action of aspirin. *Thrombosis Research*, 110(5):255–258, June 2003.

[305] K. Vanommeslaeghe and A. D. MacKerell. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.*, 52(12):3144–3154, Dec. 2012.

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

[307] D. Vassetti, M. Pagliai, and P. Procacci. Assessment of GAFF2 and OPLS-AA General Force Fields in Combination with the Water Models TIP3P, SPCE, and OPC3 for the Solvation Free Energy of Druglike Organic Molecules. *J. Chem. Theory Comput.*, 15(3):1983–1995, Mar. 2019.

[308] R. M. Venable, A. Krämer, and R. W. Pastor. Molecular Dynamics Simulations of Membrane Permeability. *Chem. Rev.*, 119(9):5954–5997, May 2019.

[309] C. Ventura, I. E. Leon, A. Asuaje, P. Martín, N. Enrique, M. Núñez, C. Cocca, and V. Milesi. Differential expression of the long and truncated Hv1 isoforms in breast-cancer cells. *J. Cell. Physiol.*, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcp.29719.

[310] V. Vivcharuk and Y. N. Kaznessis. Thermodynamic Analysis of Protegrin-1 Insertion and Permeation through a Lipid Bilayer. *J. Phys. Chem. B*, 115(49):14704–14712, Dec. 2011.

[311] I. Vorobyov, W. D. Bennett, D. P. Tieleman, T. W. Allen, and S. Noskov. The Role of Atomic Polarization in the Thermodynamics of Chloroform Partitioning to Lipid Bilayers. *J. Chem. Theory Comput.*, 8(2):618–628, Feb. 2012.

[312] I. Vorobyov, T. E. Olson, J. H. Kim, R. E. Koeppe, O. S. Andersen, and T. W. Allen. Ion-Induced Defect Permeation of Lipid Membranes. *Biophysical Journal*, 106(3):586–597, Feb. 2014.

[313] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis. *Can. J. Phys.*, 58(8):1200–1211, Aug. 1980.

[314] L. W. Votapka, C. T. Lee, and R. E. Amaro. Two Relations to Estimate Membrane Permeability Using Milestoning. *J. Phys. Chem. B*, 120(33):8606–8616, Aug. 2016.

[315] R. C. Wade and P. J. Goodford. The role of hydrogen-bonds in drug binding. *Prog. Clin. Biol. Res.*, 289:433–444, 1989.

[316] J. R. Wagner. openforcefield/openforcefields: Version 1.0.0 "Parsley", Oct. 2019.

[317] H. G. Wallnoefer, T. Fox, K. R. Liedl, and C. S. Tautermann. Dispersion dominated halogen–π interactions: Energies and locations of minima. *Phys. Chem. Chem. Phys.*, 12(45):14941–14949, 2010.

[318] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, Oct. 2006.

[319] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, July 2004.

[320] L. Wang, R. A. Friesner, and B. J. Berne. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B*, 115(30):9431–9438, Aug. 2011.

[321] R. Wang, S. Bi, V. Presser, and G. Feng. Systematic comparison of force fields for molecular dynamic simulation of Au(111)/Ionic liquid interfaces. *Fluid Phase Equilibria*, 463:106–113, May 2018.

[322] Y. Wang, S. J. Li, J. Pan, Y. Che, J. Yin, and Q. Zhao. Specific expression of the human voltage-gated proton channel Hv1 in highly metastatic breast cancer cells, promotes tumor progression and metastasis. *Biochemical and Biophysical Research Communications*, 412(2):353–359, Aug. 2011.

[323] Y. Wang, S. J. Li, X. Wu, Y. Che, and Q. Li. Clinicopathological and Biological Significance of Human Voltage-gated Proton Channel Hv1 Protein Overexpression in Breast Cancer. *J. Biol. Chem.*, 287(17):13877–13888, Apr. 2012.

[324] J. Warwicker. Simplified methods for pKa and acid pH-dependent stability estimation in proteins: Removing dielectric and counterion boundaries. *Protein Sci.*, 8(2):418–425, Feb. 1999.

[325] F. Weigend and R. Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, Aug. 2005.

[326] F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs. RI-MP2: Optimized auxiliary basis sets and demonstration of efficiency. *Chemical Physics Letters*, 294(1):143–152, Sept. 1998.

[327] A. M. A. West, R. Elber, and D. Shalloway. Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide. *J. Chem. Phys.*, 126(14):145104, Apr. 2007.

[328] B. R. White, C. R. Wagner, D. G. Truhlar, and E. A. Amin. Molecular Modeling of Geometries, Charge Distributions, and Binding Energies of Small, Druglike Molecules Containing Nitrogen Heterocycles and Exocyclic Amino Groups in the Gas Phase and in Aqueous Solution. *J. Chem. Theory Comput.*, 4(10):1718–1732, Oct. 2008.

[329] K. B. Wiberg and K. E. Laidig. Barriers to rotation adjacent to double bonds. 3. The carbon-oxygen barrier in formic acid, methyl formate, acetic acid, and methyl acetate. The origin of ester and amide resonance. *J. Am. Chem. Soc.*, 109(20):5935–5943, Sept. 1987.

[330] S. E. Wong and F. C. Lightstone. Accounting for water molecules in drug design. *Expert Opinion on Drug Discovery*, 6(1):65–74, Jan. 2011.

[331] M. L. Wood. *Investigating the Structure and Function of Ion Conducting Voltage Sensing Domains*. PhD thesis, University of California, Irvine, 2016. ProQuest ID: Wood_uci_0030D_13869. Merritt ID: ark:/13030/m5324hpp. Retrieved from `https://escholarship.org/uc/item/9j86t5s7`.

[332] M. L. Wood, E. V. Schow, J. A. Freites, S. H. White, F. Tombola, and D. J. Tobias. Water wires in atomistic models of the Hv1 proton channel. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1818(2):286–293, Feb. 2012.

[333] D. Wu and D. A. Kofke. Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation. *J. Chem. Phys.*, 123(5):054103, Aug. 2005.

[334] L.-J. Wu, G. Wu, M. R. A. Sharif, A. Baker, Y. Jia, F. H. Fahey, H. R. Luo, E. P. Feener, and D. E. Clapham. The voltage-gated proton channel Hv1 enhances brain damage from ischemic stroke. *Nat. Neurosci.*, 15(4):565–573, Apr. 2012.

[335] A.-S. Yang, M. R. Gunner, R. Sampogna, K. Sharp, and B. Honig. On the calculation of pKas in proteins. *Proteins Struct. Funct. Bioinforma.*, 15(3):252–265, 1993.

[336] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao. Enhanced sampling in molecular dynamics. *J. Chem. Phys.*, 151(7):070902, Aug. 2019.

[337] M. Zacharias, T. P. Straatsma, and J. A. McCammon. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.*, 100(12):9025–9031, June 1994.

[338] C. Zhao, L. Hong, J. D. Galpin, C. A. Ahern, and F. Tombola. Exploiting Π-Stacking Interactions to Improve Inhibition of the Hv1 Channel by Aromatic Guanidine Derivatives. *Biophysical Journal*, 114(3):374a, Feb. 2018.

[339] R. Zhao, K. Kennedy, G. A. D. Blas, G. Orta, M. A. Pavarotti, R. J. Arias, J. L. de la Vega-Beltrán, Q. Li, H. Dai, E. Perozo, L. S. Mayorga, A. Darszon, and S. A. N. Goldstein. Role of human Hv1 channels in sperm capacitation and white blood cell respiratory burst established by a designed peptide inhibitor. *PNAS*, 115(50):E11847–E11856, Dec. 2018.

[340] S. Zhu. Validation of the Generalized Force Fields GAFF, CGenFF, OPLS-AA, and PRODRGFF by Testing Against Experimental Osmotic Coefficient Data for Small Drug-Like Molecules. *J. Chem. Inf. Model.*, 59(10):4239–4247, Oct. 2019.

[341] X. Zhu, E. Mose, and N. Zimmermann. Proton channel HVCN1 is required for effector functions of mouse eosinophils. *BMC Immunology*, 14(1):24, May 2013.

[342] R. W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22(8):1420–1426, Aug. 1954.

[343] M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe, D. M. Zuckerman, and L. T. Chong. WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis. *J. Chem. Theory Comput.*, 11(2):800–809, Feb. 2015.

Appendix A

Supporting Information: Insights on small molecule binding to the Hv1 proton channel from free energy calculations with molecular dynamics simulations

A.1 Ligand parameterization for 2GBI

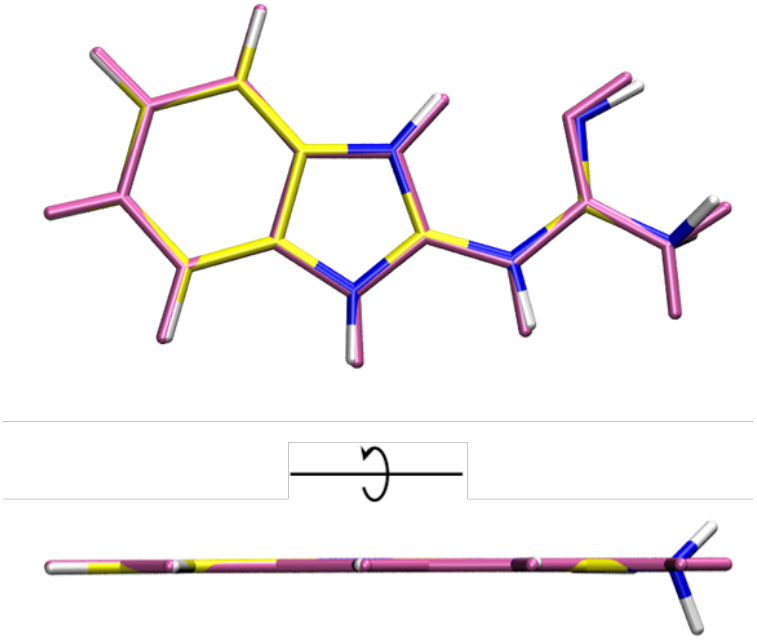


Figure A.1: A molecule overlay of 2GBI (gbi2) compares the geometry-optimized structures from quantum mechanics (MP2/6-31G*) (colored in mauve) and the unmodified CGenFF force field (colored in blue and yellow). Poor assignment of initial atom types in the force field led to a pyramidal geometry of the -NH_2 group in the guanidine region.

The CHARMM force field parameters for gbi1 were developed using CGenFF.[94, 331, 305, 306] The force field for gbi2 was generated from the CGenFF webserver, `https://cgenff.umaryland.edu/`, which performs atom typing and parameter/charge assignment by analogy. The resulting force field was modified with comparison to a reference structure from an *ab initio* QM MP2/6-31G*[120, 326, 87, 88, 246] geometry optimization. We compared the geometry-optimized structures from QM and the force field, then we conduct torsion scans around the dihedral angle labeled by the four highlighted atoms in Figure 2.1.

The original initial force field had a poor choice of assigned atom types for the nitrogen and hydrogen atoms of the terminal -NH_2 group in the guanidine region, leading to a minimized structure that is non-planar as shown in blue and yellow in Figure A.1. The mauve structure shows the QM-optimized structure for reference. Changes were made to the CGenFF force

field to redefine those two atom types going from a methylamine-like moiety to an imine-like moiety. We also lightly adjusted standard and dihedral angle parameters to improve the consistency of the MM geometry to the QM geometry.

