UNIVERSITY OF CALIFORNIA
RIVERSIDE

Molecular Recognition of Association Processes: Assessment of Diffusional Association, Binding Pathways, and Selective Affinity in Ligand-Receptor Complexes

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemistry

by

Christopher Carver Roberts

December 2014

Dissertation Committee:
Dr. Chia-en A. Chang, Chairperson
Dr. Gregory J. O. Beran
Dr. Leonard J. Mueller

The Dissertation of Christopher Carver Roberts is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to my advisor, Dr. Chia-en Chang, for the guidance and support she has provided during my graduate experience. The research environment she created provided necessary training while also allowing the freedom to explore ideas and topics that most interested me. Her support was critical to my success and progression as a graduate student. My thanks also go out to my committee members, Dr. Beran and Dr. Mueller. Their questions and feedback during my research examinations and oral presentations were invaluable for my improvement as a researcher, and pushed my ability to express my work. I would also like to thank my group members, who provided unique perspectives on my research throughout our shared time together. Finally, I'd like to thank my collaborators for illuminating interesting areas of biology and chemistry to explore.

Published materials are used as follows:

Chapter 2: Roberts, C. C.; Chang, C. A. Modeling of Enhanced Diffusional Association in Multi-enzyme Nanostructures: Effect of Molecular Scaffolds, Spatial Organization, and Concentration. *J. Chem. Theory Comput.* **2014**, (accepted).

Chapter 4: Roberts, C. C.; Chang, C. A. Ligand Binding Pathway Elucidation for Cryptophane Host–Guest Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 2010–2019.

ABSTRACT OF THE DISSERTATION

Molecular Recognition of Association Processes: Assessment of Diffusional Association, Binding Pathways, and Selective Affinity in Ligand-Receptor Complexes

by

Christopher Carver Roberts

Doctor of Philosophy, Graduate Program in Chemistry
University of California, Riverside, December 2014
Dr. Chia-en Chang, Chairperson

Molecular recognition processes guide the preferential association of molecular species in chemical and biological systems through physical differentiation and complementarity. Recognition exists at many scales, including the long-range interactions dictating diffusional association, as well as the short-range intermolecular interactions that discriminate the selectivity of a substrate to a host molecule and dictate the configurational transition to the bound state. In this work, methods are developed and employed to study three situations in molecular association processes.

Diffusional association was investigated in the intermediate transfer between enzymes in engineered spatially organized nanostructures. In particular, the glucose oxidase-horseradish peroxidase enzyme pair was investigated in several geometric arrangements of DNA-origami scaffolds. A computationally efficient Brownian dynamics software package was developed for the determination of substrate association probability over microsecond to millisecond time scales. We found that arrangements of enzymes on a

planar scaffold primarily gain efficiency from induced enzyme colocalization with moderate enhancement due to the scaffold acting as a diffusive barrier. Confinement of the enzyme system within a nanotube scaffold greatly enhanced substrate transfer, up to ten fold relative to colocalization, and up to 150 fold relative to a disorganized solution of the enzymes and substrate over the same time period. Our results have implications for the efficient engineering of synthetic enzyme cascades.

A computational method was developed for the determination of final non-covalent binding pathways for molecular complexes. Normal mode calculations were used to model coordinated natural motions of a host-guest complex, which were then utilized to connect conformational minima to form non-covalent binding pathways. Our results demonstrate that conformational transitions can be modeled and extended to construct coordinated final binding events. This approach has advantages over simulation-based methods for studying systems with slow binding processes and can help design molecules with preferred binding kinetics.

A procedure for determining receptor subtype specificity for inhibitors of the human proteasome was performed. A natural product derivative inhibitor was simulated with molecular dynamics, and molecular recognition was assessed through structural stability and energetic affinity with each receptor subtype. The determined energetic and structural preference for two of the three catalytic sites suggests potential for desirable activity as a human proteasome inhibitor.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Overview

Molecular recognition is the broad study of how molecular entities can preferentially associate by differentiation through physical interaction. Its study resides at the junction between pharmacology, biology, chemistry, physics, and computer science. Computational study of physical processes of chemical and biological systems have been ongoing for decades.[3–7] The turn of the 21st century brought about advances in computational power that have allowed computational physical science to bloom into an essential part of the study of chemistry and biology, including the study of molecular recognition processes. Computational modeling and simulation of molecular recognition processes has taken an integral predictive and explanatory role in bioengineering and pharmaceutical drug design. This work aims to use modern computational chemical theory and technology to extend the understanding of the processes involved in the association, binding, and interaction of molecular entities. Subsequent sections outline the computational and chemical theory that allow for the computational study of molecular association. The chapters that follow describe novel computer software to model specific parts of the association process, as well as several studies of association processes in chemical, biological, and bioengineered molecular systems.

Understanding the association process of two molecules requires an understanding of the chemical and physical behavior at multiple spatial and temporal scales. The first step in an association process is the diffusional encounter of two molecular entities, which occurs over a long time scale and large spatial scale. Brownian motion describes the

random diffusive motion of the molecules through a solution. Molecular recognition at this scale takes the form of intermolecular forces that bias the diffusive motion, guiding the translation of the molecules in the vicinity of one another towards desirable locations of interaction. Chapters 2 and 3 outline a novel software application applied to bioengineered protein systems to elucidate various factors that influence the successful diffusional encounter of a small molecule to its protein target.

Following the diffusional encounter of the molecules, the distinct binding process can take place. This is a subtle process guided by the thermodynamics of the interacting molecules. However, there are often energetic barriers to binding that prevent facile elucidation with more traditional computational methods, such as dynamics simulations. Thus, modeling the conformational changes as the binding process occurs, as well as the changes in the enthalpy, entropy, and free energy of the molecules, is desirable. Chapter 4 introduces such a technique, known as Hopping Minima, to model the energetics and conformational structures of intermediate steps in the binding process of a ligand-receptor complex.

Finally, the bound state of a ligand-receptor complex can be studied using dynamics simulations, energetic analysis, and structural analysis to understand the intermolecular interactions that lead to favorable or unfavorable affinity. Chapter 5 demonstrates the study of bound state dynamics in differentiating receptor subtype specificity within the human proteasome.

1.2 Modeling with molecular mechanics

1.2.1 Molecular mechanical theory

Molecular mechanics is a molecular modeling theory that utilizes classical mechanics paired with empirical parameterization to model the structure and behavior of atomic and molecular systems.[8–11] Molecular mechanics typically refers to the modeling of each atom in a molecular or atomic system as explicit parameterized hard spheres. The principles that define molecular mechanics, however, are applied to many types of models that abstract the structure or energetics of traditional all-atom molecular mechanics.

In all-atom molecular mechanics models, atoms are defined explicitly and treated as simple rigid spheres with a defined radius and a point charge embedded in its center. A potential energy function describes the atomic and molecular interactions. The potential energy function of an atomic system is defined generally as follows:

$$E = E_{bonded} + E_{nonbonded} \qquad \text{eq. 1.1}$$

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}$$

$$E_{nonbonded} = E_{electrostatic} + E_{van\ der\ Waals}$$

Molecular entities are defined by a series of atoms connected by bonded, or covalent, terms. Explicit bonds between atoms are modeled as classical harmonic springs with a reference bond length and bond strength specific to the two atomic species involved. The functional form is defined as:

$$E_{bond} = \frac{1}{2}k(l_0 - l)^2 \qquad \text{eq. 1.2}$$

where l_0 is the reference bond length, l is the measured distance between the two atoms involved in the bond, and k is the force constant corresponding to the bond strength or rigidity.

Further defining the geometry of a molecule, an angular potential is defined for each set of three atoms connected sequentially by bonds. Angular terms are most commonly defined by a harmonic potential:

$$E_{angle} = \frac{1}{2}k(\theta_0 - \theta)^2 \qquad \text{eq. 1.3}$$

where θ_0 is the reference angle, θ is the angle measured about three atoms, and k is the force constant dictating the propensity of the atoms to hold their angular conformation.

Finally, dihedral angles are defined for each set of four atoms involved in sequential bonds. For the four atoms defined by labels *i*, *j*, *k* and *l*, each dihedral angle Φ is defined as the angle between the two planes defined by atoms *i*, *j*, and *k* and atoms *j*, *k*, and *l* (Figure 1.1).

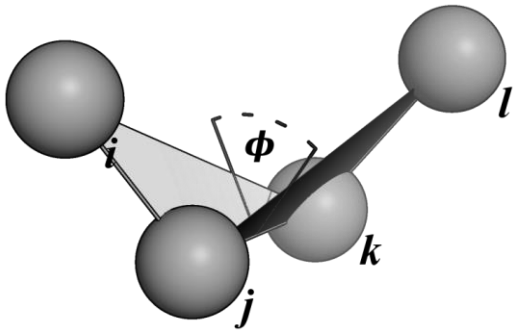

Figure 1.1. Dihedral torsion angle Φ defined between a set of four sequentially bonded atoms *i*, *j*, *k*, and *l*.

Dihedral angles play a crucial role in defining the geometry of molecular entities, as the other bonded and nonbonded potentials cannot adequately capture the forces required to differentiate conformational preference in a molecular fragment. The potential form most commonly used to model dihedral angles is:

$$E_{dihderal} = \frac{1}{2}V_n(1 + \cos(n\Phi - \gamma))^2 \qquad \text{eq. 1.4}$$

where V_n is the force constant, or barrier height, *n* is the periodicity, defining the number of potential minima and maxima with a 360 degree rotation, Φ is the measured dihedral angle, and γ translates the function dictating the angle at which minima and maxima occur. For example, a dihedral angle potential defined by n=2 and γ=180° will have minima at Φ=0° and 180°.

A second form of dihedral angle is referred to as the improper dihedral angle, describing the angle at which one atom deviates from a plane defined by three others. For four atoms *i*, *j*, *k* and *l* in which atom *i* shares a covalent bond with each of *j*, *k*, and *l*, an improper

dihedral angle Φ is defined as the angle between the planes defined by atoms *i*, *j*, and *k* and atoms *j*, *k*, and *l* (Figure 1.2).

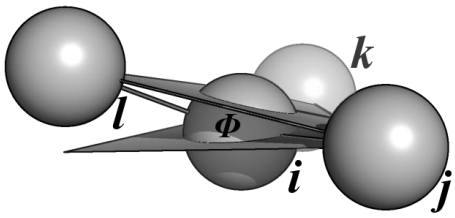

Figure 1.2. Improper dihedral angle Φ defined between a set of three sequentially bonded atoms *j*, *i*, and *k*, and a fourth atom *l* bound to atom *i*.

A potential is defined for atoms involved in improper dihedral angles to prevent out-of-plane motion. This is common to ensure a rigid planar geometry for sp2 hybridized species, such as the carbonyl of peptide bonds in proteins. The aforementioned cosine dihedral potential as well as a harmonic potential are both commonly utilized for improper dihedral potentials:

$$E_{dihderal,improper} = \frac{1}{2}k(1 + \cos(\Phi - \pi))^2 \qquad \text{eq. 1.5}$$

$$E_{dihderal,improper} = \frac{1}{2}k(\Phi)^2 \qquad \text{eq. 1.6}$$

where *k* is the force constant and Φ is the measured improper dihedral angle.

In addition to bonded molecular forces, nonbonded potentials are included to model electrostatic and van der Waals forces. Classical molecular mechanical models take the atom-centered monopole approach to electrostatic interactions. Charges are assigned to each atom as a partial electron charge that is derived from experiment or quantum mechanical calculations.[10–12] Atomic species that are not involved directly in a bond,

angle or dihedral exert nonbonded forces on one another. Electrostatic interactions are modeled using Coulomb interaction potential:

$$E_{electrostatic} = k_e \frac{q_i q_j}{l} \quad \text{eq. 1.7}$$

where q_i and q_j are the partial electron charges assigned to interacting atoms *i* and *j*, *l* is the measured distance between the two atoms, and k_e is Coulomb's constant ($(4\pi\epsilon_0)^{-1}$, where ϵ_0 is the electric permittivity of free space).

Van der Waals forces define the non-polar component of the nonbonded forces in molecular mechanics, and are modeled with the Lennard-Jones (LJ) potential.

$$E_{van\ der\ Waals} = 4\varepsilon \left[\left(\frac{\sigma}{l}\right)^{12} - \left(\frac{\sigma}{l}\right)^{6} \right] \quad \text{eq. 1.8}$$

where ε is the depth of the potential well specific to the types of interacting atoms, σ defines the distance at which the potential value is zero, and *l* is the measured distance between two interacting atoms. Each atom in a molecular mechanical model is assigned a radius derived from the van der Waals radius of the element, with empirical deviations related to specific molecular environments.[13,14] The σ parameter of the LJ potential is calculated by averaging the radii of the two atoms involved in the interaction.

1.2.2 Force fields

Each term in the potential energy function relies on empirically determined parameters to accurately describe the structure and energetics of a molecular system. Decades of experimental data have been rigorously tabulated and integrated into packages of parameters known as force fields.[15–19] Force fields often focus on specific types of molecular systems, such as proteins or nucleotides, while others attempt to generalize

parameters for use with any type of molecular entity including novel compounds.[16,18–20] While force fields can exist independently of specific computer software, they are often paired together to make the process of creating a parameterized molecular mechanical model easier and less time consuming. For example, the AMBER software package contains both force fields and a tool known as LeAP to create a complete parameterized description of the molecular mechanical system, as well as many modeling and simulation tools.[11,20,21]

1.2.3 Solvent modeling

As many chemical and biological processes occur in solution, properly modeling the solvent is important for the accuracy of a solvated model. Water as a solvent has been studied extensively due to its existence in biology and common use in chemistry. Force fields and potential functions specific to water have been actively developed and are in common use.[22,23] In molecular mechanics, the solvent is either modeled explicitly, with hard sphere models, or implicitly through the introduction of forces. Common explicit water models include TIP3P, TIP4P, TIP5P, and SPC(/E).[24,25] These models combine a molecular mechanical description of the water molecule structure with intermolecular force field parameters that attempt to accurately describe the structure and energetics of water. The explicit water models represent the most accurate molecular mechanical models of water. However, the large number of water molecules that are necessary to solvate a molecular system significantly slows the calculation of the potential energy function. In cases where this makes a particular model impractical for a desired application, an implicit water model is often used.

The Poisson-Boltzmann (PB) equation is an accurate function for modeling the electrostatic properties of a charged solute in an ionic solution.[26] The PB equation accurately models the electrostatic field of a charged solute in an implicit solvent with defined salt content:[27]

$$\nabla \cdot [\varepsilon(r)\nabla\phi(r)] = -4\pi\rho(r) - 4\pi\lambda(r)\sum_i z_i c_i e^{\left(\frac{-z_i\phi(r)}{k_b T}\right)} \qquad \text{eq. 1.9}$$

where $\varepsilon(r)$ is the dielectric constant, $\phi(r)$ is the electrostatic potential, $\rho(r)$ is the solute charge, $\lambda(r)$ is the Stern layer masking function, k_b is Boltzmann's constant, T is the temperature, z_i is the charge of ion type *i*, and c_i is the number density of ion type *i* in solution far from the solute.[20] Summation is performed over all ion types (i.e. monovalent cations, divalent anions). Formulations of the PB equation can also be solved to calculate the solvation free energy of a molecular system. Due to the computational expense of solving the PB equation, it is commonly used to investigate energies of static structures of a molecular model, rather than as a part of a dynamics simulation.

The relatively efficient generalized Born (GB) equation is used to approximate the solvation free energy:[28–31]

$$\Delta G_{solv} = -\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{GB}}\left(1 - \frac{e^{(-\kappa f_{GB})}}{\varepsilon}\right) \quad \text{eq. 1.10}$$

where q_i and q_j are the charges on two interacting particles, ε is the solvent dielectric, and κ is Debye-Huckel electrostatic salt screening parameter. The choice of the function f_{GB} typically takes the following form:

$$f_{GB} = \sqrt{l_{ij} + R_i R_j e^{\left(\frac{-{l_{ij}}^2}{4R_i R_j}\right)}} \quad \text{eq. 1.11}$$

where l_{ij} is the distance between two interacting particles, R_i and R_j are the effective Born radii of the atoms. The effective Born radii describes the degree to which a particle is buried within its parent molecule. For an isolated ionic particle, this is equal to its van der Waals radius. The calculation of these radii is imperative to the accuracy of the model.[31] GB is a commonly used electrostatic water model in dynamics simulations due to its relative simplicity and computational practicality.

The computational cost of the PB and GB equations can make their use prohibitive in many situations. A distance dependent dielectric is often used as a coarse approximation to the screening of electrostatic potential imposed by the solvent and salt.[32,33] It is implemented by scaling the Coulombic potential in eq. 1.7 by the inverse of the product of the interparticle distance and a dielectric constant or function ε:

$$E_{electrostatic,ddd} = k_e \frac{q_i q_j}{l^2 \varepsilon} \quad \text{eq. 1.12}$$

Beyond the electrostatic effects of a solvent, other implicit solvent forces may be included in a molecular model to further approximate an omitted solvent, including non-polar contributions to solvation. A non-polar solvation term, related to the surface area of a solute, is commonly combined with either PB or GB to include the effects of solvent implicitly in energetic analysis of molecular structure or interactions.[34] These joint models are commonly referred to as PBSA or GBSA, depending on the use of the PB or GB equation.

Algorithms used to simulate the dynamics of a molecular system, such as Langevin dynamics, include forces to reproduce the friction and random fluctuations a solute incurs in a solvent (Section 1.3). These models and algorithms are not mutually exclusive, and oftentimes implicit electrostatic and non-polar solvent models are combined with a dynamics algorithm like Langevin dynamics to more completely model a solvent implicitly.

1.2.4 Coarse grain modeling

While standard molecular mechanics model the atomic-level detail of a molecular system, the vast number of degrees of freedom inherent in systems with a large number of atoms can make the simulation or calculation of structural properties computationally prohibitive. For this reason it is often desirable to abstract the structure of a molecular system to reduce the number of degrees of freedom by combining atoms into larger interaction spheres. This type of model is referred to as a coarse grain molecular model. Coarse grain models typically target a specific type of molecular entity, such as proteins or hydrocarbons, and can range from simply combining non-polar carbon and hydrogen

atoms (united atom models)[12,35–37], to abstracting entire functional groups, molecular fragments, or entire molecules.[32,38,39]

Structural abstraction comes at the cost of the physical detail obtainable with the model due to fewer parameters and functional forms defining the behavior of the system.[40] Simple abstractions, such as the united atom models, provide similar physical resolution relative to all-atom models while reducing the computational demand. There are many cases where united atom models accurately approximate the behavior of the corresponding explicit atom model, however system-specific tuning is often required for accuracy.[41–45]

Further structural abstraction leads to efficient modeling of the structural dynamics and molecular interactions of large molecules. Proteins are often the subject of coarse grain modeling due to their size and long time scale dynamics. Coarse grain models of proteins often abstract the amino acid monomers into 1 or 2 interaction spheres, depending on the application.[38,40] These models retain the nanometer-scale global geometry of the protein. The lack of atom-specific interactions and the isotropic forces induced by the large spherical interaction sites limit the system to large scale structural dynamics and simplified intermolecular encounter.[46–48] However, these have less influence on the accuracy of the model when applied appropriately to simulate large scale processes, such as protein backbone fluctuations and intermolecular interactions in diffusion processes.[49,50] In the context of molecular recognition, coarse grain modeling is applicable to simulating protein-protein association.[51,52] For the association of a small molecule to a protein, coarse grain models are capable of simulating the initial diffusional

encounter and association of a molecular complex. The details of the final binding process most often requires the atom-level detail provided by a standard all-atom molecular mechanical model due to the complex chemical environment and detail required to model the structural changes upon binding .[50,53]

The potential energy function terms for coarse grain models vary with application. CG models intended for general dynamics often retain the bond, angle, and dihedral angle terms described in Equations 1.2 - 1.6 to define the bonded terms that dictate structure. Nonbonded interactions typically also retain full electrostatics, often with a distance dependent dielectric to approximate screening, and a van der Waals term similar to the LJ potential. General models, such as the MARTINI model, generalize van der Waals interactions between combinations of types of amino acids (e.g. polar, non-polar, neutral, charged).

Other models intended for the study of structural fluctuations or normal mode analysis define the structure in terms of a harmonic elastic network. Elastic network models constrain the relative position of all beads in a CG model with harmonic bonds to retain the general conformation of the starting reference structure. The harmonic bonds are defined similarly to Equation 1.2. Variations on the elastic network model include applications to the study of conformational transitions in proteins. In these cases, multi-welled potentials are used to define multiple conformational states of a protein.[46]

Another type of CG model defines the molecules in a system as rigid bodies. In this model, CG molecular structures retain the exact relative position of all particles

throughout the duration of the simulation. Forces acting on particles within a rigid body are propagated to a singular translational and rotational force for that rigid body.

$$F = \sum F_i \qquad \text{eq. 1.13}$$

$$\tau = \sum r \times F_i \qquad \text{eq. 1.14}$$

where F is translation force on the rigid body, F_i is the force acting on particle i within the rigid body, τ is the rotational force (torque) on the rigid body, and r is the displacement vector of particle i relative to the center of mass of the rigid body. These collective forces are used to translate and rotate the component particles of the rigid body in unison about the center of mass. This model can greatly accelerate simulations and is well suited to studies where structural fluctuation is not important.

Another use for rigid body modeling features the embedding of a collection of point charges as a rigid body within a single CG bead (see Chapter 3). This has the benefit of providing detailed electrostatic interactions for charged interaction sites, and introduces electrostatic interactions to net neutral interaction sites.

1.3 Techniques and applications of molecular models

1.3.1 Minimization

With a defined molecular mechanical system, and an initial set of coordinates, the potential energy function can be subject to an iterative mathematical minimization algorithm to optimize the structure with respect to the forces in the system. These algorithms use the gradient and/or hessian of the potential energy function to iteratively alter the coordinates of the system to minimize the energy in the terms of the potential

energy function. This is an important procedure to perform on a newly constructed molecular mechanical model so that the structure is free from steric clashes or other unfavorable interactions, preventing large forces in subsequent operations performed on the model.

A second-order optimization algorithm used for molecular minimization is the Newton-Raphson method.[54] This method uses contributions from the gradient and hessian of the potential energy function to quickly and accurately minimize a function to its minimum in only a few steps. The computational cost and memory requirements of calculating the hessian makes this method impractical for molecular systems with a large number of atoms.

First-order optimization algorithms include the steepest descent, conjugate gradient, and BFGS methods. The steepest descent method is a simple algorithm that minimizes a function $E(x)$ to a local energy minimum by stepping x in a direction following the negative of the gradient of $E(x)$. The algorithm is efficient at arriving near the closest local minimum, but suffers from a slow rate of convergence to the local minimum, oscillating about the solution. For this reason, the steepest descent method is most often used for coarse initial minimization of a molecular model.

The conjugate gradient method uses both the current and past gradient calculations to greatly improve the convergence rate to the local minimum. However, the conjugate gradient method can be quite slow far from a local minimum. Thus, the steepest descent and conjugate gradient methods are often used together to minimize a molecular system.[55]

The Broyden–Fletcher–Goldfarb–Shanno, or BFGS, algorithm is a quasi-Newton method, meaning it attempts to approximate the second-order Newton-Raphson algorithm.[56–59] BFGS uses the gradient calculations at each successive step to progressively approximate the Hessian of a function. The approximate hessian is used with the exact gradient at each step to minimize a function. BFGS performance and convergence is similar to the conjugate gradient method, but in some cases can converge in fewer steps. A limited memory variant of the algorithm is also available.[60]

1.3.2 Dynamics

The dynamics of a molecular system can be simulated using computational methods known as molecular dynamics (MD). At its simplest, MD calculates the time dependent behavior of a molecular mechanical system by solving Newton's equations of motion. The first MD simulations were performed in the 1950s and 1960s, and explored simple atomic liquid systems.[61,62] Advances in the 1960s and 1970s led to molecular simulations of water followed by the first molecular dynamics simulation of a protein.[5,63] Molecular dynamics is now popularly used to simulate the atomic-level dynamics of solvated biomolecular systems in the nanosecond to microsecond time scale. A variation to the standard form of molecular dynamics, called Brownian dynamics (BD), allows for the simulation of the diffusive motion of a molecular system on the microsecond to second time scale. It is often employed to investigate problems with a larger scale, in space and time, relative to standard MD. For this reason it is commonly paired with a coarse grain molecular model.

1.3.2.1 Molecular dynamics

MD calculates the time dependent behavior of a molecular system by solving Newton's equations of motion, given a potential energy function and coordinates of a molecular system. Newton's second law of motion states that the force acting on a particle is equal to the product of the particle's mass, *m*, and acceleration, *a*:

$$\boldsymbol{F}_i = m_i \boldsymbol{a}_i \qquad \text{eq. 1.15}$$

This can also be stated in relation to gradient of the potential energy function, E, with respect to a particle's coordinates *r*:

$$\boldsymbol{F}_i = -\nabla_{\mathrm{i}}\mathrm{E} = -\frac{dE}{d\boldsymbol{r}_i} \qquad \text{eq. 1.16}$$

Combining the Equations 1.15 and 1.16, the acceleration of a particle can be stated in relation to a potential energy function.

$$m_i \boldsymbol{a}_i = -\frac{dE}{d\boldsymbol{r}_i} \qquad \text{eq. 1.17}$$

Given a position *r*, velocity and acceleration can be defined in relation to position:

$$\boldsymbol{v_i} = \frac{d\boldsymbol{r}_i}{dt} \qquad \text{eq. 1.18}$$

$$\boldsymbol{a}_i = \frac{d\boldsymbol{v_i}}{dt} = \frac{d^2\boldsymbol{r}_i}{d^2t} \qquad \text{eq. 1.19}$$

Integration gives rise to the following, assuming constant acceleration over the time interval:

$$\boldsymbol{v_i} = \int \boldsymbol{a}_i dt = \boldsymbol{a}_i t + \boldsymbol{v}_{i,0} \qquad \text{eq. 1.20}$$

$$\boldsymbol{r}_i = \int \boldsymbol{v_i} dt = \int \left(\boldsymbol{a}_i t + \boldsymbol{v}_{i,0}\right) dt = \frac{\boldsymbol{a}_i t^2}{2} + \boldsymbol{v}_{i,0} t + \boldsymbol{r}_{i,0} \qquad \text{eq. 1.21}$$

Substituting Equations 1.17 into 1.21 yields a relation that defines the position of a particle in relation to the initial position, initial velocity, and a potential energy function:

$$\boldsymbol{r}_i = -\frac{1}{2m_i}\frac{dE}{d\boldsymbol{r}_i}t^2 + \boldsymbol{v}_{i,0}t + \boldsymbol{r}_{i,0} \qquad \text{eq. 1.22}$$

Thus, Equation 1.22 defines the position of a particle *i* at time *t* as a function of the initial position, $\boldsymbol{r}_{i,0}$, the initial velocity, $\boldsymbol{v}_{i,0}$, and a potential energy function *E*. A classical molecular mechanical potential energy function is commonly utilized for molecular dynamics simulations, though more complex forms are also used in practice.[64,65] The initial position of the particles in a simulation varies with the system type, requirement, and availability of experimental structures. For biomolecular simulations, experimentally determined X-ray diffraction or NMR structural coordinates are most commonly available, as evidenced by the entries in the Protein Data Bank.[66] Initial velocities are typically assigned a random distribution of values related to an assigned temperature for the system such that the total initial momentum of the system is zero.

As the potential energy function is complex and dependent on the many degrees of freedom in the system, the equation of motion must be solved numerically. Thus, numerical integration algorithms must be used to advance time with a time step, Δt, from an initial state to some final time *t*. The historically used integration algorithms include the Verlet[4], velocity-Verlet[67], and Leapfrog algorithms. While the algorithms take slightly different approaches, they all assume that accelerations, velocities, and positions can be approximated by a Taylor series expansion. The expansions are typically truncated after the acceleration term.

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{a(t)\Delta t^2}{2} + \cdots \quad \text{eq. 1.23}$$

$$v(t+\Delta t) = v(t) + a(t)\Delta t + \frac{b(t)\Delta t^2}{2} + \cdots \quad \text{eq. 1.24}$$

The Verlet algorithm does not calculate or store a velocity at any time during the simulation. Rather, it uses the difference between the positions of the current and previous time step to estimate velocity. This was particularly attractive when computer memory was a resource in short supply. However, the algorithm does suffer from issues with precision. Because of this it is no longer commonly used for MD. The algorithm is derived by first defining a change in position forward and backward in time:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{a(t)\Delta t^2}{2} \quad \text{eq. 1.25}$$

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{a(t)\Delta t^2}{2} \quad \text{eq. 1.26}$$

Combining Equation 1.26 into Equation 1.25 yields the Verlet equation:

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + a(t)\Delta t^2 \quad \text{eq. 1.27}$$

The velocity-Verlet algorithm uses Equation 1.25 to advance the position of a particle. As the name implies, velocities are calculated and stored throughout the duration of the simulation. The velocity-Verlet algorithm requires, during each step, a partial calculation of the velocity before and after the calculation of potential energy function forces:

$$v(t+\Delta t) = v(t) + \frac{1}{2}(a(t) + a(t+\delta t))\Delta t \quad \text{eq. 1.28}$$

The Leapfrog algorithm is a numerical method, similar to the velocity-Verlet algorithm. Leapfrog differs from velocity-Verlet in that it staggers the calculation of positions and

velocities by a factor of $\frac{1}{2}\Delta t$ such that, at each step, the velocity "leaps" over the position followed by the position "leaping" over the velocity.

Due to the sensitivity of the potential energy function to the position of the particles in an MD simulation, a time step, Δt, on the order of the fastest possible fluctuation of this system must be used. Because bond stretching occurs on a femtosecond timescale, a time step of 1 femtosecond is commonly used in MD simulations.[3] There exist modifications to the previously mentioned integration algorithms that constrain the bond vibrations throughout the duration of a simulation allowing for larger time steps, potentially accelerating the computation of the simulation.[68–70]

MD is often employed to calculate macroscopic observable phenomena, such as the change in energy of a molecular system during a chemical or physical process. While MD simulations calculate information about microscopic states of a molecular system, in the form of positions and momenta, statistical mechanical theory connects these microscopic states to the macroscopic observables. Due to the large number of molecular entities in a macroscopic chemical system, statistical mechanics describes macroscopic observable molecular process variables as averages over a statistical ensemble of all microscopic states of a system. A fundamental premise of statistical mechanics is the Ergodic hypothesis, which states that over a adequately long period of time, all accessible microstates of system are equiprobable.[71,72] Thus, a statistical ensemble average of a process variable is equal to the time average of the process variable in a single system given a sufficiently long sample. MD simulations are performed using one of several statistical ensembles corresponding to the state variables of the system that are held

constant during the simulation. The microcanonical ensemble, or NVE ensemble, is obtained by integrating the equations of motion with no pressure or temperature control, only holding the number of particles, system volume, and energy constant.[73] The canonical ensemble, or NVT ensemble, corresponds to system in thermal equilibrium with a heat bath, corresponding computationally to a thermostat algorithm.[6,74–77] The temperature of the system is held fixed along with the number of particles and volume of the system. In the isothermal-isobaric, or NPT, ensemble, the pressure and temperature are conserved through the simulation through the use of a thermostat and barostat.[6,76,78–82] This ensemble corresponds most closely to experimental conditions, as chemical reactions typically take place under constant pressure.

1.3.2.2 Brownian dynamics

Brownian motion describes the diffusive behavior of solute in a solution. A simulation technique known as Brownian dynamics (BD) has been developed to describe the diffusive processes of molecular systems. There exist many applications of BD to biological systems, such as protein folding, particle coagulation, and diffusion limited reactions.[83]

Brownian dynamics belongs to a class of dynamics techniques known as stochastic MD. BD is derived from Langevin dynamics.[83] Langevin dynamics is a variant of MD that uses stochastic differential equations to model the dynamics of a molecular mechanical system that has omitted solvent degrees of freedom. In particular, Langevin dynamics assumes that explicit solvent molecules are not included in the simulation, and instead includes random atomic forces to model the jostling of a solvent as well as a velocity

damping term to model the friction incurred by a solute in solution. It should, however, be noted that Langevin dynamics does not fully model a solvent implicitly, and thus must be paired with an implicit solvent model, such as GBSA. Starting from Equation 1.17, Langevin dynamics introduces two terms to the equation of motion:

$$m_i \boldsymbol{a}_i = -\frac{dE}{d\boldsymbol{r}_i} - \gamma m_i \boldsymbol{v}_i + \sqrt{2\gamma k_b T m_i} R(t) \qquad \text{eq. 1.29}$$

where γ is the damping coefficient, k_b is Boltzmann's constant, T is the temperature, and R is a zero-mean white noise vector.[84] The introduced terms dictate that the system reach and fluctuate around the temperature *T* according to the fluctuation-dissipation theory. Thus Langevin dynamics acts as a thermostat and approximates the canonical ensemble.[85] When the damping coefficient γ is set to zero, the classical equation of motion is restored. With increasing γ, the Langevin equation crosses the inertial regime and approaches the diffusive regime. At this overdamped limit, no acceleration occurs, and Langevin dynamics becomes Brownian dynamics:

$$0 = -\frac{dE}{d\boldsymbol{r}_i} - \gamma m_i \boldsymbol{v}_i + \sqrt{2\gamma k_b T m_i} R(t) \qquad \text{eq. 1.30}$$

Rearranging Equation 1.30 for velocity, then substituting into Equation 1.21 gives:

$$\boldsymbol{r}_i = \int \boldsymbol{v_i} dt = \int \left(-\frac{1}{\gamma m_i} \frac{dE}{d\boldsymbol{r}_i} + \sqrt{2\frac{k_b T}{\gamma m_i}} R(t) \right) dt \qquad \text{eq. 1.31}$$

Integration with a discrete time Euler-Maruyama scheme[86] leads to:

$$\boldsymbol{r}_i(t + \Delta t) = -\frac{1}{\gamma m_i} \frac{dE}{d\boldsymbol{r}_i} \Delta t + \sqrt{2\frac{k_b T}{\gamma m_i} \Delta t}\, \xi + \boldsymbol{r}_i(t) \qquad \text{eq. 1.32}$$

where ξ is a zero-mean, stationary Gaussian process satisfying:

$$\langle R(t) \rangle = 0 \qquad \text{eq. 1.33}$$

$$\langle R(t)R(t') \rangle = \delta(t - t') \qquad \text{eq. 1.34}$$

Equation 1.32 is typically written in terms of a damping value, $\zeta_i = \gamma m_i$, and the Einstein relation defining the diffusion coefficient $D_i = \frac{k_b T}{\zeta_i}$:

$$\boldsymbol{r}_i(t + \Delta t) = -\frac{D_i}{k_b T}\frac{dE}{d\boldsymbol{r}_i}\Delta t + \sqrt{2D_i \Delta t}\,\xi + \boldsymbol{r}_i(t) \qquad \text{eq. 1.35}$$

Thus, the Brownian dynamics of a particle *i* can be simulated using Equation 1.35 given a potential energy function, *E*, and a diffusion coefficient, D_i. The diffusion coefficient can also be derived from experiment or calculated with respect to the radius of a spherical particle, α, and the viscosity of the solvent, η:

$$D_i = \frac{k_b T}{6\pi\eta\alpha_i} \qquad \text{eq. 1.36}$$

The choice of time step, Δt, should be selected such that $\Delta t \gg \frac{m_i D_i}{k_b T}$, but small enough so the force acting on the particles is effectively constant over the time step to prevent the introduction of error.[83]

1.3.2.3 Model preparation

All-atom model preparation

Simulating the dynamics of an all-atom molecular mechanical system requires careful preparation of the structure because the potential energy function is very sensitive to small changes in the coordinates of the system. Thus, steps must be taken before a molecular dynamics simulation is run to ensure a stable system at equilibrium.

First, the initial set of coordinates must be defined. If an experimentally determined set of

coordinates for the molecule of interest exist in a database, such as the PDB, these coordinates can be used as a starting structure.[66] For proteins, if experimental structures do not exist, but molecules with similar sequence have been characterized, homology model algorithms can predict folding patterns for use as a starting point for the initial coordinates of a model.[87] Small molecules can be constructed manually using molecular visualization and editing software, such as Avogadro or VEGA ZZ.[2,88]

Regardless of the choice of initial coordinates, the protonation state of the molecule must be taken into account. For small molecules with known pKa, and a known pH of the desired model system, the protonation state can be assigned manually by adding or removing hydrogen atoms. For a large molecule, the pKa of the appropriate residues should be calculated in their local chemical environment to determine the correct protonation states.[89] The correct protonation state for a protein is important for the study of protein dynamics and protein-ligand interactions, as hydrogen bonding and electrostatic interactions play important roles in these situations.

With the initial coordinates and correct protonation state determined, a force field can be applied to the set of initial coordinates. A generalizable force field can be applied to organic compounds, such as drugs and other small molecules. Protein force fields are templated by amino acid. Therefore, when a protein force field is applied to an initial set of coordinates using software packages like AMBER, missing atoms will be added automatically in positions relative to the resolved atoms in the residues. The coordinates from both experiment and homology modeling of proteins are likely to be missing at least a few atoms. Experimental structures from X-ray crystallographic experiments do not

resolve hydrogens, and homology models may not be able to predict certain side chain coordinates. If entire amino acid side chains are added, their conformations can be more accurately predicted with software.[90] If only hydrogen atoms and/or a few missing atoms are added by the templating procedure, a constrained minimization, where only the new atoms are minimized, can be performed.

Once the initial coordinates, correct protonation state, and a force field have been assigned, the system can be solvated. Explicit ions, such as sodium and chloride, can be added to the system to render the total charge of the system neutral. If an implicit water model is desired, no further actions are needed. If an explicit water model is to be used, the water molecules can be added to the system using an algorithm that ensures the accurate placement and density of the solvent.

Starting from a solvated model, thorough minimization should be performed to stabilize all intramolecular and intermolecular forces. A set of constrained minimizations should be performed successively. For proteins, this can include minimizing hydrogen atoms, side chain atoms, water molecules, followed by a full system minimization.

Finally, before starting a production molecular dynamics simulation, the system should be brought slowly to equilibrium. This can be accomplished by running simulations with a thermostat for a set system temperature of 50K, 100K, 150K, etc. until the final desired temperature is reached. Each equilibration simulation should be run until the energy of the system is stable over time. Once the system has reached equilibrium, a production simulation can be performed.

Coarse grain model preparation

Coarse grain molecular models abstract the atomic structure of a molecular system. As such, the spatial resolution and sensitivity of the potential energy function with respect to the coordinates is greatly reduced. This means that running a CG Brownian dynamics simulation requires a less rigorous preparation procedure than an all-atom MD simulation.

The initial all-atom coordinates of a molecular system are obtained in the same way as stated above for an all-atom model. Protonation states for this all-atom structure should also be considered to ensure correct formal charges are assigned. The all-atom coordinates can then be abstracted into coarse grain beads with a desired coarse graining scheme and accordingly parameterized with a CG force field. CG models are often system specific and require specialized procedures.

1.3.3 Structural and energetic analysis

The structure and energy of a molecular system can be analyzed, either from a static structure or a set of snapshots from a completed dynamics simulation, to derive information about the system or a process.

The root mean square deviation (RMSD) of a molecular system is the average distance of particles between two states of the system. It allows for analysis of stability of a conformational state.[91] It is also used as the function of minimization for the alignment, or superposition, of two molecular structures. RMSD is defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_0)^2} \qquad \text{eq. 1.37}$$

where r_i is a set of coordinates in some state of interest *i*, and r_0 is the set of coordinates in a reference state denoted by *0*.

