Molecular Simulation Methods for Conformational Searches and Diffusivity

by

© **Kari Gaalswyk**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Interdisciplinary Program in Scientific Computing

Memorial University of Newfoundland

August 2016

St. John's, Newfoundland and Labrador, Canada

Abstract

Computer modeling is a powerful technique to provide explanations and make predictions in drug development using computational methods. Molecular conformations affect drug binding and biological activity, so the preferred conformation of a drug molecule plays an important role in design and synthesis of new drugs. We have developed a conformational search method to automatically identify low energy conformations of drug molecules in an explicit solvent. This method uses replica-exchange molecular dynamics and clustering analysis to efficiently sample conformational space and identify the most probable conformations. The method produces distinct primary conformations for a molecule in explicit solvent, implicit solvent, and gas phase. Drug development is also concerned with membrane permeation. Many drugs have intracellular targets, and the rate and mechanism of membrane permeation affects their biological behavior. Transmembrane diffusion coefficients can be calculated using Generalized Langevin methods. We have compared the velocity autocorrelation and the position autocorrelation methods using molecular dynamics simulations of various solutes in homogeneous liquids, and of a water molecule harmonically restrained at various points within a lipid bilayer. Our results indicate that known limitations when using the position autocorrelation function can potentially be resolved using the velocity autocorrelation function. The effects of the spring constant and the choice of thermostat on both methods are also discussed.

Acknowledgements

I would like to thank my supervisor Dr. Chris Rowley for his support and guidance, and Ernest Awoonor-Williams for his derivation of the expressions for diffusion using the Generalized Langevin Equation in Appendix A. I would also like to thank the School of Graduate Studies at Memorial University of Newfoundland and NSERC of Canada for funding, and Compute Canada for providing computational resources.

Statement of contribution

I was responsible for writing the code used in Chapters 2 and 3, and for writing this manuscript. I was also responsible for carrying out simulations, and analyzing data for Chapters 2 and 3. I received assistance in this from my supervisor Dr. Chris Rowley, who also assisted in experimental design, and providing simulation data for Chapter 3. Ernest Awoonor-Williams contributed the derivation of the Generalized Langevin Equation that can be located in Appendix A.

Table of contents

List of tables

List of figures

List of symbols

$C_z(t)$	position autocorrelation function
$C_v(t)$	position autocorrelation function
k_B	Boltzmann constant $1.38 \times 10^{-23} JK^{-1}$
$p_z(z)$	concentration density along z-axis
δG	concentration gradient
D	diffusion coefficient
θ_{eq}	equilibrium angle
r_{eq}	equilibrium bond length
k_θ	force constant for bond angles
k_b	force constant for bonded interactions
υ_n	force constant for dihedral interactions
γ	friction coefficient
ξ	friction exerted on solute by solvent
$\hat{C}_v$	Laplace transform of VACF
σ_{ij}	Lennard-Jones radii for a given pair of atoms
ϵ_{ij}	Lennard-Jones well depths for a given pair of atoms
m	mass
n	multiplicity
ω	normalized frequency of oscillator
q_i	partial charge of atom i
P_m	permeability coefficient
ψ	phase angle
$\mathcal{V}(\vec{r})$	potential (energy) of the system when atoms hold coordinates $\vec{r}$
w	potential of mean force
p	probability
a	radius of solute
$R(t)$	random force
J	flux
μ	reduced mass of oscillator
ΔG	relative Gibbs energy
T	temperature
φ	torsional angle
η	viscosity of solvent

List of abbreviations

DPPC	1,2-dipalmitoyl-sn-glycero-3-phosphocholine
EPR	Electron paramagnetic resonance
GAFF	Generalized Amber Force Field
GBIS	Generalized Born Implicit Solvent
NpT	Isothermal-isobaric ensemble
NVT	Isothermal-isochoric ensemble
MELD	Modeling Employing Limited Data
MD	Molecular Dynamics
PME	Particle Mesh Ewald
PBC	Periodic Boundary Conditions
PSF	Protein Structure File
QSAR	Quantitative structure activity relationship
REMD	Replica-Exchange Molecular Dynamics
RESP	Restrained Electrostatic Potential
RMSD	Root mean squared deviation
SMILES	Simplified Molecular Input Line Entry System

Chapter 1

Introduction

1.1 Introduction

Computers can be used to study the properties and behaviors of physical systems by modeling processes that govern their behavior. Using these techniques, computational methods can interpret and validate experimental results, and explore properties that are difficult to study experimentally. The rapid increase in computational power has led to an increase in capability and popularity of computational methods [1, 2]. In this thesis, we will present the development of computer modeling methods that can aid the development of new pharmaceutical drugs.

One problem that benefits from computational methods is the identification of molecular conformations. A conformation is an isomer that differs in its rotation around a single bond. Conformational isomers of molecules can occur with different probabilities because of the effects of interactions within the molecule and interactions with the environment that variably stabilize or destabilize a given conformation [3, 4]. The conformation of a molecule affects its biological activity and chemical reactivity.

Conformational analysis is particularly important in the field of drug development because the conformation of a drug affects its binding behavior and efficacy [5]. Drug receptors are highly sensitive to the structure of molecules binding to them [6], so identifying potential conformations and their relative probability is an important part of drug development. For example, crystal structure analysis of the insomnia drug suvorexant determined that the drug takes on a π-stacked horseshoe conformation when binding to the human OX_2 orexin receptor [7]. Molecular simulations indicated that this conformation is a low-free-energy state and a favorable design feature for other distinct orexin receptor antagonists.

Computer modeling can be used to identify the lowest energy conformation of a molecule in solution, which can be difficult to achieve experimentally. Automated conformational search methods are particularly useful because they can quickly generate the lowest energy conformations of many different molecules automatically [8]. These methods can be systematic [9, 10, 11], Monte Carlo based [12], use genetic algorithms [13, 14], or other methods [15, 16, 17, 18, 19]. A popular conformational search method is molecular dynamics (MD) [20, 1, 21]. Each method has its advantages and drawbacks, so methods are continually being developed and refined.

Experimentally determining molecular conformations is difficult due in part to the complexity of the system, or to the difficulty in synthesizing the compound. The process is further hampered by the fact that a molecule can have a large number of conformations. Exhaustively generating all possible conformations for a molecule and calculating their energies is computationally intensive [8]. Not all conformations are equally probable; instead, the most probable conformation relates through a Boltzmann distribution as the lowest energy conformation [18, 22]. Conformational search methods are used to identify different conformations, and calculate their relative energies.

Some of the main issues in conformational searching involve balancing accuracy with computational efficiency [18, 19]. How a model represents the particles affects the number of computations required; a coarse-grained model that groups atoms into beads only calculates inter-bead interactions, while a fine-grained model that represents individual atoms has to compute interatomic interactions. Presence and representation of a solvent also play a role; including a representation of the solvent is more accurate but requires more computations. The methods themselves vary in how they search conformational space. Some methods exhaustively scan the complete conformational space. This will identify all possible conformations, so it is a rigorous method but also becomes computational expensive for systems with a large number of possible conformations. Other methods use an algorithm to sample a representative set of configurations, which reduces the number of configurations that must be generated but introduces error due to incomplete sampling. There are also many methods to speed up the simulation by smoothing the energy surface, scaling system parameters, or running multiple copies of the system, many of which can affect the accuracy of the simulation or the number of computations required [18]. New methods are needed to overcome the limitations in current methods. These new methods must also balance accuracy and efficiency in order for them to be used in practice.

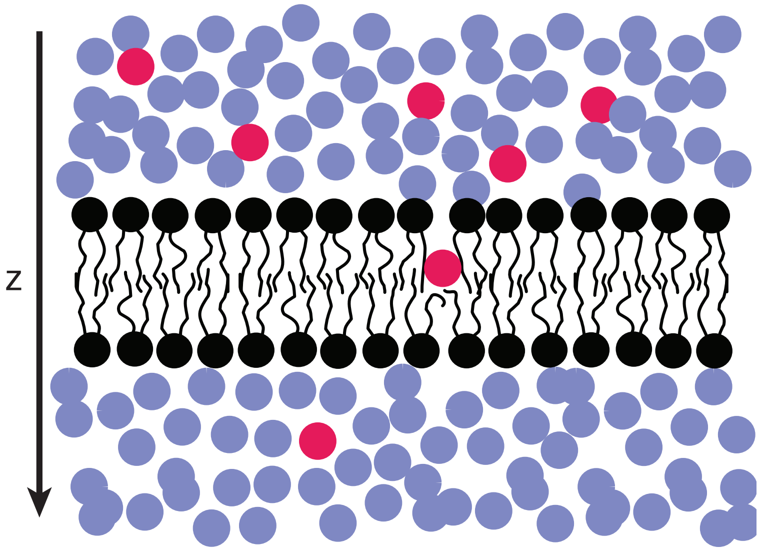

Figure 1.1: Schematic of a cell membrane. Lipid molecules form a planar bilayer (black) with aqueous phases corresponding to the cell interior and exterior (blue). If there is a concentration gradient of a solute (red) between the two phases (ΔC), there will be a net flux of the solute across the membrane. The rate of flux depends on the properties of the bilayer and solute.

A second area where computer modeling can be used is in modeling the permeation of molecules across cell membranes (Figure 1.1). Cell membranes contain and protect cellular proteins and molecules. These semi-permeable membranes allow passage of specific molecules through passive permeation [23]. The flux (J) is related to the concentration gradient across the bilayer (ΔC), and the permeability coefficient (P_m),

$$J = P_m \cdot \Delta C \tag{1.1}$$

P_m depends on the properties of the solute, the composition of the membrane, and the conditions that permeation occurs under. This permeation process is illustrated in Figure 1.2.

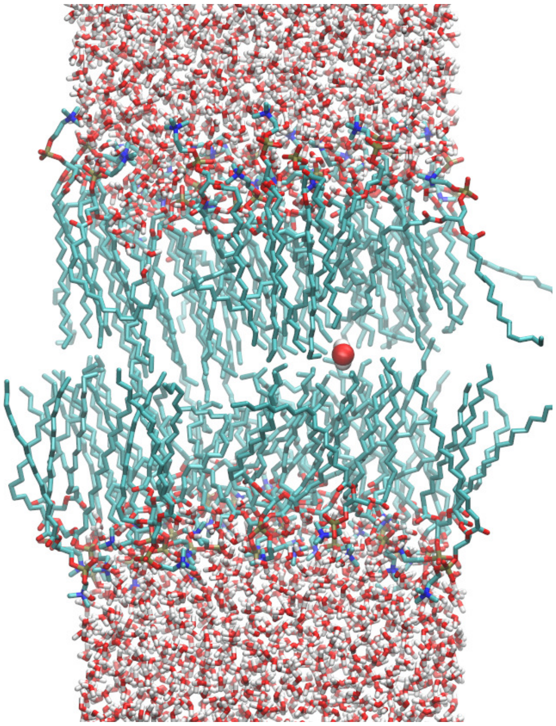

Figure 1.2: A water molecule permeating in a lipid bilayer membrane. The solute permeates through the membrane along the transmembrane coordinate that corresponds to the depth of the solute in the membrane (z).

Membrane permeability is an important factor in understanding cell function and biological barriers to drug delivery [24]. For example, the membrane permeability of the anti-psychotic drug chlorpromazine can be affected by the presence of the large unilamelar vesicle POPS and cholesterol [25]. Isothermal titration calorimetry indicated these additions change the affinity of chlorpromazine to the membrane, affecting its permeability coefficient. Many drugs have intracellular targets, but the rate and mechanism at which they permeate a membrane can be difficult to determine experimentally [23].

These two problems showcase the variety and capability of computational methods. The field of drug development has benefited greatly from advancements in computer modeling. The next sections will highlight the significance of these two computational problems.

1.2 Conformations

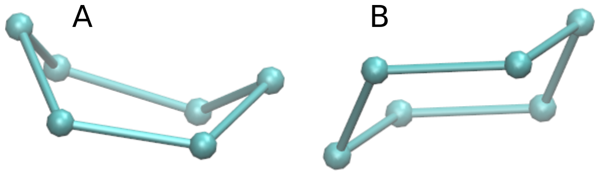

Figure 1.3: Boat (A) and chair (B) conformations of cyclohexane.

Conformations occur due to the different steric, electrostatic, and solute-solvent interactions [18, 1]. Cyclohexane has two prominent conformations, see Figure 1.3. Because of these interactions, the chair conformation is preferred over the boat conformation. These two conformations have significantly different energies. Conformational properties can be used as the basis for design, development, and synthesis of new drugs [6]. Identifying the most probable conformation is an important, but often challenging, part of the drug development process.

1.3 Transmembrane Diffusion

1.3.1 Lipid Bilayers

Lipid molecules are comprised of a polar head group that is linked to an alkyl chain by an ester linkage. For example, 1,2-dipalmitoyl-sn-glycero-3-phosphocholine (DPPC), is a lipid with a zwitterionic phosphocholine head group, a glycerol ester group, and two saturated 16-carbon alkyl chains. Figure 1.4 shows the structure of a DPPC lipid molecule.

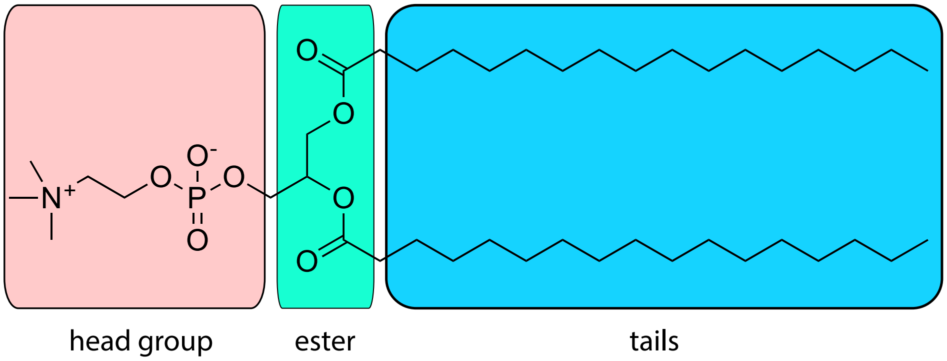

Figure 1.4: The molecular structure of a DPPC lipid. The phosphocholine head group is highlighted in red, the glycerol ester group is highlighted in green, and the alkyl tail is highlighted in blue.

In aqueous solutions, some types of lipid molecules will spontaneously form supermolecular structures like vesicles, micelles, and bilayers [26]. Lipid bilayers are planar structures comprised of two opposing monolayers. The polar head groups of the bilayer face the aqueous solutions, while the nonpolar alkyl tails form a hydrophobic membrane interior. Permeating molecules must be removed from the aqueous solvent and enter the non-polar interior, so molecules that are highly soluble in water will be energetically disfavored from permeating. The permeation of a water molecule through a model membrane is illustrated in Figure 1.2.

Lipid bilayers are of particular biological importance. Cell membranes of living organisms are predominantly comprised of phospholipid bilayers. They serve to contain cellular components and serve as a barrier to chemical species entering or exiting the cell. Transmembrane proteins selectively control the passage of specific, critical species like ions. Many other endogenous or exogenous molecules cross cell membranes by passive diffusion [24]. This is particularly important for the development of new drug molecules because many of these molecules must pass through a cell membrane through passive diffusion to reach their site of action.

1.3.2 Membrane Permeation

The process of membrane permeation can be modeled using computer simulations. The inhomogeneous solubility-diffusion model expresses P_m in terms of the potential of mean force ($w(z)$) and the diffusivity profile ($D(z)$) for a solute crossing the bilayer along the transmembrane axis, z [27, 28, 29, 30]. The permeability coefficient is expressed as an integral of these terms over an interval $[z_1, z_2]$ that spans the membrane,

$$\frac{1}{P_m} = \int_{z_1}^{z_2} \frac{e^{w(z)/k_B T}}{D(z)} \mathrm{d}z \tag{1.2}$$

There are established computational methods for calculating $w(z)$, but less effort has been devoted to the calculation of $D(z)$. Because a solute crossing a cell membrane will experience a range of chemical environments, $D(z)$ is dependent on the depth of the solute (z-position) in the membrane. Diffusion of a solute through a membrane cannot be determined using homogeneous models or calculations because the diffusion constant varies greatly as the solute moves from bulk water, through the interface, and into the membrane interior (Figure 1.2). Electron Paramagnetic Resonance (EPR) experiments have shown that the diffusion coefficient of a solute across a lipid bilayer varies considerably as a function of membrane depth [31].

Diffusion

Fundamentally, diffusion is the process by which matter spontaneously moves from a region of high concentration to a region of low concentration, and it plays a role in protein-ligand binding and membrane permeation. On a macroscopic scale, diffusion is described by Fick's Law,

$$J = -D\frac{dp_x(z)}{dz} \tag{1.3}$$

where $p_x(z)$ is the concentration along the z-axis, and D is the diffusion coefficient. Larger diffusion coefficients correspond to faster flux along a coordinate.