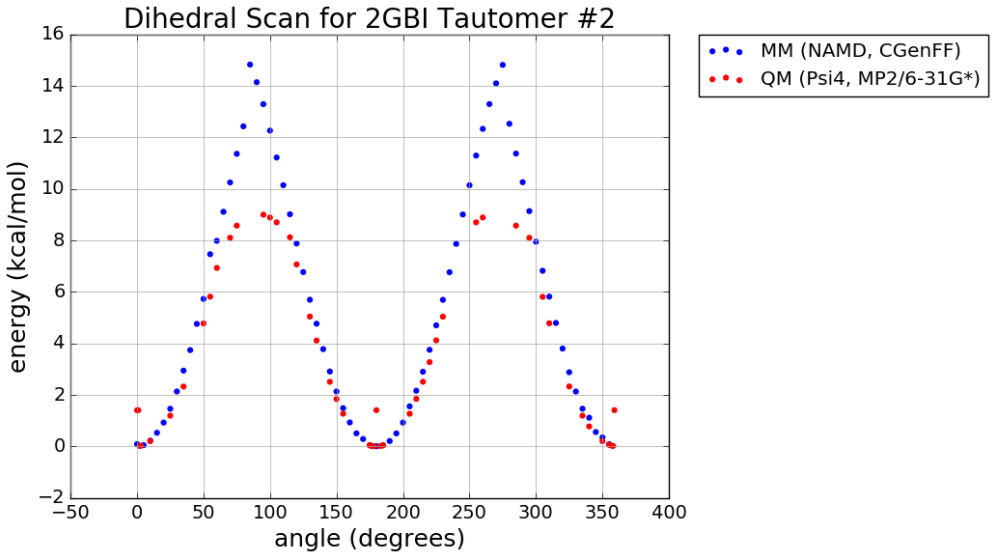


Figure A.2: Overlay of dihedral scan energy profiles from the CGenFF force field (blue) and reference quantum mechanical data (red) for gbi2.

Secondly, we computed energy profiles for torsion scans of the dihedral angle labeled in Figure 2.1. The QM result is plotted in red in Figure A.2, and the MM result is shown as the blue markers. There is reasonable agreement between QM and the force field except for the regions around the minima (0°, 180°) and maxima (90°, 270°). The deviation at the maxima is expected to have negligible influence on our protein-bound simulations as the torsion scans are similar up to 8 kcal/mol, and the likelihood for conformational states to be occupied decreases dramatically at very high energy levels. Further investigation at the 180° region (Figure A.3) showed that the small plateau of QM energy (blue curve) reflects a steric interaction between two atoms of the guanidine region (red curve) which is not captured in classical MD simulations. For that reason, we did not attempt to further modify the dihedral parameters.

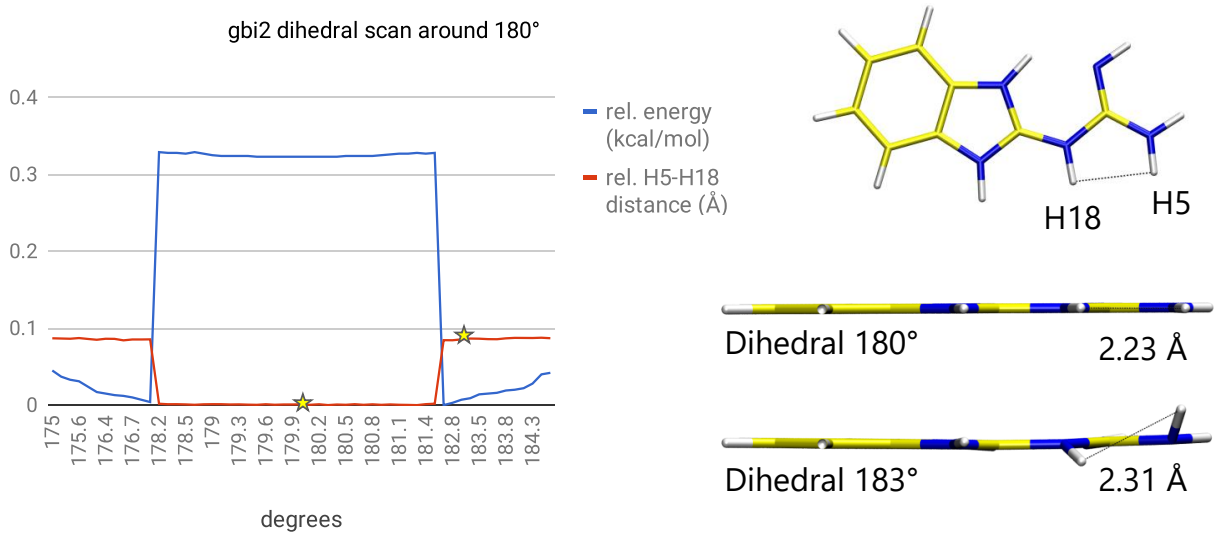


Figure A.3: Focused dihedral scan around the 180° region of gbi2. We applied the QM method B3LYP-D3MBJ/def2-TZVP[22, 170, 313, 285, 325] for a more rigorous examination compared to the traditional MP2/6-31G* reference. The increase in relative dihedral energy is directly correlated to the H18-H5 distance, suggesting the influence of steric interactions.

A.2 Fixed dihedral angles in docking

The placement and orientation of 2GBI bound to Hv1 was explored using molecular docking with AutoDock Vina.[297] We impose planarity of 2GBI, rationalized by a planar optimized geometry and a higher energy barrier to rotate the dihedral angle (Figure 2.1) away from planar 180° (Figure A.2 and Figure A.4).

When we did not enforce a planar ligand during docking, this led to a variety of geometries for possible configurations of 2GBI in Hv1, including that shown in Figure A.5. Docking tends to ignore internal strain energies in molecules so it is not necessarily unusual for it to generate highly strained candidate binding modes. Depending on the timescale for changes in binding mode, these can end up being retained throughout even lengthy MD simulations if the ligand is unable (often for steric reasons) to find something better.

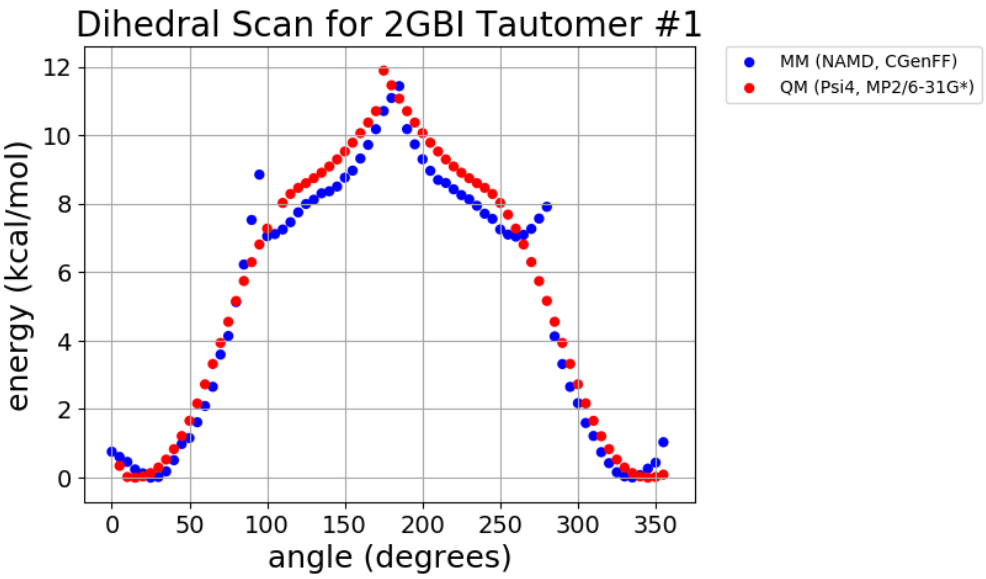


Figure A.4: Overlay of dihedral scan energy profiles from the CGenFF force field (blue) and reference quantum mechanical data (red) for gbi1.

A.3 Pose selection and refinement with MD

A "reverse clustering" approach from fixed-planar docking results gave us distinct potential binding poses in the binding site. Each pose was assessed from putative protein-ligand interactions from Hong et al.[128] We also evaluated the fit of the ligand in the Hv1 binding region from MD simulations. Regarding protein-ligand interactions, we monitored the following distances: (1) between the center of the benzo moieties on 2GBI and F150, (2) between the central carbon atoms of the guanidine moieties on 2GBI and R211, and (3) between the closest imidazolic nitrogen atom on 2GBI with the closest oxygen atom on D112. The optimal pose showed relatively stable contact distances in each of these criteria as well as a stable protein backbone RMSD and stable ligand RMSD.

After docking, we assessed hydration of the ligand-bound protein pore in 20 ns of equilibrating MD simulations to verify that the pore was well-hydrated prior to free energy calculations (Figure A.6). In Figure A.7, we show our binding pose as in Figure 2.4 with additional protein residues labeled to compare to that of Chamberlin et al.[38]

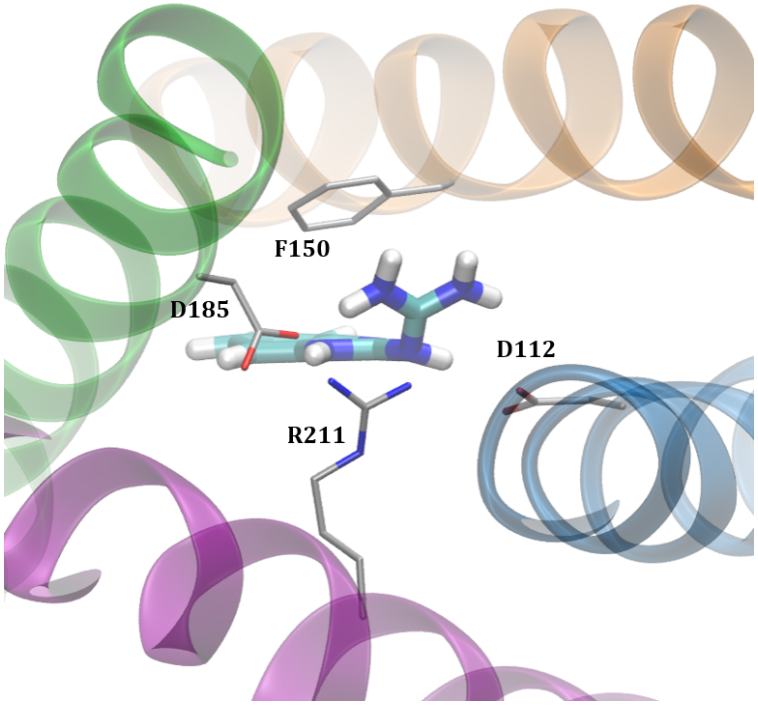


Figure A.5: Illustrative example from docking and molecular dynamics for gbi1 bound to Hv1, viewing from extracellular end of channel. 2GBI is trapped in a strained configuration with its dihedral angle at 112° (cf. energy profile in Figure A.4).

dimension	lower bound	upper bound
x	A175, backbone N	A135, backbone C
y	L124, backbone O	E153, backbone O
z	H168, backbone CA	I127, backbone N

Table A.1: Atom selections used to define coordinates of the bounding box to count pore waters in MD simulations following docking calculations.