Similarly, the root mean square fluctuation (RMSF) of a molecular system is the deviation of a single particle from a reference state over a period of time:

$$RMSF = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(r_i(t) - r_{i,0}\right)^2} \qquad \text{eq. 1.38}$$

where $r_i(t)$ is the position of particle *i* at some time *t*, $r_{i,0}$ is the reference position of particle i, and T is the total period of time to average over. RMSF allows for the quantification of the dynamics of atoms in a molecular system, and can also assess the stability or flexibility of specific parts of a molecular system over time.[92]

MMPBSA (Molecular Mechanics Poisson Boltzmann/Surface Area) and MMGBSA (Molecular Mechanics Generalized Born/Surface Area) are techniques for assessing specific intermolecular interactions and estimating the binding energy of a ligand-receptor system from the trajectory of a dynamics simulation.[93,94] The trajectory of a simulation is stripped of explicit solvent if it is present, and the molecular mechanical energy, including the implicit solvent model defined by either PBSA or GBSA, is calculated and averaged over snapshots from the simulation. For the evaluation of binding energy of a molecular complex, the calculation is performed over three separate simulations, and calculated as the difference between the average energy of the complex,

the free protein, and the free ligand. MMPBSA and MMGBSA are also useful for decomposing the interactions in a single simulation into the interatomic or intermolecular energies. This is useful for molecular recognition and drug design, as it can pinpoint specific sites of attraction or repulsion.[95]

1.3.4 Free energy and entropy

The statistical properties of a system in thermodynamic equilibrium can be expressed with a partition function. A classical partition function is a function of position and momentum of a thermodynamic state defined by temperature, volume, and other parameters. It encodes how the probabilities of microstates are partitioned based on their energies. Thermodynamic values, such as total energy, entropy, free energy, and pressure, can be derived from the partition function of a system. In modeling of molecular recognition processes, the calculation of these thermodynamic properties can allow for the investigation into the stability of a drug bound to protein receptor, or the relative affinity of a series of drug candidates to a target protein.

The canonical ensemble is commonly used in molecular modeling for investigation of a molecular system with a fixed number of particles, in a system with a specific volume and temperature. The corresponding classical canonical partition function is given by:

$$Z = N \int e^{-\beta E(x)}\, dx \qquad \text{eq. 1.39}$$

where *x* is condition of the system that defines the microstate (momentum and position), β is the Boltzmann factor $k_b T^{-1}$, k_b is the Boltzmann's constant, *T* is temperature, *E* defines the total energy function of the system, and *N* is a constant that renders this

classical form unitless. The probability of the system occupying a microstate defined by *x* is:

$$P(x) = \frac{Ne^{-\beta E(x)}}{Z} = \frac{e^{-\beta E(x)}}{\int e^{-\beta E(x)}\, dx} \qquad \text{eq. 1.40}$$

The probability function can be used to calculate ensemble averages, or expectation values, of an observable, such as the total energy, *E*, of a system:

$$\langle E \rangle = \int E(x)P(x)\, dx = \frac{\int E(x)e^{-\beta E(x)}\, dx}{\int e^{-\beta E(x)}\, dx} \qquad \text{eq. 1.41}$$

In the canonical ensemble, the free energy is defined by the Helmholtz free energy, *A*, in terms of a temperature *T* and the partition function:

$$A = -k_b T \ln Z \qquad \text{eq. 1.42}$$

The entropy can also be defined according to the probability density function:

$$S = -k_b \int P(x) \ln P(x) \qquad \text{eq. 1.43}$$

or in terms of the other thermodynamic values:

$$S = \frac{-A + \langle E \rangle}{T} \qquad \text{eq. 1.44}$$

One practical application of Equation 1.43 is to the calculation of configuration entropy of biomolecules using a dynamics trajectory. This approach analyzes the trajectory of a molecular system in internal coordinates to assess the torsional configuration entropy by calculating a probability distribution for each torsional angle.[96] Other dynamics analysis methods include Free Energy Perturbation, empirical linear regression, and mean field potential calculations.[97–99]

Approximations to the calculation of these thermodynamic values is necessary because of

the difficulty in sampling the extremely complex phase space of a molecular system. Many approximations seek to build the partition function of a molecular system through contributions of the minimum energy conformational states:

$$Z = \sum Z_i \qquad \text{eq. 1.45}$$

These approaches attempt to obtain all relevant minimum conformations through systematic conformational search and minimization algorithms.[100,101] The energy well surrounding each minimum conformation is defined through various methods, and integration of this well leads to an approximation of the partition function for a single well, Z_i. The configuration entropy is then defined as:

$$S = \sum S_i = -k_b \sum P_i(x) \ln P_i(x) \qquad \text{eq. 1.46}$$

where

$$P_i(x) = \frac{Z_i}{Z} \qquad \text{eq. 1.47}$$

Mining Minima generation 2 (M2) is a computational method for the determination of the free energy of a molecular system. Its intended use is for the calculation of the change in binding free energy of a molecular complex. M2 calculates the free energy of a molecular system through the determination of configuration integrals for a robust set of minimum conformational states of the system.[102] The Tork algorithm is used to quickly generate a large set of molecular conformations through perturbations on the structure in Bond-Angle-Torsion (BAT) coordinates, allowing for elucidation of conformations that are difficult to sample otherwise.[103] M2 combines several approaches to free energy calculations, utilizing a modified harmonic approximation. M2 calculates the normal

modes of the system in BAT coordinates through calculation and diagonalization of the potential energy function Hessian matrix, and each mode is associated with a Guassian integrand. The structure is perturbed along the normal modes, and energies are calculated to determine if the well can be treated harmonically. Harmonic treatments are improved in M2, avoiding double counting of neighboring wells by limiting integration over the normal modes. Anharmonic minima are treated separately with numerical integration.[102]

1.3.4 Docking

A primary challenge in molecular recognition, drug design, and drug discovery studies is the prediction of the position, orientation, conformation, and energetic affinity of a small molecule in the binding pocket of a molecular receptor.[104] Computational methods known as molecular docking predict bound conformational states and corresponding energetic affinities of a ligand-receptor complex. Docking methods are typically composed of a conformational search routine coupled to an energetic evaluation function. A large number of conformational states are sampled and subsequently subject to an evaluation function to determine the states with the most favorable intermolecular interaction. Various simplifications to energy evaluation and conformational search are employed to ensure docking methods are computationally practical for their intended use.[105]

The estimation of the energetic interaction between a ligand and receptor is handled by an evaluation routine known as a scoring function.[106] Scoring functions typically fall into three categories: molecular mechanical potential energy functions, empirical scoring functions, and knowledge based potentials.[107–112] Using these approaches, the chemical and steric complementarity are evaluated for a given conformation, relative position, and

relative orientation of the ligand and receptor. The output of scoring functions can include potential energy values, free energy values, binding constants, or simple rankings of the various sampled conformations.

Conformational search routines for docking algorithms modify the position and orientation of the ligand relative to the receptor in an attempt to sample a wide range of complex conformational states for evaluation with the scoring function. Docking search algorithms are defined generally by the rigidity imposed on the molecular species. For the fastest sampling of the complex conformations, rigid body constraints are placed on the ligand and receptor such that only the ligand translational and rotational degrees of freedom are modified. In these cases, a single receptor conformation is used in combination with the results of an independent ligand conformational search to start multiple complex conformational searches. In flexible docking methods, certain portions of the receptor structure, such as the side chains of proteins, are also allowed conformational degrees of freedom. Flexible docking algorithms can oftentimes be more accurate than rigid docking algorithms because of their ability to sample induced-fit bound complex conformational states.[113,114] More complex docking schemes can include minimization or dynamics procedures of the sampled complex conformation to refine the bound states.

In the context of drug discovery, it is desirable to screen large libraries of small molecules against a molecular receptor to predict drug candidates with high affinity for further empirical study. Traditionally, high throughput *in situ* screening has been used to test libraries of small molecules against a molecular target receptor. Docking methods for

in silico virtual screening were developed to replace or complement existing high throughput screening methods.[115,116] Rigid docking with an empirical or semi-empirical scoring function is often employed for virtual screening purposes.

For molecular recognition studies with molecular dynamics, docking methods can serve as a means to generate starting conformations for simulations. This has the benefit of starting the molecular system reasonably near a local energy minimum, as well as removing human bias of manually placing a ligand in a binding pocket. Docking typically is not used for direct energy evaluation for molecular recognition studies, as statistical mechanical free energy calculation methods, such as M2, can more accurately predict both the bound state conformation and binding free energy of a complex.[102]

1.4 References

(1) Groll, M.; Schellenberg, B.; Bachmann, A. S.; Archer, C. R.; Huber, R.; Powell, T. K.; Lindow, S.; Kaiser, M.; Dudler, R. A Plant Pathogen Virulence Factor Inhibits the Eukaryotic Proteasome by a Novel Mechanism. *Nature* **2008**, *452*, 755–758.

(2) Avogadro: an open-source molecular builder and visualization tool http://avogadro.openmolecules.net/ (accessed Sep 20, 2012).

(3) Ameijden, M. van. Molecular Dynamics Simulation and Visualisation in a Grid Environment. *J. Chem. Phys.* **1957**.

(4) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98–103.

(5) Stillinger, F. H. Improved Simulation of Liquid Water by Molecular Dynamics. *J. Chem. Phys.* **1974**, *60*, 1545.

(6) Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure And/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384.

(7) Pal, S.; Fichthorn, K. A. Accelerated Molecular Dynamics of Infrequent Events. *Chem. Eng. J.* **1999**, *74*, 77–83.

(8) Westheimer, F.; Newman, M. Steric Effects in Organic Chemistry. *by MS Newman, John Wiley Sons, New York* **1956**.

(9) Engler, E. M.; Andose, J. D.; Schleyer, P. V. R. Critical Evaluation of Molecular Mechanics. *J. Am. Chem. Soc.* **1973**, *95*, 8005–8025.

(10) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(11) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

(12) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

(13) Bondi, A. Van Der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(14) Alvarez, S. A Cartography of the van Der Waals Territories. *Dalton Trans.* **2013**, *42*, 8617–8636.

(15) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

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

(17) Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

(18) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

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

(20) Case, D.; Darden, T.; III, T. C. AMBER 12. *Univ. California* **2012**.

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

(22) Reimers, J. R.; Watts, R. O.; Klein, M. L. Intermolecular Potential Functions and the Properties of Water. *Chem. Phys.* **1982**, *64*, 95–114.

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

(24) Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001**, *105*, 9954–9960.

(25) Zielkiewicz, J. Structural Properties of Water: Comparison of the SPC, SPCE, TIP4P, and TIP5P Models of Water. *J. Chem. Phys.* **2005**, *123*, 104501.

(26) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10037–10041.

(27) Gruziel, M.; Grochowski, P.; Trylska, J. The Poisson-Boltzmann Model for tRNA: Assessment of the Calculation Set-up and Ionic Concentration Cutoff. *J. Comput. Chem.* **2008**, *29*, 1970–1981.

(28) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(29) Bashford, D.; Case, D. A. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

(30) Tsui, V.; Case, D. A. Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations. *Biopolymers 56*, 275–291.

(31) Onufriev, A.; Case, D. A.; Bashford, D. Effective Born Radii in the Generalized Born Approximation: The Importance of Being Perfect. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(32) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(33) Schaefer, M.; Bartels, C.; Karplus, M. Solution Conformations of Structured Peptides: Continuum Electrostatics versus Distance-Dependent Dielectric Functions. *Theor. Chem. Accounts Theory, Comput. Model. (Theoretica Chim. Acta)* **1999**, *101*, 194–204.

(34) Feig, M.; Brooks, C. L. Recent Advances in the Development and Application of Implicit Solvent Models in Biomolecule Simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(35) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comput. Chem.* **1986**, *7*, 230–252.

(36) Chiu, S.-W.; Pandit, S. A.; Scott, H. L.; Jakobsson, E. An Improved United Atom Force Field for Simulation of Mixed Lipid Bilayers. *J. Phys. Chem. B* **2009**, *113*, 2748–2763.

(37) Yang, L.; Tan, C.-H.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollman, P. A.; Luo, R. New-Generation Amber United-Atom Force Field. *J. Phys. Chem. B* **2006**, *110*, 13166–13176.

(38) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(39) Bauler, P.; Huber, G.; Leyh, T.; McCammon, J. A. Channeling by Proximity: The Catalytic Advantages of Active Site Colocalization Using Brownian Dynamics. *J. Phys. Chem. Lett.* **2010**, *1*, 1332–1335.

(40) Tozzini, V. Coarse-Grained Models for Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(41) McCoy, J. D.; Curro, J. G. Mapping of Explicit Atom onto United Atom Potentials. *Macromolecules* **1998**, *31*, 9362–9368.

(42) Chen, C.; Depa, P.; Maranas, J. K.; Garcia Sakai, V. Comparison of Explicit Atom, United Atom, and Coarse-Grained Simulations of Poly(methyl Methacrylate). *J. Chem. Phys.* **2008**, *128*, 124906.

(43) Chen, C.; Depa, P.; Sakai, V. G.; Maranas, J. K.; Lynn, J. W.; Peral, I.; Copley, J. R. D. A Comparison of United Atom, Explicit Atom, and Coarse-Grained Simulation Models for Poly(ethylene Oxide). *J. Chem. Phys.* **2006**, *124*, 234901.

(44) Paul, W.; Yoon, D. Y.; Smith, G. D. An Optimized United Atom Model for Simulations of Polymethylene Melts. *J. Chem. Phys.* **1995**, *103*, 1702.

(45) Zaman, M. H.; Shen, M.-Y.; Berry, R. S.; Freed, K. F. Computer Simulation of Met-Enkephalin Using Explicit Atom and United Atom Potentials: Similarities, Differences, and Suggestions for Improvement. *J. Phys. Chem. B* **2003**, *107*, 1685–1691.

(46) Chu, J.-W.; Voth, G. A. Coarse-Grained Free Energy Functions for Studying Protein Conformational Changes: A Double-Well Network Model. *Biophys. J.* **2007**, *93*, 3860–3871.

(47) Hills, R. D.; Brooks, C. L. Insights from Coarse-Grained Gō Models for Protein Folding and Dynamics. *Int. J. Mol. Sci.* **2009**, *10*, 889–905.

(48) Periole, X.; Cavalli, M.; Marrink, S.-J.; Ceruso, M. A. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theory Comput.* **2009**, *5*, 2531–2543.

(49) Tozzini, V.; McCammon, J. A. A Coarse Grained Model for the Dynamics of Flap Opening in HIV-1 Protease. *Chem. Phys. Lett.* **2005**, *413*, 123–128.

(50) Chang, C.-E. A.; Trylska, J.; Tozzini, V.; McCammon, J. A. Binding Pathways of Ligands to HIV-1 Protease: Coarse-Grained and Atomistic Simulations. *Chem. Biol. Drug Des.* **2007**, *69*, 5–13.

(51) Basdevant, N.; Borgis, D.; Ha-Duong, T. A Coarse-Grained Protein-Protein Potential Derived from an All-Atom Force Field. *J. Phys. Chem. B* **2007**, *111*, 9390–9399.

(52) Frembgen-Kesner, T.; Elcock, A. H. Absolute Protein-Protein Association Rate Constants from Flexible, Coarse-Grained Brownian Dynamics Simulations: The

Role of Intermolecular Hydrodynamic Interactions in Barnase-Barstar Association. *Biophys. J.* **2010**, *99*, L75–7.

(53) Kang, M.; Roberts, C.; Cheng, Y.; Chang, C. A. Gating and Intermolecular Interactions in Ligand-Protein Association: Coarse-Grained Modeling of HIV-1 Protease. *J. Chem. Theory Comput.* **2011**, *7*, 3438–3446.

(54) Hilderbrandt, R. L. Application of Newton-Raphson Optimization Techniques in Molecular Mechanics Calculations. *Comput. Chem.* **1977**, *1*, 179–186.

(55) Zimmermann, K. ORAL: All Purpose Molecular Mechanics Simulator and Energy Minimizer. *J. Comput. Chem.* **1991**, *12*, 310–319.

(56) Broyden, C. The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.* **1970**.

(57) Fletcher, R. A New Approach to Variable Metric Algorithms. *Comput. J.* **1970**.

(58) Goldfarb, D. A Family of Variable-Metric Methods Derived by Variational Means. *Math. Comput.* **1970**.

(59) Shanno, D. Conditioning of Quasi-Newton Methods for Function Minimization. *Math. Comput.* **1970**.

(60) Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208.

(61) Alder, B. J.; Wainwright, T. E. Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **1957**, *27*, 1208.

(62) Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **1959**, *31*, 459.

(63) McCammon, J. Dynamics of Folded Proteins. *Nature* **1977**.

(64) Borodin, O. Polarizable Force Field Development and Molecular Dynamics Simulations of Ionic Liquids. *J. Phys. Chem. B* **2009**, *113*, 11463–11478.

(65) Hu, H.; Elstner, M.; Hermans, J. Comparison of a QM/MM Force Field and Molecular Mechanics Force Fields in Simulations of Alanine and Glycine

"Dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in Water in Relation to the Problem of Modeling the Unfolded Peptide Backbone in Solution. *Proteins* **2003**, *50*, 451–463.

(66) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *Arch. Biochem. Biophys.* **1978**, *185*, 584–591.

(67) Swope, W. C. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76*, 637.

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

(69) Andersen, H. C. Rattle: A "velocity" Version of the Shake Algorithm for Molecular Dynamics Calculations. *J. Comput. Phys.* **1983**, *52*, 24–34.

(70) Palmer, B. J. Direct Application of Shake to the Velocity Verlet Algorithm. *J. Comput. Phys.* **1993**, *104*, 470–472.

(71) Boltzmann, L. *Vorlesungen Über Gastheorie*; 1896.

(72) Lebowitz, J.; Percus, J.; Verlet, L. Ensemble Dependence of Fluctuations with Application to Machine Computations. *Phys. Rev.* **1967**, *153*, 250–254.

(73) Cheung, P. S. Y. On the Calculation of Specific Heats, Thermal Pressure Coefficients and Compressibilities in Molecular Dynamics Simulations. *Mol. Phys.* **1977**, *33*, 519–526.

(74) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81*, 511.

(75) Tobias, D. J.; Martyna, G. J.; Klein, M. L. Molecular Dynamics Simulations of a Protein in the Canonical Ensemble. *J. Phys. Chem.* **1993**, *97*, 12959–12966.

(76) Soddemann, T.; Dünweg, B.; Kremer, K. Dissipative Particle Dynamics: A Useful Thermostat for Equilibrium and Nonequilibrium Molecular Dynamics Simulations. *Phys. Rev. E* **2003**, *68*, 046702.

(77) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(78) Nosé, S.; Klein, M. L. Constant Pressure Molecular Dynamics for Molecular Systems. *Mol. Phys.* **1983**, *50*, 1055–1076.

(79) Melchionna, S.; Ciccotti, G.; Lee Holian, B. Hoover NPT Dynamics for Systems Varying in Shape and Size. *Mol. Phys.* **1993**, *78*, 533–544.

(80) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101*, 4177.

(81) Sturgeon, J. B.; Laird, B. B. Symplectic Algorithm for Constant-Pressure Molecular Dynamics Using a Nosé–Poincaré Thermostat. *J. Chem. Phys.* **2000**, *112*, 3474.

(82) Bussi, G.; Zykova-Timan, T.; Parrinello, M. Isothermal-Isobaric Molecular Dynamics Using Stochastic Velocity Rescaling. *J. Chem. Phys.* **2009**, *130*, 074101.

(83) Ermak, D. L.; McCammon, J. A. Brownian Dynamics with Hydrodynamic Interactions. *J. Chem. Phys.* **1978**, *69*, 1352.

(84) Klann, M.; Koeppl, H. Spatial Simulations in Systems Biology: From Molecules to Cells. *Int. J. Mol. Sci.* **2012**, *13*, 7798–7827.

(85) Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide: An Interdisciplinary Guide*; Springer Science & Business Media, 2010; p. 768.

(86) Erban, R.; Chapman, S. J. Stochastic Modelling of Reaction-Diffusion Processes: Algorithms for Bimolecular Reactions. *Phys. Biol.* **2009**, *6*, 046001.

(87) Schwede, T. SWISS-MODEL: An Automated Protein Homology-Modeling Server. *Nucleic Acids Res.* **2003**, *31*, 3381–3385.

(88) Pedretti, A.; Villa, L.; Vistoli, G. VEGA - An Open Platform to Develop Chemo-Bio-Informatics Applications, Using Plug-In Architecture and Script Prpgramming. *J.C.A.M.D* **2004**, *18*, 167–173.

(89) Antosiewicz, J.; Briggs, J. M.; Elcock, A. H.; Gilson, M. K.; McCammon, J. A. Computing Ionization States of Proteins with a Detailed Charge Model. *J. Comput. Chem.* **1996**, *17*, 1633–1644.

(90) Xiang, Z.; Honig, B. Extending the Accuracy Limits of Prediction for Side-Chain Conformations. *J. Mol. Biol.* **2001**, *311*, 421–430.

(91) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Reversible Peptide Folding in Solution by Molecular Dynamics Simulation. *J. Mol. Biol.* **1998**, *280*, 925–932.

(92) Koshy, C.; Parthiban, M.; Sowdhamini, R. 100 Ns Molecular Dynamics Simulations to Study Intramolecular Conformational Changes in Bax. *J. Biomol. Struct. Dyn.* **2010**, *28*, 71–83.

(93) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc.* **2001**, *123*, 5221–5230.

(94) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.py : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 3314–3321.

(95) Massova, I.; Kollman, P. A. Combined Molecular Mechanical and Continuum Solvent Approach (MM-PBSA/GBSA) to Predict Ligand Binding. *Perspect. Drug Discov. Des.* **2000**, *18*, 113–135.

(96) Ai, R.; Qaiser Fatmi, M.; Chang, C.-E. A. T-Analyst: A Program for Efficient Analysis of Protein Conformational Changes by Torsion Angles. *J. Comput. Aided. Mol. Des.* **2010**, *24*, 819–827.

(97) Weber, P. C.; Wendoloski, J. J.; Pantoliano, M. W.; Salemme, F. R. Crystallographic and Thermodynamic Comparison of Natural and Synthetic Ligands Bound to Streptavidin. *J. Am. Chem. Soc.* **1992**, *114*, 3197–3200.

(98) Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417.

(99) Jiang, L.; Gao, Y.; Mao, F.; Liu, Z.; Lai, L. Potential of Mean Force for Protein-Protein Interaction Studies. *Proteins* **2002**, *46*, 190–196.

(100) Gō, N. Analysis of the Contribution of Internal Vibrations to the Statistical Weights of Equilibrium Conformations of Macromolecules. *J. Chem. Phys.* **1969**, *51*, 4751.

(101) Wang, J.; Szewczuk, Z.; Yue, S. Y.; Tsuda, Y.; Konishi, Y.; Purisima, E. O. Calculation of Relative Binding Free Energies and Configurational Entropies: A Structural and Thermodynamic Analysis of the Nature of Non-Polar Binding of Thrombin Inhibitors Based on hirudin55-65. *J. Mol. Biol.* **1995**, *253*, 473–492.

(102) Chang, C.-E.; Gilson, M. K. Free Energy, Entropy, and Induced Fit in Host-Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.

(103) Chang, C.-E.; Gilson, M. K. Tork: Conformational Analysis Method for Molecules and Complexes. *J. Comput. Chem.* **2003**, *24*, 1987–1998.

(104) Lybrand, T. P. Ligand—protein Docking and Rational Drug Design. *Curr. Opin. Struct. Biol.* **1995**, *5*, 224–228.

(105) Brooijmans, N.; Kuntz, I. D. Molecular Recognition and Docking Algorithms. **2003**.

(106) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.

(107) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput. Aided. Mol. Des.* **2001**, *15*, 411–428.

(108) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(109) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 425–445.

(110) Böhm, H.-J. Prediction of Binding Constants of Protein Ligands: A Fast Method for the Prioritization of Hits Obtained from de Novo Design or 3D Database Search Programs. *J. Comput. Aided. Mol. Des.* **1998**, *12*, 309.

(111) Muegge, I. A Knowledge-Based Scoring Function for Protein-Ligand Interactions: Probing the Reference State. *Perspect. Drug Discov. Des.* **2000**, *20*, 99–114.

(112) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791–804.

(113) Sandak, B.; Wolfson, H.; Nussinov, R. Flexible Docking Allowing Induced Fit in Proteins: Insights from an Open to Closed Conformational Isomers. *Proteins Struct. Funct.* **1998**.

(114) Nabuurs, S.; Wagener, M.; Vlieg, J. De. A Flexible Approach to Induced Fit Docking. *J. Med.* **2007**.

(115) Kitchen, D.; Decornez, H. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug* **2004**.

(116) Shoichet, B. Virtual Screening of Chemical Libraries. *Nature* **2004**.

Chapter 2: Modeling of Enhanced Catalysis in Multi-enzyme Nanostructures: Effect of Molecular Scaffolds, Spatial Organization, and Concentration

2.1 Introduction

Many biological catabolic and metabolic processes occur within colocalized multi-step enzymatic reaction pathways with high yield and specificity.[1–5] In these complexes, the relative orientation and position of the enzymes can allow for efficient diffusion of substrates between the active sites of enzymes in the complex.[6,7] Enzymes often associate when freely diffusing in solution, but they also commonly associate on membrane surfaces, peripherally or as integral membrane proteins. The organization on surfaces has been proposed to increase local substrate concentration or substrate dwell time, caused by reduced dimensionality, potentially leading to additional gains in efficiency.[8–11] These factors may all help create an efficient environment for catalysis, but we do not fully understand their contribution to the efficiency of multi-enzyme constructs. A fundamental understanding of the factors that lead to this efficiency can result in engineered reaction-coupled enzyme systems *in vitro* to develop novel synthetic pathways or replace costly or wasteful existing procedures.

Techniques have been developed to create various types of spatially organized reaction-coupled enzyme systems.[4,12–19] Studies have explored fixating the relative positions of enzyme systems, such as anchoring enzymes to various molecular scaffold.[20–24] Recent advances in DNA synthetic procedures have been exploited for anchoring enzymes to various types of DNA scaffolds with highly accurate positioning.[20,25–30] Such experiments

have attempted to establish the advantage of immobilizing enzyme systems and a relationship between enzyme distance and enhancement of enzyme efficiency.

Taking advantage of complimentary nucleic acid base-pairing, Wilner *et al* synthesized a flexible, hexagonal DNA scaffold with dangling DNA anchor points for two DNA-modified enzymes that share a product-reactant interdependence: glucose oxidase (GOx) and horseradish peroxidase (HRP).[20,31,32] The GOx-HRP system is an appropriate model for assessing the efficiency of spatially organized reaction-coupled enzyme systems because GOx produces H_2O_2 as a side product, a required co-substrate for HRP turnover. The authors observed enhanced enzymatic efficiency because of enhanced interenzyme substrate transfer of H_2O_2 from GOx to HRP, showing the rate in the mutual turnover of GOx and HRP is dependent on this intermediate transfer under certain conditions.[20] Fu *et al* expanded on this work by constructing a more rigid DNA origami surface with highly accurate programmable anchor points for the enzymes.[25] Enzymatic enhancement was high at small interenzyme distance, with reduced efficiency with increasing separation. The enhancement was considered due to increased local substrate concentration from restricted diffusion, attributable to the DNA scaffold. However, these conclusions remain controversial.[33]

Previous work has used analytical approaches to characterize the efficacy of spatially organized systems under various conditions.[33,34] These insightful models account for important features of spatially organized systems over macroscopic timescales but are based on theory that may not fully consider the implications of subtle geometric features of scaffolded systems on the diffusive behavior of the substrate.[35,36] Previous

computational simulations have shown promise in accurately describing diffusive behavior in the presence of coarse-grained (CG) biomolecular structures.[37–39] Bauler *et al* used Brownian dynamics (BD) simulations to help characterize the effect of interenzyme distance and relative rotational orientation on intermediate substrate transfer in generic enzyme systems with simplified geometry.[38]

In this work, we explored the geometric features of an enzyme system anchored to a planar scaffold by using parallel CG BD simulations in microsecond to millisecond timescales. We investigated the mechanisms and revealed key features that enhance the efficiency of enzymes spatially organized on a DNA origami scaffold. We modeled the GOx-HRP system and assessed the substrate transfer probability dependence with spatial organization, system concentration, and surface scaffold interaction. We introduce BD simulation software, GeomBD, optimized for assessing specific geometric effects of large biomolecular systems over long timescales. To efficiently approach simulation over millisecond timescales, we developed this software by using modern parallelization techniques and a distance-dependent time-stepping algorithm.[40]

2.2 Methods

We simulated a CG GOx-HRP scaffolded enzyme system using BD implemented in the GeomBD package. The H2O2 intermediate transfer, from the release site of GOx to the binding site of HRP, was simulated in parallel to determine the probability of binding-site association over the timespan of the turnover of the originating GOx enzyme (approximated as 300 s-1).[25,41–43] Interenzyme distance, enzyme concentration, and the presence of scaffold were investigated for their impact on intermediate transfer.

Figure 2.1: Coarse-grained model of glucose oxidase (GOx; left, orange) and horseradish peroxidase (HRP; right, purple) scaffolding system. The 2D plane represents the DNA origami scaffold model, and a black sphere, close to GOx, represents the H_2O_2 substrate.

2.2.1 Molecular Systems

The HRP isozyme C and GOx atomic structures were obtained from the protein data bank (PDB: 1ATJ and 1GPE, respectively).[44,45] CG enzyme models were created from the obtained atomic PDB structures. Each amino acid was represented by an interaction sphere centered on the alpha-carbon of each residue. The radii of the amino acids were obtained from previous work: derived from hundreds of all-atom protein structures, the radii were calculated from the maximum distance from the alpha-carbon to all other atoms and then increased by a constant, 1.4 Å, to account for van der Waals radii.[46,47] Similarly, a CG H_2O_2 molecule was created by using a single bead with radius derived from the radius of gyration of the all-atom structure, then increased by the constant 1.4 Å, for a radius of approximately 2.5 Å. The DNA origami scaffold was represented by a simple 2D rectangular plane barrier of 60 by 80 nanometers, matching the dimensions of the experimental DNA tile.[25] The structural details of the DNA are abstracted into a

smooth plane; thus the results arising from this model are generalized to generic planar scaffolds. Explicit CG representations of the DNA strands that link each enzyme to the 2D scaffold were created from all-atom representations of a randomly sequenced DNA strand created using the NAB AMBER software interface.[48] For each nucleoside base in this CG DNA representation, two beads were positioned to represent the base and the associated sugar–phosphate backbone atoms. A Lennard-Jones-like potential was used for the CG ligand–protein and ligand–DNA linker interactions, as described previously, with the form:

$$U = \frac{1}{2}\left(\left(\frac{\sigma}{r}\right)^8 - \frac{3}{2}\left(\frac{\sigma}{r}\right)^6\right)$$

where r is the distance between a ligand and an interaction bead, and σ is the sum of the radii of the two interacting spheres.[49] The ligand interaction with the 2D planar scaffold was modeled with the same potential, however with the calculation of r modified for the implicit planar geometry. The radius of the implicit field radiating from the 2D planar scaffold was set to 1 nanometer to match the thickness of the experimental DNA origami tile.

The H_2O_2 CG model was assigned an experimentally determined diffusion coefficient of 1×10^{-3} nm^2/ps.[50] Electrostatic interactions between the ligand and other molecular structures were excluded from the simulations because of the neutrally charged substrate. Excluding electrostatic interactions isolates the effects brought about by the geometry of the spatially organized system without the system specific bias of long-range guided association. As shown previously, electrostatic interactions play a significant role in

diffusive association in highly charged systems.[38] Thus, the impact of electrostatic interactions can be investigated in future studies using an embedded partial charge scheme. All structures, aside from the ligands, were held fixed in space during the simulations. Thus, no intramolecular potentials were included. A cubic simulation volume was used for all simulations. Volumetric bounds for all simulations involved periodic boundary conditions. As each simulation replicate contains one GOx enzyme, one HRP enzyme, and one substrate, the concentration of each solute was dictated by the simulation volume, with concentrations of 1, 10, and 100 nM corresponding to box edge lengths of approximately 1184, 550 and 255 nm, respectively (Figure 2.2).

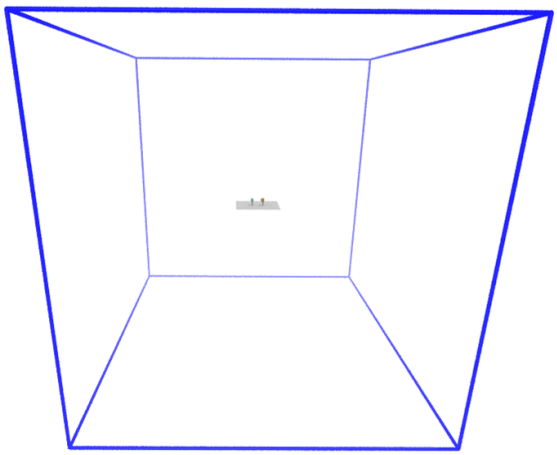

Figure 2.2: The GOx–HRP–DNA origami-scaffold model centered in the cubic simulation volume corresponding to 100 nM system concentration. The box edge length is approximately 255.13 nm.

2.2.2 Simulation Configuration

We performed a set of simulations over 1.5 μs at enzyme and substrate concentrations of 10 nM to assess the variance of interenzyme substrate transfer probability with spatial configuration. The simulations of GOx and HRP on a planar scaffold were performed at several interenzyme distances, measured from the anchor points: 10, 15, 20, and 25 nm. Because the enzyme orientations *in situ* cannot be fixed, simulation subsets were systematically performed with varying enzyme orientations. Starting from rotational states corresponding to the active sites of GOx and HRP directly facing each other, each enzyme was successively rotated 90º clockwise for a total of 16 combinations of GOx and HRP orientations (Figure 2.3). Thus, for each interenzyme distance, 16 simulation subsets were performed, each with 3000 parallel replicates for a total of 48,000 parallel simulations for each interenzyme distance. Each 3000 replicate simulation was completed in approximately 3 - 5 hours on 16 CPU cores.

To assess the role of the scaffold in the efficiency of initial ligand association probability, all simulations were repeated with the planar scaffold and DNA–linker molecules excluded while retaining the colocalized positions and orientations of the enzymes. This allowed for isolation of the effect of the scaffold beyond the effect of colocalization.

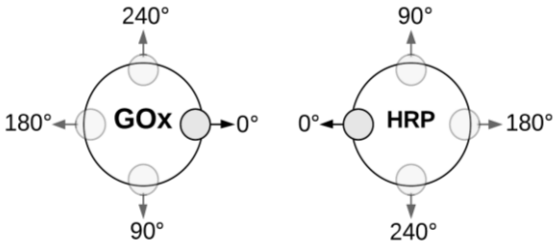

Figure 2.3. Top-down view of the rotational states for GOx and HRP varied in all subsets of simulation experiments. The grey circle represents the active site of each enzyme. Rotations were performed clockwise.

Long timescale (4 millisecond) simulations were performed to assess the efficacy of the scaffolded enzyme system relative to a solution of the substrate and an unmodified, unscaffolded target HRP enzyme. Because concentration effects are important at this timescale, simulations were performed with equal enzyme and substrate concentrations of 1, 10, and 100 nM.

Long timescale scaffolded simulations were performed with 16 orientation combinations in the same manner as described for the 1.5 us simulations. However, only a single interenzyme distance of 10 nm was simulated. Each enzyme orientation simulation set was replicated 3000 times, for 48,000 simulations per system concentration. Simulations were run in sets of 1000 replicates and completed in approximately 18 hours running on 16 CPU cores.

A model of a disorganized solution containing an unmodified HRP enzyme with a freely diffusing H_2O_2 solute was simulated. For computational practicality, the CG HRP model was fixed in the center of the simulation volume. The substrate of each of 14,000

replicates was placed randomly in the simulation and allowed to diffuse either until associated with HRP or the 4 millisecond time limit was reached. These simulations were completed in batches of 2000 replicates, requiring 17 - 20 hours running on 16 CPU cores.

2.2.3 Software Implementation and Optimization

GeomBD, our BD simulation software, was written in C++ with the CilkPlus extension for parallelization.[40] Because of the use of a generic short-range potential and omission of electrostatic interactions, the BD software is most appropriate for single-particle substrate diffusion investigations of the geometric effects of spatial configurations of large molecular systems. Though not used in the current work, electrostatic potentials are supported for single point charges centered in interaction spheres. To allow for electrostatic interaction of net neutral particles with a polarized charge distribution, rigid body partial charge embedding will be implemented for future investigation into system specific long range electrostatic guidance and its effect on diffusional association.

The software was based on the first-order Ermak-McCammon equation for propagating the Brownian trajectory.[51–53] The BD algorithm was implemented with a distance-dependent time step ranging from 0.05 to 50 ps, depending on a ligand's distance, r, from the closest point to the enzyme–DNA system:

$$r < 10\ nm, dt = 0.05\ ps$$

$$10\ nm < r < 25\ nm, dt = 0.5\ ps$$

$$25\ nm < r < 50\ nm, dt = 5.0\ ps$$

$$r > 50\ nm, dt = 50.0\ ps$$

All parallel replicates independently advanced through time. The input file allows for configuration of all simulation parameters as well as manipulation of structure coordinates for ease of spatial configuration. Planar potential fields, such as that used for the DNA origami scaffold in this work, are added to a model system through the configuration file. Multiple planar potentials can be added for more complex arrangements. Although not used in the current work, cylindrical potential fields can be included in model systems for investigation of nanotube-like systems.

2.3 Results and Discussion

2.3.1 Catalytic Enhancement on a Short Timescale

We performed a set of simulations over 1.5 μs to assess how differences in spatial configuration affect initial interenzyme substrate transfer probability of the GOx/HRP planar scaffold system. The model was sensitive to orientation and interenzyme distance and agreed with previously reported trends.[25,38] The highest probability of initial binding-site association occurred when the interenzyme distance was at its smallest value (10 nm) and the substrate release site of GOx directly faced HRP ($\theta_{GOx} = \theta_{HRP} = 0°$). Increasing the interenzyme distance affected the overall probability of substrate–binding-site association and induced an initial delay in the association distribution (Figure 2.4).

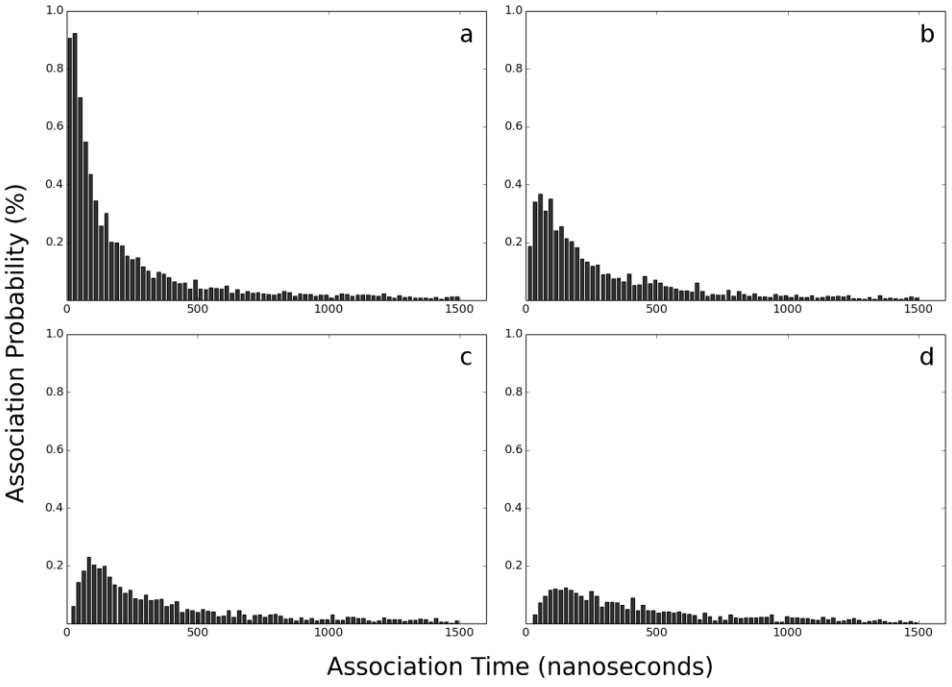

Figure 2.4: Binding time distribution for scaffolded GOx-HRP systems at different interenzyme distances. Each plot shows the percentage of ligands among 48,000 replicates (3000 x 16 orientational subsets) associated with the binding site of HRP within the first 1.5 μs with interenzyme distances of a) 10 nm, b) 15 nm, c) 20 nm, and d) 25 nm.