Diffusion is directly related to the hydrodynamic friction of the solvent through the Einstein relation,

$$D = \frac{k_B T}{m\xi} \tag{1.4}$$

where k_B is the Boltzmann constant $(1.38 \times 10^{-23} JK^{-1})$, m is the mass of the particle, and ξ represents the friction exerted on the particle by the surrounding liquid. For spherical particles with weak intermolecular interactions with the solvent, the friction can be approximated using the radius of the particle a and the viscosity of the liquid η with a mass m,

$$\xi = \frac{6\pi a\eta}{m} \tag{1.5}$$

Molecular dynamics simulations can model diffusion. The trajectories generated from a MD simulation provide an atomic-scale model that directly corresponds to the process of diffusion. Diffusion coefficients of solutes in homogeneous solutions can be calculated from a MD trajectory using the Einstein equation [32] or the Kubo relation [33]. More sophisticated techniques are needed to describe the diffusivity of solutes in heterogeneous environments where the diffusivity has a position dependence.

There are several methods to calculate diffusion across a membrane using MD simulations. These methods differ in how they compute the diffusivity from the trajectory. Two of the more common methods use Bayesian inference [34, 35] or the Generalized Langevin equation [23, 29]. Generalized Langevin methods calculate the diffusivity of a solute from a MD trajectory where the solute is harmonically restrained at a position along the z-axis. Chapter 3 of this thesis investigates these methods for use in calculating transmembrane diffusivity profiles.

1.4 Theory and Methods for Molecular Simulation

A variety of theoretical models and computational methods were used in this thesis. These methods are briefly described in the following sections.

1.4.1 Molecular Dynamics

Molecular dynamics is a technique that uses Newton's equations of motion to simulate the dynamics of a system over time [36]. These simulations must be performed numerically, where the positions of the atoms are propagated through a series of time steps,

$$r_i(t + \Delta t) = 2r_i(t) - r_i(t - \Delta t) + \frac{F_i(t)}{m_i}\Delta t^2 \tag{1.6}$$

The position of a particle i can be determined for some later time $t+ \Delta t$ given its position at the current time t by computing the forces F_i acting on that particle with a mass of m_i [21]. To limit error associated with this process, the time step for simulations of molecular systems must be small ($\approx 1 - 2$ fs).

Molecular dynamics simulations are often used to sample an isothermal-isobaric (NpT) or an isothermal-isochoric (NVT) ensemble [21]. This is accomplished by modifying the equations of motion of the dynamics so that a simulation will sample the necessary ensemble. To sample a constant temperature ensemble, the dynamics are said to be coupled to a thermostat. For an NpT simulation, the dynamics are said to be coupled to a barostat, which causes the simulation cell to vary over the course of the simulation so that system samples the ensemble consistent with the specified pressure.

Molecular dynamics has advantages over other atomistic methods such as Monte Carlo, which propagates movement using a random step direction [18, 36]. Temporal information is retained through the trajectory, allowing for computation of transport properties like diffusion, reaction rates, and protein folding times. Since the changes in the intermolecular degrees of freedom are guided, MD generates accepted configurations. This is advantageous over Monte Carlo methods, where the intermolecular degrees of freedom change randomly. This is inefficient as the resulting configurations, especially with more flexible molecules, are more likely to be rejected. Molecular dynamics is, however, more computationally expensive than other methods like Monte Carlo because of its guided step direction, especially with complex systems.

The MD simulations presented in this thesis are atomistic, meaning that atoms are represented individually. Atomistic models provide a more accurate representation of a system because they compute the individual interactions between atoms, allowing us to accurately describe molecular systems. The length of the simulation is on the nanosecond scale. Because of this fine-grained representation, atomistic models are more computationally expensive and require longer simulations. Molecular dynamics simulations can also be coarse grained, representing atoms as conglomerate beads rather than individual atoms [18].

1.4.2 Force Fields

The dynamics of the system are governed by the forces acting on the constituent atoms. These forces include bonded forces (i.e., bonds stretching, angle bending, dihedral rotations...), and non-bonded forces (i.e., electrostatic, Pauli repulsion, and London dispersion) [37]. The equations used to describe the forces on the atom are collectively referred to as the force field [22].The total potential energy function for a force field is [37],

$$\mathcal{V}(r) = \sum_{\text{bonds}} k_b(r - r_{eq})^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{\upsilon_n}{2}[1 + \cos(n\varphi - \psi)] + \\ \sum_i \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_o} \frac{1}{r_{ij}} \tag{1.7}$$

where r_{eq} is the equilibrium bond length, θ_{eq} is the equilibrium angle, k_b, k_θ , and $\mathcal{V}_n$ are the force constants, n is the multiplicity, φ is the torsional angle, and ψ is the phase angle for torsional parameters. The last summation represents the non-bonded interactions, including London dispersion forces, Pauli repulsion, and electrostatic interactions. ϵ_{ij} and σ_{ij} are the Lennard-Jones well depths and radii for a given pair of atoms, and q_i is the partial charge of atom i. At each step of a MD simulation, the energy and forces on the atoms must be calculated for the current atomic positions using this force field.

The underlying force field ultimately determines the properties of the system that are calculated using MD simulations [38]. As a result, it is essential to use a force field that accurately describes the properties of the system. The parameters of the force field are often determined using experimental data of the condensed-phase properties of small molecules or quantum mechanical calculations [39, 40, 41, 42]. Using these parameters, force fields have been developed that describe larger molecules and even biomacromolecules like lipids [37, 43, 44].

1.4.3 Periodic Boundary Conditions and Long Range Forces

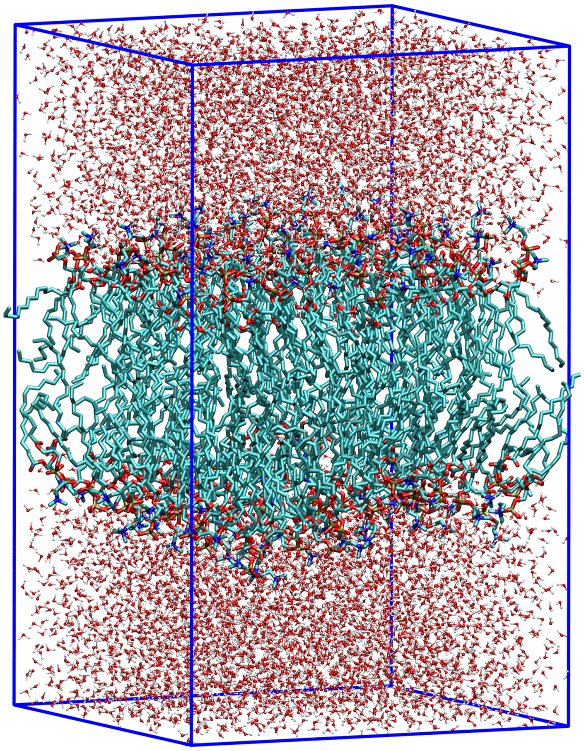

Figure 1.5: An example periodic simulation cell for a DPPC lipid bilayer. The lipid tails (blue) form a layer in the centre of the cell. The head groups of the lipids form an interface with the water molecules (red) that form solvent layers above and below the bilayer.

In order to simulate a bulk solution, periodic boundary conditions (PBC) are used. A unit cell is repeated such that a particle in one cell interacts with particles in a neighboring cell, and a particle that leaves the cell on one side reappears on the other side [38]. A periodic cell used to simulate a lipid bilayer is depicted in Figure 1.5.

Periodic systems formally have an infinite number of non-bonded interactions between the atoms comprising the system. Dispersion interactions have the form of

$\mathcal{V}(r) \propto 1/r^6$, so the strength of these interactions becomes negligible after at a relatively short distance (e.g., 10 Å). As a result, these interactions can be truncated at a fixed distance using a smoothed potential.

Electrostatic interactions cannot be as easily truncated as dispersion interactions because Coulombic interactions are very long-range ($\mathcal{V}(r) \propto 1/r$). Particle Mesh Ewald (PME) divides these interactions into short and long-range using a Gaussian distribution function [45]. The long-range component of these interactions are calculated by mapping the charges onto a grid and then the interactions are calculated using the Fast Fourier Transform [38, 22]. The remaining real space component of the electrostatic interactions are now short-range, so they can be truncated at modest distances.

There are many other modeling techniques that vary in their scale from microscopic methods, like MD, to macroscopic methods like kinetic models and fluid dynamics. Kinetic models use coupled ordinary differential equations to represent chemical reactions [46]. They can also be coupled to partial differential equations to describe fluid dynamics. These methods are often applied to population dynamics, or other kinetic rate problems [22]. Fluid dynamic models represent fluid as a continuum using the Navier–Stokes equation [47]. These models assume that the density of a fluid is high enough to describe it as continuum, and can thus specify a mean velocity and a mean kinetic energy [48]. This allows the model to easily define properties such as temperature and density at any point in the continuum. Fluid dynamics is used to describe transport phenomena and other equations of fluid motion.

1.4.4 Solvation Methods

An important part of a MD simulation is the representation of the solvent. The presence and representation of a solvent can affect the conformation of a molecule [22, 4].

Solute–solvent interactions affect the conformation by limiting the solutes movement and available conformations. A simulation in the gas phase, while computationally much simpler, is able to access conformations unavailable to a solvated molecule. The resulting trajectory does not represent the configurations a solvated system would take.

A solvent can be represented either implicitly or explicitly. The Generalized Born Implicit Solvent (GBIS) method represents the solvent as a dielectric continuum [49]. Explicit solvent models represent the solvent as discrete particles and explicitly calculate solute–solvent interactions.

1.4.5 Replica-Exchange Molecular Dynamics

One of the limitations of MD is that simulations can become stuck in a local minimum without enough energy to cross some barrier in the energy landscape. This means that the simulation does not fully sample conformational space, and thus may not find the lowest energy conformation [50].

One method to overcome this is Replica-Exchange Molecular Dynamics (REMD). Multiple copies of the system are simulated at distinct temperatures. Periodically, neighboring replicas attempt to exchange temperatures and velocities. Because replicas at higher temperature are able to overcome barriers in the energy landscape, REMD simulations are better able to sample the conformational space [51, 52].

1.4.6 Clustering Analysis

A REMD simulation returns a trajectory describing all the configurations the system occupied during the simulation. Clustering analysis is used to extract the configuration the system took on most during a simulation, which equates to the most probable and thus the lowest energy conformation. Conformations are grouped based on some

Cartesian distance metric [53]. The lowest energy conformation equates to the largest cluster.

1.5 Outline

The original research presented in this thesis is divided into two chapters. A conformational search method for explicitly solvated molecules is described in Chapter 2. Chapter 3 evaluates methods for calculating transmembrane diffusion coefficients based on the Generalized Langevin Equation.

Bibliography

[1] Wilfred F. van Gunsteren and Herman J. C. Berendsen. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie International Edition in English*, 29(9):992–1023, 1990.

[2] G. E. Moore. Cramming More Components Onto Integrated Circuits. *Electronics*, 38(8):114–117, 1965.

[3] Gordon Crippen and Timothy F. Havel. *Distance Geometry and Molecular Conformation*. John Wiley and Sons, New York, 1988.

[4] Ramu Anandakrishnan, Aleksander Drozdetski, Ross C. Walker, and Alexey V. Onufriev. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophysical Journal*, 108(5):1153–1164, 2015.

[5] Robert A Copeland. Conformational adaptation in drug–target interactions and residence time. *Future Medicinal Chemistry*, 3(12):1491–1501, 2011.

[6] R. S. Struthers, J. Rivier, and A. T. Hagler. Molecular Dynamics and Minimum Energy Conformations of GnRH and Analogs: A Methodology for Computer-aided Drug Design. *Annals of the New York Academy of Sciences*, 439(1):81–96, 1985.

[7] Jie Yin, Juan Carlos Mobarec, Peter Kolb, and Daniel M. Rosenbaum. Crystal structure of the human OX2 orexin receptor bound to the insomnia drug suvorexant. *Nature*, 519(7542):247–250, March 2015.

[8] K. Shawn Watts, Pranav Dalal, Robert B. Murphy, Woody Sherman, Rich A. Friesner, and John C. Shelley. ConfGen: A Conformational Search Method for

Efficient Generation of Bioactive Conformers. *Journal of Chemical Information and Modeling*, 50(4):534–546, 2010.

[9] Ekaterina I. Izgorodina, Ching Yeh Lin, and Michelle L. Coote. Energy-directed tree search: an efficient systematic algorithm for finding the lowest energy conformation of molecules. *Physical Chemistry Chemical Physics*, 9(20):2507–2516, 2007.

[10] Paul C. D. Hawkins, A. Geoffrey Skillman, Gregory L. Warren, Benjamin A. Ellingson, and Matthew T. Stahl. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling*, 50(4):572–584, 2010.

[11] Roberto Vera Yasset Perez-Riverol. A Parallel Systematic-Monte Carlo Algorithm for Exploring Conformational Space. *Current Topics in Medicinal Chemistry*, 12(16), 2012.

[12] Maria A. Miteva, Frederic Guyon, and Pierre Tuffry. Frog2: Efficient 3d conformation ensemble generator for small compounds. *Nucleic Acids Research*, 38(suppl 2):W622–W627, 2010.

[13] Yoshitake Sakae, Tomoyuki Hiroyasu, Mitsunori Miki, Katsuya Ishii, and Yuko Okamoto. Combination of genetic crossover and replica-exchange method for conformational search of protein systems. *arXiv:1505.05874 [cond-mat, physics:physics, q-bio]*, 2015. arXiv: 1505.05874.

[14] Adriana Supady, Volker Blum, and Carsten Baldauf. First-Principles Molecular Structure Search with a Genetic Algorithm. *Journal of Chemical Information and Modeling*, 55(11):2338–2348, 2015.

[15] Justin L. MacCallum, Alberto Perez, and Ken A. Dill. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22):6985–6990, 2015.

[16] T. J. Brunette and Oliver Brock. Guiding conformation space search with an all-atom energy potential. *Proteins*, 73(4):958–972, 2008.

[17] Daniel Cappel, Steven L. Dixon, Woody Sherman, and Jianxin Duan. Exploring conformational search protocols for ligand-based virtual screening and 3-D QSAR modeling. *Journal of Computer-Aided Molecular Design*, 29(2):165–182, 2014.

[18] Markus Christen and Wilfred F. van Gunsteren. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *Journal of Computational Chemistry*, 29(2):157–166, 2008.

[19] Kaihsu Tai. Conformational sampling for the impatient. *Biophysical Chemistry*, 107(3):213–220, 2004.

[20] S Crouzy, T B Woolf, and B Roux. A molecular dynamics study of gating in dioxolane-linked gramicidin A channels. *Biophysical Journal*, 67(4):1370–1386, 1994.

[21] Harold A. Scheraga, Mey Khalili, and Adam Liwo. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annual Review of Physical Chemistry*, 58(1):57–83, 2007.

[22] Wilfred F. van Gunsteren, Dirk Bakowies, Riccardo Baron, Indira Chandrasekhar, Markus Christen, Xavier Daura, Peter Gee, Daan P. Geerke, Alice Glttli, Philippe H. Hnenberger, Mika A. Kastenholz, Chris Oostenbrink, Merijn

Schenk, Daniel Trzesniak, Nico F. A. van der Vegt, and Haibo B. Yu. Biomolecular Modeling: Goals, Problems, Perspectives. *Angewandte Chemie International Edition*, 45(25):4064–4092, 2006.

[23] Christopher T. Lee, Jeffrey Comer, Conner Herndon, Nelson Leung, Anna Pavlova, Robert V. Swift, Chris Tung, Christopher N. Rowley, Rommie E. Amaro, Christophe Chipot, Yi Wang, and James C. Gumbart. Simulation-based approaches for determining membrane permeability of small compounds. *J. Chem. Inf. Model.*, 56(4):721–733, 2016.

[24] T.-X. Xiang and B. D. Anderson. Phospholipid surface density determines the partitioning and permeability of acetic acid in DMPC:cholesterol bilayers. *The Journal of Membrane Biology*, 148(2):157–167, 1995.

[25] Patrcia T. Martins, Adrian Velazquez-Campoy, Winchil L. C. Vaz, Renato M. S. Cardoso, Joana Valrio, and Maria Joo Moreno. Kinetics and Thermodynamics of Chlorpromazine Interaction with Lipid Bilayers: Effect of Charge and Cholesterol. *Journal of the American Chemical Society*, 134(9):4184–4195, March 2012.

[26] Jacob N. Israelachvili. *Intermolecular and Surface Forces.* Academic Press, San Diego, 3rd ed. edition, 2011.

[27] Jared M. Diamond and Yehuda Katz. Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water. *J. Membr. Biol.*, 17(1):121–154, 1974.

[28] Siewert J. Marrink and Herman J. C. Berendsen. Permeation process of small molecules across lipid membranes studied by molecular dynamics simulations. *J. Phys. Chem*, 100(41):16729–16738, 1996.

[29] Saleh Riahi and Christopher N. Rowley. Why can hydrogen sulfide permeate cell membranes? *J. Am. Chem. Soc.*, 136(43):15111–15113, 2014.

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

[31] Witold K. Subczynski, Magdalena Lomnicka, and James S. Hyde. Permeability of Nitric Oxide through Lipid Bilayer Membranes. *Free Radical Research*, 24(5):343–349, 1996.

[32] A. Einstein. *Investigations on the Theory of the Brownian Movement.* Dover Books on Physics Series. Dover Publications, 1956.

[33] R. Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.

[34] Gerhard Hummer. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7(1):34, 2005.