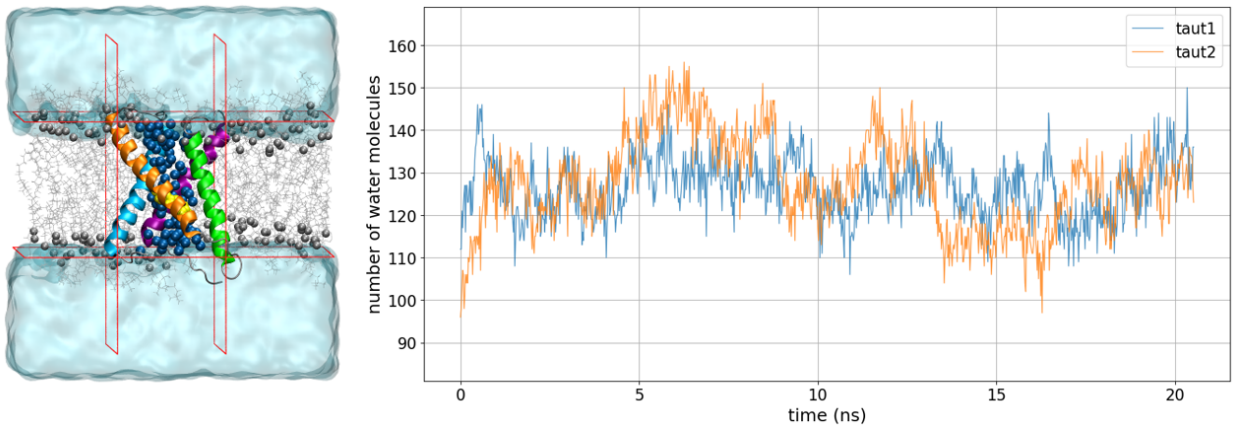


Figure A.6: Count of water molecules within the pore of the 2GBI-bound Hv1 open state in equilibration simulations prior to free energy calculations. We counted waters within the rectangular prism illustrated on the left panel, where the box dimensions were based on the atom positions (following trajectory alignment on the Hv1 backbone) as listed in Table A.1.

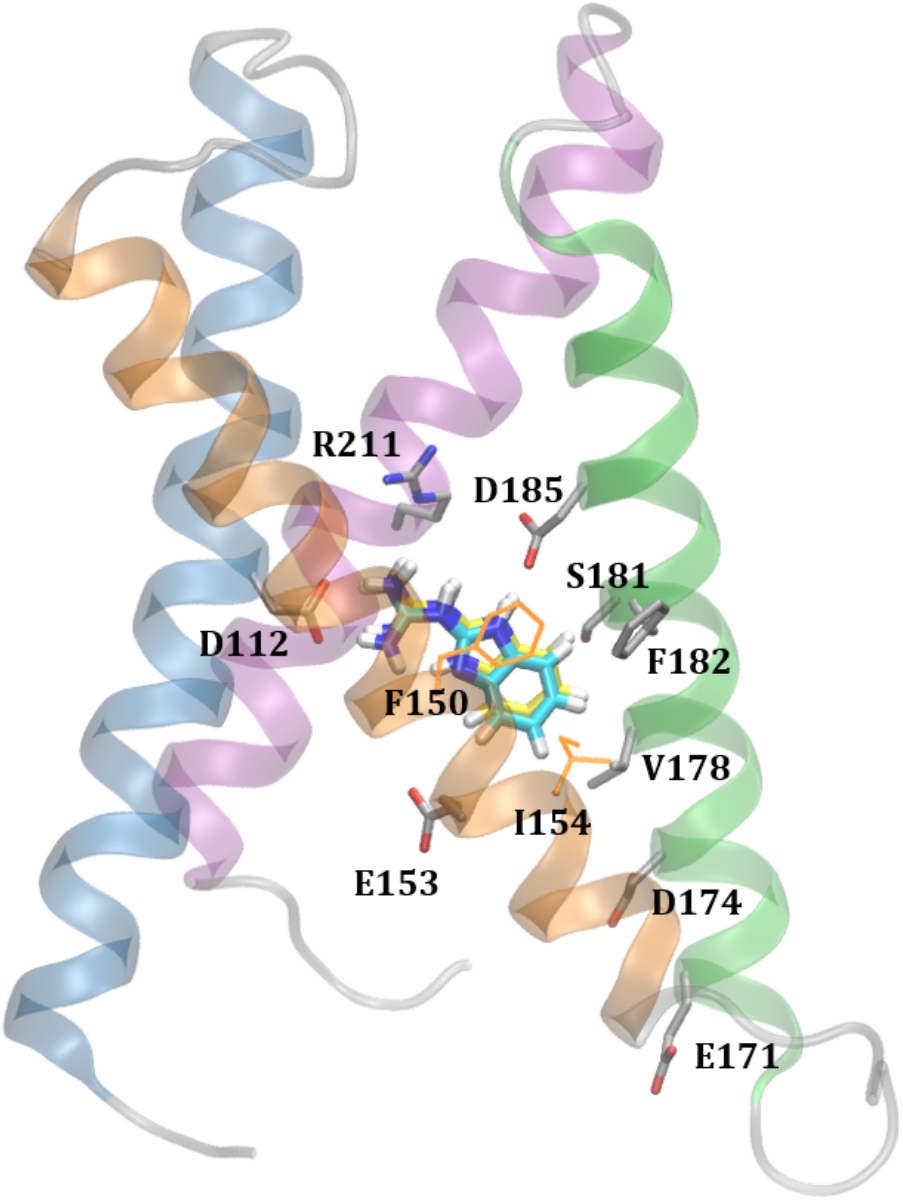


Figure A.7: Our binding pose as in Figure 2.4 with additional protein residues labeled. The extracellular side of the channel is at the top of the figure. Select residues, namely E153, D174, and E171, comprise part of the binding pocket of 2GBI in *Ciona intestinalis* Hv1 as reported by Chamberlin et al. Interactions with these residues would lead to an intracellular-pointing ligand orientation which would preclude the protein-ligand interactions with S181 or R211 as proposed by Hong et al.

A.4 Analysis of protein-ligand interactions of bound ligand before and after D112E mutation

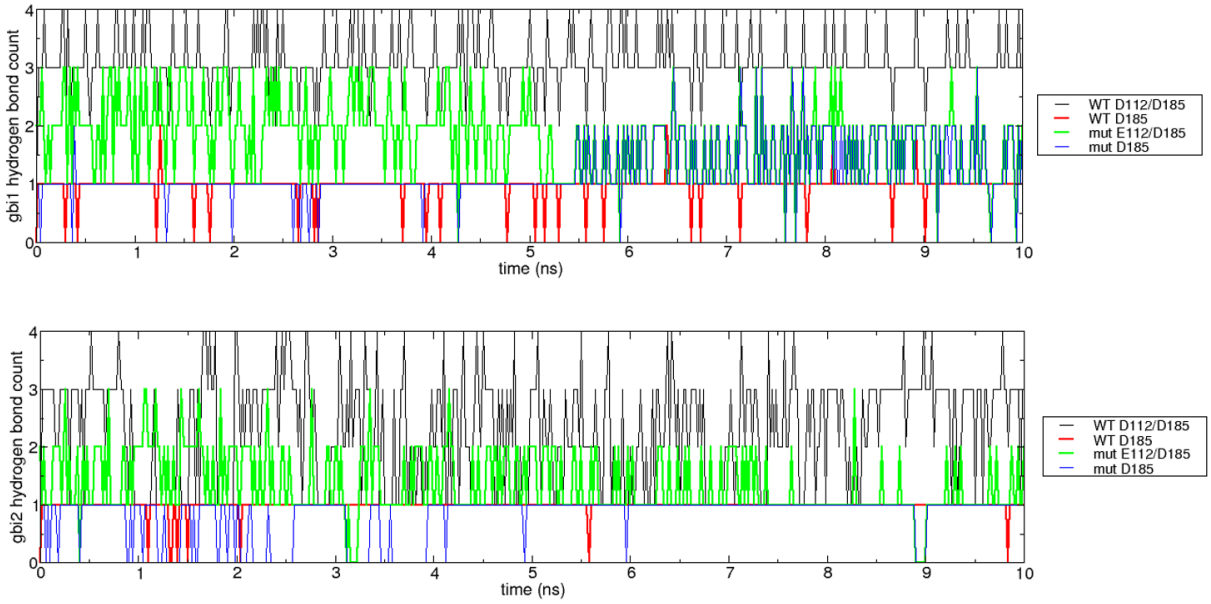


Figure A.8: Hydrogen bond count between 2GBI and Hv1 residues 112 and 185 for 10 ns of simulation with WT Hv1 and with the D112E mutant of Hv1 for (**top**) gbi1 and (**bottom**) gbi2. The black and green lines represent the total hydrogen bond count between the ligand and D112/D185 before and after mutation, respectively. The red and blue lines represent the number of hydrogen bonds between 2GBI and only D185 before and after mutation, respectively. The hydrogen bond counts to both acidic residues are generally higher before D112E mutation (black vs. green) for both tautomers. The hydrogen bonds to D185 are slightly disrupted after mutation, especially for gbi2 (bottom; red vs. blue). There is an increase in hydrogen bonding interactions between gbi1 and D185 after around 5 ns as the ligand reorients as described in the main text (Figure2.6).

A.5 Flat-bottom restraints for mutations involving Arg or Lys residues

The charge neutralization of flexible side chains during alchemical free energy calculation may lead to the loss of native electrostatic interactions with proximal side chains. This may

result in artifactual configurations of the transforming side chain that adversely affects the calculated binding free energies (Figure 2.7). To address this issue, we used the `colvars` module of NAMD[83] to add flat-bottom restraints to arginine and lysine residues partaking in mutation. This imposes an energetic penalty when the variable (here, a distance) is below some threshold x_0 or above some threshold x_1. In between the thresholds, the potential is zero and no additional force is exerted. The distances defining the flat-bottom restraint parameters are shown in the following table:

mutation	atom of mutating side chain	atom of reference side chain	x_0: lower boundary (Å)	x_1: upper boundary (Å)
R208K	R208 CZ	D123 CG	3.0	8.5
R208K	R208 CZ	E119 CD	3.0	8.5
R208K	R208 CZ	E192 CD	3.0	8.5
R208K	K208 NZ	D123 CG	3.0	8.5
R208K	K208 NZ	E119 CD	3.0	8.5
R208K	K208 NZ	E192 CD	3.0	8.5
R211S	R211 CZ	D112 CG	3.5	7.0
R211S	R211 CZ	D185 CG	3.5	7.0
R211S	R211 CZ	F150 CG	6.0	10.5

Table A.2: Atom labels and distance parameters used in flat-bottom restraints imposed in mutating Arg or Lys residues. Shaded rows represent restraints that were applied to the holo protein only.

The distance boundaries were chosen to be restrictive enough to prevent excessive side chain mobility, but also flexible enough to allow conformational sampling.