With rotation of the enzymes, the probability of substrate transfer decreased. However, because of the concave binding-site region of HRP, when only HRP was rotated 240 degrees, the binding site was partially oriented toward GOx, which led to behavior similar to unrotated HRP (Table 1). Analysis of the trajectories revealed that the general reduction in binding probability with rotation is due to the occlusion of the direct diffusion pathway from GOx to the binding site of HRP. The interenzyme distance and

orientation dependence on efficiency are system-specific and suggest that the geometric details of the enzymes used for spatial organization study are important when modeling potential production systems.[38,54]

	$\theta_{GOx} = 0°$	$\theta_{GOx} = 90°$	$\theta_{GOx} = 180°$	$\theta_{GOx} = 240°$
$\theta_{HRP} = 0°$	0.127	0.083	0.060	0.070
$\theta_{HRP} = 90°$	0.090	0.068	0.055	0.051
$\theta_{HRP} = 180°$	0.085	0.059	0.043	0.057
$\theta_{HRP} = 240°$	0.117	0.088	0.055	0.067

Table 2.1: Probabilities of direct stream binding for rotational orientations of GOx and HRP assembled on a planar scaffold at 10 nM enzyme concentration and 10 nm interenzyme distance.

2.3.2 Role of Planar Scaffold on Initial Enhancement

The impact of the DNA origami-scaffold on initial binding after release of the substrate was determined by comparing simulations of the colocalized and spatially organized GOx–HRP system assembled on a planar scaffold with an artificially colocalized GOx-HRP model without a scaffold model. The results are significant for all interenzyme distances tested and are summarized in Table 2.

Interenzyme Distance	**Colocalized Complex With Scaffold**	**Colocalized Complex Without Scaffold**	**Fold Enhancement Due to Scaffold**
10 nm	0.073	0.055	1.34
15 nm	0.049	0.034	1.45
20 nm	0.038	0.023	1.65
25 nm	0.030	0.018	1.68

Table 2.2: Fold enhancement in probability of transfer for the spatially organized GOx–HRP complex with and without a DNA origami scaffold. Simulations were performed at various interenzyme distances, averaged across enzyme orientations, over 1.5 μs.

The overall transfer probability decreased with increasing interenzyme distance, but enhancement accountable to the scaffold increased with interenzyme distance. The planar scaffold blocked the diffusion of substrate in the direction of the scaffold. The substrate-scaffold affinity temporarily retained the substrate, allowing for transient 2D diffusion of the substrate on the plane. This led to longer substrate dwell times in the vicinity of the target enzyme relative to the artificially colocalized configuration. For most substrate, this retention was followed by diffusion back to the volume occupied by the enzymes. Approximately 18% of the total substrate from the planar scaffold simulation sets interacted with the scaffold. By the end of these simulations, 1% of the total substrate remained associated with the scaffold. This value is likely underestimated due to a lack of hydrodynamic interactions between the ligand and scaffold as well as a lack of atomic features on the planar scaffold. Regardless, this finding suggests that the scaffold has the potential to increase the local concentration of the substrate, in agreement with previous theoretical and experimental conclusions.[20,34]

2.3.3 Catalytic Enhancement on a Long Timescale

The efficiency of the planar scaffold model system over time on the order of the turnover of GOx was compared to that with a model of an unorganized solution of the substrate and target enzyme, HRP. Simulations were run for 4 ms simulation time to determine the cumulative probability of binding-site association of the substrate to HRP for the two systems. Because system concentration is relevant at this timescale, simulations were run at three system concentrations. Results are summarized in Table 3. In the most dilute solution (1 nM), the initial enhanced probability of substrate transfer with the scaffold

system dominated the cumulative probability over the time period because of the relatively low rate of bulk substrate binding-site association in the unorganized solutions. The higher concentration solutions revealed nearly identical cumulative binding probabilities for the scaffolded and disorganized solution, with the binding distributions at the 100 nM concentration converging by 4 ms simulation time (Figure 2.5). Despite the convergence at higher concentration, the spike of initial binding probability within the first several microseconds after substrate release from GOx may still offer temporary kinetic advantages. Further optimization of initial interenzyme substrate transfer could result in improved kinetic advantages for spatially organized systems.

Concentration	**Complex With Scaffold**	**Disorganized HRP-H_2O_2 Solution**	**Spatial Organization Fold Enhancement**
1 nM	0.10	0.02	5
10 nM	0.25	0.19	1.32
100 nM	0.89	0.87	1.02

Table 2.3: Enhancement in transfer probability with the spatially organized enzyme complex system and disorganized solution of HRP and substrate over long timescales (4 ms). Cumulative binding-site association probabilities are compared at various system concentrations.

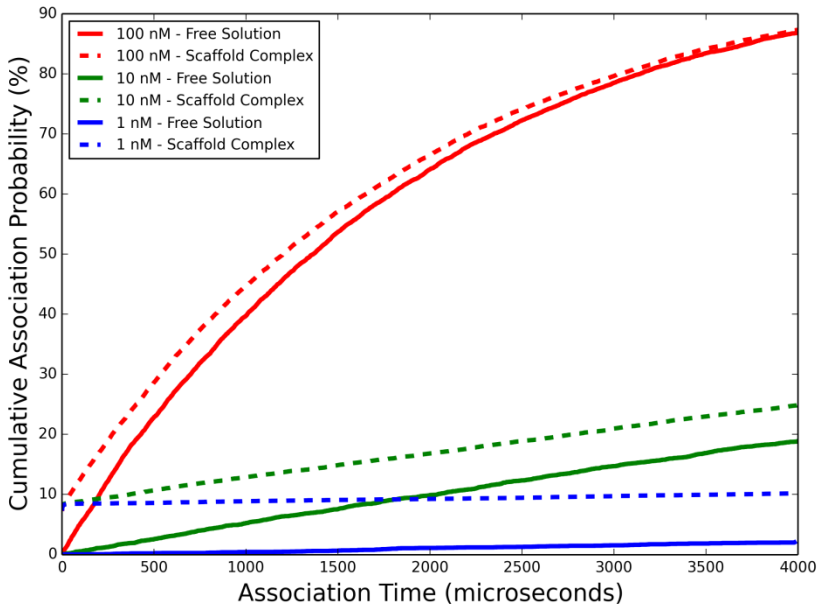

Figure 2.5: Probability of cumulative binding-site association comparison over 4 ms for spatially organized system and disorganized solutions at various system concentrations. For the scaffolded enzyme complex, the initial spike in binding-site association within the first microseconds of substrate diffusion after release from GOx may suggest a kinetic advantage of reaction-coupled enzyme systems assembled on a scaffold.

2.4 Conclusions

We used coarse grain Brownian dynamics simulations with the GeomBD package to assess enhancement in catalysis with spatial organization of GOx and HRP assembled on a planar scaffold. Our method builds on the findings of previous computational studies of interenzyme substrate transfer by including more system specific geometric detail and assessing the separate impacts of colocalization and assembly on a planar scaffold.

Colocalization of the enzymes dominated the enhancement of the planar scaffold-bound spatially organized system. Specifically, interenzyme distance and rotational orientation were the primary factors affecting the efficiency of our model enzyme system on a scaffold surface. The diffusive barrier induced by the scaffold played a significant but secondary role in the enhanced transfer. Maximum enhancement in intermediate transfer of the substrate was achieved for the enzyme system assembled on a planar scaffold with small interenzyme separation and enzyme binding sites oriented toward one another. Our results from varying enzyme orientation suggest that the geometric details of the enzymes used for spatial organization studies must be considered. As suggested, the interenzyme distance dependence on efficiency is also specific to the system geometry.[38,54]

With our model, a low system concentration is required to retain the initial benefit of spatial organization throughout millisecond timescales. At a higher concentration (100 nM), the cumulative probability of HRP–substrate association in the disorganized solutions competed with that the spatially organized system over 4 milliseconds.

More complex spatial configurations that include compartmentalization of enzyme systems, which appear to offer additional benefits to efficiency beyond that of a planar scaffold, will be explored computationally in the future.[55]

2.5 References

(1) Kahn, R. A.; Fahrendorf, T.; Halkier, B. A.; Møller, B. L. Substrate Specificity of the Cytochrome P450 Enzymes CYP79A1 and CYP71E1 Involved in the Biosynthesis of the Cyanogenic Glucoside Dhurrin in Sorghum Bicolor (L.) Moench. *Arch. Biochem. Biophys.* **1999**, *363*, 9–18.

(2) Winkel, B. S. J. Metabolic Channeling in Plants. *Annu. Rev. Plant Biol.* **2004**, *55*, 85–107.

(3) Miles, E. W. The Molecular Basis of Substrate Channeling. *J. Biol. Chem.* **1999**, *274*, 12193–12196.

(4) Kim, Y. H.; Kwon, T. K.; Park, S.; Seo, H. S.; Cheong, J. J.; Kim, C. H.; Kim, J. K.; Lee, J. S.; Choi, Y. D. Trehalose Synthesis by Sequential Reactions of Recombinant Maltooligosyltrehalose Synthase and Maltooligosyltrehalose Trehalohydrolase from Brevibacterium Helvolum. *Appl. Environ. Microbiol.* **2000**, *66*, 4620–4624.

(5) Jandt, U.; You, C.; Zhang, Y. H.-P.; Zeng, A.-P. Compartmentalization and Metabolic Channeling for Multienzymatic Biosynthesis: Practical Strategies and Modeling Approaches. *Adv. Biochem. Eng. Biotechnol.* **2013**, *137*, 41–65.

(6) Burack, W. R.; Shaw, A. S. Signal Transduction: Hanging on a Scaffold. *Curr. Opin. Cell Biol.* **2000**, *12*, 211–216.

(7) Savage, D. F.; Afonso, B.; Chen, A. H.; Silver, P. A. Spatially Ordered Dynamics of the Bacterial Carbon Fixation Machinery. *Science* **2010**, *327*, 1258–1261.

(8) Singer, S. J. The Molecular Organization of Membranes. *Annu. Rev. Biochem.* **1974**, *43*, 805–833.

(9) Adam, G.; Delbrück, M. Reduction of Dimensionality in Biological Diffusion Processes. In *Structural chemistry and molecular biology*; Rich, A.; Davidson, N., Eds.; W. H. Freeman and Co., 1968; pp. 198–215.

(10) Axelrod, D.; Wang, M. D. Reduction-of-Dimensionality Kinetics at Reaction-Limited Cell Surface Receptors. *Biophys. J.* **1994**, *66*, 588–600.

(11) Hrazdina, G.; Jensen, R. A. Spatial Organization of Enzymes in Plant Metabolic Pathways. *Annu. Rev. Plant Physiol. Plant Mol. Biol. 43*, 241–267.

(12) Bülow, L.; Ljungcrantz, P.; Mosbach, K. Preparation of a Soluble Bifunctional Enzyme by Gene Fusion. *Bio/Technology* **1985**, *3*, 821–823.

(13) Conrado, R. J.; Varner, J. D.; DeLisa, M. P. Engineering the Spatial Organization of Metabolic Enzymes: Mimicking Nature's Synergy. *Curr. Opin. Biotechnol.* **2008**, *19*, 492–499.

(14) Ljungcrantz, P.; Carlsson, H.; Mansson, M. O.; Buckel, P.; Mosbach, K.; Buelow, L. Construction of an Artificial Bifunctional Enzyme, .beta.-Galactosidase/galactose Dehydrogenase, Exhibiting Efficient Galactose Channeling. *Biochemistry* **1989**, *28*, 8786–8792.

(15) Riedel, K.; Bronnenmeier, K. Intramolecular Synergism in an Engineered Exo-Endo-1,4-Beta-Glucanase Fusion Protein. *Mol. Microbiol.* **1998**, *28*, 767–775.

(16) Mao, Q.; Schunk, T.; Gerber, B.; Erni, B. A String of Enzymes, Purification and Characterization of a Fusion Protein Comprising the Four Subunits of the Glucose Phosphotransferase System of Escherichia Coli. *J. Biol. Chem.* **1995**, *270*, 18295–18300.

(17) De Pascale, D.; Di Lernia, I.; Sasso, M. P.; Furia, A.; De Rosa, M.; Rossi, M. A Novel Thermophilic Fusion Enzyme for Trehalose Production. *Extremophiles* **2002**, *6*, 463–468.

(18) Orita, I.; Sakamoto, N.; Kato, N.; Yurimoto, H.; Sakai, Y. Bifunctional Enzyme Fusion of 3-Hexulose-6-Phosphate Synthase and 6-Phospho-3-Hexuloisomerase. *Appl. Microbiol. Biotechnol.* **2007**, *76*, 439–445.

(19) Levasseur, A.; Navarro, D.; Punt, P. J.; Belaïch, J.-P.; Asther, M.; Record, E. Construction of Engineered Bifunctional Enzymes and Their Overproduction in Aspergillus Niger for Improved Enzymatic Tools to Degrade Agricultural by-Products. *Appl. Environ. Microbiol.* **2005**, *71*, 8132–8140.

(20) Wilner, O. I.; Weizmann, Y.; Gill, R.; Lioubashevski, O.; Freeman, R.; Willner, I. Enzyme Cascades Activated on Topologically Programmed DNA Scaffolds. *Nat. Nanotechnol.* **2009**, *4*, 249–254.

(21) Delebecque, C. J.; Lindner, A. B.; Silver, P. A.; Aldaye, F. A. Organization of Intracellular Reactions with Rationally Designed RNA Assemblies. *Science* **2011**, *333*, 470–474.

(22) Dueber, J. E.; Wu, G. C.; Malmirchegini, G. R.; Moon, T. S.; Petzold, C. J.; Ullal, A. V; Prather, K. L. J.; Keasling, J. D. Synthetic Protein Scaffolds Provide Modular Control over Metabolic Flux. *Nat. Biotechnol.* **2009**, *27*, 753–759.

(23) Mosbach, K.; Mattiasson, B. Matrix-Bound Enzymes. II. Studies on a Matrix-Bound Two-Enzyme-System. *Acta Chem. Scand.* **1970**, *24*, 2093–2100.

(24) Mingardon, F.; Chanal, A.; López-Contreras, A. M.; Dray, C.; Bayer, E. A.; Fierobe, H.-P. Incorporation of Fungal Cellulases in Bacterial Minicellulosomes Yields Viable, Synergistically Acting Cellulolytic Complexes. *Appl. Environ. Microbiol.* **2007**, *73*, 3822–3832.

(25) Fu, J.; Liu, M.; Liu, Y.; Woodbury, N. W.; Yan, H. Interenzyme Substrate Diffusion for an Enzyme Cascade Organized on Spatially Addressable DNA Nanostructures. *J. Am. Chem. Soc.* **2012**, *134*, 5516–5519.

(26) Rothemund, P. W. K. Folding DNA to Create Nanoscale Shapes and Patterns. *Nature* **2006**, *440*, 297–302.

(27) Seeman, N. C. Nanomaterials Based on DNA. *Annu. Rev. Biochem.* **2010**, *79*, 65–87.

(28) Lin, C.; Liu, Y.; Yan, H. Designer DNA Nanoarchitectures. *Biochemistry* **2009**, *48*, 1663–1674.

(29) Michelotti, N.; Johnson-Buck, A.; Manzo, A. J.; Walter, N. G. Beyond DNA Origami: The Unfolding Prospects of Nucleic Acid Nanotechnology. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. 4*, 139–152.

(30) Winfree, E.; Liu, F.; Wenzler, L. A.; Seeman, N. C. Design and Self-Assembly of Two-Dimensional DNA Crystals. *Nature* **1998**, *394*, 539–544.

(31) Veitch, N. C. Horseradish Peroxidase: A Modern View of a Classic Enzyme. *Phytochemistry* **2004**, *65*, 249–259.

(32) Raba, J.; Mottola, H. A. Glucose Oxidase as an Analytical Reagent. *Crit. Rev. Anal. Chem.* **1995**, *25*, 1–42.

(33) Idan, O.; Hess, H. Diffusive Transport Phenomena in Artificial Enzyme Cascades on Scaffolds. *Nat. Nanotechnol.* **2012**, *7*, 769–770.

(34) Idan, O.; Hess, H. Origins of Activity Enhancement in Enzyme Cascades on Scaffolds. *ACS Nano* **2013**, *7*, 8658–8665.

(35) Takahashi, K.; Arjunan, S. N. V.; Tomita, M. Space in Systems Biology of Signaling Pathways--towards Intracellular Molecular Crowding in Silico. *FEBS Lett.* **2005**, *579*, 1783–1788.

(36) Ridgway, D.; Broderick, G.; Lopez-Campistrous, A.; Ru'aini, M.; Winter, P.; Hamilton, M.; Boulanger, P.; Kovalenko, A.; Ellison, M. J. Coarse-Grained Molecular Simulation of Diffusion and Reaction Kinetics in a Crowded Virtual Cytoplasm. *Biophys. J.* **2008**, *94*, 3748–3759.

(37) Klann, M. T.; Lapin, A.; Reuss, M. Stochastic Simulation of Signal Transduction: Impact of the Cellular Architecture on Diffusion. *Biophys. J.* **2009**, *96*, 5122–5129.

(38) Bauler, P.; Huber, G.; Leyh, T.; McCammon, J. A. Channeling by Proximity: The Catalytic Advantages of Active Site Colocalization Using Brownian Dynamics. *J. Phys. Chem. Lett.* **2010**, *1*, 1332–1335.

(39) Klann, M.; Koeppl, H. Spatial Simulations in Systems Biology: From Molecules to Cells. *Int. J. Mol. Sci.* **2012**, *13*, 7798–7827.

(40) Blumofe, R. D.; Joerg, C. F.; Kuszmaul, B. C.; Leiserson, C. E.; Randall, K. H.; Zhou, Y. Cilk. *ACM SIGPLAN Not.* **1995**, *30*, 207–216.

(41) Frederick, K. R.; Tung, J.; Emerick, R. S.; Masiarz, F. R.; Chamberlain, S. H.; Vasavada, A.; Rosenberg, S.; Chakraborty, S.; Schopfer, L. M.; Schopter, L. M. Glucose Oxidase from Aspergillus Niger. Cloning, Gene Sequence, Secretion from Saccharomyces Cerevisiae and Kinetic Analysis of a Yeast-Derived Enzyme. *J. Biol. Chem.* **1990**, *265*, 3793–3802.

(42) Bright, H. J.; Gibson, Q. H. The Oxidation of 1-Deuterated Glucose by Glucose Oxidase. *J. Biol. Chem.* **1967**, *242*, 994–1003.

(43) Yamanaka, S. A.; Nishida, F.; Ellerby, L. M.; Nishida, C. R.; Dunn, B.; Valentine, J. S.; Zink, J. I. Enzymatic Activity of Glucose Oxidase Encapsulated in Transparent Glass by the Sol-Gel Method. *Chem. Mater.* **1992**, *4*, 495–497.

(44) Gajhede, M.; Schuller, D. J.; Henriksen, A.; Smith, A. T.; Poulos, T. L. Crystal Structure of Horseradish Peroxidase C at 2.15 A Resolution. *Nat. Struct. Biol.* **1997**, *4*, 1032–1038.

(45) Wohlfahrt, G.; Witt, S.; Hendle, J.; Schomburg, D.; Kalisz, H. M.; Hecht, H. J. 1.8 and 1.9 A Resolution Structures of the Penicillium Amagasakiense and Aspergillus Niger Glucose Oxidases as a Basis for Modelling Substrate Complexes. *Acta Crystallogr. D. Biol. Crystallogr.* **1999**, *55*, 969–977.

(46) Reva, B. A.; Finkelstein, A. V.; Sanner, M. F.; Olson, A. J. Residue-Residue Mean-Force Potentials for Protein Structure Recognition. *Protein Eng. Des. Sel.* **1997**, *10*, 865–876.

(47) Shen, T.; Wong, C. F.; McCammon, J. A. Atomistic Brownian Dynamics Simulation of Peptide Phosphorylation. *J. Am. Chem. Soc.* **2001**, *123*, 9107–9111.

(48) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

(49) Kang, M.; Roberts, C.; Cheng, Y.; Chang, C. A. Gating and Intermolecular Interactions in Ligand-Protein Association: Coarse-Grained Modeling of HIV-1 Protease. *J. Chem. Theory Comput.* **2011**, *7*, 3438–3446.

(50) Henzler, T. Transport and Metabolic Degradation of Hydrogen Peroxide in Chara Corallina: Model Calculations and Measurements with the Pressure Probe Suggest Transport of H2O2 across Water Channels. *J. Exp. Bot.* **2000**, *51*, 2053–2066.

(51) Ermak, D. L.; McCammon, J. A. Brownian Dynamics with Hydrodynamic Interactions. *J. Chem. Phys.* **1978**, *69*, 1352.

(52) Długosz, M.; Zieliński, P.; Trylska, J. Brownian Dynamics Simulations on CPU and GPU with BD_BOX. *J. Comput. Chem.* **2011**, *32*, 2734–2744.

(53) Greives, N.; Zhou, H.-X. BDflex: A Method for Efficient Treatment of Molecular Flexibility in Calculating Protein-Ligand Binding Rate Constants from Brownian Dynamics Simulations. *J. Chem. Phys.* **2012**, *137*, 135105.

(54) Eun, C.; Kekenes-Huskey, P. M.; Metzger, V. T.; McCammon, J. A. A Model Study of Sequential Enzyme Reactions and Electrostatic Channeling. *J. Chem. Phys.* **2014**, *140*, 105101.

(55) Fu, Y.; Zeng, D.; Chao, J.; Jin, Y.; Zhang, Z.; Liu, H.; Li, D.; Ma, H.; Huang, Q.; Gothelf, K. V; Fan, C. Single-Step Rapid Assembly of DNA Origami Nanostructures for Addressable Nanoscale Bioreactors. *J. Am. Chem. Soc.* **2013**, *135*, 696–702.

Chapter 3: Colocalization in Multi-enzyme Bioreactor Scaffolds: Computational Assessment of Planar and Tubular Geometry Scaffolds

3.1 Introduction

Nanostructures have become a promising technology in chemical and biological engineering for the construction of catalytic reactors that produce desirable chemical species. Engineering of enzymes into synthetic spatially organized nanostructures has been shown to increase the efficiency of enzyme reaction cascades relative to disorganized solutions.[1–7] These systems are inspired by the high specificity and yield of biological metabolic channeling, where colocalization of the enzymes involved in a cascade improves the rate of successful intermediate substrate diffusion between enzymes.[8–11] It is thus desirable to produce nanostructures that provide an environment that optimizes the intermediate diffusion processes between a series enzymes of interest *in vitro*.

There have been many approaches to colocalization of enzymes for biocatalytic applications, such as synthetic fusion protein synthesis,[12–18] immobilization of enzyme aggregates,[1,2,6] and adhesion to nanoparticles[19] and protein scaffolds.[11,20] Recent advances in the synthesis of self-assembling DNA nanostructures, called DNA origami, have presented new avenues for the controlled design of enzyme nanostructures.[21–23] DNA origami technology allows for the construction of multidimensional self-assembling nanostructures with geometries including sheets, cubes, spheres, tubes, and more complex arrangements.[22,24–30] In addition to controlled geometry, accurate placement of molecular species on DNA origami scaffolds is "programmable" in the synthetic

procedure, utilizing complimentary single-stranded DNA spatial addressing to anchor specific molecules to specific locations with nanometer resolution.[4,31–35] As such, DNA origami represent an attractive scaffold for the construction of catalytic nanostructures with designed spatial organization.

Reaction coupled enzyme pairs are commonly used as model systems to explore catalytic nanostructure designs due to the dependence on the transfer of an intermediate for joint enzyme turnover. One such model system is the glucose oxidase (GOx) and horseradish peroxidase (HRP) enzyme pair, which relies on the transfer of hydrogen peroxide (H_2O_2) from GOx to HRP for mutual turnover to occur. Thus it is an appropriate system for assessing the effect of spatial organization due to the observable change in HRP product concentration over a given time span. Jinglin Fu et al anchored a GOx-HRP enzyme pair on a planar DNA origami scaffold, realizing a kinetic advantage relative to the free solution of the enzymes.[4] More recently, Yanming Fu et al synthesized a similar system, anchoring a GOx-HRP enzyme pair within the interior of a tubular origami.[36] While not directly quantified, the results of the nanotube system show a distinct increase in turnover efficiency for the tubular origami scaffold geometry over the planar geometry. The present study focuses on this DNA origami nanotube system, with comparisons drawn to our previous work characterizing the planar geometry scaffold GOx-HRP system (Chapter 2).

Computational modeling and simulations with the GeomBD package are used to characterize the GOx-HRP enzyme system assembled in a nanotube, on a planar scaffold, and free in a disorganized solution. The effects of nanotube geometry, enzyme

placement, and scaffold-substrate affinity are investigated for their effect on successful intermediate diffusional transfer from GOx to HRP. A new version of the GeomBD package, featuring partial charge embedding, is used to investigate the system through coarse grain (CG) Brownian dynamics (BD). Our previously reported results of the planar geometry DNA origami assembly of GOx and HRP (Chapter 2) are updated with electrostatic interactions enabled, and are presented here for comparison to the nanotube system. In addition, a fully diffusive free solution of hydrogen peroxide and its target enzyme, HRP, were modeled and simulated to assess the baseline association probability for comparison to the scaffolded systems.

3.2 Methods

3.2.1 Overview

Diffusion studies using coarse grained (CG) Brownian dynamics (BD) simulations were performed under various conditions to assess the factors that affect the probabilities of successful intermediate substrate transfer of H_2O_2 between GOx and HRP under various conditions. Successful substrate transfer is considered if a substrate diffuses to and interacts with the active site of HRP. The number of simulations resulting in successful transfer is divided by the total number of simulations performed to arrive at a final association probability.

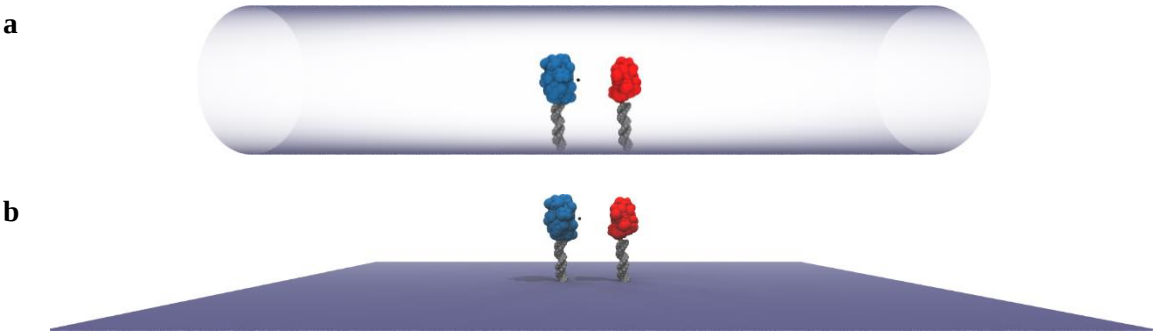

Figure 3.1. The GOx/HRP pair assembled (a) within an implicit nanotube model and (b) on a planar scaffold. The GOx and HRP enzymes, colored in blue and red respectively, and the DNA linker strands, colored in grey, are modeled with a coarse grain scheme. The substrate is marked by a black sphere near the active site of GOx.

The GOx/HRP pair was modeled as an organized assembly within a nanotube scaffold (Figure 3.1a), as an assembly on a planar scaffold (Figure 3.1b), and freely diffusing in a disorganized solution. The enzymes, enzyme-scaffold DNA strand linkers, and substrate were modeled using a coarse grain molecular model, while the scaffolds were modeled as implicit structures defined by simple geometry and potential energy functions. Simulations were performed with a new version of the GeomBD package. Rigid body force propagation has been included allowing for multi-bead substrate, and in this case, the embedding of atomic partial charges in a single-bead representation of a molecule. Neutral molecules abstracted into a single bead can now consider electrostatics and interact with other charged species. Our previously reported results of the GOx-HRP pair

assembled on a planar scaffold (Chapter 2) have been updated with electrostatic interactions enabled using this new GeomBD version.

3.2.2 Structure and Preparation

In vitro, the structure of the tubular scaffold GOx-HRP molecular system consists of a DNA origami nanotube with DNA strands linking each of GOx and HRP to the inner nanotube surface. Computationally, the nanotube scaffold was modeled implicitly as a cylindrical force field, while the enzymes and 10 nm double-stranded DNA enzyme-scaffold linkages were represented as explicit CG structures (Figure 3.1a). The planar scaffold GOx-HRP system was similarly constructed, using an implicit planar force field (Figure 3.1b).

The HRP isozyme C and GOx atomic structures were obtained from the protein data bank (PDB: 1ATJ and 1GPE, respectively).[37,38] CG models were generated from the all-atom PDB structures using the cgProtein utility in the geomBD package. In this CG model, each protein residue is represented as a single sphere, centered on the alpha carbon of the amino acid. The radii were derived from the average radius of gyration about the alpha carbon for each standard amino acid type from hundreds of unique molecular dynamics simulations.[39,40] Formal charges were assigned to the center of CG beads corresponding to charged amino acid residues. DNA strands linking the scaffold and enzymes were created from all-atom representations of a randomly sequenced DNA strand created using the NAB AMBER software interface.[41] For each nucleoside base in this CG DNA representation, two beads were positioned to represent the base and the associated sugar–phosphate backbone atoms. Bead radii for the DNA CG model were calculated as the

radius of gyration of the abstracted atoms about their average position, accounting for their atomic radii. Formal charges were placed on the phosphate beads. The H_2O_2 substrate was abstracted into a single particle. Utilizing improvements to the GeomBD package, atomic partial charges for the atoms in H_2O_2 were embedded in the substrate bead as a rigid body. Electrostatic forces on each point charge are propagated to translating the CG particle and internally rotating the point charges. This allows the neutral, but polar, H_2O_2 substrate to interact electrostatically with the proteins, linkers, and scaffold.

The scaffolds in both the planar and tubular geometries were modeled implicitly. Because the structures are highly abstracted, our models are generalizable to generic planar and tubular scaffolds. The structures were defined using geometric expressions coupled with polar and non-polar potential energy functions that dictate the interactions with the substrate. The geometry of the planar scaffold was 80 nanometers by 60 nanometers, matching the dimensions of the experimental structure.[4] The same dimensions were used for the "rolled" nanotube scaffold, resulting in a length of 80 nanometers and diameter of approximately 20 nanometers.[36]

3.2.3 Model Parameters

The substrate, H_2O_2, was assigned an experimentally determined diffusion coefficient of 1 x 10^{-3} nm^2/s.[42] For simulations that included a scaffold model, no diffusion coefficients were assigned for the enzymes and scaffold, as they were held fixed in space. However, for the simulation of a free solution of HRP and hydrogen peroxide, the HRP enzyme was assigned a diffusion coefficient according to the Stokes-Einstein equation. The radius of gyration of HRP was used as the radius for this calculation.

Intermolecular interactions include terms for both van der Waals and electrostatic interactions. Van der Waals interactions were modeled using a Lennard-Jones like potential:

$$E = \varepsilon\, a\left[\left(\frac{\sigma}{l}\right)^8 - \frac{3}{2}\left(\frac{\sigma}{l}\right)^6\right] \qquad \text{eq. 3.1}$$

where l is the distance between two interacting particles, σ is the sum of the radii of the two interacting spheres, ε is the well depth of the potential, and a is a proportionality constant scaling the function to match the desired well depth. This function is similar to the 12-6 Lennard Jones equation, but accounts for the structural abstraction of the CG model by softening the repulsive term, allowing for more particle interpenetration. The ligand-enzyme and ligand-linker interactions were assigned well depths of 0.267 kcal/mol, matching the previously parameterized form of the equation.[43] The non-polar ligand-scaffold interactions were also defined by Equation 3.1. The potential radiates equally from the area and edges of the planar and cylindrical geometries. The well depth of this function was varied through ε to assess the impact of affinity on substrate transfer.

Electrostatic interactions were modeled with a Coulombic potential, with a distance dependent dielectric of 15, of the form:

$$E = k_e \frac{q_i q_j}{l^2 15} \qquad \text{eq. 3.2}$$

where q_i and q_j are the charges assigned to interacting species *i* and *j*, *l* is the measured distance between the two atoms, and k_e is Coulomb's constant ($k_e = (4\pi\epsilon_0)^{-1}$, where ϵ_0 is the electric permittivity of free space). The distance dependent dielectric is utilized to account for electrostatic screening.[44] Electrostatic forces exerted on a partial charge embedded within a single CG bead were propagated to angular and linear forces on the bead. This results in translation of the bead and internal rotation of the partial charge distribution. The polar ligand-scaffold interactions were defined using Equation 3.2, where the charge of the scaffold was set to -1. Like the non-polar potential function, the polar potential radiates equally from the area and edges of the geometries.

3.2.4 Simulation Configurations

Simulations were performed for 10 microseconds to determine the probability of successful substrate transfer from the active site of GOx to that of HRP in several spatially organized configurations. All simulations were performed within a cubic simulation volume corresponding to an enzyme and substrate concentration of 10 nM (edge length equal to 549.65 nm). System temperature was set to 298K. Solvent viscosity was set to that of water.

Free Solution Model

A fully diffusive free solution of a single HRP enzyme and hydrogen peroxide substrate were simulated using rigid body BD to assess the association probability of H_2O_2 to HRP over the course of 10 microseconds. Each molecule was randomly placed in the simulation volume. The simulations were performed 16,000 times. Each simulation took approximately 8 minutes to complete on 6 CPU cores.

Artificially Colocalized Enzyme Model

A series of simulations was performed with the GOx and HRP enzymes colocalized and artificially fixed in space with defined interenzyme distances and orientations. These simulations sets were performed to determine the effect of colocalization on intermediate substrate transfer relative to a free solution without the influence of a scaffold. The interenzyme distance was set to 10, 20, and 30 nanometers to assess the effect of interenzyme distance on interenzyme substrate transfer probability. Enzyme orientation was not considered as an explicit simulation parameter because the orientations of the enzymes are not fixed *in vitro*. Instead, each enzyme was successively rotated 90 degrees, creating 16 orientation states for each simulation configuration. Each simulation configuration was replicated a total of 16,000 times. Thus, for the GOx-HRP colocalized enzyme pair, 16 orientation combinations were simulated with 1000 replicates, resulting in a total of 16,000 replicates for each configuration.

Planar Scaffold Model

We previously reported the results of simulations assessing the interenzyme substrate transfer probability of hydrogen peroxide from GOx to HRP while assembled on a planar scaffold (Chapter 2). These simulations had been performed with the first generation of the GeomBD software, which did not support partial charge embedding for electrostatic interactions of net neutral molecules, like the substrate in these studies. With the introduction of this feature, we revisited this system to assess the effect of electrostatic interactions on the interenzyme substrate transfer probability. The GOx-HRP pair was assembled on a planar scaffold to determine the substrate transfer probability relative to the free solution and colocalized enzyme pair. The planar scaffold geometry was defined as 80 nanometers by 60 nanometers with interaction potentials defining a 2 nanometer thickness. The implicit scaffold was centered in the simulation volume, with the 80 nanometer length running along the X-axis. The enzymes were assembled on this scaffold, approximately 8 nanometers above the plane, with their separation running along the X-axis. Explicit CG DNA linkers were positioned to span the gap between the enzymes and the scaffold. The interenzyme distance was simulated with values of 10, 20, and 30 nanometers to determine its impact on substrate transfer. Like the colocalized enzyme pair model, each enzyme was rotated 90 degrees, creating 16 orientation states for each simulation configuration. Thus, for the GOx/HRP enzyme pair assembled with the planar scaffold model, 16 orientation combinations were simulated with 1000 replicates each, resulting in a total of 16,000 replicates for each configuration.

Each 1000 replicate simulation required 0.75 to 2.5 hours on 16 CPU cores. Simulations with a strong ligand-scaffold affinity proceeded more slowly due to the increased residence time on the scaffold, restricting the time step to its smallest value.

Tubular Scaffold Model

The GOx-HRP pair was assembled in a tubular scaffold to determine the substrate transfer probability relative to the free solution, colocalized enzyme pair, and the enzyme pair assembled on a planar scaffold. The tubular scaffold had the same geometry as the planar scaffold, but was "rolled" into a tube geometry. The tube had an 80 nanometer length and 60 nanometer circumference resulting in a 20 nanometer diameter. The scaffold model was centered at the origin with the length of the tube running along the X-axis. The enzymes and DNA linkers were positioned within the tubular scaffold, approximately 8 nanometers above the bottom of the inside of the tube, with their separation running along the X-axis. Explicit CG DNA linkers were positioned to span the gap between the enzymes and the scaffold. These nanotube scaffold simulations were run for the combinations of the following parameters: nanotube radii of 10nm, 15nm, and 20nm; nanotube affinity ranging from a well depth of 0.267 to 2.5 kcal/mol; enzyme positions with interenzyme distances of 10, 20, and 30nm.

Sixteen orientation combinations were simulated with 1000 replicates each, resulting in a total of 16,000 replicates for each configuration. Each 1000 replicate simulation required 1.5 to 3 hours on 16 CPU cores. Simulations with a strong ligand-scaffold affinity proceeded more slowly due to the increased residence time on the scaffold, restricting the time step to its smallest value.

Tubular Scaffold with Multiple Target HRP Enzymes

A novel system was constructed with two HRP target enzymes assembled with a single GOx enzyme inside a nanotube. In this system, the HRP/GOx/HRP trio, the GOx enzyme was mounted in the center, lengthwise, of the nanotube. Both HRP enzymes were placed on either side of the GOx enzyme, with a specified interenzyme distance. The effect of multiple target enzymes was assessed for its ability to impact association probability. The interenzyme distance was varied with values of 10, 20, and 30nm. In this system, 64 states of enzyme orientation were simulated each with 250 replicates for a total of 16,000 simulation replicates. Each 250 replicate simulation required 1.5 to 3.5 hours on 16 CPU cores.

3.3 Results and Discussion

Simulations of the GOx/HRP system were performed with the enzymes organized within a nanotube scaffold, on a planar scaffold, and artificially colocalized with no scaffold model. As well, a model free solution of HRP and a hydrogen peroxide substrate were simulated for comparison to all spatially organized and scaffolded simulations.

3.3.1 Free Solution

Simulations of the substrate, hydrogen peroxide, and the GOx and HRP enzymes were simulated to assess the probability of substrate-HRP association in a free solution. Both HRP and a substrate were randomly placed in a cubic simulation volume corresponding to substrate and enzyme concentrations of 10 nM. Both the enzyme and the substrate diffused freely throughout the course of 16,000 simulations with a simulation length of

10 microseconds. The probability of substrate association was 0.42% with an average binding time of 5.6 microseconds.

