Multiscale Modeling of Peptides and Star Polymeric Systems

A Dissertation Presented

by

Amber Carr

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

May 2013

Stony Brook University
The Graduate School

Amber Carr

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Carlos L. Simmerling – Dissertation Advisor
Professor, Department of Chemistry

Fernando O. Raineri – Chairperson of Defense
Lecturer, Department of Chemistry

Daniel P. Raleigh – Committee Member
Professor, Department of Chemistry

Hans W. Horn – Outside Member
Research Staff Member, IBM Almaden Research Center

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Multiscale Modeling of Peptides and Star Polymeric Systems

by

Amber Carr

Doctor of Philosophy

in

Chemistry

Stony Brook University

2013

In the field of structural biology, the synergy between theoretical and experimental approaches has lead to great advances in our understanding of the structural ensembles and dynamic behavior of biomolecules. Despite the successes of biomolecular simulation within the context of these goals, there remain limitations in the ability of this methodology to accurately model protein structure and dynamics, and to solve problems efficiently in terms of the time required by the work and the computational resources required. Biomolecules such as proteins exist on a rugged free energy landscape, which limits the extent of conformational space that can be visited in a single simulation. The development of enhanced sampling techniques employs manipulation of the formulation of the system's energy function and equations of motion in order to efficiently sample slow events such as large conformational changes and rare events such as transitions between two states, and to increase the volume of conformational space that is available to the system under study.

This work outlines two methods in the computational study of protein folding which aim to enhance conformational sampling while reducing thecomputational demands of the simulation. One common strategy of enhancing conformational sampling that has been incorporated into many simulation algorithms is to periodically afford the simulated molecule the opportunity to escape from energy minima and to thereby sample a much larger volume of phase space than by conventional methods. In the self-guided Langevin (SGLD) formalism, low-frequency modes of motion of the protein are enhanced in order to allow the protein to cross potential energy barriers. We have applied SGLD to three model peptides in implicit solvent in order to examine the effect of the method's two adjustable parameters on the peptides' resulting structural ensembles and folding rates. The model systems are of similar sizes but differing topologies, which allows for examination of transferral of parameters between systems.

Another strategy of enhancing sampling is an extension of parallel-tempering Monte Carlo to molecular dynamics. In this method, known as replica-exchange molecular dynamics (REMD), periodic attempts are made to exchange structures that are simulated at different temperatures, and a random walk in temperature space is achieved in order to surmount

conformational barriers in the energy landscape. Variants of this technique have been developed over the years in order to increase the efficiency of REMD simulations of biomolecules. In particular, approaches have been developed in which a structural reservoir is used to decouple the high-temperature search for structures from the exchanges and annealing which occur at lower temperatures. It has been shown that the contents of this reservoir need not comprise a Boltzmann-weighted ensemble; any ensemble of structures may be used as long as its probability distribution is known. Expanding on this method, we have developed an algorithm to further enhance the efficiency of reservoir REMD through the inclusion of a weight factor that relates the relative probabilities of the highest-temperature replica structure and the structure in the reservoir under exchange. In this work, we outline attempts to apply this method to the model system alanine dipeptide, and discuss the results obtained using a coarse-grained model that considers only the potential energy of the dipeptide as a function of its dihedral angles and does not consider its atomistic degrees of freedom.

Finally, the application of simulation methodology to a non-biological self-assembling polymeric system on the nanoscale is demonstrated in this work, and its potential application to the field of targeted drug delivery is discussed. Diblock star copolymers are self-assembling nanoscale systems that have shown great potential in the field of targeted drug delivery in the human body. Intriguingly, these star polymer systems bear many important similarities in structure and composition to proteins, being composed of linear polymeric chains of repeating units which self-assemble with hydrophobicity as the driving force. These similarities allow for the application of many of the techniques of molecular modeling and simulation developed for proteins to these systems. At present, experimental imaging of star diblock copolymers and nanogel star copolymers, particularly in complex with drug molecules, has been limited, providing computational studies with the opportunity to predict the structures of these molecules in atomic detail, as well as their dynamic behavior. In this work, we describe a comparative study of three star block copolymer systems with varying hydrophobicity in their core regions. The goal of this work is to provide atomic-level information on star polymer structure and dynamic behavior, including the size and shape of the polymer, the details of its bonding patterns, and its potential for aggregation. Additionally, the kinetics of drug uptake and delivery, as well as the degradation profile of the delivery material, may also be examined. Because theoretical methods, in contrast to experiment, are often less expensive and more time-efficient, their systematic application may offer strategies at the molecular level by which to modify formulations of drug and polymer for optimal compatibility and delivery efficiency.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This work would not have been possible without the continuous support of so many people. I thank my advisor, Professor Carlos Simmerling, for giving me the freedom to explore ideas, and to make the mistakes which allowed me to truly test the limits of my knowledge and grow as a scientist. His boundless enthusiasm for our work kept me going through setbacks, and set an example of the perseverance necessary to see a project through to fruition. My dissertation committee was of great assistance to me over the years in providing feedback on my work, and I sincerely thank them for their time. Professor Daniel Raleigh provided many helpful suggestions and asked difficult questions along the way which helped me to improve my work and presentations. Professor Fernando Raineri has been an inspirational teacher and mentor for the past ten years; his Physical Chemistry course set me on the path to obtaining this degree, and through emulating his example of kindness and generosity, I learned to teach and mentor others in turn. Dr. Hans Horn provided me with very helpful feedback on the dissertation, and his friendship and sense of humor were a great support to me in the weeks leading up to my defense.