A correction must be applied to the computed free energies to unbias the effects of the restraints. The thermodynamic cycle for correcting the relative binding free energies is depicted in Figure A.9.

The free energy computations require eight additional simulations to estimate ΔG_7, $-\Delta G_6$, ΔG_9, and $-\Delta G_8$. Each of these terms is determined by calculating the distances between each restrained pair of side chains (the forward work), calculating the distances between each non-restrained pair of side chains (the reverse work), computing the associated energies

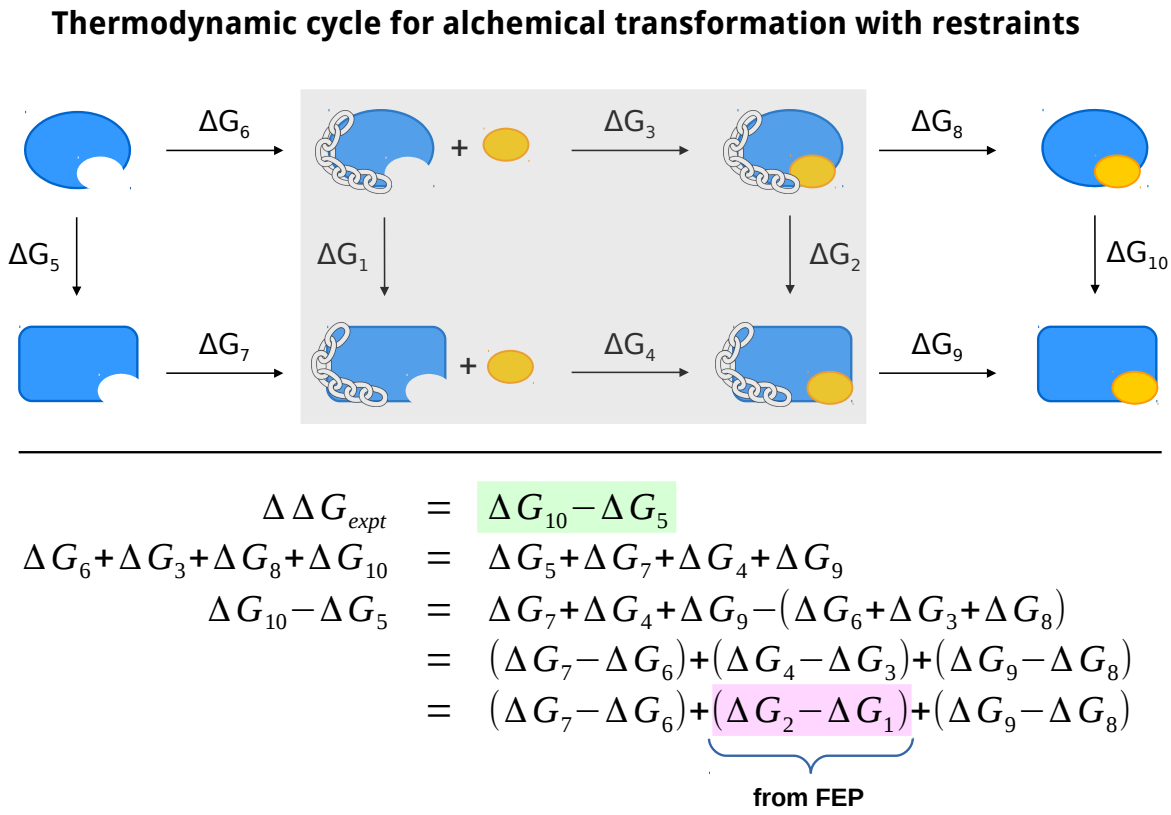


Figure A.9: Thermodynamic cycle for computing relative binding free energies in an alchemical transformation, modified to account for the addition of side chain restraints. The small orange oval represents the ligand, the large blue oval represents the wild type protein, and the blue rectangle represents the mutated protein. The shaded region represents the standard thermodynamic cycle, such as that shown in Figure 2.2, with the incorporation of flat-bottom restraints on the mutating side chain. The processes represented by each component are as follows: ΔG_1, alchemical free energy of mutation in apo protein (restrained); ΔG_2, alchemical free energy of mutation in holo protein (restrained); ΔG_3, artificial binding free energy of wild type protein (restrained); ΔG_4, artificial binding free energy of mutant protein (restrained); ΔG_5, alchemical free energy of mutation in apo protein (unrestrained); ΔG_6, free energy to add restraints to apo wild type protein; ΔG_7, free energy to add restraints to apo mutant protein; ΔG_8, free energy to remove restraints from holo wild type protein; ΔG_9, free energy to remove restraints from holo mutant protein; ΔG_{10}, alchemical free energy of mutation in holo protein (unrestrained). The effects of the side chain restraints can be removed from the computed relative binding free energy ($\Delta G_2 - \Delta G_1$) by the four additional terms of ΔG_7, $-\Delta G_6$, ΔG_9, and $-\Delta G_8$.

from the flat-bottom potential, subsampling based on statistical inefficiency, then calculating the free energy difference from the BAR estimator.[23, 273] To calculate the free energy to

remove the restraints for correcting the relative binding free energies, the forward work represents the energy to remove the restraints from a restrained simulation, and the reverse work represents the energy to introduce restraints into a non-restrained simulation. For well-chosen restraint distances such that the side chain distances fall within the limits of the restraints even when no restraints are applied, the ΔG correction term goes to zero. As such, we defined our restraint distances so that the additional ΔG terms of the correction were zero.

A.6 Potential of mean force calculation for F182 rotation after V178A mutation

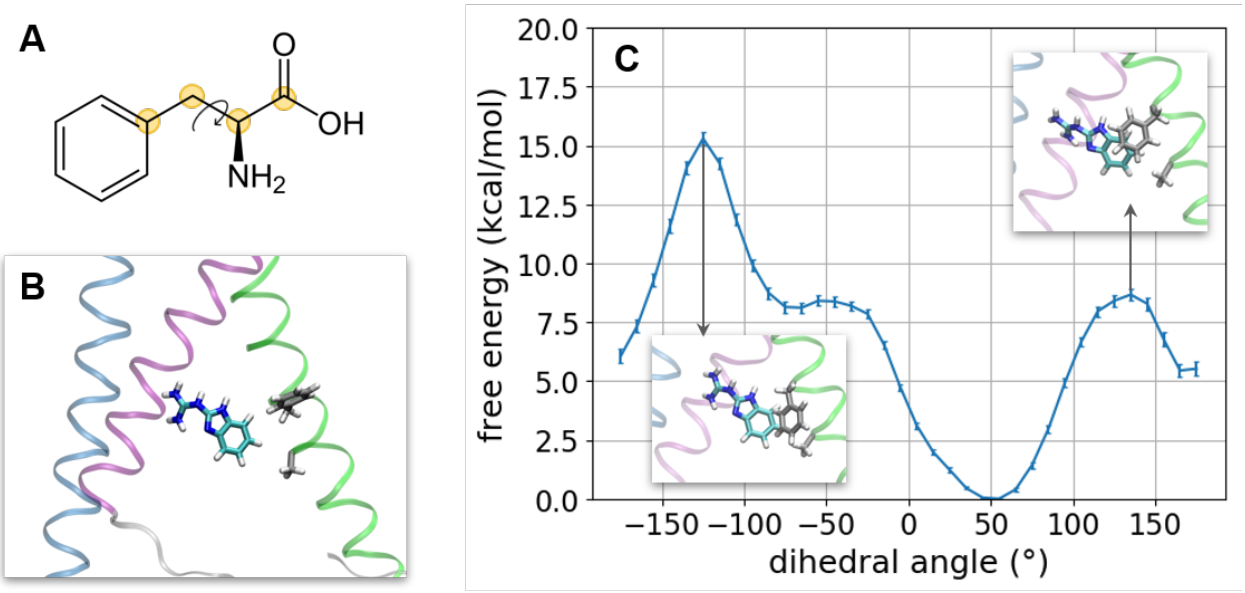


Figure A.10: Consideration of potential ligand stabilization by reorientation of F182 after V178A mutation. (**A**) The highlighted four atoms define the dihedral angle of interest around which we consider rotation of the F182 side chain. (**B**) Representative snapshot from equilibrium MD simulations with F182 dihedral angle at 46°. S2 is not shown for clarity, and colors are consistent with those in Figure 2.3. This shows approximately the most energetically favorable configuration of this phenylalanine residue. (**C**) Potential of mean force for rotation of the F182 dihedral angle. Two insets show the position of F182 at -130° and 140°.

The V178A mutation is expected to contribute to a slightly more favorable 2GBI binding free energy relative to wild type Hv1 according to mutagenesis experiments. Our computed values suggested a slightly more unfavorable binding free energy. We examined the effect of secondary conformational changes in the protein that may result after the V178A mutation. Specifically, we focused on F182, which is in proximity to both V178 and 2GBI (we consider gbi1 only). We used umbrella sampling[296] to compute the potential of mean force (PMF) for rotation of the side chain, which could lead to better interactions with the bound ligand.

The dihedral angle, defined by the four atoms shown in Figure A.10A, was restrained in increments of 10° from -180° to 180° using a harmonic potential with a force constant of 0.1 kcal/mol/deg^2. Each configuration was minimized for 5000 steps with the conjugate gradient algorithm, then run for 5 ns with the MD simulation settings as described in the main text. We used the MBAR algorithm[273] to generate the PMF and associated error bars. It does not appear that reorientation of F182 as considered here contributes to the favorable relative binding free energy of V178A.

Appendix B

Supporting Information: Potential of mean force calculations for the membrane permeation of Hv1 channel blockers

B.1 pK_a shift analysis

To estimate whether these solutes remain protonated as they permeate the bilayer, we computed the pK_a shift profile across the membrane for each molecule. This involved additionally calculating PMFs for the neutral solutes. Structures of the positively charged molecules and their neutral analogs are shown in Figure B.1.

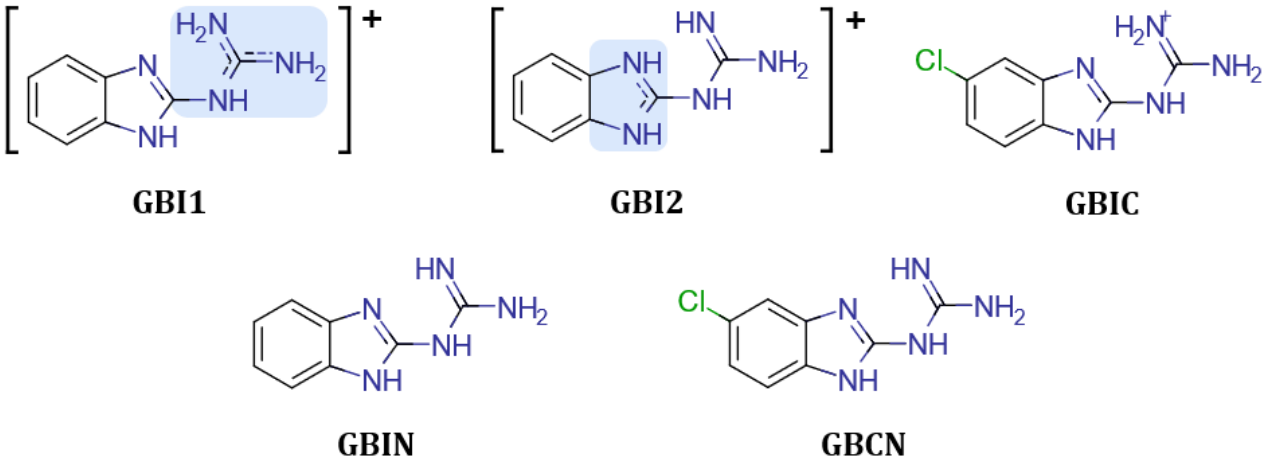


Figure B.1: GBI1 and GBI2 refer to distinct tautomeric states of the positively charged molecule, 2-guanidinobenzimidazole. The shaded blue regions have higher positive charge density. GBIN is the neutral analog to 2GBI. GBIC refers to 5-chloro-2-guanidinobenzimidazole, and GBCN denotes its neutral counterpart.

