UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational Studies on Biomolecular Diffusion and Electrostatics

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

in

Chemistry

by

Nuo Wang

Committee in charge:

Professor James A. McCammon, Chair
Professor Olga K. Dudko
Professor Michael Galperin
Professor Elizabeth A. Komives
Professor John H. Weare

2015

The dissertation of Nuo Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2015

DEDICATION

For my parents, who taught me what to pursue in life.

For Chris, who always makes me laugh.

EPIGRAPH

I admire the elegance of your method of calculation. It must be nice to ride through these fields upon the horse of true mathematics, while the likes of us have to make our way laboriously on foot.

– Albert Einstein (in correspondence with Tullio Levi-Civita)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my dad, who encouraged me to study science. He always believed in science, the good things that it can bring to humanity and that it is one of the most meaningful and valuable subjects to be studied. I would like to thank my mom for being a cheery and strong person who only brings wit and happiness to other people's lives. I've never known anyone else as optimistic as her in my life.

I would like to thank Chris for being patient with me and supportive to me in the past 4 years. He is an expert at making me laugh. He is also someone that I can talk to. He helped me stay strong and move forward. I would like to thank Chris (again), his parents and our friends at UCSD for teaching me a significant amount about life and people, family and friendship.

I would like to thank Sijia Dong for being who she is, always being there for me when I have something to share and discuss. She is my comrade.

I would like to thank Andy for being the most understanding and generous adviser that I've known. Andy gave me a lot of freedom to explore the world of science, which was a huge luxury during my graduate study.

I would like to thank Paul, Patrick, Olivia, Gary, Changsun, Pete, Steffen, Yinglong, Alisha, Chris, Clarisse, Joe, Levi, Mehrnoosh, Leo, Vince, Sara, Yi, Juanma, Patricia and Ryosuke in the McCammon lab for providing a collaborative and fun working environment and having enlightening scientific discussions with me. I would like to thank Patti, Robert, Brian and the rest of the McCammon lab for all of the supports that you've given me.

I would like to thank Professor Bo Li and his student Shenggao for their great help in our collaborations. I would like to thank Professors Katja Lindenberg and Michael Galperin for their valuable discussions and supports in the first year of my graduate study.

Last, I would also like to thank my doctoral committee, Professors Andy McCam-

mon, Olga Dudko, Michael Galperin, Elizabeth Komives and John Weare for their help along the way to the successful completion of my Ph.D. degree.

Chapter 2 is a minimally modified reprint of the material as it appears in Nuo Wang, Gary A. Huber, and J. Andrew McCammon, "Assessing the Two-Body Diffusion Tensor Calculated by the Bead Models," Journal of Chemical Physics, 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is a minimally modified reprint of the material as it appears in Nuo Wang, Shenggao Zhou, Peter M. Kekenes-Huskey, Bo Li, and J. Andrew McCammon, "Poisson-Boltzmann versus Size-Modified Poisson-Boltzmann Electrostatics Applied to Lipid Bilayers," Journal of Physical Chemistry B, 2014. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is a minimally modified reprint of the material as it appears in Nuo Wang, and J. Andrew McCammon, "Substrate Channeling Between the Human Dihydrofolate Reductase and Thymidylate Synthase," Protein Science, 2015. The dissertation author was the primary investigator and author of this paper.

VITA

2010	B. S. in Biochemistry, minor in Computer Science, University of Hong Kong, Hong Kong, China
2010-2011	Graduate Teaching Assistant, University of California, San Diego, La Jolla, U.S.A.
2012	M. S. in Chemistry, University of California, San Diego, La Jolla, U.S.A.
2015	Ph. D. in Chemistry, University of California, San Diego, La Jolla, U.S.A. (With a specialization in Theoretical and Computational Chemistry)

PUBLICATIONS

N. Wang, G. A. Huber and J. A. McCammon. Assessing the Two-Body Diffusion Tensor Calculated by the Bead Models. J. Chem. Phys., 138(20):204117, 2013.

N. Wang, S. Zhou, P. M. Kekenes-Huskey, B. Li and J. A. McCammon. Poisson-Boltzmann versus Size-Modified Poisson-Boltzmann Electrostatics Applied to Lipid Bilayers. J. Phys. Chem. B, 118(51):14827-14832, 2014.

N. Wang and J. A. McCammon. Substrate Channeling Between the Human Dihydrofolate Reductase and Thymidylate Synthase. Protein Science, 2015

ABSTRACT OF THE DISSERTATION

Computational Studies on Biomolecular Diffusion and Electrostatics

by

Nuo Wang

Doctor of Philosophy in Chemistry

University of California, San Diego, 2015

Professor James A. McCammon, Chair

As human understandings of physics, chemistry and biology converge and the development of computers proceeds, computational chemistry or computational biophysics has become a substantial field of research. It serves to explore the fundamentals of life and also has extended applications in the field of medicine. Among the many aspects of computational chemistry, this Ph. D. work focuses on the numerical methods for studying diffusion and electrostatics of biomolecules at the nanoscale. Diffusion and electrostatics are two independent subjects in terms of their physics, but closely related in applications. In living cells, the mechanism of diffusion powers a ligand to move towards its binding target. And electrostatic forces between the ligand and the

target or the ligand and the environment guide the direction of the diffusion, the correct binding orientation and, together with other molecular forces, ensure the stability of the bound complex. More abstractly, diffusion describes the stochastic manner biomolecules move on their energy landscape and electrostatic forces are a major contributor to the shape of the energy landscape. This Ph. D. work aims to acquire a good understanding of both biomolecular diffusion and electrostatics and how the two are used together in numerical calculations. Three projects are presented. The first project is a proof of concept of the bead-model approach to calculate the diffusion tensor. The second project is the benchmark for a new electrostatics method, the size-modified Poisson-Boltzmann equation. The third project is an application that combines diffusion and electrostatics to calculate the substrate channeling efficiency between the human thymidylate synthase and dihydrofolate reductase.

Chapter 1

A Brief Overview of the Computational Methods of Biomolecular Diffusion and Electrostatics

This chapter provides the background information on the field of study of this thesis and some of the key theoretical foundations for Chapters 2, 3 and 4. Specifically, Section 1.2.1 introduces the diffusion tensor, which is the focus of study in Chapter 2. Section 1.2.2 illustrates the Langevin equation, which is used for the diffusion simulations in Chapter 4. Sections 1.3.1 and 1.3.2 explain the Poisson-Boltzmann equation, which is the basis for the development of its improvement, the size-modified Poisson-Boltzmann equation, in Chapter 3.

1.1 Background

1.1.1 The Importance of the Diffusion Process and Electrostatic Interactions in Living Organisms

In living organisms, biomolecules are being constantly transported within a cell and across cells. This flux of biomolecules is a fundamental process of life.[1] There is the energy-consuming active transport by the membrane transporters[2] or cellular vesicles;[3] there is also the passive transport by diffusion down the concentration gradient of biomolecules that does not require energy input but instead increases the entropy of the system. The journey of a biomolecule from its location of creation to its location of function is usually a combination of active transport and passive diffusion. For example, once a protein enters its destination organelle through active transport, it still needs to encounter its reaction partner in the organelle by the mechanism of diffusion. Diffusion is a ubiquitous and crucial part of the cellular and tissue-wide biomolecule trafficking that needs to be decoded for a better understanding of the machineries in biology Biomolecular diffusion is driven by the concentration gradient but also heavily influenced by external forces, especially the long-range electrostatic forces. The electrostatic influence on biomolecular diffusion is both at the mesoscopic[4,5] and the microscopic scales.[6,7] There are cases where electrostatic forces help to guide the diffusional encounter and the interaction between a ligand and its binding target, and effectively increasing the reaction rate between the two.[8] But more generally, the effect that electrostatic forces have on the diffusion pattern of biomolecules is more elusive and involves a system of proteins, it cannot be summarized by the increase or decrease of a single reaction rate.[4,5] Setting aside diffusion, electrostatics itself is also a major energy component of biomolecular systems and contributes to the shape of the energy landscape of the system. The energetics of a biomolecule highly determines its structure and function. Electrostatics can guide

protein folding,[9] it can also play a key role in enzymatic reactions.[10] Being able to calculate and understand the electrostatic features of a biomolecular system is thus vital to the unveiling of the mysteries in biology.

1.1.2 The Value of Computational Chemistry

As much as technologies have been invented and improved at an unprecedented rate in the past century, the number of experiments that humans can conduct to observe the phenomena in nature, such as those of biological systems, cannot encompass everything that humans would like to understand. On the other hand, physicists and chemists have developed a set of self-consistent theories from the quantum level to the classical level that covers almost all of the fundamental mechanisms of nature. Ideally, the complexities of nature, such as life, that are built upon these fundamental mechanisms can be theorized or modeled by building upon the fundamental physical laws, such as Newton's law. However, in reality, when theories are applied to comprehend a biological system that is reasonably-sized and at a relevant time-scale, more often than not, no analytical solution can be obtained from the theories, the only way is to solve the problem numerically by modern computers. This line of works is called computational chemistry or computational biophysics. The foundation of computational chemistry is rooted in the discovery of quantum mechanics in the early twentieth century.[11] It became an accessible scientific method to the general researchers since the invention of the first microprocessor in 1971.[12] Computational chemistry or biophysics can be applied at various scales, from climate change[13] to the function of a single protein.[14] Computational chemistry is used to provide insights on the puzzle pieces that cannot be collected by experiments and guide the direction of further experiments, which in return confirm the computational results. This computation-experiment cycle is an especially promising workflow in the computer-aided drug design field and helped in the successful development of a number of drug

molecules.[15,16] In this thesis, computational chemistry is used to study the currently experimentally inaccessible details of biomolecular diffusion and electrostatics.

1.2 Biomolecular Diffusion

One of the most common macroscopic examples of diffusion is a drop of dye spreading along its concentration gradient in a beaker of water. The microscopic explanation of this lies in the existence of atoms and their thermal motions. This example only involves diffusion in the physical space, while diffusion in general can happen in any degree of freedom, i.e. position and momentum, or collective degree of freedom of a system. All of the works in this thesis are in the classical regime. In classical mechanics, everything is deterministic, however the theory of diffusion models a system using probability. The non-deterministic behavior of diffusion comes from the neglect of a large number of degrees of freedom. This neglect is due to the facts that first, it is too expensive and tedious to keep track of all of the degrees of freedom of a realistically-sized physical system; second, typically, only a fraction of the degrees of freedom is of interest. In the case of biomolecular systems, the degrees of freedom that are usually neglected are those of the solvent in a solution of biomolecules of interest. The probabilistic perturbations from the neglected degrees of freedom require the degrees of freedom of interest to be modeled as stochastic variables that take defined ranges of values and have defined probability distributions over their ranges. An ensemble of the realizations of such stochastic variables over time is called a stochastic process. The equilibrium state properties of a system with incomplete information are calculated as the ensemble averages of all of the possible configurations of the stochastic variables weighted by the corresponding probability of occurrence. The dynamics of the system is calculated by the deterministic energy function of the biomolecule together with the stochastic energy

input from the solvent molecules. There are multiple mathematical models to implement the stochasticity of a system. One that is relevant to a lot of chemistry and physics problems is called the Markov process. The stochastic process of stochastic variable x is a Markov process if at any time t_n, the probability of x to take value x_n is uniquely determined by the immediate previous state of the system x_{n-1} at t_{n-1} and is not affected by the state of the system at any time prior to t_{n-1}. In other words, a Markov process has no memory. Brownian dynamics or Brownian motion that is used in this thesis is a subcategory of Markov process. The discovery of Brownian motion is credited to botanist Robert Brown who observed the random motions of pollen in water in 1827. Between 1905 and 1906, Albert Einstein and Marian Smoluchowski made theoretical explanations of this phenomenon based on the existence of atoms and molecules. Brownian motion is typically expressed in two ways: one is the Langevin equation, which is a stochastic equation of motion of the stochastic variables; the other is the Fokker-Planck equation, which is a deterministic equation for the probability density of the stochastic variables. The two equations are equivalent to each other. Here the Fokker-Planck equation is derived first and then its relationship to the Langevin equation is shown.

1.2.1 The Fokker-Planck Equation

The derivation in this section follows two books, one by N. G. van Kampen,[17] one by W. J. Coffey et al.[18] All of the stochastic variables are in general multi-dimensional vectors, but here they are presented as one-dimensional variables for simplicity. Mathematically, a Markov process is defined as:

$$P_{1|n-1}(x_n,t_n|x_1,t_1;\ldots;x_{n-1},t_{n-1}) = P_{1|1}(x_n,t_n|x_{n-1},t_{n-1}) \tag{1.1}$$

where $P_{a|b}$ is the conditional probability for a stochastic variable to take certain values at a different times given their values fixed at b other times. From the definition of a Markov process, the joint probability P_3 for stochastic variable x to take values x_1, x_2, x_3 at times $t_1 < t_2 < t_3$ is:

$$P_3(x_1,t_1;x_2,t_2;x_3,t_3) = P_1(x_1,t_1)P_{1|1}(x_2,t_2|x_1,t_1)P_{1|1}(x_3,t_3|x_2,t_2) \tag{1.2}$$

Integrating the above equation over x_2 and dividing $P_1(x_1,t_1)$ on both sides gives the **Chapman-Kolmogorov equation**:

$$P_{1|1}(x_3,t_3|x_1,t_1) = \int_{-\infty}^{\infty} P_{1|1}(x_3,t_3|x_2,t_2)P_{1|1}(x_2,t_2|x_1,t_1)dx_2 \tag{1.3}$$

The Chapman-Kolmogorov equation is the starting point of the derivation of the Fokker-Planck equation.

Now consider an arbitrary function $f(x)$ of the stochastic variable x. At $x = \pm\infty$, $f(x)$ equals 0 and all of its derivatives exist. Write the following expression:

$$\begin{aligned}&\int_{-\infty}^{\infty} f(x_3)\frac{\partial P_{1|1}(x_3,t_3|x_1,t_1)}{\partial t_3}dx_3 = \\ &\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{\infty} f(x_3)\left(P_{1|1}(x_3,t_3+\Delta t|x_1,t_1) - P_{1|1}(x_3,t_3|x_1,t_1)\right)dx_3\end{aligned} \tag{1.4}$$

The naming of the variables in Equation (1.4) is arbitrary but times $t_1 < t_3$. Using the Chapman-Kolmogorov equation on $P_{1|1}(x_3,t_3+\Delta t|x_1,t_1)$, the right hand side of Equation (1.4) becomes:

$$\begin{aligned}&\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_3)P_{1|1}(x_3,t_3+\Delta t|x_2,t_3)P_{1|1}(x_2,t_3|x_1,t_1)dx_2dx_3 \\ -&\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{\infty} f(x_3)P_{1|1}(x_3,t_3|x_1,t_1)dx_3\end{aligned} \tag{1.5}$$

Expand $f(x_3)$ in a Taylor series around $x_3 = x_2$ and keep only the first three terms (this will be justified later):

$$f(x_3) \approx f(x_2) + (x_3 - x_2)f'(x_2) + \frac{(x_3 - x_2)^2}{2!}f''(x_2) \tag{1.6}$$

Plugging Equations (1.5) and (1.6) into Equation (1.4) and substituting x_3 as x_2 for the terms where they are not integrated together:

$$\begin{aligned}
&\int_{-\infty}^{\infty} f(x_2)\frac{\partial P_{1|1}(x_2,t_3|x_1,t_1)}{\partial t_3}dx_2 \\
&= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(f(x_2) + (x_3 - x_2)f'(x_2) + \frac{(x_3 - x_2)^2}{2!}f''(x_2)\right)\right. \\
&\left. P_{1|1}(x_3,t_3+\Delta t|x_2,t_3)P_{1|1}(x_2,t_3|x_1,t_1)dx_2dx_3\right] \\
&- \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{\infty} f(x_2)P_{1|1}(x_2,t_3|x_1,t_1)dx_2 \qquad (1.7) \\
&= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left((x_3 - x_2)f'(x_2) + s\frac{(x_3 - x_2)^2}{2!}f''(x_2)\right)\right. \\
&\left. P_{1|1}(x_3,t_3+\Delta t|x_2,t_3)P_{1|1}(x_2,t_3|x_1,t_1)dx_2dx_3\right] + \\
&\lim_{\Delta t\to 0}\frac{1}{\Delta t}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_2)P_{1|1}(x_3,t_3+\Delta t|x_2,t_3)P_{1|1}(x_2,t_3|x_1,t_1)dx_2dx_3\right. \\
&\left. - \int_{-\infty}^{\infty} f(x_2)P_{1|1}(x_2,t_3|x_1,t_1)dx_2\right] \qquad (1.8)
\end{aligned}$$

The last $\lim_{\Delta t\to 0}\frac{1}{\Delta t}[\cdots]$ term in Equation (1.8) becomes zero because $\int_{-\infty}^{\infty} P_{1|1}(x_3,t_3 + \Delta t|x_2,t_3)dx_3 = 1$. With this, exchange the order of the integral over x_2 and the limit, and define:

$$D_n(x_2,t_3) = \lim_{\Delta t\to 0}\frac{\int_{-\infty}^{\infty}(x_3 - x_2)^n P_{1|1}(x_3,t_3+\Delta t|x_2,t_3)dx_3}{n!\Delta t} \tag{1.9}$$

Equation (1.8) becomes:

$$\int_{-\infty}^{\infty} f(x_2)\frac{\partial P_{1|1}(x_2,t_3|x_1,t_1)}{\partial t_3}dx_2$$

$$= \int_{-\infty}^{\infty} \Big(D_1(x_2,t_3) f'(x_2) + D_2(x_2,t_3) f''(x_2) \Big) P_{1|1}(x_2,t_3|x_1,t_1) dx_2 \tag{1.10}$$

Integrating by parts for the terms containing $f'(x_2)$ and $f''(x_2)$ in Equation (1.10), Equation (1.10) becomes:

$$\begin{aligned} &\int_{-\infty}^{\infty} f(x_2) \frac{\partial P_{1|1}(x_2,t_3|x_1,t_1)}{\partial t_3} dx_2 \\ &= \int_{-\infty}^{\infty} f(x_2) \Bigg[-\frac{\partial}{\partial x_2} \Big(D_1(x_2,t_3) P_{1|1}(x_2,t_3|x_1,t_1) \Big) \\ &\quad + \frac{\partial^2}{\partial x_2^2} \Big(D_2(x_2,t_3) P_{1|1}(x_2,t_3|x_1,t_1) \Big) \Bigg] dx_2 \end{aligned} \tag{1.11}$$

For Equation (1.11) to fulfill any arbitrary function $f(x)$, the following condition has to be met:

$$\begin{aligned} &\frac{\partial P_{1|1}(x_2,t_3|x_1,t_1)}{\partial t_3} = \\ &-\frac{\partial}{\partial x_2} \Big(D_1(x_2,t_3) P_{1|1}(x_2,t_3|x_1,t_1) \Big) + \frac{\partial^2}{\partial x_2^2} \Big(D_2(x_2,t_3) P_{1|1}(x_2,t_3|x_1,t_1) \Big) \end{aligned} \tag{1.12}$$

Simplifying the variable naming in Equation (1.12), a clean form of the **Fokker-Planck equation** is obtained:

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x} \Big(D_1(x,t) P(x,t) \Big) + \frac{\partial^2}{\partial x^2} \Big(D_2(x,t) P(x,t) \Big) \tag{1.13}$$

In Equation (1.6), the terms after the third one are truncated. This is because of the approximation that variable x is perturbed by white noise. Referring to Equation (1.9), when $n = 1$, D_1 has the unit of velocity, it is physically the drift velocity of the stochastic variables under an external potential field. If there is no external field, $D_1 = 0$, which is the mean of the white noise. When $n = 2$, D_2 is the variance of the white noise. With a

change of symbols:

$$D_2(x,t) = \lim_{\Delta t \to 0} \frac{\int_{-\infty}^{\infty} (x-y)^2 P_{1|1}(x,t+\Delta t|y,t)dy}{2!\Delta t}$$
$$= \lim_{\Delta t \to 0} \frac{2D\Delta t}{2!\Delta t} = D \tag{1.14}$$

where D is the **diffusion tensor**. It is usually approximated by a constant, the **diffusion constant**, for an isotropic solution with spherical solutes that do not interact with each other. However, it can be more accurately represented by an n by n square matrix, where n is the number of degrees of freedom of interest of the system. Chapter 2 contains a detailed description of one of the ways to realize the diffusion tensor. The higher moments ($n > 2$) of the white noise are zero.

When it comes to applying the Fokker-Planck equation to study the diffusion of macro-biomolecules, such as proteins and nucleic acids, a specialized version of the equation, the **Smoluchowski equation**, is usually used. The Smoluchowski equation can be expressed in the same abstract form as Equation (1.13), but it differs from the Fokker-Planck equation in two key aspects:

1. The Smoluchowski equation only models the Brownian motion at the time scale much longer than the velocity relaxation time, where the Brownian particles have no acceleration, i.e. the **over-damped** regime. This is true for macro-biomolecules, which are significantly bigger than the solvent water molecules that provide the random perturbations.

2. Generalized variable x in Equation (1.13) represents a vector of both position and momentum; But because that the velocity of the Brownian particles is effectively constant under the Smoluchowski time-scale, the Smoluchowski equation is an equation of the particle positions only.

As in most cases, when the position coordinates in the Smoluchowski equation

are the x, y, z coordinates of the Brownian particles in the physical space, the probability density in the Smoluchowski equation is naturally the concentration of the particles. Solving the Smoluchowski equation with the appropriate initial and boundary conditions grants the time-dependent concentration profile of the Brownian particles of interest. A good example of the application of the Smoluchowski equation is its use in simulating the calcium sparks in cardiomyocytes.[19] In this paper, the solution of the Smoluchowski equation is able to provide direct information on the duration and the magnitude of the sparks, as well as the localization and density of Ca^{2+} near the Ca^{2+} handling proteins.

Finite element method is usually used to solve the Smoluchowski equation. An example is program SMOL[20] that utilizes the Finite Element Tool Kit (FETK).[21] A newer Smoluchowski solver Smolfin[22] instead makes use of the finite element libraries from the FEniCS project.[23]

Although the works done in this thesis do not involve solving the Fokker-Planck or the Smoluchowski equation, Chapter 2 discusses the bead model approach and its validity to calculate realistic multi-dimensional diffusion tensors that can be used as input in solving the Fokker-Planck or Smoluchowski equation. Specifically, the bead model calculated diffusion tensor can be used to substitute the D_2 term in Equation (1.13).[24]

1.2.2 The Langevin Equation

The Langevin equation is similar to the Newton's second law of motion, $F = ma$, but with the addition of the random force. Here the **Langevin equation** will be directly given:

$$m\frac{d^2x}{dt^2} = -\gamma\frac{dx}{dt} - \frac{dU(x)}{dx} + \eta \tag{1.15}$$

x and m are the position and mass of the diffusing particle. $\gamma = 6\pi\lambda r$ is the friction coefficient of the solvent, where λ is viscosity, r is the radius of the diffusing particle. γ

is related to the diffusion constant D by the Stokes-Einstein relation $D = k_B T/\gamma$. $U(x)$ is the external potential field on the system, it causes the "drifting" of the diffusing particles. η is the stochastic force that obeys Gaussian distribution, i.e. white noise, it has the following properties:

$$\langle \eta(t) \rangle = 0$$
$$\langle \eta(t_1)\eta(t_2) \rangle = 2k_B T\gamma\delta(t_2 - t_1)$$

These properties correspond to that of D_1 and D_2 when $n = 1$ and $n = 2$ in Equation (1.9).

Here, a brief proof of the equivalency of the Fokker-Planck equation and the Langevin equation is shown. The derivation in this section follows Zwanzig's Nonequilibrium Statistical Mechanics[25] and the lecture notes by Lennart Sjögren at the University of Gothenburg.[26]

Writing down a general Langevin equation of variable a:

$$\frac{da}{dt} = \nu(a) + F(t) \tag{1.16}$$

where $\nu(a)$ is a given function of a. $F(t)$ is Gaussian noise with zero mean and variance $\langle F(t_1)F(t_2) \rangle = 2B\delta(t_2 - t_1)$, B is a constant. $f(a,t)$, the density of matter at coordinate a, is a conserved function of a, it obeys the continuity equation:

$$\nabla \cdot J(a,t) + \frac{\partial f(a,t)}{\partial t} = 0 \tag{1.17}$$

$J(a,t) = f(a,t)(\partial a/\partial t)$ is the flux of the density of matter. Plugging Equation (1.16)

into Equation (1.17):

$$\frac{\partial f(a,t)}{\partial t} = -\frac{\partial}{\partial a}\left(\nu(a)f(a,t) + F(t)f(a,t)\right) \tag{1.18}$$

Define $L_0 = \frac{\partial}{\partial a}\nu(a)$, $L_1 = \frac{\partial}{\partial a}F(t)$, Equation (1.18) becomes:

$$\frac{\partial f(a,t)}{\partial t} = (-L_0 - L_1)f(a,t) \tag{1.19}$$

Represent f by $f(a,t) = e^{-L_0 t}\sigma(a,t)$. Plugging it into Equation (1.19):

$$\frac{\partial \sigma(a,t)}{\partial t} = -e^{L_0 t}L_1(t)e^{-L_0 t}\sigma(a,t) = -V(t)\sigma(a,t) \tag{1.20}$$

where $V(t) = e^{L_0 t}L_1(t)e^{-L_0 t}$. The solution of Equation (1.20) is:

$$\sigma(a,t) = e^{-\int_0^t V(t')dt'}\sigma(a,0) = e^{iX}\sigma(a,0) \tag{1.21}$$

where $X(t) = i\int_0^t V(t')dt'$. Taking the average of $\sigma(a,t)$ over Gaussian noise $F(t)$:

$$\langle\sigma(a,t)\rangle_{F(t)} = \left\langle e^{iX}\right\rangle_{F(t)}\sigma(a,0) \tag{1.22}$$

$\langle\sigma(a,t)\rangle_{F(t)}$ takes the form of the characteristic function of the variable $X = X(V(F(t)))$, which also behaves as Gaussian noise. The mean of X is 0, the variance of X is:

$$\begin{aligned}
\left\langle X(t)^2\right\rangle_{F(t)} &= \int_0^t\int_0^t \langle V(t_1)V(t_2)\rangle_{F(t)}\,dt_1dt_2 \qquad (1.23)\\
&= \int_0^t\int_0^t \left\langle e^{L_0t_1}\frac{\partial}{\partial a}F(t_1)e^{-L_0t_1}e^{L_0t_2}\frac{\partial}{\partial a}F(t_2)e^{-L_0t_2}\right\rangle_{F(t)} dt_1dt_2\\
&= \int_0^t\int_0^t e^{L_0t_1}\frac{\partial}{\partial a}e^{-L_0t_1}e^{L_0t_2}\frac{\partial}{\partial a}e^{-L_0t_2}\,2B\delta(t_1-t_2)\,dt_1dt_2\\
&= 2B\int_0^t e^{L_0t_1}\frac{\partial^2}{\partial a^2}e^{-L_0t_1}dt_1 \qquad (1.24)
\end{aligned}$$

For Gaussian variable $X(t)$ with 0 mean, $\left\langle e^{iX} \right\rangle = e^{-\left\langle X(t)^2 \right\rangle/2}$, so:

$$\langle \sigma(a,t) \rangle_{F(t)} = e^{-\left\langle X(t)^2 \right\rangle/2} \, \sigma(0,t) \tag{1.25}$$

$$\frac{\partial \langle \sigma(a,t) \rangle_{F(t)}}{\partial t} = B e^{L_0 t} \frac{\partial^2}{\partial a^2} e^{-L_0 t} \langle \sigma(a,t) \rangle_{F(t)} \tag{1.26}$$

Plugging $\sigma(a,t) = e^{L_0 t} f(a,t)$ into Equation (1.26), rewriting $\langle f(a,t) \rangle_{F(t)}$ as probability density $P(a,t)$ and using the full form of L_0, a Fokker-Planck-like equation is obtained :

$$\frac{\partial}{\partial t} P(a,t) = -\frac{\partial}{\partial a} \left(\nu(a) P(a,t) \right) + B \frac{\partial^2}{\partial a^2} P(a,t) \tag{1.27}$$

In the most general case, $a = [x_1, x_2, ...x_n, p_1, p_2, ..., p_n]$ is a vector of all of the degrees of freedom of the system, x is position and p is momentum, Equation (1.27) is the Fokker-Planck equation. However, when a is the particle position only, comparing Equation (1.16) to Equation (1.15), it can be seen that the particle acceleration is absent. In this case, Equation (1.16) is the **over-damped Langevin equation** and Equation (1.27) is the Smoluchowski equation with external potential field $\nu(a)$.