[35] Jeffrey Comer, Christophe Chipot, and Fernando D. Gonzlez-Nilo. Calculating position-dependent diffusivity in biased molecular dynamics simulations. *J. Chem. Theory Comput.*, 9(2):876–882, 2013.

[36] Adam Liwo, Cezary Czaplewski, Stanisaw Odziej, and Harold A Scheraga. Computational techniques for efficient conformational sampling of proteins. *Current Opinion in Structural Biology*, 18(2):134–139, 2008.

[37] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and

David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[38] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable Molecular Dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[39] William L. Jorgensen and Julian Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.*, 102(19):6665–6670, 2005.

[40] Saleh Riahi and Christopher N. Rowley. A Drude Polarizable Model for Liquid Hydrogen Sulfide. *J. Phys. Chem. B*, 117(17):5222–5229, 2013.

[41] Saleh Riahi and Christopher N. Rowley. Solvation of hydrogen sulfide in liquid water and at the water–vapor interface using a polarizable force field. *J. Phys. Chem. B*, 118(5):1373–1380, 2014.

[42] Archita N. S. Adluri, Jennifer N. Murphy, Tiffany Tozer, and Christopher N. Rowley. Polarizable force field with a σ-hole for liquid and aqueous bromomethane. *J. Phys. Chem. B*, 119(42):13422–13432, 2015.

[43] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, 2010.

[44] Jing Huang and Alexander D. MacKerell. CHARMM36 all-atom additive protein

force field: validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, September 2013.

[45] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[46] David R. Mott, Elaine S. Oran, and Bram van Leer. A Quasi-Steady-State Solver for the Stiff Ordinary Differential Equations of Reaction Kinetics. *Journal of Computational Physics*, 164(2):407–428, 2000.

[47] Yilong Bai, Jianxiang Wang, Daining Fang, Shiyi Chen, Moran Wang, and Zhenhua Xia. Mechanics for the World: Proceedings of the 23rd International Congress of Theoretical and Applied Mechanics, ICTAM2012multiscale Fluid Mechanics and Modeling. *Procedia IUTAM*, 10:100–114, 2014.

[48] Jiri Blazek. *Computational Fluid Dynamics: Principles and Applications*. Butterworth-Heinemann, 2015.

[49] M. Bhandarkar, A. Bhatele, E. Bohm, R. Brunner, F. Buelens, C. Chipot, A. Dalke, S. Dixit, G. Fiorin, P. Freddolino, P. Grayson, J. Gullingsrud, A. Gursoy, D. Hardy, C. Harrison, J. Hnin, W. Humphrey, D. Hurwitz, N. Krawetz, S. Kumar, D. Kunzman, J. Lai, C. Lee, R. McGreevy, C. Mei, M. Nelson, J. Phillips, O. Sarood, A. Shinozaki, D. Tanner, D. Wells, G. Zheng, and F. Zhu. *NAMD User's Guide*. University of Illinois and Beckman Institute, 2015.

[50] David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[51] Oren M. Becker, Alexander D. MacKerell Jr., Benoit Roux, and Masakatsu Watanabe, editors. *Computational Biochemistry and Biophysics*. Marcel Dekker, Inc., New York, 2001.

[52] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1–2):141–151, 1999.

[53] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

Chapter 2

An Explicit-Solvent Conformation Search Method Using Open Software

This chapter is based on an article published in PeerJ:

Kari Gaalswyk and Christopher N. Rowley. An explicit-solvent conformation search method using open software. *PeerJ*, 4:e2088, 2016.

2.1 Introduction

Many molecules can exist in multiple conformational isomers. Conformational isomers have the same chemical bonds, but differ in their 3D geometry because they hold different torsional angles [1]. The conformation of a molecule can affect chemical reactivity, molecular binding, and biological activity [2, 3]. Conformations differ in stability because they experience different steric, electrostatic, and solute-solvent interactions. The probability, p, of a molecule existing in a conformation with index i, is related to its relative Gibbs energies through the Boltzmann distribution,

$$p_i = \frac{\exp(-\Delta G_i/k_B T)}{\sum_j \exp(-\Delta G_j/k_B T)} \tag{2.1}$$

where k_B is the Boltzmann constant, T is the temperature, and ΔG is the relative Gibbs energy of the conformation. The denominator enumerates over all conformations.

Alternatively, the probability of a conformation can be expressed in classical statistical thermodynamics in terms of integrals over phase space,

$$p_i = \frac{\int_i \exp(-\mathcal{V}(\vec{r})/k_B T)\mathrm{d}\vec{r}}{\int \exp(-\mathcal{V}(\vec{r})/k_B T)\mathrm{d}\vec{r}} \tag{2.2}$$

The integral over configurational space in the numerator is restricted to coordinates corresponding to conformation i. The denominator is an integral over all configurational space. $\mathcal{V}(\vec{r})$ is the potential of the system at when the atoms hold coordinates $\vec{r}$.

Computational chemistry has enabled conformational analysis to be performed systematically and quantitatively with algorithms to generate different conformations and calculate their relative stability. Automated conformational search algorithms can generate possible conformations, and molecular mechanical or quantum methods can determine their relative energies.

Conformational search methods can be classified as either exhaustive/systematic or heuristic. Exhaustive methods scan all, or a significant portion of the configuration space. Subspaces corresponding to high energy structures can be eliminated without a loss in quality using *a priori* knowledge regarding the structure of the configuration space to be searched [4]. These methods are usually limited to small molecules due to the computational cost of searching so much of the configuration space. Heuristic methods generate a representative set of conformations by only visiting a small fraction of configuration space [5]. These methods can be divided into non-step and step methods. Non-step methods generate a series of system configurations that are independent of each other. Step methods generate a complete system configuration in a stepwise manner by a) using configurations of molecular fragments, or b) using the previous configuration [4].

2.1.1 Solvent Effects

A solvent can also affect the conformation of a molecule by effects like solvent-solute hydrogen bonding, dipole-dipole interactions, etc. [4] Incorporating the effect of solvation can complicate conformation searches. It is common to perform a conformation in the gas phase, neglecting solvent effects altogether. Alternatively, the solvent can be included in the simulation either implicitly or explicitly.

Implicit models approximate the solvent as a dielectric continuum interacting with the molecular surface [6]. Depending on the model used, the computational cost of calculating the solvation can be modest, allowing solvation effects to be included in the conformation search. A common and efficient implicit solvent method used with molecular mechanical models is the Generalized Born Implicit Solvent (GBIS) method [7]. A limitation of this type of model is that features like solute-solvent hydrogen bonding and solute-induced changes in the solvent structure are difficult to describe accurately when the solvent is described as a continuum.

Explicit solvation methods surround the solute with a number of solvent molecules that are represented as discrete particles. Provided that this model accurately describes solvent molecules and their interactions with the solute, some of the limitations in accuracy associated with implicit solvent models can be overcome. Although the accuracy of these models is potentially an improvement over continuum models, the inclusion of explicit solvent molecules presents challenges in conformation searches. Some conformational search algorithms that arbitrarily change dihedral angles cannot be used in an explicit solvent because an abrupt change in a solute dihedral angle can cause an overlap with solvent molecules.

A significant drawback of explicit solvent representations is that the computational cost of these simulations is increased considerably due to the additional computations needed to describe the interactions involving solvent molecules. Longer simulations are also needed to thoroughly sample the configurations of the solvent; the stability of each conformation is the result of a time average over an ensemble of possible solvent configurations (i.e., its Gibbs/Helmholtz energy), rather than the potential energy of one minimum-energy structure.

2.1.2 Previous Work

Many conformational search methods have been developed. Sakae et al. used a combination of genetic algorithms and replica exchange [8]. They employed a two point crossover, where consecutive amino acid residues were selected at random from each pair, and then the dihedral angles were exchanged between them. Superior conformations were selected using the Metropolis criterion, and these were then subjected to replica-exchange. Supady et al. also used a genetic algorithm where the parents were chosen using a combination of three energy-based probability metrics [9].

One example of a systematic method is the tree searching method of Izgorodina et al. [10]. The method optimizes all individual rotations, and then ranks their energies. It then eliminates those with relative energies greater than the second lowest energy conformation from the previous round, and performs optimizations on only the remaining subset. After a set number of rotations, the lowest ranked conformation is selected. Brunette and Brock developed what they called a model-based search, and compared it to traditional Monte Carlo [11]. The model-based search characterizes regions of space as funnels by creation an energy-based tree where the root of the tree corresponds to the bottom of the funnel. The funnel structure illustrates the properties of the energy landscape and the sample relationships. Cappel et al. tested the effects of conformational search protocols on 3D quantitative structure activity relationship (QSAR) and ligand based virtual screening [12].

Perez-Riverol et al. developed a parallel hybrid method that follows a systematic search approach combined with Monte Carlo-based simulations [13]. The method was intended to generate libraries of rigid conformations for use with virtual screening experiments.

Some methods have been extended to incorporate physical data. MacCallum et al. developed a physics-based Bayesian computational method [14] to find preferred structures of proteins. Their Modeling Employing Limited Data (MELD) method identifies low energy conformations from replica-exchange molecular dynamics simulations that are subject to biases that are based on experimental observations.

2.1.3 Conformation Searches Using Molecular Dynamics

Molecular dynamics (MD) simulations are a popular method for sampling the conformational space of a molecule. Equations of motion are propagated in a series of short time steps that generates a trajectory describing the motion of the system. These simulations are usually coupled to a thermostat to sample a canonical or isothermal–isobaric ensemble for the appropriate thermodynamic state. This approach is naturally compatible with explicit solvent models because the dynamics will naturally sample the solvent configurations. For a sufficiently long MD simulation, the conformational states of the molecule will be sampled with a probability that reflects their relative Gibbs/Helmholtz energies. This is in contrast to many conformational search methods that can search for low *potential* energy conformations.

One of the limitations of MD is that very long simulations may be needed to sample the conformational states of a molecule with the correct weighting. This occurs because MD simulations will only rarely cross high barriers between minima, so a simulation at standard or physiological temperatures may be trapped in its initial conformation and will not sample the full set of available conformations.

Replica Exchange Molecular Dynamics (REMD) enhances the sampling efficiency of conventional MD by simulating multiple copies of the system at a range of temperatures. Each replica samples an ensemble of configurations occupied at its corresponding temperature. Periodically, attempts are made to exchange the configurations of neighboring systems (see Figure 2.1). The acceptance or rejection of these exchanges is determined by an algorithm analogous to the Metropolis Monte Carlo algorithm, which ensures that each replica samples its correct thermodynamic distribution. This type of simulation is well suited for parallel computing because replicas can be divided between many computing nodes. Exchanges between the replicas are only attempted after hundreds or thousands of MD steps, so communication overhead between replicas is low compared to a single parallel MD simulation.

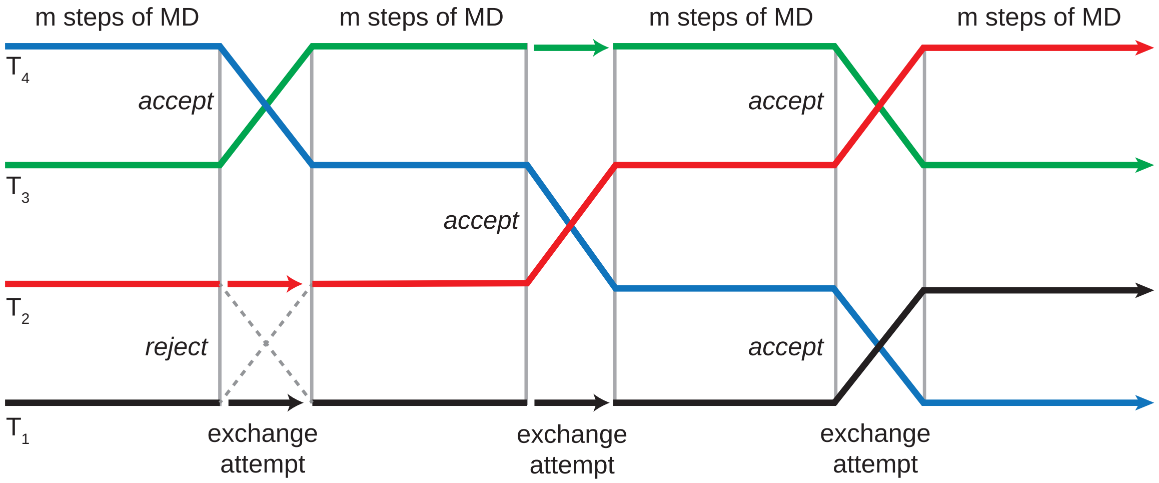

Figure 2.1: Schematic of exchange attempts between four replicas simulated at temperatures T_1, T_2, T_3, and T_4. After a large number of exchanges, each replica will have been simulated at the full range of temperatures. The lowest temperature replica will have contributions from each simulation.

REMD simulations can sample the conformational space of a molecule more completely because the higher temperature replicas can cross barriers more readily. Analysis of the statistical convergence of REMD simulations has shown that when there are significant barriers to conformational isomerization, an REMD simulation of m replicas is more efficient than a single-temperature simulation running m times longer [15]. The lowest temperature replica is typically the temperature of interest. Exchanges allow each replica to be simulated at each temperature in the set. Barriers that prevent complete sampling at low temperatures can be overcome readily at high temperatures.

After a sufficiently long REMD simulation, the trajectory for this replica will contain a correctly-weighted distribution of the conformations available at this temperature. This trajectory must be analyzed to group the structures sampled into distinct conformations.

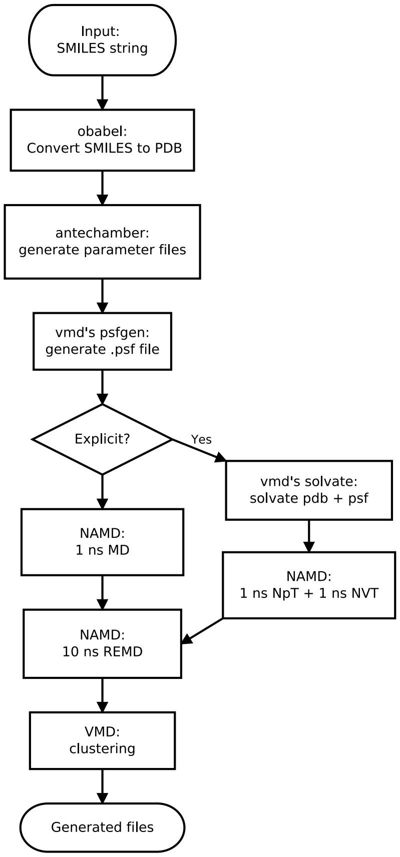

Figure 2.2: The work-flow for the conformation search method presented in this paper. A parent script executes OpenBabel, VMD, and NAMD to generate the set of lowest energy conformations.

2.1.4 Cluster Analysis

The product of an REMD simulation is a trajectory for each temperature. For a sufficiently long simulation where the simulations were able to cross barriers freely, the configurations will be sampled according to their equilibrium probability. A discrete set of conformations must be identified from this trajectory. Cluster analysis can be used to identify discrete conformations in this ensemble by identifying groups of conformations that have similar geometries according to a chosen metric. Clustering works by assigning a metric to each configuration, measuring the distance between pairs of these configurations, and then grouping similar configurations into conformations based on this distance metric. Cluster analysis allows common conformations to be identified from the configurations of a trajectory using little to no *a priori* knowledge.

2.1.5 Work Undertaken

In this paper, we present the implementation of a work flow for conformation searches using REMD and cluster analysis (see Figure 2.2). This method supports conformation searches for molecules in the gas phase, implicit solvents, and explicit solvents. The method is implemented by integrating open source software using Python scripting. Examples of the conformations search results for two drug molecules are presented.

2.2 Theory

2.2.1 Replica Exchange Molecular Dynamics

In replica exchange molecular dynamics, m non-interacting replicas of the system are run, each at its own temperature, T_m . Periodically, replicas i and j exchange coordinates and velocities according to a criterion derived from the Boltzmann distribution [16, 17]. In the implementation used here, exchanges are only attempted between replicas with neighboring temperatures in the series. Exchange attempts for replica i alternate between attempts to exchange with the $i-1$ replica and the $i+1$ replica. The exchanges are accepted or rejected based on an algorithm that ensures detailed balance, similar to the Metropolis criterion [18]. By this criterion, the probability of accepting an exchange is,

$$P_{acc} = \min\left[1, \exp\left(\frac{1}{k_B}\left(\frac{1}{T_i} - \frac{1}{T_j}\right)(\mathcal{V}(\vec{r}_i) - \mathcal{V}(\vec{r}_j))\right)\right] \tag{2.3}$$

where $\mathcal{V}$ is the potential energy, and $\vec{r}_i$ specifies the positions of the N particles in system i. A conformational exchange is accepted if this probability is greater than a random number between 0 and 1, which is taken from a uniform distribution. In a successful exchange, the coordinates of the particles of the two replicas are swapped. When the momenta of the particles are swapped, they are also scaled by a factor of $\sqrt{\frac{T_i}{T_{i+1}}}$ to generate a correct Maxwell distribution of velocities. The process of REMD is illustrated in the following pseudocode.