Working with the Computational Chemistry group at the IBM Almaden Research Center was the opportunity of a lifetime. It was truly an honor to be advised at IBM by Dr. Bill Swope, and I cannot thank him enough for his infinite patience, humor, kindness, and enthusiasm in all things, especially in sharing his excellent wine. I thank Dr. Julia Rice for being an incredible role model for women in science, and Dr. Jed Pitera for his many helpful suggestions on the polymer project. Dr. Gavin Jones and Amanda Parker showed me the ropes when I arrived at IBM, and made every day (and every ride on the Light Rail) a lot of fun.

The members of the Simmerling lab, past and present, have been my second family over the years. I thank Dr. Asim Okur and Dr. Daniel Roe for initiating very interesting projects before leaving the lab, aspects of which haunt my nightmares to this day. I still appreciate the friendliness shown to me by Dr. Salma Rafi, Melinda Layten, and Dr. Kun Song when I first entered the lab. Dr. Christina Bergonzo, Dr. Lin Fu, Kevin Hauser, Agnes Huang, Koushik Kasavajhala, Colleen Kirkup, Haoquan Li, Eric Cheng-Tsung Lai, James Maier, Carmenza Martinez, Hai Nguyen, Dr. Sally Pias, and Dr. Miranda Yi Shang have all helped me to solve problems in both work and life. It was a pleasure working with all of you, and it is an honor to have you as my friends. I especially thank Dr. AJ Campbell, Dr. Fangyu Ding, and Dr. Lauren Wickstrom for the encouragement and support that they have shown me through the years. Thank you for being the most wonderful friends!

It is difficult to leave Stony Brook at a moment when the members of our field have come together to form such a vibrant community. The current members of the Laufer Center for Physical and Quantitative Biology, as well as the members of the previous Joint Computational Biology Group, have all played an important role in the development of my work and exposure

to new ideas in the field. I thank the members of the Dill Group for their friendship and support over the past year; it has truly been a pleasure getting to know each of you. I thank Professor Jin Wang for introducing me to protein folding, and I thank his past group members Dr. Saul Lapidus, Dr. David Lepzelter, and Dr. Qiang Lu for their friendship and scientific input. Professor Rob Rizzo and Professor David Green have both been very helpful and supportive over the years, and I thank the members of their groups for their friendship.

I have been incredibly fortunate to have had many amazing teachers and mentors over the years at Stony Brook. In the Department of Chemistry, Professor George Stell advised me in my first research experience, which led to my M.S. degree and set me on a lifelong path of worrying about phase transitions and critical phenomena. Professor Bob Schnieder served on the committee for my M.S., and has been very supportive throughout my doctoral work. I thank Professor Clare Grey for her outstanding teaching and mentorship, and Dr. Mohammad Akhtar for giving me my first opportunity to work as a Teaching Assistant. In the Department of Physics, Professor Emilio Mendez first taught me how to think like a scientist, and his excellent teaching and kind nature remain an inspiration a decade later. At the Center for Science and Mathematics Education, I thank Professor Linda Padwa and Professor Keith Sheppard for their mentorship and for providing me with many opportunities to improve my teaching. I am greatly indebted to the Department of Chemistry for providing me with the opportunity to pursue a doctorate, and particular thanks are due to Katherine Hughes, Heidi Ciolfi, and Norma Reyes for handling my many registration emergencies. At the Laufer Center, I thank Eileen Dowd and Nancy Rohring for their help and support.

I thank my many friends who, although now scattered across the globe, are never further away than an email or phone call. I thank my family for always encouraging me to continue my education and for giving me the freedom to pursue my dreams. I thank Nathan for giving meaning to it all.

“Could the search for ultimate truth really have revealed so hideous and visceral-looking an object?”

-- Perutz MF (1964) The Hemoglobin Molecule. *Scientific American* 211(5):64-76.

1. Introduction

1.1 Challenges in the Study of Self-Assembling Biological Systems

A defining characteristic of many complex systems is that of emergence: the existence of collective principles of organization such that the system appears to develop novel behavior that is not necessarily predictable from the laws governing its microscopic constituents [1,2,3,4]. Emergent phenomena are prevalent in biology, from the self-assembly of proteins from amino acid precursors to the co-evolution of neighboring ecosystems. In many cases, the constituents of these systems and their behavior on the microscopic level have been firmly established, but exactly how these constituents cooperate to yield emergent phenomena is not known. Much work thus remains to be done in establishing not only the fundamental structures which constitute these systems, but also the organizational principles which guide their behavior.

One of the major breakthroughs of science in the twentieth century was the elucidation of the structure and function of the molecules that carry out the processes of life, including DNA, RNA, and proteins [5,6,7,8,9,10]. Proteins are major players in a diverse array of biological processes, and as such, are of great diversity of function: they catalyze reactions, maintain the structural integrity of cells, permit cell motility, and function as signals and receptors. One defining feature of proteins is their interrelation of structure and function. The correct folding of a protein into a single, specific three-dimensional conformation is vital to its proper function, and misfolded and aggregated proteins are believed to be major factors in the development of disorders such as Alzheimer’s, Parkinson’s, and Huntington’s diseases [11,12,13]. Much current research in the field of structural biology thus aims to formulate hypotheses concerning the folding of proteins, from the relationship of the amino acid sequence to its resultant structure, to the robustness of specific sequences and conformations to perturbations such as genetic mutation.