As is mentioned in the previous section, the Smoluchowski equation, instead of the more general Fokker-Planck equation, is usually used to study the diffusion of macro-biomolecules. For the same reason, the over-damped Langevin equation is usually used for macro-biomolecules instead of the Langevin equation.

Although both of the over-damped Langevin equation and the Smoluchowski equation can be used to calculate reaction rates and to observe the localization of the diffusing particles, the numerical solutions of the two equations that are used to generate these quantities are different. The Smoluchowski equation offers the concentration map, $P(x)$, of the diffusing particle over the entire system while the over-damped Langevin equation, as an equation of motion, generates individual trajectories, $x(t)$, of the diffusing

particle. The Smoluchowski solution gives the ensemble information of the system at a lower resolution while the over-damped Langevin equation provides the instances of the ensemble at a higher resolution. The lower resolution of the Smoluchowski approach lies in the facts that first, the system needs to be discretized for its numerical solution and second, the diffusing particles are modeled as concentrations at each point in space and the information of their shapes cannot be included. Neither of these two approximations is necessary in the Langevin approach. If one simulates enough numbers of trajectories by the over-damped Langevin equation, they should obtain the Smoluchowski solution by averaging over all of the trajectories.[27] Which equation to use is usually a choice between computational efficiency and accuracy. Note that, from the numerical calculation and programming point of view, as soon as more than one diffusing species and time-dependent boundary conditions, e.g. opening or closing of a channel, are involved, the over-damped Langevin approach becomes significantly easier to apply.

A number of research papers used the over-damped Langevin equation to calculate the reaction rates between biomolecules.[6,7,28] People also used the equation to observe other interesting biological phenomena, such as biomolecular crowding effects.[29,30] A few known Langevin dynamics or Brownian dynamics simulation software are UHBD,[31] SDA,[32] BrownDye,[33] BD_BOX.[34] Some molecular dynamics software, for example GROMACS,[35] can also be used to simulate Langevin dynamics.

The Chapter 4 in this thesis uses software BrownDye to simulate thousands of stochastic trajectories by the over-damped Langevin equation to calculate the reaction rate of a small ligand that is being released as the product from one enzyme active site and reacts as the substrate with another enzyme active site. The "substrate channeling efficiency" is then calculated, i.e. the number of substrates that reacted with the active site over the total number of substrates produced (that includes the reacted and the escaped substrates). Substrate channeling is an important catalytic strategy of the multi-enzyme

complexes in vivo, and numerical diffusion simulation is a powerful tool to tackle such problems.[36]

1.3 Biomolecular Electrostatics

Biomolecules always carry partial charges and the electrostatic interactions between these partial charges contribute a significant amount of energy, the electrostatic energy, to the total energy of biomolecular systems. Usually, the charges on the atoms in the biomolecules are represented by point charges centered at their atomic center. When it comes to representing the charges on the water and the ions in the solvent, usually two approaches are taken, the **explicit solvent model** and the **implicit solvent model**.[37]

In the explicit solvent model, the charges on the water and the ions are represented exactly as that of the biomolecule. However, in the implicit solvent model, the net charge on the water molecules is taken as zero and the charges on ions are represented by a continuous charge density. The electrostatic screening of the polar water molecules is modeled by a high dielectric constant, usually taken as 78 or 80, versus the low dielectric constant in the biomolecule, usually taken as 1 or 2. The implicit solvent model is invented because first, again, it is not always necessary to keep track of the full details of the solvent degrees of freedom. And second, in biomolecular simulations, a large number of solvent molecules needs to be added to reasonably represent a solvated biomolecule and computing the dynamics or energetics of the solvent's degrees of freedom takes a significant amount of computing power. It is more economical if these computations can be saved at the cost of a small decrease in accuracy. The implicit solvent model is frequently used in the diffusion simulations to study biological problems at long time-scale or large length-scale.[38]

To solve for the electrostatic properties of the biomolecular systems in the implicit

solvent model, the Poisson's equation and its approximations are used. The popular approximations are the Poisson-Boltzmann equation[39] and the generalized Born model,[40] which is a further approximation of the Poisson-Boltzmann equation. In this thesis, the Poisson-Boltzmann equation is exclusively used and it will be explained in more details in the following text.

1.3.1 The Poisson-Boltzmann Equation

The Poisson-Boltzmann equation was invented independently by Louis George Gouy and David Leonard Chapman in 1910 and 1913. An informal derivation of the equation is given below starting from the Poisson's equation.

Poisson's equation is the fundamental equation for calculating the electrostatic potential of a system with a given charge distribution ρ. The differential form of Gauss's law is $\nabla \cdot (\varepsilon E) = \rho$ and the electric field $E = -\nabla\phi$, where ε is the dielectric constant and ϕ is the electrostatic potential. Combining the two equations results in the **Poisson's equation** for electrostatics:

$$\nabla \cdot \varepsilon \nabla \phi = -\rho \tag{1.28}$$

In the implicit solvent model, for biomolecules in a box of pure water, the Poisson's equation applies. In this case, $\rho = \rho(x)$ is the partial charge distribution on the biomolecule, while the partial charges on the atoms in water are considered as a whole to be zero and are not represented in $\rho(x)$. Dielectric constant ε takes on different values depending on the location:

$$\varepsilon(x) = \begin{cases} 2 & \text{x} \in \text{biomolecule} \\ 78 & \text{x} \in \text{solvent} \end{cases} \tag{1.29}$$

Equation (1.29) describes a sudden jump in ε at the solute-solvent boundary. Alternatively

a smooth interpolation between 78 and 2 can be used across the boundary and the "boundary" will no longer be a two dimensional surface but a three dimensional shell surrounding the biomolecule. Note that values 2 and 78 are the typical dielectric constants used for protein and water, alternative values are also allowed. When Equation (1.28) is solved by numerical methods, the electrostatic potential $\phi(x)$ over space is obtained. From $\phi(x)$, the electrostatic free energy of the system can be calculated.

If the biomolecule is solvated in a box of water that contains ionic species, like in the cellular environment, the charges on the ions need to be included in $\rho(x)$ for the accurate calculation of $\phi(x)$. In this case, the **Poisson-Boltzmann equation** is used instead:

$$\nabla \cdot \varepsilon \nabla \phi = -\left(\rho + \sum_i z_i c_i^{\text{bulk}} e^{-z_i \phi/(k_B T)} \right) \tag{1.30}$$

where index i represents the i^{th} ionic species in the solution; z_i is the charge carried by one ion i in the unit of elementary charge e; c_i^{bulk} is the bulk concentration for ion species i; k_B and T are Boltzmann constant and temperature. The Poisson-Boltzmann equation is strictly a steady-state equation that assumes that the ion distribution in the steady-state obeys the Boltzmann distribution $e^{-z_i \phi/(k_B T)}$.

The Smoluchowski equation can be used to prove the validity of the Boltzmann distribution of ions in the steady-state. The ion distribution achieves steady-state in solution by diffusion. In the steady-state, only the position of the ions matters and the ion density in space no longer changes with time. Take Equation (1.13) as the Smoluchowski equation, $D_2 = D$ is the diffusion constant; $D_1 = -(D/k_B T)(\partial z \phi / \partial x)$ is the drift velocity of the ion of charge z under electrostatic potential ϕ, ϕ is generated together by the charges on the biomolecule and the ions themselves. Equation (1.13) can

be written as:

$$\begin{aligned}
\frac{\partial c}{\partial t} &= -\frac{\partial}{\partial x}\left(-\frac{D}{k_B T}\frac{\partial z\phi}{\partial x}c(x)\right) + \frac{\partial^2}{\partial x^2}\left(Dc(x)\right) \\
\frac{\partial c}{\partial t} &= \frac{\partial}{\partial x}\left(\frac{D}{k_B T}\frac{\partial z\phi}{\partial x}c(x) + D\frac{\partial}{\partial x}c(x)\right) \\
\frac{\partial c}{\partial t} &= \frac{\partial}{\partial x}De^{-z\phi/(k_B T)}\frac{\partial}{\partial x}\left(e^{z\phi/(k_B T)}c(x)\right)
\end{aligned}$$

From Equation (1.17), $\partial c/\partial t = -\partial J/\partial x$, J is flux. And in the steady-state, the net flux of ions in solution is zero, so:

$$\begin{aligned}
J = -De^{-z\phi/(k_B T)}\frac{\partial}{\partial x}\left(e^{z\phi/(k_B T)}c(x)\right) &= 0 \\
\frac{\partial}{\partial x}\left(e^{z\phi/(k_B T)}c(x)\right) &= 0 \\
c(x) = c^{\text{bulk}}e^{-z\phi/(k_B T)} &
\end{aligned} \tag{1.31}$$

c^{bulk} appears due to the boundary condition used in the integration. At the "boundary", the bulk solution, the ion concentration equals to its bulk value. Equation (1.31) is exactly the **Boltzmann distribution**. And it is concluded that in steady-state, ions obey the Boltzmann distribution. Solving the Poisson-Boltzmann equation gives the electrostatic potential ϕ of the biomolecular system in an ionic solution. Some common Poisson-Boltzmann equation solvers are UHBD,[31] APBS[41] and DelPhi.[42] VISM is a newer development that uses a more physics-based approach to generate molecular surfaces.[43]

1.3.2 Solvation and Binding Free Energy

The Poisson-Boltzmann equation and the generalized Born model have been widely used in the calculation of the electrostatic component of the solvation free energy and binding free energy of biomolecules.[44,45] Molecular mechanics/generalized Born

surface area (MM/GBSA) and molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) are the names of the specific computational procedures.[46,47] Accurately calculating the free energies of biomolecular systems helps the understanding of their structure and function. Particularly, the calculation of binding free energy is essential in computer-aided drug discovery. MM/GBSA and MM/PBSA are relatively fast computational methods used to filter out the top drug binders out of hundreds of candidates for the further selection by the computationally more demanding and more accurate methods like thermodynamic integration.[48,49]

In MM/GBSA and MM/PBSA, the expression for the **solvation free energy** of a biomolecule is:[50]

$$\Delta G^{\text{solvation}} = (G^{\text{elec}}_{\text{solvated}} - G^{\text{elec}}_{\text{ref}}) + G^{\text{nonpolar}}_{\text{SASA}} \tag{1.32}$$

where $G^{\text{elec}}_{\text{solvated}}$ is the electrostatic interaction energy between the biomolecule and the solvent; $G^{\text{elec}}_{\text{ref}}$ is the electrostatic interaction energy between the biomolecule and an environment of the same dielectric constant as the biomolecule; $G^{\text{nonpolar}}_{\text{SASA}} = \gamma A$ is the nonpolar solvation free energy component calculated by the product of surface tension γ and the solvent accessible surface area (SASA) A of the biomolecule. The determination of the optimum location of the solvent accessible surface is itself a sub-field of study, software like VISM (variational implicit-solvent model) specifically aims at tackling this problem,[43,51] more details can be found in the references.

The expression for the **binding free energy** between biomolecules A and B is:[50]

$$\Delta G^{\text{binding}} = G_{\text{AB}} - G_{\text{A}} - G_{\text{B}} \tag{1.33}$$

$$G_{\text{A}} = \Delta G^{\text{solvation}}_{\text{A}} + G^{\text{Coul}}_{\text{A}} + G^{\text{vdW}}_{\text{A}} - TS_{\text{A}} \tag{1.34}$$

where AB is the bound complex of biomolecules A and B. Equation (1.34) also applies for G_{AB} and G_{B} in Equation (1.33). In Equation (1.34), $\Delta G^{\text{solvation}}_{\text{A}}$ is taken from Equation

(1.32); $G_{\mathrm{A}}^{\mathrm{Coul}}$ and $G_{\mathrm{A}}^{\mathrm{vdW}}$ are the Coulombic electrostatic interaction energy and the van der Waals interaction energy between the atoms within the biomolecule; T is temperature; S_{A} is the entropy of the biomolecule. S_{A} can be ignored, taken as only the translational and rotational entropy of a rigid biomolecule, or calculated by normal mode analysis[52] from a molecular dynamics trajectory depending on the nature of the biomolecular system and the desired accuracy.

In MM/GBSA and MM/PBSA, the atomic charges, radii and the van der Waals parameters of the biomolecule are assigned according to the choice of the molecular mechanics (MM), aka molecular dynamics, force field. Terms $\Delta G_{\mathrm{A}}^{\mathrm{solvation}}$, $G_{\mathrm{A}}^{\mathrm{Coul}}$ and $G_{\mathrm{A}}^{\mathrm{vdW}}$ in Equation (1.34) are all calculated based on the MM force field parameters. S_{A}'s dependency on the MM parameters is decided by how it is calculated. "GB" and "PB" reflect how the electrostatic terms in $\Delta G_{\mathrm{A}}^{\mathrm{solvation}}$ is calculated. "SA" reflects how the nonplolar term in $\Delta G_{\mathrm{A}}^{\mathrm{solvation}}$ is calculated.

All of the terms in Equations (1.32) and (1.34) can be calculated separately. And then they are combined to produce the final binding free energy between two biomolecules. Chapter 3 of this thesis studies a method that is potentially more accurate than the Poisson-Boltzmann equation in the calculation of the electrostatic terms in $\Delta G_{\mathrm{A}}^{\mathrm{solvation}}$.

The Poisson-Boltzmann equation is an approximate way to model the ionic solution. One of the approximations made is that the ions are point charges and they experience no steric hindrance from each other, i.e. no repulsive energy is reflected in the energy term $z\phi$ in Equation (1.31). The problem caused by this approximation is that even for only moderately charged biomolecular surfaces, there tends to be an over accumulation of counterions near the biomolecular surface, which leads to over-screening of the electrostatic potential exerted by the biomolecule. This problem is traditionally accounted for by using an ad hoc Stern layer,[53] while in Chapter 3, a more physics-based

method, **the size-modified Poisson-Boltzmann equation**, that uses the lattice gas theory to model the size of the ions[54] into the Poisson-Boltzmann equation is illustrated in the realistic lipid bilayer systems.[55]

1.4 References

1. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4th edition, 2002.

2. H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. *Molecular Cell Biology*. W. H. Freeman, New York, 5th edition, 2003.

3. S. Schuh. An actin-dependent mechanism for long-range vesicle transport. *Nature Cell Biol.*, 13:1431–1436, 2011.

4. P. M. Kekenes-Huskey, T. Liao, A. K. Gillette, J. E. Hake, Y. Zhang, A. P. Michailova, A. D. McCulloch, and J. A. McCammon. Molecular and subcellular-scale modeling of nucleotide diffusion in the cardiac myofilament lattice. *Biophys. J.*, 105(9): 2130–2140, 2013.

5. P. M. Kekenes-Huskey, A. K. Gillette, and J. A. McCammon. Predicting the influence of long-range molecular interactions on macroscopic-scale diffusion by homogenization of the smoluchowski equation. *J. Chem. Phys.*, 140(17):174106, 2014.

6. J. A. McCammon, S. H. Northrup, and S. A. Allison. Diffusional dynamics of ligand-receptor association. *J. Phys. Chem.*, 90(17):3901–3905, 1986.

7. R. R. Gabdoulline and R. C. Wade. Simulation of the diffusional association of barnase and barstar. *Biophys. J.*, 72(5):1917–1929, 1997.

8. E. D. Getzoff, D. E. Cabelli, C. L. Fisher, H. E. Parge, M. S. Viezzoli, L. Banci, and R. A. Hallewell. Faster superoxide dismutase mutants designed by enhancing electrostatic guidance. *Nature*, 358:347–351, 1992.

9. M. Roca, B. Messer, and A. Warshel. Electrostatic contributions to protein stability and folding energy. *FEBS Lett.*, 581(10):2065–2071, 2007.

10. C. T. Liu, J. P. Layfield, R. J. Stewart, J. B. French, P. Hanoian, J. B. Asbury, S. Hammes-Schiffer, and S. J. Benkovic. Probing the electrostatics of active site microenvironments along the catalytic cycle for Escherichia coli dihydrofolate reductase. *J. Am. Chem. Soc.*, 136(29):10349–10360, 2014.

11. C. J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. John Wiley & Sons, England, 2nd edition, 2004.

12. Wikipedia. Intel 4004 - Wikipedia, the free encyclopedia, 2015. URL https://en.wikipedia.org/wiki/Intel_4004. [Online; accessed 3-Aug-2015].

13. B. Helmuth, J. G. Kingsolver, and E. Carrington. Biophysics, physiological ecology, and climate change: Does mechanism matter? *Annu. Rev. Physiology*, 67:177–201, 2005.

14. M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Strut. Biol.*, 9:646–652, 2002.

15. A. C. Anderson. The process of structure-based drug design. *Chemistry & Biology*, 10(9):787–797, 2003.

16. G. S. Chen and J. W. Chern. *Drug Discovery Research: New Frontiers in the Post-Genomic Era*. John Wiley & Sons, Hoboken, New Jersey, 1st edition, 2007.

17. N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, Oxford, UK, 3rd edition, 2007.

18. W. T. Coffey, Y. P. Kalmykov, and J. T. Waldron. *The Langevin Equation: With Applications to Stochastic Problems in Physics, Chemistry and Electrical Engineering*. World Scientific Publishing Company, Singapore, 2nd edition, 2004.

19. J. Hake, A. G. Edwards, Z. Yu, P. M. Kekenes-Huskey, A. P. Michailova1, J. A. McCammon, M. J. Holst, M. Hoshijima, and A. D. McCulloch. Modelling cardiac calcium sparks in a three-dimensional reconstruction of a calcium release unit. *J. Physiology*, 590(18):4403–4422, 2012.

20. Y. Cheng, J. K. Suen, D. Zhang, S. D. Bond, Y. Zhang, Y. Song, N. A. Baker, C. L. Bajaj, M. J. Holst, and J. A. McCammon. Finite element analysis of the time-dependent Smoluchowski equation for acetylcholinesterase reaction rate calculations. *Biophys. J.*, 92(10):3397–3406, 2007.

21. R. E. Bank and M. Holst. A new paradigm for parallel adaptive meshing algorithms. *SIAM Rev.*, 45(2):291–323, 2003.

22. P. M. Kekenes-Huskey, A. Gillette, and J. Hake. Finite-element estimation of protein-ligand association rates with post-encounter effects: Applications to calcium binding in troponin C and SERCA. *Comp. Sci. & Disc.*, 5:014015, 2012.

23. A. Logg, K.-A. Mardal, and G. Wells. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, Berlin, Heidelberg, 1st edition, 2012.

24. N. Wang, G. A. Huber, and J. A. McCammon. Assessing the two-body diffusion

tensor calculated by the bead models. *J Chem. Phys.*, 138(20):204117, 2013.

25. R. Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, New York, 1st edition, 2001.

26. Lennart Sjögren. Non-equilibrium stochastic processes. Chapter 7 Brownian motion: Fokker-Planck equation. Lectures at the University of Gothenburg.

27. V. T. Metzger, C. Eun, P. M. Kekenes-Huskey, G. A. Huber, and J. A. McCammon. Electrostatic channeling in P. falciparum DHFR-TS: Brownian dynamics and Smoluchowski modeling. *Biophys. J*, 107(10):2394–2402, 2014.

28. P. Mereghetti, D. Kokh, J. A. McCammon, and R. C. Wade. Diffusion and association processes in biological systems: Theory, computation and experiment. *BMC Biophys.*, 4(1):2, 2011.

29. G. Wieczorek and P. Zielenkiewicz. Influence of macromolecular crowding on protein-protein association rates - a Brownian dynamics study. *Biophys. J*, 95(11): 5030–5036, 2008.

30. S. R. McGuffee and A. H. Elcock. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comp. Biol.*, 6(3): e1000694, 2010.

31. J. Madura, J. Briggs, R. Wade, M. Davis, B. Luty, A. Ilin, J. Antosiewicz, M. Gilson, B. Bagheri, L. Scott, and J. A. McCammon. Electrostatics and diffusion of molecules in solution - simulations with the University-of-Houston Brownian dynamics program. *Comput. Phys. Commun.*, 91(1):57–95, 1995.

32. M. Martinez, N. J. Bruce, J. Romanowska, D. B. Kokh, M. Ozboyaci, X. Yu, M. A. Ozturk, S. Richter, and R. C. Wade. SDA 7: A modular and parallel implementation of the simulation of diffusional association software. 36(21):1631–1645, 2015.

33. G. A. Huber and J. A. McCammon. Browndye: A software package for Brownian dynamics. *Comput. Phys. Commun.*, 181(11):1896–1905, 2010.

34. M. Dlugosz, P. Zielinski, and J. Trylska. Brownian dynamics simulations on CPU and GPU with BD_BOX. *J. Comp. Chem.*, 32(12):2734–2744, 2011.

35. S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

36. N. Wang and J. A. McCammon. Substrate channeling between the human dihydrofolate reductase and thymidylate synthase. *Protein Sci.*, 2015.

37. L. Y. Zhang, E. Gallicchio, R. A. Friesner, and R. M. Levy. Solvent models for protein-ligand binding: Comparison of implicit solvent Poisson and surface generalized Born models with explicit solvent simulations. *J. Comp. Chem.*, 22(6):591–607, 2001.

38. M.-Y. Shen and K. F. Freed. Long time dynamics of met-enkephalin: Comparison of explicit and implicit solvent models. *Biophys. J.*, 82(4):1791–1808, 2002.

39. M. Holst. *Multilevel Methods for the Poisson-Boltzmann Equation*. PhD thesis, University of Illinois at Urbana-Champaign, 1993.

40. V. Tsui and D. A. Case. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers*, 56(4):275–291, 2000.

41. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA*, 98(18):10037–10041, 2001.

42. L. Li, C. Li, S. Sarkar, J. Zhang, S. Witham, Z. Zhang, L. Wang, N. Smith, M. Petukh, and E. Alexov. DelPhi: A comprehensive suite for DelPhi software and associated resources. *BMC Biophys.*, 5(9), 2012.

43. S. Zhou, L. T. Cheng, J. Dzubiella, B. Li, and J. A. McCammon. Variational implicit solvation with Poisson-Boltzmann theory. *J. Chem. Theory Comput.*, 10(4): 1454–1467, 2014.

44. F. Fogolari, A. Brigo, and H. Molinari. The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *J. Mol. Recognition*, 15(6):377–392, 2002.

45. N. A. Baker. Reviews in computational chemistry, volume 21. Chapter 5. Biomolecular applications of Poisson-Boltzmann methods. 2005.

46. G. Rastelli, A. Del Rio, G. Degliesposti, and M. Sgobba. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comp. Chem.*, 31(4):797–810, 2009.

47. T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.*, 51(1):69–82, 2011.

48. T. P. Straatsma and H. J. C. Berendsen. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *J. Phys. Chem.*, 89(9):5876, 1988.

49. M. J. Mitchell and J. A. McCammon. Free energy difference calculations by thermo-

dynamic integration: Difficulties in obtaining a precise value. *J. Comp. Chem.*, 12 (2):271–275, 1991.

50. F. Dong, B. Olsen, and N. A. Baker. Computational methods for biomolecular electrostatics. *Methods Cell Biol.*, 84:843–870, 2008.

51. Z. Wang, J. Che, L. T. Cheng, J. Dzubiella, B. Li, and J. A. McCammon. Level-set variational implicit-solvent modeling of biomolecules with the Coulomb-field approximation. *J Chem Theory Comput.*, 8(2):386–397, 2012.

52. I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava. Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. *Chem Rev.*, 110(3):1463–1497, 2010.

53. O. Stern. The theory of the electrolytic double-layer. *Z. Elektrochem. Angew. Phys. Chem.*, 30:508–516, 1924.

54. I. Borukhov, D. Andelman, and H. Orland. Steric effects in electrolytes: A modified poisson-boltzmann equation. *Phys. Rev. Lett.*, 79(3):435–438, 1997.

55. N. Wang, S. Zhou, P. M. Kekenes-Huskey, B. Li, and J. A. McCammon. Poisson-Boltzmann versus size-modified Poisson-Boltzmann electrostatics applied to lipid bilayers. *J. Phys. Chem. B*, 118(51):14827–14832, 2014.

Chapter 2

Assessing the Two-Body Diffusion Tensor Calculated by the Bead Models

2.1 Abstract

The diffusion tensor of complex macromolecules in Stokes flow is often approximated by the bead models. The bead models are known to reproduce the experimental diffusion coefficients of a single macromolecule, but the accuracy of their calculation of the whole multi-body diffusion tensor, which is important for Brownian dynamics simulations, has not been closely investigated. As a first step, we assess the accuracy of the bead model calculated diffusion tensor of two spheres. Our results show that the bead models produce very accurate diffusion tensors for two spheres where a reasonable number of beads are used and there is no bead overlap.