Algorithm 1: Algorithm for Replica-Exchange Molecular Dynamics

```
Function REMD (cycles c, replicas n, steps m)
    for c cycles do
        for $a \leftarrow 0$ to $n$ do
            perform $m$ steps of NVT MD;
        for neighboring pairs of replicas $\{i, i+1\}$ do
            choose random z $\in$ (0,1) ;
            $P_{acc} = \min\left[1, \exp\left(\frac{1}{k_B}\left(\frac{1}{T_i} - \frac{1}{T_{i+1}}\right)(\mathcal{V}(\vec{r}_i) - \mathcal{V}(\vec{r}_{i+1}))\right)\right]$ ;
            if $z < P_{acc}$ then
                $\vec{r}_i \leftrightarrow \vec{r}_{i+1}$ ;
                $\vec{p}_i \leftrightarrow \vec{p}_{i+1}$ ;
```

2.2.2 Cluster Analysis

Configurations in the REMD trajectory are grouped into clusters that correspond to distinct conformations. The lowest energy conformation will correspond to the cluster with the greatest number of configurations. The process of clustering conformations involves using some pattern proximity function to measure the similarity between pairs of conformations. This clustering algorithm groups these configurations according to this function [19].

In this work, the solute root mean square deviation (RMSD) metric is used to identify the highly probable conformations from the REMD trajectory. The RMSD provides a metric for the quality threshold of the similarity of two solute configurations. It is calculated from the Cartesian coordinates of the two configurations ${r_k}^{(i)}$ ${r_k}^{(j)}$ each having n atoms using [20],

$$d_{ij} = \left[\frac{1}{N} \sum_{k=1}^{n} \left| r_k^{(i)} - r_k^{(j)} \right|^2 \right]^{1/2} \tag{2.4}$$

The quality threshold clustering algorithm groups objects such that the diameter of a cluster does not exceed a set threshold diameter. The number of clusters (N) and the maximum diameter must be specified by the user prior to the clustering analysis. A candidate cluster is formed by selecting a frame from the trajectory (a conformation) as the centroid. The algorithm iterates through the rest of the configurations in the trajectory, and the conformation with the smallest RMSD with respect to the centroid is added to the cluster. Configurations are added to this cluster until there is no remaining configuration with an RMSD less than the threshold. The clustered configurations are removed from consideration for further clusters, and a new cluster is initiated. This process is repeated until N clusters have been generated.

2.3 Computational Work Flow

The first section describes a work flow that was developed to perform an explicitly-solvated conformational search of small drug molecules. In the second section, applications of the work flow are described, and the results are compared to gas phase and Generalized Born Implicit Solvent (GBIS) implementations.

Our method automatically performs conformational searches in the gas phase, implicit aqueous solvent, and explicit aqueous solvent for each solute structure. The work flow makes use of several open source programs, as illustrated in Figure 2.2. The conformation search work flow can be divided into 5 steps.

1. Generation of initial 3D molecular structure.
2. Solvation of solute (for explicit solvent method only).
3. Equilibration MD simulation.
4. REMD simulation.
5. Cluster analysis.

1. Structure Generation

The initial 3D structure is generated using the OBBuilder class of OpenBabel version 2.3.2. OpenBabel is a chemistry file translation program that is capable of converting between various file formats, but can also automatically generate 2D and 3D chemical structures and perform simple conformation searches [21]. Our work-flow uses OpenBabel to converts the SMILES string input, which is an ASCII string representation of a molecular structure, into an initial 3D structure that is saved in Protein Data Bank (pdb) format. OpenBabel supports many other chemical file formats, so alternative input formats can also be used. To generate a reasonable initial conformation, a conformation search is performed using the OBConformerSearch class of OpenBabel. This algorithm uses rotor keys, which are arrays of values specifying the possible rotations around all rotatable bonds [22]. Structures for each combination of rotor keys are generated and the potential energies for these conformations are calculated. The lowest energy structure for a rotor key is identified [23]. Once all possible conformations have been generated, the algorithm selects the one with the lowest energy. The Generalized Amber Force Field (GAFF) is used for all OpenBabel MM calculations [24]. Solvation effects are not included in this model.

One drawback of OpenBabel is that the current version can generate wrong stereoisomers for chiral centers in fused rings for some molecules. In these cases, the user should check the initial structure to ensure that the correct stereoisomers is modeled.

2. Solvation of Solute

The Antechamber utility of the Ambertools suite is used to generate the necessary topology (.rtf) and parameter (.prm) files of the solute [25]. This utility automatically detects the connectivity, atom types, and bond multiplicity of organic molecules and generates the parameter file and topology files based on the Generalized Amber Force Field (GAFF). The psfgen plugin of VMD is used to generate a Protein Structure File (PSF) for the molecule from the RTF file. For simulations with an explicit solvent, the Solvate plugin of VMD is used to add a 10 Å layer of water in each direction from the furthest atom from the origin in that direction. This creates a periodic unit cell that is sufficiently large so that solute-solute interactions and finite-size effects are small. For ionic molecules, the autoionize VMD plugin is used to add Na^+ or Cl^- ions such that the net charge of the simulation cell is zero.

3. Equilibration

For simulations with an explicit solvent, MD simulations are performed with NAMD to equilibrate the system prior to the conformational search. For the gas phase and GBIS models, a 1 ns MD simulation using a Langevin thermostat is performed. For the explicit solvent simulations, a 1 ns isothermal-isochoric (NVT) simulation is followed by a 1 ns isothermal-isobaric ensemble (NpT) simulation A Langevin thermostat and a Langevin piston barostat are used to regulate the temperature and pressure of the system, respectively.

To simplify visualization and analysis, the center of mass of the solute is restrained to remain at the center of the simulation cell using a weak harmonic restraining force. This restraint is imposed with the Colvar (Collective Variables) module of NAMD using a force constant of 5.0 kcal Å^{-2}.

4. Replica Exchange MD

Using the equilibrated system, a replica exchange MD simulation is performed to sample the configurational space of the system. A total of 24 replicas are simulated, with a range of temperatures between 298 and 500 K. The temperatures of the replicas are spaced according to a geometric series [26, 16]. A 1 ns equilibration followed by a 10 ns sampling simulation is performed for each replica. Configurations are saved and exchanges are attempted every 1000 time steps. The REMD simulations were performed at constant volume, which was the final volume of the NpT equilibration simulation.

5. Cluster Analysis

The trajectory of the lowest temperature replica is analyzed by clustering analysis to identify the most probable conformations. The positions of the solute atoms in each frame of the trajectory are rotated and translated to minimize the RMSD. The *cluster* routine of the *measure* module of VMD is used to identify highly-weighted conformations. This routine uses the quality threshold clustering algorithm, with the RMSD as the metric. An RMSD cutoff of 1.0 Å was used. In this work flow, 5 clusters are generated. The clusters are sorted in order of the largest to smallest numbers of configurations included, the first of which is the most important as it represents the most probable conformation for the lowest temperature replica. The configurations that are part of each cluster are saved to separate trajectory files. The conformation is defined by the set of configurations grouped into this trajectory file.

2.4 Implementation and Usage

The work flow is implemented in a Python script that calls external programs and processes the data from these programs. This script is responsible for handling user input and integrating the work flow into the a PBS-type queuing system. PBS is a distributed workload management system, which is responsible for queuing, scheduling, and monitoring the computational workload on a system [27]. The program is executed by the command,

```
python fluxionalize.py -p [number of processors, default is 2]
-n [name, default is  ‘‘test’’]
-l [location/directory, default is current working directory]
-c [number of clusters to save in {[}name{]}_out per instance, default is 1]
-i [input]
```

When the calculation has completed, the following files/directories will have been generated in the specified/default location:

[name]_out contains the conformation pdb files for each instance

[name].out the logfile from the queue containing all the runtime command line outputs

[name].tar.gz contains all the files used and generated by the work flow, compressed for space

OpenBabel is used to parse the molecular structure provided by the user and convert it to an initial 3D conformation, so any of the input formats supported by OpenBabel can be used. The examples presented here use SMILES (Simplified Molecular

Input Line Entry System) strings as the input. SMILES denotes chemical structure as ASCII-type strings. If using a SMILES string, the input for the fluxionalize.py script is in the form of -i '[SMILES string]'. For other files types, the input is in the form of: -i [file]. In this case, if no name is specified with the -n option, then the file name is used in its place.

2.4.1 Availability

The code and required source files are available freely from GitHub at `https://github.com/RowleyGroup/fluxionalize`.

2.5 Technical Details

The current version of this code uses OpenBabel 2.3.2 [21] and VMD 1.9.1 [28]. All MD and REMD simulations were performed using NAMD 2.10 [29]. Bonds containing hydrogen were constrained using the SHAKE algorithm [30]. Lennard-Jones interactions were truncated using a smoothed cutoff potential between 9 Å and 10 Å. A Langevin thermostat with a damping coefficient of 1 ps^{-1} was used. The simulation time step was 1 fs. Generalized born model simulations used a dielectric constant of 78.5 and an ion concentration of 0.2 M. For the simulations with an explicit solvent, water molecules were described using the TIP3P model [31]. The molecule and solvent were simulated under cubic periodic boundary conditions. The electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method with a 1 Å grid spacing [29]. Isothermal–isobaric MD simulations used a Nosé–Hoover Langevin piston barostat with a pressure of 101.325 kPa, a decay period of 100 fs, and an oscillation period of 2000 fs.

The potential energy terms for the solute were described using the General Amber

Force Field (GAFF) [24]. Atomic charges are assigned using the restrained electrostatic potential fit (RESP) charge fitting method [32], where the atomic charges were fit to the AM1-BCC model [33].

2.6 Examples

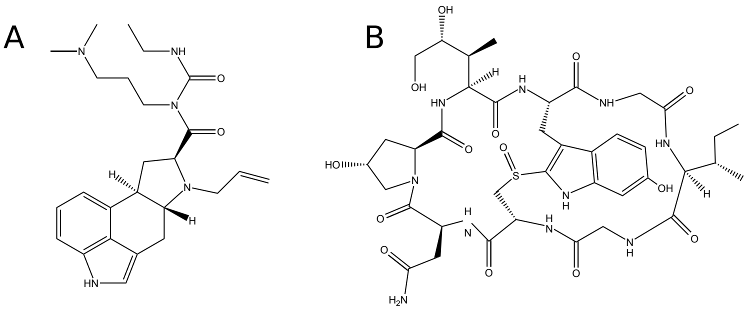

Figure 2.3: Chemical structures of molecules used to demonstrate conformation search work-flow. (a) Cabergoline and (b) α-Amanitin are mid-sized pharmaceuticals with significant conformational flexibility. The intramolecular and solute-solvent interactions result in complex conformation distributions.

To demonstrate the capabilities and performance of our method, conformation searches were performed on two drug molecules: α-amanitin and the neutral state of cabergoline (Figure 2.3) [34] [35]. α-Amanitin serves as a good example of the effectiveness of the work-flow. There are significant differences between the primary conformations in the gas phase, implicit solvent, and explicit solvent models. The most probable conformations derived from these models are overlaid in Figure 2.4. The gas phase structure is more compact than the explicit solvent structure, which is consistent with the tendency of gas phase molecules to form intramolecular interactions, while solution structures can extend to interact with the solvent. The implicit solvent model structure is more similar to the explicit solvent structure, but is still distinct from the explicit solvent structure. Figure 2.5 shows the four most probable conformations from the explicit solvent simulations. The clustering algorithm successfully categorized conformations with different configurations of the fused rings and orientations of the pendant chains.

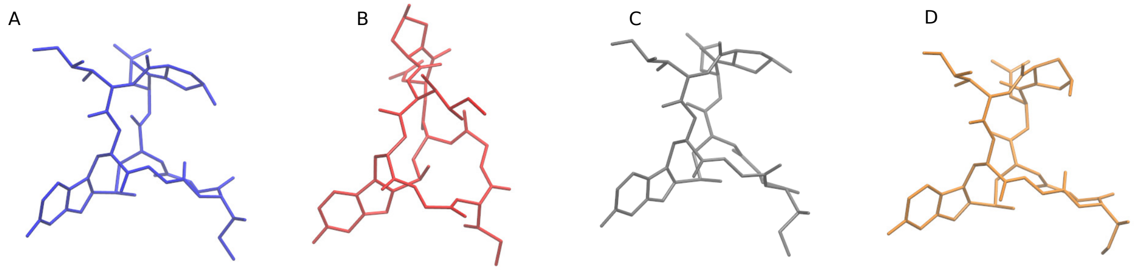

Figure 2.4: Comparison of the most probable explicitly solvated α-amanitin conformations where a) is the most probable, and b) is the second most probable, and so forth.

Cabergoline has a simpler chemical structure, containing no long chains and a more rigid ring structure. The most probable conformations with the explicit solvent (see Figure 2.6 (b)) are all quite similar; the RMSD values are under 0.98. Significant differences are apparent in the primary conformations of the explicit, GBIS, and gas phase simulations (see Figure 2.6 (a)). In particular, the configuration of the alkyl chains are sensitive to the effect of solvation. Generally, more rigid molecules will likely be less sensitive to solvation effects.

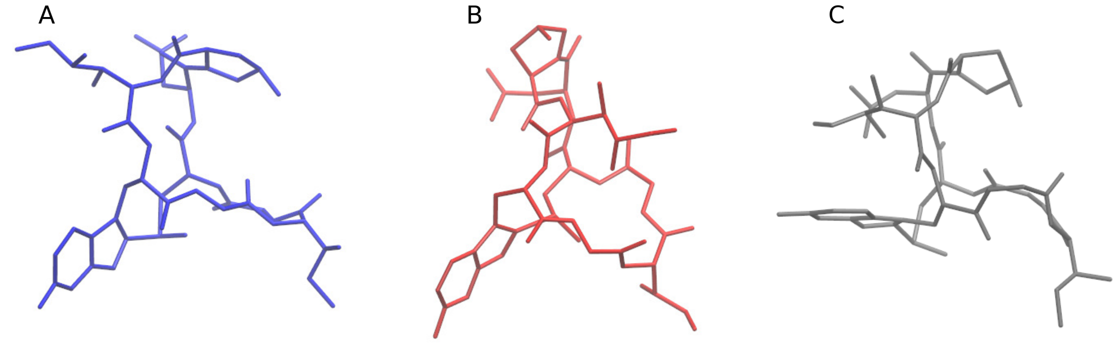

Figure 2.5: Most probable α-amanitin conformations. The explicitly solvated (a) and GBIS (c) conformations show the effect of the solvent, as compared to the more compact conformation in the gas phase (b).

Cabergoline contains two nitrogen centers that are formally chiral. Some conformation search algorithms have difficulty with type of moiety because the chirality of these centers can be switched by inversion of the nitrogen center. These inversion moves must be explicitly implemented into the structure generation algorithm of the method. Because the method presented here uses REMD, these inversions occur thermally, so conformations corresponding to these inverted configurations are identified automatically.

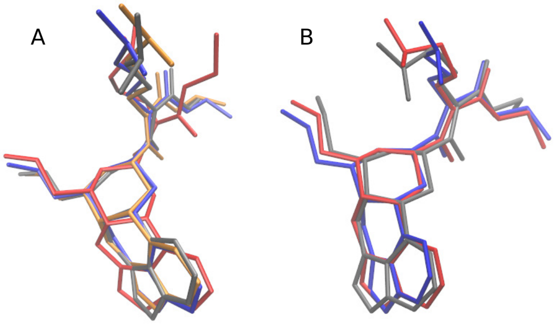

Figure 2.6: The lowest energy conformations of cabergoline calculated using the implicit and explicit solvent models. a) Most probable conformations, where the explicit solvent is blue, gas phase is red, and GBIS is grey. b) Most probable conformations calculated using explicit solvent models. In order of most to least probable: blue, red, grey, orange.

The computational cost of these simulations is moderate. The most computationally-intensive step is the REMD simulations in the explicit solvent. These simulations completed after approximately 80 hours when run on 72 2100 MHz AMD Opteron 6172 processors. Although the computational resources needed for REMD conformational searches are considerably greater than for the high-throughput heuristic methods that are currently used in high-throughput screening, these calculations are currently tractable. As the cost of these simulations scales well, this type of simulation could become routine when computational resources are widely available.

The average acceptance rates for the exchanges in the REMD simulations are collected in Table 2.1. The acceptance probabilities of the gas phase and implicit solvent models were high ($> 80\%$). REMD in an explicit solvent was found to be an efficient means to sample the configuration space, with acceptance probabilities of 27% and 31% for the simulations of α-amanitin and cabergoline, respectively. REMD can be inefficient for simulations in explicit solvents because the acceptance probability decreases with the heat capacity of the system, which is proportional to the number of atoms in the system [36].

Molecule	Simulation	Average Acceptance Rate
α-amanitin	Explicit	0.27
	Gas Phase	0.83
	GBIS	0.84
cabergoline	Explicit	0.31
	Gas Phase	0.88
	GBIS	0.88

Table 2.1: Acceptance rates of exchanges for replica exchange simulations, averaged over all replicas. The gas phase and GBIS simulations have very high acceptance rates, but the explicit solvent simulations have much lower acceptance

For large molecules that must be enclosed in a large solvent box, a prohibitively high number of replicas would be needed to ensure a sufficiently exchange probability. For small and medium sized molecules, like the ones used here, the simulation cell is small enough so that the exchange acceptance probability is > 0.25.