Much of the development of protein folding theory and methodology has been based upon work that was performed earlier in the 20th century on non-biological polymeric systems [14,15,16,17]. Proteins are nanoscale polymeric structures that are formed through the spontaneous and concerted self-assembly of a set of 20 amino acid monomeric units, or residues. Although this set of amino acids and their resultant chemical interactions are, broadly speaking, more complex than those of typical non-biological polymeric systems, there remains enough similarity between the two systems that the insights and tools that are gained from work in one field may be applied to the other. As will be described below, many of the problems, goals, advantages, and limitations of computational studies of both protein and polymeric systems are shared.

For example, structure and activity in biological and self-assembling polymeric systems span a wide range of lengths and times, ranging from the nanoscale to the macroscale [18]. How, then, do we begin to examine these processes in proteins and polymers, and to unravel the relationship between their microscale physical laws, mesoscale structure, and macroscale dynamics and activity in order to uncover the layers of complexity of the system and their interrelation? How do experimentalists treat both the atomic-scale and ensemble behavior of these structures, and how do theorists create models that accurately address the levels of complexity that these systems entail?

The goal of this work is to outline two new methods in the computational study of protein folding which aim to provide solutions to a current problem in the field, that of efficiently observing the accessible conformations of a protein during the timescale of a simulation. Additionally, an application of simulation methodology to a non-biological self-assembling polymeric system on the nanoscale is demonstrated, and its potential application to the field of targeted drug delivery is discussed.

1.2 Protein Structure

In a living cell, proteins play the role of messengers, sentinels, motors, and catalysts. The ability of proteins to execute such a diverse range of tasks is due to the high specificity that they possess for the molecules with which they interact. This specificity is based in large part on the three-dimensional structure of the protein, which dictates its function as well as its chemical propensity to interact with other biomolecules. Understanding protein structure is therefore essential to discerning its biological function. As elucidated by Fischer in the early 1900s, proteins are polymers of 20 amino acid subunits, or residues [19]. The number of residues in a protein chain ranges from tens to thousands; the particular sequence of amino acids that constitutes each protein chain is encoded by the genome, and is unique for each protein, as discovered by Sanger in the early 1950s [20,21].

Proteins are described as having four levels of structure: primary, secondary, tertiary, and quaternary. The linear sequence of amino acids is known as the protein primary structure. Amino acids have a backbone, which consists of a central α-carbon linking an amide group and a carbonyl group. Connected to each α-carbon is the chemical moiety known as the sidechain, which gives each amino acid its unique chemical identity. Amino acids are connected by a peptide bond, which covalently links the amino group to the carbonyl group (Figure 1.1).

Although the peptide bond is rigid, rotations around the α-carbon allow for conformational flexibility.

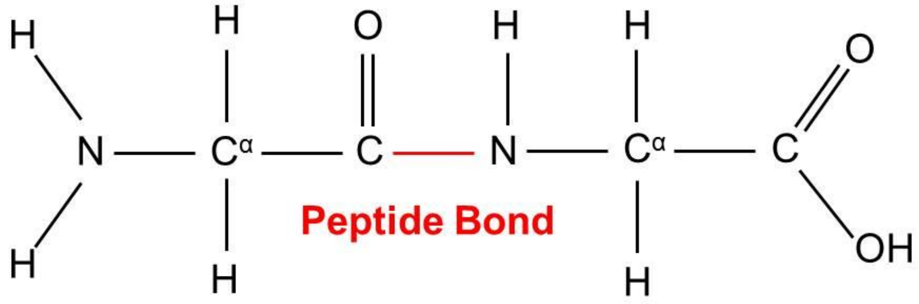

Figure 1.1: Peptide bond between two glycine amino acid residues.

In order to understand how the conformational flexibility conferred by the protein primary structure gives rise to specific structures, it is important to have an understanding of the interactions that stabilize proteins. Protein stability depends upon a balance of both short-range and long-range interactions. Within the protein, covalent bonds maintain the integrity of the polypeptide chain, while hydrogen bonding between electronegative atoms, van der Waals interactions between dipoles, and electrostatic interactions between charged atoms affect the protein's three-dimensional shape. Hydrophobic interactions between residues of the polypeptide chain and the surrounding aqueous solvent provide the driving force of protein folding [22].

Protein secondary structure consists of distinct motifs formed by regular arrangement of the protein backbone with varying sidechains [5,10]. Helices, which can be right-handed or left-handed, are stabilized by hydrogen bonds formed between the carbonyl and amide groups of the polypeptide backbone. Extended, slightly twisted β-strands coalesce into β-sheets by means of hydrogen bonding between chains (Figure 1.2). Polyproline helices lack hydrogen bonds, and are instead stabilized solely by van der Waals interactions.

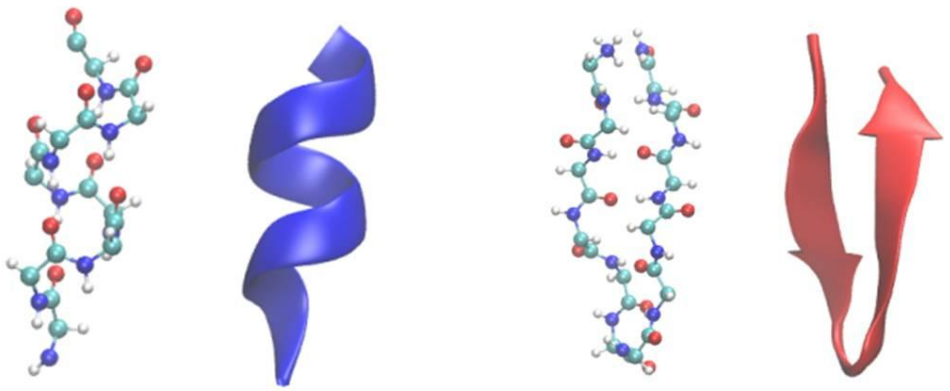

Figure 1.2: α-helix from trp-cage miniprotein (PDB 1L2Y) and β-strand from trpzip2 miniprotein (PDB 1LE1). Figure created using Visual Molecular Dynamics software.