2.2 Introduction

Calculating the diffusion tensor of the macromolecules is an essential step in their Brownian dynamics simulations.[1] An analytical form of the diffusion tensor of macromolecules based on their exact complex structure is impossible to obtain and the tensor is commonly approximated numerically by the bead models. The bead model was inspired by Kirkwood's fundamental works on the hydrodynamic properties of macromolecules.[2,3,4] Its initial form was proposed by Bloomfield et al;[5,6,7] and later it was extensively studied and implemented in scientific software packages by the García de la Torre group.[8,9,10,11,12]

The bead model belongs to the type of diffusion tensor calculation methods that approximate the complex structure of macromolecules by smaller geometrically simple frictional elements for which an exact or approximate analytical diffusion tensor is available. And then these methods tackle the solvable frictional element-level hydrodynamic problems in place of the analytically unsolvable macromolecule-level problem. We summarize that all of the bead models have two key components: the way the macromolecules are approximated by a collection of beads, which we call the bead representation, and the way in which one formulates the diffusion tensor between beads. We also want to point out that the frictional elements can be points as in Kirkwood's model, spherical beads as in the bead models, triangular patches as in a boundary element method,[13] or possibly other geometrical shapes. Here we study the bead model because it is more widely used in Brownian dynamics simulations than the boundary element method and it gives more accurate results than Kirkwood's "point model".

Generally, the bead models can be applied to two types of Brownian dynamics simulations. The first type is the Brownian dynamics simulation of one macromolecule,[14,15] where the macromolecule is set to be at least partly flexible. Each (bio)chemical unit,

which can be an atom, amino acid, or even a protein domain, is approximated as one bead, and only the bead-level diffusion tensor needs to be calculated. The second type is the Brownian dynamics simulation of the diffusional association of multiple macromolecules,[16,17,18] where the macromolecules are modeled as rigid bodies and (preferably) beads are packed on the surface of the macromolecules without considering their chemical structures.[11] Aside from calculating the bead-level diffusion tensor, an extra step is needed to convert the bead-level diffusion tensor into the diffusion tensor of the macromolecules. Note that the above two types of simulations can be combined with one another and do not necessarily have to incorporate the bead models or any kind of diffusion tensor calculations. The hydrodynamic interactions of the macromolecules are often ignored in order to save computational time.

The calculation of the one-body diffusion coefficients, i.e. the diagonal elements of the diffusion tensor, using the bead models has been shown to be computationally manageable and accurate.[11,12,19] However, the performance of the bead models in calculating the diffusion coefficients and the off-diagonal elements of the diffusion tensors for two or more bodies has not been closely investigated.

In this work, we focus on the application of the bead models in the diffusional association type of Brownian dynamics simulations. As a first step, we assess the accuracy of the bead model calculation of the whole diffusion tensor of two spheres. We chose the two-sphere system because first, it is the simplest two-body system and each sphere represents one of the two biomolecules in a two-body diffusional association process. Second, the analytical and series solution of the diffusion tensor for the two-sphere system exists,[20] to which we can easily compare our bead model calculation results. Specifically, to build our bead models, we chose a type of bead representation called the shell (S) representation,[5,10] which uses a shell of beads to represent the surface of the spheres. The bead-level diffusion tensor was then calculated by the popular modified

Rotne-Prager (RP) tensor.[8,21,22] From here on, "bead model" and "SRP model" both refer to the bead models consisting of the shell representation and the RP tensor.

2.3 The Mathematical Formulation of the Bead Model

2.3.1 The Two-Sphere Diffusion Tensor $\mathbb{D}$

The system studied here is two rigid non-overlapping equal-sized spheres in Stokes flow. The velocities/angular velocities of the spheres and the external forces/torques acting on the spheres are related to each other through the diffusion tensor $\mathbb{D}$ and the resistance tensor $\mathbb{R}$:

$$V = \mathbb{D} \cdot F \tag{2.1}$$

$$F = \mathbb{R} \cdot V \tag{2.2}$$

$$\mathbb{D} = \begin{bmatrix} \mathbb{D}_{11}^{tt} & \mathbb{D}_{12}^{tt} & \mathbb{D}_{11}^{tr} & \mathbb{D}_{12}^{tr} \\ \mathbb{D}_{21}^{tt} & \mathbb{D}_{22}^{tt} & \mathbb{D}_{21}^{tr} & \mathbb{D}_{22}^{tr} \\ \mathbb{D}_{11}^{rt} & \mathbb{D}_{12}^{rt} & \mathbb{D}_{11}^{rr} & \mathbb{D}_{12}^{rr} \\ \mathbb{D}_{21}^{rt} & \mathbb{D}_{22}^{rt} & \mathbb{D}_{21}^{rr} & \mathbb{D}_{22}^{rr} \end{bmatrix} \tag{2.3}$$

$$F = \begin{bmatrix} F_1 & F_2 & T_1 & T_2 \end{bmatrix}^T \tag{2.4}$$

$$V = \begin{bmatrix} V_1 & V_2 & \Omega_1 & \Omega_2 \end{bmatrix}^T \tag{2.5}$$

Both $\mathbb{D}$ and $\mathbb{R}$ are in the form of symmetric 12 by 12 matrices. $\mathbb{D}$ is the inverse of $\mathbb{R}$ and vice versa. Each sub-matrix in Equation (2.3) is itself a 3 by 3 (x,y,z by x,y,z) matrix. The subscripts 1 and 2 refer to sphere 1 and 2, the superscripts t and r refer to the translational and rotational motions. In Equations (2.4) and (2.5), F_i, T_i, V_i and Ω_i are the external force, external torque, velocity and angular velocity of sphere i, $i \in \{1,2\}$.

Except for the 12 diagonal elements in $\mathbb{D}$, the element on the column i and the row j in $\mathbb{D}_{cd}^{ab}$, $i,j \in \{x,y,z\}$, $a,b \in \{r,t\}$, $c,d \in \{1,2\}$, models the coupling between the a motion of sphere c along the i-axis and the b motion of sphere d along the j-axis. The diagonal elements are not coupling terms, they model the translational or rotational motion of a sphere along an axis.

We define a "block" of the diffusion tensor to be an arbitrary group of terms in the tensor (the locations of those terms in the tensor do not have to form a connected rectangular block); for example, the tt block refers to all terms with the superscript tt. In the Numerical Results section, we assess the accuracy of the whole tensor and its different blocks separately.

The analytical and series expressions of the terms in $\mathbb{D}$ and $\mathbb{R}$ for two spheres have been derived over the years and are summarized in a paper by Jeffrey and Onishi (JO).[20] Here, we call it the JO solution set; it will be considered to give the exact values for $\mathbb{D}$ and all of our SRP model calculation results will be compared against it to assess the model's accuracy.

2.3.2 The Bead-level Diffusion Tensor D

One of the two key components of a bead model is the way in which one formulates the diffusion tensor between beads. In this work, for a two-sphere system, we always use the same number of spherical beads for each sphere. If each sphere is represented by N beads and the system contains $2N$ beads in total, then the grand diffusion tensor D and the grand resistance tensor R for all of the beads are:

$$v = D \cdot f \tag{2.6}$$

$$f = R \cdot v \tag{2.7}$$

$$D = \begin{bmatrix} D_{11} & \dots & D_{1,2N} \\ \vdots & \ddots & \vdots \\ D_{2N,1} & \dots & D_{2N,2N} \end{bmatrix} \tag{2.8}$$

Similar to Equations (2.4) and (2.5), f and v are the force/torque and velocity/angular velocity vectors of all of the beads. Theoretically, in this coupled $2N$-body system, each D_{ij} is a 6 by 6 matrix containing four blocks with superscripts tt, tr, rt, rr, and depends on the coordinates and sizes of all of the beads. Unfortunately, such a complicated grand diffusion tensor is unsolvable. To make the calculations feasible, we only use approximate diffusion tensors for the many-bead system. Particularly, in the modified Rotne-Prager (RP) tensor,[8,21,22] only the two-body hydrodynamic interactions are taken into account to the 3rd order of the inverse of the inter-bead distances. The rotational motions of the beads are also ignored. The form of the RP tensor is as follows:

When $i = j$

$$D_{ii} = \frac{1}{6\pi\mu a_i} I \tag{2.9}$$

When $i \neq j$

$$D_{ij} = \frac{1}{8\pi\mu r_{ij}^3} \left[\left(r_{ij}^2 I + r_{ij} \otimes r_{ij} \right) + \frac{a_i^2 + a_j^2}{r_{ij}^2} \left(\frac{r_{ij}^2}{3} I - r_{ij} \otimes r_{ij} \right) \right]$$

$$(a_i + a_j \leq r_{ij}) \tag{2.10}$$

$$D_{ij} = \frac{1}{6\pi\mu a} \left[\left(1 - \frac{9}{32} \frac{r_{ij}}{a} \right) I + \frac{3}{32} \frac{r_{ij} \otimes r_{ij}}{a r_{ij}} \right]$$

$$(a_i + a_j > r_{ij}) \tag{2.11}$$

where a_i is the radius of bead i, $a = (a_i + a_j)/2$, r_{ij} is the vector pointing from the center of bead i to the center of bead j, r_{ij} is the length of r_{ij}, $\otimes$ represents outer product, μ is the solvent viscosity, and I is a 3 by 3 identity matrix. Note that Equation (2.11) only

applies to beads with the same radius.

Aside from the RP tensor we use here, there is the lower order Oseen tensor,[23] the higher order Reuland-Felderhof-Jones,[24] Mazur-van Saarloos,[25] Goldstein[26] tensors and the JO solution set (not exact for the many-bead system). However, the RP tensor appears to be the best compromise between accuracy and computational power and it is the most commonly used diffusion tensor in Brownian dynamics simulations of biomolecules. So in this work, we use only the RP tensor to calculate the bead-level diffusion tensors.

2.3.3 From D to $\mathbb{D}$

To apply the bead models to the diffusional association of rigid macromolecules, $\mathbb{D}$ needs to be calculated from D. During the D to $\mathbb{D}$ calculation procedure, all of the inter-bead hydrodynamic interactions are summed up to get the total hydrodynamic interaction between the two spheres. One of the steps is the inversion of D. The computational complexity of the conventional matrix inversion operation for a n by n matrix is $O(n^3)$, making it the most time-consuming step in the bead model diffusion tensor calculations. The calculation procedure can be found in previous works.[8,27,28] Here we used an equivalent but slightly different approach described below.

We first define a "12-case" velocity matrix, $V^{\text{12-case}}$, for the two-sphere system. The twelve cases correspond to the twelve degrees of freedom of the system; in order they are: the translation of sphere 1 along x, y and z-axis, the translation of sphere 2 along x, y and z-axis, the rotation of sphere 1 about x, y and z-axis, the rotation of sphere 2 about x, y and z-axis. In the i-th case, we let the system move by unit speed/angular speed only along its i-th degree of freedom. From the twelve cases, we obtain twelve velocity/angular velocity vectors of the system; together we write them into $V^{\text{12-case}}$ and

it is numerically equal to an identity matrix:

$$V^{\text{12-case}} = \begin{bmatrix} V^{\text{case 1}} & \cdots & V^{\text{case 12}} \end{bmatrix}_{12\times 12} = I \tag{2.12}$$

And then from Equation (2.2):

$$\begin{aligned} \mathbb{R} \cdot V^{\text{12-case}} &= F^{\text{12-case}} \\ \mathbb{R} &= F^{\text{12-case}} \\ \mathbb{D} &= (F^{\text{12-case}})^{-1} \end{aligned} \tag{2.13}$$

$F^{\text{12-case}}$ is the matrix that contains the forces and torques experienced by the two spheres in the twelve cases. $\mathbb{R}$ and $F^{\text{12-case}}$ are only numerically identical; matrix I is not explicitly written out in Equation (2.13) but it bears units.

Having obtained Equation (2.13), the next step is to calculate $F^{\text{12-case}}$, invert it and get $\mathbb{D}$. Since the two spheres are rigid, for each case, we can easily calculate the velocity vector, $v^{\text{case } i}$, of all the beads given the velocity/angular velocity vector, $V^{\text{case } i}$, of the spheres. Note that since we use the RP tensor, the angular velocities of the beads do not enter $v^{\text{case } i}$.

$$v^{\text{case } i} = \begin{bmatrix} V_1^{\text{case } i} + \Omega_1^{\text{case } i} \times r_1 \\ \vdots \\ V_1^{\text{case } i} + \Omega_1^{\text{case } i} \times r_N \\ V_2^{\text{case } i} + \Omega_2^{\text{case } i} \times r_{N+1} \\ \vdots \\ V_2^{\text{case } i} + \Omega_2^{\text{case } i} \times r_{2N} \end{bmatrix} \tag{2.14}$$

where r_i is the vector that starts at the center of the sphere that bead i is on and ends at the center of bead i. After calculating the twelve $v^{\text{case } i}$, we write them together as $(v^{\text{12-case}})_{6N\times 12}$. The grand resistance tensor R is the inverse of the grand diffusion tensor

D, which is calculated by the RP tensor, and we have:

$$\begin{aligned} f^{\text{12-case}} &= R \cdot v^{\text{12-case}} \\ &= D^{-1} \cdot v^{\text{12-case}} \end{aligned} \tag{2.15}$$

Here we obtain $f^{\text{12-case}}$, which stores the twelve force vectors of the beads in the twelve cases. For each case, we can calculate the force/torque vector $F^{\text{case } i}$ of the spheres from force vector $f^{\text{case } i}$ of the beads by:

$$F^{\text{case } i} = \begin{bmatrix} \sum_{j=1}^{N} f_j^{\text{case } i} \\ \sum_{j=N+1}^{2N} f_j^{\text{case } i} \\ \sum_{j=1}^{N} r_j \times f_j^{\text{case } i} \\ \sum_{j=N+1}^{2N} r_j \times f_j^{\text{case } i} \end{bmatrix} \tag{2.16}$$

where subscript j is the index of the beads. This way we obtain $F^{\text{12-case}}$ from $f^{\text{12-case}}$. And last, we obtain $\mathbb{D}$ as the numerical inverse of $F^{\text{12-case}}$.

2.4 Generating Bead Representations For Spheres

The other key component of the bead models is the way in which one approximates the macromolecules by a collection of spherical beads, or as we call it, a bead representation. There are several types of bead representations.[10] Here we work with one type of bead representation only, the shell representation (most literature calls it "shell model", but to avoid ambiguity, we call it the "shell representation" here). It only assigns beads to the surface of the spheres and it best represents the physical picture of

the hydrodynamics of the rigid nonporous spheres because hydrodynamic interactions are fully screened in the interior of the spheres.

2.4.1 Building the Shell Representations

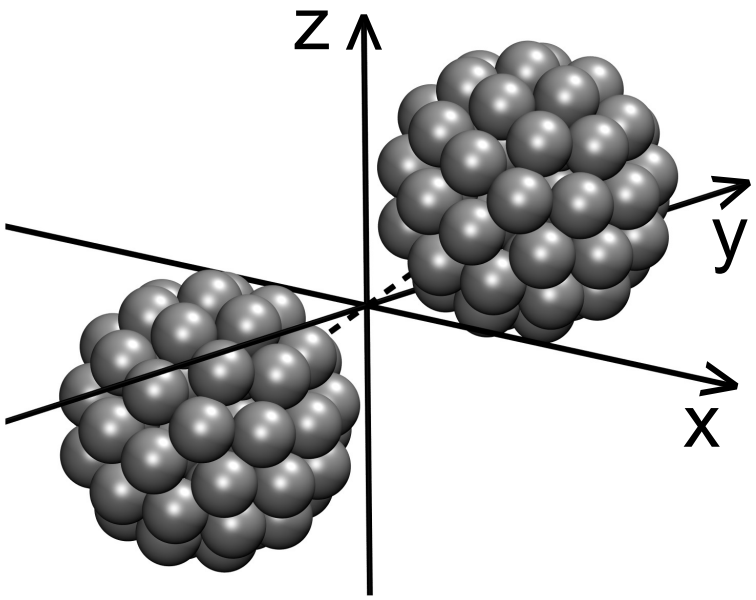

Figure 2.1: The shell representation for two spheres. Both spheres are placed along the vector (1,1,1), the dashed line, at an equal distance from the origin. In this example, both spheres have the same radius and each sphere surface consists of 50 beads.

To build a N-bead shell representation for a single sphere, we restrict the N beads to be tangential to the sphere surface from the inside and apply a pseudo-repulsive force between all pairs of beads. We iteratively minimize the pseudo-potential energy of the system and adjust the bead radius (which is same for all of the beads) such that no beads overlap and at least one pair of beads is tangential to each other.

The shell representation we use here is different from the commonly seen shell representations.[5,29] In our shell building method, the surface of the sphere is roughly evenly divided by the beads, the shell of beads is smooth, and we can build a shell representation with an arbitrary number of beads. We believe that such shell representation

will more evenly account for the hydrodynamic interactions at every part of the sphere surface. A view of our shell representation for the two-sphere system is shown in Figure 2.1. We put the two spheres along vector (1,1,1) instead of an axis to avoid zero entries in the JO solution set calculated diffusion tensor and so the accuracy of all of the entries in the diffusion tensor can be assessed.

2.4.2 Calibrating the Shell Representations

Before attempting to calculate the two-sphere diffusion tensor using the bead models we made, we want to make sure that they can be used to calculate the single-sphere (ss) diffusion tensor accurately. The single-sphere diffusion tensor follows Stokes' law; its unitless form is:

$$\mathbb{D}_{\text{ss}} = I \tag{2.17}$$

where I is a 6 by 6 identity matrix. The diagonal elements of the $\mathbb{D}_{\text{ss}}$ calculated by our bead models that use the shell representations built in the previous section are all bigger than 1. And the off-diagonal coupling terms are non-zero; but they have very small magnitudes ($< 10^{-5}$). This is because, according to the way we build the shell representations, there are holes in between the beads and the bead centers are located on the inside of the sphere. These two factors lead to underestimated frictional forces experienced by the sphere.

We calibrate the bead models calculated $\mathbb{D}_{\text{ss}}$ to be as close to the Stokes' law as possible by scaling the shell representations. According to Equations (2.9)-(2.11), the grand diffusion tensor, Equation (2.8), is inversely proportional to the length scale of the shell representation, as such, we can shrink or stretch the shell representation of a bead model by multiplying a scaling factor c to all of the bead coordinates and radii to make its calculated $\mathbb{D}_{\text{ss}}$ as close to I as possible.

We first define the optimized scaling factor, c_{opt}, which scales the bead model calculated $\mathbb{D}_{ss}$ to have the smallest standard deviation from I. For comparison, we also defined another two scaling factors, c_t and c_r, that respectively scales the calculated average translational and rotational diffusion coefficients to 1.

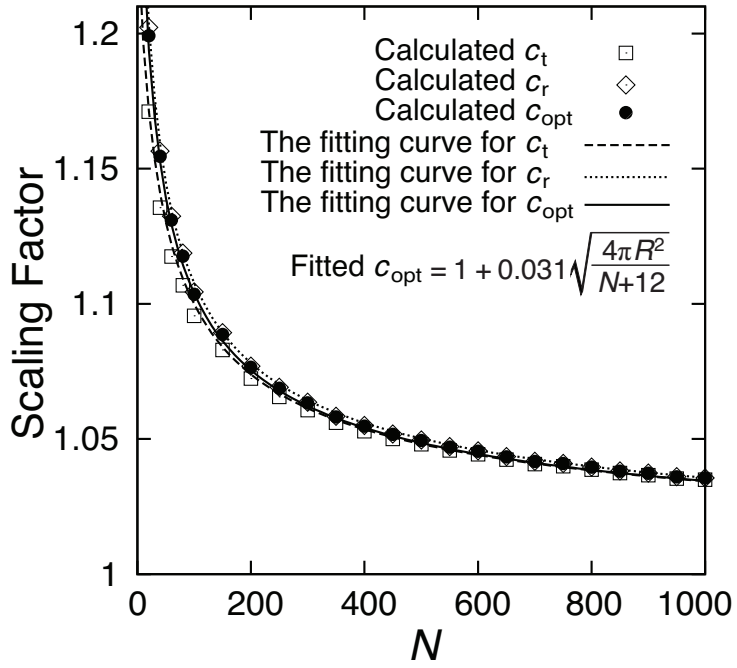

Figure 2.2: The values of the scaling factors for the bead models with different shell representations. The scaling factors calibrate the bead model calculated single-sphere diffusion tensors to be as close to the Stokes' law as possible. When a larger number of beads are used, the shell representation requires less scaling (scaling factor closer to 1), and therefore a smaller degree of calibration. Scaling factors c_t, c_r and c_{opt} are as described in the text. Their values are calculated for 23 bead models and plotted with respect to N, the number of beads in a bead model. The R is the radius of the sphere, set to be 10 unit length here.

The values of the scaling factors only depend on two parameters, the number of beads used, N, and the sphere radius, R (Figure 2.2). The value of c_{opt} falls between c_t and c_r and all three scaling factors are very close to each other at all N. After scaling with c_{opt}, for all 23 bead models used Figure 2.2, the average diffusion coefficient is less than 2% different from the Stokes' law. We can also see that, when more beads are used (larger N), less scaling, therefore less calibration, is needed (scaling factors closer to 1, 1

means no scaling). This means that the bead models with more beads give more accurate $\mathbb{D}_{ss}$ without scaling.

According to our original shell representation building approach, for the two-sphere system, no bead overlap occurs as long as the spheres do not overlap. However, after scaling, the shell representations get stretched out (all the scaling factors are bigger than 1). The beads on different spheres will overlap when the two spheres are close. The smaller the number of beads used, the more the overlap when the two spheres are nearly touching.

In the following Numerical Results section, all of the bead models used are scaled by c_{opt}.

2.5 Numerical Results

For all of the test cases shown in this section, the two-sphere diffusion tensor was calculated by both the JO solution set and the SRP model, and we denote the two differently calculated diffusion tensors as $\mathbb{D}_{\text{JO}}$ and $\mathbb{D}_{\text{SRP}}$. Then, we measured the accuracy of the diffusion tensor by Frobenius norm and Pearson's Correlation Coefficient. Note that all of the lengths and tensors calculated are converted to be unitless for easy comparison.[20]

First, we want to compare the differences between the magnitudes of $\mathbb{D}_{\text{JO}}$ and $\mathbb{D}_{\text{SRP}}$. The magnitude of individual tensors can be calculated from their Frobenius norm, see Equation (2.18). Likewise the magnitude of the difference between $\mathbb{D}_{\text{JO}}$ and $\mathbb{D}_{\text{SRP}}$ can be calculated from the Frobenius norm of the difference of the two tensors. We standardize the magnitude difference by the norm percent deviation (NPD), see Equation

(2.19), to make the difference comparable across different system configurations.

$$\|\mathbb{D}\| = \sqrt{\sum_{i,j=1}^{12} d_{ij}^2} \tag{2.18}$$

$$\text{NPD} = \frac{\|\mathbb{D}_{\text{SRP}} - \mathbb{D}_{\text{JO}}\|}{\|\mathbb{D}_{\text{JO}}\|} \tag{2.19}$$

where d_{ij} denotes the element in the diffusion tensor and NPD $\in [0, +\infty)$. NPD tells us the average accuracy of the elements in $\mathbb{D}_{\text{SRP}}$, but it does not say anything about the accuracy of the individual elements. To address whether the accuracies of the individual elements in $\mathbb{D}_{\text{SRP}}$ are similar, we calculate the Pearson's correlation coefficient (PCC) between the two tensors. PCC measures the degree of linear correlation, or proportionality, between $\mathbb{D}_{\text{SRP}}$ and $\mathbb{D}_{\text{JO}}$ and is defined by Equation (2.20):

$$\text{PCC} = \frac{\sum_{i,j=1}^{12} \left((d_{\text{JO}})_{ij} - \bar{d}_{\text{JO}}\right)\left((d_{\text{SRP}})_{ij} - \bar{d}_{\text{SRP}}\right)}{\sqrt{\sum_{i,j=1}^{12} \left((d_{\text{JO}})_{ij} - \bar{d}_{\text{JO}}\right)^2}\sqrt{\sum_{i,j=1}^{12} \left((d_{\text{SRP}})_{ij} - \bar{d}_{\text{SRP}}\right)^2}} \tag{2.20}$$

where $\bar{d}_{\text{JO}}$ and $\bar{d}_{\text{SRP}}$ denote the mean of the elements in $\mathbb{D}_{\text{JO}}$ and $\mathbb{D}_{\text{SRP}}$ and PCC $\in [-1, 1]$. When PCC = 0, the two tensors have no linear correlation and the accuracies of the elements in $\mathbb{D}_{\text{SRP}}$, i.e. the differences between each element in $\mathbb{D}_{\text{SRP}}$ and its corresponding element in $\mathbb{D}_{\text{JO}}$, have large standard deviation. When PCC = ± 1, $\mathbb{D}_{\text{SRP}} = \pm\alpha\mathbb{D}_{\text{JO}}$ and NPD$= |1 \mp \alpha|$, α is a positive constant, the accuracies of the elements in $\mathbb{D}_{\text{SRP}}$ are the same, despite the fact that they can be equally high or equally low. Note that the signs in the diffusion tensor have physical meanings, any PCC ≤ 0 indicates a wrongly calculated $\mathbb{D}_{\text{SRP}}$. An accurate bead model calculation should give an NPD value close to 0 and a PCC value close to 1. Note that both the NPD and PCC values can be calculated for any block of $\mathbb{D}_{\text{SRP}}$ with respect to the corresponding block of $\mathbb{D}_{\text{JO}}$.

2.5.1 The Whole Tensor

We first assessed the accuracy of the whole diffusion tensor calculated by six different bead models at various sphere separations (Figure 2.3). We see that as the two spheres get further apart or as more beads are used in the calculation, the NPD value gets closer to 0 and the PCC value gets closer to 1 consistently, meaning an increasing accuracy of the bead model calculated diffusion tensors. However, the degree of improvement of the NPD and PCC values also gets smaller at the same time.

In Figure 2.3, for all of the bead models used, when there are bead overlaps ($d <$ the sphere separations marked by the filled dots), the NPD values are much bigger than 0 and the PCC values are less close to 1, meaning a poor accuracy of the bead model calculated diffusion tensors. Also, with overlaps, the NPD and PCC curves show irregular behaviors, e.g. the 50-bead model curve crossing over the 100-bead model curve in both Figure 2.3(A) and (B). But as the sphere separation gets larger, the NPD and PCC values sharply tend towards 0 and 1. When the gap between the bead shells fits exactly one bead ($d =$ the sphere separations marked by the filled triangles), the NPD and PCC values have basically converged.