The initial coordinate (.pdb) files for the explicitly solvated structures, and for the gas phase and implicitly solvated structures can be found on the Github. Also available are the coordinate (.pdb) files for the four most probable explicitly solvated conformations (see Figure 2.4, and Figure 2.6 (b)), the coordinate files for the most probable conformations in gas phase and implicit solvent (see Figure 2.5 and Figure 2.6(a)), and the SMILES strings for α-amanitin and cabergoline.

2.7 Conclusions

In this chapter, we described a work-flow for performing conformational searches using REMD and clustering analysis for molecules in the gas phase, implicit solvents, and explicit solvents. The work-flow consists of five primary steps: generation of a 3D structure, solvation of the solute (for the explicit solvent method), an equilibration MD simulation, a REMD simulation, and cluster analysis. This method is implemented in Python scripting by integrating several open source packages (i.e., OpenBabel, VMD, and NAMD). The work-flow makes use of the greater conformation sampling achieved by REMD, and then performs cluster analysis to find the most probable conformations sampled in the trajectory. Two drug molecules were used as examples of the work-flow, which show significant differences between conformations in the gas phase, implicit solvent, and explicit solvent. This work-flow has the potential to be applicable to many fields such as drug design, cheminformatics, and molecular structure studies.

Bibliography

[1] Gordon Crippen and Timothy F. Havel. *Distance Geometry and Molecular Conformation.* John Wiley and Sons, New York, 1988.

[2] R. S. Struthers, J. Rivier, and A. T. Hagler. Molecular Dynamics and Minimum Energy Conformations of GnRH and Analogs: A Methodology for Computer-aided Drug Design. *Annals of the New York Academy of Sciences*, 439(1):81–96, 1985.

[3] Robert A Copeland. Conformational adaptation in drug–target interactions and residence time. *Future Medicinal Chemistry*, 3(12):1491–1501, 2011.

[4] Markus Christen and Wilfred F. van Gunsteren. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *Journal of Computational Chemistry*, 29(2):157–166, 2008.

[5] Wilfred F. van Gunsteren, Dirk Bakowies, Riccardo Baron, Indira Chandrasekhar, Markus Christen, Xavier Daura, Peter Gee, Daan P. Geerke, Alice Glttli, Philippe H. Hnenberger, Mika A. Kastenholz, Chris Oostenbrink, Merijn Schenk, Daniel Trzesniak, Nico F. A. van der Vegt, and Haibo B. Yu. Biomolecular Modeling: Goals, Problems, Perspectives. *Angewandte Chemie International Edition*, 45(25):4064–4092, 2006.

[6] Ramu Anandakrishnan, Aleksander Drozdetski, Ross C. Walker, and Alexey V. Onufriev. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophysical Journal*, 108(5):1153–1164, 2015.

[7] M. Bhandarkar, A. Bhatele, E. Bohm, R. Brunner, F. Buelens, C. Chipot,

A. Dalke, S. Dixit, G. Fiorin, P. Freddolino, P. Grayson, J. Gullingsrud, A. Gursoy, D. Hardy, C. Harrison, J. Hnin, W. Humphrey, D. Hurwitz, N. Krawetz, S. Kumar, D. Kunzman, J. Lai, C. Lee, R. McGreevy, C. Mei, M. Nelson, J. Phillips, O. Sarood, A. Shinozaki, D. Tanner, D. Wells, G. Zheng, and F. Zhu. *NAMD User's Guide*. University of Illinois and Beckman Institute, 2015.

[8] Yoshitake Sakae, Tomoyuki Hiroyasu, Mitsunori Miki, Katsuya Ishii, and Yuko Okamoto. Combination of genetic crossover and replica-exchange method for conformational search of protein systems. *arXiv:1505.05874 [cond-mat, physics:physics, q-bio]*, 2015. arXiv: 1505.05874.

[9] Adriana Supady, Volker Blum, and Carsten Baldauf. First-Principles Molecular Structure Search with a Genetic Algorithm. *Journal of Chemical Information and Modeling*, 55(11):2338–2348, 2015.

[10] Ekaterina I. Izgorodina, Ching Yeh Lin, and Michelle L. Coote. Energy-directed tree search: an efficient systematic algorithm for finding the lowest energy conformation of molecules. *Physical Chemistry Chemical Physics*, 9(20):2507–2516, 2007.

[11] T. J. Brunette and Oliver Brock. Guiding conformation space search with an all-atom energy potential. *Proteins*, 73(4):958–972, 2008.

[12] Daniel Cappel, Steven L. Dixon, Woody Sherman, and Jianxin Duan. Exploring conformational search protocols for ligand-based virtual screening and 3-D QSAR modeling. *Journal of Computer-Aided Molecular Design*, 29(2):165–182, 2014.

[13] Roberto Vera Yasset Perez-Riverol. A Parallel Systematic-Monte Carlo Algorithm for Exploring Conformational Space. *Current Topics in Medicinal Chemistry*, 12(16), 2012.

[14] Justin L. MacCallum, Alberto Perez, and Ken A. Dill. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22):6985–6990, 2015.

[15] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1–2):141–151, 1999.

[16] David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[17] Ayori Mitsutake and Yuko Okamoto. Replica-exchange simulated tempering method for simulations of frustrated systems. *Chemical Physics Letters*, 332(12):131–138, 2000.

[18] Daan Frenkel and Berend Smit. Chapter 14 - accelerating monte carlo sampling. In Daan Frenkel and Berend Smit, editors, *Understanding Molecular Simulation (Second Edition)*, pages 389–408. Academic Press, San Diego, second edition edition, 2002.

[19] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[20] Oren M. Becker, Alexander D. MacKerell Jr., Benoit Roux, and Masakatsu Watanabe, editors. *Computational Biochemistry and Biophysics*. Marcel Dekker, Inc., New York, 2001.

[21] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011.

[22] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open babel: Conformer searching. http://openbabel.org/dev-api/group__conformer.shtml, 2012. Accessed:25-01-2016.

[23] T. Vandermeersch. forcefield.cpp. http://openbabel.sourcearchive.com/documentation/2.3.0plus-pdfsg-2ubuntu1/forcefield_8cpp_source.html, 2006. Accessed:25-01-2016.

[24] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[25] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260, 2006.

[26] David A. Kofke. Erratum: On the acceptance probability of replica-exchange Monte Carlo trials [J. Chem. Phys. 117, 6911 (2002)]. *The Journal of Chemical Physics*, 120(22):10852–10852, 2004.

[27] Anne Urban. *PBS Professional User's Guide.* Altair Engineering, Inc., 2010. http://www.pbsgridworks.jp/%28S%28gafuzx45nni4lyydiywwe345%29%29/documentation/support/PBSProUserGuide10.4.pdf.

[28] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[29] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and

Klaus Schulten. Scalable Molecular Dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[30] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman 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, 1977.

[31] William L. Jorgensen. Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *Journal of the American Chemical Society*, 103(2):335–340, 1981.

[32] Junmei Wang, Piotr Cieplak, and Peter A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000.

[33] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.

[34] David A. Bushnell, Patrick Cramer, and Roger D. Kornberg. Structural basis of transcription: -AmanitinRNA polymerase II cocrystal at 2.8 resolution. *Proceedings of the National Academy of Sciences*, 99(3):1218–1222, 2002.

[35] Najam A. Sharif, Marsha A. McLaughlin, Curtis R. Kelly, Parvaneh Katoli, Colene Drace, Shahid Husain, Craig Crosson, Carol Toris, Gui-Lin Zhan, and Carl Camras. Cabergoline: Pharmacology, ocular hypotensive studies in multiple

species, and aqueous humor dynamic modulation in the Cynomolgus monkey eyes. *Experimental Eye Research*, 88(3):386–397, 2009.

[36] Martin Lingenheil, Robert Denschlag, Gerald Mathias, and Paul Tavan. Efficiency of exchange schemes in replica exchange. *Chemical Physics Letters*, 478(1–3):80–84, 2009.

Chapter 3

Generalized Langevin Methods for the Calculation of Diffusion Coefficients

3.1 Introduction

The rate of diffusion of a solute is fundamental to biochemical transport processes like protein-ligand binding and membrane permeation. The diffusivity of a solute is particularly significant to the permeation of solutes through lipid bilayer membranes. In the inhomogeneous solubility diffusion model, the diffusion coefficient of a solute can be estimated from the potential of mean force ($w(z)$) and the diffusion coefficients ($D(z)$) of the solute as a function of its position along the transmembrane axis (z) [1, 2, 3, 4, 5]. In this model, the permeability coefficient (P_m) can be expressed as an integral over an interval of z that spans the bilayer.

$$\frac{1}{P_m} = \int_{z_1}^{z_2} \frac{e^{w(z)/k_B T}}{D(z)} \mathrm{d}z \tag{3.1}$$

There are several mature methods for calculating $w(z)$ using molecular simulations [6, 4], but the calculation of the $D(z)$ profile has received less attention. The Einstein [7] or Kubo [8] relations can be used to calculate the diffusion coefficient of a solute in a homogeneous solution by analysis of a molecular dynamics (MD) trajectory, but these methods have limited applicability for inhomogeneous systems like a bilayer. In these systems, the variation of the solute's diffusivity is large because the frictional environment varies dramatically as the solute moves from bulk water, through the interface, and into the membrane interior.

The fluctuation-dissipation equation provides one method to calculate the diffusion coefficient from the autocorrelation of the force exerted on a solute constrained at a point along the reaction coordinate. This method requires the imposition of a constraint on the z-position of the solute and the calculation of the force autocorrelation function, which can converge slowly. More elaborate statistical techniques have also been developed to interpret diffusion coefficients from MD simulations [9, 10, 11, 12], although implementing and applying these methods can be an involved process.

The Generalized Langevin Equation (GLE) provides distinct methods for calculating diffusion coefficients. If a degree of freedom of a system is restrained using a harmonic potential, its motion can be described by the GLE solution for a harmonic oscillator in a frictional bath. The diffusion coefficient of the solute along this coordinate can then be calculated by analysis of this trajectory. Roux and coworkers developed techniques to calculate these properties from the velocity autocorrelation function (VACF) [13, 14, 15]. Hummer later derived a simplified expression to calculate $D(z)$ from the position autocorrelation function (PACF) [9]. Although the PACF-based method has been used to calculate $D(z)$ in several membrane permeation studies, Lee et al. showed that there are some practical issues associated with this method [4].

In this paper, we present a comparison of these GLE-based methods for calculating the diffusivity of small molecule solutes in various liquids and with various simulation conditions. A general, automatic method for the calculation of the diffusion coefficient from the VACF is developed. The diffusivity profile of a water molecule across a lipid bilayer is calculated using the PACF and VACF methods.

3.2 Theory

3.2.1 GLE Methods for Calculating Diffusion Coefficients

The GLE of a harmonic oscillator takes the form,

$$m\ddot{z}(t) = -kz(t) - \int_0^t \dot{z}(\tau)\ \zeta(t-\tau)d\tau + R(t) \tag{3.2}$$

Here, k is the spring constant of the oscillator, $\zeta(t)$ is the dynamic friction kernel (or memory kernel), and $R(t)$ is the random force.

The effect of the balance of the system (i.e., the solution and the bilayer) on the oscillator is introduced through the friction and the random force terms. The analytical solutions for the PACF ($C_z(t)$) and VACF ($C_v(t)$) of a harmonic oscillator in a dissipative bath are [16],

$$C_z(t) = \mathrm{var}(z) e^{-\gamma(\bar{\omega})t/2\mu} \left[\cos(\Omega t) + \frac{\gamma(\bar{\omega})}{2\mu\Omega} \sin(\Omega t) \right] \tag{3.3}$$

$$C_v(t) = \mathrm{var}(\dot{z}) e^{-\gamma(\bar{\omega})t/2\mu} \left[\cos(\Omega t) - \frac{\gamma(\bar{\omega})}{2\mu\Omega} \sin(\Omega t) \right] \tag{3.4}$$

Here, γ is the friction coefficient, $\bar{\omega}$ is the renormalized frequency of the oscillator, μ is the reduced mass of the oscillator, and $\mathrm{var}(z)$ is the variance of z.

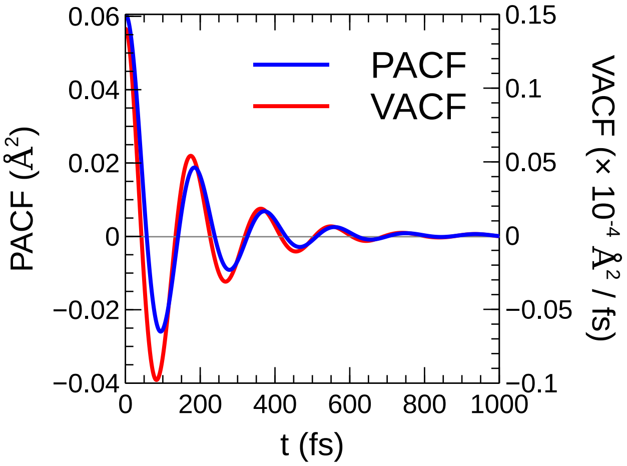

Figure 3.1: Position autocorrelation and velocity autocorrelation functions of a H_2O molecule in liquid hexane restrained with a harmonic potential with a spring constant of k = 10 kcal mol^{-1} Å^{-2}. For this system, both functions are exponentially-decaying oscillatory functions.

From this equation, the autocorrelation functions of a harmonically-restrained solute in a condensed phase are expected to be damped oscillatory functions, with the rate of decay depending on the friction imposed on the restrained degree of freedom. This form is apparent in the PACF and VACF of a harmonically-restrained water molecule in liquid hexane (Figure 3.1), although the oscillatory nature of the ACFs are not always apparent if the rate of decay is high.

These autocorrelation functions provide a connection between the dynamics of a restrained solute and the friction it experiences from the solvent. Extending the work of Straub and Berne [17], Roux and coworkers derived an expression for the diffusion coefficient of the solute from the GLE of a harmonic oscillator [13, 14, 15],

$$D(z_i = \langle z \rangle_i) = \lim_{s \to 0} \frac{-\hat{C}_v(s; z_i) \langle \delta z^2 \rangle_i \langle \dot{z}^2 \rangle_i}{\hat{C}_v(s; z_i) \left[s \langle \delta z^2 \rangle_i + \langle \dot{z} \rangle_i / s \right] - \langle \delta z^2 \rangle_i \langle \dot{z}^2 \rangle_i} \tag{3.5}$$

$\langle z^2 \rangle$ and $\langle \dot{z}^2 \rangle$ are the variances of the position and velocity of the oscillator, respectively. $\hat{C}_v$ is the Laplace transform of the velocity autocorrelation function. The diffusion coefficient is the limit of this equation as $s \to 0$, where s is the coefficient of the Laplace transform. As a practical matter, this limit cannot be taken directly because of a singularity at $s = 0$, so the limit must be extrapolated from $D(s)$ in the range where the function is well-behaved.

Hummer derived a simpler form of this equation that uses the PACF instead of the VACF [9],

$$D(z_i = \langle z \rangle_i) = \frac{\mathrm{var}(z)^2}{\int_0^\infty C_z(t)\,\mathrm{d}t}. \tag{3.6}$$

An advantage of this form is that is it not necessary to store the velocity time series, compute a Laplace transform, or perform the extrapolation step; the diffusion coefficient can be calculated in a straightforward way from the variance and PACF alone. The derivation of these expressions from the GLE is presented in Appendix A.

3.2.2 Practical Calculation of Correlation Functions

The presented GLE-based methods for calculating the diffusion coefficients both require the calculation of a correlation function for the motion of the solute when the solute is restrained to some position along the transmembrane coordinate (z) by a harmonic potential,

$$\mathcal{V}(r) = \frac{1}{2}k(z - z_0)^2 \tag{3.7}$$

where k is the spring constant of the restraint, and z_0 is the reference position. The diffusion coefficient can be calculated at different positions along the coordinate by selecting different reference positions for the restraint.

Beginning from an equilibrium configuration for the membrane system, a time series of the z-coordinate of the center of mass of the solute is collected by performing a MD simulation. The simulation must be sufficiently long to allow the calculation of correlation functions that are well-converged for the relaxation time of the system. This typically requires a simulation that is at least 1 ns in duration. These simulations are performed with restraint reference positions at intervals that span the membrane (e.g., $z_0 = -40, -39, ..., 0, ..., 39, 40$ Å).

The time series of the position, z, or the velocity, $\dot{z}$, along the z-coordinate can be used to calculate the position or velocity autocorrelation functions, respectively. The PACF is denoted as $C_z(t)$ while the VACF is denoted as $C_v(t)$. The correlation functions for a time series with regularly spaced intervals can be calculated most directly by a summation over the trajectory [18],

$$C_z(t) = \langle \delta z(0)\delta z(t)\rangle = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}} \delta z(i)\delta z(t + i) \tag{3.8}$$

where $\delta z(t) = z(t) - \langle z \rangle$.

Because these correlation functions converge to zero after the relaxation time of the system, it should be sufficient to calculate the correlation function over a short interval (e.g., 0 – 5 ps). For particularly long time series, the correlation functions could be calculated more efficiently using Fourier transforms [19, 20].

These GLE-based methods are attractive for calculating transmembrane diffusivity because many biomolecular simulation codes (e.g., NAMD and GROMACS) natively support the imposition of a harmonic restraint like Eq. 3.7 and for the time series generated by this simulation to be saved to disk. This procedure is identical to that used to perform an umbrella sampling simulation to calculate the potential of mean force for the permeation of a solute. In principle, both properties could be calculated from the same data, although in practice there are some issues regarding this practice (vide infra).