3.3.2 Artificially Colocalized Enzymes

The GOx-HRP enzyme pair was colocalized artificially in a solution of the same volume and concentration as the free solution simulation. Colocalization enhanced the interenzyme substrate transfer by 46.5, 20, and 11 fold for interenzyme distances of 10, 20, and 30 nanometers relative to the free solution over the 10 microsecond simulation duration. (Figure 3.2).

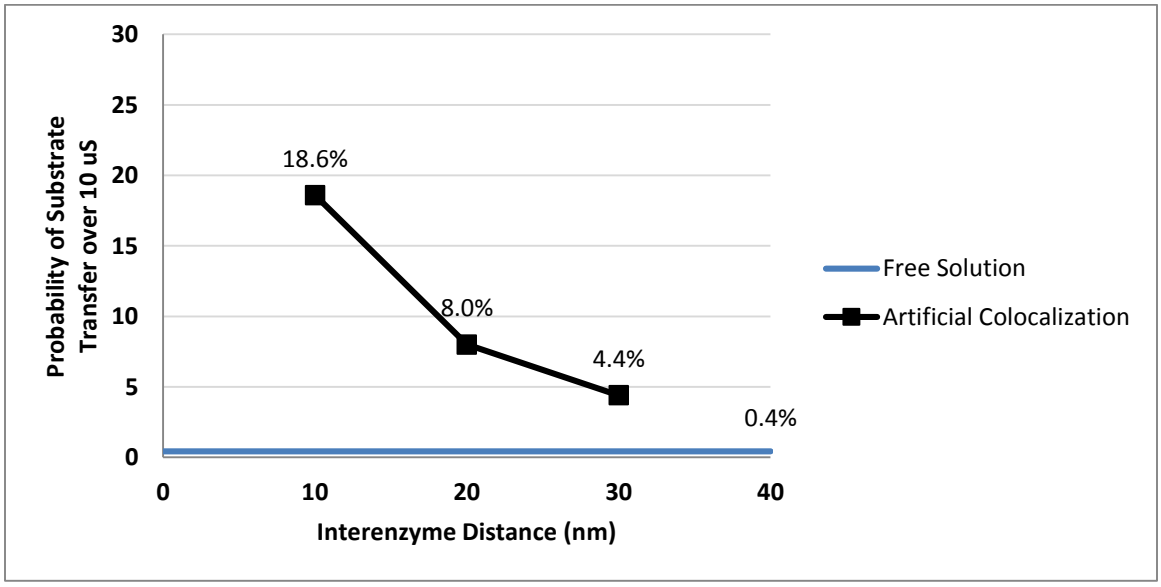

Figure 3.2. Substrate transfer probability, over 10 microseconds, for the GOx/HRP pair artificially colocalized by fixation in the solution volume. Simulation result of a free solution of GOx, HRP, and the substrate is presented for comparison.

3.3.3 Planar Geometry Scaffold

The GOx-HRP enzyme pair was spatially organized on a planar scaffold under the same conditions as the free solution and artificially colocalized systems. The scaffold imparted additional enhancement in substrate transfer. Relative to the artificially colocalized system, the substrate transfer probability increased by 1.2, 1.4, and 1.5 fold for interenzyme distances of 10, 20, and 30 nanometers. Relative to the free solution, the enzymes on a planar scaffold increased 54, 27, and 16 fold for interenzyme distances of 10, 20, and 30 nanometers (Figure 3.3).

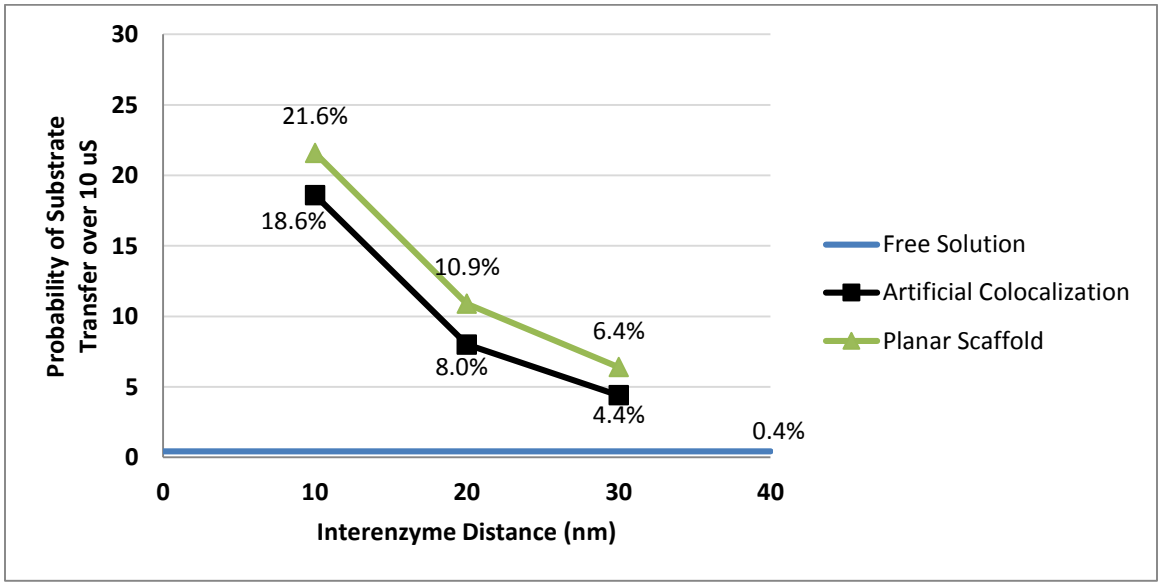

Figure 3.3. Substrate transfer probability, over 10 microseconds, for the GOx/HRP pair assembled on a planar scaffold. Simulation results of the artificially colocalized GOx/HRP pair are presented for comparison. The scaffold-substrate affinity well depths were set to the default value of 0.267 kcal/mol for the planar scaffold simulations.

As previously stated, the GOx-HRP enzyme pair assembled on a planar scaffold was previously characterized with a version of GeomBD that did not support partial charge embedding, and thus did not include electrostatic interactions in the simulations. These results, performed over the course of 1.5 microseconds, are presented in Figure 3.4 for comparison to the new results with electrostatic interactions included. The new results were performed over the course of 10 microseconds, but only the data for the first 1.5 microseconds is presented in Figure 3.4 for direct comparison to the previous results.

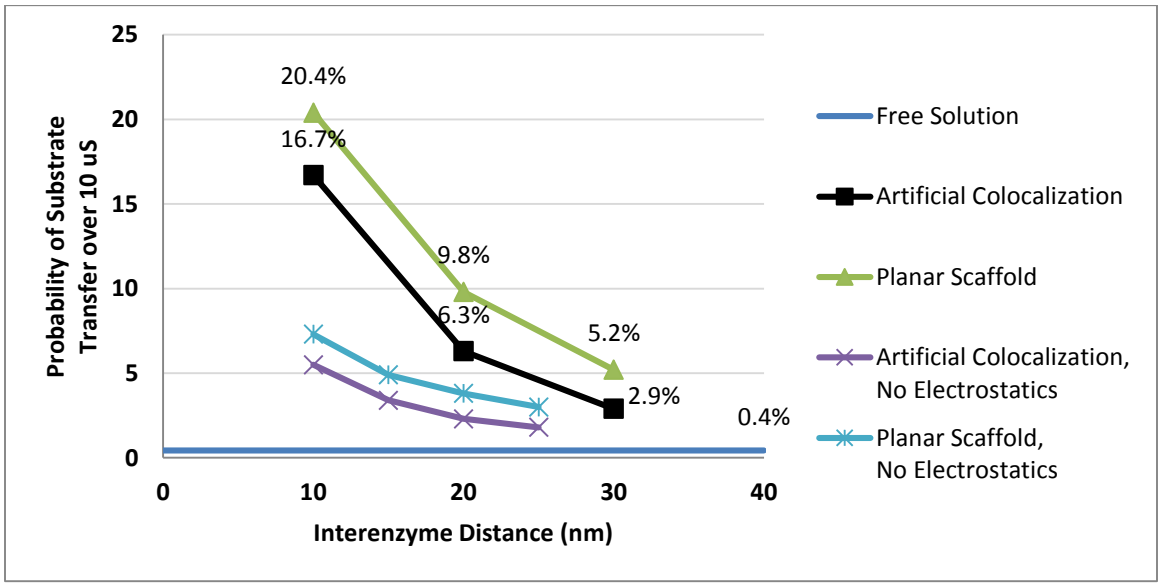

Figure 3.4. Data comparison between previous and current simulation results of the substrate transfer probability, over 1.5 microseconds, for the GOx/HRP pair assembled on a planar scaffold. The artificially colocalized systems from both data sets are presented to demonstrate the systematic increase of colocalized enhancement with electrostatic interactions.

It is clear from these results that electrostatic interactions are imperative for the proper assessment of protein-substrate association, even for molecular substrate with a net neutral charge. Including electrostatics increases the effect of colocalization most significantly, as evidenced by the increase in substrate transfer probability for both colocalized simulation sets. This effect is further emphasized by the steeper increase in substrate transfer probability with decreasing interenzyme distance, relative to the systems without electrostatic interactions.

3.3.4 Tubular Geometry Scaffold

The nanotube scaffold increased the successful transfer of the hydrogen peroxide intermediate significantly across all configurations relative to the planar scaffold assembly.

Effect of Interenzyme Distance

Varying the interenzyme distance over 10, 20, and 30 nanometers, the interenzyme substrate transfer probability for the nanotube scaffold system increased by 3, 5, and 6 fold, respectively, relative to the planar scaffold system. Comparison to the artificially colocalized simulations shows that confinement within the nanotube enhances successful substrate diffusion beyond simple colocalization by 3, 6, and 9 fold for interenzyme distances of 10, 20, and 30 nanometers, respectively (Figure 3.5).

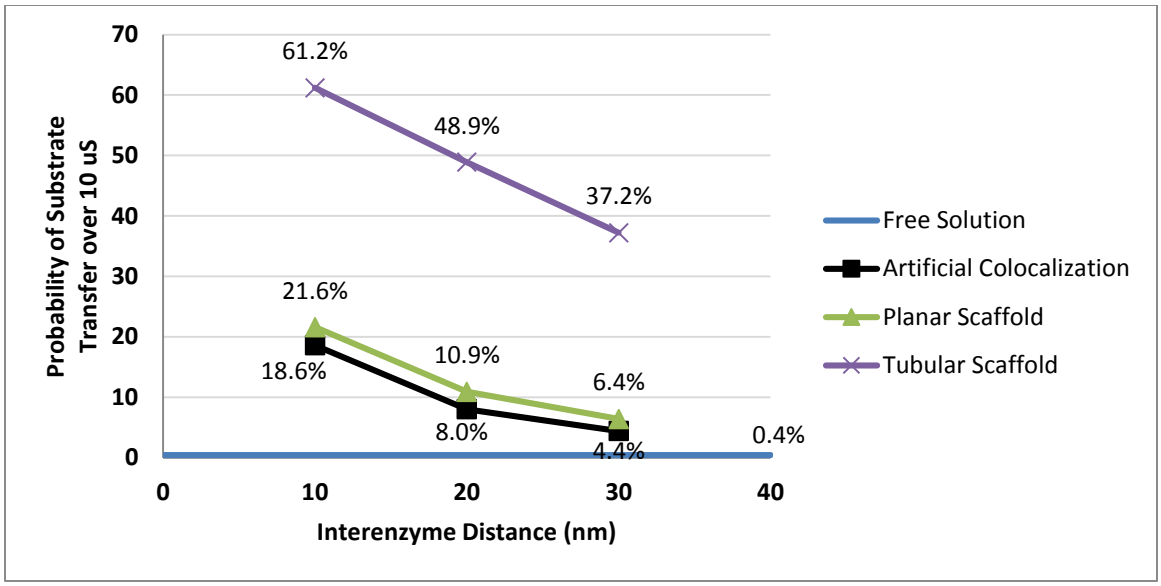

Figure 3.5. Substrate transfer probability, over 10 microseconds, for the GOx/HRP pair assembled within a nanotube scaffold and on a planar scaffold. Simulation results of the GOx/HRP pair colocalized, but with the scaffold model removed, are presented for comparison. In addition, the baseline probability for free solution substrate association over the same time period is provided for comparison. The scaffold-substrate affinity well depths were set to the default value of 0.267 kcal/mol.

Effect of Nanotube Radius

The geometry of the nanotube scaffold was explored for its ability to affect successful intermediate substrate transfer from GOx to HRP. Simulations were performed with a nanotube radius of 10, 15, and 20 nanometers and a single interenzyme distance of 30 nanometers. The nanotube radius increase from 10 nm to 15 nm reduced the substrate transfer probability to 72% of its original value, while increasing the radius from 10 nm

to 20 nm reduced the substrate transfer probability to 51% of its original value (Figure 3.6).

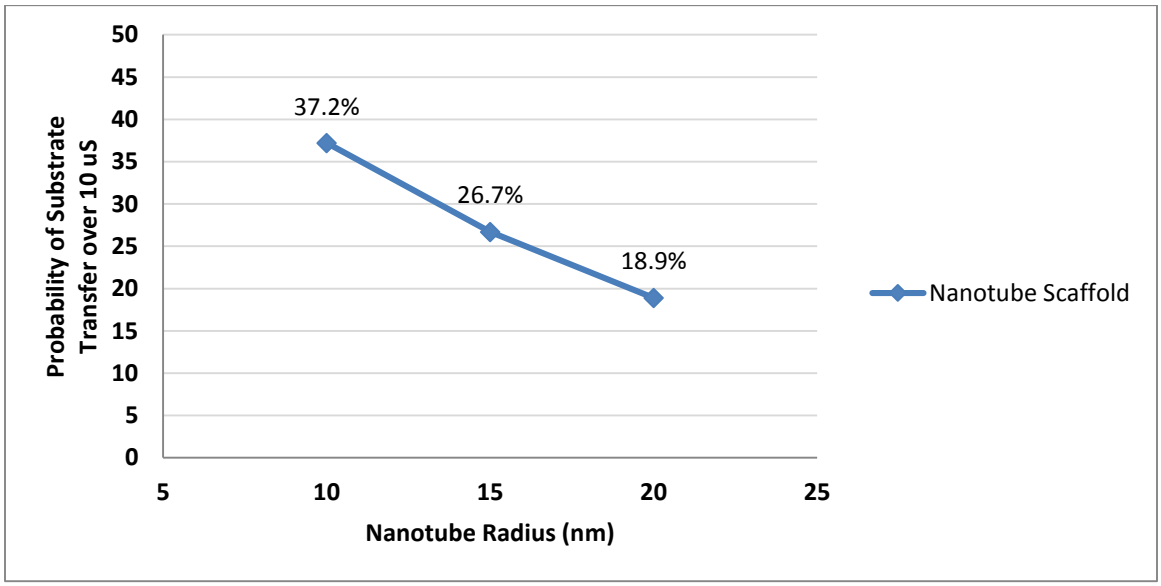

Figure 3.6. Substrate transfer probability as it varies with nanotube radius, over 10 microseconds, for the GOx/HRP pair assembled within a nanotube scaffold. Simulations were performed at an interenzyme distance of 30 nanometers. The scaffold-substrate affinity well depths were set to the default value of 0.267 kcal/mol.

Substrate-Scaffold Affinity Dependence

The well depth of the non-polar interaction potential between the substrate and scaffold were varied to determine the effect of substrate-scaffold affinity on interenzyme substrate transfer. With increasing affinity, the substrate transfer probability drops due to sequestration of the substrate on the scaffold (Figure 3.7). For the case of the nanotube scaffold, substrate association probability remains effectively the same when the well depth was doubled from 0.267 kcal/mol to 0.534 kcal/mol. The probability decreased by

7% when increased from 0.267 kcal/mol to 1.07 kcal/mol, and by 38% when increased from 0.267 kcal/mol to 2.163 kcal/mol. For the GOx/HRP pair assembled on the planar scaffold, the probability of substrate transfer dropped by 4% when the well depth was increased from 0.267 kcal/mol to 0.534 kcal/mol, by 6% when increased from 0.267 kcal/mol to 1.07 kcal/mol, and by 23% when increased from 0.267 kcal/mol to 2.163 kcal/mol. This further emphasizes the importance of the scaffold in the efficiency of substrate transfer in the GOx/HRP nanotube system.

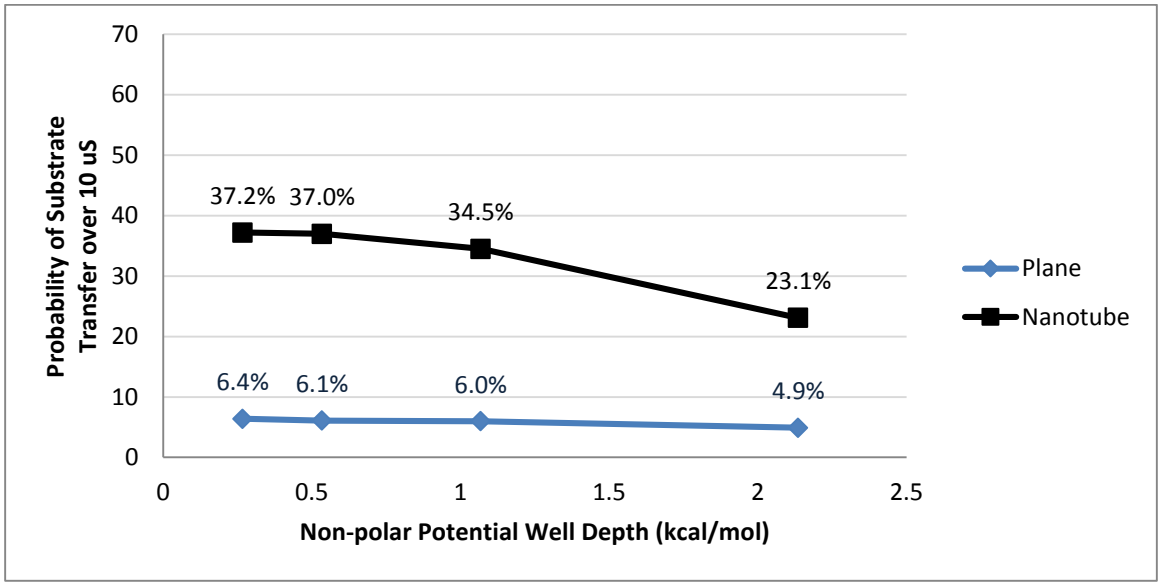

Figure 3.7. Substrate transfer probability as it varies with substrate-scaffold non-polar affinity, over 10 microseconds, with a fixed interenzyme distance of 30 nanometers. The nanotube radius was set to 10 nanometers. The scaffold-substrate affinity well depths were set to the default value of 0.267 kcal/mol.

3.3.5 Tubular Scaffold with Multiple Target HRP Enzymes

A novel spatially organized enzyme system, the HRP/GOx/HRP trio, was created by assembling a GOx enzyme with two target HRP enzymes within a nanotube scaffold. The GOx enzyme was centered in the nanotube, and both HRP enzymes were placed on either side of the GOx enzyme. The interenzyme distance between the GOx enzyme and each HRP enzyme was varied in the same manner as the GOx/HRP pair assembled in a nanotube. For the interenzyme distances of 10, 20, and 30 nanometers, the probability of interenzyme transfer increased by 31%, 51%, and 73%, respectively, relative to the GOx-HRP pair assembled in the nanotube scaffold (Figure 3.8). This trend also demonstrates that increasing interenzyme distance has less of a detrimental effect on interenzyme substrate transfer in the HRP/GOx/HRP trio than with the GOx/HRP pair.

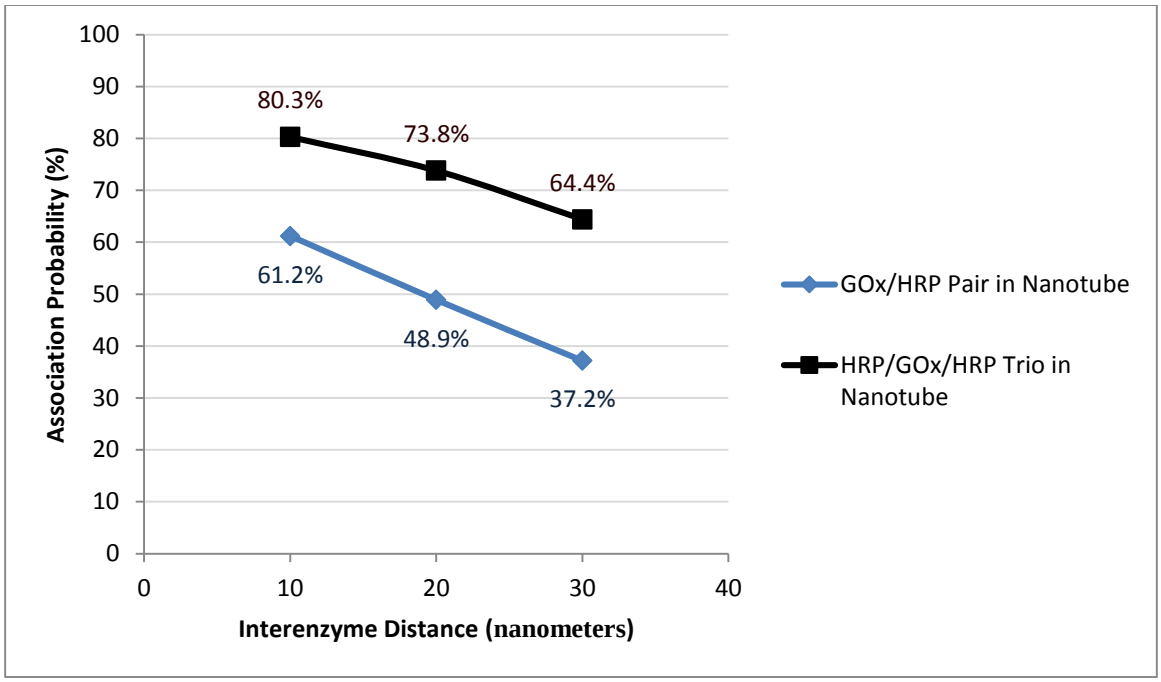

Figure 3.8. Substrate transfer probability comparison between the GOx/HRP pair and HRP/GOx/HRP trio. These systems were assembled within a nanotube and simulated over 10 microseconds. The nanotube radius was set to 10 nanometers, and the scaffold-substrate affinity well depths were set to the default value of 0.267 kcal/mol for all simulations.

However, the GOx enzyme may be shielded in this configuration due to occlusion by the HRP enzymes on either side. This could reduce the probability of diffusional encounter between GOx and its primary substrate, preventing or delaying the initiation of the GOx/HRP cascade. This may have a detrimental effect on the overall enzyme efficiency. Future investigation into the association of GOx and HRP to their respective co-substrates is thus warranted.

3.4 Conclusions

Simulations were performed to determine the probability of successful substrate transfer from GOx to HRP when the enzymes were spatially organized on a scaffold, colocalized artificially in solution, and disorganized in solution. An improved version of GeomBD was utilized for these simulations which included both non-polar and electrostatic interactions. The introduction of coarse grain partial charge embedding in the newest revision of GeomBD was shown to increase the substrate transfer probability for the GOx/HRP pair assembled on a planar scaffold by at least 2.5 fold relative to our previously reported results that excluded electrostatic interactions (Chapter 2). Our results suggest that the primary effect leading to enhanced throughput of the GOx/HRP pair assembled on a planar scaffold is colocalization, rather than the effect of the scaffold as a diffusive barrier.

The simulations performed with the GOx/HRP pair within a cylindrical nanotube scaffold were shown to increase the probability of interenzyme substrate transfer by 3 to 6 fold relative to the planar scaffold system. In addition, confinement within the nanotube led to a 3 to 9 fold increase in substrate transfer beyond the effects of colocalization. Thus, confinement is the primary means of efficiency in the nanotube system. Increasing the radius of the nanotube scaffold led to a decrease in substrate transfer, but still retained high efficiency relative to the planar scaffold.

Increasing scaffold-substrate affinity sequestered the substrate away from the enzymes, resulting in a reduction of substrate transfer of up to 23% for the planar scaffold in the range of affinities assessed. The effect of increasing the substrate-scaffold affinity in the

nanotube scaffold system led to a maximum of 38% reduction of substrate transfer probability in the range of affinities simulated. While increasing substrate affinity lowers the probability of substrate transfer on a short timescale, high substrate-scaffold affinity may lead to advantages from increased local concentration on a longer timescale.

The effect of adding a second target HRP enzyme to the system increases the interenzyme substrate transfer markedly. Because both GOx and HRP require co-substrate association for catalytic turnover, in addition to hydrogen peroxide transfer, the trio configuration may hinder cosubstrate association with GOx, potentially delaying the initiation of the chain reaction. Future research into the co-substrate association steps of all GOx/HRP pair and HRP/GOx/HRP trio configurations assembled within a nanotube scaffold would allow for thorough assessment of the most efficient configurations. In these future studies, we suggest varying the length of the nanotube to assess the impact on the accessibility of the GOx and HRP enzymes to their respective co-substrate.

3.5 References

(1) Mosbach, K.; Mattiasson, B. Matrix-Bound Enzymes. II. Studies on a Matrix-Bound Two-Enzyme-System. *Acta Chem. Scand.* **1970**, *24*, 2093–2100.

(2) Mateo, C.; Chmura, A.; Rustler, S.; van Rantwijk, F.; Stolz, A.; Sheldon, R. A. Synthesis of Enantiomerically Pure (S)-Mandelic Acid Using an Oxynitrilase–nitrilase Bienzymatic Cascade: A Nitrilase Surprisingly Shows Nitrile Hydratase Activity. *Tetrahedron: Asymmetry* **2006**, *17*, 320–323.

(3) You, C.; Myung, S.; Zhang, Y.-H. P. Facilitated Substrate Channeling in a Self-Assembled Trifunctional Enzyme Complex. *Angew. Chemie* **2012**, *124*, 8917–8920.

(4) Fu, J.; Liu, M.; Liu, Y.; Woodbury, N. W.; Yan, H. Interenzyme Substrate Diffusion for an Enzyme Cascade Organized on Spatially Addressable DNA Nanostructures. *J. Am. Chem. Soc.* **2012**, *134*, 5516–5519.

(5) Conrado, R. J.; Wu, G. C.; Boock, J. T.; Xu, H.; Chen, S. Y.; Lebar, T.; Turnšek, J.; Tomšič, N.; Avbelj, M.; Gaber, R.; Koprivnjak, T.; Mori, J.; Glavnik, V.; Vovk, I.; Benčina, M.; Hodnik, V.; Anderluh, G.; Dueber, J. E.; Jerala, R.; DeLisa, M. P. DNA-Guided Assembly of Biosynthetic Pathways Promotes Improved Catalytic Efficiency. *Nucleic Acids Res.* **2012**, *40*, 1879–1889.

(6) Scism, R. A.; Bachmann, B. O. Five-Component Cascade Synthesis of Nucleotide Analogues in an Engineered Self-Immobilized Enzyme Aggregate. *Chembiochem* **2010**, *11*, 67–70.

(7) Bäumler, H.; Georgieva, R. Coupled Enzyme Reactions in Multicompartment Microparticles. *Biomacromolecules* **2010**, *11*, 1480–1487.

(8) Winkel, B. S. J. Metabolic Channeling in Plants. *Annu. Rev. Plant Biol.* **2004**, *55*, 85–107.

(9) Ovádi, J.; Srere, P. A. Metabolic Consequences of Enzyme Interactions. *Cell Biochem. Funct.* **1996**, *14*, 249–258.

(10) Bauler, P.; Huber, G.; Leyh, T.; McCammon, J. A. Channeling by Proximity: The Catalytic Advantages of Active Site Colocalization Using Brownian Dynamics. *J. Phys. Chem. Lett.* **2010**, *1*, 1332–1335.

(11) Dueber, J. E.; Wu, G. C.; Malmirchegini, G. R.; Moon, T. S.; Petzold, C. J.; Ullal, A. V; Prather, K. L. J.; Keasling, J. D. Synthetic Protein Scaffolds Provide Modular Control over Metabolic Flux. *Nat. Biotechnol.* **2009**, *27*, 753–759.

(12) Ljungcrantz, P.; Carlsson, H.; Mansson, M. O.; Buckel, P.; Mosbach, K.; Buelow, L. Construction of an Artificial Bifunctional Enzyme, .beta.-Galactosidase/galactose Dehydrogenase, Exhibiting Efficient Galactose Channeling. *Biochemistry* **1989**, *28*, 8786–8792.

(13) Riedel, K.; Bronnenmeier, K. Intramolecular Synergism in an Engineered Exo-Endo-1,4-Beta-Glucanase Fusion Protein. *Mol. Microbiol.* **1998**, *28*, 767–775.

(14) Mao, Q.; Schunk, T.; Gerber, B.; Erni, B. A String of Enzymes, Purification and Characterization of a Fusion Protein Comprising the Four Subunits of the Glucose Phosphotransferase System of Escherichia Coli. *J. Biol. Chem.* **1995**, *270*, 18295–18300.

(15) Kim, Y. H.; Kwon, T. K.; Park, S.; Seo, H. S.; Cheong, J. J.; Kim, C. H.; Kim, J. K.; Lee, J. S.; Choi, Y. D. Trehalose Synthesis by Sequential Reactions of Recombinant Maltooligosyltrehalose Synthase and Maltooligosyltrehalose Trehalohydrolase from Brevibacterium Helvolum. *Appl. Environ. Microbiol.* **2000**, *66*, 4620–4624.

(16) De Pascale, D.; Di Lernia, I.; Sasso, M. P.; Furia, A.; De Rosa, M.; Rossi, M. A Novel Thermophilic Fusion Enzyme for Trehalose Production. *Extremophiles* **2002**, *6*, 463–468.

(17) Orita, I.; Sakamoto, N.; Kato, N.; Yurimoto, H.; Sakai, Y. Bifunctional Enzyme Fusion of 3-Hexulose-6-Phosphate Synthase and 6-Phospho-3-Hexuloisomerase. *Appl. Microbiol. Biotechnol.* **2007**, *76*, 439–445.

(18) Levasseur, A.; Navarro, D.; Punt, P. J.; Belaïch, J.-P.; Asther, M.; Record, E. Construction of Engineered Bifunctional Enzymes and Their Overproduction in Aspergillus Niger for Improved Enzymatic Tools to Degrade Agricultural by-Products. *Appl. Environ. Microbiol.* **2005**, *71*, 8132–8140.

(19) Jia, F.; Narasimhan, B.; Mallapragada, S. K. Biomimetic Multienzyme Complexes Based on Nanoscale Platforms. *AIChE J.* **2013**, *59*, 355–360.

(20) You, C.; Zhang, Y.-H. P. Self-Assembly of Synthetic Metabolons through Synthetic Protein Scaffolds: One-Step Purification, Co-Immobilization, and Substrate Channeling. *ACS Synth. Biol.* **2013**, *2*, 102–110.

(21) Winfree, E.; Liu, F.; Wenzler, L. A.; Seeman, N. C. Design and Self-Assembly of Two-Dimensional DNA Crystals. *Nature* **1998**, *394*, 539–544.

(22) Rothemund, P. W. K. Folding DNA to Create Nanoscale Shapes and Patterns. *Nature* **2006**, *440*, 297–302.

(23) Gu, H.; Chao, J.; Xiao, S.-J.; Seeman, N. C. Dynamic Patterning Programmed by DNA Tiles Captured on a DNA Origami Substrate. *Nat. Nanotechnol.* **2009**, *4*, 245–248.

(24) Douglas, S. M.; Dietz, H.; Liedl, T.; Högberg, B.; Graf, F.; Shih, W. M. Self-Assembly of DNA into Nanoscale Three-Dimensional Shapes. *Nature* **2009**, *459*, 414–418.

(25) Andersen, E. S.; Dong, M.; Nielsen, M. M.; Jahn, K.; Subramani, R.; Mamdouh, W.; Golas, M. M.; Sander, B.; Stark, H.; Oliveira, C. L. P.; Pedersen, J. S.; Birkedal, V.; Besenbacher, F.; Gothelf, K. V; Kjems, J. Self-Assembly of a Nanoscale DNA Box with a Controllable Lid. *Nature* **2009**, *459*, 73–76.

(26) Dietz, H.; Douglas, S. M.; Shih, W. M. Folding DNA into Twisted and Curved Nanoscale Shapes. *Science* **2009**, *325*, 725–730.

(27) Castro, C. E.; Kilchherr, F.; Kim, D.-N.; Shiao, E. L.; Wauer, T.; Wortmann, P.; Bathe, M.; Dietz, H. A Primer to Scaffolded DNA Origami. *Nat. Methods* **2011**, *8*, 221–229.

(28) Han, D.; Pal, S.; Nangreave, J.; Deng, Z.; Liu, Y.; Yan, H. DNA Origami with Complex Curvatures in Three-Dimensional Space. *Science* **2011**, *332*, 342–346.

(29) Shen, X.; Song, C.; Wang, J.; Shi, D.; Wang, Z.; Liu, N.; Ding, B. Rolling up Gold Nanoparticle-Dressed DNA Origami into Three-Dimensional Plasmonic Chiral Nanostructures. *J. Am. Chem. Soc.* **2012**, *134*, 146–149.

(30) Wei, B.; Dai, M.; Yin, P. Complex Shapes Self-Assembled from Single-Stranded DNA Tiles. *Nature* **2012**, *485*, 623–626.

(31) Wilner, O. I.; Weizmann, Y.; Gill, R.; Lioubashevski, O.; Freeman, R.; Willner, I. Enzyme Cascades Activated on Topologically Programmed DNA Scaffolds. *Nat. Nanotechnol.* **2009**, *4*, 249–254.

(32) Wilner, O. I.; Willner, I. Functionalized DNA Nanostructures. *Chem. Rev.* **2012**, *112*, 2528–2556.

(33) Seeman, N. C. Nanomaterials Based on DNA. *Annu. Rev. Biochem.* **2010**, *79*, 65–87.

(34) Lin, C.; Liu, Y.; Yan, H. Designer DNA Nanoarchitectures. *Biochemistry* **2009**, *48*, 1663–1674.

(35) Michelotti, N.; Johnson-Buck, A.; Manzo, A. J.; Walter, N. G. Beyond DNA Origami: The Unfolding Prospects of Nucleic Acid Nanotechnology. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. 4*, 139–152.

(36) Fu, Y.; Zeng, D.; Chao, J.; Jin, Y.; Zhang, Z.; Liu, H.; Li, D.; Ma, H.; Huang, Q.; Gothelf, K. V; Fan, C. Single-Step Rapid Assembly of DNA Origami Nanostructures for Addressable Nanoscale Bioreactors. *J. Am. Chem. Soc.* **2013**, *135*, 696–702.

(37) Gajhede, M.; Schuller, D. J.; Henriksen, A.; Smith, A. T.; Poulos, T. L. Crystal Structure of Horseradish Peroxidase C at 2.15 A Resolution. *Nat. Struct. Biol.* **1997**, *4*, 1032–1038.

(38) Wohlfahrt, G.; Witt, S.; Hendle, J.; Schomburg, D.; Kalisz, H. M.; Hecht, H. J. 1.8 and 1.9 A Resolution Structures of the Penicillium Amagasakiense and Aspergillus Niger Glucose Oxidases as a Basis for Modelling Substrate Complexes. *Acta Crystallogr. D. Biol. Crystallogr.* **1999**, *55*, 969–977.

(39) Reva, B. A.; Finkelstein, A. V.; Sanner, M. F.; Olson, A. J. Residue-Residue Mean-Force Potentials for Protein Structure Recognition. *Protein Eng. Des. Sel.* **1997**, *10*, 865–876.

(40) Shen, T.; Wong, C. F.; McCammon, J. A. Atomistic Brownian Dynamics Simulation of Peptide Phosphorylation. *J. Am. Chem. Soc.* **2001**, *123*, 9107–9111.

(41) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

(42) Henzler, T. Transport and Metabolic Degradation of Hydrogen Peroxide in Chara Corallina: Model Calculations and Measurements with the Pressure Probe Suggest Transport of H2O2 across Water Channels. *J. Exp. Bot.* **2000**, *51*, 2053–2066.

(43) Kang, M.; Roberts, C.; Cheng, Y.; Chang, C. A. Gating and Intermolecular Interactions in Ligand-Protein Association: Coarse-Grained Modeling of HIV-1 Protease. *J. Chem. Theory Comput.* **2011**, *7*, 3438–3446.

(44) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

Chapter 4: Ligand binding pathway elucidation for cryptophane host-guest complexes

4.1 Introduction

The mechanisms of molecular association are critical to the understanding of many problems in chemistry, biology, and medicine.[1–4] In addition to studying the final ligand–receptor bound complexes, investigating binding processes helps elucidate fundamental mechanisms, including allostery, induced fit, and gated control associations. An understanding of the binding pathway of a molecular complex can lend thermodynamic and kinetic information which is relevant in medical and pharmaceutical research, and can aid in the development of new or improved drugs and drug carriers. The kinetic behavior, in particular, is relevant to the efficacy of drugs *in vivo*.[5,6] This study focuses on chemical host-guest systems as approachable ligand-receptor models to gain insight into the thermodynamics and kinetics of binding.[7,8] Supramolecular systems are ideal systems for method development and optimization because they are simple but preserve all the important characteristics of more complicated ligand–protein systems.

The standard free energy change of binding is an alternative way of expressing its equilibrium constant, $K_{eq}=\exp(-\Delta G°/RT)$, which in turn is the ratio of the rate constants for association (k_{on}) and dissociation (k_{off}), $K_{eq} \approx k_{on}/k_{off}$.[9–11] A free energy barrier is coupled with association or dissociation rates, which may result from entropy loss, desolvation penalty, and conformational changes. Important and elegant work has been done on the gated theories and statistical thermodynamics of non-covalent association.[2,12–14] However, real molecular systems are usually quite complicated, and the

theoretical studies and analytical solutions may not be able to provide direct links to real molecular systems of practical interest. Computational modeling is a powerful tool to explain fundamental processes of molecular binding.

Computational methods capable of establishing binding pathways traditionally rely on various forms of molecular dynamics (MD) or Monte Carlo (MC) simulations.[15–25] Standard simulations are computationally expensive, and in some cases impractical, for this determination because of the relatively long timescales under which binding processes typically occur.[8] In addition, many repeated simulations are required to adequately sample possible binding pathways.[26] Dedicated hardware has been engineered to make this direct simulation route more practical, but at a high monetary cost.[27,28] This has been more commonly addressed, at the cost of accuracy, through the use of accelerated dynamics simulations or through a reduction in system complexity with coarse-grained models and/or Brownian dynamics simulations.[29–40] The computation of ligand binding free energy barriers is even more challenging. One popular method is to apply steered MD simulations to gradually "pull" a ligand from the ligand binding site to the solvent or use simulations with various sampling techniques to obtain the potential of mean force (PMF).[20,22,23,41–48] Other methods apply metadynamics or the adaptive biasing force (ABF) methods to bias the simulations.[31,38,49,50] These methods generate a free energy profile along a ligand dissociation/association pathway; thus, an energy barrier may be observed from the PMF plot. However, because the pathways are not known beforehand, one may need to perform very thorough sampling.[18,45,51–55] Moreover, both the ligand and receptor may undergo conformational changes during ligand association,

so adequate sampling to compute an accurate binding free energy landscape can be difficult.