As a side note here, since none of the shell representations for the sphere are symmetric along all the lines that pass through the sphere center, different orientations of the shell representation for each bead model lead to slightly different $\mathbb{D}_{\mathrm{SRP}}$ tensors. For each bead model calculation in Figure 2.3, we took 10 random orientations of the bead model to calculate the NPD and PCC values. However, the standard deviations of the 10 NPD and the 10 PCC values for the 20-bead model when the two bead shells are exactly touching are only 0.0002 and 0.000004. For models with more beads and at larger separations, the standard deviations are essentially zero. Since different orientations give little differences in most cases, from here on, we will only use one orientation for each bead model calculation.

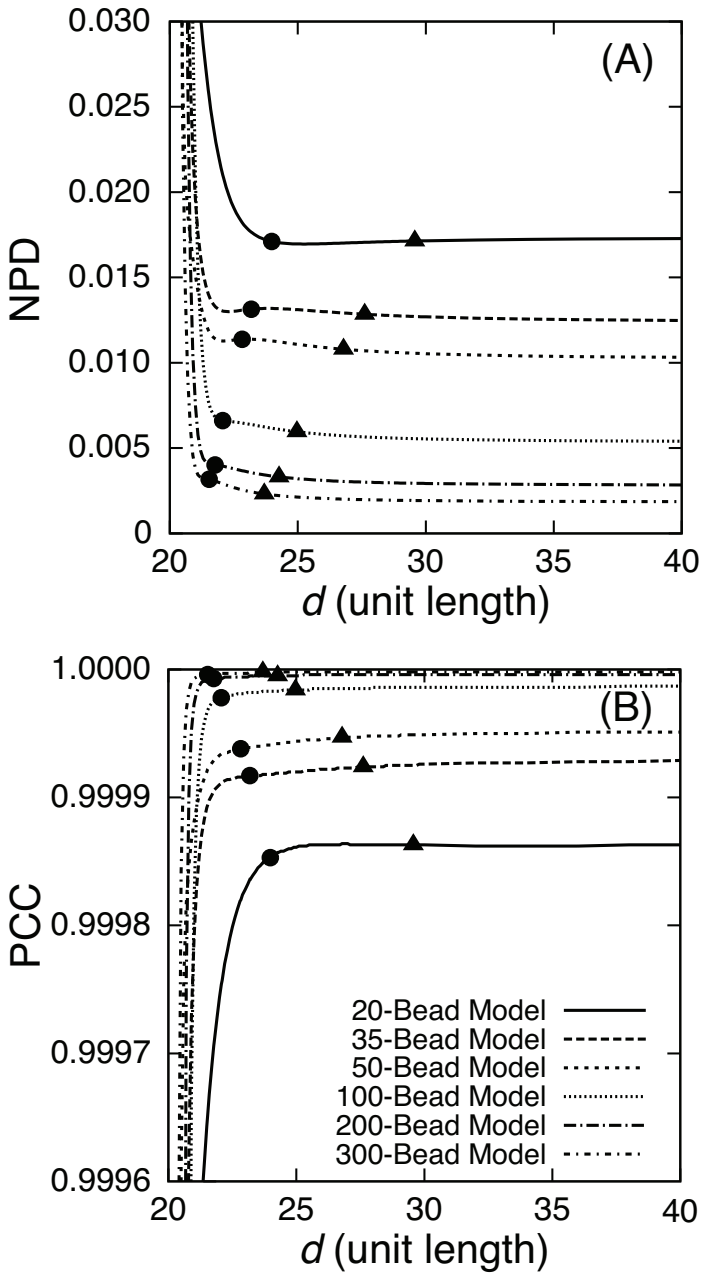

Figure 2.3: The accuracy of different bead models at different sphere separations. The bead model calculated diffusion tensor becomes more accurate (the NPD value approaches 0 and the PCC value approaches 1) as the distance between the spheres increases or as more beads are used in the calculation. The sphere separation, d, is the distance between the centers of two spheres. The two spheres here both have a radius of 10 in unit length. For each bead model, the filled dots mark the distance at which the gap between the outer surfaces of the two bead shells is 0. The filled triangles mark the distance at which the gap between the outer surfaces of the two bead shells can fit exactly one bead.

2.5.2 Different Tensor Blocks

Here, we used the 50-bead model for the investigation of the accuracy of different tensor blocks. In Figure 2.4, we present only the NPD values because, similarly to Figure 2.3, the NPD and PCC values change with d according to the same trend; presenting both of them would be repetitive.

In Figure 2.4, the tensor blocks are sorted by the type of motion coupling, the coupling between the translational motions (tt), the coupling between the rotational motions (rr), and the coupling between the translational and the rotational motions (tr/rt). For each type of motion coupling, the bead model always calculates the coupling between the motions of different spheres more accurately than the coupling between the motions of the same sphere. For example, in Figure 2.4(A), the tensor block consists of $\mathbb{D}^{tt}_{12}$ and $\mathbb{D}^{tt}_{21}$ achieves lower NPD values than the tensor block consists of the off-diagonal elements of $\mathbb{D}^{tt}_{11}$ and $\mathbb{D}^{tt}_{22}$. Aside from this observation, the NPD value of the $\mathbb{D}^{tt}_{12}$, $\mathbb{D}^{tt}_{21}$ block converges to near 0 at large d, while the NPD value of the off-diagonal $\mathbb{D}^{tt}_{11}$, $\mathbb{D}^{tt}_{22}$ block keeps increasing as d increases, meaning that the elements in the block only get more and more different from the exact solution as d increases. Similar results are seen in Figure 2.4(B) and (C).

The NPD values of the diagonal elements in $\mathbb{D}$ are presented separately because they model the translational or rotational motions of the single spheres, not the couplings between the motions. The accuracy of the bead model calculated diagonal elements falls between the accuracies of the couplings terms between spheres and the couplings terms within each sphere; as can be seen from their median NPD values in Figure 2.4(A) and (B). The NPD values of the whole tensor $\mathbb{D}$ are also plotted in Figure 2.4 as a reference.

In Figure 2.4, bead overlaps again lead to inaccuracies (large NPD values) and irregular NPD trends. However, except for the NPD values of the off-diagonal 11/22 sub-blocks, the NPD values of the tensor blocks improve and converge quickly as d

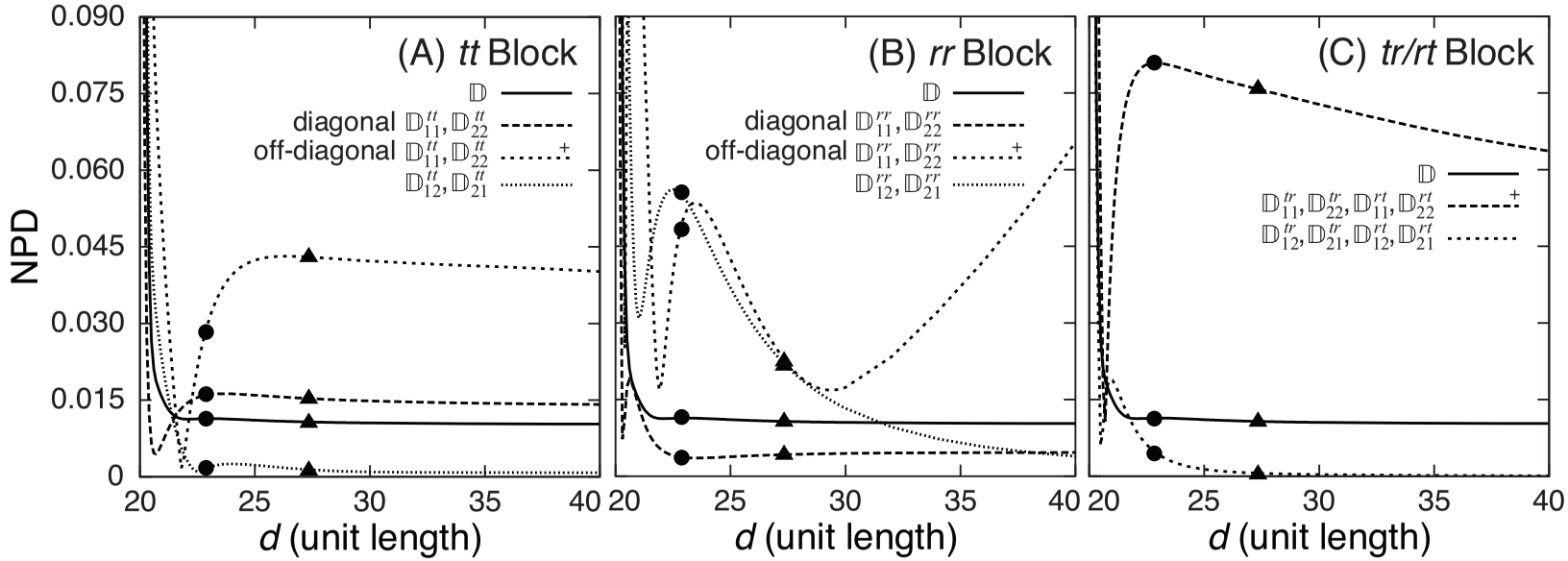

Figure 2.4: The accuracy of the different blocks of the bead model calculated diffusion tensor. The NPD values of the tensor blocks of the 50-bead model calculated diffusion tensors in Figure 2.3 are plotted (the filled dots and triangles here have the same meanings as in Figure 2.3). For the each type of motion coupling (tt, rr and tr/rt), the bead model always calculates the coupling between the motions of different spheres more accurately than the coupling between the motions of the same sphere. Consequently, the 12/21 sub-blocks always achieve lower NPD values than the non-diagonal 11/22 sub-blocks. The NPD values of the non-diagonal 11/22 sub-blocks, marked with plus signs in the legends, never converge and keep increasing as d increases.

increases and bead overlaps disappear.

2.6 Discussion

In this paper, we assess the accuracy of the whole diffusion tensor calculated by the bead models through the calculation of NPD and PCC values instead of assessing the accuracy of only the diagonal elements of the tensor.[11,12] In the Brownian dynamics simulations that take hydrodynamics into account, the whole diffusion tensor should be used,[1] therefore it is important to assess the accuracy of all its entries.

The presented NPD and PCC measurements have direct physical meanings. In Brownian dynamics simulations, an NPD value closer to 0 means a more accurate average propagation speed of the system by $\mathbb{D}_{\mathrm{SRP}}$; A PCC value closer to 1 means a more accurate shape of the propagation path generated by $\mathbb{D}_{\mathrm{SRP}}$. If we perform Brownian dynamics simulations of the same system with the same sequence of random kicks, but two different diffusion tensors $\mathbb{D}_{\mathrm{SRP}}$ and $\mathbb{D}_{\mathrm{JO}}$ and we obtain two trajectories, when NPD = 0 and PCC = 1, $\mathbb{D}_{\mathrm{SRP}} = \mathbb{D}_{\mathrm{JO}}$ and the two trajectories can be exactly superimposed onto each other. When NPD > 0 and PCC = 1, $\mathbb{D}_{\mathrm{SRP}} = \alpha\mathbb{D}_{\mathrm{JO}}$ and NPD = $|1-\alpha|$. For this case, the two trajectories will have the same shape, but the $\mathbb{D}_{\mathrm{JO}}$ trajectory needs to be enlarged/shrunk by a factor of α to be superimposed onto the $\mathbb{D}_{\mathrm{SRP}}$ trajectory; This is because the system propagated by $\mathbb{D}_{\mathrm{SRP}}$ will move a factor of α faster/slower than the system propagated by $\mathbb{D}_{\mathrm{JO}}$. When NPD = 0 and PCC < 1, on average the systems in the two trajectories move at the same speed, but they follow two different propagation paths that cannot be superimposed onto each other by simple enlarging or shrinking.

The accuracy of the bead model calculated diffusion tensors gets better when more beads are used in the calculation. The reason for this is that when more beads are used, the bead model becomes equivalent to the boundary element method,[13] which becomes

equivalent to the exact solution to the Stokes equations as the boundary elements become infinitesimally small. The accuracy also gets better when the two spheres are further apart. Because the RP tensor was derived for well-separated beads and as the spheres get further apart, the beads on different spheres get further apart, the bead configuration fits more into the regime where the RP tensor was derived.

The magnitude accuracy (reflected by the NPD values) and proportionality accuracy (reflected by the PCC values) improve and decline according to the same trend. This means that, e.g. in simulations, the two-sphere system trajectory generated by the bead models with more beads will follow a more accurate path with a more accurate speed. In Figure 2.3, all six bead models acquired a less than 2% difference between the magnitudes of $\mathbb{D}_{\mathrm{SRP}}$ and $\mathbb{D}_{\mathrm{JO}}$, and a PCC value bigger than 0.9998, meaning that all of the $\mathbb{D}_{\mathrm{SRP}}$ are nearly proportional to $\mathbb{D}_{\mathrm{JO}}$. This is a quite satisfying result, because, even with a small number of beads (20 beads), the $\mathbb{D}_{\mathrm{SRP}}$ is still fairly accurate.

The hydrodynamic interactions between the beads on the same sphere are always more poorly accounted for than those between the beads on different spheres. In the two-sphere system, the hydrodynamic interactions between the different degrees of freedom of the same sphere should diminish quickly as the two spheres move apart. But because of the intrinsic asymmetry of the shell representations, these hydrodynamic interactions never disappear. And this leads to the big inaccuracy of the off-diagonal elements with subscripts 11 or 22 at large sphere separation. Fortunately, these elements are usually small enough that they do not contribute significantly to the overall error of $\mathbb{D}_{\mathrm{SRP}}$. The RP tensor does not predict the hydrodynamic interactions between non-overlapping, but closely placed, beads as accurately as those of distantly placed beads, primarily because the derivation of the RP tensor assumes that the beads are well separated.

It has been reported that the bead models using the RP tensor do not produce accurate enough rotational diffusion coefficients because the rotational motion is only

accounted for at the center of the beads.[19,28,30] Our bead models with scaling greatly improved the accuracy of the rotational tensor block and the whole tensor. The term $0.031(4\pi R^2/(N+12))^{1/2}$ in the fitted formula of the scaling factor shown in Figure 2.2 makes up for the unaccounted hydrodynamic interactions due to the holes between the beads on a sphere surface of area $4\pi R^2$ and the unaccounted bead volumes. Note that when using scaling in our calculations, the ranking of the average accuracy of the three types of coupling terms, tt, rr and tr/rt, depends on the choice of the scaling factor. We chose c_{opt} in our calculations because it optimizes the average accuracy of all terms in $\mathbb{D}$. On the other hand, e.g. c_{t} optimizes the average accuracy of the tt terms only and leaves the rr and tr/rt terms less accurate.

For a two-sphere system, the bead models give consistent accuracy except for at small sphere separations. In general, for any bead model used here, when the bead shells of the two spheres overlap, the accuracy of $\mathbb{D}_{\text{SRP}}$ is always fairly low (Figure 2.3), and the accuracies of different tensor blocks show irregular improvements and declinations (Figure 2.4). Using the RP tensor on overlapping beads requires extra precaution,[31] and it may lead to unexpected results as shown here. Avoiding the bead overlapping situation helps us calculate more accurate diffusion tensors. In the diffusional association Brownian dynamics simulations of two proteins modeled by beads, one could use less beads at large separations to save computation times and more beads at small separations to avoid overlaps.

The method of Brownian dynamics simulations with hydrodynamics calculated by the bead models was first developed more than 30 years ago.[1] Until today, the full hydrodynamics of the simulated biomolecules is still often ignored due to the size and complexity of the biomolecules. Usually, only the diagonal elements of the bead model calculated diffusion tensors are used. And in terms of the diffusional association simulations, the diffusion tensors are often pre-calculated for isolated biomolecules

instead of calculated on-the-fly for the whole simulation system involving multiple biomolecules. This is mainly due to the challenge of quickly inverting and decomposing the diffusion tensors for large simulation systems. Both the conventional matrix inversion operation in the bead model calculations and the matrix decomposition operation in the Brownian dynamics simulations scale as $O(N^3)$, where N is the number of frictional elements in the system. Significant computational power is required to improve the bead model accuracy. Faster matrix operation methods exist[32] and should be applied in the Brownian dynamics simulations with the bead models for practical reasons.

2.7 Conclusions

In general, our bead models produce very accurate two-sphere diffusion tensors when the bead shells do not overlap and a reasonable number of beads are used (more than around 20 beads for each sphere). Through this work, we gained confidence in the accuracy of the bead model calculated diffusion tensors for two bodies with simple geometry. Tests with realistic globular biomolecules should be performed as the next investigation on the capability of the bead models. Continuous efforts are being made in developing faster matrix operation algorithms for diffusion tensors, as well as more accurate ways to calculate diffusion tensors from the bead models. We believe that, with these efforts, we can eventually use the bead models to calculate realistic diffusion tensors in the Brownian dynamics simulations of biomolecules.

2.8 Acknowledgments

This work has been supported by the National Institutes of Health, National Science Foundation, the Howard Hughes Medical Institute, the Center for Theoretical

Biological Physics at UCSD, and the National Biomedical Computational Resource.

Chapter 2 is a minimally modified reprint of the material as it appears in Nuo Wang, Gary A. Huber, and J. Andrew McCammon, "Assessing the Two-Body Diffusion Tensor Calculated by the Bead Models," Journal of Chemical Physics, 2013. The dissertation author was the primary investigator and author of this paper.

2.9 References

1. D. L. Ermak and J. A. McCammon. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.*, 69(4):1352–1360, 1978.

2. J. G. Kirkwood and J. Riseman. The intrinsic viscosities and diffusion constants of flexible macromolecules in solution. *J. Chem. Phys.*, 16(6):565–573, 1948.

3. J. G. Kirkwood. The statistical mechanical theory of irreversible processes in solutions of flexible macromolecules. visco-elastic behavior. *Recl. Trav. Chim. Pays-Bas*, 68(7):649–660, 1949.

4. J. G. Kirkwood. The general theory of irreversible processes in solutions of macromolecules. *J. Poly. Sci.*, 12(1):1–14, 1954.

5. V. A. Bloomfield, K. E. Van Holde, and W. O. Dalton. Frictional coefficients of multisubunit structures. I. Theory. *Biopolymers*, 5(2):135–148, 1967.

6. V. A. Bloomfield, K. E. Van Holde, and W. O. Dalton. Frictional coefficients of multisubunit structures. II. Application to proteins and viruses. *Biopolymers*, 5: 149–159, 1967.

7. D. P. Filson and V. A. Bloomfield. Shell model calculations of rotational diffusion coefficients. *Biochemistry*, 6(6):1650–1658, 1967.

8. J. M. García Bernal and V. A. Bloomfield. Hydrodynamic properties of macromolecular complexes. I. Translation. *Biopolymers*, 16(8):1747–1763, 1977.

9. J. García de la Torre, S. Navarro, M. C. López Martínez, F. G. Diaz, and J. J. López Cascales. HYDRO: A computer program for the prediction of hydrodynamic properties of macromolecules. *Biophys. J.*, 67(2):530–531, 1994.

10. B. Carrasco and J. García de la Torre. Hydrodynamic properties of rigid particles: Comparison of different modeling and computational procedures. *Biophys. J.*, 76(6): 3044–3057, 1999.

11. J. García de la Torre, M. L. Huertas, and B. Carrasco. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J.*, 78(2): 719–730, 2000.

12. A. Ortega, D. Amorós, and J. García de la Torre. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophys. J.*, 101(4):892–898, 2011.

13. S. Aragon. A precise boundary element method for macromolecular transport properties. *J. Comp, Chem.*, 25(9):1191–1205, 2004.

14. R. C. Wade, M. E. Davis, B. A. Luty, J. D. Madura, and J. A. McCammon. Gating of the active site of triose phosphate isomerase: Brownian dynamics simulations of flexible peptide loops in the enzyme. *Biophys. J.*, 64(1):9–15, 1993.

15. C. E. Chang, T. Y. Shen, J. Trylska, V. Tozzini, and J. A. McCammon. Gated binding of ligands to hiv-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophys. J.*, 90(11):3880–3885, 2006.

16. S. H. Northrup, S. A. Allison, and J. A. McCammon. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *J. Chem. Phys.*, 80(4):1517–1524, 1984.

17. R. Gabdoulline and R. Wade. Simulation of the diffusional association of barnase and barstar. *Biophys. J.*, 72(5):1917–1929, 1997.

18. G. A. Huber and J. A. McCammon. Browndye: A software package for brownian dynamics. *Comput. Phys. Commun.*, 181(11):1896–1905, 2010.

19. B. Carrasco and J. García de la Torre. Improved hydrodynamic interaction in macromolecular bead models. *J. Chem. Phys.*, 111(10):4817–4826, 1999.

20. D. J. Jeffrey and Y. Onishi. Calculation of the resistance and mobility functions for two unequal rigid spheres in low-reynolds-number flow. *J. Fluid Mech.*, 139: 261–290, 1984.

21. J. Rotne and S. Prager. Variational treatment of hydrodynamic interaction in polymers. *J. Chem. Phys.*, 50(11):4831–4837, 1969.

22. H. Yamakawa. Transport properties of polymer chains in dilute solution: Hydrodynamic interaction. *J. Chem. Phys.*, 53(1):436–443, 1970.

23. C. W. Oseen. *Hydrodynamik*. Leipzig: Akademische Verlag, 1927.

24. P. Reuland, B.U. Felderhof, and R.B. Jones. Hydrodynamic interaction of two spherically symmetric polymers. *Physica A*, 93(3-4):465–475, 1978.

25. P. Mazur and W. van Saarloos. Many-sphere hydrodynamic interactions and mobilities in a suspension. *Physica A*, 115(1-2):21–57, 1982.

26. R. F. Goldstein. Macromolecular diffusion constants: A calculational strategy. *J. Chem. Phys.*, 83(5):2390–2397, 1985.

27. J. M. García Bernal and J. García de la Torre. Transport properties and hydrodynamic centers of rigid macromolecules with arbitrary shapes. *Biopolymers*, 19(4):751–766, 1980.

28. J. García de la Torre and V. Rodes. Effects from bead size and hydrodynamic interactions on the translational and rotational coefficients of macromolecular bead models. *J. Chem. Phys.*, 79(5):2454–2460, 1983.

29. J. García de la Torre. Building hydrodynamic bead-shell models for rigid bioparticles of arbitrary shape. *Biophys. Chem.*, 94(3):265–274, 2001.

30. J. García de la Torre, G. del Rio Echenique, and A. Ortega. Improved calculation of rotational diffusion and intrinsic viscosity of bead models for macromolecules and nanoparticles. *J. Phys. Chem. B*, 111(5):955–961, 2007.

31. B. Carrasco, J. García de la Torre, and P. Zipper. Calculation of hydrodynamic properties of macromolecular bead models with overlapping spheres. *Eur. Biophys. J.*, 28(6):510–515, 1999.

32. R. R. Schmidt, J. G. Hernández Cifre, and J. García de la Torre. Comparison of brownian dynamics algorithms with hydrodynamic interaction. *J. Chem. Phys.*, 135 (8):084116, 2011.

Chapter 3

Poisson-Boltzmann vs. Size-modified Poisson-Boltzmann Electrostatics Applied to Lipid Bilayers

3.1 Abstract

Mean-field methods, such as the Poisson-Boltzmann equation (PBE), are often used to calculate the electrostatic properties of molecular systems. In the past two decades, an enhancement of the PBE, the size-modified Poisson-Boltzmann equation (SMPBE), has been reported. Here, the PBE and the SMPBE are reevaluated for realistic molecular systems, namely, lipid bilayers, under eight different sets of input parameters. The SMPBE appears to reproduce the molecular dynamics simulation results better than the PBE only under specific parameter sets but in general, it performs no better than the Stern layer correction of the PBE. These results emphasize the need for careful discussions of the accuracy of mean-field calculations on realistic systems with respect to the choice of parameters and call for reconsideration of the cost-efficiency and the

significance of the current SMPBE formulation.

3.2 Introduction

Mean-field methods provide a way to coarse grain the electrostatic interactions between the solvent and the biomolecules of interest.[1] They are commonly used to generate the electrostatic potentials needed for biomolecular diffusion simulations[2,3,4,5] and are used in calculations of the solvation free energies of biomolecules.[6,7]

The hundred-year-old Poisson-Boltzmann equation (PBE) is currently the flagship of the mean-field methods.[8,9,10] Despite the success of the PBE,[11,12] it is built upon several approximations; for example, the ions in the PBE are considered infinitesimally small. This approximation runs into trouble near highly-charged biomolecular surfaces, such as those of typical enzyme active sites.[13] The reason is that, under the strong electrostatic attractions near highly-charged surfaces, the concentration of the counterion can exceed its maximally allowed value without the restriction from steric hindrance; this will then lead to an overestimation of electrostatic screening and an underestimation of the electrostatic potential exerted by the biomolecule.[14,15] To remedy the lack of ion steric hindrance, or ion-size effects, in the PBE, a Stern layer, or ion-exclusion layer, outside of the molecular surface has been used, within which no ion is allowed.[16]

To improve the PBE, the size-modified Poisson-Boltzmann equation (SMPBE), which incorporates the finite ion sizes through a more physical lattice gas formulation, was developed and has been applied to biomolecular systems.[14,17,18,19,20] The computational cost of the SMPBE is comparable to that of the PBE when there are only two ion sizes.[18] However, in its generalized form that can handle an arbitrary number of different ion sizes, the SMPBE takes much longer to solve.[19,21]

It is known that mean-field methods are parameter-dependent and that there is

no standard way of choosing input parameters that guarantee the optimum accuracy of prediction.[22,23,24,25] The SMPBE is reported to reproduce experimental[18] and molecular dynamics simulation[20] results better than the PBE, but there still lacks a comprehensive comparison between these two mean-field methods using a variety of input parameters to determine whether the SMPBE generally outperforms the PBE.

This paper discusses the accuracy of prediction and the cost-efficiency of the SMPBE with respect to the PBE using several different parameter sets. Given that the mean-field methods are parameter-dependent, this work also discusses the existence of a single parameter set with which the mean-field methods give consistent high accuracy of prediction across different ionic strengths and surface charge densities. Specifically, the equilibrium ion distributions outside of neutral and charged lipid bilayers are calculated by the PBE and the SMPBE using eight different parameter sets and are compared to the MD simulation resultsq. The long-term goal of this work is to identify appropriate tools for modeling diffusion dynamics of molecules that are of biological or medical interest.[26,27] In these diffusion processes, the system evolves from a nonequilibrium state to the equilibrium state and this work is an initial step to test the ability of the SMPBE in predicting ion distributions in the equilibrium state. Lipid bilayers are considered here due to their profound importance in the activity of membrane proteins and cross-membrane signaling pathways.[28,29]

3.3 Methods

3.3.1 The SMPBE Implementation

The SMPBE formulation used here is capable of calculating the electrostatic potential and the ion distributions for molecular systems containing an arbitrary number of ion species with different ion sizes.[19,21] A brief description of the formulation is

included in Section 3.8.1. This formulation is an extension of the earlier foundational SMPBE theory by Borukhov et al,[14] which, in its original form, can only handle one same-size cation-anion pair.