We applied the thermodynamic cycle[175] in Figure B.2. The orange ovals represent the positively charged form of the solute, and the blue ovals represent the neutral form. The deprotonation free energy is computed from the difference in potentials of mean force from the neutral solute minus the positively charged solute using the expression:

$$\Delta\Delta G_{\text{deprot}} = W_{\text{neutral}}(z) - W_{\text{charged}}(z)\,. \tag{B.1}$$

The neutral solute PMF was computed using the same MD simulation settings and protocol as described in the main text. Given the relationship of free energy ΔG to the equilibrium

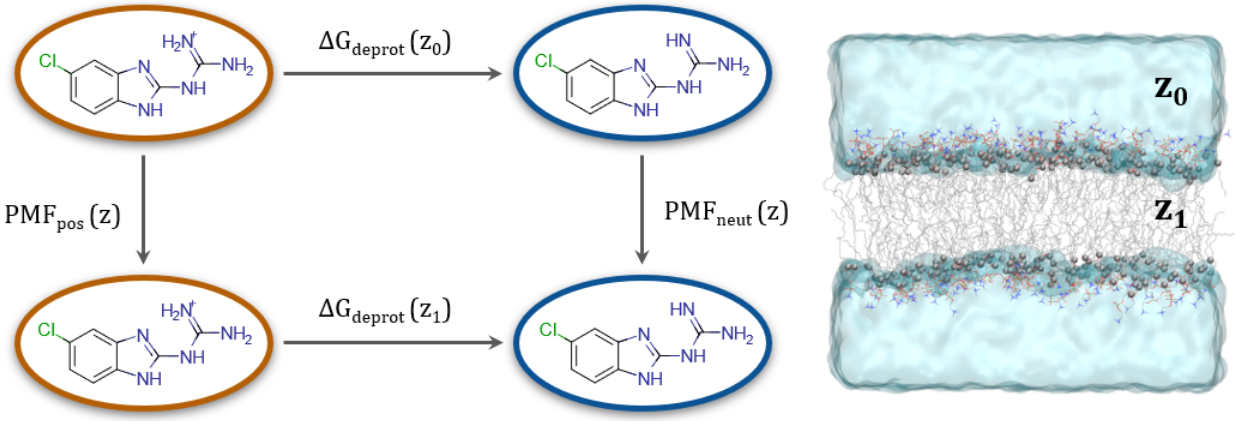


Figure B.2: Thermodynamic cycle applied to calculate the deprotonation free energy for a solute to permeate the membrane. The deprotonation free energy is used to calculate the pK_a shift profile.

constant K_a, the final pK_a shift profile takes the expression:

$$\Delta pK_a = \frac{\Delta\Delta G_{\text{deprot}}}{2.3 k_B T} \,. \tag{B.2}$$

The PMFs for the neutral molecules are plotted alongside the positively charged solute PMFs in Figure B.3. The primary distinction between the PMFs of the neutral and charged forms is that there is considerably less energetic stabilization as the solute is in the lipid headgroup region ($|z|$ = 15-20 Å) for the neutral compared to the cationic solutes. This may be attributed to the fact that fewer electrostatic interactions are present between the imidazole or guanidine moieties and the negatively charged phosphate groups of the POPC molecules. For GBCN, the neutral variant of GBIC, the error bars are considerably larger and the PMF appears less smooth close to the membrane center. In our simulations, we observe inconsistent solute flip-flop events (vide infra) while sampling the membrane interior, suggesting that these results are not sufficiently converged even after 1.48 microseconds of total simulation time for this solute. We do not extend our pK_a shift analysis for this case.

From these PMFs, the pK_a shift profile was calculated for 2GBI and shown in Figure B.4.

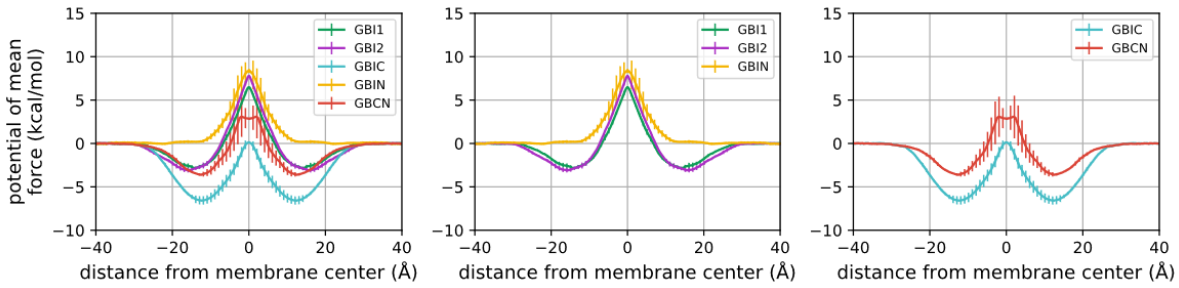


Figure B.3: Potentials of mean force of five solutes to permeate the membrane. The left panel shows all results together, and the middle and right panels compare specific subsets for 2GBI-based structures and ClGBI-based structures, respectively. Error bars denote deviation from symmetry.

The green curve represents GBI1, and the purple curve shows the pK_a shift profile for GBI2. Both results are similar, with a predicted increase in pK_a up to 2.5 pK units in the lipid headgroup region. This data seems to indicate that interactions with the acidic headgroups of the membrane with the nitrogen centers of the imidazole or guanidine moieties of 2GBI causes the molecular basicity to increase, which is consistent with fundamental acid-base principles that the basicity of an amine would increase by electron-donating groups.[223] Consequently, we draw the conclusion that the positively charged 2GBI would be less likely to donate a proton during the membrane permeation process and would retain its protonation state throughout.

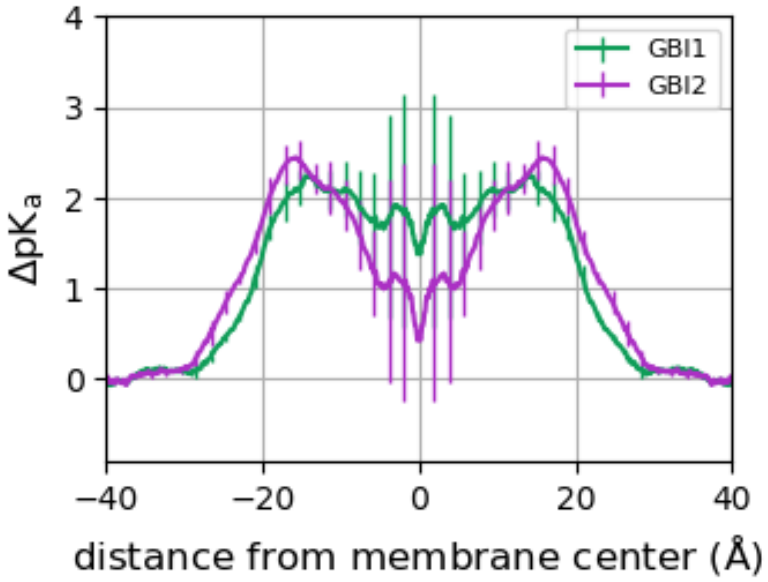


Figure B.4: pK_a shift profiles for two tautomers of 2GBI as a function of z in membrane permeation.

B.2 Flip-flop of neutral solutes in membrane center

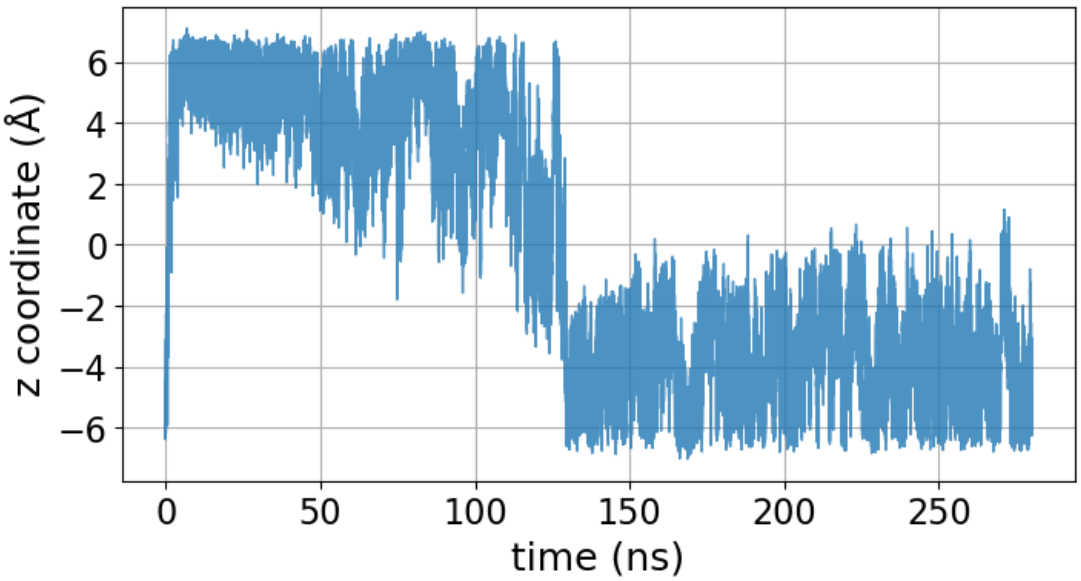


Figure B.5: A neutral solute may flip-flop in the hydrophobic interior of the membrane only 2 times over a 280 ns enhanced sampling simulation. This timescale would be even longer for positively charged solutes. Here, we plot the solute distance from the membrane center as a function of the amount of simulation time for the simulation of GBIN in an interior membrane window (a 12 Å range of $|z| = 0$-6 Å). The solute may associate with one leaflet at longer times relative to another, leading to an inconsistent sampling of phase space, making it more difficult to obtain converged results.