Tertiary protein structure is the three-dimensional structure of the entire protein, including the coalescence of the individual structural domains described by the secondary structure. The structural propensities of each of the 20 amino acid residues are determined by the chemical characteristics of their sidechains, and structures are stabilized by hydrophobic or ionic interactions between them. Polar, hydrophilic amino acids prefer surface regions where they can easily participate in hydrogen bonds with the polypeptide chain as well as with water. Nonpolar, hydrophobic residues prefer β-structure, as their bulky sidechains are more easily accommodated in this arrangement, and the chemical nature of these sidechains results in their location in the protein interior. The burial of hydrophobic side chains is a major driving force for protein folding, as discussed in Section 1.3. Figure 1.3a shows top and bottom views of the hemoglobin molecule, which comprises α-helices connected by disordered loops. Each hemoglobin molecule is a tetramer constructed of four hemoglobin protein subunits, here shown in red, orange, blue, and cyan. Hemoglobin possesses quaternary structure due to its oligomeric construction. The interactions stabilizing protein quaternary structure are the same as those which stabilize tertiary structure. Figure 1.3b shows front (red) and back (blue) views of the green fluorescent protein, which contains β-sheets that have coalesced into a β-barrel structure. .

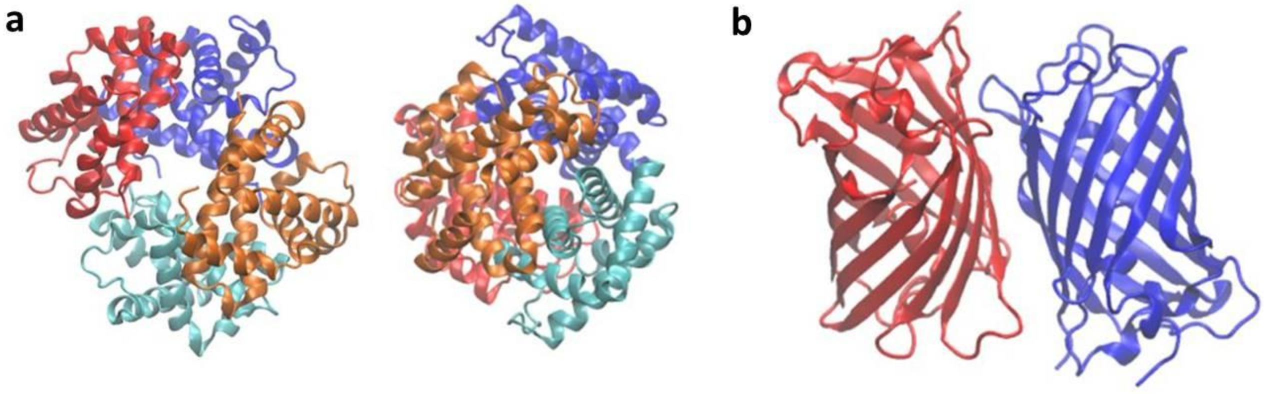

Figure 1.3: Structures of a) hemoglobin (PDB 1GZX) and b) green fluorescent protein (PDB 1GFL). Figure created using Visual Molecular Dynamics software.

1.3 The Protein Folding Problem

Having discussed the interactions that stabilize protein structure and drive protein folding, the question arises as to the process that a linear amino acid chain undergoes so that its requisite interactions produce the correct three-dimensional structure. For many large proteins, correct folding requires the assistance of molecular chaperones, different classes of which function to guide folding, prevent protein aggregation and misfolding, and to translocate proteins across membranes [23]. Relatively small globular proteins, however, are able to spontaneously fold *in vivo* without the assistance of chaperones. In 1961, Anfinsen noted that the small globular protein bovine ribonuclease was able to spontaneously re-fold into its functional structure after having been gently denatured [24]. This finding, known as Anfinsen's dogma, confirmed that all of the information necessary for a protein to fold is contained in the protein's primary amino acid sequence, and implies that under the conditions *in vivo* at which folding occurs, the native state is a unique, stable, and kinetically accessible minimum of the free energy.

How is this free energy minimum located? Experimentally, the folded state of globular proteins is known to be only marginally stable, as the difference in free energy between the folded and unfolded states often ranges from 5-15 kcal/mol. This marginal stability suggests that the system is near the critical point of a first-order phase transition, and that the folding process is a competition between entropy and enthalpy, according to the following relation

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

where G is the Gibbs free energy, H is the enthalpy, and S is the entropy. The hydrophobic effect is believed to be the main driving force in protein folding in aqueous solvent, and unlike the electrostatic interactions that drive the formation of some elements of tertiary structure, it is entropic in origin. Water is a structured liquid, as it is able to form hydrogen bonds with itself to form an ordered structure. The introduction of a hydrophobic solute, such as a protein, into water disrupts the hydrogen bonding network of the bulk water, and leads to the reorientation of a solvation shell around the protein which has restricted mobility and reduced entropy. The burial of the hydrophobic amino acids of the protein into the protein interior and away from the water reduces this entropic loss and provides the driving force for protein folding. The protein itself loses conformational entropy as it folds from an extended chain to a more compact structure, but also undergoes favorable enthalpic gains through interactions such as hydrogen bonds and salt bridges to find the native state.