This SMPBE formulation is implemented in the Adaptive Poisson-Boltzmann Solver (APBS),[30] and is freely available upon request. A brief explanation of the SMPBE routine in APBS is included in Section 3.8.2.

3.3.2 Numerical Calculations

Molecular dynamics (MD) simulations can generate the equilibrium ion distributions in a molecular system at higher accuracy and resolution than the mean-field methods, but also at significantly greater computational expense.[1] In this work, the MD simulation results are used as a standard to assess the accuracy of the PBE and the SMPBE predictions.

Molecular Dynamics Simulations

The dimensions of the lipid bilayer systems used in the MD simulations and the mean-field calculations are illustrated in Section 3.8.5 Figure 3.4. In this work, eight lipid bilayer MD systems are simulated (Table 3.1).

1-palmitoyl-2-oleoyl phosphatidylcholine (POPC) is a partially charged, but overall neutral, lipid molecule and each 1-palmitoyl-2-oleoyl phosphatidylserine (POPS) lipid molecule carries one negative charge. All of the systems in Table 3.1 are built by the online application CHARMM-GUI Membrane Builder.[31] The MD simulations are performed by NAMD2.9 software package[32] with CHARMM36[33] lipid force field[34] using mostly the default system settings offered by the CHARMM-GUI Membrane Builder (see Section 3.8.3). Each system in Table 3.1 is simulated for 100 to 150ns. All of the POPC simulations are 150ns long and reach convergence after 20ns. All

Table 3.1: The composition of the molecular dynamics systems. POPC: 1-palmitoyl-2-oleoyl phosphatidylcholine; POPS: 1-palmitoyl-2-oleoyl phosphatidylserine. The bulk salt concentration is the averaged salt (NaCl, KCl) concentration at $|z|$=60Å (z, the perpendicular distance to the bilayer center, is defined in Section 3.8.5 Figure 3.4) over the converged portion of the MD trajectories.

System	Lipid	Cation	Anion	Bulk salt concentration
1	48 POPC	10 Na^+	10 Cl^-	0.060 M
2	48 POPC	20 Na^+	20 Cl^-	0.126 M
3	48 POPC	40 Na^+	40 Cl^-	0.268 M
4	48 POPC	10 K^+	10 Cl^-	0.060 M
5	48 POPC	20 K^+	20 Cl^-	0.134 M
6	48 POPC	40 K^+	40 Cl^-	0.275 M
7	48 POPS	68 Na^+	20 Cl^-	0.170 M
8	48 POPS	68 K^+	20 Cl^-	0.169 M

of the POPS simulations are 100ns long and converge after 30ns. The converged MD trajectories are utilized for all of the data analysis (see Seciton 3.8.3 for the convergence criteria).

Mean-field Calculations

To test the accuracy of a mean-field method in predicting the ion distributions in a MD system listed in Table 3.1, three MD frames are evenly sampled from the converged MD trajectory and a mean-field calculation is performed for each frame; the final mean-field results are averaged over all three calculations. Note that APBS has no periodic boundary conditions in its finite difference routine. To avoid boundary effects, the lipid bilayer structures obtained from the MD simulations are manually extended 0.25 times along the positive and negative x,y-axis, while the mean-field results are only collected above and below the original size of the lipid bilayer as in the MD simulations, see Section 3.8.5 Figure 3.4. A 161(x) by 161(y) by 193(z) finite difference grid is used for all of the mean-field calculations following the APBS electrostatic focusing scheme30 and the grid spacing is 0.43Å(x) by 0.43Å(y) by 0.47Å(z), a reasonable resolution for calculating ion

distributions. The lipid atomic charges and radii are taken from CHARMM36 charges and CHARMM36 van der Waals radii. Both the PBE and the SMPBE are solved in their nonlinear form. The majority of the input parameters for the mean-field methods follow the default settings of the APBS input preparation program PDB2PQR[35] except for the ones mentioned in the next section. See Section 3.8.4 for an example of the APBS input file. Data analysis in this work utilizes GROMACS4.5.5,[36] VMD1.9.1[37] and customized scripts.

3.4 Explanations on the Eight Mean-Field Parameter Sets

The use of different mean-field method input parameters is key to the discussion of the accuracy of predictions of the SMPBE with respect to those of the PBE. Among all of the mean-field calculation input parameters or settings, two are selected for the discussion. First, the ion size is chosen mainly because it is a particularly important input parameter for the SMPBE and it is also used to define the size of the Stern layer. Second, the definition of molecular surface is chosen for its direct relationship to ion-accessibility, thus the ion distributions. Specifically, four definitions of molecular surface, Figure 3.1, and two sets of ion sizes, Table 3.2, are tested, yielding a permutation of eight different sets of input parameters.

Table 3.2: The two sets of ion radii. The van der Waals (VDW) radii are taken from the CHARMM36 ion VDW radii. The radial distribution function (RDF) radii are the first-peak positions of the ion-lipid carbonyl oxygen RDFs from the MD simulations.

	VDW radius	RDF radius
Na^+	1.4Å	2.3Å
K^+	1.8Å	2.6Å
Cl^+	2.3Å	4.0Å

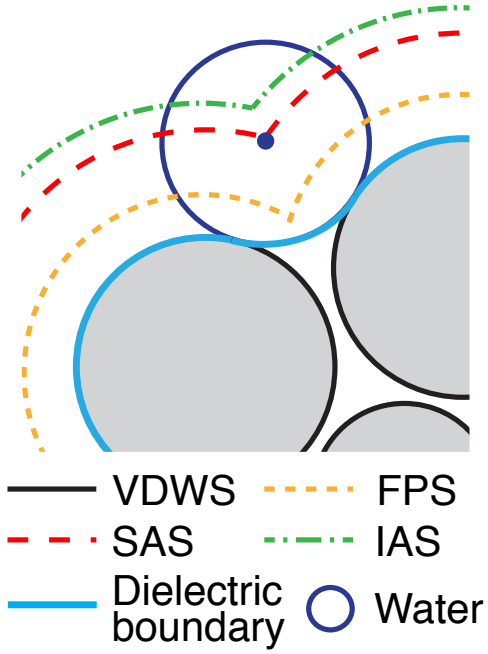

Figure 3.1: The definitions of molecular surface. VDWS: the van der Waals surface, drawn around the lipid atom centers with the CHARMM36 VDW radii, i.e. the Lennard-Jones $r_{min}/2$ values. FPS: the "first-peak" surface, drawn according to the MD cation-lipid carbonyl oxygen radial distribution function (RDF) first-peak positions. SAS: the solvent-accessible surface, drawn by the center of the water probe of radius 1.4Å[38] rolling across VDWS. IAS: the ion-accessible surface, drawn in the same way as the SAS, only that the probe radius is that of the counterion. The dielectric boundaries in all of the mean-field calculations are the same and are defined by the solid blue line (or equivalently the solvent-exclusion surface).

Three key points need to be clarified. First, despite the multiple definitions of the "molecular surface" in Figure 3.1, only the van der Waals surface (VDWS) is commonly recognized as the molecular surface. Second, the reason for the flexible use of the term "molecular surface" in this work is as follows. Supposedly, the SMPBE, without the help of the Stern layer, is able to predict solvation properties comparably accurately or more accurately than the PBE with a Stern layer. However, as it is shown later in the Results section, unexpectedly, using the SMPBE outside of the VDWS without a Stern layer leads to large ion concentration overestimation. To carry out further investigations, several alternative "molecular surfaces" that are further away from the lipid molecules than the

VDWS are tested. Finally, the seemingly peculiar use of the FPS is inspired by the fact that the MD RDF first-peak positions are often smaller than the sum of the VDW radii of the cation and its association partner, the lipid carbonyl oxygen.[39] FPS is generated by uniformly expanding the VDWS until the minimum FPS-lipid carbonyl oxygen center distance is equal to the counterion-lipid carbonyl oxygen MD radial distribution function (RDF) first-peak position. The VDWS is expanded by 0.6Å for the Na^+ FPS and 0.9Å for the K^+ FPS. Using the FPS in the mean-field methods could theoretically reproduce the MD results better. However, this turns out not to be the case as shown in the Results section. The RDF radius is chosen because it was used in a previous work[20] and an alternative set of ion radii is needed to show how the PBE and the SMPBE results change with respect to the change of ion sizes.

3.5 Results and Discussion

3.5.1 The PBE vs. the SMPBE

Figure 3.2 shows an array of calculations on the number of Na^+ ions bound per POPC molecule. The corresponding results for the K^+ ion are shown in Section 3.8.5 Figure 3.5. The "number of ions bound" is calculated by integrating the concentration of ions within 25Å to the lipid bilayer center. An example of the detailed position-dependent ion concentration is shown in Section 3.8.5 Figure 3.6.

By a crude observation, the PBE and the SMPBE reproduce MD results the best in Figure 3.2(g), while the PBES (the PBE with Stern layer; the size of the Stern layer is always set to be the radius of the cation) performs the best in Figure 3.2(a). However, it is important to note that the PBES calculation in Figure 3.2(a) is actually the same as the PBE calculation in Figure 3.2(g). This is because that the ion-accessible surface (IAS) used in Figure 3.2(g) exactly overlaps with the outer surface of the Stern layer in Figure

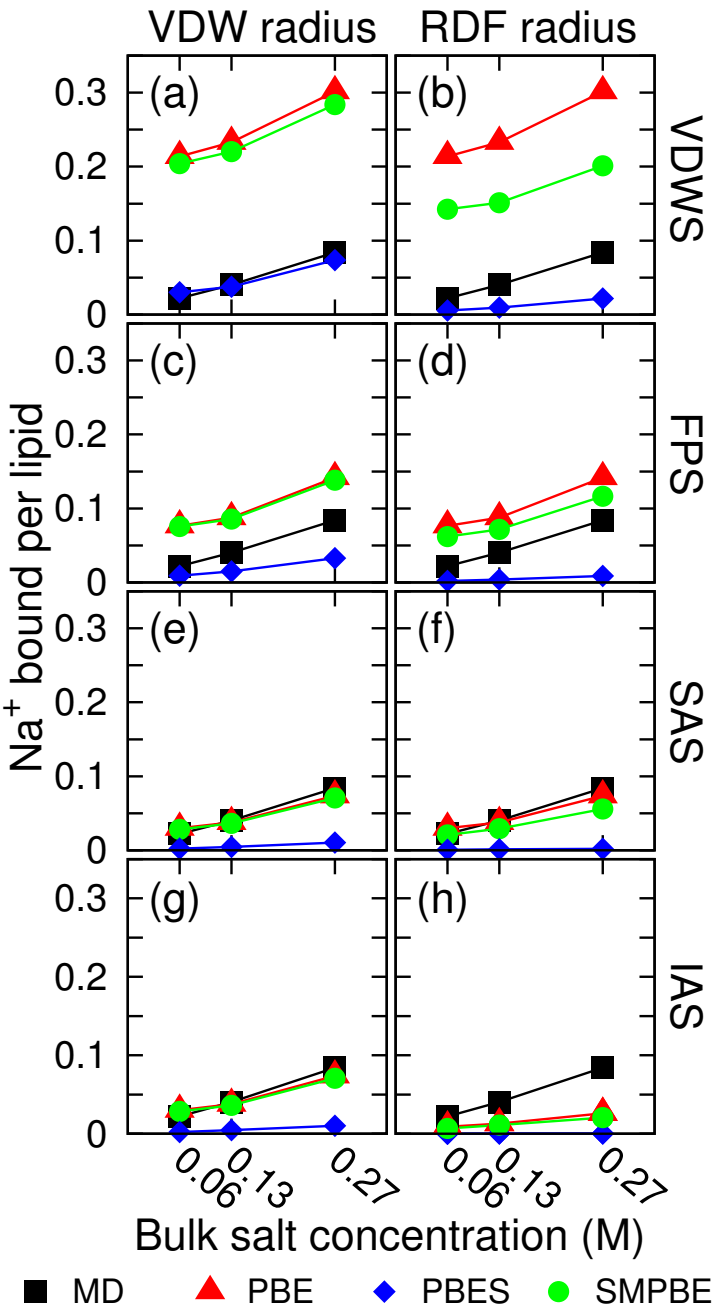

Figure 3.2: The number of Na^+ bound per POPC molecule at three NaCl bulk concentrations (systems 1 to 3 in Table 3.1). MD: molecular dynamics; PBE: nonlinear Poisson-Boltzmann equation (without Stern layer); PBES: PBE with Stern layer; SMPBE: size-modified Poisson-Boltzmann equation (without Stern layer). Subfigures (a) to (h) use eight different parameter sets; each is a combination of a molecular surface (VDWS, FPS, SAS, IAS, see Figure 3.1) and an ion radius set (VDW radius, RDF radius, see Table 3.2). (g) turns out to be identical to (e) because the Na^+ VDW radius happens to be the same as the radius of water, 1.4Å , and so the SAS in (e) is the exactly same as the IAS in (g).

3.2(a). Also, it should be reemphasized that the VDWS is the physical molecular surface and it is only fair to compare the SMPBE results without the use of a Stern layer to the PBE results with the use of a Stern layer. With this said, it can be concluded from Figure 3.2(a) and (b) that when the SMPBE is used outside of the physical molecular surface without a Stern layer, it significantly overestimates the MD ion concentrations compared to the PBE calculation with the aid of a Stern layer - the size-modified correction of the PBE does not seem to be as effective as the "Stern-layer-modified" correction. This is a surprising result considering the fact that the POPC bilayer has zero net surface charge and yet the capability of the SMPBE is not enough to limit the ion concentrations to the correct range. Furthermore, caution should be taken to distinguish the real source of the corrections or errors. Given the poor performance of the SMPBE in Figure 3.2(a), it is clear that the reproduction of the MD results by the SMPBE in Figure 3.2(g) mainly owes to the introduction of an effective Stern layer through the use of the IAS instead of the SMPBE formulation itself. And the reason for the underestimation of the MD results by the PBES in Figure 3.2(g) is the redundant use of a second Stern layer on top of the IAS.

Fortunately, as expected, at least the use of the SMPBE consistently alleviates the overestimation of ion concentrations by the PBE. The strength of SMPBE correction increases as the molecular surface shrinks (IAS to VDWS), which causes ions to experience a stronger electrostatic potential, and as the ion radii increase (VDW radius to RDF radius), which results in more steric exclusion between ions in the SMPBE formulation. However, in general, the SMPBE gives similar results to the PBE (Figure 3.2(a), (c), (e), (f), (g) and (h)). A previous study has shown that the electrostatic free energies predicted by the PBE and the SMPBE can be quite different,[40] but in this work, the PBE and the SMPBE electrostatic potentials are shown to also be very similar with most parameter sets, Section 3.8.5 Figure 3.7. As mentioned in the introduction, the SMPBE has a comparable computational cost to the PBE when two ion sizes are used,[18] but the cost

significantly increases when the SMPBE is generalized to handle an arbitrary number of different ion sizes because the SMPBE can no longer be expressed analytically.[19,21] If the simple Newton's method is used to solve the generalized SMPBE, the SMPBE is up to hundreds of times slower than the PBE. Without developing a more efficient SMPBE solver, only in the two-ion-size case is SMPBE a cost-efficient choice for its degree of correction on the PBE. However, even with a faster SMPBE solver, the Stern layer remains a more effective correction than the current SMPBE formulation at no extra computational cost.

3.5.2 The Existence of the Best Parameter Set

From Figure 3.2, it can be said that the VDWS combined with a Stern layer of the size of the VDW radius of the cation is the parameter set that gives the best PBE prediction of ion distributions. However, Figure 3.3 shows that this does not hold true for a different lipid bilayer surface charge density (the corresponding figure for the K^+ ion is the Section 3.8.5 Figure 3.8).

The surface charge density of the simulated POPS bilayer is about $0.28 C/m^2$ or $1.73 e/nm^2$ (for reference, the surface charge density of DNA is about $1 e/nm^2$), a fairly high value. The gray shade positions in Figure 3.3(a) and (b) show that the thickness of the Stern layer for the POPS bilayer needs to be smaller than what is used for the POPC bilayer (the VDW radius of Na^+) for the PBE to reproduce the MD ion concentrations. Figure 3.3 reveals that the same parameter set does not give consistent accuracy of prediction when the surface charge density changes. This is also true, although not as significantly, when the bulk ionic strength, or the bulk ion concentration, varies. When looked at closely, the slopes of the PBE, the PBES and the SMPBE plots in Figure 3.2 are always equal to or smaller than that of the MD, leading to inconsistent performances of a parameter set across different ionic strengths. Additionally, the same parameter set

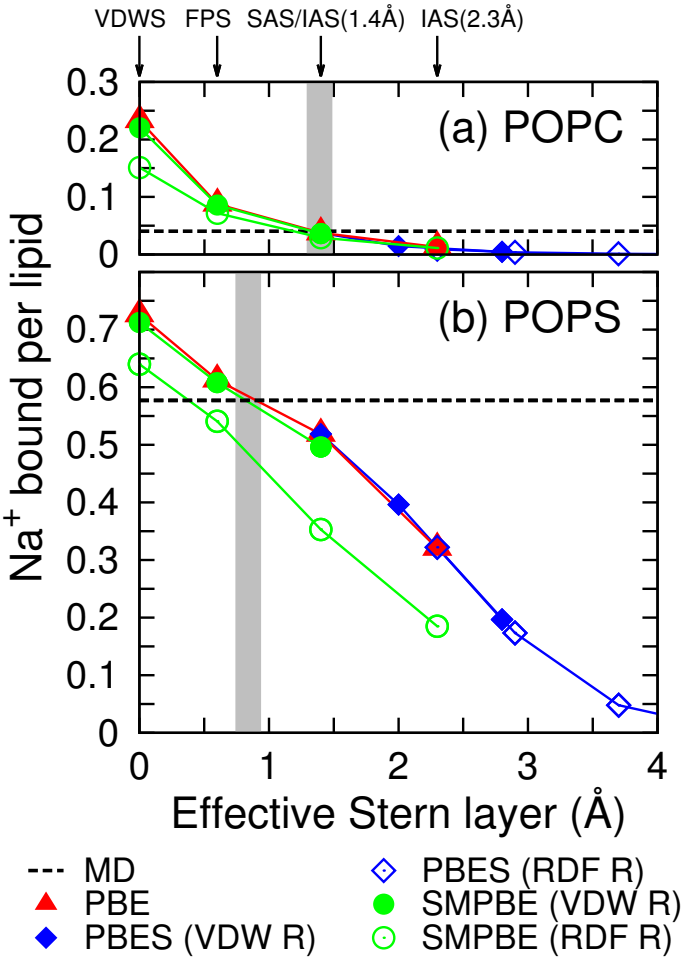

Figure 3.3: The number of Na^+ bound per lipid as a function of the "effective Stern layer thickness", the minimum distance between the VDWS and the "molecular surface" used for each calculation. (a) NaCl of bulk concentration 0.13M outside of the POPC bilayer, system 2 in Table 3.1. (b) NaCl of bulk concentration 0.17M outside of the POPS bilayer, system 7 in Table 3.1; VDW R and RDF R are shorthand for VDW radius and RDF radius. Gray shades mark where the mean-field calculations using the VDW radius agree with MD.

does not give the same performance for different ions. For example, for K^+, the IAS with VDW radius does not perform as well as it does for Na^+, i.e. the gray shade in Figure 3.8 is not exactly overlapping with the IAS + VDW radius data point like it does in Figure 3.3. The inconsistency of the performance of a parameter set is partly due to the fact that mean-field methods are much more simplified than the MD simulations and do not describe a number of ionic effects such as ion-ion correlations and fluctuations;[7]

they also ignore the ion-lipid specific interactions modeled in the MD force fields.[41,42]

The parameter inconsistency can be circumvented if empirical guidelines are available on how to choose parameters based on the characteristics of a system. For example, one can summarize a relationship between the surface charge density and the effective Stern layer thickness. Two data points in this relationship are generated here: at zero surface charge density, the VDW radius can be used as the size of the Stern layer; and at surface charge density $1.73e/nm^2$, roughly 0.8Å for Na^+ and 0.6Å for K^+ should be used as the size of the Stern layer. More data points need to be generated for a more complete knowledge of this relationship and the lipid-type specific ion absorption effects should be considered.

Last but not least, even though theoretically the PBE and the SMPBE become more distinguishable under a stronger electrostatic potential, the strong electrostatic potential exerted by the higher surface charge density of the POPS bilayer is not yet enough to significantly separate the PBE and the SMPBE results when the common VDW ion radii are used, Figure 3.3(b), putting the capability of the SMPBE under further doubt.

3.6 Conclusions

First, when it comes to studying ion equilibrium distributions, the SMPBE only offers a slight advantage over the PBE for the systems considered here. It is neither faster nor more accurate than the simple Stern layer ion-size correction of the PBE. There may, however, be other physical regimes where the SMPBE has a strong advantage. Further improvements are needed for the SMPBE to be an impactful correction to the PBE. Second, no single parameter set gives consistent mean-field method accuracy of prediction across systems of varying surface charge densities and ionic strengths.

Empirical rules may be able to be summarized to guide the choice of parameters of the mean-field calculations.

3.7 Acknowledgments

This work is supported by the National Institutes of Health (grants GM31749, GM096188), the National Science Foundation (grants MCB-1020765, DMS-0811259) and the NSF Center for Theoretical Biological Physics (grant PHY-0822283).

Chapter 3 is a minimally modified reprint of the material as it appears in Nuo Wang, Shenggao Zhou, Peter M. Kekenes-Huskey, Bo Li, and J. Andrew McCammon, "Poisson-Boltzmann versus Size-Modified Poisson-Boltzmann Electrostatics Applied to Lipid Bilayers," Journal of Physical Chemistry B, 2014. The dissertation author was the primary investigator and author of this paper.

3.8 Supporting Information

3.8.1 A Brief Description of the SMPBE Formulation

Assume that the system contains M ion species; their radii are r_1, ..., r_M and their "linear sizes" are a_1, ..., a_M. The volume of ion i is $v_i = a_i^3 = 4\pi r_i^3/3$. The ion concentrations at location $\mathbf{x}$, $\mathbf{x}$ is a vector, are $c_1(\mathbf{x})$, ..., $c_M(\mathbf{x})$. Similarly, the size, volume and concentration of the solvent molecule are denoted as r_0, a_0, v_0 and $c_0(\mathbf{x})$. With the above definitions, we have:

$$c_0(\mathbf{x}) = v_0^{-1}\left[\lambda - \sum_{i=1}^{M} v_i c_i(\mathbf{x})\right] \tag{3.1}$$

Where $\lambda = 0.64$ is the maximum packing fraction for randomly placed uni-sized spheres; it is used here as a reasonable approximation.

The way the ion sizes are incorporated into the PBE formulation is through the construction of the phenomenological lattice gas free energy of the system:

$$F(c_1,...,c_M) = \int \left\{ \frac{1}{2}\sum_{i=1}^{M} q_i c_i \psi + \frac{1}{2} Q\psi \right\} dV + \int \left\{ \frac{1}{\beta}\sum_{i=0}^{M} c_i[ln(v_i c_i) - 1] - \sum_{i=1}^{M} \mu_i c_i \right\} dV \tag{3.2}$$

Where q_i is the charge of the ith ion species, $\psi = \psi(\mathbf{x})$ is the local electrostatic potential, $Q = Q(\mathbf{x})$ is the local fixed-charge density from the biomolecule, $\beta^{-1} = k_B T$, k_B is the Boltzmann constant, T is the temperature and μ_i is the chemical potential for ion species i. Both integrations are over all space. The first term is the potential energy term and the second is the entropy term.

At equilibrium, the free energy should be minimized. If c_is are considered as the only independent variables in Equation (3.2), Equation (3.2) can be minimized with respect to each c_i constrained by the Poisson equation. After minimization, the following equation is obtained:

$$\frac{v_i}{v_0} ln(v_0 c_0) - ln(v_i c_i) = \beta(q_i \psi - \mu_i) \tag{3.3}$$

In the bulk phase, Equation (3.3) gives $\mu_i = \beta^{-1} ln[(v_i c_i^{\text{bulk}})/(v_0 c_0^{\text{bulk}})^{v_i/v_0}]$, where μ_i is a function of the bulk ion concentrations. Plugging μ_i back into Equation (3.3) for non-bulk regions gives:

$$c_i(\mathbf{x}) = \left[\frac{c_0(\mathbf{x})}{c_0^{\text{bulk}}} \right]^{\frac{v_i}{v_0}} \left[c_i^{\text{bulk}} e^{-\beta\psi} \right] \tag{3.4}$$

Plugging Equation (3.4) into Equation (3.1) will result in an equation containing c_0 as

the only unknown:

$$v_0 c_0(\mathbf{x}) + \sum_{i=1}^{M} \left\{ v_i \left[\frac{c_0(\mathbf{x})}{c_0^{\text{bulk}}} \right]^{\frac{v_i}{v_0}} \left[c_i^{\text{bulk}} e^{-\beta\psi} \right] \right\} - \lambda = 0 \tag{3.5}$$

where c_0, as the root of the equation, can be solved by numerical methods like the Newton's method. Once c_0 is obtained, all of the c_is can be calculated from Equation (3.4). Equation (3.4) is called the size-modified Boltzmann distribution as it contains the Boltzmann distribution term, $c_i^{\text{bulk}} e^{-\beta\psi}$, and a multiplicative correction term modeling the ion-size effects. The SMPBE is a PBE with the Boltzmann distribution term substituted by the size-modified Boltzmann distribution term in Equation (3.4).

3.8.2 The SMPBE Routein in APBS

For each SMPBE calculation, a 3-dimentional finite difference grid is created covering the biomolecule and its surrounding solvent region. The electrostatic potential $\psi(\mathbf{x})$ and ion concentrations $c_i(\mathbf{x})$, where $\mathbf{x}$ represents each grid point, are solved iteratively. The calculation starts with an initial guess of $\psi(\mathbf{x})$, which is uniformly zero. This initial guess is used in Equation (3.5) as the input to solve for $c_0(\mathbf{x})$, which gives the $c_i(\mathbf{x})$s through Equation (3.4). With the $c_i(\mathbf{x})$s, the potential is updated by solving the Poisson equation:

$$\nabla^2 \psi(\mathbf{x}) = Q(\mathbf{x}) + \sum_{i=1}^{M} q_i c_i(\mathbf{x}) \tag{3.6}$$

This process goes on iteratively until the maximum relative difference between the ψ obtained from two consecutive iteration steps is below an acceptable error. And the resulting ψ and c_i will be considered as the converged solution.