Because transition/intermediate states are too rare, standard unbiased dynamics-based simulations have difficulty finding accurate pathways and free energy landscapes. Many methods, for example transition path sampling (TPS), nudged elastic band (NEB) and string methods, have been developed to determine proper transition pathways between adjacent stable states without making *a priori* assumption about transition paths.[24,49,56–65] They provide accurate coordinates and energy barriers of transition states, and can be applied to study chemical reactions or conformation transitions. However, applying these rigorous but computationally intensive methods to protein-ligand association remains a daunting task; especially binding pathways which cover large configuration space and involve several states.[66]

This article describes a new methodological approach called Hopping Minima for the determination of conformational transitions and association/dissociation pathways by connecting local energy minima using the molecular system's natural motions. The natural motions of a molecular system are modeled from their normal modes, which are exaggerated into the coordinated motions of the molecule or host-guest complex. The method is similar to the superposition approach, termed reaction path Hamiltonian superposition approach (RPHSA), developed by Strodel and Wales; however, there are several key differences between the two methods.[67] Both methods begin with using conformational search engine to find local energy minima and utilize harmonic approximations. Our objective is to rapidly identify connected intermediate states to

determine binding pathways from numerous local energy minima. Because we aim to work on much more complex molecular systems, we have developed an automated algorithm to bridge minima in the HM method using internal coordinates. The RPHSA method obtains more detailed free energy surface (FES), but due to the concern of computation time, our method provides a simpler approximation for the FES of a ligand-receptor binding pathway.

We present a proof of concept implementation of the method which uses the conformational search program Tork to perform conformational searches and a set of programs to carry out the detection of conformational transitions and building of the binding pathways.[68–71] The method is exemplified by utilizing our implementation to connect distinct conformations of the alanine dipeptide. Alanine dipeptide has served as a standard model system for various computational works, including the development of new sampling methods, benchmark of force field, and determination of transition paths between two stable states.[56,58,66,67] Unlike existing works that constructed detailed free energy surfaces of a transition pathway, our work shows a quick method to illustrate possible pathways of conformational changes. We then extend the implementation to a set of host-guest systems: two cryptophane hosts with two cation guests. The cryptophanes are a class of spherically-shaped organic molecular capsules with hollow internal cavities. They are commonly studied in the context of molecular recognition because of their propensity to bind atoms and small molecules.[72–75] For example, a ^{129}Xe-cryptophane biosensor was designed for potential use as a magnetic resonance imaging (MRI) contrast agent.[76] A well-studied member of the cryptophanes,

cryptophane-E, has proven to form supramolecular complexes with small neutral and cationic guest molecules. Cryptophane-ES, a derivative of cryptophane-E with the methoxy gating arms replaced with thiomethyl gates, has been shown to have similar binding characteristics.[72] The association and dissociation rates of a guest binding to cryptophane-E can be 10^3 to 10^4 times faster than with the same guest binding to cryptophane-ES, but the net binding affinities (ΔG) are essentially the same. However, it is not clear how such a tiny difference in the arms of the host can significantly affect the kinetics but not the thermodynamics. We therefore applied the method presented in this work to crytophanes-E and -ES with trimethylammonium and tetramethylammonium cation guests to reveal interesting guest binding pathways and shed light on the nature of the interactions between cryptophanes and their guests. The work also characterizes changes in free energy, entropy and enthalpy of minimum conformations along ligand-receptor association pathways.

4.2 Methods

4.2.1 Overview

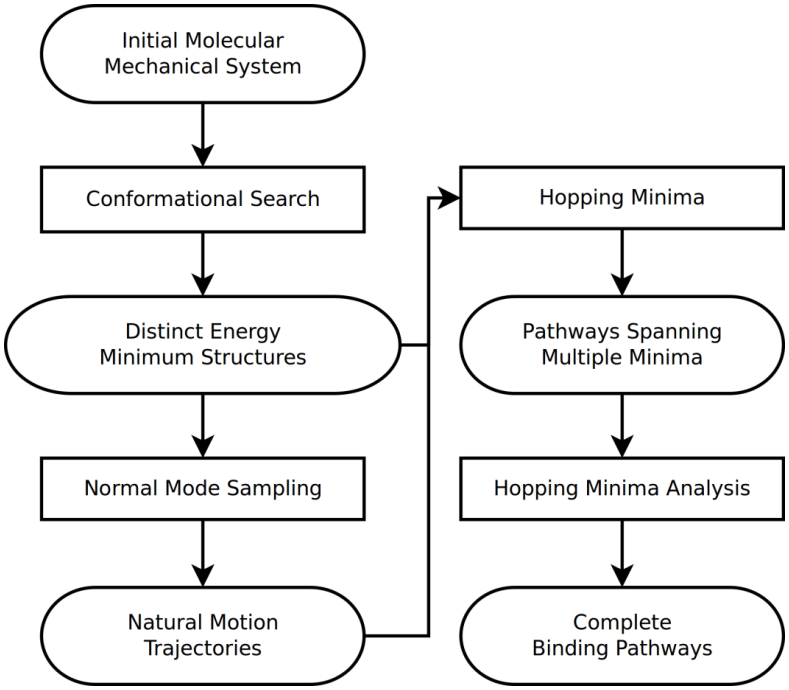

Figure 4.1. Overview of the hopping minima method. Rectangles represent distinct steps in the method, while ovals represent input and output structures and trajectories of structures. Arrows direct the work flow through the method.

The procedure, termed “Hopping Minima,” has been summarized as follows (Figure 4.1). An extended conformational search procedure is performed on the molecular system. Free energies of the individual minima are calculated using the Mining Minima generation 2 (M2) program. Natural motions are then sampled using a modified M2 program.[68] The second derivative (Hessian) matrix of the potential energy function is computed and diagonalized to obtain eigenvalues and eigenvectors in Bond-Angle-

Torsion (BAT) coordinates for each local minimum.[70,71] The eigenvalues are sorted and only eigenvectors corresponding to the first 3 to 10 small eigenvalues are considered. Soft modes, or normal modes with a small eigenvalue, represent the most relevant natural motions of a molecular system for our purposes as they have the largest associated ranges of vibration. The system's coordinates are scanned along the normal modes and snapshots are saved as a trajectory during the scanning. Transitions between conformational minima are determined by structurally comparing the saved trajectories to the minimum conformations. If a motion approaches multiple local energy minima, the trajectory is saved as a molecular path. For molecular complexes, multiple paths can be combined to reveal possible ligand binding pathways. If local energy minima are numerous, several steps and analyses are carried out in order to determine tractable pathways and/or pathways that connect more popular (low energy) local energy minima.

This method relies on the principles of the energy landscape theory.[77] As such, only the local energy minimum conformations are explicitly sampled in the conformational search procedure. We rely on the natural motion sampling to capture the high energy transition state structures.

4.2.2 Finding local energy minima

Tork is a conformational search utility that distorts a molecular system's initial conformation along internal coordinates, minimizing along the way in order to find new minima. This process is automated, but conceptually similar to previous manual methods of exploring conformational space.[71] The results of a given search are dependent on the conformation used to seed the search. Thus, many conformations, including both high

and low energy conformations, should be used as input to thoroughly sample the conformational space of a molecular system.

Unlike the free energy calculations where one can examine if the integration has converged, it is less straightforward to tell whether a conformational search is complete. Conformational search should be targeted to structures in and around conformations of relevance, if the algorithm allows. Searches should be performed until relevant conformational states are well represented. Because we aim to study association pathways, many local energy minima should be available in addition to the global energy minimum. For complexes, we are most notably interested in those local energy minima in which the ligand is in the intermediately bound and unbound states. To more efficiently find local energy minima, multiple intermediate partially bound or unbound conformations are used as initial structures for successive runs of conformational search. Root mean square deviation (RMSD) based alignment of all conformations can help reveal areas around the host with sparse ligand population.[78] Minimum conformations with a guest molecule near the sparsely populated areas can be used to start additional searches. Because the search will result in the same conformations multiple times, repeat conformations are filtered out. Eliminating repeats is nontrivial when a molecular system has symmetries, so a symmetry detection algorithm is applied.[79] For example, for our Me_4N^+ guest, simply rotating a methyl group by 120 degree is considered the same conformation.

4.2.3 Free energy calculation

The M2 method is used to compute the free energy for all conformations in a system via a modified harmonic approximation. The M2 method calculates the configuration integral as a sum of contributions from the conformational states, which is Boltzmann weighted by each conformation.[68] The configuration integral for each local minimum is calculated upon the harmonic approximation. The M2 method builds upon the harmonic approximation by correcting for anharmonicities in BAT coordinates.[70,71] To compute the binding free energy, the calculation is carried out for three species: the complex, free host, and free guest. These values are then subtracted to find the absolute binding free energy for the complex. For the HM method, the free energy is calculated for all sampled conformational states of the complex only with no Boltzmann sum applied.

4.2.4 Natural motion sampling

Natural motions are modeled using the normal modes of the host-guest complex. Although the mass of the atom is not considered and it is not equivalent to classical normal modes in mechanics, we term each set of eigenvalue and eigenvectors a "normal mode". Internal BAT coordinates are built for the structures, instead of using standard Cartesian coordinates. With a set of local energy minima of intermediate states available, the second derivative (Hessian) matrix of the potential energy function is computed for each local minimum and diagonalized to obtain eigenvalues and eigenvectors. The eigenvalues are sorted and Gaussian integrands associated with the harmonic approximation are constructed.[70,71] The molecular system's coordinates are scanned along the soft normal modes in increments of 0.05 radians or Angstroms until the eigenvectors

reach the nine standard deviation (9σ) point of the Gaussian. The scanned conformations are recorded and the results output as a trajectory. In this study, we scan the lowest 5 normal modes to obtain 5 paths of natural motions for a given local energy minimum.

4.2.5 Identifying paths

There are two RMSD-based methods for comparing similarity in molecular structure between natural motion trajectories and conformational minima: Cartesian coordinate RMSD and dihedral angle RMSD. Cartesian coordinate RMSD compares absolute position of atoms between the natural motion trajectory and conformational minima. Alignment of all structures to the same reference minimum is required. Because of this alignment, two atomic selections are required: an alignment selection, and an RMSD comparison selection. Cartesian coordinate RMSD is appropriate for all types of systems, whether a single molecule or complex of molecules. Dihedral angle RMSD compares all, or a subset of, dihedral angles containing only heavy atoms. This comparison is fast relative to standard Cartesian coordinate RMSD because alignment of the structures is not required. Dihedral angle RMSD is appropriate for single molecules only, as it is unable to capture relative positions and orientations of multiple molecules.

For the alanine dipeptide, dihedral RMSD and Cartesian RMSD are both employed using alignment and RMSD selections of all heavy atoms and amide hydrogens. For the cryptophane complexes, we selected atoms in the two 9-membered rings, the least flexible portions of the host molecule, for alignment.

The Hopping Minima application performs the Cartesian RMSD comparison as follows: for a natural motion trajectory t_i calculated from minimum conformation m_i, each frame

in t_i, $t_{i,f}$, is aligned to m_i. All conformational minima, m_j ($j \neq i$), are then aligned to m_i, and the cartesian RMSD comparison is performed between $t_{i,f}$ and m_j. The dihedral RMSD comparison similarly cycles through all $t_{i,f}$ and calculates the dihedral RMSD between minima m_j ($j \neq i$) and $t_{i,f}$. For either type of comparison, if the RMSD value is within a user-supplied tolerance, then m_i and m_j are considered to have a conformational transition described by t_i, and is saved as a separate path trajectory.

RMSD tolerance is the main variable that affects resultant conformational transitions. RMSD tests with more lenient tolerance values result in numerous pathways, and most of them are not of interest. Therefore we introduced several techniques to our implementation to reduce redundancy in minimum conformations and to filter out highly similar or unlikely pathways. These techniques are provided to help the user reduce noise in the results of calculations with lenient RMSD tolerance values.

4.2.5.1 Thinning

After a thorough conformational search, numerous distinct conformations, including high energy conformations, may populate spatially small regions near an equilibrium basin. Therefore, before we build any paths, or binding pathways, low-energy regionally representative conformations are determined in order to limit the search results (Figure 4.2). A thinning cutoff parameter is provided by the user that dictates that no two ligands will be within the cutoff distances from one another. The thinning step starts with sorting the conformations based on their computed free energy. The conformations are sequentially analyzed. Each minimum conformation's position is compared with its predecessors. If any previous conformation is within a thinning cutoff, then the

conformation will be ignored; otherwise, the new conformation will be reported as a new representative conformation.

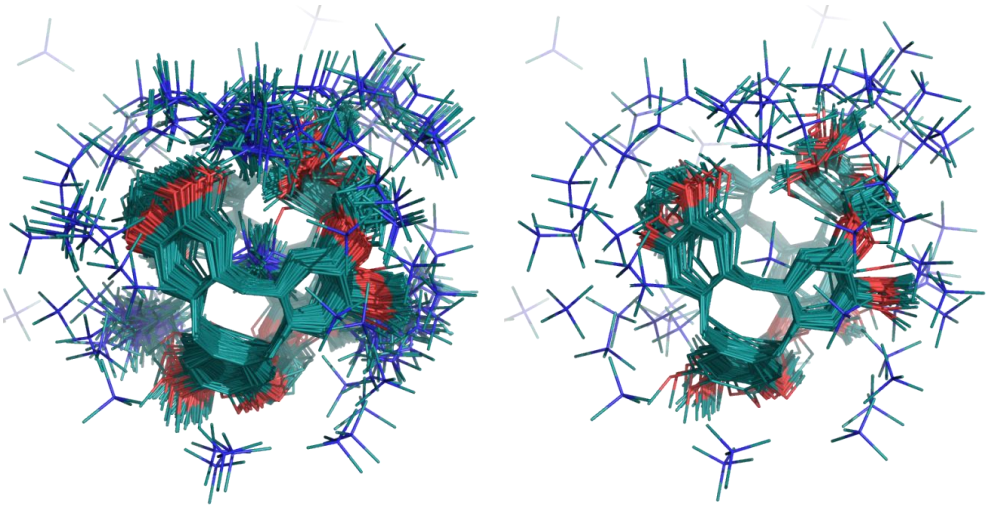

Figure 4.2: Thinning filter. (Left) An aligned set of conformational minima for a host-guest complex, showing the many bound and unbound ligand states. (Right) The set of minima after thinning, showing the regionally representative conformational minima. A thinning cutoff of 1.5A was used, thus no ligand is within a 1.5Å radius from another.

4.2.5.2 Intersection threshold

Because the natural motion paths may not intersect directly with the minimum conformations, and to complement the thinning procedure, we included an intersection threshold parameter (Figure 4.3). This parameter is provided by the user as a distance in angstroms that the ligand in the natural motion path must be from a given ligand position in a minimum conformation for the two to be considered intersecting.

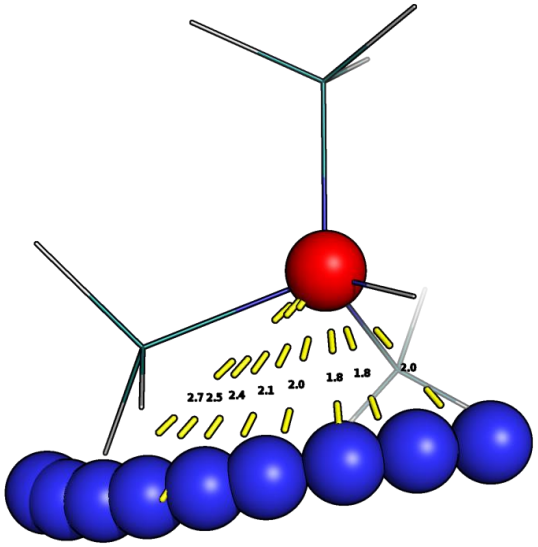

Figure 4.3: Intersection cutoff parameter**.** Distance calculation between the center of each ligand in a path from the natural motion sampling (centers of each ligand shown as a blue sphere) and the center of a minimum ligand conformation (center shown as a red sphere). The intersection between the path and the minimum conformation is determined by the intersection threshold parameter. In this example, the distances range from 1.8 to 2.7Å, thus any intersection threshold parameter greater or equal to 1.8Å would yield an intersection.

4.2.5.3 Minimum path length

To remove any particularly short paths, we implemented a minimum path length parameter, a value provided by the user in angstroms (Figure 4.4). After a path is determined to pass between at least two conformational minimum conformations, the total length of the path is calculated. If the length is less than the minimum path length parameter, then the path is discarded.

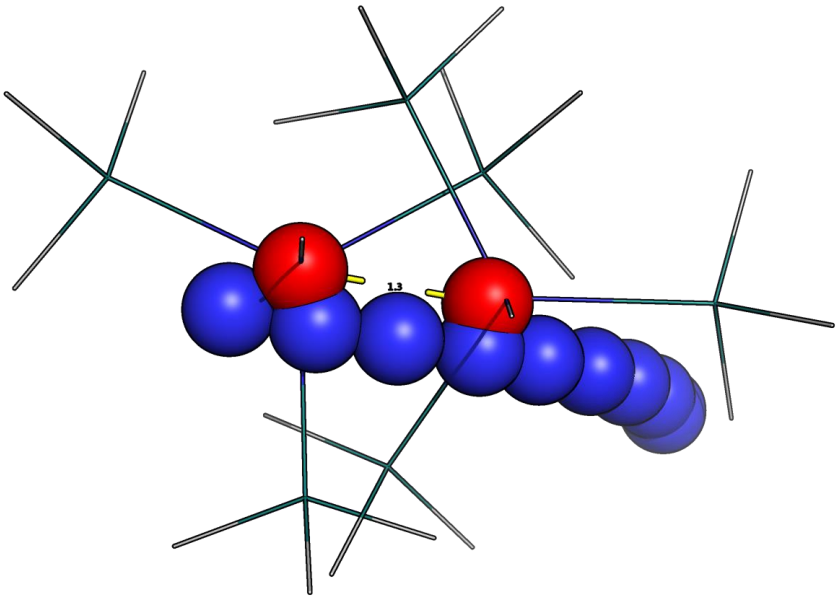

Figure 4.4: Minimum path length. For a path (represented here only as blue spheres, the center of each ligand in the path) intersecting two more minimum conformations (shown in full representation, with centers shown with red spheres), the total distance from the first to the last intersected minimum is calculated. If this distance is less than the minimum path length parameter, the path is discarded. For this example, the distance between the two intersected minima is 1.3Å apart. Thus, a minimum path length parameter greater than 1.3Å would exclude this path from being recorded.

4.2.5.4 Scan filtering

As multiple normal modes are scanned for each conformational minimum, paths formed from a given minimum may be very similar, resulting in similar paths intersecting the same local minima. Thus if two or more motions, sampled from the same minimum conformation, yield similar paths, we will only keep the one built from the lowest normal mode.

4.2.6 Combining multiple paths

To illustrate guest association processes and possible binding pathways, individual minima-connecting paths will be combined to form a more complete and longer binding pathway (see Figures 4.8 – 4.11). A script was written that aids in determining combinations of relevant paths for the formation of binding pathways. This script prompts the user for minima of interest and returns information on paths involving those minima. The output is sorted by the potential energy of the path, and provides information such as maximum, minimum, and average potential energy of the natural motion paths, and the index of local energy minima that the path intersects. This allows the user to inspect paths that may possibly join to form a binding pathway, and quickly select paths of low potential energy. Note that high potential energy suggests that a normal mode scanning path may contain steric clashes or overly distorted structures. Although these paths are often not relevant, occasionally a rare and high potential energy pathway may be of interest. Thus, we do not explicitly exclude the high energy paths.

A strategy that we found useful was to construct the pathway of unbinding. Any paths that linked a bound state conformation to an intermediate state conformation were used as the starting point for construction. These intermediate state conformations were then used to find paths leading to further unbound conformations. This process is repeated until there are no additional linking paths.

4.2.7 Computational details

The alanine dipeptide was parameterized using AMBER FF03 forcefield.[80] The initial structure of the free cryptophane-E host was obtained from the Cambridge Structural

Database, ID SEDPOG.[81] As only this experimental structure was available, all other structures were created manually. We created the sulfurous derivative, cryptophane-ES, using the cryptophane-E initial structure as a template and manually constructed the ligands using Avogadro[82]. The DREIDING force field was used to describe all complex structures with parameters assigned by the program Vdock, and partial charges computed with the program Vcharge with the VC/2004 parameter set.[83–85] The solvation energy, W, is computed with the generalized Born model using the water dielectric constant of 80.0 for the alanine dipeptide system and the tetrachloroethane dielectric constant of 8.42 for the complexes.[86] For the cryptophane hosts with an open cavity, the dielectric cavity radius of each atom is set to the mean of the size of chlorine (~1.8 Å) and the atom's van der Waals radius.[87] All local energy conformations were energy-minimized by the conjugate gradient method and then the Newton-Raphson method until the energy gradient is < 10-3 kcal/mol/Å. Once the free cryptophane hosts and free guest molecules were fully constructed, we created the complex systems by combining each combination of host and guest and minimizing the structures using the aforementioned minimization methods. When running the Hopping Minima application, the alanine dipeptide required no filtering methods. RMSD tolerances of 0.2 and 0.3 Å and radians were used to explore conformational transitions of the alanine dipeptide. For the four complexes, the thinning cutoff was set to 1.5 Å, except for cryptophane-E in complex with trimethylammonium, where a value of 2.5 Å was used. The RMSD tolerance was set to 2.5 Å for all complex systems. All calculations were performed as single processes on an Intel Xeon 2.67GHz quad core processor computer using Ubuntu Linux 10.10. Each conformational search

run with Tork took about 10 minutes, with 10 to 15 runs performed in total on each complex system and at least 6 on each free host structure. All conformational minima were obtained for the alanine dipeptide after 3 Tork search runs. The free energy calculations took 6-8 hours for the 160-255 cryptophane-E host-guest complexes, 16 hours for the 144-174 cryptophane-ES host-guest complexes, 4 hours for the 120 free cryptophane-E host conformations, and 6 hours for the 197 free cryptophane-ES host conformations. The alanine dipeptide and ligand free energies were calculated in less than one minute due to having few minimum conformations. RMSD-based identification of conformational transitions took less than 20 minutes to complete for all systems. Each manual path construction time for the complexes took an hour or less to complete.

4.3 Results and Discussion

Because alanine dipeptide has been used as a model system by many computational methods, we first apply the HM method to the system as an example to demonstrate the procedures of the method. Five consecutive conformational searches were performed on the alanine dipeptide, at which point the conformational search converged. Additional searches were performed on the identified minima to ensure that the conformational space was thoroughly sampled. The HM application was applied using the 16 conformations that were found. Utilizing either RMSD calculation method, with no thinning of the minima, five unique conformational transitions were found for the system that connect two minimum states within 0.2 angstroms. Larger RMSD tolerance values revealed conformational transitions between additional minima. As seen in Figure 4.5, an RMSD tolerance value of 0.3 angstroms yielded two conformational transitions that were

combined to form a transition spanning 3 minima. This transition is driven by rotation around the psi and phi dihedral angles. The first motion primarily rotates the psi angle, while the second motion rotates the psi and phi angles in concert. These results demonstrate that for small single molecule systems, a single normal mode can accurately describe specific conformational transitions.

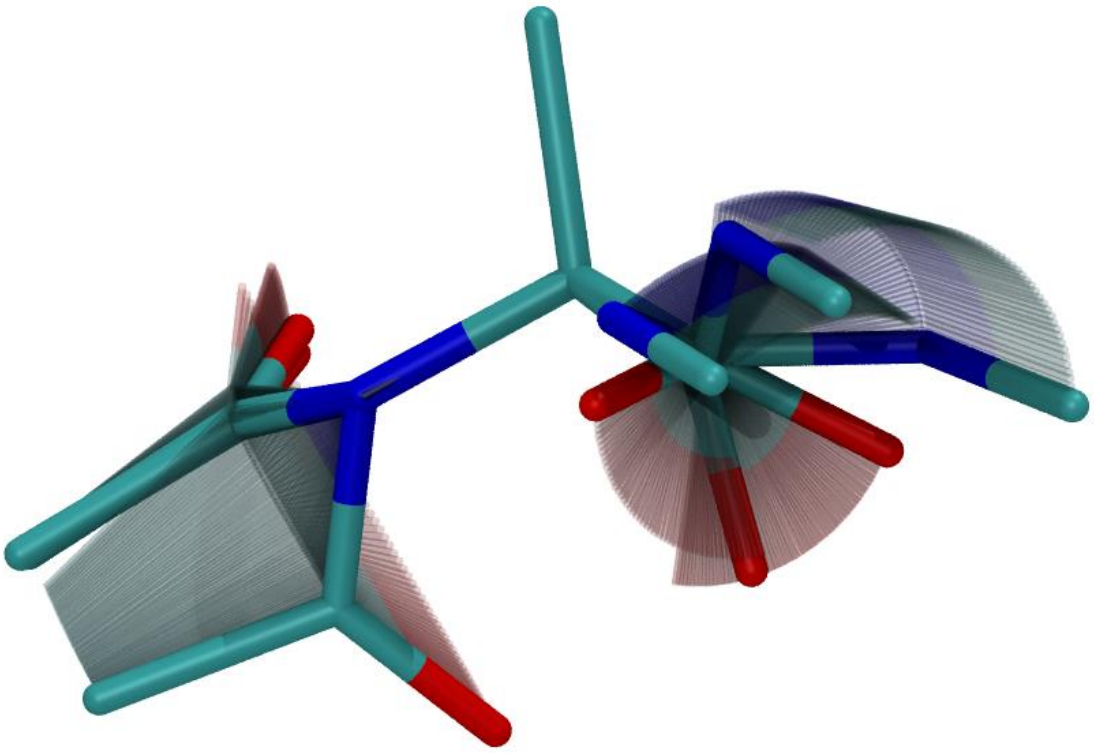

Figure 4.5: Conformational transitions of the Alanine dipeptide. Two conformational transitions connect three local energy minimum conformations of the alanine dipeptide. The major dihedral angles of rotation are the phi and psi angles of the alanine backbone. The first motion primarily rotates the psi angle, while the second motion rotates the psi and phi angles in concert.

Table 4.1 shows the binding free energies, entropies, and mean energy components for the four cryptophane systems. We first computed the binding free energy on each system using the M2 method to validate the molecular mechanical force field parameters and partial charges model employed. For easier comparison, we separated the systems into two groups based on the guest molecules. The relative binding free energies, $\Delta\Delta G$, are in a reasonable range when compared with experiments, although the absolute binding free energies are not as accurate as other reported values.[68,88,89] All complexes are stabilized upon binding primarily by van der Waals interactions, as assessed experimentally with a neutral ligand.[73] However, with the force field utilized for these systems, the cryptophane-ES system over stabilized the guest molecules through van der Waals interaction, resulting in a free energy calculation that did not perfectly match experiments. While the electrostatic interaction between the hosts and guests are significantly favorable, the desolvation penalty negates this impact. Because the free energies of only local minimum conformations are available to the HM method, the free energy barrier to binding cannot be known exactly. However, with multiple constructed binding pathways for a complex, one can analyze the free energies of each conformational minimum along the pathway to rule out unlikey paths that contain unfavorable conformational transitions.

Guest	Host	ΔG_{cal}	$\Delta\Delta G_{cal}$	ΔG_{exp}	$\Delta\Delta G_{exp}$	$-T\Delta S_{conf}$	ΔE	ΔW_{GB}	ΔG^*_{cal}	$\Delta\Delta G^*_{cal}$	k_{on}
Me_3NH^+	**Cryp. E**	-8.5	0	-4.3	0	10.6	-27.5	9.1	2.4		76.6
	Cryp. ES	-9.3	-0.8	-4.1	0.2	12.2	-43.3	22.3	14.2	11.8	2.32×10^{-3}
Me_4N^+	**Cryp. E**	-9.5	0	-7.6	0	12.4	-30.3	8.8	8.9		7.2
	Cryp. ES	-12.4	-2.9	-7.3	0.3	11.9	-41.7	17.9	27.7	18.8	2.88×10^{-3}

Table 4.1: Free energies and mean energies for cryptophane-cation systems. The calculated binding free energy (ΔG_{cal}) and experimental binding free energy (ΔG_{exp}) are presented for comparison, along with relative binding free energies for each ($\Delta\Delta G$). The calculated entropy term ($-T\Delta S_{conf}$), contribution to solvation free energy (ΔW_{GB}), and mean potential energy contributions (ΔE) for electrostatics, van der Waals, and bonded terms. The calculated binding free energy (ΔG^*_{cal}), difference in the binding free energy between hosts for each guest molecule ($\Delta\Delta G^*_{cal}$), and experimental association constant (k_{on}) from reference [72] are presented. All data are in units of kcal/mol, except for the association constant (units of $M^{-1}s^{-1}$).

The results of the conformational search of free hosts, cryptophane-E and cryptophane-ES, were analyzed for trends in gating, flexibility, and the degree each host exhibited structural preorganization for binding. Identical conformational search procedures yielded more conformational minima for cryptophane-ES relative to cryptophane-E. However, cryptophane-E had more conformations in the low energy states relative to its global minimum, while cryptophane-ES had higher population of conformations in the

high energy states. For example, cryptophane-E had 30 out of 120 total conformations within 10RT (where R is the gas constant, T is the temperature, 300K), or ~6 kcal/mol, of its global minimum conformation. Cryptophane-ES had 14 out of 197 conformations within 10RT. As seen in Figure 4.6, cryptophane-E displays a distinctly more open cavity structure relative to cryptophane-ES. Cryptophane-E conformations also show a higher propensity for gating arms to be in the "open" position, while the gating arms of cryptophane-ES tend to be in the closed position. Cryptophane-ES does exhibit these "open" conformations, but they are higher energy transition state structures to binding. The closed arm conformations from cryptophane-ES may be due to the larger radius of the sulfur atom than the oxygen atom, resulting in stronger intra-molecular interactions between the two arms. The distribution of partial charges is also different in cryptophane-ES and cryptophane-E, which contributes further to the overall shape of the cavity.

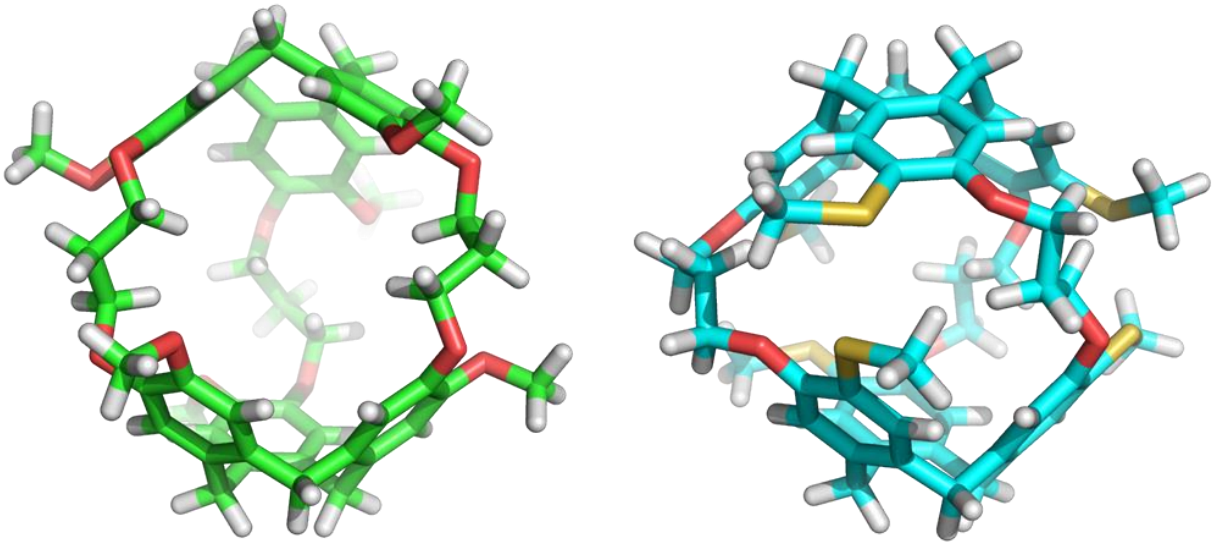

Figure 4.6: Structural comparison of cryptophane-E and cryptophane-ES. The global energy minimum conformation of cryptophane-E (left, green) has distinctly open gating arms and open cavity while those of cryptophane-ES (right, cyan) have a more compact structure with closed gating arms.

4.3.1 Analysis of the binding pathways

Upon analysis of the binding pathways, we found that the cryptophane-E host was preorganized for binding and had little structural movement during the course of the binding process. Cryptophane-ES, unexpectedly and on the contrary, underwent large structural changes to accommodate the entry of the ligand to the interior of the host structure. The consistency in the entropic penalties matched these flexibility trends, staying constant for cryptophane-E and fluctuating for cryptophane-ES before dropping for both systems in the final step of the binding process. Interestingly, the guest molecule quickly lost configurational entropy, mainly from translation/rotation entropy loss while

approaching the host surface. However, it did not fully gain favorable intermolecular attractions before reaching the final bound states, resulting in unfavorable free energy intermediates (Figure 4.7).

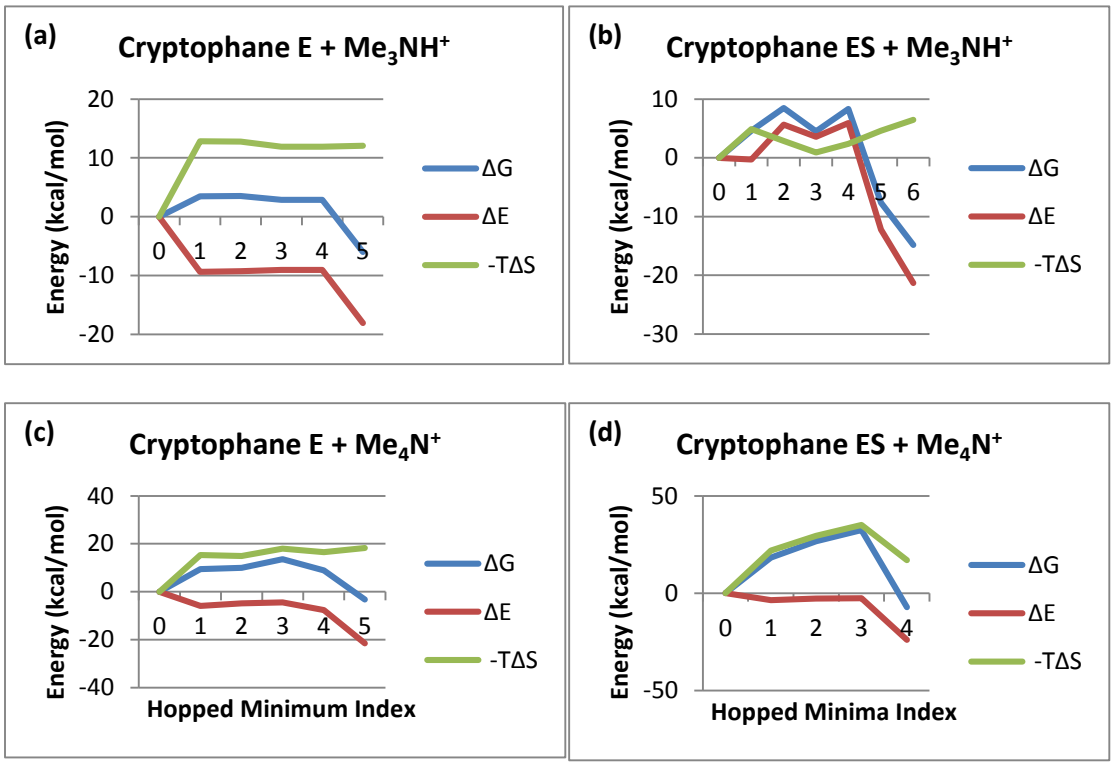

Figure 4.7: Free energy distributions for pathways developed for the cryptophane-cation systems. Energetic breakdown for the lowest free energy barrier binding pathways developed for the four test systems. All values are calculated relative to a reference energy (point 0), which is calculated as the free energy of the host and ligand when not interacting.

In the following sections we discuss the results of the four host-guest systems in more detail. While many high potential energy and high free energy pathways can be sampled for each system, only those with reasonable free energy barriers are discussed and presented.

4.3.1.1 Trimethylammonium and Cryptophane-E

Two low free energy barrier binding pathways were found for the complex of cryptophane-E with the trimethylammonium guest (Figure 4.8). Of the four complexes, this system exhibited the lowest free energy transition state and the most favorable enthalpy upon binding. Rather than approaching the host directly from free space, the pathways had ligands that first associated with the surface of the host molecule, then approached the window from the surface. The initial contacts provide intermolecular attractions. However, these interactions are not large enough to overcome the configurational entropy loss. (Figure 4.7a).

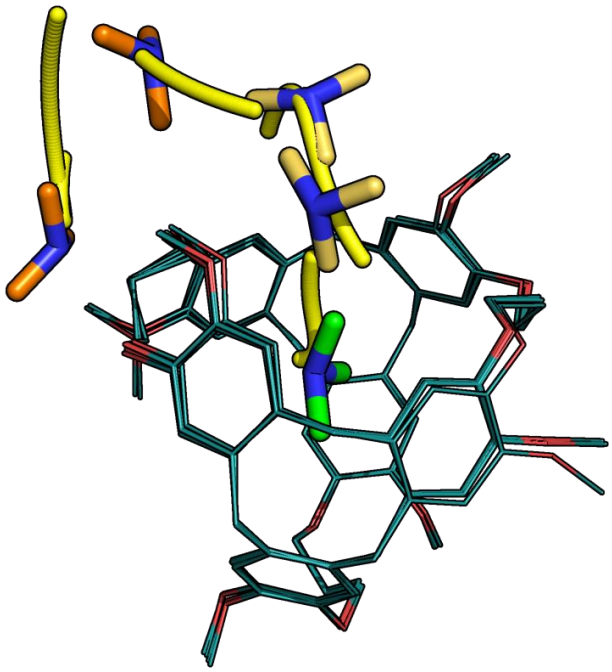

Figure 4.8: Hopping Minima resulting binding pathway for Cryptophane-E with Trimethylammonium guest. Four natural motion paths, represented by yellow traces, connect five distinct minimum states, with free energies of 664.2, 664.3, 663.6, 663.6, and 654.7 kcal/mol, in order of decreasing distance from the center of the host.

4.3.1.2 Trimethylammonium and Cryptophane-ES

Three binding pathways were sampled for the complex of cryptophane-ES with trimethylammonium.(Figure 4.8) The three pathways sampled binding processes through each of the three host windows and approached the binding site from free space. While the conformational search found guests on the host surface, our method did not connect any pathways from these local energy minima. The gating arms were in the closed position when the guest approached the surface, therefore the guest may need to return to

free space and rebind to induce gate opening. Positive enthalpies were observed upon association, which dropped dramatically upon binding (Figure 4.7b). The intermediate state conformations, specifically those conformations with ligands near the gating arms of the host, suffered van der Waals repulsions and/or structural perturbations that resulted in penalties in the calculated enthalpy and free energy.

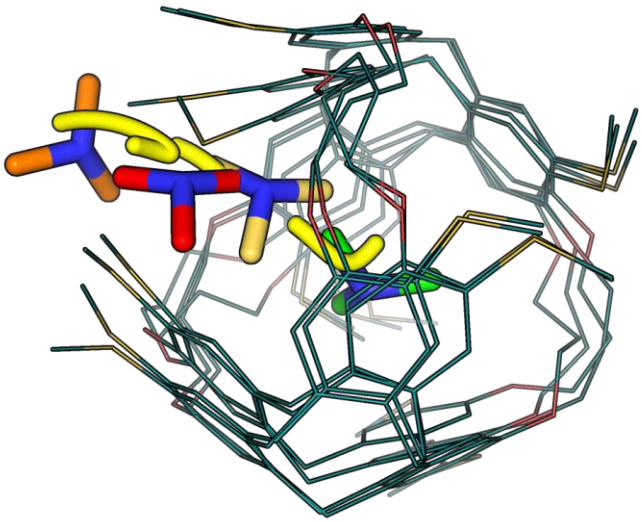

Figure 4.9: Hopping Minima resulting binding pathway for Cryptophane-ES with Trimethylammonium guest. Three natural motion paths, represented by yellow traces, connect six distinct minimum states, with free energies of 652.5, 656.3, 640.4, and 633.2 kcal/mol, in order of decreasing distance from the host window.