In order for the thermodynamically favorable native state to be reached, it must also be kinetically accessible on a reasonable timescale. In 1969, Levinthal recognized that the search for the native structure potentially involved an exponentially large number of configurations [25]. For example, a polypeptide chain with 100 amino acids, each of which has two possible stable conformations, has 2^{100} possible conformations available to it. Assuming that a transition between conformational states occurs on the fastest possible timescale of one picosecond, corresponding to the period of a single bond rotation, a single folding event would require approximately 2^{100} picoseconds, or 10^{10} years, to sample all available conformations on the way to the fold of greatest thermodynamic stability. Levinthal's paradox recognizes that the attainment of a single native state is evidence of its stability, but as the estimate of folding time given above demonstrates, the protein does not have enough time to prove that the native structure is in fact the most stable among all those that are possibly available to it. As most proteins are able to fold on a timescale of microseconds to seconds, the amino acid chain clearly does not systematically attempt all possible conformations until it finds the one that is most energetically favorable.

The resolution of Levinthal's paradox lies in part in the application of aspects of energy landscape theory to the problem. Levinthal's statement regarded all protein interactions to be equally probable, corresponding to a random search. Levinthal himself postulated the existence of a specific, well-defined folding pathway for each protein, the end of which is the protein's native fold. This pathway is narrow in conformation space, such that extensive sampling is avoided, and the protein is driven to its free energy minimum. This suggestion leads to the conclusion that the native state might not correspond to the global free energy minimum of the protein, but rather, to a locally accessible metastable minimum. The existence of an energy bias toward the native state reduces the folding time scale to a realistic value.

Within this context, Wolynes suggested that, through the process of evolution, biologically functional proteins exhibit minimal frustration [26]. This frustration can be either energetic, which is dependent upon constraints of its sequence, or topological, which is dependent upon its structural topology. Onuchic proposed that, rather than following a single folding pathway, the native state might be the endpoint of an ensemble of convergent kinetic pathways [27]. The free energy landscape of a protein is thereby pictured as being biased toward the native state with few local minima to act as kinetic traps. This view of the protein folding problem necessitates the study of ensembles of molecules in order to obtain a complete picture of the energy landscape, as statistical theories allow for the determination of the probability distribution for sampling the energy surface.

Having established that the protein folding problem involves both thermodynamic and kinetic control, single molecules as well as statistically significant ensembles of these molecules, how do we develop a theoretical framework and experimental techniques to treat a system of such daunting complexity?

1.4 Experimental and Computational Approaches

Theoretical and experimental methods used to examine protein structure and the process of protein folding focus on capturing both static and dynamic pictures of molecular structure. These techniques thereby aim to extract snapshots of instantaneous structures, the kinetic rates at which events occur, and average thermodynamic properties over a statistical ensemble of protein structures. Single-molecule experiments have the potential to yield information about individual molecular trajectories, whereas ensemble experiments provide averaged quantities over the entire ensemble. The insight gained from these methods allows hypotheses to be formulated concerning the structure of proteins and their relation to function, with the ultimate goal of contributing to the understanding of disease, its prevention, and its cure. As in any field, the synergy between theory and experiment serves to provide a broader perspective than would be possible using just one approach. Although experiment and theory each have certain shortcomings, the advantages of one can in some cases be used to overcome the limitations of the other. When used together in a complementary fashion, theory and experiment can provide a more complete picture of the structure and dynamics of a protein.

In order to obtain static pictures of protein molecules, X-ray crystallography is often used. X-ray crystallography of biological molecules was pioneered by Kendrew and Perutz, with their elucidation of the structures of hemoglobin and myoglobin [28,29]. Using this method, X-rays are passed through a crystalline protein sample and produce a two-dimensional diffraction pattern that is then converted to a three-dimensional model using Fourier transforms. Although this technique is currently able to produce high-resolution (~1Å) atomic structures, the difficulty of obtaining a protein crystal, as well as the possible presence of crystallization artifacts in the sample, have the potential to skew the results. Additionally, the resulting picture is predominantly static, and does not suggest the dynamic movement of the molecule. Attempts to relate the distribution of electron density to conformational fluctuations using B-factors are not always straightforward, as they may be indicative of errors in structure refinement.

In NMR spectroscopy of proteins, pioneered by Wüthrich [30], aqueous samples of a protein are subjected to a large magnetic field, and the interaction of each atom's nuclear spin with the field allows for calculation of the relative locations of each atom. Refinement of NMR structures in principle produces an ensemble of structures in solution, but difficulty remains in obtaining a large number of experimental observations for each residue.

Other spectroscopic tools that can be used to probe the conformations of proteins include circular dichroism (CD), tryptophan fluorescence, fluorescence resonance energy transfer, pulsed electron paramagnetic resonance, and temperature-jump spectroscopy. Ongoing development of these techniques may improve their applicability to proteins in the future, but there remain fundamental limitations of these methods, such as the ability to probe only spectroscopically observable properties [31]. Due to these limitations, the need remains to develop techniques that will allow for the atomic-level observation of single molecules as well as ensembles, and time resolution as well as ensemble averages.

The possibility of performing “computer experiments” in the form of numerical simulations on physical and chemical systems was realized in the early 1950s, ushering in a new methodology for validating theories in physics and chemistry. During the Second World War, electronic computers were built and developed to perform calculations involved in the development of nuclear weapons, ballistics calculations, and code breaking. These computers became available for unclassified research in the early 1950’s, and one of the first areas of research to which these computers were applied was the simulation of dense liquids. Although many theories at the time treated the properties of dense liquids, few experiments other than painstaking mechanical simulation [32] were available to test the validity of these theories.