The SMPBE routine in APBS can be easily used by adding the following keywords into the APBS input file:

smpb1 step {step} eps {eps}

{step} is the step size of the Newton's method, which is used to solve the SMPBE. {step} is a number between 0 and 1. The smaller the {step}, the more numerically stable, but slower, the method and vice versa. {esp} is the value that the maximum relative difference of the electrostatic potentials among all grid points between two consecutive Newton iteration steps have to go below for the confirmation of convergence. Mathematically, this convergence criterion is given by $max\left(\left|\frac{\psi^k(\mathbf{x})-\psi^{k-1}(\mathbf{x})}{\psi^{k-1}(\mathbf{x})}\right|\right) < \{\text{esp}\}$, where $\psi^k(\mathbf{x})$ is the electrostatic potential at grid point $\mathbf{x}$ in step k. In this work, {step} is set to be 0.01 or smaller, {esp} is set to be 0.01.

3.8.3 The MD Simulation Settings

The equilibration for the MD simulations use the default steps generated by CHARMM-GUI Membrane Builder. The production simulation conditions are: NPT ensemble with constant temperature set at 300K by the Langevin thermostat with damping coefficient 1/ps and constant pressure set at 1atm by the Nose-Hoover Langevin piston barostat with oscillation period 50fs and oscillation decay time 25fs. The simulation time step is 2fs. All the bonds involving hydrogen atoms are fixed, non-bonded interactions are subjected to cut-off at 12Å. Periodic boundary conditions are applied on all sides of the simulation box. Particle Mesh Ewald (PME) method is used to calculate the electrostatic forces in the system.

The convergence of the MD simulations are determined by two criteria, the convergence of the area per lipid value and the convergence of the number of ions bound to the lipid bilayer surface. The number of ions bound is calculated by counting the total number of ions within $|z|$=25Å (z is defined in Figure 3.4).

3.8.4 APBS Input File Example

```
read
    mol pqr lipid.pqr
end
elec name lipidBilayer
    mg-auto
    dime 161 161 193
    cglen 100 100 120
    fglen 70 70 90
    cgcent 0 0 0
    fgcent 0 0 0
    mol 1
    npbe
    bcfl mdh
    pdie 2.00
    sdie 78.54
    srfm smol
    chgm spl4
    sdens 10.00
    swin 0.30
    temp 300
    calcenergy total
    write pot dx potentialFile
    write conc dx concentrationFile
    srad 1.4
    smpb1 step 0.01 eps 0.01
    ion charge 1 conc 0.13 radius 1.4
    ion charge -1 conc 0.13 radius 2.3
end
quit
```

3.8.5 Other Supporting Figures

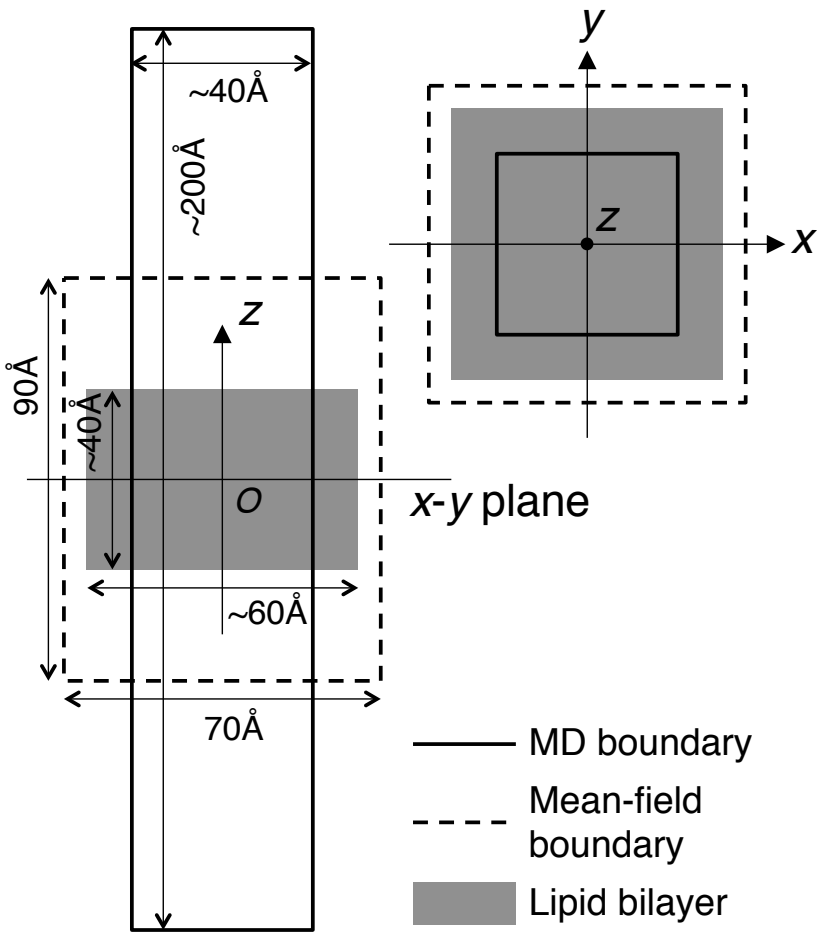

Figure 3.4: The dimensions of the lipid bilayer systems used in the MD simulations and the mean-field calculations. The x and y dimensions of the lipid bilayers are always equal.

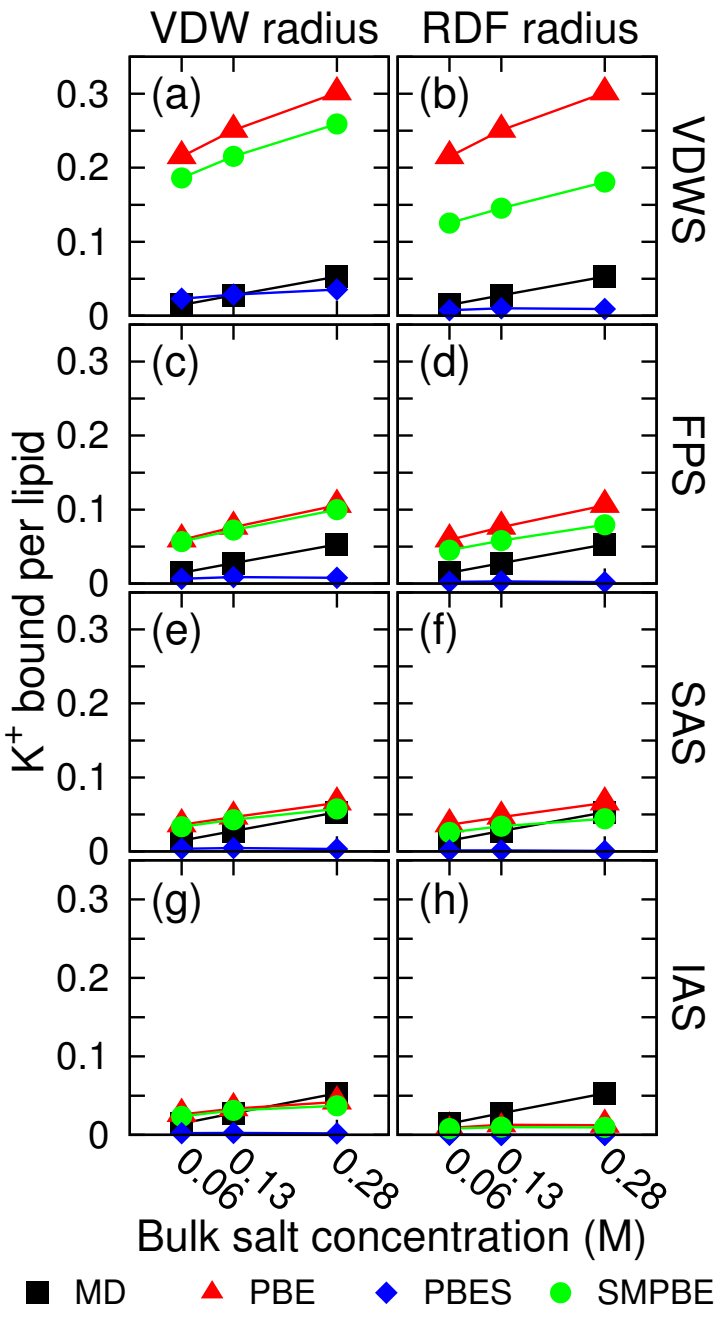

Figure 3.5: The number of K^+ bound per POPC molecule at three KCl bulk concentrations (systems 4 to 6 in Table 3.1). MD: molecular dynamics; PBE: nonlinear Poisson-Boltzmann equation (without Stern layer); PBES: PBE with Stern layer; SMPBE: size-modified Poisson-Boltzmann equation (without Stern layer). Subfigures (a) to (h) use eight different parameter sets; each is a combination of a molecular surface (VDWS, FPS, SAS, IAS, see Figure 3.1) and an ion radius set (VDW radius, RDF radius, see Table 3.2).

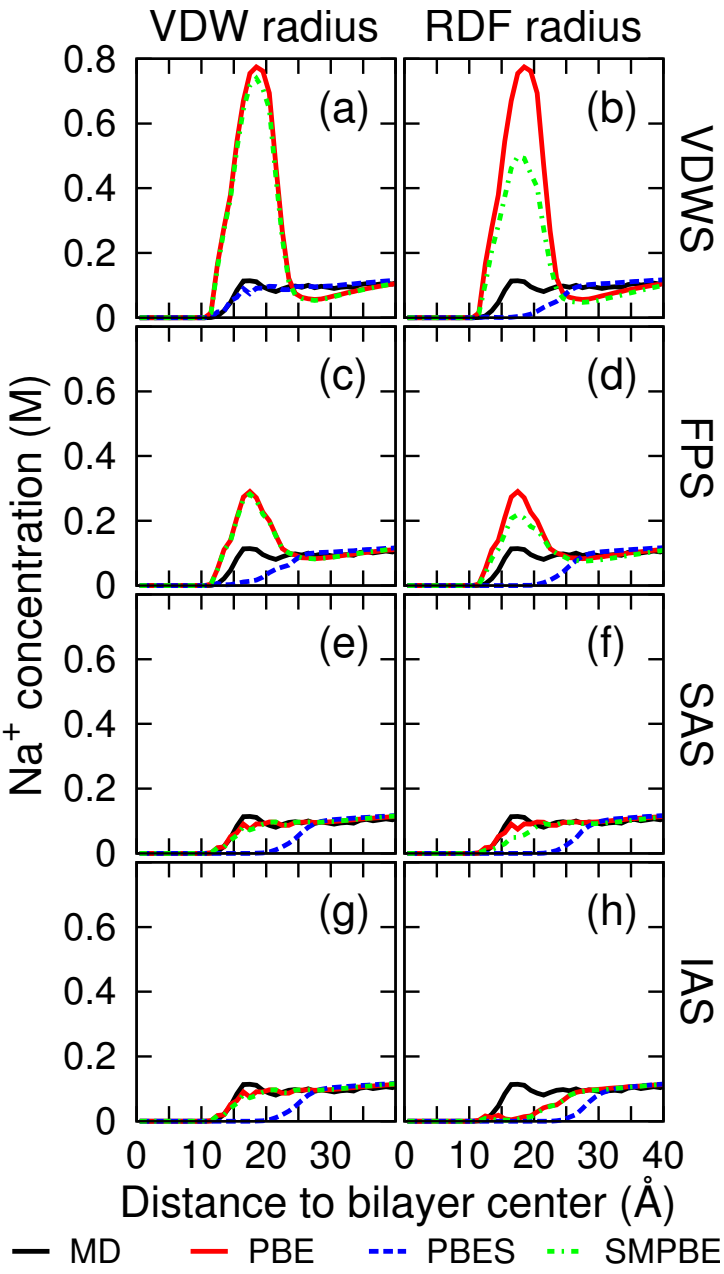

Figure 3.6: The Na^+ distribution along the perpendicular direction to a POPC lipid bilayer surface (system 2 Table 3.1). MD: molecular dynamics; PBE: nonlinear Poisson-Boltzmann equation (without Stern layer); PBES: PBE with Stern layer; SMPBE: size-modified Poisson-Boltzmann equation (without Stern layer). Subfigures (a) to (h) use eight different parameter sets; each is a combination of a molecular surface (VDWS, FPS, SAS, IAS, see Figure 3.1) and an ion radius set (VDW radius, RDF radius, see Table 3.2).

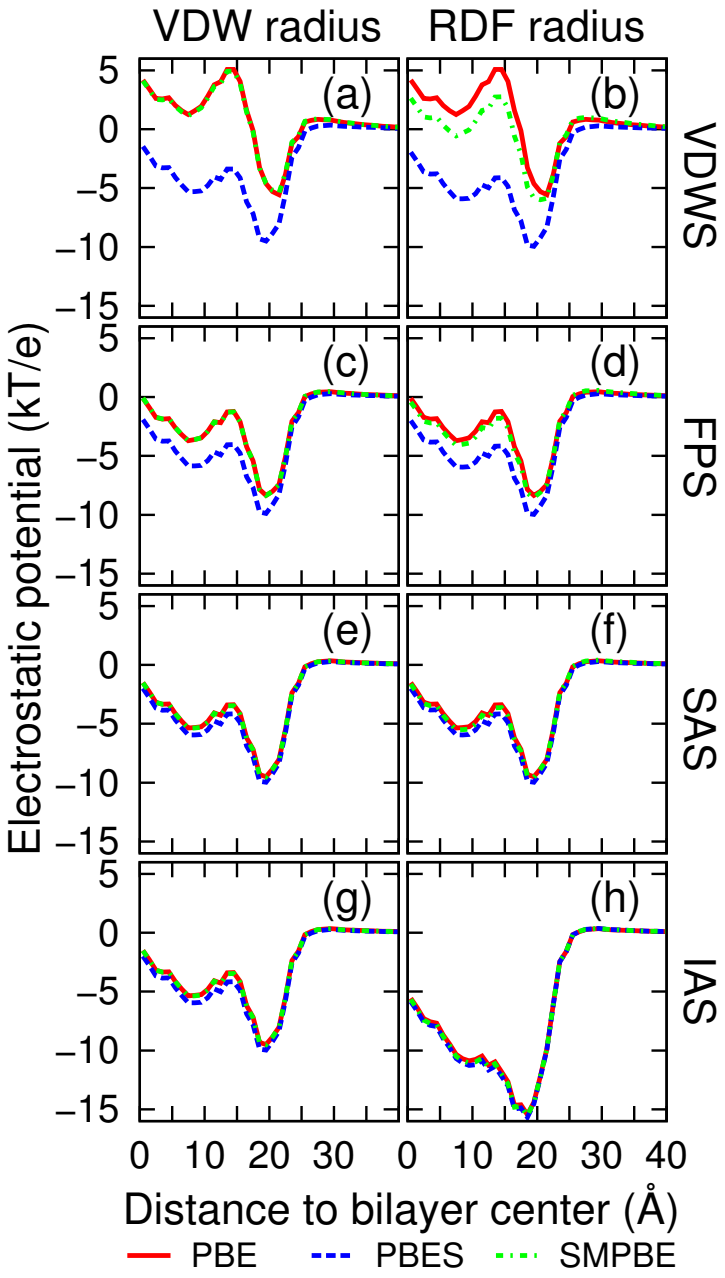

Figure 3.7: The electrostatic potential along the perpendicular direction to a POPC lipid bilayer surface (system 2 Table 3.1). PBE: nonlinear Poisson-Boltzmann equation (without Stern layer); PBES: PBE with Stern layer; SMPBE: size-modified Poisson-Boltzmann equation (without Stern layer). Subfigures (a) to (h) use eight different parameter sets; each is a combination of a molecular surface (VDWS, FPS, SAS, IAS, see Figure 3.1) and an ion radius set (VDW radius, RDF radius, see Table 3.2).

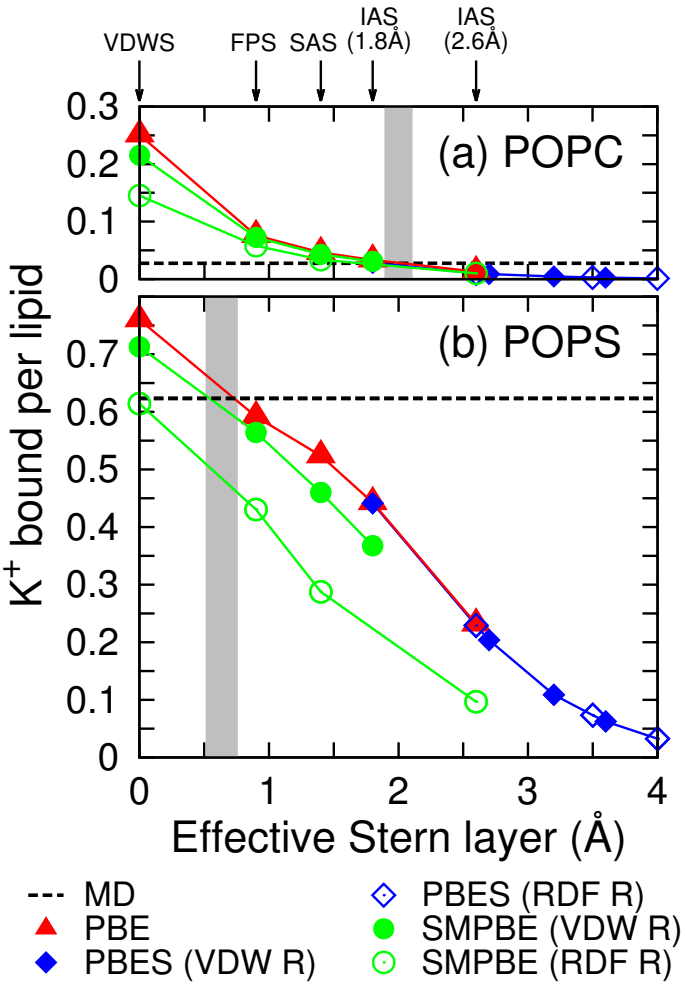

Figure 3.8: The number of K^+ bound per lipid as a function of the "effective Stern layer thickness", the minimum distance between the VDWS and the "molecular surface" used for each calculation. (a) KCl of bulk concentration 0.13M outside of the POPC bilayer, system 5 in Table 3.1. (b) KCl of bulk concentration 0.17M outside of the POPS bilayer, system 8 in Table 3.1. VDW R and RDF R are shorthand for VDW radius and RDF radius. Gray shades mark where the mean-field calculations using VDW radius agree with MD.

3.9 References

1. W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. van der Vegt, and H. B. Yu. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed.*, 45 (25):4064–4092, 2006.

2. J. A. McCammon, S. H. Northrup, and S. A. Allison. Diffusional dynamics of ligand-receptor association. *J. Chem. Phys.*, 90(17):3901–3905, 1986.

3. K. A. Sharp, R. Fine, and B. Honig. Computer simulations of the diffusion of a substrate to an active site of an enzyme. *Science*, 236(4807):1460–1463, 1987.

4. J. A. McCammon. Theory of biomolecular recognition. *Curr. Opin. Struct. Biol.*, 8 (2):245–249, 1998.

5. C. Eun, P. M. Kekenes-Huskey, V. T. Metzger, and J. A. McCammon. A model study of sequential enzyme reactions and electrostatic channeling. *J. Chem. Phys.*, 140: 105101, 2014.

6. J. P. Bardhan. Biomolecular electrostatics - i want your solvation (model). *Comput. Sci. Discovery.*, 5:013001, 2012.

7. P. Ren, J. Chun, D. G. Thomas, M. J. Schnieders, M. Marucho, J. Zhang, and N. A. Baker. Biomolecular electrostatics and solvation: a computational perspective. *Q. Rev. Biophys.*, 45(4):427–491, 2012.

8. M. E. Davis and J. A. McCammon. Solving the finite difference linearized Poisson-Boltzmann equation: A comparison of relaxation and conjugate gradient methods. *J. Comput. Chem.*, 10(3):386–391, 1989.

9. M. J. Holst and F. Saied. Multigrid solution of the Poisson-Boltzmann equation. *J. Comput. Chem.*, 14(1):105–113, 2004.

10. A. H. Boschitsch and M. O. Fenley. A fast and robust Poisson-Boltzmann solver based on adaptive cartesian grids. *J. Chem. Theory Comput.*, 7(5):1524–1540, 2011.

11. F. Dong, B. Olsen, and N. A. Baker. Computational methods for biomolecular electrostatics. *Methods Cell Biol.*, 84:843–870, 2008.

12. K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, 19:301–332, 1990.

13. G. J. Bartlett, C. T. Porter, N. Borkakoti, and J. M. Thornton. Analysis of catalytic residues in enzyme active sites. *J Mol. Biol.*, 324(1):105–121, 2002.

14. I. Borukhov, D. Andelman, and H. Orland. Steric effects in electrolytes: A modified Poisson-Boltzmann equation. *Phys. Rev. Lett.*, 79(3):435–438, 1997.

15. V. L. Shapovalov and G. Brezesinski. Breakdown of the gouy-chapman model for highly charged langmuir monolayers: Counterion size effect. *J. Phys. Chem. B.*, 110 (20):10032–10040, 2006.

16. O. Stern. The theory of the electrolytic double-layer. *Z. Elektrochem. Angew. Phys. Chem.*, 30:508–516, 1924.

17. V. Kralj-Iglič and A. Iglič. A simple statistical mechanical approach to the free energy of the electric double layer including the excluded volume effect. *J. Phys. II France*, 6:477–491, 1996.

18. V. Chu, Y. Bai, J. Lipfert, D. Herschlag, and S. Doniach. Evaluation of ion binding to dna duplexes using a size-modified Poisson-Boltzmann theory. *Biophys. J.*, 93(9): 3202–3209, 2007.

19. S. Zhou, Z. Wang, and B. Li. Mean-field description of ionic size effects with nonuniform ionic sizes: A numerical approach. *Phys. Rev. E.*, 84(2):021901, 2011.

20. S. Kirmizialtin, A. Silalahi, R. Elber, and M. Fenley. The ionic atmosphere around a-rna: Poisson-Boltzmann and molecular dynamics simulations. *Biophys. J.*, 102(4): 829–838, 2012.

21. B. Li, P. Liu, Z. Xu, and S. Zhou. Ionic size effects: Generalized boltzmann distributions, counterion stratification, and modified debye length. *Nonlinearity*, 26 (10):2899–2922, 2013.

22. F. Dong, M. Vijayakumar, and H. X. Zhou. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. *Biophys. J.*, 85(1):49–60, 2003.

23. X. Pang and H. X. Zhou. Poisson-Boltzmann calculations: van der Waals or molecular surface? *Commun. Comput. Phys.*, 13(1):1–12, 2013.

24. R. C. Harris, A. H. Boschitsch, and M. O. Fenley. Influence of grid spacing in Poisson-Boltzmann equation binding energy estimation. *J. Chem. Theory Comput.*, 9(8):3677–3685, 2013.

25. R. C. Harris, A. H. Boschitsch, and M. O. Fenley. Sensitivities to parameterization in the size-modified Poisson-Boltzmann equation. *J. Chem. Phys.*, 140:075102, 2014.

26. A. Spaar, C. Dammer, R. R. Gabdoulline, R. C. Wade, and V. Helms. Diffusional encounter of barnase and barstar. *Biophys. J.*, 90(6):1913–1924, 2006.

27. Y. Cheng, P. Kekenes-Huskey, J. E. Hake, M. J. Holst, and J. A. McCammon.

Michailova, a. p. multi-scale continuum modeling of biological processes: from molecular electro-diffusion to sub-cellular signaling transduction. *Comput. Sci. Discovery*, 5(1):015002, 2012.

28. K. Simons and D. Toomre. Lipid rafts and signal transduction. *Nat. Rev. Mol. Cell Biol.*, 1(1):31–39, 2000.

29. R. Phillips, T. Ursell, P. Wiggins, and P. Sens. Emerging roles for lipids in shaping membrane-protein function. *Nature*, 459:379–385, 2009.

30. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.*, 98(18):10037–10041, 2001.

31. S. Jo, J. B. Lim, J. B. Klauda, and W. Im. CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes. *2009*, 97(1):50–58, Biophys. J.

32. J. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comp. Chem.*, 26(16):1781–1802, 2005.

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

34. J. Klauda, R. Venable, J. Freites, J. OâĂŹConnor, D. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. Jr. MacKerell, and R. Pastor. Update of the CHARMM all-atom additive force field for lipids: Validation on six lipid types. *J. Phys. Chem. B.*, 114 (23):7830–7843, 2010.

35. T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker. Nucleic acids res. *2007*, 35:W522–W525, 2007.

36. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–4, 2008.

37. W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

38. F. M. Richards. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

39. R. A. Böckmann, A. Hac, T. Heimburg, and H. Grubmüller. Effect of sodium

chloride on a lipid bilayer. *Biophys. J.*, 85(3):1647–1655, 2003.

40. A. R. J. Silalahi, A. H. Boschitsch, R. C. Harris, and M. O. Fenley. Comparing the predictions of the nonlinear Poisson-Boltzmann equation and the ion size-modified Poisson-Boltzmann equation for a low-dielectric charged spherical cavity in an aqueous salt solution. *J. Chem. Theory Comput.*, 6(12):3631–3639, 2010.

41. S. A. Pandit, D. Bostick, and M. L. Berkowitz. Molecular dynamics simulation of a dipalmitoylphosphatidylcholine bilayer with NaCl. *Biophys. J.*, 84(6):3743–3750, 2003.

42. S. J. Lee, Y. Song, and N. A. Baker. Molecular dynamics simulations of asymmetric NaCl and KCl solutions separated by phosphatidylcholine bilayers: Potential drops and structural changes induced by strong Na^+-lipid interactions and finite size effects. *Biophys. J.*, 94(9):3565–357, 2008.

Chapter 4

Substrate Channeling Between the Human Dihydrofolate Reductase and Thymidylate Synthase

4.1 Abstract

In vivo, as an advanced catalytic strategy, transient non-covalently bound multi-enzyme complexes can be formed to facilitate the relay of substrates, i.e. substrate channeling, between sequential enzymatic reactions and to enhance the throughput of multi-step enzymatic pathways. The human thymidylate synthase and dihydrofolate reductase catalyze two consecutive reactions in the folate metabolism pathway, and experiments have shown that they are very likely to bind in the same multi-enzyme complex in vivo. While reports on the protozoa thymidylate synthase-dihydrofolate reductase bifunctional enzyme give substantial evidences of substrate channeling along a surface "electrostatic highway", attention has not been paid to whether the human thymidylate synthase and dihydrofolate reductase, if they are in contact with each other

in the multi-enzyme complex, are capable of substrate channeling employing surface electrostatics. This work utilizes protein-protein docking, electrostatics calculations and Brownian dynamics to explore the existence and mechanism of the substrate channeling between the human thymidylate synthase and dihydrofolate reductase. The results show that the bound human thymidylate synthase and dihydrofolate reductase are capable of substrate channeling and the formation of the surface "electrostatic highway". The substrate channeling efficiency between the two can be reasonably high and comparable to that of the protozoa.