4.3.1.3 Tetramethylammonium and Cryptophane-E

One association pathway was determined for cryptophane-E with tetramethylammonium, involving a ligand path to the window of the host starting from the surface of the host molecule (Figure 4.10). The pathway had a negative enthalpy upon association with significant entropic compensation (Figure 4.7c). The last step of the binding process, connecting the tetramethylammonium in the window to the center of the host cavity, was not found. Because the guest has a larger size that limits its flexibility while in the window and final bound state, the trajectories sampled following the normal mode motions show small ranges of vibration. Therefore, the Hopping Minima method could not connect the two minima using the parameters applied for building other paths. However, the final step is not complicated as it only has one direction to move forward; thus it does not affect the illustration of the binding pathway.

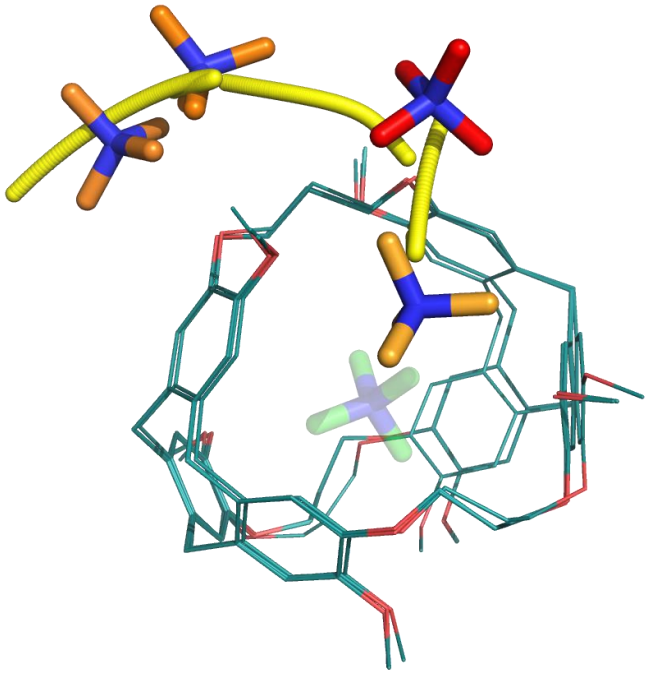

Figure 4.10: Hopping Minima resulting association pathway for Cryptophane-E with Tetramethylammonium guest. Three natural motion paths, represented by yellow traces, connect four distinct minimum states, with free energies of 698.5, 699.1, 702.6 and 697.9 kcal/mol, in order of decreasing distance from the host window. The lowest free energy bound state ligand conformation, with a value of 685.8 kcal/mol, is included to visualize free energy trends.

4.3.1.4 Tetramethylammonium and Cryptophane-ES

Two binding pathways were generated for the complex of cryptophane-ES with the tetramethylammonium guest. Both sampled pathways approached the binding site from free space (Figure 4.11). Analysis of the minima involved in these paths revealed free energy values for the intermediate state conformations that deviate from that expected based on experimental free energy barrier values. The results suggest that correctly

computing energies for intermediate states may require more accurate force field parameters than those in the free and bound conformations. Our results again show positive changes in enthalpies and free energies for the intermediate conformations were due to steric clashes resulting in van der Waals repulsion and unfavorable structural perturbation. High repulsion (>40 kcal/mol) was observed in several minima hopping paths that approached the host window. The repulsive potentials were more pronounced in this system than in cryptophane-ES with trimethylammonium due to the increased size of the ligand in this system.

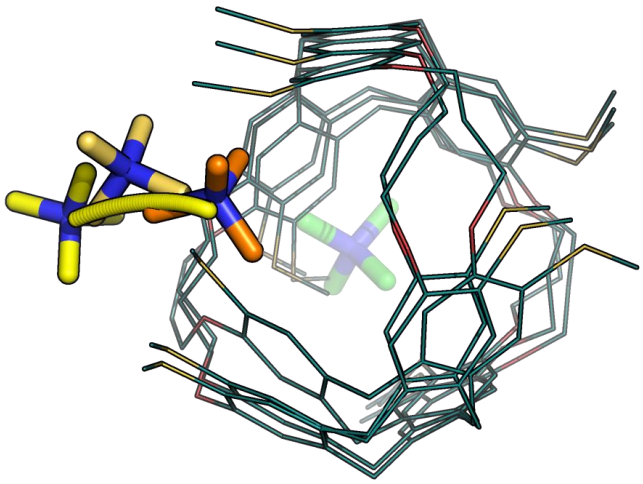

Figure 4.11: Hopping Minima resulting association pathway for Cryptophane-ES with Tetramethylammonium guest. A single natural motion path, represented by a yellow trace, connects three distinct minimum states, with free energies of 687.66, 695.99, and 701.74 kcal/mol, in order of decreasing distance from host window. The lowest free energy bound state ligand conformation is included to visualize free energy trends.

4.4 Comments

The alanine dipeptide test system demonstrated the efficacy and speed of the HM method. After parameterization of the system, only a few minutes were required to carry out all required calculations, and multiple transitions between several different conformations were found. Thus, the HM method may be an attractive and competitive option for quickly exploring conformational transitions in single molecules without the need for time-consuming dynamics or the need to specify starting and ending conformational states.

The results of modeling binding pathways of the cryptophane-guest systems are promising, provide insight into molecular recognition, and show potential for use in a wide range of applications, such as designing drug carriers and inhibitors with preferred binding kinetics. Actual guest binding processes can take milliseconds to seconds; therefore, applying classical molecular dynamics simulations to sample binding pathways can be very expensive computationally. As with all methods that rely on molecular mechanical force fields, our method is dependent on the force field for its accuracy. Although we validated our force field parameters by reproducing experimental binding affinities using the M2 free energy calculation method, the free energy barriers found from Hopping Minima, particularly for the cryptophane-ES systems, show the sensitivity of the method to the choice of force field parameters. Our models reveal that when a guest is passing the window of a host, the intermediate states have very small spatial restrictions, and small errors in force field parameters may result in a local minimum with unrealistically large computed free energy.

While the AMBER FF03 molecular mechanical force field, DREIDING force field and VC/2004 partial charge models were utilized for the series of test systems, any molecular mechanical model can be employed.[90–94] Other implicit solvent models, in addition to the generalized Born model, can be used as well.

Thorough conformational search is a key to successfully finding binding pathways and building binding free energy landscapes. This study used the Tork conformational search tool which has advantages in efficiently finding conformations for chemical hosts with macro-cyclic rings. Nevertheless, any conformational search method used together with a duplicate conformation filtering algorithm that accounts for symmetry can be used. For example, other conformational search methods, such as the basin-hopping approach or commercial packages also can be used.[95–97] The method can also be scaled up, with regards to system size. For larger molecular systems, such as protein-ligand complexes, significant speed increases can be obtained by fixing regions of the protein that are not critical to ligand binding while allowing the rest of the atoms to be flexible.

For the cryptophane host-guest systems, our experiments were setup such that the RMSD intersection tests between conformational minima and natural motions used only the distance between the center nitrogen atom of the ligands. This measurement works efficiently for small molecules with symmetry, such as the guests in this work. For more complicated ligand structures, more atoms should be included in the RMSD selection. Additionally, host atoms could be added to the selection to capture conformational transitions that occur during ligand translation. Because the Hopping Minima method is based on connecting distinct local energy minima, occasionally binding processes from

the free to intermediate bound states are sampled, but a complete binding pathway to the final bound state could not be found. Molecular dynamics simulations could be employed to sample the final binding event for the route to bridge the gap. Molecular dynamics or existing methods for transition path samplings, such as the RPHSA, string and TPS methods, can be further applied to the distinct local energy minima found in a binding pathway to smooth the transitions between each energy state. If one is interested in applying steered or targeted molecular dynamics in explicit solvent, the binding pathway could be used as a physically based reaction coordinate, thus avoiding the potential for introducing unnatural bias to the simulation.

4.5 Conclusions

The presented method allows for the computational modeling of molecular transitions as well as binding pathways for host-guest systems. Investigating binding pathways helps elucidate fundamental mechanisms, including allostery, conformational changes, gated control associations, and can guide molecular design. Our study also illustrated that the gating effects and molecular distortion contribute mainly to slow association rate in guest binding to the cryptophane-ES system. Our analysis reveals the balance or imbalance between enthalpy and configuration entropy changes during a ligand binding process. The method relies on multiple filters and analysis procedures, and thorough minimum conformational search of a host-guest complex. The method has a distinct computational cost advantage over existing methods. Calculations performed on the alanine dipeptide provided evidence of the efficacy of the method to capture conformational transitions. Using four cryptophane complexes as example systems, the method demonstrated the effectiveness of sampling multiple binding pathways for each system and deepened our understanding of cryptophane-guest associations.

4.6 References

(1) Lauffenburger, D. A.; Linderman, J. *Receptors: Models for Binding, Trafficking, and Signaling (Google eBook)*; Oxford University Press, 1996; p. 376.

(2) Janin, J. The Kinetics of Protein-Protein Recognition. *Proteins* **1997**, *28*, 153–161.

(3) Swinney, D. C. Kinetic Binding Mechanisms: Their Contribution to an Optimal Therapeutic Index. *Label-Free Technol. Drug Discov.* **2011**, 284–302.

(4) Schreiber, G.; Haran, G.; Zhou, H.-X. Fundamental Aspects of Protein-Protein Association Kinetics. *Chem. Rev.* **2009**, *109*, 839–860.

(5) Markgren, P. O.; Lindgren, M. T.; Gertow, K.; Karlsson, R.; Hämäläinen, M.; Danielson, U. H. Determination of Interaction Kinetic Constants for HIV-1 Protease Inhibitors Using Optical Biosensor Technology. *Anal. Biochem.* **2001**, *291*, 207–218.

(6) Dong, B.; Zhou, Q.; Zhao, J.; Zhou, A.; Harty, R. N.; Bose, S.; Banerjee, A.; Slee, R.; Guenther, J.; Williams, B. R. G.; Wiedmer, T.; Sims, P. J.; Silverman, R. H. Binding Kinetics of Darunavir to Human Immunodeficiency Virus Type 1 Protease Explain the Potent Antiviral Activity and High Genetic Barrier. *J. Virol.* **2007**, *78*, 8983–8993.

(7) Helgeson, R. C.; Hayden, A. E.; Houk, K. N. Controlled Gating of a Hemicarcerand by Disulfide-Dithiol Interchange. *J. Org. Chem.* **2010**, *75*, 570–575.

(8) Rieth, S.; Hermann, K.; Wang, B.-Y.; Badjić, J. D. Controlling the Dynamics of Molecular Encapsulation and Gating. *Chem. Soc. Rev.* **2010**, *40*, 1609–1622.

(9) Berry, R. S.; Rice, S. A.; Ross, J. *Physical Chemistry*; Oxford University Press, USA, 2000; p. 1080.

(10) Dill, K. A.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*; Garland Science, 2010; p. 720.

(11) Wales, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*; Cambridge University Press, 2003; p. 681.

(12) Schlosshauer, M.; Baker, D. A General Expression for Bimolecular Association Rates with Orientational Constraints. *J. Phys. Chem. B* **2002**, *106*, 12079–12083.

(13) Cai, L.; Zhou, H.-X. Theory and Simulation on the Kinetics of Protein-Ligand Binding Coupled to Conformational Change. *J. Chem. Phys.* **2011**, *134*, 105101.

(14) Northrup, S. H.; Zarrin, F.; McCammon, J. A. Rate Theory for Gated Diffusion-Influenced Ligand Binding to Proteins. *J. Phys. Chem.* **1982**, *86*, 2314–2321.

(15) Simunovic, M.; Zagrovic, B.; Tomić, S. Mechanism and Thermodynamics of Ligand Binding to Auxin Amidohydrolase. *J. Mol. Recognit. 24*, 854–861.

(16) Vashisth, H.; Abrams, C. F. Ligand Escape Pathways and (un)binding Free Energy Calculations for the Hexameric Insulin-Phenol Complex. *Biophys. J.* **2008**, *95*, 4193–4204.

(17) Zuckerman, D. *Statistical Physics of Biomolecules : An Introduction*; CRC: Boca Raton Fla. [u.a.], 2010.

(18) Held, M.; Noé, F. Calculating Kinetics and Pathways of Protein-Ligand Association. *Eur. J. Cell Biol.* **2012**, *91*, 357–364.

(19) Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J. Am. Chem. Soc.* **2009**, *131*, 11811–11818.

(20) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10184–10189.

(21) Ahmad, M.; Gu, W.; Helms, V. Mechanism of Fast Peptide Recognition by SH3 Domains. *Angew. Chem. Int. Ed. Engl.* **2008**, *47*, 7626–7630.

(22) Huang, D.; Caflisch, A. The Free Energy Landscape of Small Molecule Unbinding. *PLoS Comput. Biol.* **2011**, *7*, e1002002.

(23) Ramadugu, S. K.; Kashyap, H. K.; Ghirlanda, G.; Margulis, C. Binding Free Energy Calculations of Dimannose to Cyanovirin-N High and Low Affinity Sites Using the Jarzynski's Equality and Umbrella Sampling. *Glycobiology* **2011**, *21*, 1477–1478.

(24) Peters, B. Recent Advances in Transition Path Sampling: Accurate Reaction Coordinates, Likelihood Maximisation and Diffusive Barrier-Crossing Dynamics. *Mol. Simul.* **2010**, *36*, 1265–1281.

(25) Elber, R. Long-Timescale Simulation Methods. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151–156.

(26) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13118–13123.

(27) Shaw, D. E.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Dror, R. O.; Piana, S.; Shan, Y.; Towles, B.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B. Millisecond-Scale Molecular Dynamics Simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09*; ACM Press: New York, New York, USA, 2009; p. 1.

(28) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133*, 9181–9183.

(29) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9*, 71.

(30) Pal, S.; Fichthorn, K. A. Accelerated Molecular Dynamics of Infrequent Events. *Chem. Eng. J.* **1999**, *74*, 77–83.

(31) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(32) Chang, C.-E. A.; Trylska, J.; Tozzini, V.; McCammon, J. A. Binding Pathways of Ligands to HIV-1 Protease: Coarse-Grained and Atomistic Simulations. *Chem. Biol. Drug Des.* **2007**, *69*, 5–13.

(33) Spaar, A.; Dammer, C.; Gabdoulline, R. R.; Wade, R. C.; Helms, V. Diffusional Encounter of Barnase and Barstar. *Biophys. J.* **2006**, *90*, 1913–1924.

(34) Frembgen-Kesner, T.; Elcock, A. H. Absolute Protein-Protein Association Rate Constants from Flexible, Coarse-Grained Brownian Dynamics Simulations: The Role of Intermolecular Hydrodynamic Interactions in Barnase-Barstar Association. *Biophys. J.* **2010**, *99*, L75–7.

(35) Kim, Y. C.; Hummer, G. Coarse-Grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. *J. Mol. Biol.* **2008**, *375*, 1416–1433.

(36) Mereghetti, P.; Gabdoulline, R. R.; Wade, R. C. Brownian Dynamics Simulation of Protein Solutions: Structural and Dynamical Properties. *Biophys. J.* **2010**, *99*, 3782–3791.

(37) Huber, G. A.; McCammon, J. A. Browndye: A Software Package for Brownian Dynamics. *Comput. Phys. Commun.* **2010**, *181*, 1896–1905.

(38) Wang, J.-C.; Pal, S.; Fichthorn, K. Accelerated Molecular Dynamics of Rare Events Using the Local Boost Method. *Phys. Rev. B* **2001**, *63*, 085403.

(39) Carmichael, S. P.; Shell, M. S. A New Multiscale Algorithm and Its Application to Coarse-Grained Peptide Models for Self-Assembly. *J. Phys. Chem. B* **2012**, *116*, 8383–8393.

(40) Zheng, W. Coarse-Grained Modeling of the Structural States and Transition Underlying the Powerstroke of Dynein Motor Domain. *J. Chem. Phys.* **2012**, *136*, 155103.

(41) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

(42) Gan, W.; Roux, B. Binding Specificity of SH2 Domains: Insight from Free Energy Simulations. *Proteins* **2009**, *74*, 996–1007.

(43) Wang, J.; Zhang, K.; Lu, H.; Wang, E. Quantifying the Kinetic Paths of Flexible Biomolecular Recognition. *Biophys. J.* **2006**, *91*, 866–872.

(44) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free Energy Calculation from Steered Molecular Dynamics Simulations Using Jarzynski's Equality. *J. Chem. Phys.* **2003**, *119*, 3559.

(45) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical Free Energy Methods for Drug Discovery: Progress and Challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150–160.

(46) Elenewski, J. E.; Hackett, J. C. Free Energy Landscape of the Retinol/serum Retinol Binding Protein Complex: A Biological Host-Guest System. *J. Phys. Chem. B* **2010**, *114*, 11315–11322.

(47) Zhang, D.; Gullingsrud, J.; McCammon, J. A. Potentials of Mean Force for Acetylcholine Unbinding from the alpha7 Nicotinic Acetylcholine Receptor Ligand-Binding Domain. *J. Am. Chem. Soc.* **2006**, *128*, 3019–3026.

(48) Isralewitz, B.; Gao, M.; Schulten, K. Steered Molecular Dynamics and Mechanical Functions of Proteins. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224–230.

(49) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(50) Chipot, C.; Hénin, J. Exploring the Free-Energy Landscape of a Short Peptide Using an Average Force. *J. Chem. Phys.* **2005**, *123*, 244906.

(51) Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.

(52) Lettieri, S.; Zuckerman, D. M. Accelerating Molecular Monte Carlo Simulations Using Distance and Orientation-Dependent Energy Tables: Tuning from Atomistic Accuracy to Smoothed "Coarse-Grained" Models. *J. Comput. Chem.* **2012**, *33*, 268–275.

(53) Lettieri, S.; Mamonov, A. B.; Zuckerman, D. M. Extending Fragment-Based Free Energy Calculations with Library Monte Carlo Simulation: Annealing in Interaction Space. *J. Comput. Chem.* **2011**, *32*, 1135–1143.

(54) Lv, C.; Zheng, L.; Yang, W. Generalized Essential Energy Space Random Walks to More Effectively Accelerate Solute Sampling in Aqueous Environment. *J. Chem. Phys.* **2012**, *136*, 044103.

(55) Zheng, L.; Yang, W. Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. *J. Chem. Theory Comput.* **2012**, *8*, 810–823.

(56) Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. Transition Pathways in Complex Systems: Application of the Finite-Temperature String Method to the Alanine Dipeptide. *J. Chem. Phys.* **2005**, *123*, 134109.

(57) E, W.; Ren, W.; Vanden-Eijnden, E. Simplified and Improved String Method for Computing the Minimum Energy Paths in Barrier-Crossing Events. *J. Chem. Phys.* **2007**, *126*, 164103.

(58) Maragliano, L.; Vanden-Eijnden, E. On-the-Fly String Method for Minimum Free Energy Paths Calculation. *Chem. Phys. Lett.* **2007**, *446*, 182–190.

(59) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in Free Energy Space. *J. Chem. Phys.* **2007**, *126*, 054103.

(60) Henkelman, G.; Jónsson, H. Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points. *J. Chem. Phys.* **2000**, *113*, 9978.

(61) Bolhuis, P.; Chandler, D.; Dellago, C.; Geissler, P. Transition Path Sampling: Throwing Ropes over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*.

(62) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. Transtion Path Sampling. *Adv. Phys.* **2002**, *123*, 1–78.

(63) Fischer, S.; Karplus, M. Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. *Chem. Phys. Lett.* **1992**, *194*, 252–261.

(64) Maragliano, L.; Vanden-Eijnden, E.; Roux, B. Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials. *J. Chem. Theory Comput.* **2009**, *5*, 2589–2594.

(65) Noé, F.; Fischer, S. Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.

(66) Pan, A. C.; Sezer, D.; Roux, B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J. Phys. Chem. B* **2008**, *112*, 3432–3440.

(67) Strodel, B.; Wales, D. J. Free Energy Surfaces from an Extended Harmonic Superposition Approach and Kinetics for Alanine Dipeptide. *Chem. Phys. Lett.* **2008**, *466*, 105–115.

(68) Chang, C.-E.; Gilson, M. K. Free Energy, Entropy, and Induced Fit in Host-Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.

(69) Chang, C.-E.; Gilson, M. K. Tork: Conformational Analysis Method for Molecules and Complexes. *J. Comput. Chem.* **2003**, *24*, 1987–1998.

(70) Chang, C.-E.; Potter, M. J.; Gilson, M. K. Calculation of Molecular Configuration Integrals. *J. Phys. Chem. B* **2003**, *107*, 1048–1055.

(71) Isogai, Y. Enkephalin: Conformational Analysis by Means of Empirical Energy Calculations. *Proc. Natl. Acad. Sci.* **1977**, *74*, 414–418.

(72) Garcia, C.; Humilière, D.; Riva, N.; Collet, A.; Dutasta, J.-P. Kinetic and Thermodynamic Consequences of the Substitution of SMe for OMe Substituents of Cryptophane Hosts on the Binding of Neutral and Cationic Guests. *Org. Biomol. Chem.* **2003**, *1*, 2207.

(73) Canceill, J.; Cesario, M.; Collet, A.; Guilhem, J.; Lacombe, L.; Lozach, B.; Pascard, C. Structure and Properties of the Cryptophane-E/CHCl3 Complex, a Stable van Der Waals Molecule. *Angew. Chemie Int. Ed. English* **1989**, *28*, 1246–1248.

(74) Holman, K. T. Cryptophanes: Molecular Containers. *Encycl. Supramol. Chem.* **2004**, 340–348.

(75) Spence, M. M.; Rubin, S. M.; Dimitrov, I. E.; Ruiz, E. J.; Wemmer, D. E.; Pines, A.; Yao, S. Q.; Tian, F.; Schultz, P. G. Functionalized Xenon as a Biosensor. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10654–10657.

(76) Aaron, J. A.; Chambers, J. M.; Jude, K. M.; Di Costanzo, L.; Dmochowski, I. J.; Christianson, D. W. Structure of a 129Xe-Cryptophane Biosensor Complexed with Human Carbonic Anhydrase II. *J. Am. Chem. Soc.* **2008**, *130*, 6942–6943.

(77) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.

(78) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. Fast Determination of the Optimal Rotational Matrix for Macromolecular Superpositions. *J. Comput. Chem.* **2010**, *31*, 1561–1563.

(79) Chen, W.; Huang, J.; Gilson, M. K. Identification of Symmetries in Molecules and Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1301–1313.

(80) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

(81) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **2002**, *58*, 380–388.

(82) Avogadro: an open-source molecular builder and visualization tool http://avogadro.openmolecules.net/ (accessed Sep 20, 2012).

(83) Kairys, V.; Gilson, M. K. Enhanced Docking with the Mining Minima Optimizer: Acceleration and Side-Chain Flexibility. *J. Comput. Chem.* **2002**, *23*, 1656–1670.

(84) Gilson, M. K.; Gilson, H. S. R.; Potter, M. J. Fast Assignment of Accurate Partial Atomic Charges: An Electronegativity Equalization Method That Accounts for Alternate Resonance Forms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982–1997.

(85) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(86) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(87) Luo, R.; Head, M. S.; Given, J. A.; Gilson, M. K. Nucleic Acid Base-Pairing and N-Methylacetamide Self-Association in Chloroform: Affinity and Conformation. *Biophys. Chem.* **1999**, *78*, 183–193.

(88) Moghaddam, S.; Yang, C.; Rekharsky, M.; Ko, Y. H.; Kim, K.; Inoue, Y.; Gilson, M. K. New Ultrahigh Affinity Host-Guest Complexes of cucurbit[7]uril with bicyclo[2.2.2]octane and Adamantane Guests: Thermodynamic Analysis and Evaluation of M2 Affinity Calculations. *J. Am. Chem. Soc.* **2011**, *133*, 3570–3581.

(89) Chen, W.; Chang, C.-E.; Gilson, M. K. Calculation of Cyclodextrin Binding Affinities: Energy, Entropy, and Implications for Drug Design. *Biophys. J.* **2004**, *87*, 3035–3049.

(90) Muddana, H. S.; Gilson, M. K. Calculation of Host-Guest Binding Affinities Using a Quantum-Mechanical Energy Model. *J. Chem. Theory Comput.* **2012**, *8*, 2023–2033.

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

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

(93) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

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

(95) Kolossváry, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.

(96) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

(97) Wales, D. J. Global Optimization of Clusters, Crystals, and Biomolecules. *Science (80-.).* **1999**, *285*, 1368–1372.

Chapter 5: Differentiating Receptor Subtype Specificity in Proteasome Inhibitors

5.1 Introduction

The proteasome is a cellular supramolecular catalytic protein complex that plays an essential role in cellular proteolysis, maintaining homeostasis and controlling signaling pathways.[1–4] The fully assembled constitutive 26S proteasome is characterized by a barrel-like catalytic 20S core particle (20S CP) capped at both ends by regulatory domains, the 19S regulatory particles (19S RP), that control the entrance of proteins to the catalytic sites.[5] This complex segregation of proteolytic enzymes from the cytosolic environment ensures orderly, specific protein degradation which is critical to normal cellular function.[2] Disruption of the proteasome regulation pathway is implicated in several forms of cancer as well as the pathogenesis of human diseases.[6,7] Targeting the proteasome of malignant or diseased cells with designed inhibitors has been revealed as a successful avenue for the treatment of specific types of cancer, such as multiple myeloma.[3,8–10] Inhibition of the proteasome active sites blocks protein degradation, inducing apoptosis, particularly in tumor cells.[11]

Structurally, the 20S CP consists of four homologous rings of proteins with the domain sequence $\alpha_7\beta_7\beta_7\alpha_7$.[1,12] The outer α rings are composed of seven non-catalytic α domains, while the inner β rings each contain three catalytic domains (β1, β2, β5) and four non-catalytic domains (β3, β4, β6, β7). In the mature 26S proteasome, each catalytic domain has an active nucelophilic N-terminal threonine residue that is responsible for peptide bond hydrolysis.[13] However, each catalytic domain has unique activity, specifically caspase-like (β1), trypsin-like (β2), and chymotrypsin-like (β5) activities.[14,15] The

specificity of the domains is the result of molecular recognition processes brought about by subtly different chemical environments in the binding pocket surrounding the active threonine residues in each domain. Understanding the specificity of novel proteasome inhibitors to each of the receptor subtypes can predict their potential to block protein degradation, and thus promote apoptosis.[11]

A class of irreversible proteasome inhibitors, known as the syrbactins, has been discovered and researched in recent years.[9,16] This class of inhibitors is characterized by a molecular structure derived from the natural products syringolin A and glidobactin A. One such derivative, referred to here as TIR-199 (Figure 5.1), has shown promise as a future cancer therapeutic drug.

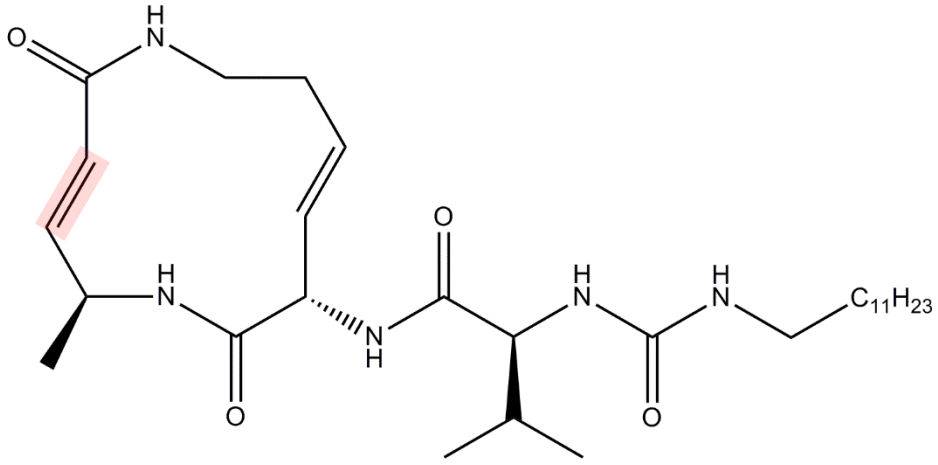

Figure 5.1. Two dimensional structure of the syrbactin derivative, TIR-199. The twelve-membered ring is characteristic of syrbactin molecules. The highlighted double bond is involved in covalent bonding to the N-terminal threonine of a proteasome active site.

In this study, a computational affinity assessment of the syrbactin-derivative proteasome inhibitor TIR-199 covalently bound to each of the three receptor subtypes is presented. Energetic and structural analysis is performed to elucidate the primary environmental differences between the receptor subtypes that affect the subtype specificity of the drug.

5.2 Method

5.2.1 Overview

The human proteasome and syrbactin proteasome inhibitor TIR-199 were modeled computationally. The ligand TIR-199 covalently binds *in vivo* via a Michael addition reaction between the outer syrbactin ring double bond (Figure 5.1) and the N-terminal threonine hydroxyl oxygen of a proteasome active site. As such, the ligand was modeled and simulated in the bound state. Covalent docking of the inhibitor in each of the three binding pockets was performed to explore the conformational states of the covalent complex. Docking poses were used as starting conformations for molecular dynamics simulations.[17] Energetic decomposition of MMGSA interaction energies was then performed on the resulting trajectories, as well as an analysis of the conformational stability of the ligand.

5.2.2 Homology Modeling

Homology models of the human proteasome were created, using known mammalian crystal structures as the reference coordinates.[1] The sequence of each β-subtype protein was provided as input to the online SWISS-MODEL homology modeling software, along with the reference PDB structure and appropriate chain identifier.[18] The bovine

proteasome crystal structure (PDB id: 1IRU) was selected for use as the reference structure due to its sequence similarity (>= 99.6%) to that of Homo sapiens.

5.2.3 Structure Preparation

Molecular recognition takes place within a distinct binding pocket at each catalytic site in the proteasome, which includes both the active catalytic enzyme as well as the domain directly adjacent to the active site (Figure 5.2). Therefore, three binding pocket PDB structures were constructed from the homology models for use in all subsequent calculations, simulations, and analyses: caspase-like binding pocket (β1 and β2), trypsin-like binding pocket (β2 and β3), and chymotrypsin-like binding pocket (β5 and β6).

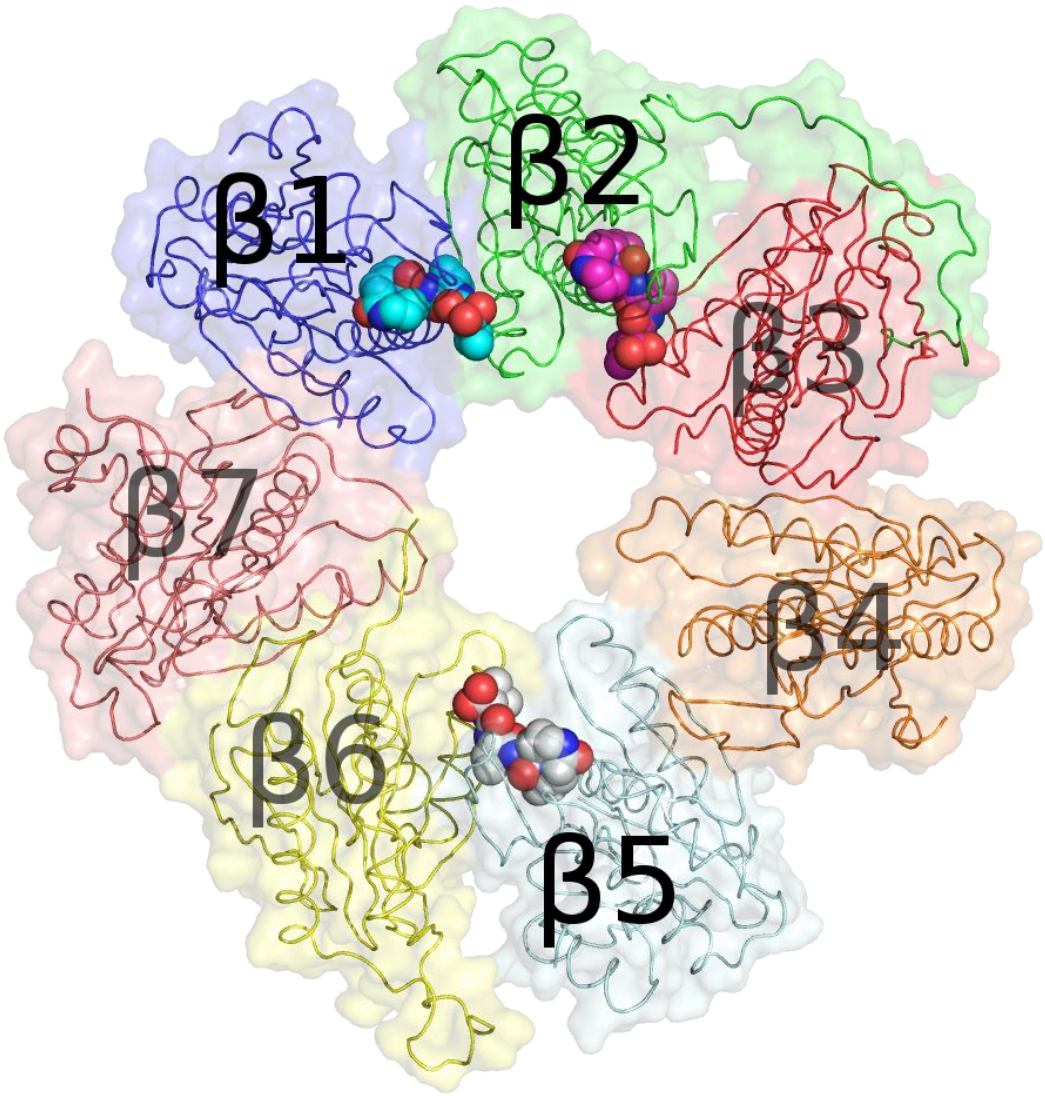

Figure 5.2. A proteasome β ring taken from PDB entry 2ZCY.[9,19] The catalytic active sites are shown with a covalently bound syringolin A ligand.

Using the crystal structure coordinates of the primary syrbactin ring of syringolin A from PDB entry 2ZCY as a starting point, the structure of TIR-199 was constructed manually using the Avogadro molecular modeling software.[9,19] All modeling and simulations

procedures were performed on the ligand in the bound state. Modeling of this covalent system is complicated by the fact that separate forcefields must be used for the protein and ligand, and subsequently merged into a single system. To accommodate this merging process, the structure of the syrbactin ligand was constructed with a lone attached threonine oxygen acting as a placeholder for constructing the final bound-state complex.

5.2.4 Docking

Covalent docking of the ligand TIR-199 to the three distinct proteasome binding pockets was performed using AutoDock 4.2.[20] The ligand was prepared as an AutoDock PDBQT formatted file using the AutoDockTools graphical interface. The three prepared binding pockets were converted to PDBQT files, and the grid and docking parameter files were generated using AutoDockTools.

Modifications were then made to the parameter files, ligand structure, and protein structures to accommodate covalent docking sessions. The ligand structure file was modified, replacing the threonine oxygen placeholder atom attached to the syrbactin ring with a special anchoring "Z" type atom. Covalent anchoring is achieved with a potential well that only affects the Z-type atom, funneling this anchor atom into the zero-point energy at the well minimum.

To the parameter files, a “map” command was added to generate and include the special Z atom type potential map in the docking session. For example, the grid and dock parameter files for the caspase-like binding pocket were modified from:

```
map Caspase.C.map                    # atom-specific affinity map
map Caspase.HD.map                   # atom-specific affinity map
map Caspase.OA.map                   # atom-specific affinity map
map Caspase.N.map                    # atom-specific affinity map
```

to:

```
map Caspase.C.map                    # atom-specific affinity map
map Caspase.HD.map                   # atom-specific affinity map
map Caspase.OA.map                   # atom-specific affinity map
map Caspase.N.map                    # atom-specific affinity map
map Caspase.Z.map                    # atom-specific affinity map
```

The “ligand_type” command was modified to include the Z atom by appending the letter Z to the command:

```
ligand_types C HD OA N Z
```

A “covalentmap” command was added to the grid parameter file to define the potential well properties and potential minimum position for anchoring the Z atom of the ligand:

```
covalentmap 13.0 1000.0 18.6940 15.9240 58.8120
```

The first two terms of the command define the potential well half-width and maximum potential energy outside the well. These were configured to achieve a smooth energetic transition from the edges of the docking region to the well minimum at the anchor point. The well minimum anchor position was set as the binding site coordinate of the N-terminal threonine’s hydroxyl oxygen. To prevent steric clashes in the docking session between the anchor atom on the ligand and the actual hydroxyl group on the N-terminal threonine, the N-terminal threonine hydroxyl group was removed from the binding pocket structure files. These modifications have the effect of anchoring the ligand to the

appropriate location in the active site in each binding pocket while the rest of the molecule can freely explore the surrounding chemical environment (Figure 5.3).

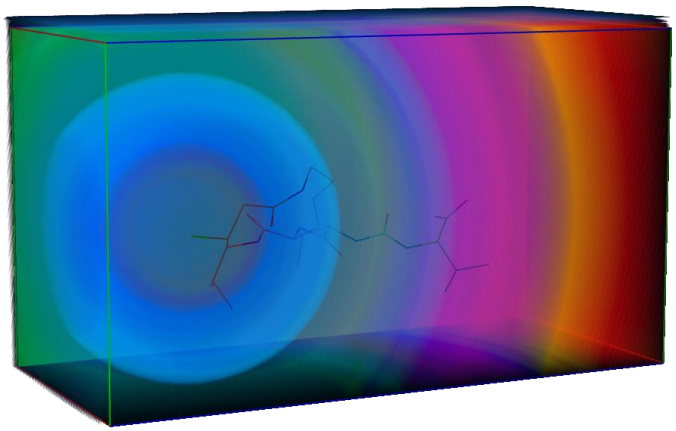

Figure 5.3. Radial potential generated within the rectangular docking bounds by the AutoDock "covalentmap" command. A syrbactin ligand is depicted with the Z-type anchoring atom positioned at the potential well zero-point energy minimum. The potential energy increases radially from the minimum value of 0 kcal/mol (blue) to the maximum value of 1000 kcal/mol (red).

Docking sessions were run with a population size of 150, with 2.5e7 maximum energy evaluations, and 2.7e4 maximum generations. The three lowest energy bound-state complex conformations for each binding pocket were exported to PDB format.

5.2.5 Molecular Mechanical System Preparation

The ligand structure, with the placeholder oxygen attached to the syrbactin ring, was parameterized with the AMBER GAFF force field using the ANTECHAMBER, PARMCHK, and TLeAP software utilities in the AMBER10 software package.[21–23] The threonine oxygen placeholder bound to the prepared ligand structure was temporarily

parameterized as a hydroxyl functional group for later modification upon covalent attachment to each receptor structure. Each of the three binding pocket PDB structures was parameterized with the AMBER FF99SB protein force field using the TLeAP software. The protein structures were minimized to ensure a stable starting conformation. The hydrogen atoms were first minimized with the rest of the protein held fixed, followed by a minimization of the side chain atoms.