The first simulation of a liquid was carried out by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller [33] on the MANIAC computer at Los Alamos, using the Metropolis Monte Carlo method. Developed by Metropolis and Ulam [34] in the context of neutron scattering calculations, the Monte Carlo method uses probabilistic sampling to provide solution to intractable deterministic problems. The first simulations using the molecular dynamics algorithm, in which an equation of motion is integrated and a model for the interatomic forces, or forcefield, is used to describe the interatomic interactions, were performed in 1957 by Alder and Wainwright [35] at Livermore. Their research produced a time-dependent trajectory of motion of a collection of hard spheres. In 1964, Rahman reported the first MD simulation of a real liquid (argon) [36], and the first simulation of liquid water was undertaken in 1974 by Stillinger and Rahman [37].

The development of empirical potential energy functions for biological molecules was undertaken in the 1970s by a few different groups [38,39,40,41]. In 1975, Levitt and Warshel published a quantum mechanical/molecular mechanics (QM/MM) study of bovine pancreatic trypsin inhibitor using a coarse-grained model with stochastic dynamics and used the results to present a general model for protein folding [42]. MD simulation of the same protein was undertaken in 1977 by McCammon, Gelin, and Karplus, and the results outlined the importance of fluctuations to the dynamic nature of the protein [43]. Since these seminal works were published, the field of computational structural biology has produced a significant body of work that has led to great insights in problems concerning the relationship between biological structure and function. The vast expansion of computer resources that has been made available for this problem has allowed simulation to grow from a methodology used in support of theoretical and experimental approaches to a discipline in its own right.

Protein folding, in particular, has been viewed as one of science's current “grand challenge” problems, and recent years have seen the development of world-class supercomputers designed specifically for tackling this problem [44,45]. Currently, work in the field of computational protein folding focuses on three major areas: the *de novo* prediction of native protein structure from its amino acid sequence, the thermodynamic problem of how the interatomic forces encoded in an amino acid sequence give rise to a stable native structure, and the kinetic question of how the complex system of a protein is able to locate its native conformation so quickly [46,47,48,49].

Despite the successes of biomolecular simulation within the context of these goals, there remain limitations in the ability of this methodology to accurately model protein structure and dynamics, and to solve problems efficiently in terms of the time required by the work and the computational resources required. First, the forcefields that are used to model the macromolecular system are only approximations to the real physics, and are often chosen for computational tractability rather than accuracy. Inaccuracies in the forcefields and solvent models result in pathological problems in the simulation structures, and although these inaccuracies are usually discovered and corrected, they still point to fundamental deficiencies in the models. Secondly, the sizes of the biomolecular systems under study are limited due to losses in computational efficiency that arise when certain size limits are reached. This size limitation precludes the study of large systems of biological interest, and often prohibits the simulation of systems in an aqueous solvent that is modeled in atomic detail.

Finally, as discussed above, proteins exist on a rugged free energy landscape, which limits the extent of conformational space that can be visited in a single simulation. The development of enhanced sampling techniques, which is the focus of Chapters 2 and 3 of this dissertation, aims to bypass this limitation by manipulating the formulation of the system's energy function and the equations of motion. Improvement of the sampling problem should serve to increase the size of the systems that are able to be studied in a computationally efficient way.

Applications of molecular dynamics techniques outside of the field of protein structure determination include studies of the thermodynamic and kinetic properties of novel materials. One of the most important goals of simulation studies of biological, as well as non-living material is to be able to understand the mechanisms driving their function. Applying this knowledge in a clinical setting may result in the ability to treat and prevent disease. Chapter 4 of this dissertation discusses all-atom molecular dynamics simulations of diblock star copolymers, which are self-assembling nanoscale systems that have shown great potential in the field of targeted drug delivery in the human body. Intriguingly, these star polymer systems bear many important similarities in structure and composition to proteins, being composed of linear polymeric chains of repeating units which self-assemble with hydrophobicity as the driving force [17]. These similarities allow for the application of many of the techniques of molecular modeling and simulation developed for proteins to these systems.

1.5 Polymeric Nanoparticles with Drug Delivery Applications

Current estimates place the percentage of small drug and drug-like molecules that are hydrophobic at 40% [50]. Due to the poor aqueous solubility of these molecules, recent efforts in drug formulation and delivery have focused on increasing their solubilization. Several lipid-based materials have emerged as effective solubilizing agents, as they are able to dissolve highly hydrophobic drugs in their lipid bilayer, as well as hydrophilic therapeutics in their interior (Figure 1.4a) [51]. Additionally, these liposomal formulations are often able to release the drug to a specific target, which minimizes the number and severity of patient side-effects and allows for more efficient administration of the drug [52].