3.2.3 Practical Calculation of Diffusion from VACF

Calculation of $D(z)$ using the VACF-based method requires a procedure to find the limit as $s \rightarrow 0$ in Eq. 3.5. The limit cannot be taken directly due to a singularity at $s = 0$. There are two additional singularities in $D(s)$, and the function is only well-behaved in the interval between them, so the value of $D(s = 0)$ must be extrapolated from the range between the 2nd and 3rd singularities. To facilitate routine calculation of $D(z)$ using the VACF method, we developed an algorithm to extrapolate $D(z, s = 0)$ from $D(z, s)$. The full details of the algorithm are described in Appendix B.

3.3 Technical Details

Molecular dynamics simulations of the solutes in the homogeneous solvents were performed using NAMD 2.10 [21]. The SHAKE algorithm was used to constrain bonds containing hydrogen [22]. Lennard-Jones interactions were truncated at 12 Å using a smoothed cutoff potential. A Langevin thermostat at 298.15 K with a damping coefficient of 1 ps^{-1} was used for equilibration. The time series for the calculation of the diffusion coefficients using the GLE-based methods were performed under (microcanonical) NVE conditions. The simulation time step was 2 fs. The electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method with a $32 \times\ 32 \times\ 32$ grid [23]. Diffusion coefficients were calculated from the average of three trajectories. For each simulation, a 1 ns MD simulation was performed under isothermal-isobaric (NpT) conditions to equilibrate the system before a 10 ns production simulation under NVE conditions. In order to calculate the diffusion coefficient using the GLE-based methods, a harmonic restraint along the z axis was imposed on the center of mass of the solute with a spring constant of 10 kcal mol^{-1} Å^{-2}.

The force field of Fischer and Lago was used in simulations of O_2 [24]. The TIP3P model was used in the simulations involving water [25]. The CHARMM General Force Field was used to describe the aliphatic solvents [26].

A lipid bilayer system was used comprising 64 DPPC lipids arranged into a symmetric bilayer, with the bilayer running along the xy plane. The bilayer was surrounded by a solvent layer containing 4551 water molecules. The dimensions of the simulation cell were roughly 44 Å × 44 Å × 114 Å.

A 40 ns steered MD simulation was performed to generate the initial configurations for the restrained simulations. A water molecule from solution was pulled along the z-axis over the course of this simulation. Configurations were selected at 1 Å separations relative to the center of mass of the bilayer, spanning a -40 Å to 40 Å interval. The permeating water molecule was restrained to these reference positions using a harmonic restraint with a 10 kcal mol^{-1} Å^{-2} spring constant. 10 ns MD simulations were performed to equilibrate these restrained simulations. The time series used to calculate the $D(z)$ profiles were calculated from 2 ns MD simulations in the microcanonical ensemble. 2 ns equilibration simulations were performed to generate distinct starting points for the successive series. The profiles presented in Figure 3.6 are calculated from the average of the each of these 3 simulations, symmetrized about the center of the bilayer. The temperature of the system was 314 K.

3.3.1 Implementation of Diffusivity Calculations

The diffusivity calculations using the PACF and VACF GLE methods from MD time series has been implemented in our code, ACFCalculator, which is freely available under the GNU Public License [27]. This program can read time series files generated from CHARMM, NAMD, and GROMACS.

3.4 Results and Discussion

3.4.1 Validation of GLE Methods with Homogeneous Liquids

The diffusivities of H_2O and O_2 molecules in bulk liquids of water (TIP3P model), pentane, and hexane were calculated using the GLE methods (Figure 3.2). These diffusivities were also calculated using the Einstein equation,

$$D = \frac{1}{6t}\langle|r(t) - r(0)|\rangle. \tag{3.9}$$

The diffusivities calculated using the GLE methods can be compared to the diffusivities calculated using the Einstein equation, which provides an independent and rigorous comparison. The diffusivities calculated using these three methods are compared in Figure 3.3.

Generally, all three methods yield comparable values, although the VACF method

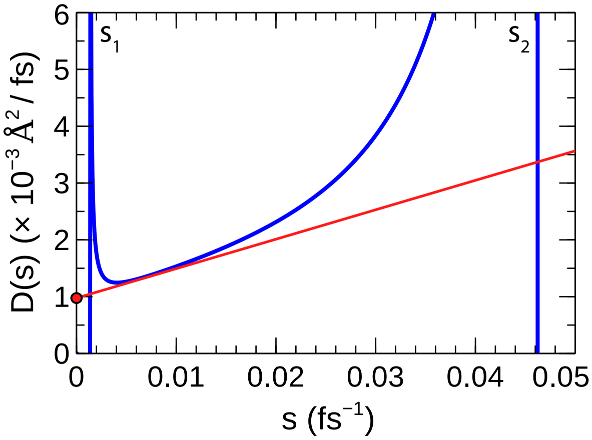

Figure 3.2: D(s) curve (Eq. 3.5) for a harmonically-restrained O_2 molecule in liquid hexane ($k = 10$ kcal mol^{-1} Å^{-2}). The diffusion coefficient ($D(s = 0) = 1.04 \times 10^{-3}$ Å^2 fs^{-1}) is estimated by linearly extrapolating $D(s)$ from the region of lowest curvature between singularities s_1 and s_2.

is generally in closer agreement with the Einstein method, and has the smallest standard deviation of all three methods. The GLE methods appear to be effective for these solutes for both aqueous and paraffinic solutions, which are representative of the environments in a lipid-bilayer.

Effect of Thermostat Friction of Calculated Diffusivity

Molecular dynamics simulations are often performed using stochastic thermostats. For example, a Langevin thermostat can be used to sample the canonical (NVT) ensemble of the system by introducing artificial frictional and random forces on the dynamics of the molecules. These forces change the dynamics of the molecules, so transport properties like diffusion will be affected. To assess the effect of this, the self-diffusion coefficient of TIP3P-model water was calculated using the GLE methods from simulations with a Langevin thermostat and frictional coefficients in the range commonly used in biomolecular MD simulations ($\gamma = 1$ ps^{-1} – 10 ps^{-1}). We compared this to the diffusion coefficients calculated from a NVE MD simulation, which lacks

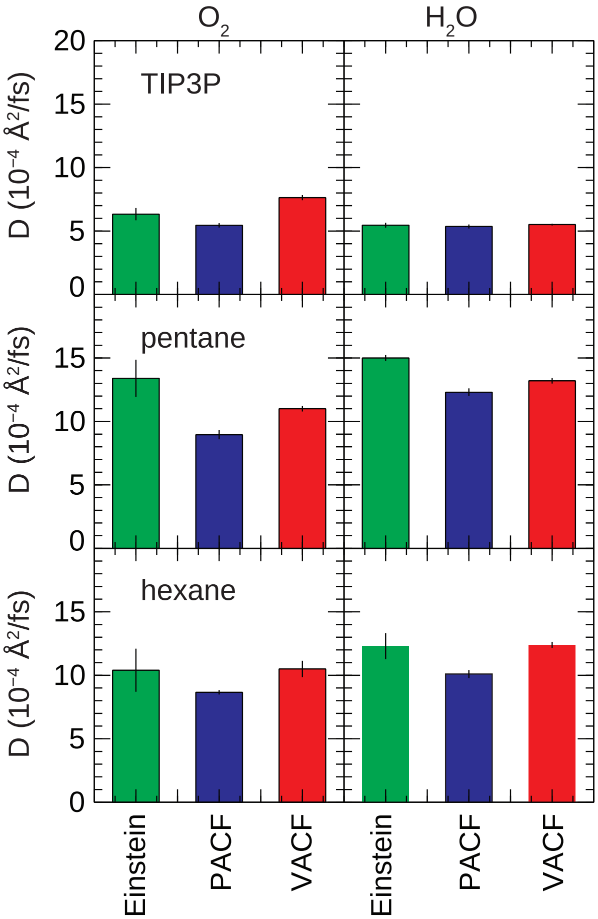

Figure 3.3: The diffusivity of O_2 (left) and H_2O (right) in liquid water (TIP3P model), pentane, and hexane calculated the PACF and VACF GLE methods. For reference the diffusivity calculated from the RMSD using the Einstein equation is also shown.

any artificial thermostat forces, and from an NVT MD simulation using the Lowe–Andersen thermostat, which is a stochastic thermostat that has a smaller effect on diffusion coefficients. The diffusivity of TIP3P-model water calculated using these methods is presented in Figure 3.4.

Both the PACF and VACF methods show a decline in the diffusion coefficient as the thermostat frequency is increased. This can be attributed to the damping of the dynamics of the oscillating solute due to the frictional force imposed by the thermostat. The PACF method is more sensitive to this effect; the diffusion coefficient calculated using the PACF method drops to 2.0×10^{-4} Å^2/fs when $\gamma = 10$ ps^{-1}, but it only drops to 4.0×10^{-4} Å^2/fs when using the VACF method. This difference is due to the application of the Laplace transform in the VACF method, which reduces the influence of the correlation function at long times. The effect of thermostat frequency has a larger effect on the correlation functions at longer times because the dynamics have been subjected to these forces for a longer period.

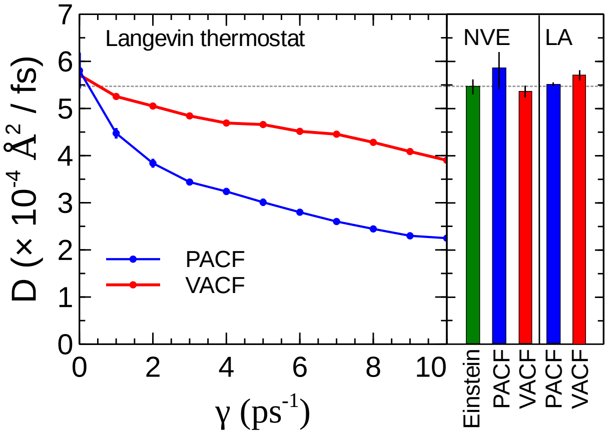

Figure 3.4: The effect of the Langevin thermostat frictional coefficients (γ) on the diffusion coefficients of TIP3P-model water (left). The diffusion coefficients calculated with both the PACF and VACF GLE methods are decreased as the friction of the Langevin thermostat are increased. The reference values for a simulation performed under NVE conditions (i.e., no thermostat) and those performed using a Lowe–Andersen thermostat (LA) are shown for comparison (right). The dotted line indicates the diffusivity calculated using the Einstein equation and the NVE simulation.

Simulations to calculate the potential of mean force must use a thermostat in order to sample the correct ensemble. If the same simulation is to be used to calculate $D(z)$ as is used to calculate $w(z)$, it is important to use a small frictional coefficient, or to use a thermostat that does not have a large effect on diffusivity (e.g., Nosé–Hoover or Lowe–Andersen).

Effect of Restraint Spring Constant on Calculated Diffusivity

Several factors affect the choice of the spring constant to restrain the solute. The assumption underlying these GLE-based methods is that the dynamics of the solute can be described as a harmonic oscillator in a frictional bath. In heterogeneous environments like a lipid bilayer, the underlying free energy surface can be rough, so the harmonic restraining force should be strong enough to dominate over these forces. Likewise, to calculate a diffusion coefficient at an arbitrary point along this coordinate, the restraining force must be sufficiently large so that the average position along the coordinate is close to the reference position (e.g. $\langle z \rangle_i \approx z_{0,i}$.

To test for systematic errors that could result from the choice of the spring constant, simulations were performed to calculate the diffusivity of water using the GLE-based methods with spring constants ranging from 1 kcal mol^{-1} Å^{-2} to 50 kcal mol^{-1} Å^{-2}(Figure 3.5 a). The PACF method shows some variation with the spring constant, but there is no systematic trend, and all the simulations performed with the various spring constants give a prediction in reasonable agreement with the diffusivity calculated using the Einstein method.

The VACF method is more sensitive to the spring constant. There is a roughly linear decrease in the calculated diffusivity when the spring constant is increased. The origin of this trend is apparent Figure 3.5 b. The variance of the simulation and the Laplace transform of the VACF both affect the locations of singularities and curvature of $D(s)$. Simulations performed with larger spring constants have greater curvature in $D(s)$, and an extrapolation range that is located at a larger value of s, making the linear extrapolation technique used here less reliable. As the shape of $D(s)$ is sensitive to both the spring constant restraining the solute and the dynamic friction function of the surroundings, this method should be used with caution.

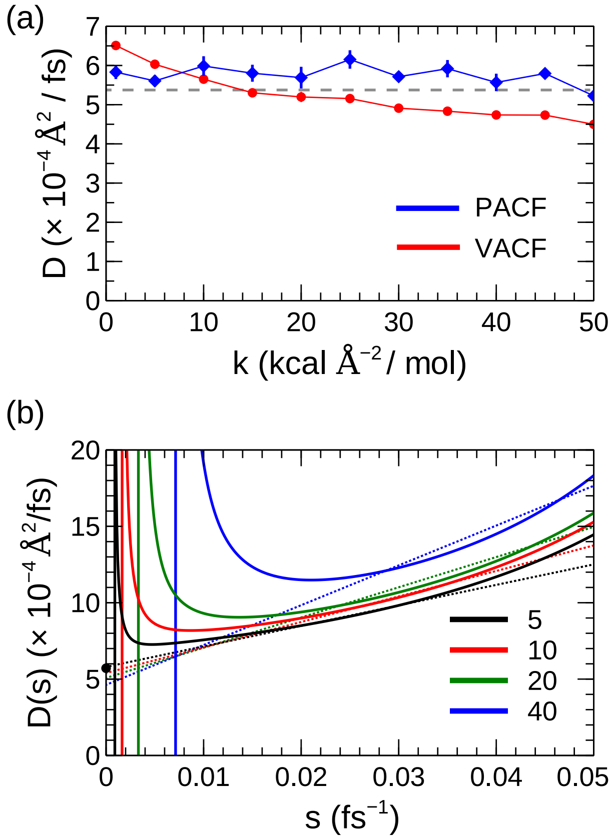

Figure 3.5: (a) The diffusivity of TIP3P-model water calculated using the GLE methods with various spring constants for the harmonic restraining force (k). The diffusivity calculated using the Einstein method is indicated by the dashed gray line. (b) The $D(s)$ profiles calculated using the VACF method (Eq. 3.5) corresponding to select values of the spring constant, k. The dotted lines indicate the extrapolation. The reference diffusion coefficient is indicated by a black dot on the y-axis.

3.4.2 Transmembrane Diffusivity Profiles

To compare these methods for calculating the diffusivity of solutes permeating through a lipid bilayer, simulations were performed where a water molecule was restrained at positions that span the bilayer in the interval $z = [-40$ Å, 40Å$]$ along the transmembrane axis. The solute was restrained with a harmonic spring constant of $k = 10$ kcal mol^{-1} Å^{-2}. These profiles are presented in Figure 3.6.

The PACF and VACF methods predict similar diffusion coefficients in solution

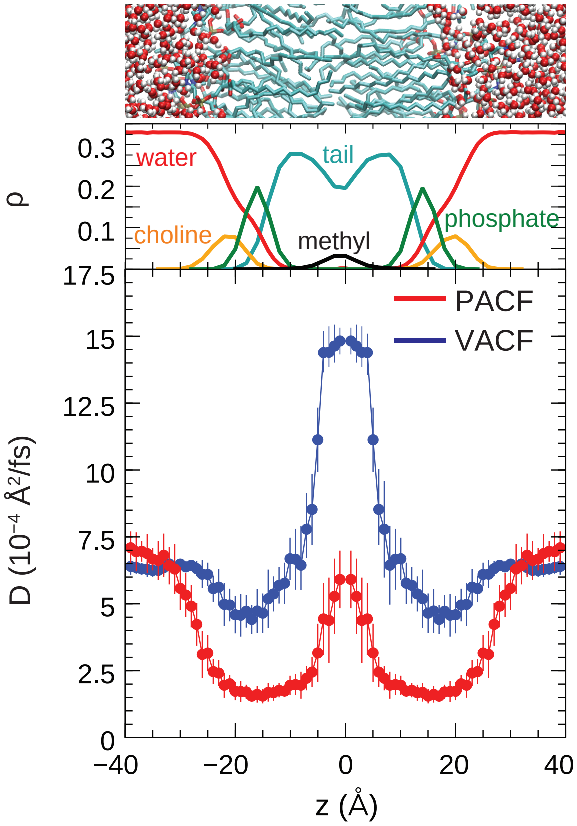

Figure 3.6: Diffusion coefficients of a water molecule permeating through a DPPC lipid bilayer calculated using the PACF and VACF GLE methods. The upper panel shows the density contribution of the various membrane components in the bilayer. The VACF method predicts systematically higher rates of diffusion inside the bilayer. The profiles are symmetrized about the center of the bilayer. Error bars were calculated from the standard deviations of the three independent simulations.

($|z| > 35$ Å), but differ inside the bilayer. The PACF method predicts low diffusivity in the upper regions of the lipid tails (5 Å$< |z| < 20$ Å). Both methods predict relatively high diffusivities in the disordered region at the centre of the bilayer (5 Å$< |z| < 20$ Å), although the VACF method predicts much higher diffusivities (1.5×10^{-3} Å^2/fs).