Appendix C

Supporting Information: Benchmark assessment of molecular geometries and energies from small molecule force fields

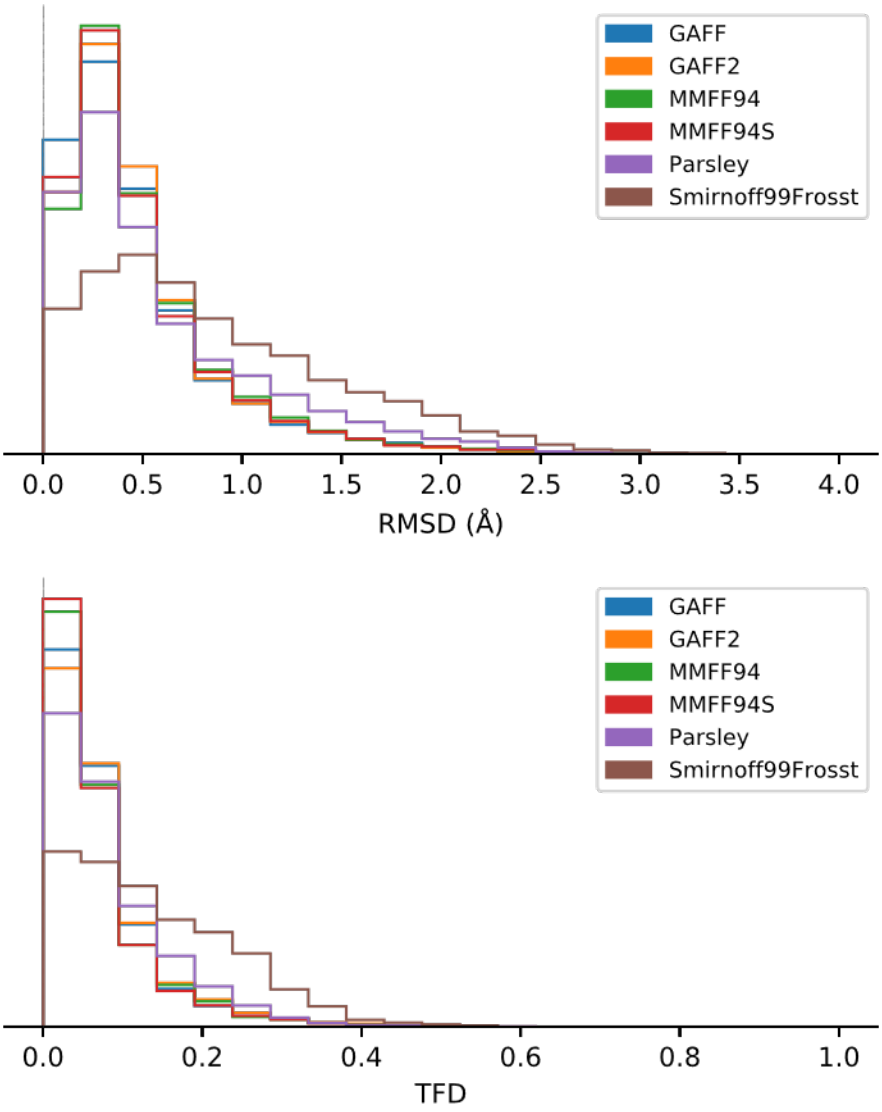


Figure C.1: Histograms of the RMSD and TFD values between force field structures as compared to QM structures. Values closer to zero indicate higher geometric similarity for both RMSD and TFD. This figure is the same as that of Figure 3 in the main text but without smoothing applied from kernel density estimation.

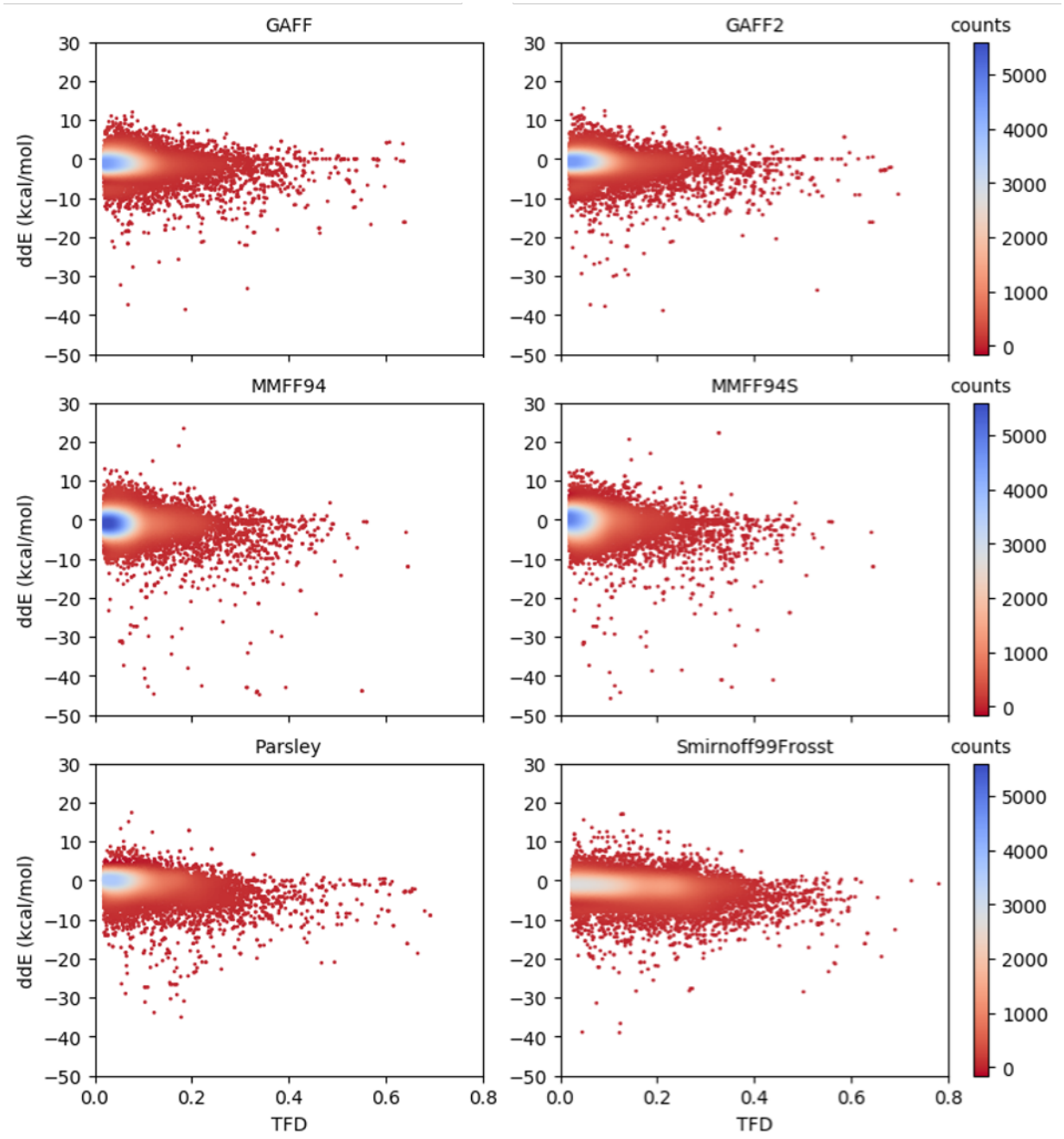


Figure C.2: Scatter plots of relative conformer energies versus TFD scores. The points are colored by the interpolated density of points in a certain area. Blue indicates region of high density, that is, high compactness of points in that area. A force field having better agreement in both relative energies and geometries with the QM reference would have more points around the origin (ddE = 0, TFD = 0). The vertical axis is represented on linear scale; the same plots with logarithmic scaling can be found in the main text (Figure 6).

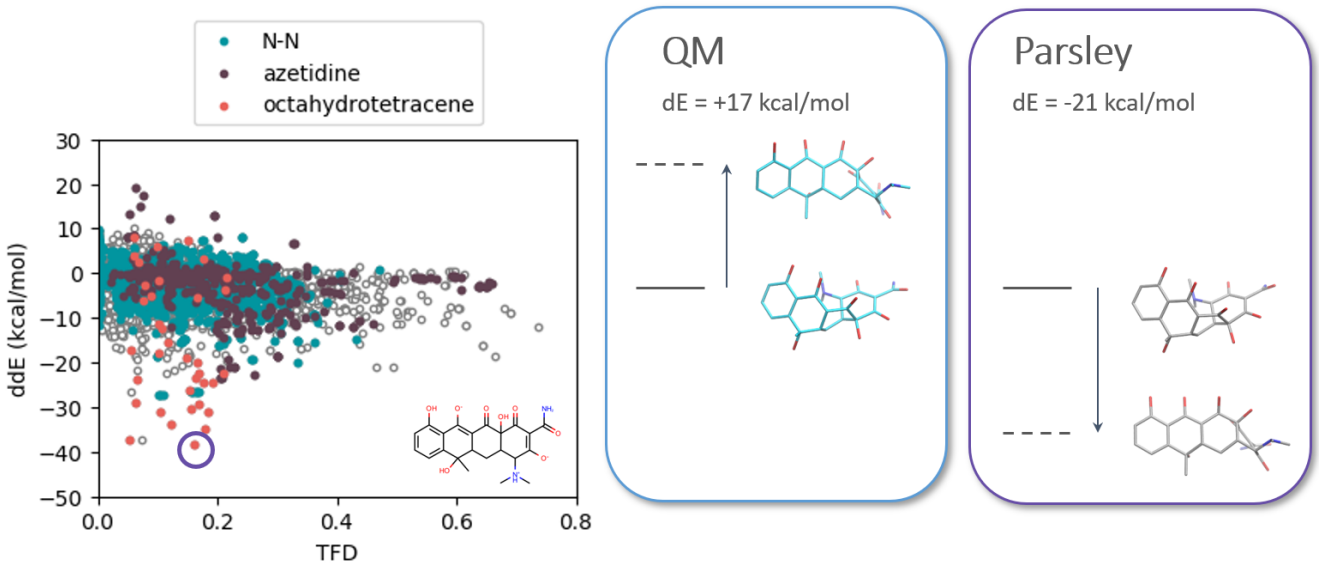


Figure C.3: An example of an outlying ddE result of the substituted octahydrotetracene structure (circled in the scatter plot). The two panels on the right show the four structures and energy differences that go into computing the ddE value of -38 kcal/mol. Energy bands are not drawn to scale. The solid band represents the energy of the reference conformer as it has the lowest QM energy for all the conformers of this molecule. The dashed band shows the energy of the example structure relative to the reference conformer. QM predicts that this conformation should be higher in energy by +17 kcal/mol, but Parsley evaluates the energy difference to be -21 kcal/mol relative to the reference.