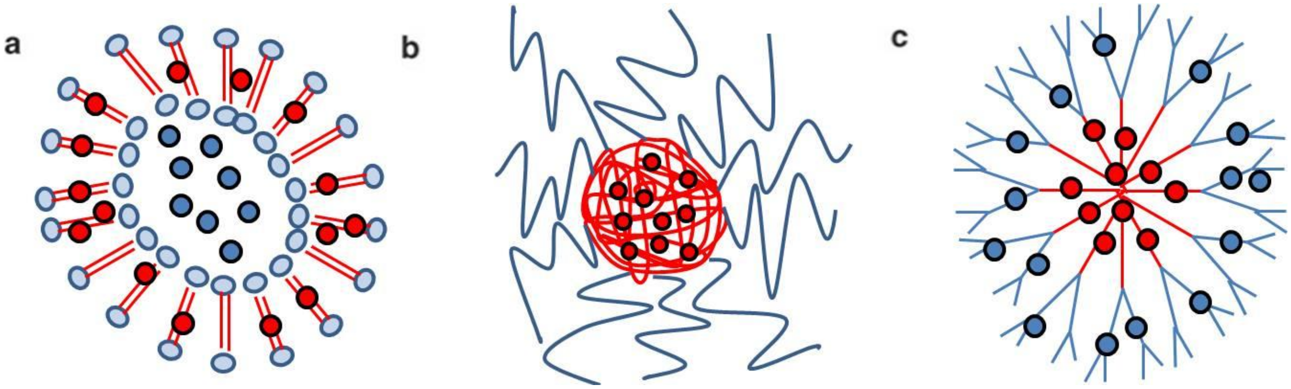

Figure 1.4: Schematic of nanoparticles currently in use and under development for drug delivery: a) liposome, b) polymeric micelle, c) dendrimer. Blue areas represent hydrophilic regions, and blue spheres represent hydrophilic drugs. Red spheres represent hydrophobic regions, and red spheres represent hydrophobic drugs.

As a result of these advantages, the FDA has approved several liposome-based formulations for the treatment of various disorders, including Myocet® for metastatic breast cancer, Abelcet® for fungal infections, and DaunoXome® for HIV-related Kaposi sarcoma [53]. A number of additional nanomedicines are currently undergoing clinical trials or are in preclinical development [53,54]. Despite the initial success of liposomal drug carriers, liposomal formulations are limited in their ability to solubilize highly hydrophobic drug molecules, and these molecules are often rapidly released from the lipid bilayer following their entry into the bloodstream. Once released into the bloodstream, liposomes are often rapidly recognized through opsonization, which is a mechanism by which foreign materials in the bloodstream are flagged for recognition and removal by macrophages. The rapid exit from the bloodstream by liposomal drug carriers often leads to an unfavorable therapeutic index [53,54].

The desire to create a vehicle for drug delivery with the solubility of a liposomal formulation, but with greater control of drug release, has led to the investigation of compounds that exhibit increased stability in aqueous solvent while retaining the ability to solubilize highly hydrophobic materials. Block copolymers have emerged as a material with a great deal of potential for accomplishing these goals [54,55,56]. Block copolymers consist of segments of chemically distinct mononers that are covalently bonded to one another. Diblock copolymers consist of two different subunits, such as polycaprolactone and polyethylene glycol, which are covalently linked. When placed into a solvent that is selective for one of the blocks, diblock copolymers undergo microphase separation and self-assemble into micellar structures (see Figure 1.4b). In these micelles, the insoluble block forms the core, while the soluble block comprises the outer shell, or corona. The size and topology of the resultant structure depends on the relative sizes of the two blocks, their chemical identities, and their interactions with the solvent.

Spherical or cylindrical micelles are commonly produced; however, recent synthetic advances have led to an increase in topological complexity in the form of liposomes, brush copolymers, dendrimers, and star block copolymers (Figure 1.4c) [56,57].

1.6 Star Diblock Copolymers

Star block copolymers retain a micellar architecture, but one end of each amphiphilic arm is tethered to a central group, thereby forming a unimolecular micelle. As the individual amphiphilic blocks are covalently fixed, the polymer exhibits an enhanced stability when compared to untethered micelles, which exist in equilibrium between bound and unbound monomers. The stable micellar architecture of the star copolymer makes it a singularly effective platform for drug delivery. Use of an amphiphilic diblock in a polar solvent such as the human bloodstream creates a nonpolar interior core with a hydrophilic exterior that remains solvated and protects the hydrophobic interior [56]. The micelle formed by an amphiphilic diblock star copolymer thereby has the potential to sequester a hydrophobic drug in its interior, increasing its apparent solubility and allowing the drug to be stably and effectively transported to its ultimate target [53]. The targeting of the drug cargo, which will be discussed below, also increases the specificity of the drug's activity and is believed to decrease the occurrence of side effects in the patient [58]. Effective encapsulation and transport by the star copolymer system is dependent upon the solubility of the drug molecule in the hydrophobic core, as well as the stability of the core monomer self-association in aqueous solvent. Due to the large number and chemical diversity of drug-like molecules that have the potential to be carried by star copolymer systems, a large number of combinations of drug molecules and hydrophobic polymer carriers have been experimentally tested in order to determine the most effective pairs [53]. We now turn to a discussion of the most common materials used for the hydrophobic interior.

Polyethers such as poly(ethylene oxide)-poly(propylene oxide) (PEO-b-PPO) block copolymers have been extensively investigated for drug delivery, which has lead to FDA approval of several of their derivatives, including Pluronic® and Tetronic®, for use as drug carriers [59]. These materials, however, exhibit instability even when cross-linked, dissociating quickly after injection and causing an undesirable burst release of the encapsulated drug. Due to the shortcomings of carriers based on PPO, other materials, such as polyamino acids and polycarbonates, have been examined as alternatives [56]. The most promising class of alternative materials that has been examined is the polyesters. Poly(ε-caprolactone) (PCL) is more hydrophobic than PPO, and studies have indicated that this material exhibits a greater stability, particularly for neurotrophic agents [60]. Poly(lactide) micelles created from block copolymers of both pure stereoisomer forms L-lactide and D-lactide are mixed together exhibit improvements in stability, likely due to the formation of a stereocomplex between the pure stereoisomer chains [56]. The potential for polyvalerolactone (PVL) to be used as a drug excipient has also been shown experimentally [61]. Due to their potential for high stability when introduced *in vivo,* simulations of drug delivery vehicles based on these materials will be the focus of the work outlined below in Chapter 4.