3.4.3 Slow Decay of the PACF

The difference between the PACF and VACF diffusion profiles can be rationalized by examining the correlation functions used to calculate the diffusivities. Lee et al.

showed that that the PACF of a solute harmonically restrained inside a lipid bilayer can decay far slower than those in bulk liquids [4]. In some cases, the PACF can have a significantly non-zero value even after 5 ps. The denominator of Eq. 3.6 should formally be integrated until $C_z(t)$ converges to zero, although in practice, it is typically only computed for some predefined interval. This causes calculated diffusion coefficient to become sensitive to the bounds chosen for the integration. The diffusion coefficients calculated using Eq. 3.6 can be significantly decreased for bilayer depths where the PACF decays slowly.

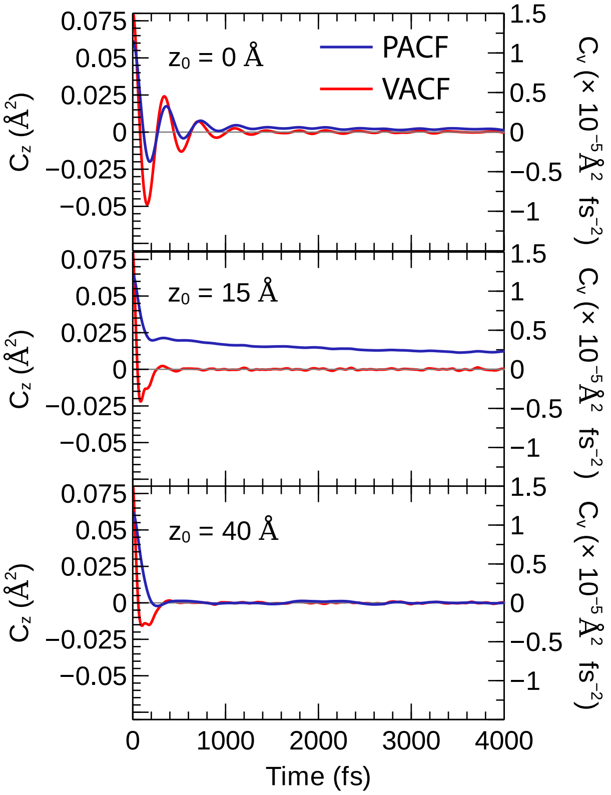

Figure 3.7: Position autocorrelation functions of a water molecule restrained at different depths in a DPPC bilayer. The PACF converges to zero in less than 1000 fs when the solute is in the bulk solution ($z_0 = 40$ Å), but does not converge to zero even after 4 ps when the solute is immersed in the bilayer ($z_0 = 0$ Å and $z_0 = 15$ Å) The red curves shows the VACF. The VACF converges to zero for all solute depths.

This slow decay is apparent in the PACFs of a water molecule restrained at different reference positions along the z-axis of a DPPC lipid bilayer system (Figure 3.7). When the solute is in the bulk water above the bilayer ($z_0 = 40$ Å), the PACF converges to zero in less than 1000 fs. The PACF holds a significant non-zero value when the solute is immersed in the bilayer ($z_0 = 15$ Å).

The PACF indicates that there is a degree of correlation in the position of the solute after an elapsed time. In most bulk liquids, the frictional forces are sufficiently strong so that the position of the restrained solute is no longer correlated after 1–2 ps (i.e., its current position is independent of its previous). The long tails on the PACFs for simulations inside the bilayer indicate that the position of the solute can have significant correlations for much longer intervals than in bulk liquids.

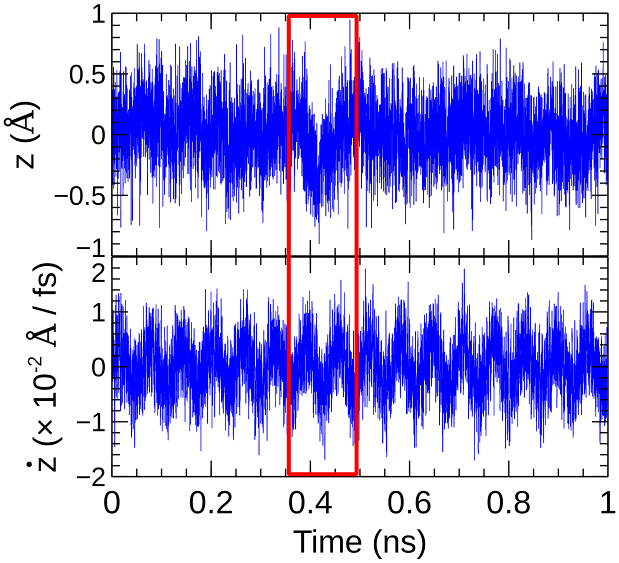

Figure 3.8: Time series of a water molecule restrained to oscillate around the center of a DPPC bilayer by a harmonic potential ($k = 10$ kcal mol^{-1} Å^{-2}). Fluctuations in the position occur with a life time on the order of 200 ps during this 1 ns simulation.

These long-timescale correlations are apparent in the time series of a simulation of a water molecule restrained at the center of the bilayer (Figure 3.8). Analysis of the time series shows fluctuations that can extend over 100 ps. These long-timescale fluctuations can be attributed to slow rearrangements of the bilayer that hinder the oscillation of the solute and the changes in the hydration state of hydrogen-bonding solutes. This type of long-timescale correlation of the position of a solute restrained inside lipid bilayers has been previously noted by Neale et al. as a challenge in calculating the PMF of solute permeation [28, 29, 6].

The VACF does not exhibit the same slow decay. The VACF converges to zero in less than 2000 fs for all reference positions, including the solution, interface, and bilayer center. In the lipid tail region ($z_0 = 0$ Å), the PACF only decays to 10% of its initial value after 4 ps, while the VACF decays to a value near 0 after only 500 fs. Physically, the VACF does not show long term correlations because even if a long-timescale fluctuation occurs in the position of the solute, the oscillations in the velocity of the solute are not strongly affected. This suggests that the VACF-based method to calculate $D(z)$ could resolve the issues related to the slow decay that affect the PACF-based method.

3.5 Conclusions

Two methods based on the Generalized Langevin Equation were examined for their utility in calculating transmembrane diffusivity profiles. The first method uses the position autocorrelation function (PACF) of a solute harmonically restrained at a chosen bilayer depth, while the second method uses the velocity autocorrelation function (VACF) of a solute also harmonically restrained. The VACF method requires the diffusivity to be extrapolated from an equation involving the Laplace transform of the VACF, so an algorithm was developed to calculate solute diffusivity automatically using this method.

When tested on simulations of bulk liquids, the VACF method predicted diffusivities that were in closer agreement with the reference Einstein method than with the PACF method, and had a smaller standard error. The PACF method was also more sensitive to the application of a stochastic thermostat to the simulation, so this method should only be used with thermostats that do not strongly affect transport properties. On the other hand, the PACF method was less sensitive to the spring constant chosen for the harmonic restraining force, although the VACF method presented here predicted systematically lower diffusion coefficients if higher spring constants were used. Generally, the VACF method should be used cautiously, and checks should be performed to ensure the extrapolation technique is accurate for a given simulation.

The methods predicted significantly different diffusivities for a water molecule permeating a DPPC lipid bilayer. When the solute is immersed in the bilayer, the PACF can have a very slow decay due to long-timescale fluctuations. This spuriously lowers the calculated diffusion coefficients, particularly in depth of the bilayer corresponding to the lipid tails. In contrast, the VACF decays quickly at all bilayer depths, so it is not affected by these long-timescale fluctuations. The intramembrane diffusion coefficients calculated using the VACF method are systematically higher than those calculated using the PACF method, suggesting that permeabilities calculated using the PACF method are underestimated.

Bibliography

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

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

[3] Siewert J. Marrink and Herman J. C. Berendsen. Permeation process of small molecules across lipid membranes studied by molecular dynamics simulations. *J. Phys. Chem*, 100(41):16729–16738, 1996.

[4] Christopher T. Lee, Jeffrey Comer, Conner Herndon, Nelson Leung, Anna Pavlova, Robert V. Swift, Chris Tung, Christopher N. Rowley, Rommie E. Amaro, Christophe Chipot, Yi Wang, and James C. Gumbart. Simulation-based approaches for determining membrane permeability of small compounds. *J. Chem. Inf. Model.*, 56(4):721–733, 2016.

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

[6] Chris Nealea and Rgis Poms. Sampling errors in free energy simulations of small molecules in lipid bilayers. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, pages –, 2016.

[7] A. Einstein. *Investigations on the Theory of the Brownian Movement.* Dover Books on Physics Series. Dover Publications, 1956.

[8] R. Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.

[9] Gerhard Hummer. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7(1):34, 2005.

[10] Sergei V. Krivov and Martin Karplus. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. USA*, 105(37):13841–13846, 2008.

[11] Jeffrey Comer, Christophe Chipot, and Fernando D. Gonzlez-Nilo. Calculating position-dependent diffusivity in biased molecular dynamics simulations. *J. Chem. Theory Comput.*, 9(2):876–882, 2013.

[12] Mauro L. Mugnai and Ron Elber. Extracting the diffusion tensor from molecular dynamics simulation with milestoning. *J. Chem. Phys.*, 142(1):014105, 2015.

[13] T. B. Woolf and B. Roux. Molecular dynamics simulation of the gramicidin channel in a phospholipid bilayer. *Proc. Natl. Acad. Sci. USA*, 91:11631–11635, 1994.

[14] Thomas B. Woolf and Benoit Roux. Conformational flexibility of o-phosphorylcholine and o-phosphorylethanolamine: A molecular dynamics study of solvation effects. *J. Am. Chem. Soc.*, 116(13):5916–5926, 1994.

[15] Mark F. Schumaker, Rgis Poms, and Benot Roux. A combined molecular dynamics and diffusion model of single proton conduction through gramicidin. *Biophys. J.*, 79(6):2840–2857, 2000.

[16] Mark E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation.* Oxford Graduate Texts. Oxford University Press, Oxford, 2010.

[17] John E. Straub, Michal. Borkovec, and Bruce J. Berne. Calculation of dynamic friction on intramolecular degrees of freedom. *The Journal of Physical Chemistry*, 91(19):4995–4998, 1987.

[18] M.P. Allen and D.J. Tildesley. *Computer Simulation of Liquids*. Oxford Science Publications, Clarendon Press, Oxford, 1989.

[19] Norbert Wiener. Generalized harmonic analysis. *Acta Mathematica*, 55(1):117–258, 1930.

[20] A. Khintchine. Korrelationstheorie der stationären stochastischen prozesse. *Mathematische Annalen*, 109(1):604–615, 1934.

[21] M. Bhandarkar, A. Bhatele, E. Bohm, R. Brunner, F. Buelens, C. Chipot, A. Dalke, S. Dixit, G. Fiorin, P. Freddolino, P. Grayson, J. Gullingsrud, A. Gursoy, D. Hardy, C. Harrison, J. Hnin, W. Humphrey, D. Hurwitz, N. Krawetz, S. Kumar, D. Kunzman, J. Lai, C. Lee, R. McGreevy, C. Mei, M. Nelson, J. Phillips, O. Sarood, A. Shinozaki, D. Tanner, D. Wells, G. Zheng, and F. Zhu. *NAMD User's Guide*. University of Illinois and Beckman Institute, 2015.

[22] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman 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, 1977.

[23] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[24] Johann Fischer and Santiago Lago. Thermodynamic perturbation theory for molecular liquid mixtures. *J. Chem. Phys.*, 78(9):5750–5758, 1983.

[25] William L. Jorgensen. Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *Journal of the American Chemical Society*, 103(2):335–340, 1981.

[26] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, 2010.

[27] K. Gaalswyk and C. N. Rowley. Acfcalculator. `https://github.com/RowleyGroup/ACFCalculator`, 2015.

[28] Chris Neale, W.F. Drew Bennett, D. Peter Tieleman, and Régis Pomès. Statistical convergence of equilibrium properties in simulations of molecular solutes embedded in lipid bilayers. *J. Chem. Theory. Comput.*, 7:4175–4188, 2011.

[29] Chris Neale, Chris Madill, Sarah Rauscher, and Régis Pomès. Accelerating convergence in molecular dynamics simulations of solutes in lipid membranes by conducting a random walk along the bilayer normal. *J. Chem. Theory. Comput.*, 9(8):3686–3703, 2013.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

Computer modeling methods play an important role in the development of new pharmaceutical drugs, and can be used to study a variety of properties and behaviors of chemical systems. In this thesis we have addressed two significant problems in drug development that benefit from computer modeling: molecular conformations and transmembrane diffusion.

In Chapter 2, a conformational search method for explicitly solvated molecules was presented. Molecular conformations affect the biological activity and binding affinity of drug molecules. The presence and representation of a solvent can have significant effects on a conformation. Many conformational search methods only describe a molecule in the gas phase or with an implicit solvent. We have developed a work-flow for performing a conformational search on explicitly solvated molecules using replica-exchange molecular dynamics and clustering analysis. Replica-exchange molecular dynamics has enhanced conformational sampling over conventional molecular dynamics because of exchanges with higher temperature replicas. Clustering analysis effectively identifies the most probable conformation from a REMD trajectory. The work-flow makes use of several open-source software packages and integrates them using Python scripting. Two drug molecules were used as examples. There were significant differences in the lowest energy conformations generated for these molecules in an explicit solvent, an implicit solvent, and in the gas phase.

Chapter 3 discussed Generalized Langevin methods for calculating transmembrane diffusion coefficients. Membrane permeation is a fundamental biochemical transport process, and relies on the rate of diffusion of a solute. While diffusion can be calculated simply for a solute in a homogeneous system, the diffusivity of the solute varies considerably when it is at different depths inside the membrane. Methods based on the Generalized Langevin Equation (GLE) can provide position-dependent diffusivities from a molecular dynamics simulation where the system is restrained to a position along the z-axis by a harmonic potential.

The Generalized Langevin methods presented in this thesis express the diffusion coefficient using the velocity autocorrelation function (VACF) or using the position autocorrelation function (PACF). For the permeation of a solute through a lipid bilayer, the diffusion coefficients calculated using these methods provided significantly different results. The PACF method is sensitive to long correlations of the solute due to inhomogeneities in the bilayer, resulting in underestimations of the diffusion coefficient. The method based on the VACF does not have this issue and predicts higher rates of diffusion inside the bilayer. Our implementation of the VACF method generally predicts diffusion coefficients in closer agreement with the standard rates computed using the Einstein equation, and is less sensitive to an applied Langevin thermostat. It has the drawback of being more sensitive to the spring constant of the restraint potential.

4.2 Future Work

Our work-flow for the conformational search method has the potential to be applicable to many fields such as drug design, cheminformatics, and molecular structure studies. Further work on the method could include other solvents to better represent potential systems. The molecules tested here were both small drug molecules, and an implementation of this work-flow that can work with larger molecules would be beneficial. An important part of any conformational search method is balancing efficiency with accuracy, and improvements to the computational speed while maintaining the benefits of REMD and clustering analysis in this work-flow would increase its applicability.

Currently our implementation of the VACF method is limited by its sensitivity to the spring constant, and further work is needed to fix these issues. The method is also sensitive to the extrapolation technique; within the center of the membrane the D(s) function may not have a well-defined linear section, causing our method to over-estimate the diffusion coefficient. Fixing issues with the extrapolation will improve the accuracy of this method. As with any computational method, improving the efficiency is an important area for improvement. This method would benefit from a more efficient extrapolation technique, and from improvements to its root finding algorithm.

Appendices

Appendix A

Derivation of Expression for D(s) from the GLE

The Generalized Langevin Equation (GLE) for a harmonic oscillator in a frictional bath has the form,

$$m\ddot{z}(t) = -\left(\frac{\partial \mathcal{V}}{\partial z}\right) - \int_0^t \dot{z}(\tau)\ \zeta(t-\tau)d\tau + R(t) \tag{A.1}$$

For a harmonic oscillator, this simplifies to:

$$m\ddot{z}(t) = -k \cdot z(t) - \int_0^t \dot{z}(\tau)\ \zeta(t-\tau)\mathrm{d}\tau + R(t) \tag{A.2}$$

where k is the spring constant of the restraining harmonic potential.

To derive an expression for the diffusion coefficient of the solute, we multiply Eq. (11) through by $\dot{z}(0)$ (i.e., initial velocity),

$$m\ \dot{z}(0)\ \ddot{z}(t) = -\dot{z}(0)\ k\ z(t) - \int_0^t \dot{z}(0)\ \dot{z}(\tau)\ \zeta(t-\tau)d\tau + R(t)\ \dot{z}(0) \tag{A.3}$$

$$m\ \dot{z}(0)\ \left[\frac{d}{dt}\dot{z}(t)\right] = -k\ \dot{z}(0)\ z(t) - \int_0^t \dot{z}(0)\ \dot{z}(\tau)\ \zeta(t-\tau)d\tau + R(t)\ \dot{z}(0) \tag{A.4}$$

Replacing $z(t)$ with an integral of $\dot{z}(t)$ gives:

$$m\ \dot{z}(0)\ \frac{\mathrm{d}}{\mathrm{d}t}\dot{z}(t) = -k\ \dot{z}(0)\ \int_0^t \dot{z}(t)\mathrm{d}t - \int_0^t \dot{z}(0)\cdot\dot{z}(\tau)\ \zeta(t-\tau)\mathrm{d}\tau + R(t)\ \dot{z}(0) \tag{A.5}$$

By taking the ensemble average of both sides of the equation, we obtain:

$$m\ \frac{d}{dt}\ \langle\dot{z}(0)\ \dot{z}(t)\rangle = -k\ \int_0^t \langle\dot{z}(0)\cdot\dot{z}(t)\rangle\,\mathrm{d}t - \int_0^t \langle\dot{z}(0)\cdot\dot{z}(\tau)\rangle\ \zeta(t-\tau)d\tau + \langle R(t)\cdot\dot{z}(0)\rangle \tag{A.6}$$

Random force times initial velocity will average to zero: i.e., $\langle R(t)\ \dot{z}(0)\rangle = 0$.