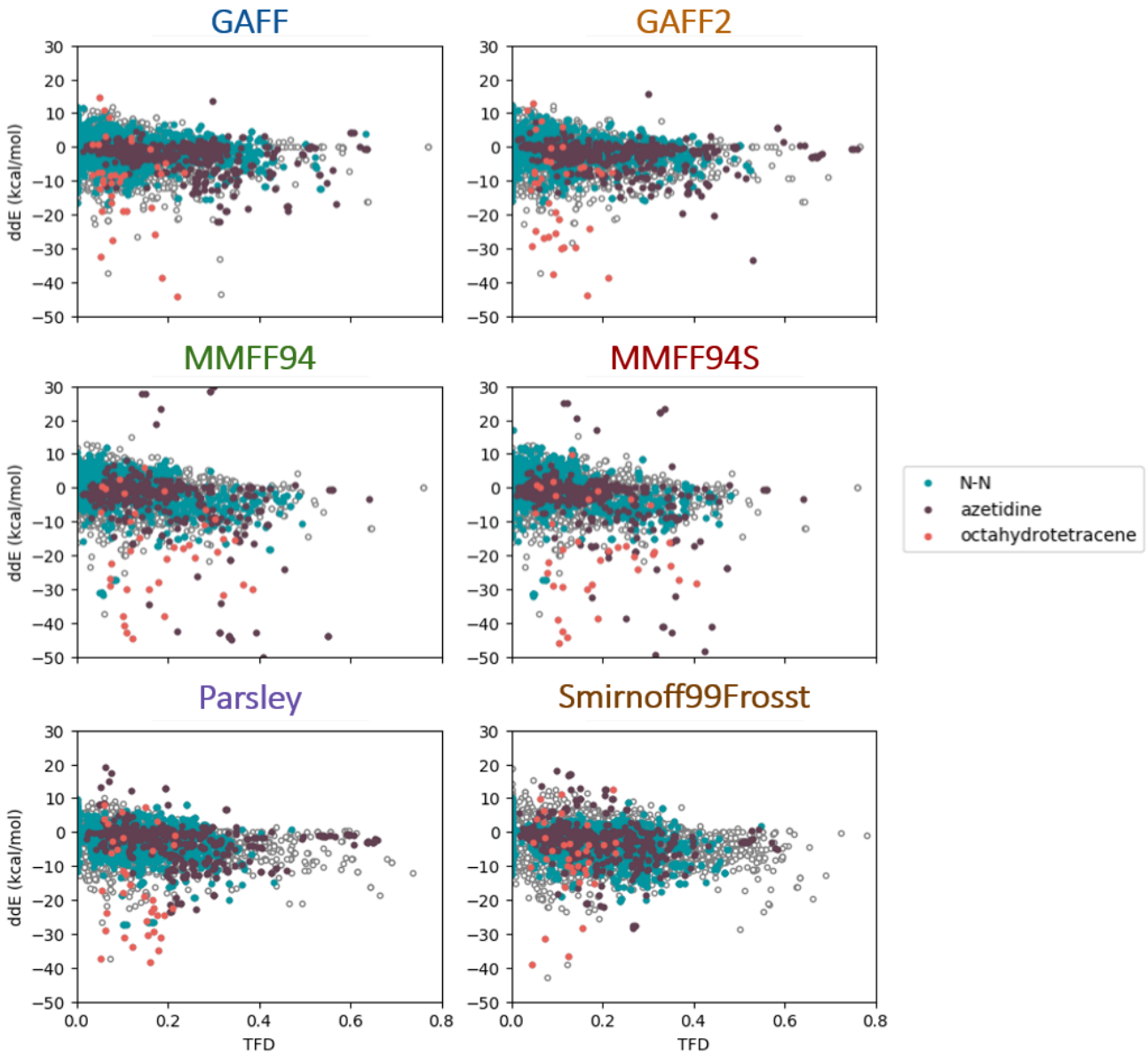


Figure C.4: Scatter plots of the all force field of relative conformer energies versus TFD scores. Colors highlight particular chemical groups that appear to be systematic outliers in energies or geometries.

Appendix D

Supporting Information: Assessing the conformational equilibrium of carboxylic acid via QM and MD studies on acetic acid

D.1 Relative energetic stabilities of trihydrated acetic acid from *ab initio* QM calculations

As a preliminary study to our pentahydrated acetic acid configurations, we evaluated the relative energies of the *syn* and *anti* conformations of acetic acid in the presence of explicit waters. We considered a microhydrated state with three explicit waters arranged to maximize hydrogen bonding to acetic acid.

D.1.1 Initial placement of water molecules for trihydrated acetic acid

The trihydrate structures for *syn* and *anti* acetic acid were manually placed using Pymol.[267] Two water molecules were placed on either side of the carbonyl group, and one was placed in line with the hydroxyl group (Figure D.1). Configurations were roughly placed to maximize intermolecular hydrogen bonding contacts between acetic acid and the water molecules. The initial and final geometries are included as XYZ files in the supporting information.

D.1.2 Description of QM calculations

For each system, a quick MM optimization was performed via OpenEye's OEChem Python Toolkit[222] using the MMFF94S force field.[111, 112, 116, 113, 117, 114, 115] Subsequently, a two-stage gas phase optimization was completed with the Psi4[301] software package. Geometries obtained from the first level of theory, MP2/def2-SV(P),[207, 325] were used as inputs for QM optimization on the subsequent level of theory, B3LYP-D3MBJ/def2-TZVP.[22, 170, 313, 285, 325] Finally, QM optimization with solvation was performed using the Turbomole version 7.1[300] software package with COSMO-TPSSh-D3BJ/def2-TZVP.[152, 151, 75, 276,

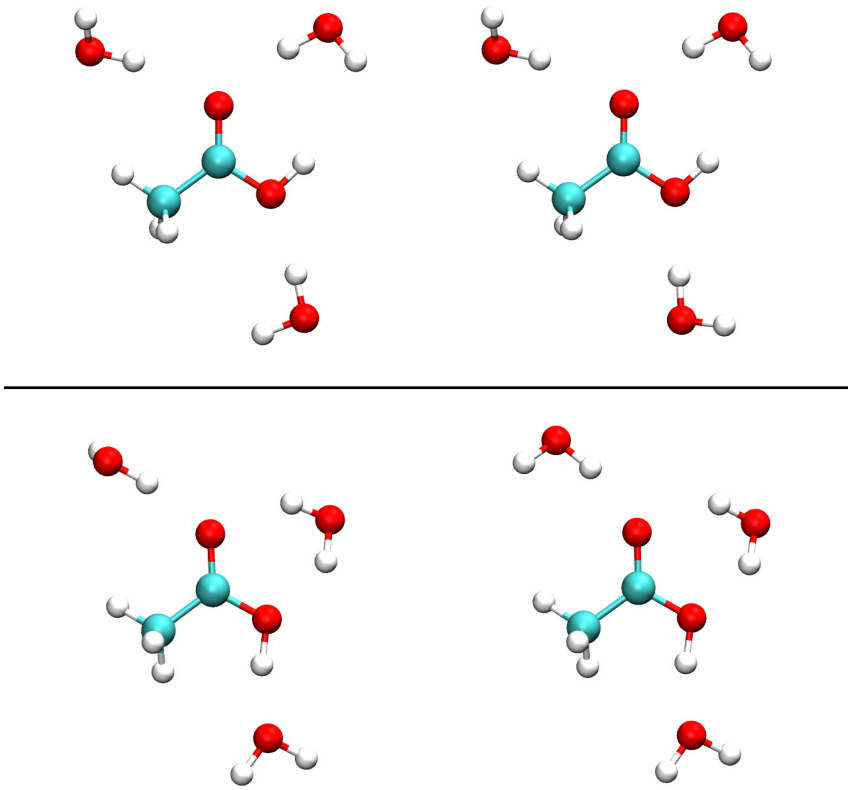


Figure D.1: Various starting conformations for three water molecules surrounding acetic acid in its *syn* (top) and *anti* (bottom) forms.

284, 105, 106, 325] At the request of a reviewer, we also ran a reference comparison with the SCS-MP2 method[104] with COSMO and the def2-TZVP basis set to evaluate alongside our COSMO-TPSSh-D3BJ calculations.

D.1.3 Evaluation of results

Ab initio QM calculations with COSMO on acetic acid with three explicit water molecules suggest an energetically preferred *anti* conformation in aqueous solution. Two arrangements of the water molecules were considered for each acetic acid conformer. In gas phase, the two *anti* configurations minimized to the same structure (Figure D.3(b)), and the minimized *syn* configurations differed in energy by about 0.11 kcal/mol. We proceeded with the lower

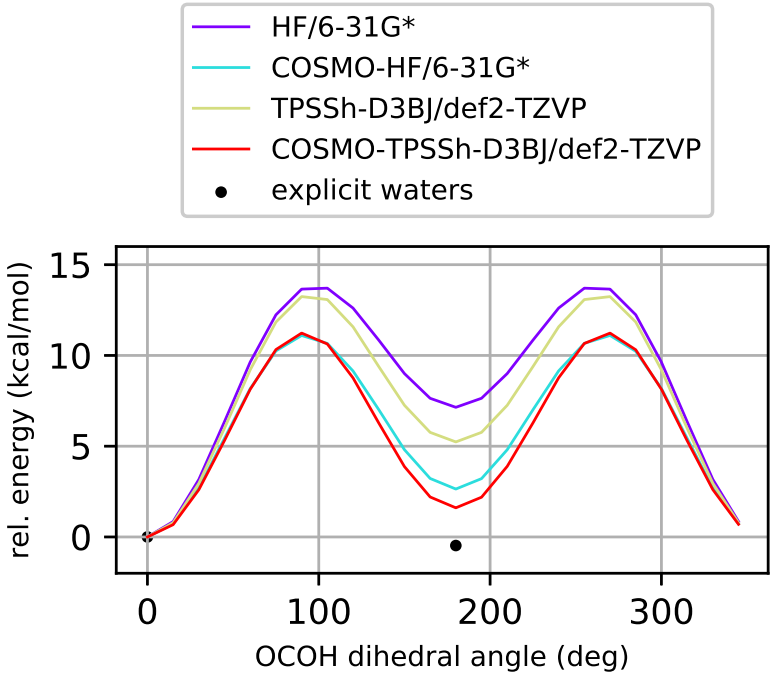


Figure D.2: QM torsion drive of acetic acid carboxyl dihedral angle for HF and TPSSh methods. Here, we include as reference two black points representing final energies from geometry optimizations of trihydrated acetic acid with COSMO-TPSSh-D3BJ/def2-TZVP.

energy *syn* configuration (Figure D.3(a)). These structures were optimized with inclusion of implicit solvent using COSMO and the TPSSh-D3BJ/def2-TZVP method. Although these acetic acid systems are surrounded by only three water molecules, observed optimized acetic acid geometries agree extremely well with past experimental and theoretical studies[213, 263] of other solvated structures (see next section). Specifically, bond lengths differed from past work by no more than 0.03 Å, and angles were within one degree. In the COSMO-optimized structures, we observe that the *anti* tetramer is the lower-energy conformation by 0.47 kcal/mol. Compared to the previously discussed torsion drive without explicit water molecules, the addition of explicit waters stabilized the *anti* form by 2 kcal/mol more than implicit solvent alone. Even with the SCS-MP2 method, the energy difference between the *syn* and *anti* trihydrate structure is minimal, showing the *syn* structure favored by 0.12 kcal/mol. The two TPSSh energy values are shown as black points in Figure D.2.