4.2 Introduction

Metabolons, transient non-covalently bound multi-enzyme complexes, are thought to be important for the organization of enzymes and the regulation of their reactions in vivo. Metabolons are usually composed of enzymes that catalyze the sequential reactions in an enzymatic pathway. One of the advantages of metabolon formation is substrate channeling.[1] Substrate channeling is the direct transportation of a reaction intermediate from one enzyme active site to the next without prior release into the bulk solution. It can facilitate the enzymatic reactions in several ways. For example, it can reduce the lag time between two reactions and it can protect labile intermediates from the aqueous environment.[2] While the existence of metabolons has been controversial,[3] convincing evidence has been been established for substrate channeling.[2,4,5] Among covalently bound enzymes, channeling has been observed in the protozoa bifunctional dihydrofolate reductase-thymidylate synthase enzyme.[6]

Thymidylate synthase (TS) and dihydrofolate reductase (DHFR) are ubiquitous enzymes across organisms. TS is a homodimer; it catalyzes the reductive methylation of deoxyuridine monophosphate by $H_2C{\cdot}H_4$folate to deoxythymidine monophosphate and

H_2folate.[7] DHFR is a monomer and it catalyzes the reduction of H_2folate to H_4folate, replenishing the intra-cellular H_4folate pool.[8] The production and regulation of DHFR and TS are closely coupled to the cell cycle because they participate in the sole de novo synthesis pathway of thymidylate, a building block of DNA.[9,10] Insufficient DHFR and TS activity leads to thymidylate deficiency, halted cell cycle and eventual cell death.[11] Because of the crucial role that DHFR and TS play in living organisms, they have attracted much biological and biomedical interest. For example, they are drug targets in cancer treatments.[12] Species-specific inhibitors of DHFR and TS are also used to eliminate pathogens in the human body, such as the parasitic protozoa that cause severe epidemics like malaria.[13,14,15,16].

While the protozoa DHFR and TS are known to be able to channel their intermediate substrate H_2folate,[6] such possibility for the human counterparts is unknown. The protozoa DHFR and TS are fused bifunctional enzymes translated from a single gene. Each monomer of the TS homodimer is covalently attached to one DHFR monomer (supporting information Figure 4.4). The negatively charged H_2folate produced by the protozoa TS is channeled directly to DHFR through a sequence of positively charged surface residues connecting the two active sites, a so-called "electrostatic highway".[17,18,19] The human DHFR and TS are separate enzymes and each has its own gene.[20,21] However, it has been reported that the human DHFR and TS are distributed in the same cellular compartments[22,23] and may even participate in binding in the same metabolic multi-enzyme complex, e.g. the replitase complex.[24,25] Also, an experimental paper shows that, in vitro, one human TS dimer binds to up to six human DHFR monomers.[26] Further more, there exists example of surface electrostatic substrate channeling between non-covalently bound proteins in mammals, the porcine citrate synthase and malate dehydrogenase.[27,28,29] Given the evidence above, it is possible that some form of substrate channeling exists between the bound human DHFR and TS in vivo. In this work, the

simplest scenario of such possibility is explored computationally, where one human DHFR monomer binds to one human TS dimer.

Computational work on the protozoa DHFR-TS bifunctional enzyme substrate channeling has previously been conducted[30,31] where the substrate channeling efficiency under different ionic strengths and charge mutations are calculated by Brownian dynamics methods.[32,33,34] Here, a similar approach is taken, but since the human DHFR and TS are separate proteins, their possible bound-states are first predicted by protein-protein docking software.[35] The substrate channeling behaviors of these bound-state structures are then observed by electrostatics and Brownian dynamics calculations. It is found that there exist human DHFR-TS bound-states that are capable of channeling its substrate through surface electrostatics, and that their substrate channeling efficiencies can be as high as that of protozoa.

4.3 Results

4.3.1 Human DHFR-TS Binding Poses

The bound-states of the separated human DHFR and TS proteins are first predicted by the rigid-body protein-protein docking software ClusPro.[36,37] In the ClusPro run, the top 1000 lowest energy docking poses of DHFR-TS are grouped into 30 clusters and the lowest energy poses of each cluster form the final 30 docking poses (Figure 4.1). 22 of the 30 poses are on the top side of TS and 8 are on the bottom side. The "top side" is defined as the side through which the protozoa DHFR binds to the protozoa TS. Among the poses on the same side of TS, no two share similar orientations relative to TS.

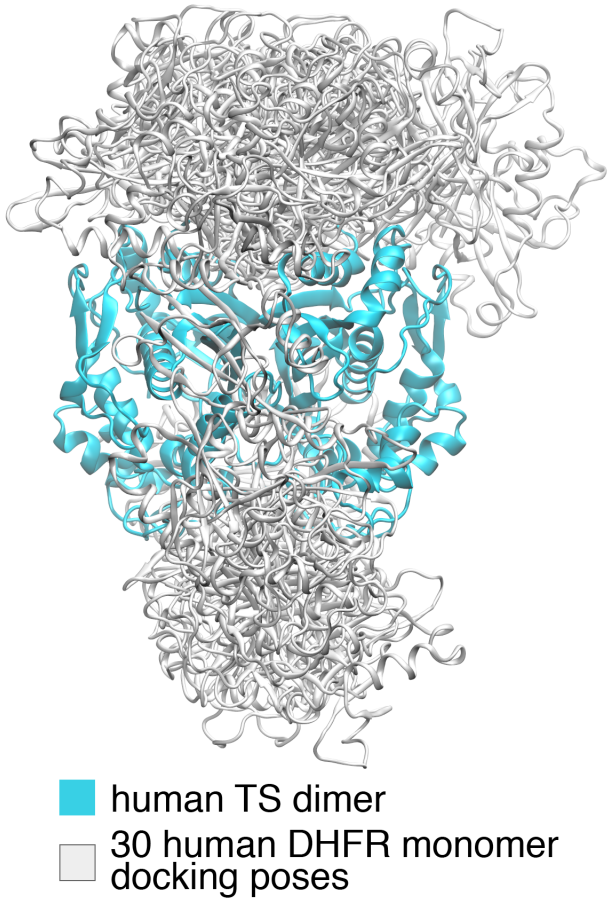

Figure 4.1: The 30 ClusPro docking poses. The cyan domain is the human TS dimer (PDB ID: 1HVY). The overlapping white domains are the ClusPro human DHFR (PDB ID: 1DHF) docking poses.

4.3.2 The "Electrostatic Highway"

The main substrate channeling mechanism of the protozoa DHFR-TSs is the so called "electrostatic highway".[19] For the discussions in this work, the "electrostatic highway" is defined as a continuous region of positive electrostatic potential on the molecular surface that connects the DHFR and TS active sites.

In each human DHFR-TS binding pose, the DHFR active site can take up substrates originated from either of the two TS active sites and so two different "electrostatic highways" may be formed. 30 ClusPro poses offer 60 DHFR-TS active site pairs. The

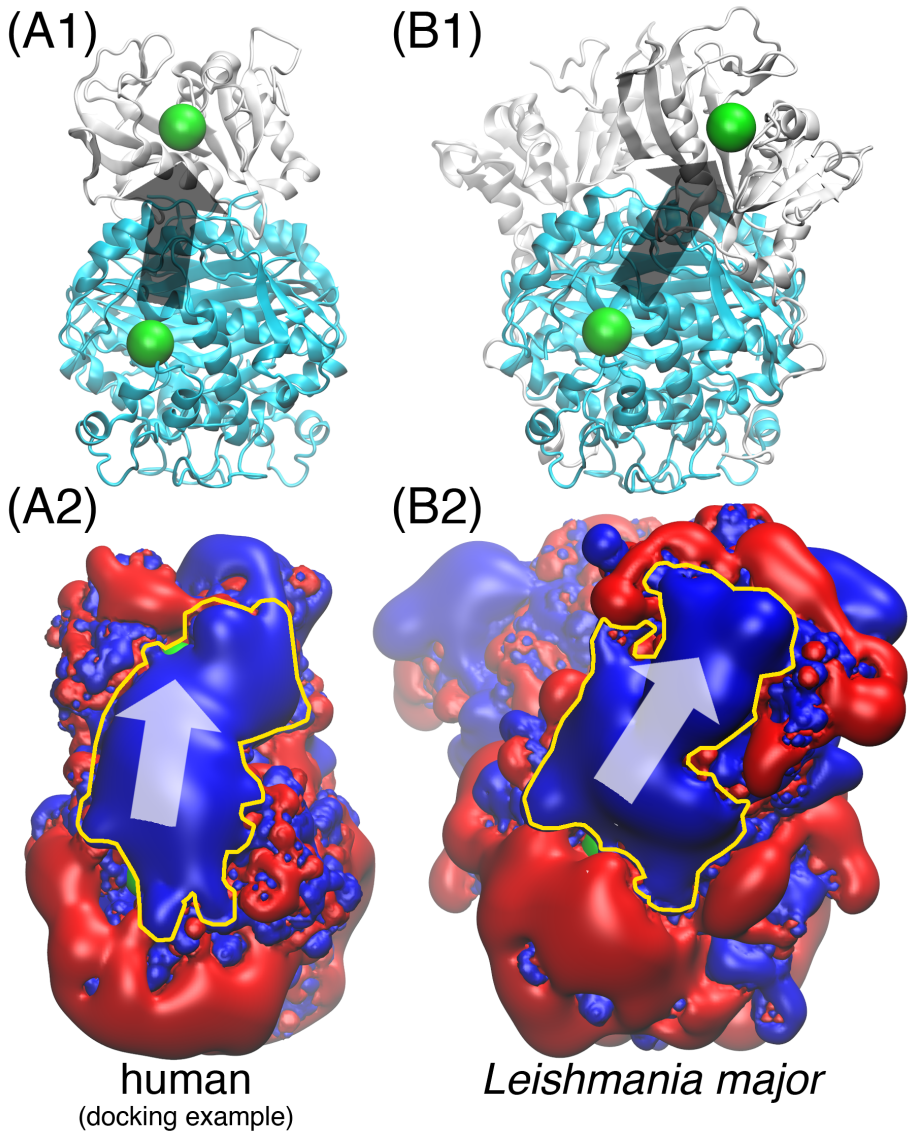

Figure 4.2: The human DHFR-TS "electrostatic highway". In (A1) and (B1), the cyan domain is the TS dimer; the white domain is DHFR (one monomer in (A1), two monomers in (B1)); the green spheres mark the location of the enzyme active sites. (A2) and (B2) display the electrostatics of the corresponding proteins in (A1) and (B1) with red and blue being the $-1kT/e$ and $1kT/e$ electrostatic potential iso-surfaces; The yellow contours highlight the "electrostatic highways". The arrows in all four subfigures indicate the direction of substrate channeling. All of the proteins are displayed 90 degrees sideways compared to the front view in Figure 4.1. The human DHFR and TS structures are the same as those used in Figure 4.1; the *Leishmania major* DHFR-TS structure is taken from the work by Knighton.[19] The docked human DHFR-TS poses are capable of forming "electrostatic highways" for substrate channeling in a similar way to the *Leishmania major* DHFR-TS, which is a well-demonstrated case.

"electrostatic highway" is observed in 17 out of the 60 pairs. And in 1 of the 30 poses, a "highway" is formed between the DHFR active site and both of the TS active sites. An example of the human DHFR-TS "electrostatic highway" is shown in Figure 4.2. It is compared to the well-studied *Leishmania major* "electrostatic highway".[19,30,31] Both "highways" not only connect the DHFR and TS active sites, but also cover the whole region between them completely. More illustrations of the "electrostatic highway" from the 17 DHFR-TS active site pairs can be found in the supporting information Figure 4.5. The electrostatic potentials in Figure 4.2 are calculated by software APBS.[38]

4.3.3 Channeling Efficiency

After observing the "electrostatic highway" in a number of human DHFR-TS docking poses, the "highways" effects on substrate channeling is studied. Substrate channeling is quantified through the substrate channeling efficiencies calculated by Brownian dynamics software BrownDye.[34] The substrate channeling efficiency is the percentage of substrates that started off at a TS active site and reach the DHFR active site by diffusion before escaping into the bulk solution. In the DHFR-TS system, the diffusion of the $-2e$ charged substrate is heavily influenced by the electrostatic potential field exerted by the proteins. The "electrostatic highway", which is a protein surface region with positive electrostatic potential, guides the diffusion of the substrate by opposite-charge attractions and achieves the observed substrate channeling.

Figure 4.3(A) shows that the channeling efficiency of the $-2e$ substrate is statistically significantly higher when there is an "electrostatic highway" connecting the DHFR and TS active sites. The channeling efficiency can be as large as 60%, comparable to the roughly 70% channeling efficiency of *Leishmania major*, which is known for its strong substrate channeling, calculated by BrownDye in previous work.[31] Figure 4.3(B) shows that if the substrate is neutral, the channeling efficiency distribution for the

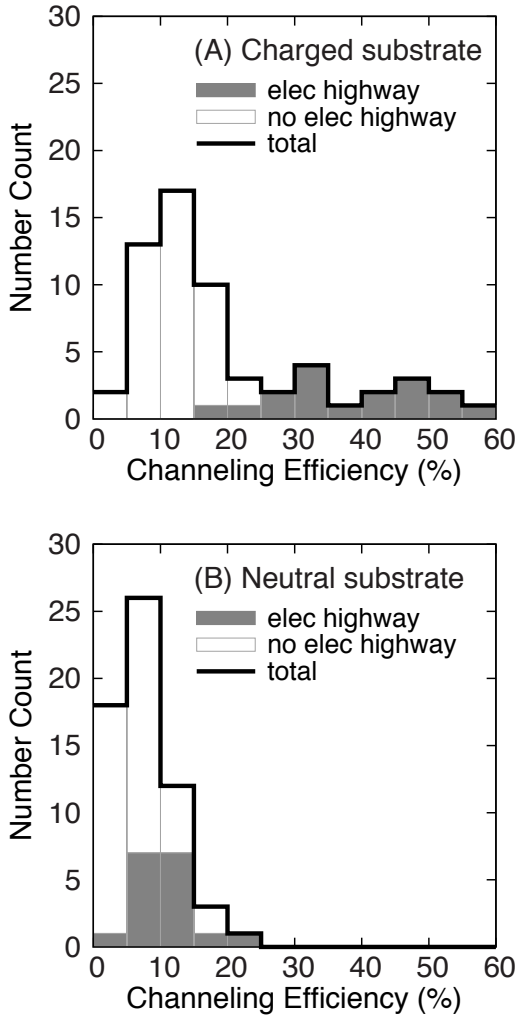

Figure 4.3: The influence of the "electrostatic highway" on the channeling of charged substrates. Among the 60 DHFR-TS active site pairs from the 30 ClusPro poses, "elec highway" (gray) represents the 17 pairs with observed "highways", and "no elec highway" (white) represents the other 43 pairs. This plot is a histogram with bin size of 5% and the first bin is (0%, 5%]. (A) For charged substrates, channeling is significantly enhanced by an "electrostatic highway". (B) When the substrate is neutral, the existence of an "electrostatic highway" does not have a noticeable impact on the substrate channeling efficiency.

DHFR-TS active site pairs connected by "electrostatic highways" (gray) overlaps with the distribution for the pairs that have no "highway" present (white). The "electrostatic highway" makes little contribution to the channeling of neutral substrates.

4.4 Discussion

The human DHFR-TS substrate channeling through "electrostatic highways" is studied with a simple model of one DHFR bound to one TS. The one-to-one binding between DHFR and TS is likely not the exact realistic case in vivo. First, multiple DHFRs may bind to one TS.[26] Second, other proteins may bind to DHFR and TS, e.g. those participating in the DNA synthesis .[24,25] Third, the human DHFR and TS are not guaranteed to bind to each other (although significant channeling may still occur since the two proteins are colocalized in the same cellular compartment[39]). This study does not aim to convince the readers that the electrostatic channeling observed here is the catalytic strategy that the human DHFR and TS use, but rather that the human DHFR-TS has the capability of electrostatic substrate channeling if their in vivo binding conformations allow.

Rigid-body protein-protein docking is employed here with the goal of searching for the possible formation of an "electrostatic highway" in the human DHFR-TS bound-states. It is only a minor goal of the docking calculations to predict the correct and stable binding poses between DHFR and TS. Without higher level calculations, such as molecular dynamics free energy calculations,[40] the ranking of the stability of a binding pose cannot be conclusive. Note that the binding pose for the native protozoa DHFR-TS (supporting information Figure 4.4(B)) is different from what has been predicted by docking (Figure 4.1). However, if the human DHFR is aligned to the *Leishmania major* binding pose by VMD MultiSeq,[41,42] an "electrostatic highway" is observed (supporting information Figure 4.6). A likely reason why ClusPro does not consider the protozoa binding pose for the human proteins is a lack of steric complementarity. *Leishmania major* DHFR and TS each has about 35 residues on the binding interface while if the human proteins are docked similarly, there are only about 10 interface residues from each

protein. This shows the limitation of rigid-body protein-protein docking because induced conformational changes could happen upon DHFR-TS binding.

The phrase "electrostatic highway" needs some clarification because, instead of literally speeding up the diffusion of the substrate, the "highway" serves as a "trap" to the oppositely charged substrate, limiting its escape into the bulk. In the Brownian dynamics calculations by BrownDye, the bulk extends to infinity and the substrate concentration in the bulk is zero, the chance of the substrate ever returning to DHFR quickly decreases to effectively zero as it diffuses away from the protein. Assuming the substrate reacts with the DHFR active site on contact, an "electrostatic highway" permanently "speeds up" the reaction by decreasing the number of escaped substrates. On the other hand, in small closed compartments, the "electrostatic highway" will only decrease the time it takes for the DHFR reaction to reach its full speed before the equilibrium of the bulk substrate concentration, and will not have a permanent speed-up of the DHFR reaction. Also, the "electrostatic highway" in the DHFR-TS system is itself not directional.[30] A directional electrostatic channeling can theoretically be achieved by a monotonically increasing surface density of the charged amino acids along the path between the two active sites, but this is not observed for DHFR-TS.

The enzyme specificity constants for the substrates for both of the human DHFR and TS are between 1 to 10 $\mu M^{-1}s^{-1}$.[43,44] However, the relative reaction rates of the two enzymes do not affect the results of the BrownDye simulations, because the BrownDye simulations here observe the channeling efficiency at the initial time point of the DHFR-TS reaction system way before it reaches equilibrium. To be more specific, at the initial moment, the substrates for TS are just added and the first few TS products are just produced, the concentrations of the products of TS in bulk are effectively zero. On the other hand, if one wants to simulate the concentration profiles of the DHFR and TS substrates and products over time, different reaction rates of the two enzymes will lead to

different behaviors of the system.

The BrownDye Brownian dynamics calculations in this work are set up specifically to study the effect of electrostatics on substrate channeling; the calculations consider mainly the electrostatic and the two-body hydrodynamic forces between the protein and the substrate. The protein and the substrate are also modeled as rigid bodies. Although electrostatics has been widely considered to be the cause of substrate channeling of the protozoa DHFR-TS,[18,19,30,31] the BrownDye results do not rule out the possibilities of other channeling mechanisms. In fact, in the protozoa system, experiments have brought up the possibilities for more complicated electrostatic channeling mechanisms, on top of the simple "electrostatic highway trapping", or additional non-electrostatic channeling mechanisms.[45,46] Comprehensive mutation experiments on the charged residues along the electrostatic highway of the human DHFR-TS are needed and it is hoped that the type of computational study in this work can be used to compare to those experiments to test whether and how mutations can alter the observed electrostatic restriction of substrate channeling and diffusion. Tests with molecular dynamics simulations that are capable of modeling the protein-ligand interactions more accurately and modeling the flexibility of proteins are also desirable.

In conclusion, with protein-protein docking, it is found that there exist bound-state conformations of the human DHFR and TS proteins where a continuous positive surface potential region, an "electrostatic highway", connecting the TS and DHFR active sites is formed. This "electrostatic highway" is formed in a similar way to what has been observed in the protozoa DHFR-TS, which are known for substrate channeling through surface electrostatics. Brownian dynamics simulations have further shown that with the "electrostatic highway" a significantly greater number of negatively charged substrates are passed from the TS to the DHFR binding site without escaping into the bulk. The human DHFR-TS have the capability of electrostatic substrate channeling if their in vivo

binding conformations allow. However, higher level computations are needed for more thorough analysis and more convincing evidence on the DHFR-TS electrostatic substrate channeling hypothesis.

4.5 Materials and Methods

4.5.1 Protein Structures

The human DHFR and TS structures are taken from the Protein Data Bank (PDB)[47] (PDB IDs 1DHF[20] and 1HVY[21]). Because 1DHF is missing ligand NADPH, the NADPH binding pose in PDB ID 2W3M[48] is aligned into the 1DHF structure by the VMD[41] tool MultiSeq.[42]

The protein PDB structures submitted to the online docking server are prepared by Maestro Protein Preparation Wizard,[49] and the protonation state of histidine residues is predicted by program PROPKA[50] embedded in the Protein Preparation Wizard. Because the docking pose PDB files output by the protein-protein docking software omit the hydrogens, program Reduce,[51] due to its batch processing capability, is used to add the hydrogens back into the docking poses before the following electrostatics and Brownian dynamics calculations.

4.5.2 Protein-Protein Docking

Online rigid protein-protein docking server ClusPro[36,37] is used to dock one human DHFR monomer onto one human TS dimer. ClusPro is chosen because it is the best-performing protein-protein docking software in CAPRI[52] rounds 22 to 27.[53,54] All of the docking jobs are submitted using the web server's default settings. The 30 ClusPro poses are scored using their "Balanced" scoring function.

The validity of ClusPro is tested by manually separating the known protozoa *Leishmania major* DHFR-TS structure (obtained from the authors of[19]) and docking the DHFR back onto the TS. ClusPro predicts the *Leishmania major* DHFR and TS native binding poses correctly. This result is included in the supporting information Figure 4.7.

4.5.3 Brownian Dynamics

Software BrownDye[34] is used to quantify the substrate channeling efficiency of the human DHFR-TS binding poses. To calculate the channeling efficiency between a TS active site and a DHFR active site, BrownDye is set up to run 10,000 Brownian dynamics[32] trajectories of substrate H_2folate starting at the TS active site diffusing under the influence of stochastic forces and the hydrodynamic and electrostatic forces between the substrate and the protein. The substrate channeling efficiency equals to the number of trajectories where the substrate reacted with the DHFR active site before escaping into the bulk over the total number of trajectories, 10,000. A Brownian dynamics trajectory is terminated on the first occasion of substrate reaction or escape. Substrate reaction is defined by the substrate diffusing into a spherical region of radius 12.5Å centered around the DHFR active site (Figure 4.8); Substrate escape is defined by the substrate leaving a large spherical boundary in the bulk solution centered around the protein (Figure 4.9). This "escape sphere" is significantly larger than the size of the protein and its radius is analytically determined.[55] More simulation details can be found in the supporting information Figure 4.8 and Figure 4.9.

4.5.4 Electrostatic Calculations

To calculate the electrostatic forces in BrownDye and to visualize the "electrostatic highway" along the DHFR-TS surface, the electrostatic potential field of the

DHFR-TS system is needed. The Adaptive Poisson-Boltzmann Solver (APBS) software is used.[38,56] APBS takes in the charges and atomic radii of a protein and outputs the electrostatic potential field generated by the protein. The protein charge and atomic radii used in APBS are also taken from the CHARMM27 force field[57] and the protein is solvated in 0.15M NaCl solution. An example of the APBS input file is attached in Section 4.7.2.

4.6 Acknowledgements

This work is supported by the National Institutes of Health (NIH) (grant GM31749), the National Science Foundation (NSF) (grant MCB-1020765), the National Biomedical Computation Resource (NBCR), the Howard Hughes Medical Institute (HHMI), and the NSF supercomputer centers. Nuo Wang would also like to thank Dr. Shenggao Zhou for the very helpful discussions on electrostatics calculations.

Chapter 4 is a minimally modified reprint of the material as it appears in Nuo Wang, and J. Andrew McCammon, "Substrate Channeling Between the Human Dihydrofolate Reductase and Thymidylate Synthase," Protein Science, 2015. The dissertation author was the primary investigator and author of this paper.

4.7 Supporting Information

4.7.1 Other Supporting Figures

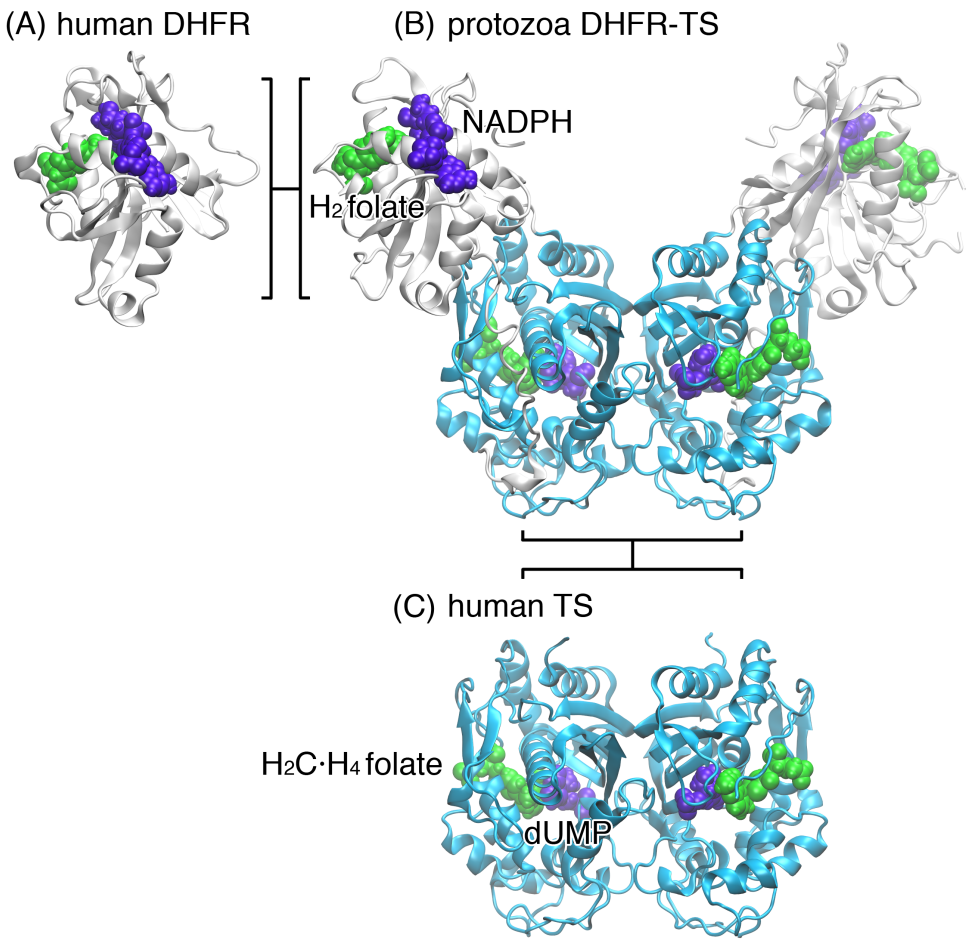

Figure 4.4: The three dimensional structures of the human and protozoa *Leishmania major* DHFR and TS enzymes. DHFR and TS are represented by the white and cyan cartoons respectively. The substrates are shown in the van der Waals representation in purple and green. The *Leishmania major* DHFR-TS,[19] (B), is a single covalently connected bifunctional enzyme; two *Leishmania major* DHFR-TSs form a homodimer joined through the TS domains. The human DHFR (PDB ID: 1DHF) and TS (PDB ID: 1HVY), (A) and (C), are individual enzymes; DHFR is a monomer, TS is a homodimer. The human DHFR and TS have the same overall tertiary structures compared to that of the protozoa.