The velocity autocorrelation function (VACF) is defined as $C_v = \langle\dot{z}(0)\ \dot{z}(t)\rangle$, thus Eq. (A.6) becomes:

$$m\ \frac{\mathrm{d}}{\mathrm{d}t}[C_v] = -k\ \int_0^t C_v\ \mathrm{d}t - \int_0^t C_v\ \zeta(t-\tau)\mathrm{d}\tau \tag{A.7}$$

A Laplace transform of both sides of the equation (i.e., Eq. (A.7)) gives:

$$m\ \mathcal{L}\left(\frac{\mathrm{d}}{\mathrm{d}}[C_v]\right) = -k\ \mathcal{L}\left(\int_0^t C_v\ \mathrm{d}t\right) - \mathcal{L}\left(\int_0^t C_v\ \zeta(t-\tau)d\tau\right) \tag{A.8}$$

Taking advantage of some Laplace transform identities, the above equation simplifies to:

$$m\ [s\ \hat{C}_v(s) - C_v(0)] = -k\ \frac{1}{s}\hat{C}_v(s)\ \ - \hat{C}_v(s)\ \hat{\zeta}(s) \tag{A.9}$$

where $\hat{C}_v = \mathcal{L}(C_v)$. Rearranging the above equation, Eq. (A.9), we can solve for $\hat{\zeta}$:

$$\hat{\zeta}(s) = \frac{-m\ [s\ \hat{C}_v(s) - C_v(0)] - \frac{k}{s}\hat{C}_v(s)}{\hat{C}_v(s)} \tag{A.10}$$

We can then take the reciprocal of the above expression and multiply by k_BT:

$$\frac{k_BT}{\hat{\zeta}(s)} = \frac{-\hat{C}_v(s)k_BT}{m[s\hat{C}_v(s) - C_v(0)] + \dfrac{k}{s}\hat{C}_v(s)} \tag{A.11}$$

D can be related to ζ using Einstein relation $D = \frac{k_BT}{\zeta}$, and taking the limit as s approaches 0:

$$\mathrm{D} = \lim_{s\to 0}\ \frac{k_BT}{\hat{\zeta}(s)} = \lim_{s\to 0}\ \frac{-\hat{C}_v(s)\ k_BT}{m\ [s\ \hat{C}_v(s) - C_v(0)] + \dfrac{k}{s}\ \hat{C}_v(s)} \tag{A.12}$$

where D is the diffusion constant.

This equation can be simplified further using the following identities: $C_v(0) = \langle \dot{z}(0) \cdot \dot{z}(0) \rangle = \langle \dot{z}^2 \rangle$; $k = \dfrac{k_BT}{\langle z^2 \rangle}$; $m = \dfrac{k_BT}{\langle \dot{z}^2 \rangle}$

Using these relations, Eq. (A.12) becomes,

$$\mathrm{D} = \lim_{s\to 0}\ \frac{-\hat{C}_v(s)\ k_BT}{\dfrac{k_BT}{\langle \dot{z}^2 \rangle}[s\ \hat{C}_v(s) - \langle \dot{z}^2 \rangle] + \dfrac{k_BT}{\langle z^2 \rangle}\ \dfrac{1}{s}\ \hat{C}_v(s)} \tag{A.13}$$

We can remove the dependence on k_BT by multiplying the denominator and the numerator by the product of the variance of position and velocity:,

$$\mathrm{D} = \lim_{s\to 0} \frac{-\hat{C}_v(s)\ k_BT}{\frac{k_BT}{\langle \dot{z}^2\rangle}[s\ \hat{C}_v(s) - \langle \dot{z}^2\rangle] + \frac{k_BT}{\langle z^2\rangle}\ \frac{1}{s}\ \hat{C}_v(s)} \times \frac{\langle \dot{z}^2\rangle\langle z^2\rangle}{\langle \dot{z}^2\rangle\langle z^2\rangle} \tag{A.14}$$

and simplifying Eq. (A.14) to:

$$\mathrm{D} = \lim_{s\to 0} \frac{-\hat{C}_v(s)\ \langle \dot{z}^2\rangle\langle z^2\rangle}{s\ \hat{C}_v(s)\langle z^2\rangle - \langle \dot{z}^2\rangle\langle z^2\rangle + \frac{1}{s}\ \hat{C}_v(s)\langle \dot{z}^2\rangle} \tag{A.15}$$

$$\mathrm{D(s)} = \lim_{s\to 0} \frac{-\hat{C}_v(s)\ \langle \dot{z}^2\rangle\langle z^2\rangle}{\hat{C}_v(s)\left[s\ \langle z^2\rangle + \frac{1}{s}\ \langle \dot{z}^2\rangle\right] - \langle \dot{z}^2\rangle\langle z^2\rangle} \tag{A.16}$$

The limit of this expression, Eq. (A.16), can not be taken directly because as $s \to 0$, $\hat{C}_v(s) = 0$. However, an alternative expression can be derived by replacing $\hat{C}_v(s)$ with $\hat{C}_z(s)$. We arrive at this using the definition of the position in terms of the integral over the velocity,

$$z(t) - z(0) = \int_0^t \dot{z}(t')\ \mathrm{d}t' \tag{A.17}$$

The mean square displacement along the z axis can be given in terms of a correlation function by squaring this expression and taking the ensemble average:

$$\langle [z(t) - z(0)]^2 \rangle = \int_0^t \int_0^t \langle \dot{z}(t') \cdot \dot{z}(t'') \rangle \mathrm{d}t'\mathrm{d}t'' \tag{A.18}$$

The integral can be separated into two integrals by a change of variables,[1]

$$\left\langle [z(t) - z(0)]^2 \right\rangle = \int_{-t}^0 \int_0^{t+\tau} \langle \dot{z}(0) \cdot \dot{z}(\tau) \rangle \mathrm{d}\tau\mathrm{d}t'' + \int_0^t \int_\tau^t \langle \dot{z}(0) \cdot \dot{z}(\tau) \rangle \mathrm{d}\tau\mathrm{d}t'' \tag{A.19}$$

$$\langle [z(t) - z(0)]^2 \rangle = 2 \int_0^t (t - \tau) \langle \dot{z}(0) \cdot \dot{z}(\tau) \rangle \mathrm{d}\tau \tag{A.20}$$

For timescales beyond the relaxation time of the system, the integral will converge to a finite value and become independent of t, yielding,

$$\langle [z(t) - z(0)]^2 \rangle = 2t \int_0^{\infty} \langle \dot{z}(0) \cdot \dot{z}(\tau) \rangle \mathrm{d}\tau \tag{A.21}$$

$$\left\langle [z(t) - z(0)]^2 \right\rangle = \int_0^t \int_0^t \langle \dot{z}(t') \cdot \dot{z}(t'') \rangle \mathrm{d}t' \mathrm{d}t'' \tag{A.22}$$

$$\left\langle [z(t) - z(0)]^2 \right\rangle = \int_0^t \int_0^t \langle \dot{z}(t') \cdot \dot{z}(t'') \rangle \mathrm{d}t' \mathrm{d}t'' \tag{A.23}$$

Using the above simple relation based on the classical equations of motion, the expression for the mean squared displacement can be written as,

$$\left\langle [z(t) - z(0)]^2 \right\rangle = \left(\int_0^t dt' \langle \dot{z}(t') \rangle \right)^2 \tag{A.24}$$

Taking the time derivative of both sides of the equation leads to,

$$\frac{\partial}{\partial t} \langle [z(t) - z(0)]^2 \rangle = \frac{\partial}{\partial t} \left(\int_0^t dt' \langle \dot{z}(t') \rangle \right)^2 \tag{A.25}$$

Expansion of the LHS yields,

$$\frac{\partial}{\partial t} \langle z^2(t) - 2z(t)z(0) + z^2(0) \rangle = 2 \int_0^t dt' \langle \dot{z}(t') \cdot \dot{z}(t) \rangle \tag{A.26}$$

This equation can be simplified by noting that the average values of $z(0)^2$ and $z(t)^2$ will be zero for a harmonic oscillator,

$$\frac{\partial}{\partial t} \langle [-2z(t) \cdot z(0)] \rangle = 2 \int_0^t \mathrm{d}t' \langle \dot{z}(t') \cdot \dot{z}(t) \rangle \tag{A.27}$$

This further simplifies to,

$$\frac{\partial}{\partial t}\langle z(t)\cdot z(0)\rangle = -\int_0^t \mathrm{d}t'\langle \dot{z}(t')\cdot \dot{z}(t)\rangle \tag{A.28}$$

Applying the property of time translational invariance to the RHS of the above expression leads to,

$$\frac{\partial}{\partial t}\langle z(t)\cdot z(0)\rangle = -\int_0^t \mathrm{d}t'\langle \dot{z}(t'-t)\cdot \dot{z}(0)\rangle \tag{A.29}$$

If we let $\tau = t' - t$; $d\tau = dt'$. Eq. (A.29) becomes,

$$\frac{\partial}{\partial t}\langle z(t)\cdot z(0)\rangle = -\int_0^t \langle \dot{z}(\tau)\dot{z}(0)\rangle \ \mathrm{d}\tau \tag{A.30}$$

To simplify the expression, we define $C_z(t)$=$\langle z(t)\cdot z(0)\rangle$ and $C_v(\tau)$=$\langle \dot{z}(\tau)\cdot \dot{z}(0)\rangle$, then,

$$\frac{\partial}{\partial t}C_z(t) = -\int_0^t C_v(\tau)\ \mathrm{d}\tau \tag{A.31}$$

The Laplace transform of the above expression leads to,

$$s\ \hat{C}_z(s) - C_z(0) = -\frac{1}{s}\ \hat{C}_v(s) \tag{A.32}$$

$$\hat{C}_v(s) = s\ C_z(0) - s^2\ \hat{C}_z(s) \tag{A.33}$$

Therefore $\hat{C}_v(s) = s\,\langle z^2\rangle - s^2\ \hat{C}_z(s)$; [N.B.: $C_z(0) = \langle z(0)\cdot z(0)\rangle = \langle z^2\rangle$]

Substituting $\hat{C}_v(s)$ into Eq. (A.16) results in,

$$\mathrm{D(s)} = \lim_{s\to 0}\ \frac{-[s\,\langle z^2\rangle - s^2\ \hat{C}_z(s)]\ \langle \dot{z}^2\rangle\langle z^2\rangle}{\left[s\,\langle z^2\rangle - s^2\ \hat{C}_z(s)\right]\left[s\ \langle z^2\rangle + \frac{1}{s}\ \langle \dot{z}^2\rangle\right] - \langle \dot{z}^2\rangle\langle z^2\rangle} \tag{A.34}$$

Which simplifies to,

$$\mathrm{D(s)} = \lim_{s\to 0}\ \frac{-\langle \dot{z}^2\rangle\langle z^2\rangle\ [\langle z^2\rangle - s\ \hat{C}_z(s)]}{\hat{C}_z(s)\,[-s^2\ \langle z^2\rangle - \langle \dot{z}^2\rangle] + s\langle \dot{z}^2\rangle^2} \tag{A.35}$$

$$\mathrm{D(s)} = \lim_{s\to 0}\ \frac{\langle \dot{z}^2\rangle\langle z^2\rangle\ [\langle z^2\rangle - s\ \hat{C}_z(s)]}{\hat{C}_z(s)\,[s^2\ \langle z^2\rangle + \langle \dot{z}^2\rangle] - s\langle \dot{z}^2\rangle^2} \tag{A.36}$$

Noting that $\lim_{s\to 0}\hat{C}_z(s) = \int_0^\infty C_z(t)\mathrm{d}t$, Eq. (A.36) gives

$$\mathrm{D(s)} = \lim_{s\to 0}\ \frac{\langle \dot{z}^2\rangle\langle z^2\rangle^2}{\hat{C}_z(s)\langle \dot{z}^2\rangle}\ = \lim_{s\to 0}\ \frac{\langle z^2\rangle^2}{\hat{C}_z(s)}\ = \frac{\mathrm{var}(z)^2}{\int_0^\infty C_z(t)\ \mathrm{d}t} \tag{A.37}$$

Appendix B

Description of D(s) Extrapolation Method

The calculation of the diffusion coefficient of the solute from the velocity autocorrelation function requires a numerical solution to the equation,

$$D(z_i = \langle z \rangle_i) = \lim_{s \to 0} \frac{-\hat{C}_v(s; z_i)\langle \delta z^2 \rangle_i \langle \dot{z}^2 \rangle_i}{\hat{C}_v(s; z_i)\left[s\langle \delta z^2 \rangle_i + \langle \dot{z} \rangle_i / s\right] - \langle \delta z^2 \rangle_i \langle \dot{z}^2 \rangle_i} \tag{B.1}$$

The singularity at $s = 0$ requires that the value of $D(s = 0)$ be extrapolated numerically from an interval of the equation that is well-behaved. This interval lies between the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ singularities of $D(s)$, denoted s_1 and s_2, (Figure B.1).

To identify the locations of s_1 and s_2, the roots of the denominator are found numerically. The method first finds the location of the minimum of the denominator, $s_{\min}$, using Brent's method [2]. s_1 and s_2 correspond to the roots in the denominator of Eq. 3.5, which are located numerically using Tom's 748 method to find the roots in the $[0.00001, s_{\min}]$ and $[s_{\min}, 1]$ intervals, respectively [3]. The minimum of $D(s)$ ($s_{\min}$) in the $[s_1, s_2]$ interval is then found using Brent's method.

A region of low curvature in $D(s)$ is found by calculating the second derivative

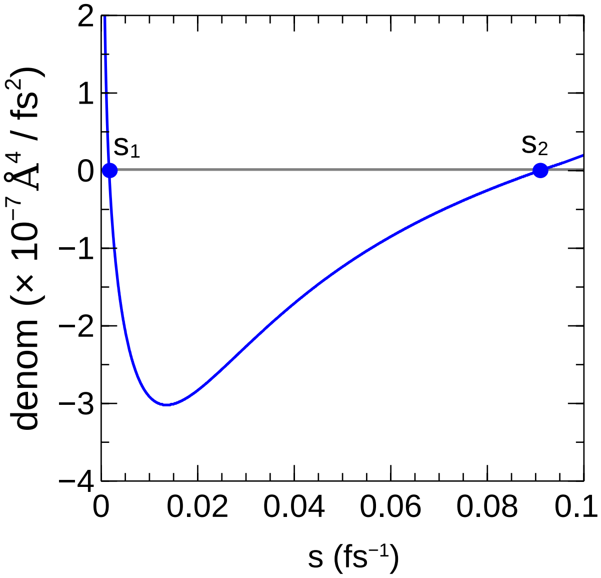

Figure B.1: Denominator of Eq. 3.5 for a harmonically-restrained water molecule in a simulation cell of liquid water (TIP3P model, k=10 kcal mol^{-1} Å^{-2}). The 2$^{\text{nd}}$ and 3$^{\text{rd}}$ singularities of $D(s)$ (s_1 and s_2 respectively) are identified by finding the roots in this equation.

of $D(s)$ over the interval $[s_{\min}, s_2]$ by two iterations of numerical differentiation. A smoothing algorithm is applied at each iteration to reduce the effect of numerical error. The position of minimum curvature in $D(s)$ is then identified by a numerical search. The value of $D(s = 0)$ is extrapolated from this point of minimum curvature.

To improve the numerical stability of the algorithm, a segment with low curvature is extended around the point of minimum curvature. This segment of the curve is fit to a linear equation using least squares regression analysis. $D(s = 0)$ is determined from the y-intercept of this linear approximation of $D(s)$. The coefficient of determination from this fit (R^2) is an indicator of how reliable the extrapolation is.

This method relies on the effective linear extrapolation of $D(s)$ to the y-intercept. A segment where the curvature is low is essential for an accurate extrapolation. The researcher should plot $D(s)$ to ensure the function has a nearly-linear segment between the singularities. The ACFCalculator program will indicate the magnitude of curvature where the extrapolation is being performed. If this value is high, the simulation should be repeated with a smaller spring constant.

Bibliography

[1] Jean Pierre Boon and Sidney Yip. *Molecular Hydrodynamics*. McGraw-Hill, 1980.

[2] R. P. Brent. An algorithm with guaranteed convergence for finding a zero of a function. *Comput. J.*, 14(4):422–425, 1971.

[3] G. E. Alefeld, F. A. Potra, and Yixun Shi. Algorithm 748: Enclosing Zeros of Continuous Functions. *ACM Trans. Math. Softw.*, 21(3):327–344, 1995.