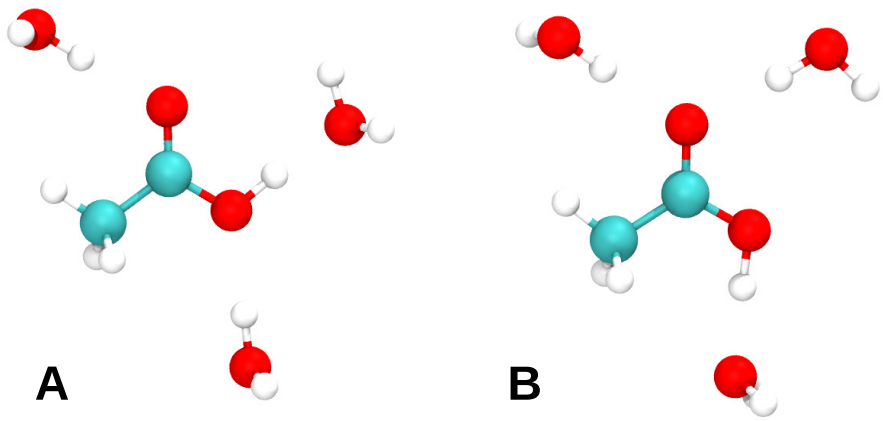


Figure D.3: Acetic acid conformation with three water molecules for (a) *syn* and (b) *anti* forms.

The water molecules in the trihydrated clusters in this section were placed to maximize the hydrogen bonding network between acetic acid and water to promote solvent stabilization. However, other arrangements of water molecules could lead to lower energy configurations. This means that results could be artifacts of water placement. Hence, we proceeded with hydrated configurations of acetic acid sampled from MD simulations.

D.1.4 Comparison of acetic acid geometries

The optimized acetic acid geometries in this work are in excellent agreement with former experimental and theoretical studies.

	COSMO solvation, trihydrate[a]	IEF-PCM solvation, free[b]	RISM-SCF/MCSCF solvation, free[c]
syn			
C–C	1.499	1.507	1.504
C=O	1.224	1.218	1.212
C–O	1.331	1.359	1.332
C–H	1.092, 1.087	1.095	not reported
O–H	1.017	0.974	0.972
C–C=O	125.0	125.8	125.1
C–C–O	113.3	111.5	113.0
C–O–H	110.8	108.2	110.5
anti			
C–C	1.501	1.513	1.506
C=O	1.231	1.214	1.213
C–O	1.319	1.363	1.322
C–H	1.092, 1.087	1.096	not reported
O–H	1.013	0.972	0.969
C–C=O	123.1	125.3	123.3
C–C–O	118.0	115.9	117.7
C–O–H	113.0	110.8	113.9

Table D.1: QM-optimized geometric parameters for acetic acid trihydrate in implicit solvent with COSMO (first column), compared to other previous work on solvated acetic acid geometries. Units of bond lengths are in Å, and angles are in degrees.

[a]Level of theory: TPSSh-D3BJ/def2-TZVP
[b]Level of theory:[213] B97D/aug-cc-pVTZ
[c]Level of theory:[263] RHF/DZP

D.2 Relative energies of pentahydrated acetic acid configurations

We present an alternate view of the violin plots for pentahydrated acetic acid referred to in the main text. The energies are shown as a scatter plot in Figure D.4.

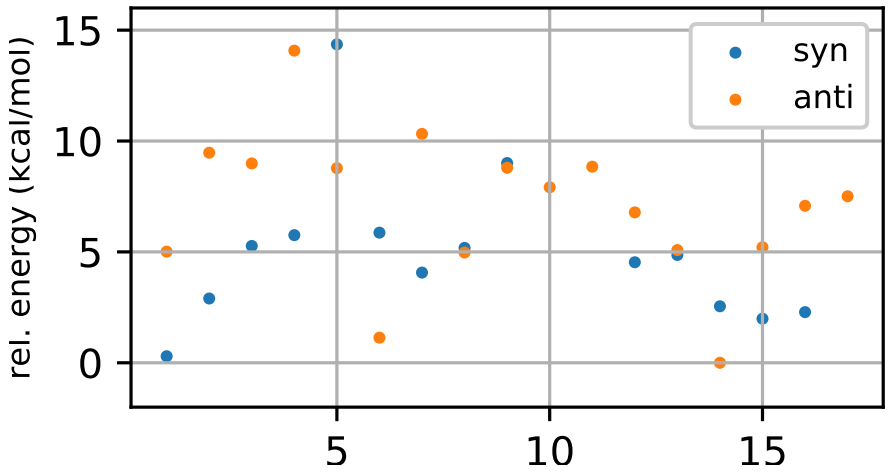


Figure D.4: Energies of pentahydrated acetic acid configurations from MD simulations. The *syn* energies are shown in blue and the *anti* energies are shown in orange. The x-axis represents arbitrary snapshot labels from MD and do not correspond between the *syn* and *anti* structures.

D.3 Comment on choice of method used for QM studies

We briefly note that the relative energies of the microhydrated acetic acid structures may be sensitive to the choice of method and basis set used in the QM calculations. As an illustrative example, we compute the relative energies of water hexamers in four arrangements: prism, cage, book, and ring (Figure D.5).

The relative energies for these hexamers are compared for B3LYP-D3/cc-pVTZ, TPSSh-D3BJ/def2-TZVP, MP2/CBS limit[19], CCSD(T)/CBS limit[19], and ZPVE+CCSD(T)/CBS limit[19]. The first two methods listed were used in the acetic acid study. We also consider these methods with COSMO implicit solvation as a point of comparison though implicit solvent was not included in the other methods.

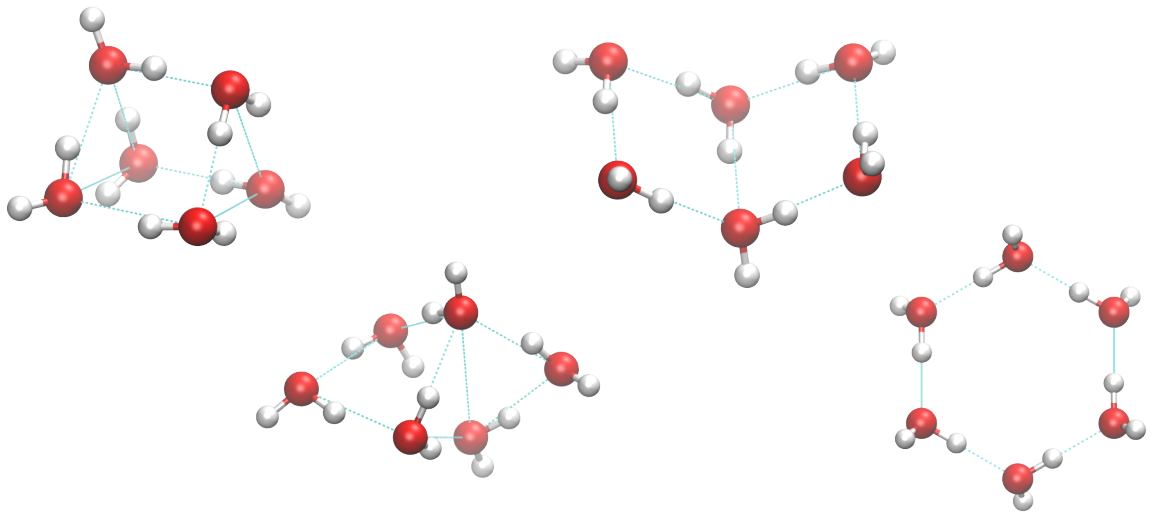


Figure D.5: Conformations of water hexamers, from left to right: prism, cage, book, and ring. Dashed lines are included for viewing guides only.

Here, considering CCSD(T)/CBS with zero-point vibrational energy corrections as the gold standard, we see that TPSSh/def2-TZVP reproduces the conformer relative energies more reliably than the B3LYP-D3/cc-pVTZ method (Figure D.6 and Table D.2). The relative energies determined by the latter method have the correct qualitative rankings for both methods in gas phase calculations, though the energy differences are exaggerated in the B3LYP method. Including COSMO with either method does not maintain the relative order of conformer energies, though we lack a proper reference for gold standard energy calculations with implicit solvation.

B3LYP and TPSSh are both hybrid functionals but on different rungs of Jacob's ladder of DFT. Both are paired with triple-zeta basis sets, cc-pVTZ and def2-TZVP, respectively. Many factors could lead to the difference between these two methods (*e.g.,* choice of functional, level of dispersion, forms of the basis functions, etc.). We do not conduct an exhaustive investigation of potential factors for these differences but demonstrate our findings to inform those pursuing related work.

The input structures for these calculations were obtained from MP2/aug-cc-pVTZ geometry optimizations provided in the supporting information of ref. 56, as was used in the protocol of ref. 19.

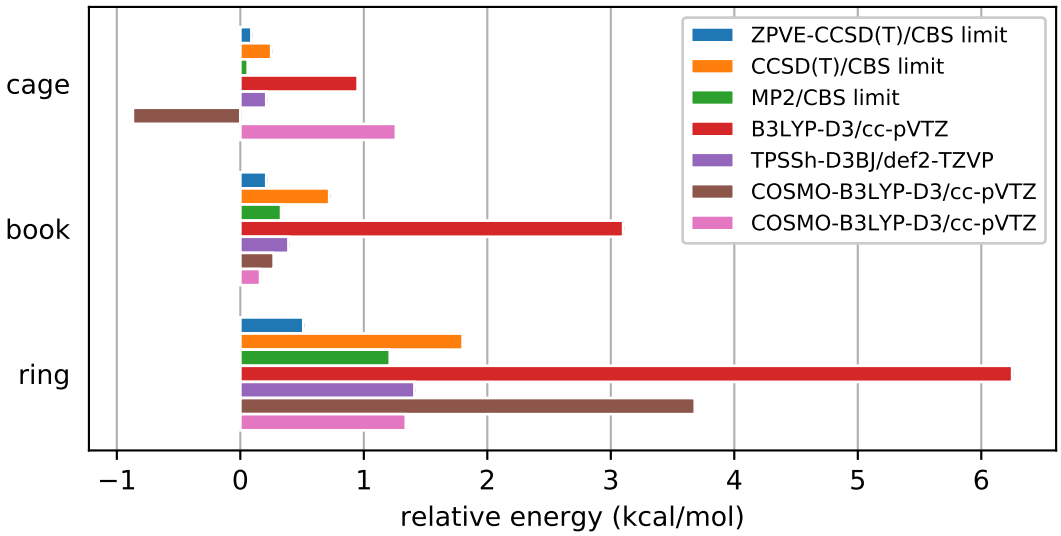


Figure D.6: Relative energy differences between various conformations of water hexamers. Energies are plotted relative to the "prism" conformation.

method	prism	cage	book	ring
ZPVE-CCSD(T)/CBS limit [a,b]	0.00	0.09	0.21	0.51
CCSD(T)/CBS limit [a]	0.00	0.25	0.72	1.80
MP2/CBS limit [a]	0.00	0.06	0.33	1.21
B3LYP-D3/cc-pVTZ	0.00	0.95	3.10	6.25
TPSSh-D3BJ/def2-TZVP	0.00	0.21	0.39	1.41
COSMO-B3LYP-D3/cc-pVTZ [c]	0.00	-0.87	0.27	3.68
COSMO-TPSSh-D3BJ/def2-TZVP [c]	0.00	1.26	0.16	1.34

Table D.2: Summary of relative energy differences between various conformations of water hexamers. Energies are listed in units of kcal/mol, relative to the "prism" conformation. CBS refers to the complete basis set limit.

[a] Ref. 19 (structures: prism, cage, book-1, cyclic-chair)
[b] These values include corrections for the zero-point vibrational energy (ZPVE).
[c] Calculations included implicit solvation with COSMO.