In order to form a covalent bond between the parameterized ligand and protein structures, an AMBER force field modification file, or FRCMOD, was created to bridge the bonded parameter terms between the GAFF (ligand) and FF99SB (protein) atom types. The parameters for the bond, angle, and dihedral terms were taken from the GAFF force field:

```
BOND
CT-oh     648.00   1.214

ANGLE
n -c3-c2    66.341      111.760
c3-oh-CT    62.1        113.935
CT-CT-oh    50.0        108.836
H1-CT-oh    50.0        110.856

DIHE
c3-oh-CT-H1        1     0.167         29.480            1.000
c3-oh-CT-CT        1     0.167        -88.000            3.000
```

These terms create bond, angle, and torsion parameters that bridge the ligand structure and the N-terminal threonine residue. Ligand atom types are notated in lower case while the protein atoms are uppercase. This FRCMOD file was loaded into the XLeAP visual parameterization software, along with the parameterized structures of the ligand and one of each of the three binding pockets. The N-terminal threonine hydroxyl group was deleted manually from the loaded structures, and the hydrogen was deleted from the temporary hydroxyl group on the syrbactin ring. A bond was then manually drawn

between the syrbactin-ring oxygen (atom type *oh*) and the N-terminal threonine side chain carbon atom (atom type *CT*) of each binding site to establish a covalent bond. The generated topologies were used in conjunction with bound-state complex conformations generated by the docking procedures to define nine total systems: three conformational states for each of the three binding pocket complexes. No explicit water molecules were added to the system. Instead, the generalized Born implicit water model was used for all minimization and simulation procedures.

The nine systems were initially prepared for molecular dynamics simulations through a set of minimization procedures with the NAMD2 simulation software.[24,25] The first minimization was performed with an entirely fixed protein structure, allowing the ligand to adjust its conformation relative to the receptor. This was necessary because the conformations of the complexes were defined according to the scoring function of the docking software, and needed to be minimized according to the molecular mechanical potential energy function. A second minimization was performed with the backbone atoms of the proteins constrained, allowing mutual side chain and ligand conformational adjustment. A final short minimization was run with the system unconstrained to ensure no backbone atoms were in unfavorable positions.

Langevin dynamics was used to simulate all systems, and was selected to control temperature as well as impart random solvent jostling to complement the generalized Born electrostatic solvation model. A series of three equilibration simulations were performed on each of the nine systems. Ten thousand simulation steps were performed at 100K, 200K, and 298K. Production simulations of each system were performed for 10

nanoseconds simulation time at 298K. Computations were executed on the XSEDE national super computer system. For all simulations, a selection of protein atoms within 10Å of the ligand's initial positions defined the unconstrained, mobile region of the simulation, while the rest of the protein structure was held fixed.

5.2.6 Energetic Analysis with MMGBSA

Trajectories of the molecular dynamics simulations were subject to molecular mechanics/generalized Born/surface area (MMGBSA) energetic analysis.[23,26] The last 5 nanoseconds of each simulation trajectory were isolated and converted to AMBER MDCRD formatted trajectory files. Using a self-authored analysis automation script, MMGBSA-Decomp.py, the MMGBSA interaction energy was calculated for every frame in each trajectory, and the energy decomposed into per-residue interaction energies between each binding pocket residue and the ligand. The script separates the per-residue interaction into polar interactions, combining Coulombic and generalized Born interaction energies, and non-polar interactions, combining van der Waals and non-polar solvation solvent-accessible surface area (SASA) interaction energies.

5.3 Results

5.3.1 Structural Analysis

Simulation of the caspase-like binding receptor with the bound TIR-199 ligand revealed two stable ligand conformations. Steric complementarity and non-polar interactions stabilized the tail region of the ligand in the lower cleft of the binding pocket (Figure 5.4, left). A folded ligand conformation, characterized by the tail region folded back towards the syrbactin ring, was sampled in all simulations (Figure 5.4, right). The root mean

square deviations (RMSD) of the trajectories, relative to their starting positions, were 0.736Å±0.235, 0.634Å±0.367, and 0.768Å±0.237. This analysis revealed that the ligand moved away from its initial conformation at the beginning of the simulation. It also confirmed that the tail region was mobile and alternated between the aforementioned conformations throughout each of the simulations.

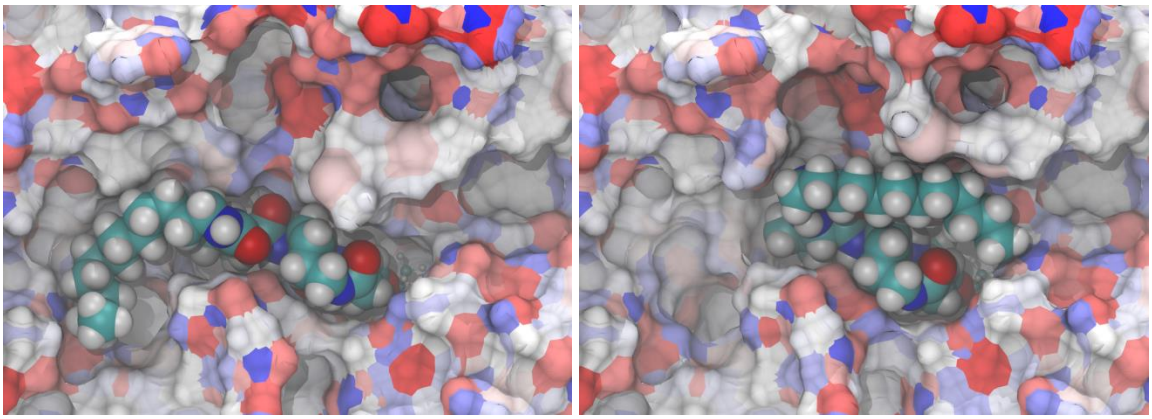

Figure 5.4. Most commonly sampled conformations of TIR-199 in the caspase-like receptor. The protein receptor is colored according to the partial charge distribution.

The simulation of the ligand in the trypsin-like receptor resulted in three similar trajectories. The tail region in all simulations was highly mobile and interacted temporarily with many regions in the binding pocket. The presence of charged residues around the entire binding site prevented a high residency in any one region. A deep pocket extends behind the binding site of this receptor. However, the pocket contains charged and polar residues, preventing favorable interaction with the hydrophobic tail region of the ligand. The large size and polar nature of the pocket suggests the pocket may be occupied by water molecules *in vivo* and *in situ*, rendering the pocket inaccessible

to the hydrophobic tail. No interaction was observed between the tail region of the ligand and this deep, polar pocket in any of the three simulations. RMSD calculations showed that one starting conformation resulted in high mobility of the free portions of the ligand with a value of 0.831Å±0.296. The other two starting conformations resulted in RMSD values of 0.611Å±0.282 and 0.680Å±0.396. On average, the coordinates of the ligand in the trypsin-like receptor were only slightly more stable on average than the caspase-like receptor.

The trajectories resulting from the ligand-chymotrypsin-like receptor complex revealed a uniquely stable conformational state. The chymotrypsin-like receptor contains a deep pocket behind the binding site. The initial conformation with the highest predicted affinity from the docking runs was defined by the ligand's hydrophobic tail extending into this pocket (Figure 5.5). This interaction stabilized the ligand conformation for the majority of the simulation, and resulted in the smallest RMSD value (0.397Å±0.370) of all simulations performed. The large size and few polar interaction sites that exist in the deep pocket suggest that, similar to the trypsin-like receptor, water molecules may inhabit the region prior to binding. However, the chymotrypsin-like deep pocket contains a channel through the rear of the pocket to the bulk solution that is adequately sized to accommodate the exit of water molecules, potentially facilitating desolvation.

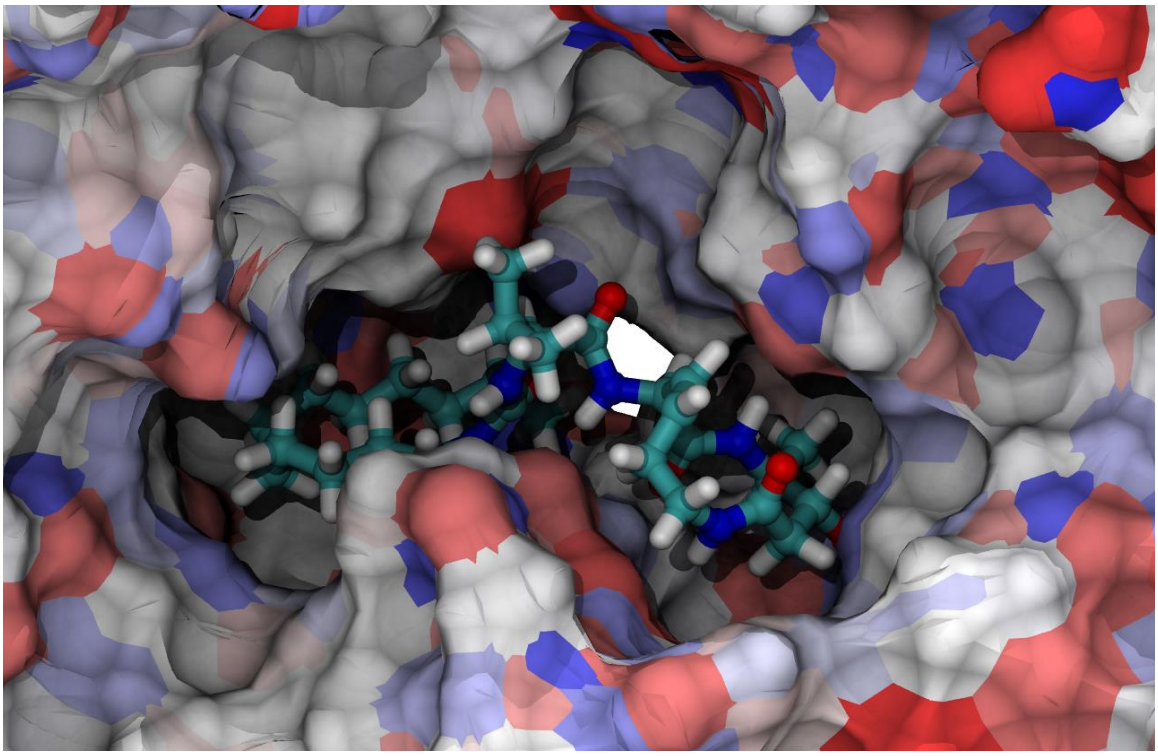

Figure 5.5. Lowest energy docked and simulated conformation of TIR-199 in the chymotrypsin-like receptor. The hydrophobic tail region of the TIR-199 ligand extended deep into the binding pocket, and remained stable for the majority of the simulation. The protein receptor is colored according to the partial charge distribution.

The other starting conformations for the chymotrypsin-like receptor resulted in simulation trajectories with highly mobile ligand tail regions, similar to the caspase-like binding pocket simulation results. The tail regions transiently explored many conformations, interacting with the superficial regions of the receptor. The tail region most commonly extended to and interacted with the hydrophobic clefts (Figure 5.6). In several cases, the hydrophobic tail region extended into the solvent, and then folded over itself into a globular structure. The RMSD values for these two simulations were

0.583Å±0.220 and 0.587Å±0.284. On average, the coordinates of the ligand in this pocket were significantly more stable than either of the two binding pockets.

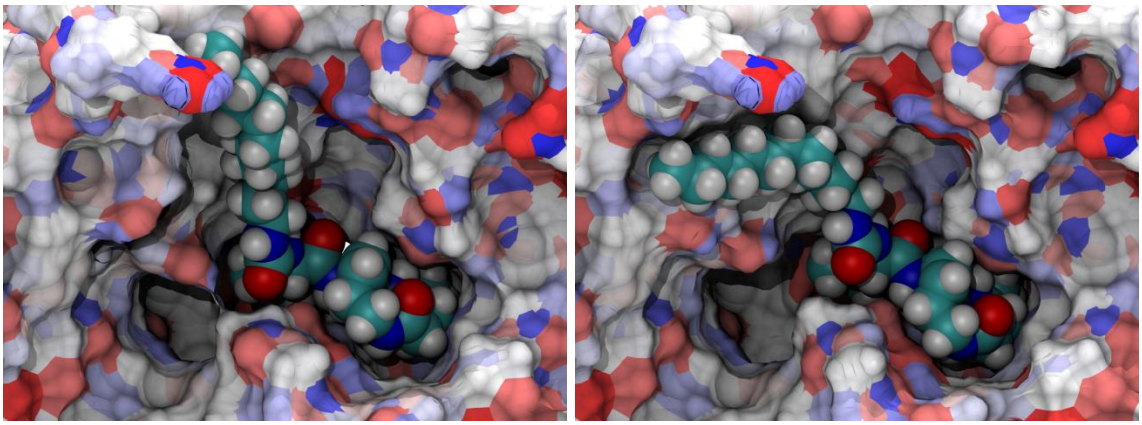

Figure 5.6. Stable alternate conformational states of TIR-199 in the chymotrypsin-like receptor. The ligand interacts with hydrophobic clefts. The protein receptor is colored according to the partial charge distribution

5.3.2 Energetic Analysis

The average affinity of TIR-199 in the caspase-like, trypsin-like, and chymotrypsin-like binding pockets is presented in Figure 5.7. The simulation with the most favorable average intermolecular interaction for each receptor is presented in Figure 5.8, demonstrating the sensitivity of simulation results to the initial conformation.

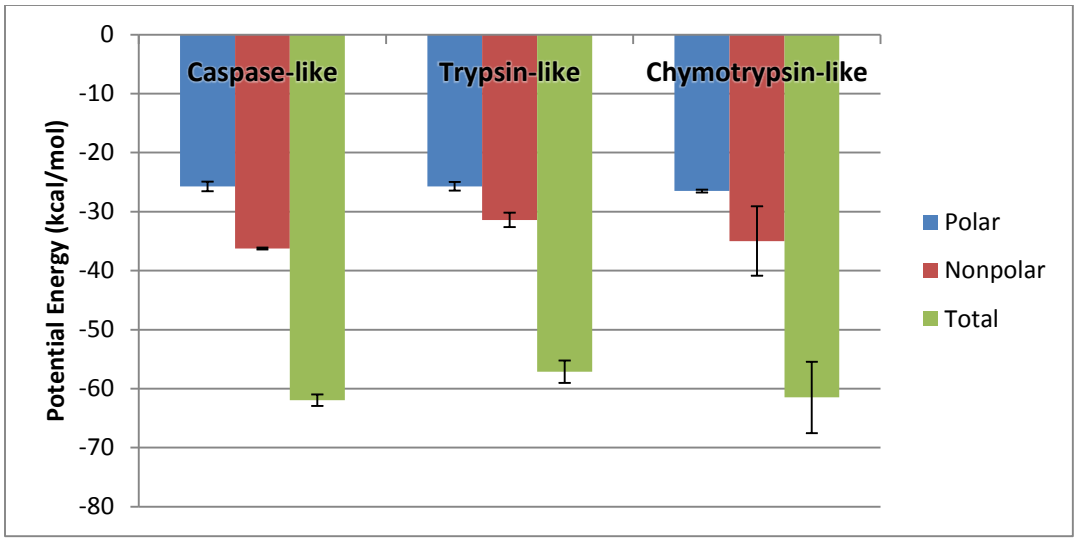

Figure 5.7. Average MMGBSA interaction between the ligand TIR-199 and each receptor. The values for each receptor are averaged across each of three molecular dynamics simulations with unique ligand starting conformations.

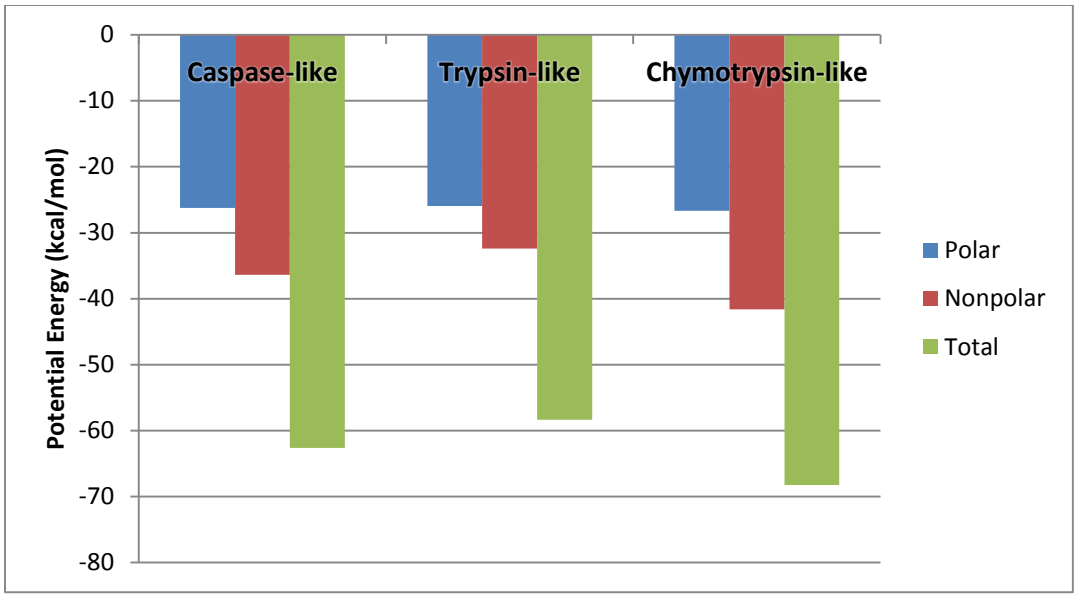

Figure 5.8. MMGBSA interaction between the ligand TIR-199 and receptors, averaged over the simulation trajectory with the most favorable interaction.

The polar contribution to the interaction was consistent across the three receptors, with little variance in the polar affinity between each of the three conformational states of each binding pocket. The polar interactions in these complexes were decomposed into per-residue interactions. This revealed that polar interactions occur primarily between the receptors and the syrbactin ring of the ligand. The primary protein residues of polar interaction were amino acids 16 – 35 and 44 – 56 in each of the mature catalytic subunits. The consistency of the interactions is due to the polar complementarity of the ring with the receptor, which is held fixed in the favorable position by the covalent bond. The differentiating interaction between the three binding pockets was thus the non-polar affinity, occurring between the receptor and the flexible non-polar tail region of the ligand.

The affinity of the ligand to the trypsin-like binding pocket was low relative to other receptor subtypes due to a lack of non-polar stabilization of the tail region. As mentioned previously in the structural analysis, the hydrophobic tail region was highly mobile, primarily interacting with residues 127 – 131 and 166 – 170 of the β2 catalytic subunit, as well as residues 94 – 106, 122 – 131 of the neighboring β3 subunit.

The caspase-like receptor stabilized the tail region of the ligand in the clefts of the binding site. This led to a significantly higher affinity relative to the trypsin-like binding site. The ligand interacted with residues 125 – 135 and residues 112 – 120 of the neighboring catalytic β2 subunit (Figure 5.4). The residues of interaction in this neighboring catalytic subunit are very close to the β2 active site N-terminal threonine.

The ligand conformation in the most stable and highest-affinity simulation had a high residency time with the hydrophobic tail of the ligand nearest to the β2 active site. These interactions may affect the association or binding process of another ligand molecule to the trypsin-like binding site due to conformational perturbation of the trypsin-like active site by the long tail-like region of the ligand.

The high non-polar affinity of TIR-199 to the chymotrypsin-like receptor is brought about through complementarity between the hydrophobic hydrocarbon tail of the ligand and a deep pocket behind the binding site (Figure 5.5). The primary protein residues of non-polar interaction in this pocket were amino acids 121 – 146 of the neighboring mature β6 subunit. This interaction may act as a mechanism of stabilization for the initial encounter and subsequent covalent binding process. Of the three initial ligand conformations, a unique rotation after the peptide bond was required for the ligand to extend into the deep pocket. This also had the effect of lowering the internal energy of the ligand by 2.5 - 3 kcal/mol relative to the other starting conformations. These advantages suggest a non-covalent preference to the chymotrypsin-like receptor. The other starting conformations for the chymotrypsin-like binding pocket resulted in affinity similar to that of the caspase-like receptor, primarily interacting with residues 92 – 109, residues of the binding pocket clefts, in the neighboring β6 subunit (Figure 5.6).

Future modification to the TIR-199 structure for increased stability and affinity may include functional modification to the hydrophobic tail region of the ligand. Due to the presence of polar interaction sites in the binding pocket of all receptor subtypes, polar functionalization near the end of the tail could provide further stabilization through

hydrogen bonding or electrostatic interaction with surrounding polar and charged residues. In particular, this could dramatically improve conformational stability and affinity to the trypsin-like receptor. However, this could potentially disrupt the uniquely favorable interaction of the tail region with the deep pocket of the chymotrypsin-like receptor, of affect other pharmacologically relevant phenomena.

5.4 Conclusions

The affinity of the syrbactin-derivative TIR-199 was assessed in the bound state of each of the three catalytic proteasome receptor subtypes. A distinct energetic and structural preference was found for the chymotrypsin-like receptor. The deep pocket behind the binding site of the chymotrypsin-like receptor led to exceptionally high affinity and conformational stability throughout the simulated trajectory when starting in this conformation. When bound to the trypsin-like and caspase-like receptors, the ligand had low structural stability. The polar residues surrounding the active site of the trypsin-like receptor energetically penalized the ligand in this receptor relative to the caspase-like receptor. Thus, our model suggests a distinct preference to the chymotrypsin-like subtype, and to a lesser degree the caspase-like receptor subtype, of the human proteasome. Inhibition of the chymotrypsin-like receptor has traditionally been the target of inhibition for drug development due to its ability to substantially reduce proteolytic activity in eukaryotic proteasomes. However, simultaneous inhibition of the chymotrypsin-like and caspase-like receptors is required to inhibit proteolytic activity by greater than 50%.[11] Thus our computational procedure predicts appropriate capacity for inhibitory action on the human proteasome.

5.5 References

(1) Unno, M.; Mizushima, T.; Morimoto, Y.; Tomisugi, Y.; Tanaka, K.; Yasuoka, N.; Tsukihara, T. The Structure of the Mammalian 20S Proteasome at 2.75 A Resolution. *Structure* **2002**, *10*, 609–618.

(2) Adams, J. The Proteasome: A Suitable Antineoplastic Target. *Nat. Rev. Cancer* **2004**, *4*, 349–360.

(3) De Bettignies, G.; Coux, O. Proteasome Inhibitors: Dozens of Molecules and Still Counting. *Biochimie* **2010**, *92*, 1530–1545.

(4) Naujokat, C.; Hoffmann, S. Role and Function of the 26S Proteasome in Proliferation and Apoptosis. *Lab. Investig.* **2002**, *82*, 965–980.

(5) Golab, J.; Bauer, T.; Daniel, V.; Naujokat, C. Role of the Ubiquitin–proteasome Pathway in the Diagnosis of Human Diseases. *Clin. Chim. acta* **2004**.

(6) Schwartz, A. L.; Ciechanover, A. The Ubiquitin-Proteasome Pathway and Pathogenesis of Human Diseases. *Annu. Rev. Med.* **1999**, *50*, 57–74.

(7) A, C. The Ubiquitin Proteolytic System and Pathogenesis of Human Diseases: A Novel Platform for Mechanism-Based Drug Targeting. **2003**.

(8) Orlowski, R. Z. The Role of the Ubiquitin-Proteasome Pathway in Apoptosis. *Cell Death Differ.* **1999**, *6*, 303–313.

(9) Groll, M.; Schellenberg, B.; Bachmann, A. S.; Archer, C. R.; Huber, R.; Powell, T. K.; Lindow, S.; Kaiser, M.; Dudler, R. A Plant Pathogen Virulence Factor Inhibits the Eukaryotic Proteasome by a Novel Mechanism. *Nature* **2008**, *452*, 755–758.

(10) Adams, J.; Kauffman, M. Development of the Proteasome Inhibitor Velcade™ (Bortezomib). **2009**.

(11) Kisselev, A. F.; Callard, A.; Goldberg, A. L. Importance of the Different Proteolytic Sites of the Proteasome and the Efficacy of Inhibitors Varies with the Protein Substrate. *J. Biol. Chem.* **2006**, *281*, 8582–8590.

(12) Groll, M.; Ditzel, L.; Löwe, J.; Stock, D. Structure of 20S Proteasome from Yeast at 2.4 A Resolution. *NATURE- ...* **1997**.

(13) Seemuller, E.; Lupas, A.; Baumeister, W. Autocatalytic Processing of the 20S Proteasome. *Nature* **1996**, *382*, 468–471.

(14) Rivett, A. The Multicatalytic Proteinase. Multiple Proteolytic Activities. *J. Biol. Chem.* **1989**.

(15) Kisselev, A. F.; Akopian, T. N.; Castillo, V.; Goldberg, A. L. Proteasome Active Sites Allosterically Regulate Each Other, Suggesting a Cyclical Bite-Chew Mechanism for Protein Breakdown. *Mol. Cell* **1999**, *4*, 395–402.

(16) Clerc, J.; Groll, M.; Illich, D. J.; Bachmann, A. S.; Huber, R.; Schellenberg, B.; Dudler, R.; Kaiser, M. Synthetic and Structural Studies on Syringolin A and B Reveal Critical Determinants of Selectivity and Potency of Proteasome Inhibition. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6507–6512.

(17) Exploring the Molecular Basis of the Enantioselective Binding of Penicillin G Acylase towards a Series of 2-Aryloxyalkanoic Acids: A Docking and Molecular Dynamics. *J. Mol. …* **2007**.

(18) Schwede, T. SWISS-MODEL: An Automated Protein Homology-Modeling Server. *Nucleic Acids Res.* **2003**, *31*, 3381–3385.

(19) Avogadro: an open-source molecular builder and visualization tool http://avogadro.openmolecules.net/ (accessed Sep 20, 2012).

(20) Morris, G.; Huey, R. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. …* **2009**.

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

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

(23) Case, D.; Darden, T.; Cheatham, T. Amber 10. **2008**.

(24) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(25) Kalé, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: Greater Scalability for Parallel Molecular Dynamics. *J. Comput. Phys.* **1999**, *151*, 283–312.

(26) Tsui, V.; Case, D. A. Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations. *Biopolymers 56*, 275–291.

Chapter 6: Molecular Simulation, Modeling, and Analysis Software Documentation

The original software introduced in Chapters 2-4 are here documented and discussed in technical detail. Both the operation and programmatic composition of the software are defined. In addition, modeling and analysis software that automate tedious or time consuming tasks are presented.

6.1 GeomBD

6.1.1 Overview

GeomBD is a software package for simulating coarse grain (CG) Brownian dynamics (BD) of a small substrate molecule in a spatially organized biomolecular environment. It simulates thousands of replicate small molecule diffusion pathways, in parallel, and calculates the probability of association between the small substrate molecule and one or more target molecular binding sites. In GeomBD simulations, the biomolecular environment is held fixed in space at all times, while the substrate molecule diffuses freely as rigid bodies throughout the environment. The software was designed to analyze the efficiency of intermediate substrate transfer between enzymes in spatially organized engineered bioreactor systems, but is generally applicable to the investigation of biological systems.

The software is compatible with POSIX compliant operating systems, such as Linux, Mac OS X, and BSD. The code is targeted for compilation with the Intel C++ Compiler, but can also be compiled with the GNU compiler collection (GCC) with the CilkPlus extension.[1]

6.1.2 Usage

The execution of a GeomBD simulation requires molecular structure files and an input file that dictates the parameters of the simulation. A GeomBD molecular structure file contains a description of the particles in a single molecule. Each particle in a molecule is described in terms of the position, charge, radius, and mass. The structural description allows for complex multi-particle rigid body structures that are subject to translational and rotational diffusion and intermolecular forces. Flexibility in the structural description allows for complex multi-particle structural descriptions as well as single particles with embedded charge distributions. A GeomBD input file specifies simulation parameters, such as temperature of the system and solvent viscosity, and computational parameters, such as the number of threads to utilize for parallel computing. In addition, the input file imports molecular structure files into the simulation system and provides a set of commands to modify the coordinates of each imported molecule for simulation system setup.

6.1.2.1 Structure preparation

GeomBD contains two utilities for creating CG protein and DNA structures from all-atom representations of molecules. The protein CG utility, CGProtein, and the DNA CG utility, CGDNA, both accept protein data bank (PDB) file formatted input on the command line interface, given the "-i" flag. The output is printed to STDOUT and can be routed to an output file:

```
cgProtein –i Protein.pdb > Protein_CG.gbd
cgDNA –i DNA.pdb > DNA_CG.gbd
```

The CG structure that is created by these utilities is a format specific to the GeomBD software package. It is similar to the PDB file format, but is more flexible and contains additional fields for the charge, radius, and mass of each CG particle. The format is insensitive to whitespace, and allows for flexible numerical precision in the values specified.

While utilities are provided to generate files in the GeomBD structure format, small molecule structures must be created manually. The format is simple enough that a structure can easily be constructed by hand. For the computational experiments outlined in Chapter 2, hydrogen peroxide was modeled as a single neutral particle. The corresponding GeomBD structure file contained a single line:

```
particle   0.0 0.0 0.0  0.0  2.5  34.0
```

The string "particle" defines an atom or CG bead. The three numbers that follow specify the Cartesian X, Y, and Z coordinates. The three values that follow the coordinates are the charge, radius, and mass respectively. All spatial units are specified in Angstroms, charge units in electron charges, and mass in AMU.

With the addition of rigid body partial charge embedding to the GeomBD package, the simulations in Chapter 3 included additional lines for the positions of the partial charges within the CG sphere:

```
particle    0.000  0.000  0.000   0.0     2.5   0.0
particle   -0.173 -0.585  0.430  -0.2528  0.0  16.0
particle    0.451 -0.526  1.206   0.2528  0.0   1.0
particle    0.173  0.584 -0.430  -0.2528  0.0  16.0
particle   -0.451  0.525 -1.206   0.2528  0.0   1.0
```

The CG particle is listed first, and acts as the van der Waals interaction sphere with a radius of 2.5Å. The embedded partial charges were then specified as particles

corresponding to the atomic positions and masses in the all-atom structure. The radii for these embedded particles were set to 0 to indicate they should not be considered as van der Waals interaction sites. These charged particles are essentially embedded within the uncharged van der Waals interaction sphere.

As previously mentioned, rigid body multi-particle structures can also be specified. For example, if a rigid body explicit all-atom representation of hydrogen peroxide was desired, the previous example could be modified to remove the CG bead and specify radii for each atomic particle:

```
particle   -0.173 -0.585  0.430  -0.2528  1.4  16.0
particle    0.451 -0.526  1.206   0.2528  0.8   1.0
particle    0.173  0.584 -0.430  -0.2528  1.4  16.0
particle   -0.451  0.525 -1.206   0.2528  0.8   1.0
```

6.1.2.2 Input file

Explicit Structure Commands

Molecules in GeomBD structure format must be individually imported by the input file into the simulation with the “body” command. The body command is followed directly by a filename for the structure:

```
body Protein_CG.gbd
```

Multiple instances of the same structure file can be imported into the simulation system. The “substrate” command is analogous to the “body” command, but specifies the molecule as the diffusing body of the simulation. It requires a third argument specifying the number of replicate simulations desired. The “substrate” command is required for the simulation to run. For example, 1000 substrate replicate simulations are specified as:

```
substrate Ligand_CG.gbd 1000
```

Modifications to the coordinates of an imported body or substrate set can be performed using “center”, “translate”, and “rotate” commands. These commands are performed on the most recently imported molecule. The “center” command translates the body such that the center of mass is located at the origin. The “translate” command accepts three arguments corresponding to displacements along the X, Y, and Z axes. The rotate command accepts three arguments, in radians, corresponding to rotations about the X, Y, and Z axes. In addition to the structural modification commands, the “D” command can also be issued to explicitly set the diffusion coefficient of the most recently imported substrate in units of $Å^2\,ps^{-1}$.

Implicit Structure Commands

In addition to explicitly importing a structure file, planar and tubular implicit structures can be defined for the simulation system in the input file. These implicit structures are used when a large molecular structure with defined planar or tubular structure are impractical to model and simulate with particles. Instead, the structures are defined as force fields, implicitly defining the geometry through polar and non-polar forces. For example, a nanotube can be introduced into a simulation with the “tube” command, followed by a set of geometry and force parameters:

```
tube 1 0 0   0 100 0   1000   100   10   0.1  -1.0
```

The first three parameters define the orientation of the tube, where (1 0 0) specifies X-axis orientation, meaning the hollow portion of the tube is aligned with the X-axis. The following three parameters specified the position of the center of the tube. The next parameter specifies the length of the tube, followed by a parameter specifying the radius

of the tube. The last three parameters specify force parameters, corresponding respectively to the non-polar field radius, effectively specifying the thickness of the tube, the non-polar field well-depth, specifying the non-polar ligand-tube affinity, and the electrostatic charge of the implicit field.

An implicit plane is specified in a similar manner:

```
plane 0 1 0   0 0 0   800 0 600   10   0.1  -1.0
```

The first three parameters correspond to the vector normal to the plane, where (0 1 0) corresponds to an XZ plane. The next three parameters define the position of the center of the plane. The three parameters that follow specify the dimensions of the plane. The last three parameters configure the non-polar field radius, non-polar field well-depth, and the polar charge of the implicit field.

Simulation Configuration Commands

Additional input file commands configure various aspects of the simulation.

The “temperature” command sets the system temperature, in Kelvin. If this command is not included in the input file, the value defaults to 298K.

The “viscosity” command sets the viscosity of the solvent, in centipoise. This is included in the calculation of diffusion coefficients for the bodies in a simulation that do not have diffusion coefficients set explicitly in the input file. If this is not specified, the value defaults to the viscosity of water.

The “bind” command defines a binding site in the simulation. This command accepts one argument corresponding to the index of a particle in the system to be considered a binding site. Any interaction between a substrate and a particle defined as an active site

with the “bind” command is considered a substrate binding event. This will halt that substrate replica’s simulation and increment the count of bound substrate. The number of bound replicates is used to calculate the final probability of association. Multiple binding sites are supported.

The “boundary” command defines the rectangular periodic boundary dimensions of the system. It accepts three parameters for the length, width, and height of the simulation volume. The resultant volume is used to define the system concentration.

The “time” command defines the maximum length of time to be simulated, in picoseconds, for all replicates. Any substrate that reaches this time limit without interacting with a binding site is considered unbound for the final binding probability calculation.

The “threads” command defines the number of threads to be used for parallel computing. This command accepts an integer value. If not assigned, threads will be created equal to the number of available CPUs on the host computer.

Execution

GeomBD simulations are run from the command line. Three command line arguments are required, and must be specified by argument flags. For example:

```
geomBD –i INPUT –o TRAJECTORY.pqr –l LOG
```

This example command would execute a GeomBD simulation with a system specified by the inputfile *INPUT*. Logging information would be output to a file named *LOG*, and a structural trajectory file would be output to `TRAJECTORY`*.pqr*. As the file extension

implies, output trajectory files are in the PQR format.[2] Trajectories can be viewed using the Visual Molecular Dynamics[3] (VMD) software by issuing the following command:

```
vmd Trajectory.pqr –pdb Trajectory.pqr
```

Because each replicate simulation advances in time independently at different rates, trajectories are asynchronous with respect to time position of replicates in each frame. The output log file contains information regarding the binding processes during a simulation. The binding time of each successfully bound substrate is reported. As well, the probability of association is continuously output through the simulation to give the user an active view of the simulation progress. The final probability of association, as well as the elapsed real-world simulation execution time, is reported at the end of the log file.

6.1.3 Composition

GeomBD is written in C++, and makes use of the CilkPlus extension of the Intel C/C++ compiler for parallel execution of various parts of the simulation. The source code can also be compiled with the Gnu Compiler Collection (GCC) with the CilkPlus extension. The C++ classes contained in the code base are described below.

6.1.3.1 StringParser

A simple base class that provides a text parsing interface.

Member Properties

None

Member Functions

```
bool StringParser::parseNextValue(string *buffer, string *value);
```

This function accepts a string pointer `buffer`, stores the next white-space separated token in `value`, and front-truncates `buffer` to remove the token and any surrounding white-space.

6.1.3.2 Particle

`Particle` is a simple storage class for an individual particle description. Every child particle of a rigid body is represented by an instance of this class.

Member Properties

```
double q; //partial electron charge
double r; //radius, A
double m; //mass, amu
vertex R; //position
vertex F; //force
```

6.1.3.3 Body

`Body` is a class that stores instances of `Particle` that represent the components of a rigid body. This class also calculates and stores a description of the assembled rigid body. In addition, member functions are provided to manipulate the coordinates of the child `Particle` objects in unison about the rigid body center of mass.

Parent Class

`StringParser`

Member Properties

```
//particle pointer storage
vector<Particle*> particles;
int    N;     //number of child particles
double m;     //total mass
double I;     //moment of inertia
double r;     //diffusive radius
double r_max; //maximum protrusion
double D;     //translation diffusion coefficient
double Da;    //rotational diffusion coefficient
vertex R;     //linear position
vertex Ra;    //angular position
vertex F;     //linear force
vertex Fa;    //angular force
bool   bound; //bound state
//restraints
bool fixed;
//time stepping
double dt;           //current time step
double t;            //current time
double l_min;        //minimum distance to interaction
vector<double> dt_l; //time step thresholds
```

Member Functions

```
Body::Body(string filename);
```

The constructor initializes all default values and calls member functions to parse the structure file. `Particle` objects are allocated when requested.

```
void Body::bake();
```

After the creation of all particles in the constructor, the `bake` function is called. The rigid body properties are defined here. The total mass is summed and the moment of inertia is calculated. The property `r_max` is the maximum protrusion distance on the rigid body. The `Particle` with position most distant from the center of mass is determined, and its distance from the center of mass, plus its radius, defines the value of `r_max`.

```
void Body::translate(double dx, double dy, double dz);
```

The `translate` function offsets the center of mass position of a body by the factors `dx`, `dy`, and `dz`. In addition, all constituent particles are translated by the same offsets.

```
void Body::rotate(double dax, double day, double daz);
```

The `rotate` function increments the angular position value of the body by `dax`, `day`, and `daz`. The positions of all constituent particles are transformed according to the rotations.

6.1.3.4 BarrierPlane

`BarrierPlane` is a simple storage class for a the parameters that define the geometry and forces exerted by an implicit planar body. An electrostatic field and Lennard-Jones like field exude from the defined two dimensional plane.

Member Properties

```
bool Ax;  // normal vector
bool Ay;  //   N=(Ax,Ay,Az)
bool Az;  //
vertex R; // center
vertex d; // dimensions
double a; // tube radius
double r; // field radius
double e; // field well depth, vdw
double q; // field charge
```

6.1.3.5 BarrierTube

`BarrierTube` is a simple storage class for a the parameters that define the geometry and forces exerted by an implicit tubular body. An electrostatic field and Lennard-Jones like field exude from the defined cylindrical geometry.

Member Properties

```
bool Ax;  // orientation vector
bool Ay;  //   V=(Ax,Ay,Az)
bool Az;  //
vertex R; // center
double L; // length
double a; // tube radius
double r; // field radius
double e; // field well depth, vdw
double q; // field charge
```

6.1.3.6 Model

`Model` is the primary class in GeomBD. It is responsible for parsing the input file and storing all molecular bodies and simulation parameters. The class contains the Brownian dynamics simulation algorithm and propagates time for all substrate replicates in the simulation system. This class also handles logging and trajectory output.