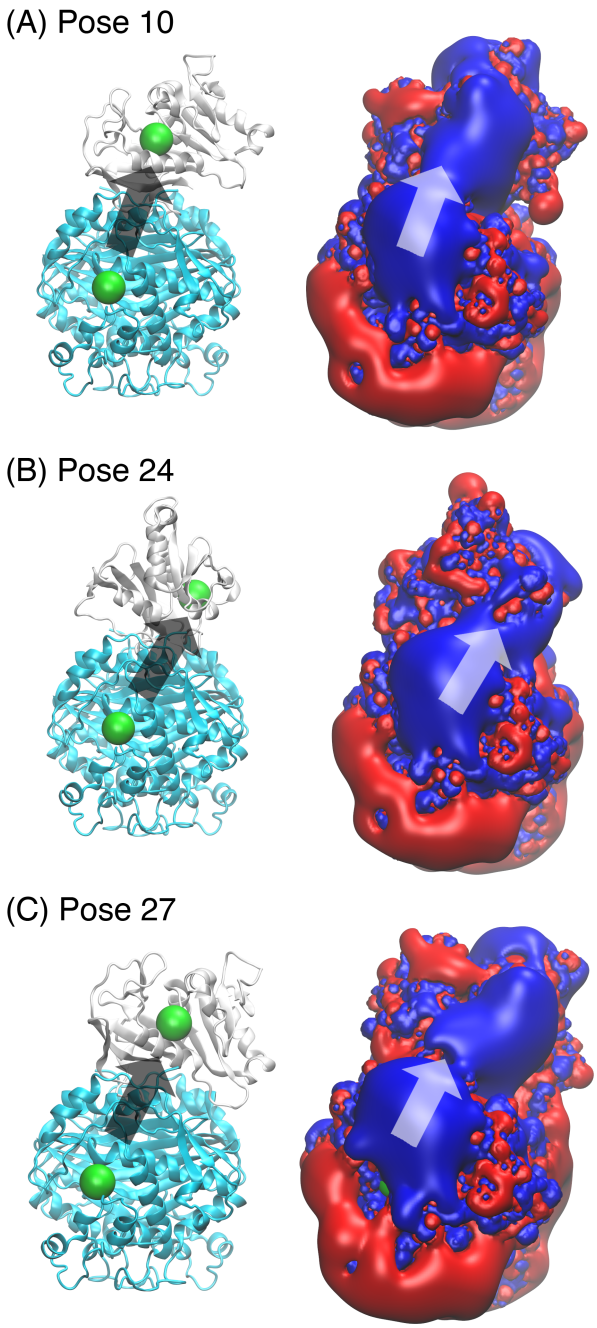

Figure 4.5: Three more examples of the "electrostatic highways" formed in the 30 ClusPro docking poses. All of the coloring and representations are the same as Figure 4.2, only that the yellow contours for the "electrostatic highways" are not drawn. Since it has not been mentioned, the pose shown in Figure 4.2 is the 9^{th} pose.

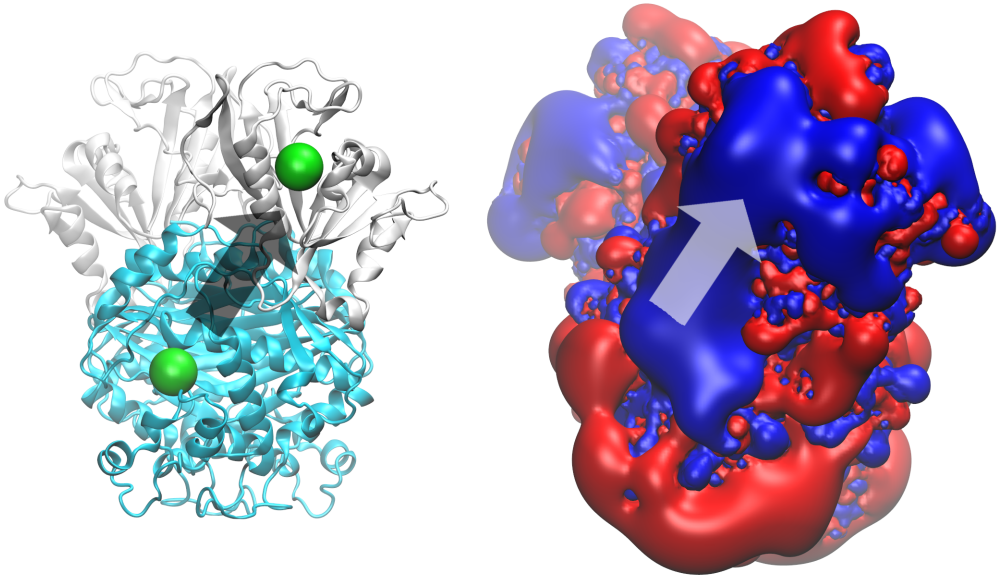

Figure 4.6: The bound human DHFR-TS is capable of forming an "electrostatic highway" in the *Leishmania major* binding conformation. All of the coloring and representations are the same as Figure 4.2. The human DHFR and TS are aligned onto the *Leishmania major* binding conformation, the resulting structure is shown on the left; the electrostatic potential of this structure is calculated and shown on the right.

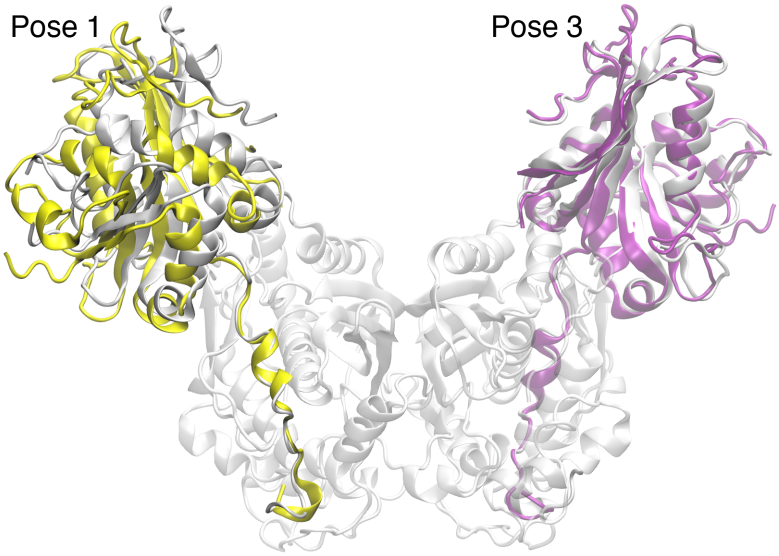

Figure 4.7: The partial proof of validity of ClusPro. The *Leishmania major* DHFR domain is manually separated from its TS and re-docked onto the *Leishmania major* TS. ClusPro gives 30 poses; the first (yellow) and third (pink) ranked docking poses correctly predict the two DHFR binding conformations on TS; the protein colored in white represents the native structure of the Leishmania major DHFR-TS. The RMSDs between the native and docking pose α-carbon atoms are less than 5Å.

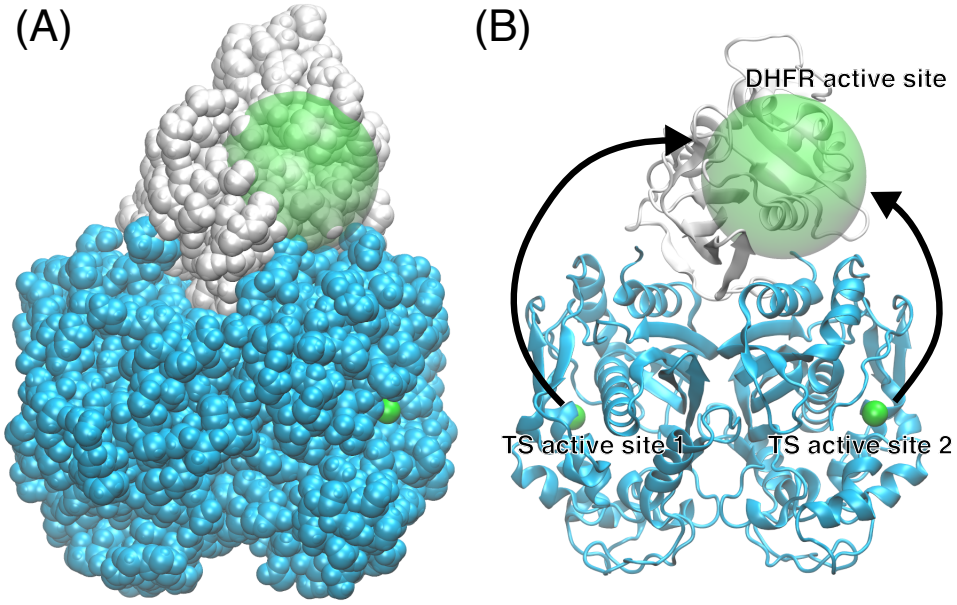

Figure 4.8: The BrownDye setup. The protein structure shown is one of the 30 ClusPro docking poses. (A) The van der Waals representation. (B) The cartoon representation. Both the substrate and the protein are considered as rigid bodies. The atomic structure of the protein is used, the charge and radius of the protein atoms are taken from CHARMM27 force field. Specifically, the atomic radii are the $r_{\min}/2$ values in the Lennard-Jones potential in CHARMM27. The substrate H_2folate is approximated by a 2Å sphere with -2e charge (solid green spheres). The starting position of the substrate, the "active site" of a TS monomer, is a point displaced by 2Å from the geometrical center of the H_2folate binding conformation in TS. The displacement is made along the line passing the geometrical center of that TS monomer and the geometrical center of H_2folate in the direction away from the TS center. The reaction of the substrate with the DHFR active site is defined as the substrate entering a spherical region of radius 12.5Å (transparent green sphere) centered around the "active site" of DHFR, which is a point defined in the same manner as the TS "active site". The escape of the substrate from the system is defined in Figure 4.9. Most of these settings are taken from the previous work[31] and the channeling efficiency of protozoa *Leishmania major* calculated using these settings is 73.6%, in agreement with the previous publication.[31] Note that for the docking poses, the channeling efficiency between each TS active site and the single DHFR active site is calculated. But for the native *Leishmania major* structure, due to symmetry, only one TS active site is chosen, and both of the DHFR active sites (there are two DHFRs, not one, in the native *Leishmania major* DHFR-TS) accept substrates.

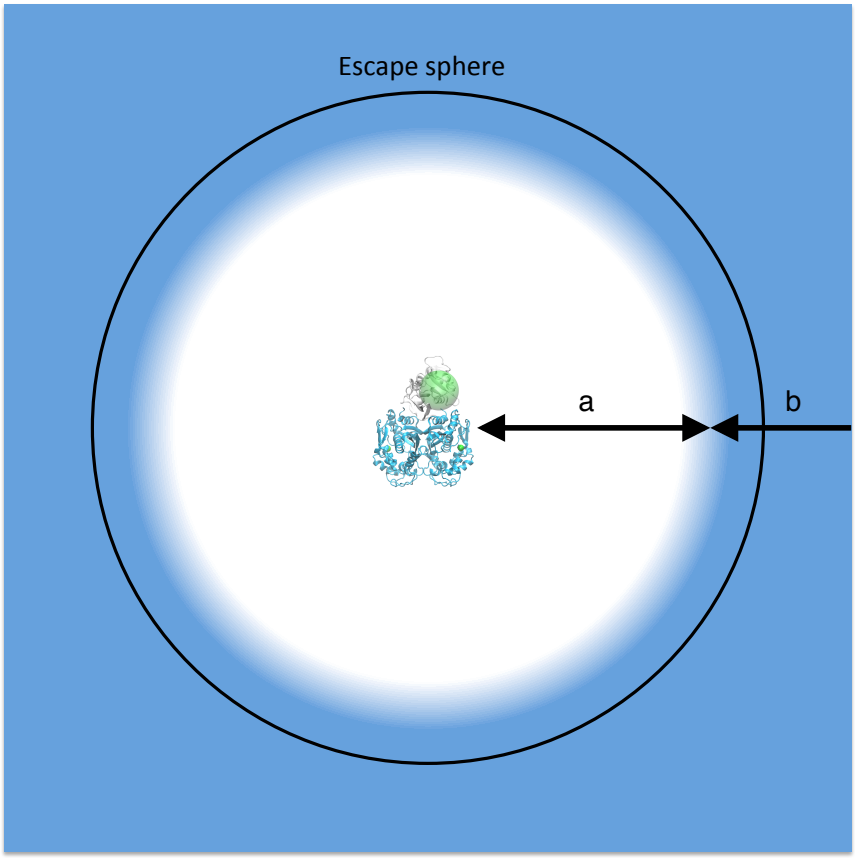

Figure 4.9: The definition of substrate escape. Figure 4.8(B) shows the path that the substrate takes in a successful reaction, however, the substrate always has a possibility to diffuse away from the protein and escape into the bulk solution. **a** (white) represents the immediate region around the protein where the electrostatic forces exerted by the protein on the substrate is dependent on the relative orientation between the two. **b** (blue) represents the space away from the protein where the electrostatic forces that the substrate experiences from the protein can be approximated as orientation-independent. The "escape sphere" is set in the **b** region. Once the substrate reaches the escape sphere, there is a probability that the substrate escapes forever, otherwise the substrate diffuses back to the **a** region. The mathematical expression of this probability has previously been described.[55] In our simulations, when the substrate reaches the escape sphere, it will escape (and the trajectory will be terminated) according to this probability.

4.7.2 APBS Input File Example

```
read
    mol pqr enzyme.pqr
end
elec
    mg-auto
    dime 385 385 385
    cglen 240 240 240
    fglen 190 190 190
    cgcent mol 1
    fgcent mol 1
    mol 1
    lpbe
    bcfl sdh
    pdie 2.00
    sdie 78.54
    srfm smol
    chgm spl2
    sdens 10.00
    srad 1.40
    swin 0.30
    temp 298.15
    ion charge 1 conc 0.15 radius 1.36
    ion charge -1 conc 0.15 radius 2.27
    calcenergy total
    calcforce no
    write pot dx enzyme-pot
end
quit
```

4.8 References

1. P. A. Srere. Complexes of sequential metabolic enzymes. *Annual Review of Biochemistry*, 56:89–124, 1987.

2. X. Huang, H. M. Holden, and F. M. Raushel. Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annual Review of Biochemistry*, 70:149–180, 2001.

3. M. Williamson. *How Proteins Work*, chapter 9.4 The Metabolon Concept, pages 338–344. Garland Science, 2011.

4. K. S. Anderson. Fundamental mechanisms of substrate channeling. *Methods in Enzymology*, 308:111–145, 1999.

5. E. W. Miles, S. Rhee, and D. R. Davies. The molecular basis of substrate channeling. *The Journal of Biological Chemistry*, 274:12193–12196, 1999.

6. K. M. Ivanetich and D. V. Santi. Bifunctional thymidylate synthase-dihydrofolate reductase in protozoa. *FASEB J.*, 4(6):1591–1597, 1990.

7. C. W. Carreras and D. V. Santi. The catalytic mechanism and structure of thymidylate synthase. *Annu. Rev. Biochem.*, 64:721–762, 1995.

8. J. R. Schnell, H. J. Dyson, and P. E. Wright. Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.*, 33:119–140, 2004.

9. L. F. Johnson. *Recombinant DNA And Cell Proliferation*, chapter 2 Expression of Dihydrofolate Reductase and Thymidylate Synthase Genes in Mammalian Cells, pages 25–48. Elsevier, 2012.

10. A. Zick, I. Onn, R. Bezalel, H. Margalit, and J. Shlomai. Assigning functions to genes: Identification of S-phase expressed genes in Leishmania major based on post-transcriptional control elements. *Nucleic Acids Research*, 33(13):4235–4242, 2005.

11. N. Sienkiewicz, S. Jarostawski, S. Wyllie, and A. H. Fairlamb. Chemical and genetic validation of dihydrofolate reductase-thymidylate synthase as a drug target in African trypanosomes. *Molecular Microbiology*, 69(2):520–533, 2008.

12. J. Walling. From methotrexate to pemetrexed and beyond. a review of the pharmacodynamic and clinical properties of antifolates. *Investigational New Drugs*, 24(1): 37–77, 2006.

13. I. H. Gilbert. Inhibitors of dihydrofolate reductase in leishmania and trypanosomes.

Biochimica et Biophysica Acta - Molecular Basis of Disease, 1587(2-3):249–257, 2002.

14. P. K. Rathod and M. A. Phillips. Prized malaria drug target nailed. *Nature Structural Biology*, 10:316–318, 2003.

15. A. C. Anderson. Targeting DHFR in parasitic protozoa. *Drug Discovery Today*, 10 (2):121–128, 2005.

16. Y. Yuthavonga, B. Tarnchompooa, T. Vilaivanb, P. Chitnumsuba, S. Kamchonwongpaisana, S. A. Charmanc, D. N. McLennanc, K. L. Whitec, L. Vivasd, E. Bongardd, C. Thongphanchanga, S. Taweechaia, J. Vanichtanankula, R. Rattanajaka, U. Arwona, P. Fantauzzie, J. Yuvaniyamaf, W. N. Charmanc, and D. Matthewse. Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target. *P.N.A.S.*, 109(42):16823–16828, 2012.

17. T. D. Meek, E. P. Garvey, and D. V. Santi. Purification and characterization of the bifunctional thymidylate synthetase-dihydrofolate reductase from methotrexate-resistant Leishmania tropica. *Biochemistry*, 24(3):678–686, 1985.

18. P.-H. Liang and K. S. Anderson. Substrate channeling and domain-domain interactions in bifunctional thymidylate synthase-dihydrofolate reductase. *Biochemistry*, 37 (35):12195–12205, 1998.

19. D. R. Knighton, C.-C. Kan, E. Howland, C. A. Janson, Z. Hostomska, K. M. Welsh, and D. A. Matthews. Structure of and kinetic channelling in bifunctional dihydrofolate reductase-thymidylate synthase. *Nature Structural Biology*, 1:186–194, 1994.

20. J. F. Davies, T. J. Delcamp, N. J. Prendergast, V. A. Ashford, J. H. Freisheim, and J. Kraut. Crystal structures of recombinant human dihydrofolate reductase complexed with folate and 5-deazafolate. *Biochemistry*, 29(40):9467–9479, 1990.

21. J. Phan, S. Koli, W. Minor, R. B. Dunlap, S. H. Berger, and L. Lebioda. Human thymidylate synthase is in the closed conformation when complexed with dUMP and raltitrexed, an antifolate drug. *Biochemistry*, 40(7):1897–1902, 2001.

22. A. S. Tibbetts and D. R. Appling. Compartmentalization of mammalian folate-mediated one-carbon metabolism. *Annual Review of Nutrition*, 30:57–81, 2010.

23. P. J. Stover and M. S. Field. Trafficking of intracellular folates. *Adv. Nutr.*, 2:325–331, 2011.

24. G. Prem veer Reddy and A. B. Pardee. Multienzyme complex for metabolic channeling in mammalian DNA replication. *P.N.A.S.*, 77(6):3312–3316, 1980.

25. S. Murthy and G. Prem veer Reddy. Replitase: Complete machinery for DNA synthesis. *Journal of Cellular Physiology*, 209(3):711–717, 2006.

26. A. Antosiewicz, E. Senkara, and J. Cieśla. Quartz crystal microbalance with dissipation and microscale thermophoresis as tools for investigation of protein complex formation between thymidylate synthesis cycle enzymes. *Biosens. Bioelectron.*, 15 (64):36–42, 2015.

27. A. H. Elcock and J. A. McCammon. Evidence for electrostatic channeling in a fusion protein of malate dehydrogenase and citrate synthase. *Biochemistry*, 35(39): 12652–12658, 1996.

28. K. Shatalin, S. Lebreton, M. Rault-Leonardon, C. Vélot, and P. A. Srere. Electrostatic channeling of oxaloacetate in a fusion protein of porcine citrate synthase and porcine mitochondrial malate dehydrogenase. *Biochemistry*, 38(3):881–889, 1999.

29. F. Wu and S. Minteer. Krebs cycle metabolon: Structural evidence of substrate channeling revealed by cross-linking and mass spectrometry. *Angew. Chem. Int. Ed.*, 54(6):1851–1854, 2014.

30. A. H. Elcock, M. J. Potter, D. A. Matthews, D. R. Knighton, and J. A. McCammon. Electrostatic channeling in the bifunctional enzyme dihydrofolate reductase-thymidylate synthase. *Journal of Molecular Biology*, 262(3):370–374, 1996.

31. V. T. Metzger, C. Eun, P. M. Kekenes-Huskey, G. A. Huber, and J. A. McCammon. Electrostatic channeling in P. falciparum DHFR-TS: Brownian dynamics and Smoluchowski modeling. *Biophysical J.*, 107(10):2394–2402, 2014.

32. D. L. Ermak and J. A. McCammon. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.*, 69:1352–1360, 1978.

33. J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Lutye, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, and J. A. McCammon. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics program. *Computer Physics Communications*, 91(1-3):57–95, 1995.

34. G. A. Huber and J. A. McCammon. Browndye: A software package for brownian dynamics. *Comput. Phys. Commun.*, 181(11):1896–1905, 2010.

35. I. A. Vakser. Protein-protein docking: From interaction to interactome. *Biophys. J.*, 107(8):1785–1793, 2014.

36. D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, 65(2):392–406, 2006.

37. D. Kozakov, D. Beglov, T. Bohnuud, S. E. Mottarella, B. Xia, D. R. Hall, and S. Vajda. How good is automated protein docking? *Proteins: Structure, Function, and Bioinformatics*, 81(12):2159–2166, 2013.

38. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *P.N.A.S.*, 98(18): 10037–10041, 2001.

39. C. C. Roberts and C. A. Chang. Modeling of enhanced catalysis in multienzyme nanostructures: Effect of molecular scaffolds, spatial organization, and concentration. *J. Chem. Theory Comput.*, 11(1):286–292, 2015.

40. M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9:646–652, 2002.

41. W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *J. Molec. Graphics*, 14(1):33–38, 1996.

42. E. Roberts, J. Eargle, D. Wright, and Z. Luthey-Schulten. MultiSeq: Unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, 7:382, 2006.

43. B. I. Schweitzer, S. Srimatkandada, H. Gritsman, R. Sheridan, R. Venkataraghavan, and J. R. Bertino. Probing the role of two hydrophobic active site residues in the human dihydrofolate reductase by site-directed mutagenesis. *J. Biol. Chem.*, 264 (34):20786–20795, 1989.

44. O. M. H. Salo-Ahen, A. Tochowicz, C. Pozzi, D. Cardinale, S. Ferrari, Y. Boum, S. Mangani, R. M. Stroud, P. Saxena, H. Myllykallio, M. P. Costi, G. Ponterini, and R. C. Wade. Hotspots in an obligate homodimeric anticancer target. structural and functional effects of interfacial mutations in human thymidylate synthase. *J. Med. Chem.*, 58(8):3572–3581, 2015.

45. C. E. Atreya, E. F. Johnson, J. Williamson, S.-Y. Chang, P.-H. Liang, and K. S. Anderson. Probing electrostatic channeling in protozoal bifunctional thymidylate synthase-dihydrofolate reductase using site-directed mutagenesis. *The Journal of Biological Chemistry*, 278:28901–28911, 2003.

46. H. Sharma, M. J. Landau, M. A. Vargo, K. A. Spasov, and K. S. Anderson. First three-dimensional structure of Toxoplasma gondii thymidylate synthase-dihydrofolate reductase: Insights for catalysis, interdomain interactions, and substrate channeling. *Biochemistry*, 52(41):7305–7317, 2013.

47. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1): 235–242, 2000.

48. A. K. W. Leung, L. J. Ross, S. Zywno-Van Ginkel, R. C. Reynolds, L. E. Seitz, V. Pathak, W. W. Barrow, E. L. White, W. J. Suling, J. R. Piper, and D. W. Borhani. Structural basis for selective inhibition of mycobacterium avium dihydrofolate reductase by a lipophilic antifolate, 2009. URL http://www.rcsb.org/pdb/explore.do?structureId=2w3m.

49. *Maestro, version 10.1*. Schrödinger, LLC, New York, NY, 2015. Schrödinger Release 2015-1.

50. M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen. PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. *J. Chem. Theory Comput.*, 7(2):525–537, 2011.

51. J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: Using hydrogen atom contacts in the choice of sidechain amide orientation. *J. Mol. Biol.*, 285(4):1735–1747, 1999.

52. J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. E. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2–9, 2003.

53. M. F. Lensink and S. J. Wodak. Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2082–2095, 2013.

54. J. P. G. L. M. Rodrigues and A. M. J. J. Bonvin. Integrative computational modeling of protein interactions. *FEBS Journal*, 281(8):1988–2003,, 2014.

55. B. A. Luty, J. A. McCammon, and H. X. Zhou. Diffusive reaction rates from brownian dynamics simulations: Replacing the outer cutoff surface by an analytical treatment. *J. Chem. Phys.*, 97:5682, 1992.

56. T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker. PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, 35:W522–W525, 2007.

57. N. Foloppe and A. D. MacKerell, Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comp. Chem.*, 21(2):86–04,, 2000.