Parent Class

```
StringParser
```

Member Properties

```
//explicit bodies
vector<Body*> bodies;
vector<Body*> substrate;
//implicit bodies
vector<BarrierPlane*> barrierPlanes;
vector<BarrierTube*>  barrierTubes;
//binding site particle indices
vector<int> bindingSites;
// simulation parameters
int    Nbodies;      //# fixed bodies
int    Nsubstrate;   //# substrate bodies
double T;            //Temperature, K
double viscosity;    //kcal.ps/A^3
double cutoff;       //A, fixed value of 20A
double t_limit;      //simulation time limit, picoseconds
vertex boundary;     //rectangular simulation bounds
//normal random number generator
VSLStreamStatePtr rngNormal; //IntelMKL random number generator
vertex *noise;               //storage for random distribution
//file output
string log_filename; //log output filename
fstream log;         //log file stream
string out_filename; //trajectory filename
```

Member Functions

```
Model::Model(string intputfn, string outputfn, string logfn);
```

The constructor initializes all default values, initializes the random number generator. It also calls member functions to parse the input file, and carries out memory allocation for the model.

```
void Model::parseInputFile(string filename);
```

This method reads and interprets the input file. A call is made to the `parseTerm` member method for each input command that is interpreted, allowing for subclasses to expand the input file vocabulary. This method allocates `Body` and implicit `BarrierTube` and `BarrierPlane` objects, modifies structures, and sets simulation parameters as requested by the contents of the input file.

```
virtual void Model::parseTerm(string *term,
                              string *args, Body *active);
```

The `parseTerm` method offers an interface to the input file parsing process for subclasses of `Model`. The first token from an input file line is passed into the member in `term`, while the rest of the input file line is stored in `args`. The most recently loaded `Body` object is provided in `active` for structural modification commands.

```
void Model::bake();
```

The `bake` method finalizes the definition of a model. The `Nbodies` and `Nsubstrate` values are set appropriately, memory is allocated for the normal random number distribution `noise`, and diffusion coefficients are calculated for all bodies in the simulation. This is called at the end of the `Model` constructor.

```
void Model::calculateDiffusionCoefficients();
```

The diffusion coefficients of substrate rigid bodies are calculated according to the Stokes-Einstein equation:

$$D = \frac{k_b T}{6\pi\eta r} \qquad \text{eq. 6.1}$$

$$D_r = \frac{k_b T}{8\pi\eta r^3} \qquad \text{eq. 6.2}$$

where k_b is Boltzmann's constant, T is the temperature of the system, η is the solvent viscosity, and r is the diffusive radius of the `Body`.

```
void Model::calculateTimestepCutoffs();
```

A variable time step for each substrate is implemented to allow substrate bodies that are diffusing at large distances from all other bodies to advance through time more quickly. The time step for each substrate is varied between 0.05 picoseconds and 50.0 picoseconds, by factors of 10, according to the minimum distance between the substrate and all other bodies. The cutoffs for the use of a time step are calculated according to:

$$\rho_{dt} = cutoff + \sqrt{3 \cdot 2 \cdot D \cdot (2\alpha)^2 \cdot dt} \qquad \text{eq. 6.3}$$

where *dt* is a timestep, *D* is the translational diffusion coefficient of the substrate, *cutoff* is the non-bonded force cutoff, and α is a reasonable estimate of the largest value obtained in the normal random distributions. A value of 7Å is used for α based on the statistical maximum of millions of calculated random distributions.

The timestep cutoffs for each substrate are stored as the square of the value in the `Body` object's `dt_l` vector.

```
void Model::calculateForces(int i);
```

The intermolecular forces between a substrate replicate *i* and each immobile rigid `Body` stored in `bodies` are calculated in the `calculateForces` method. The force variables for the substrate `Body` and each child `Particle` object are all cleared before starting the calculation. Iterating over all `bodies`, a distance calculation is performed between the centers of mass of the substrate *i* and body *j*. From this distance, the `r_max` of substrate *i* and body *j* are subtracted. If this value is less than the `cutoff`, force calculations will be performed between substrate *i* and body *j*. Force calculations are performed between the constituent particles in each body. The distance between two interacting particles, particle *k* and particle *l*, is first calculated. Coulombic and Lennard-Jones-like forces are then calculated for pairs of particles with distance values below the `cutoff` value (Chapter 2).

Next, the forces between a substrate replicate *i* are calculated for any implicit bodies. The closest point to the substrate on the tubular or planar geometry of `BarrierTube` and `BarrierPlane` is calculated. A distance calculation is then performed between the substrate center of mass and the closest point. Distances beyond the cutoff value are excluded from force calculations. The Coulombic and Lennard-Jones-like forces are then calculated.

Finally, the individual `Particle` forces are propagated to the parent rigid body:

$$F = \sum F_i \qquad \text{eq. 6.4}$$

$$\tau = \sum r \times F_i \qquad \text{eq. 6.5}$$

where F is translation force on the rigid `Body`, F_i is the force acting on particle i within the rigid body, τ is the rotational force on the rigid body, and r is the displacement vector of particle *i* relative to the center of mass of the rigid body. Throughout all the aforementioned calculations, the minimum distance between the substrate and all explicit and implicit structures is determined. This minimum distance determines the integration time step for the substrate, where a longer minimum distance corresponds to a larger time step (Chapter 3).

```
void Model::advanceTime(int i);
```

The position and rotational state of substrate body *i* are propagated in time according to Equation 1.29. The `calculateForce` function is then called for substrate *i* to calculate the forces on the substrate particle in the new position.

```
void Model::run();
```

The main loop of the simulation is in the `run` function, which repeats until there are no active simulating bodies. With each step of the main loop, a parallel Cilk `for` loop calls the `advanceTime` function on each substrate. This has the effect of parallel time advancement and force calculations for all bodies, significantly increasing computational efficiency. Every 5000 steps, logging and trajectory data are written to their respective files.

```
void Model::writePQR();
```

Trajectory data is output in the PQR format. Molecules defined in each `Body` are separated into individual chains.

6.1.4 Analysis tools

The primary output of a GeomBD simulation is the association probability. As previously stated, this is explicitly written to the end of a log file upon completion of a simulation. However, more information can be obtained from simulation results. Two Python[4,5] scripts, making use of the Matplotlib[6] graphing module, are supplied with the GeomBD software to visually display the binding distribution of a simulation or set of simulations.

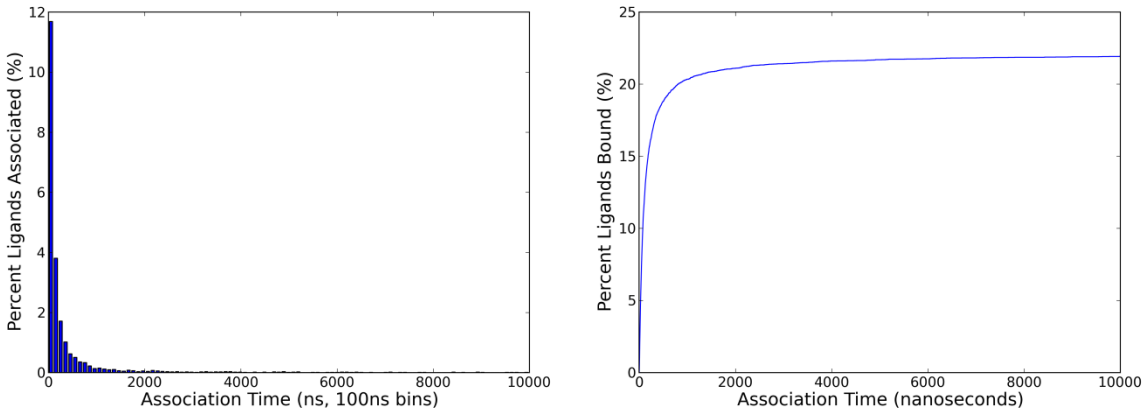

Figure 6.1: Comparison of the output of the plotHistogram.py (left) and plotCumulative.py (right) scripts on the same set of simulation data. The histogram was produced with a bin width of 100 nanoseconds.

The "plotHistogram.py" accepts command line input arguments of one or more GeomBD log file. This script reads in all binding times in each log file, bins the binding times

according to user preference, and produces a histogram showing the distribution of the probability of binding. Similarly, the “plotCumulative.py” script reads in all binding times from one or more log files. However, instead of binning and plotting as a histogram, this script calculates and plots a cumulative probability as a function of time. The graphs created by both scripts can provide visual insight into the binding kinetics and dynamics of a biomolecular system.

6.1.5 Derivative Application – RBBD

A software derivative of GeomBD, called RBBD, was developed for general purpose rigid body Brownian Dynamics (RBBD). All classes and functions remain intact. In RBBD, the concept of parallel replicate simulations is removed. Instead, all bodies in the simulation are mobile, interactive, and diffuse throughout the simulation at the same rate. RBBD was constructed to evaluate the association probability of substrate-protein association in freely diffusing solutions. Parallelization of force calculation and time advancement replaces replicate parallelization.

6.2 Python Molecular Mechanics Framework Module

6.2.1 Overview

The Python Molecular Mechanics Framework Module, or PYMM, is a Python module framework for the programmatic description of all-atom molecular mechanical systems (see Chapter 1 for molecular mechanics theory). It is used as a foundation for creating computational methods or utilities that interact with molecular mechanical systems. PYMM contains a set of Python classes that generically describe the topology and parameters of a molecular system, including the details of all atoms, bonds, angles, dihedral angles, and non-bonded potentials. A set of interpreter modules parse various file formats of molecular mechanical systems, including AMBER and Mining Minima Generation 2 (M2) topologies and parameters, and produce a set of Python objects that fully describe the system. This makes interaction with the data in a molecular mechanics system quick and easy, and requires no end-user knowledge of proprietary file formats. Operations can then be performed on the system, such as energy calculations, MMGBSA calculations, structural minimizations, RMSD calculations, and alignment procedures. First and second derivatives of the potential energy function can be calculated.

6.2.2 Composition

PYMM is a Python-C module. The majority of code in the module is written in C using the C Python interface, defining classes and functions programmatically. The primary Python module interface is stored within the *mm/* directory. All C files are stored within the *mm/c/* directory.

The primary class, *model,* for the molecular mechanical system description, is defined in *model.c* and *model.h*. The model class stores pointers to all atomic and topological objects, as well as pointers to all functions, such as *minimize* and *calculate_energy*.

The *align.c, qcprot.c, and qcprot.h* files provide an interface to QCProt RMSD alignment algorithm. It is utilized for the calculation of RMSD and alignment via RMSD minimization.[7] It provides the function *align* to the *model* class.

The *energy* source files calculate the individual components of potential energy function. Energy, force, and hessian calculations are supported for the various components of the bonded and non-bonded potentials. It provides the function *calculate_energy* to the *model* class.

The Generalized Born radii, energy, and force calculations are supported via the *gb.c* file.[8] It provides the function *calculate_energy_gb* to the *model* class.

The solvent accessible surface area is calculated in *sasa.c*. The non-polar solvation energy contribution is calculated according to the linear scaling scheme of AMBER.[9] It provides the function *calculate_energy_sasa* to the *model* class.

A basic steepest descent minimization algorithm is provided in *minimize*.c. It provides the function *minimize* to the *model* class.

Double precision storage classes that act as a bridge between C and Python are defined in *vector.h, array.h,* and *matrix.h*:.

File Format Interpreters

Two interpreter modules are available for AMBER and M2 topology and parameter files. The interpreters are written in Python, making the writing of an interpreter module quick and facile relative to a systems language like C++.

The AMBER interpreter is stored in *mm/amber/__init__.py* and provides an interface to the PRMTOP topology file format through the *prmtop* class. The AMBER v7+ formatted coordinate files are supported with the *inpcrd* class, while the trajectory files are interpreted with the *mdcrd* class. An example script that calculates the MMGBSA interaction energy for a protein-ligand complex is given below.

```
import mm, mm.amber


top = mm.amber.prmtop('Complex.prmtop')
crd = mm.amber.mdcrd('Complex.mdcrd', top.natoms)
model = mm.amber.model(top, crd)
model.gb = 1 # enable GB calculation


print('system contains %d atoms, %d trajectory frames.' %
          (model.natoms, model.trajectory_in.nframes))


for i in range(model.trajectory_in.nframes):
  print('frame', i)
  model.set_frame(i)
  model.calculate_energy()
  model.calculate_energy_sasa()
  model.print_energies()
```

An M2 interpreter module is provided under *mm/m2/*. The *m2* module provides access to M2 topology files with the *top* class, coordinate files with the *crd* class, and trajectory files with the *trj* class. An example script is given below that calculates the force vector and hessian matrix of an M2 formatted system.

```
import mm.m2


top = mm.m2.top('Receptor.top', 'Ligand.top')  # combine complex
crd = mm.m2.trj('Complex.trj') # load m2 trajectory of complex
model = mm.m2.model(top, crd)
model.hesse = 1      # enable hessian calculation
model.gb = 1         # enable GB


print('system contains %d atoms.' % model.natoms)
print('coordinates contain %d frame(s).' % crd.nframes)


model.calculate_energy()
model.print_energies()
```

Due to the generalized nature of PYMM, it is not the most computationally efficient software. However, the design makes it ideal for creating small modeling utilities and tools. It is also an ideal framework for rapidly prototyping computational methods.

6.3 Hopping Minima

6.3.1 Overview

The Hopping Minima (HM) method is a procedure, and corresponding set of programs, for the determination of association/dissociation pathways. HM connects local energy minima using the molecular system's natural motions. The natural motions of the molecular system are modeled from their normal modes, which are exaggerated into the coordinated motions of a host-guest complex. These coordinated motions are then used to bridge multiple conformational minima of the complex forming an association pathway.

6.3.2 Usage

To begin, the HM method requires an established Mining Minima generation 2 (M2) molecular system with thorough conformational search of a host-guest complex.[10,11] Free energy calculations should be performed on the conformational minima.

With a set of conformational states for a complex, the first HM program, called *HoppingMinimaScan*, is performed on the conformational states to calculate the natural motions of the complex. HoppingMinimaScan is a modified version of the Mining Minima Generation 2 (M2) free energy calculation program.[12] The mode scanning algorithm of M2 is modified to exaggerate the calculated normal modes of the complex three fold into broad motions. These motions are output to a trajectory.

The *HoppingMinima* script can then be run. This is the primary script for the HM method. It requires a set of conformational minima with corresponding free energy values, along with sampled natural motions. After an alignment procedure, the script calculates intersections between the conformational minima and the natural motions of

the complex. Intersections between two or more minima are output to a trajectory as potential association/dissociation paths.

A final HM script, called *HoppingMinimaAnalysis*, is an interactive tool for linking multiple paths generated by the *HoppingMinima* script into a full association/dissociation process. It provides information about potential path connections, and ranks them according to potential energy of the structural transitions during the natural motion, as well as free energies of the minima involved in the path. This energy information is useful for ruling out energetically unlikely paths.

6.3.2 Composition

The *HoppingMinima* script is written in Python using the PYMM Python module. The PYMM *m2* module is used to interpret and interact with the M2 molecular mechanical system. The *HoppingMinima* script utilizes the QCProt alignment functions from PYMM for initial alignment of all natural motions and conformational minima. Intersections between the conformational minima and frames from the natural motions are compared using the QCProt RMSD calculation routine.

The *HoppingMinimAnalysis* script is written in Python. It compares the minima involved in the paths output by *HoppingMinima* to determine which paths involve common minima. These are then clustered and sorted according to the free energies of the minima and the maximum potential energy of the natural motions. The user can interactively search for paths containing specific minima.

6.4 MMGBSA-Decomp.py

6.4.1 Overview

The *MMGBSA-Decomp.py* Python script is a script to automate the execution and analysis of MMGBSA energetic decomposition calculations on a protein-ligand complex using AMBER version 10.[13] Starting from a completed molecular dynamics simulation trajectory with explicit waters removed, the MMGBSA interaction energies are calculated and decomposed into pairwise interaction energies between the ligand and each residue in the protein. This is particularly useful in molecular recognition studies for understanding the most important interactions in a protein-ligand complex, and can suggest avenues for drug design and optimization.

6.4.2 Usage

The *MMGBSA-Decomp.py* is a command line utility that requires an input file as its only command line argument. The input file contains information about the protein-ligand complex AMBER topology, coordinates, and dynamics trajectory. It specifies which residues in the topology correspond to the ligand and protein. Additionally, the input file specifies which frames within the dynamics trajectory should be analyzed. An example input file is given on the next page (note, the variables defined are case sensitive).

```
prmtopFile = 'Complex.prmtop'
coordFile  = 'Complex.rst7'
mdcrdFile  = 'MDTrajectory.mdcrd'

# specify our selection ranges
lres = [1, 1]    # resid 1, our ligand
rres = [2, 412]  # resid 2 to 412, our protein
cres = [1, 412]  # resid 1 to 412

# specify trajectory frame info, base 0
fstart = 0    # start with frame 0
fend = 5000   # end at frame 5000
fstep = 20    # calculate every 20th frame
```

The pairwise MMGBSA interaction energies are output into three separate files: *mmgbsa.polar.csv*, *mmgbsa.nonpolar.csv*, and *mmgbsa.total.csv*. Each row in the file represents a separate frame from the trajectory. Column *i* on each row correspond to interactions between the ligand and i^{th} protein residue. The polar energies, including the Coulombic and Generalized Born energies, are output to the *mmgbsa.polar.csv* output file, while the van der Waals and SASA energies are combined and output to the *mmgbsa.nonpolar.csv* output file. All energies are totaled and output to the *mmgbsa.total.csv* output file.

6.4.3 Composition

The *MMGBSA-Decomp.py* script automates the calculation of MMGBSA interaction energies, as well as the extraction and analysis of the output data. The topology and dynamics trajectory are parsed and interpreted using simplified components from PYMM. Each trajectory frames is extracted into a temporary file by the *mmgbsa_decomp* class of the script, and an AMBER MMGBSA calculation is subsequently executed using the frame as the input coordinates. The *decomp_log* class parses and interprets the energetic data, which is returned in a simplified form to the *mmgbsa_decomp* script for output to the log files.

6.5 libFicus C++ Molecular Mechanics Framework

6.5.1 Overview

libFicus is a C++ molecular mechanics library. It provides an object oriented framework for writing applications utilizing molecular mechanical theory. The class inheritance of C++ is used extensively to abstract geometric and energetic calculations. In addition to providing a framework for the description of molecular mechanical systems, it provides access to alignment routines, dynamics algorithms, and robust FORTRAN minimization algorithms. The objective of this library is to provide a flexible, generic, computationally efficient molecular mechanics interface for the development of molecular mechanics methods, or to integrate molecular mechanics into an existing method.

6.5.2 Composition

6.5.2.1 Atom

`Atom` is class provides storage of atomic properties, such as mass, radius, and charge. It also provides pointers into the main coordinate and potential gradient arrays for the atom.

Member Properties

```
int       id;         //index identifier
Molecule *molecule;   //parent molecule
double    charge;     //partial charge
double    mass;       //atomic mass (amu)
double    radius;     //vdw radius
string    name;       //element identity
string    type;       //storage for ff type info
bool      fixed;      //atom should be fixed in space
double    *position;  //cartesian position x, y, z
double    *gradient;  //gradient of potential energy
```

6.5.2.2 Bond

`Bond` is a base class that represents a covalent bond between two atoms. This base class calculates the distance between the two atoms and provides a generic subclasssable interface to calculating a potential energy function and its derivatives.

Member Properties

```
Atom *atomi;   //pointer to first atom in bond
Atom *atomj;   //pointer to second atom in bond
double rij[3]; //displacement vector between atoms
double lij2;   //squared distance
double lij;    //distance
double U;      //potential energy
double dU;     //  1st derivative
double ddU;    //  2nd derivative
```

Member Functions

```
Bond::Bond(Atom *ai, Atom *aj);
```

The constructor initializes all default values and sets the atom pointers.

```
void Bond::CalculateLength();
```

The atomic displacement vector `rij` is calculated, and the distance `lij` and its square `lij2` are recorded. The square is recorded for efficiency purposes for use in potential energy calculations, if needed.

```
virtual void Bond::CalculatePotential(bool firstDerivative,
                                      bool secondDerivative);
```

A virtual function is provided to subclasses for calculation of specific potential energy forms describing bond forces. In this base class, nothing is executed in this function.

```
void Bond::PopulateGradients();
```

The gradient of the bond's potential energy function that has been recorded to `dU` is propagated to the potential gradient storage in each of `atomi` and `atomj`.

```
void Bond::PopulateHessian(double **hessian);
```

The contributions to the global `hessian`, stored in `ddU` by subclasses, are applied in this member function.

Provided Subclasses

`BondHarmonic` provides a harmonic bond subclass. The potential is calculated as:

$$U_{bond} = \frac{1}{2}k(lij - l0)^2 \qquad \text{eq. 6-x}$$

where `k` is the force constant and `l0` is the reference bond length provided to the object constructor, and `lij` is the atomic distance calculated by the base class.

`BondQuartic` provides a quartic potential bond subclass. The potential is calculated as:

$$U_{bond} = \frac{k1}{2}(lij - l0)^2 + \frac{k2}{3}(lij - l0)^3 + \frac{k3}{4}(lij - l0)^4 \qquad \text{eq. 6.6}$$

where `k1`, `k2`, and `k3` are the force constants and `l0` is the reference bond length provided to the object constructor, and `lij` is the atomic distance calculated by the base class.

`BondMorse` provides a quartic potential bond subclass. The potential is calculated as:

$$U_{bond} = \mathrm{E0}\left(1 - e^{-a(lij - l0)}\right)^2 \qquad \text{eq. 6.7}$$

where `E0` is the potential well depth, `a` is the well width, and `l0` is the reference bond length provided to the object constructor, and `lij` is the atomic distance calculated by the base class.

6.5.2.3 Angle

`Angle` is is a base class defining the angle between three atoms. This base class calculates the angle.

Member Properties

```
Atom *atomi;          //pointer to first atom in bond
Atom *atomj;          //pointer to second atom in bond
Atom *atomk;          //pointer to third atom in bond
double rij[3];        //i,j displacement vector
double rkj[3];        //k,j displacement vector
double lij, lij2;     //i,j distance, and square
double lkj, lkj2;     //k,j distance, and square
double lik;           //i,j distance
double cos_theta;     //cosine of the i,j,k angle
double sin_theta;     //sine of the i,j,k angle
double theta;         //the i,j,k angle
double U;             //potential energy
double dUdcos;        //1st derivative w.r.t. cosine of angle
double dcosdr[3][3];  //1st derivative of cosine theta
double ddUddcos;      //2nd derivative w.r.t. cosine of angle
double ddcosddr[9][9];//2nd derivative of cosine theta
```

Member Functions

```
Angle::Angle(Atom *ai, Atom *aj, Atom *ak);
```

The constructor initializes all default values and sets the atom pointers.

```
void Angle::CalculateAngle();
```

The atomic displacement vectors, atomic distances and their squares are calculated. This information is then used to calculate the cosine and sine of the angle, as well as the angle itself. The square is recorded for efficiency purposes for use in potential energy calculations, if needed.

```
virtual void Bond::CalculatePotential(bool firstDerivative,
                                      bool secondDerivative);
```

A virtual function is provided to subclasses for calculation of specific potential energy forms describing bond forces. In this base class, only the derivatives of the cosine of

the angle are calculated. Subclasses are responsible for calculating the derivative of their potential form with respect to the cosine of the angle.

```
void Angle::PopulateGradients();
```

The gradient of the bond's potential energy function that has been recorded to `dUdcos` is propagated to the potential gradient storage in each of `atomi`, `atomj`, and `atomk`.

```
void Angle::PopulateHessian(double **hessian);
```

The contributions to the global `hessian`, stored in `ddU` by subclasses, are applied in this member function.

Provided Subclasses

`AngleHarmonic` provides a harmonic bond subclass. The potential is calculated as:

$$U_{angle} = \frac{1}{2}k(theta - theta0)^2 \qquad \text{eq. 6.8}$$

where `k` is the force constant and `theta0` is the reference angle provided to the object constructor, and `theta` is the angle calculated by the base class.

`AngleQuartic` provides a quartic potential bond subclass. The potential is calculated as:

$$U_{angle} = \frac{k1}{2}(\theta - t0)^2 + \frac{k2}{3}(\theta - t0)^3 + \frac{k3}{4}(\theta - t0)^4 \qquad \text{eq. 6.9}$$

where `k1`, `k2`, and `k3` are the force constants and `t0` is the reference angle provided to the object constructor, and θ represents the angle variable `theta` calculated by the base class.

6.5.2.4 Torsion

`Torsion` is base class calculates the dihedral angle between four atoms.[14]

Member Properties

```
//pointers to atoms in dihedral angle
Atom *atomi;
Atom *atomj;
Atom *atomk;
Atom *atoml;
// ref: blondel and karplus
//   distance vectors
double F[3], G[3], H[3], A[3], B[3] , C[3];
//   lengths
double lF, lG, lH, lA, lB;
//   dihedral angle
double phi;
//   cosine and sine of angle
double cos_phi, sin_phi;
//potential energy and derivatives
double U;
double dUdphi;
double dphidF[3], dphidH[3], dphidG[3];
double ddUddphi;
double ddphiddF[3][3], ddphiddH[3][3],
       ddphiddG[3][3], ddphidFdG[3][3],
       ddphidGdH[3][3];
```

Member Functions

```
Torsion::Torsion(Atom *ai, Atom *aj, Atom *ak, Atom *al);
```

The constructor initializes all default values and sets the atom pointers.

```
void Torsion::CalculateAngle();
```

The dihedral angle is calculated between the four atoms defined in the torsion object.[14]

```
virtual void Torsion::CalculatePotential(bool firstDerivative,
                                         bool secondDerivative);
```

A virtual function is provided to subclasses for calculation of specific potential energy forms describing bond forces. In this base class, only the derivatives of the angle with

respect to the displacement vectors are calculated. Subclasses are responsible for calculating the derivatives of the potential energy function with respect to the angle.

```
void Torsion::PopulateGradients();
```

The gradient of the potential energy function that has been recorded to `dUphi` is propagated to the potential gradient storage in each of `atomi` and `atomj`.

```
void Torsion::PopulateHessian(double **hessian);
```

The contributions to the global `hessian`, stored in `ddUddphi` by subclasses, are applied in this member function.

Provided Subclasses

`TorsionHarmonic` provides a harmonic torsion subclass. The potential is calculated as:

$$U_{torsion} = \frac{1}{2}k(phi - phi0)^2 \qquad \text{eq. 6.10}$$

where `k` is the force constant and `phi0` is the reference dihedral angle provided to the object constructor, and `phi` is the dihedral angle calculated by the base class.

`TorsionHarmonic` provides a harmonic torsion subclass. The potential is calculated as:

$$U_{torsion} = k(1 + \cos(n \cdot phi - phase)) \qquad \text{eq. 6.11}$$

where `k` is the force constant, `n` is the multiplicity of the potential, and `phi0` is the reference dihedral angle provided to the object constructor, and `phi` is the dihedral angle calculated by the base class.

6.5.2.5 NBPair

`NBPair` is a class for defining the interaction between two unbounded atoms. This class is storage class for use with `NBPotential` objects, and should not be subclassed.

Member Properties

```
Atom *atomi;   //pointer to first atom in bond
Atom *atomj;   //pointer to second atom in bond
double rij[3]; //displacement vector between atoms
double lij2;   //squared distance
double lij;    //distance
double U;      //potential energy
double dU;     //  1st derivative
double ddU;    //  2nd derivative
```

Member Functions

```
NBPair::NBPair(Atom *ai, Atom *aj);
```

The constructor initializes all default values and sets the atom pointers.

```
void NBPair::CalculateLength();
```

The atomic displacement vector `rij` is calculated, and the distance `lij` and its square `lij2` are recorded. The square is recorded for efficiency purposes for use in potential energy calculations, if needed.

```
void NBPair::PopulateGradients();
```

The gradient of the potential energy function that has been recorded to `dU` is propagated to the potential gradient storage in each of `atomi` and `atomj`.

```
void NBPair::PopulateHessian(double **hessian);
```

The contributions to the global `hessian`, stored in `ddU` by subclasses, are applied in this member function.

6.5.2.6 NBPotential

`NBPotential` is is a generic interface to non-bonded potential energy functions. It provides virtual functions for initializing a non-bonded calculator and for calculating the potential energy and its derivatives between a pair of atoms.

Member Properties

```
//cumulative potential energy storage
double U;
//exclusion paramters provided if needed
bool nbexclude11; //exclude self calculation
bool nbexclude12; //exclude atoms in bonds
bool nbexclude13; //exclude atoms in angles
bool nbexclude14; //exclude atoms in torsions
// scale nonbonded potential for atoms in torsions
double scale14;
//enable or distable intra- and intermolecular forces
bool intra;
bool inter;
```

Member Functions

```
NBPair::NBPair(Atom *ai, Atom *aj);
```

The constructor initializes all default values and sets the atom pointers.

```
virtual void InitializeCalculation();
```

An interface is provided for initializing a potential energy calculation. This is called directly before calculating potential energy for individual atom pairs.

```
virtual void CalculatePotential(NBPair *nbpair,
                                bool firstDerivative,
                                bool secondDerivative);
```

A virtual function is provided for subclasses to calculate the potential energy and forces between two atoms in an `NBPair`. Propagation of forces to the individual atoms is handled by the `NBPair` class. In this base class, nothing is calculated.

6.5.2.7 Model

`Model` is a container class for the total description of a molecular mechanical system. It stores pointers to all objects of the system, and coordinates calculations of energy and forces.

Member Properties

```
//object storage
vector<Molecule*> molecules;
vector<Atom*> atoms;
vector<Bond*> bonds;
vector<Angle*> angles;
vector<Torsion*> dihedrals;
vector<Torsion*> impropers;
vector<NBPotential*> nbpotentials;
//coordinates for all atoms in system
double *coordinates; // cartesian coordinates
//potential energy
double U;          //total energy
double Ubond;      //cumulative bond energy
double Uangle;     //cumulative angle energy
double Utorsion;   //cumulative torsion energy
//gradient of total potential, 3xN vector
double *gradient;
//hessian of total potential, 3Nx3N triangle matrix
double **hessian;
```

Member Functions

```
Model::Model(bool includeHessian);
```

The constructor initializes all default values and sets the atom pointers. The constructor argument `includeHessian` determines whether a hessian matrix is calculated. If hessian calculations are desired, this value should be `False`, as the hessian can consume a significant amount of memory.

```
void Model::AllocateArrays();
```

When a system has been constructed, this method is called to allocate adequate memory for calculating energies and forces. This method also assigns the appropriate coordinate and gradient pointers in each atom object into the global `coordinates` and `gradient` arrays.

```
void Model::CreateNBExclusions();
```

According to the parameters of any allocated `NBPotential` objects, a non-bonded calculation exclusion list is created here. If desired, this should be called directly after allocating memory for the system.

```
void Model::CalculatePotential(bool firstDerivative,
                               bool secondDerivative);
```

The potential energy calculation routines in each topological object and non-bonded potential are dispatched here. The total of each energy term is tallied into the `Model` objects storage variables.

Provided Subclasses

`ModelBAT` provides additional functionality for working with the model in terms of internal Bond-Angle-Torsion coordinates.

6.5.2.8 Optimizer

`Optimizer` provides a generic interface for numerical minimization algorithms.

Member Properties

```
Model   *model;  //model to perform optimization upon
```

Member Functions

```
Optimizer::Optimizer(Model *model);
```

The constructor sets the `model` pointer.

```
virtual void Optimizer::Run(int steps);
```

The Run function provides a generic interface for subclasses to override and implement a numerical optimization algorithm.

```
void Model::CreateNBExclusions();
```

Provided Subclasses

L-BFGS and truncated Newton-Raphson minimization algorithms are provided in `OptimizerLBFGS` and `OptimizerTNR` subclasses, respectively. These subclasses utilize the LBFGS and TNPACK FORTRAN libraries and call their functions directly.[15–17] These classes act as a data translation layer between the coordinate and force data stored in a Model object and the FORTRAN functions.

6.6 References

(1) Blumofe, R. D.; Joerg, C. F.; Kuszmaul, B. C.; Leiserson, C. E.; Randall, K. H.; Zhou, Y. Cilk. *ACM SIGPLAN Not.* **1995**, *30*, 207–216.

(2) Dolinsky, T.; Czodrowski, P.; Li, H. PDB2PQR: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic acids …* **2007**.

(3) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**.

(4) Rossum, G. Van. Python Programming Language. *USENIX Annu. Tech. Conf.* **2007**.

(5) Sanner, M. Python: A Programming Language for Software Integration and Development. *J Mol Graph Model* **1999**.

(6) Hunter, J. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**.

(7) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. Fast Determination of the Optimal Rotational Matrix for Macromolecular Superpositions. *J. Comput. Chem.* **2010**, *31*, 1561–1563.

(8) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(9) Case, D.; Darden, T. AMBER 11. *… California, San …* **2010**.

(10) Chang, C.-E.; Gilson, M. K. Tork: Conformational Analysis Method for Molecules and Complexes. *J. Comput. Chem.* **2003**, *24*, 1987–1998.

(11) Chang, C.-E.; Gilson, M. K. Free Energy, Entropy, and Induced Fit in Host-Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.

(12) Chen, W.; Chang, C.-E.; Gilson, M. K. Calculation of Cyclodextrin Binding Affinities: Energy, Entropy, and Implications for Drug Design. *Biophys. J.* **2004**, *87*, 3035–3049.

(13) Case, D.; Darden, T.; Cheatham, T. Amber 10. **2008**.

(14) Blondel, A.; Karplus, M. New Formulation for Derivatives of Torsion Angles and Improper Torsion Angles in Molecular Mechanics: Elimination of Singularities. *J. Comput. Chem.* **1996**, *17*, 1132–1141.

(15) Schlick, T.; Fogelson, A. TNPACK—A Truncated Newton Minimization Package for Large-Scale Problems: I. Algorithm and Usage. *ACM Trans. Math. Softw. (…* **1992**.

(16) Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208.

(17) Zhu, C.; Byrd, R.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans. Math. …* **1997**.

Chapter 7: Conclusion and Future Directions in Substrate-Enzyme Association Research

7.1 Conclusion

Molecular recognition processes guide the preferential association of molecular species in chemical and biological systems through physical differentiation and complementarity. Molecular recognition was studied at several scales by creating and applying novel computational techniques.

The long-range interactions dictating diffusional association were studied through the use of a new computational package, GeomBD. Diffusional association was investigated in the intermediate transfer between enzymes in engineered spatially organized nanostructures. In particular, the glucose oxidase-horseradish peroxidase enzyme pair was investigated on several geometric arrangements of DNA-origami scaffolds. We found that arrangement of enzymes on a planar scaffold primarily gains efficiency from induced enzyme colocalization with moderate enhancement due to the scaffold acting as a diffusive barrier. However, confinement of the enzyme system within a nanotube scaffold greatly enhances substrate transfer, up to ten fold relative to colocalization, and up to 150 fold relative to a disorganized solution of the enzymes and substrate over the same time period. Our results have implications for the efficient engineering of synthetic enzyme cascades.

The short-range intermolecular interactions that discriminate the selectivity of a substrate to a host molecule and dictate the configurational transition to the bound state was studied using the Hopping Minima software. Normal mode calculations were used to

construct coordinated natural motions of a host-guest complex. These motions were then utilized to connect conformational minima and combined to form non-covalent binding pathways. Our results demonstrate that conformational transitions can be modeled and extended to construct coordinated final binding events, allowing for the study of systems with binding kinetics that are traditionally impractical to sample.

Finally, a bound-state affinity study was performed for determining receptor subtype specificity. Specifically, a natural product derivative proteasome inhibitor was simulated with molecular dynamics, and molecular recognition was assessed through structural stability and energetic affinity with each receptor subtype. The determined energetic and structural preference suggests desirable activity as a human proteasome inhibitor.

7.2 Future Work

The computational study of substrate-enzyme association has been limited by the high computational cost of simulating molecular systems over large time scales relative to the scope of traditional dynamics techniques. Approaching such systems typically requires compromise over the level of structural and theoretical detail included in the computational model. With the increasing availability of high powered computer clusters and parallel development tools, I plan to improve the scalability of the GeomBD package (Chapter 2 and 3) and include additional theoretical and structural detail for the assessment of substrate association probability calculations. Modeling tools will be implemented to increase electrostatic detail of macromolecules with highly abstracted coarse grain structures. Multi-scale structures will also be introduced to allow for increased level of structural details in important regions of macromolecules, such as

catalytic active sites. In addition to the assessment of diffusional association, the package will also be expanded to allow for determining the probability of deep pocket binding after superficial association with the binding site.

Utilizing the aforementioned augmentations to the GeomBD package, the simulations outlined in Chapter 3 will be revisited.

7.2.1 Multi-scale Structure

Bridging multiple levels of detail, I plan to implement a tool for the GeomBD package to integrate multiple modeling scales of macromolecular structures into a single multi-scale structural model. For example, a 1-bead-per-residue coarse grain protein model could be combined with an all-atom protein structure. In this way, areas of the protein important for association, such as the binding site, could be modeled in full atomic detail while the rest of the protein is modeled more coarsely. Support for well-established coarse graining schemes, such as MARTINI, will be included in the future for inclusion in multi-scale structural models.[1,2]

As well, a multi-scale electrostatic model will be implemented for coarse grain models using the RESPAC coarse grain partial charge method.[3] Highly abstracted coarse grain models of macromolecules will gain detailed electrostatic interactions with the substrate, more accurately modeling their interactions.

7.2.2 Deep Pocket Binding

The current implementation of GeomBD is intended for modeling the diffusional encounter of a substrate with a target binding site. Due to the coarse grain nature of the structural models in use, only superficial association (collision with binding site residues)

is considered. In many cases, diffusional association is only the first step to the binding of a substrate to an active site, and a substrate is required to diffuse deeper into the binding pocket and rearrange its conformation to fully bind. For example, acetylcholine esterase has a catalytic active site contained at the bottom of a deep gorge-like pocket.[4] Superficial association is followed by progression of the substrate through the pocket to the catalytic site. Therefore, we propose an extension to the GeomBD package for calculating the probability of a full binding event after superficial association with a macromolecule. Using the results of diffusional association simulations, a more detailed structural model of the binding pocket will be used to start new simulations. Thousands of replicate rigid body simulations will be run simultaneously from each superficial association starting point to determine the probability of substrate progression deeper into the binding pocket and, finally, to the bound state. It is possible that in some cases multiple diffusional barriers exist within deep binding pockets. In these cases multiple sequential simulation sets will be performed to assess the distinct probabilities of passing diffusional barriers. Both the forward and backward barrier transition probabilities can be calculated to provide insight into kinetic differences between association and dissociation rates. Several sets of simulations will be performed on the well-studied acetylcholinesterase enzyme for development of this technique. Simulations will be performed over 100 microseconds, as the turnover time of acetylcholinesterase at high substrate concentration is approximately 63 microseconds.[5] A coarse grain model with multi-bead representations for protein residues, such as MARTINI, will be used in an attempt to more fully capture the specific substrate-enzyme interactions.

Several coarse grain Brownian dynamics techniques will be explored. Because flexibility in both the ligand and macromolecule structure is often necessary for final bound state conformations to be achieved, we will investigate two flexible structure modeling approaches. We will explore a harmonic network approach to introducing a flexible structure to ligand and protein side chains. This would retain the relative position of each particle within a molecular entity relative to its starting reference conformational state with a network of harmonic bonds between each pair of particles.

7.3 References

(1) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(2) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(3) Terakawa, T.; Takada, S. RESPAC: Method to Determine Partial Charges in Coarse-Grained Protein Model and Its Application to DNA-Binding Proteins. *J. Chem. Theory …* **2014**.

(4) Sussman, J.; Harel, M.; Frolow, F.; Oefner, C.; Goldman, A.; Toker, L.; Silman, I. Atomic Structure of Acetylcholinesterase from Torpedo Californica: A Prototypic Acetylcholine-Binding Protein. *Science (80-.).* **1991**, *253*, 872–879.

(5) Nolte, H.-J.; Rosenberry, T. L.; Neumann, E. Effective Charge on Acetylcholinesterase Active Sites Determined from the Ionic Strength Dependence of Association Rate Constants with Cationic Ligands. *Biochemistry* **1980**, *19*, 3705–3711.