Of equal importance to the material chosen for the hydrophobic core, the polymer that comprises the surrounding hydrophilic corona must exhibit a high solubility in the solvent in order to maintain the core's stability and structural integrity. The most widely used monomer for the corona is poly(ethylene glycol) (PEG), which is nonionic and completely miscible with water at room temperature. Its high miscibility causes its surfaces to become saturated with water molecules, which not only contributes to solubility, but also prevents opsonization [56]. Nanoparticles that are coated in PEG therefore remain concealed in the bloodstream for extended periods of time without being recognized or removed. Such particles are commonly known as "stealth" molecules. PEG thereby acts to enhance the pharmacokinetics of many drugs that are currently on the market. The addition of PEG to a liposomal or polymer formulation is known as "PEGylation." PEGylation has been used to improve the formulation of several liposomal drugs, including Pegasys®, which is a reformulation of interferon alpha used in treatment of chronic hepatitis C and hepatitis B that enhances the half-life of the drug [53].

The properties of the coronal layer also largely control the transport and targeting of the polymeric vehicle, as the polymer chains comprising the corona extend into the bloodstream and interact with biological molecules, such as proteins and antibodies. In active targeting, molecules that interact with cell surface receptors are attached to the end group of the hydrophilic chain of the block copolymer [56]. For example, the addition of folic acid to a hydrophilic block containing PEG targets tumor cells, which over-express folate receptors in many kinds of cancers. The folic acid binds to the folate receptor, triggering a receptor-mediated endocytosis process that internalizes the folic acid-receptor complex as an endosome [58]. Because many cells are able to eliminate drugs that reach the cytosol, some mechanisms of drug delivery need to target DNA as closely as possible to ensure that the drug reaches its target. Such control over drug delivery is obtained through the use of materials that are sensitive to pH or temperature, such as polyethyleneimine (PEI) or poly(N-isopropylacrylamide) (PNiPAAm) [60]. The systems simulated in Chapter 4 of this work use PEG in the coronal layer due to its properties of stealth and targetability, which have led to its documented success in drug formulations that are already on the market.

The reformulation of liposomal drugs with block copolymers has lead to improvements in drug pharmacokinetics and delivery. The anti-cancer agent paclitaxel was originally solubilized in a polyethoxylated castor oil known as CremophorEL (Taxol®). Although highly effective against cancer, Taxol® exhibited several dose-limiting toxicities in patients. Taxol® was reformulated with a PEG-b-PDLLA copolymer (Genexol-PM®), and clinical trials of this new formulation have reported a 6000-fold increase in solubility, along with the toleration of higher doses of the drug with fewer patient side effects. Several therapeutic agents currently in preclinical development are composed of drugs that are successfully solubilized in liposomes, such as doxorubicin and daunorubicin, with polymeric micellar vehicles [50,53].

At present, experimental imaging of star diblock copolymers and nanogel star copolymers, particularly in complex with drug molecules, has been limited, providing computational studies with the opportunity to predict the structures of these molecules in atomic detail, as well as their dynamic behavior. In Chapter 4 of this work, we describe a comparative study of three star block copolymer systems with varying hydrophobicity in their core regions. The goal of this work is to provide atomic-level information on star polymer structure and dynamic behavior, including the size and shape of the polymer, the details of its bonding patterns, and its potential for aggregation. Additionally, the kinetics of drug uptake and delivery, as well as the degradation profile of the delivery material, may also be examined. Because theoretical methods, in contrast to experiment, are often less expensive and more time-efficient, their systematic application may offer strategies at the molecular level by which to modify formulations of drug and polymer for optimal compatibility and delivery efficiency [62,63,64].

2. Methods

2.1 Introduction to Molecular Dynamics and Stochastic Dynamics Simulations

As discussed in the introduction to this work, the three current goals of simulation studies of proteins are the correct prediction of their functional structure based on the identities of their constituent molecules, the description of the thermodynamic properties of ensembles of molecules, and the delineation of the kinetic pathways taken by proteins on their energy landscape as they reach their functional conformation. The first two goals employ simulation in order to sample the configuration space that is available to a biomolecule under certain thermodynamic conditions. In these cases, simulations are used to obtain a description of the system once equilibrium has been reached, requiring not only that conformation space is sampled, but with the additional requirement the each state be weighted by its corresponding Boltzmann factor. In order to achieve this equilibrium sampling, non-deterministic stochastic simulations can be undertaken. When an accurate description of the time evolution of the system, as well as its dynamic properties, is required, a molecular dynamics simulation is undertaken, in which Newtonian mechanics is used to propagate the system in time.

Thus, as a complement to experiment, simulations of biomolecules may be used to obtain both equilibrium and time-dependent properties of both single molecules and ensembles. In molecular dynamics and stochastic simulations, the potential energy of the system is described in terms of a function of atomic positions known as a force field, and the value of the system Hamiltonian is numerically solved over a discrete time step. The forces on all of the particles in the system are computed, and the appropriate equations of motions are integrated to obtain the motion of the system over the specified timestep. According to the ergodic hypothesis, the time average of a thermodynamic property of a trajectory is equivalent to its ensemble average. From the time evolution of the system, simulations may thus reveal both the spatial and temporal extent of conformational sampling, as well as the instantaneous and ensemble averages of thermodynamic properties.

It is the goal of this chapter to outline the steps taken in an MD or stochastic simulation, including the modeling of the biomolecule, its motion, and its environment. The details of this discussion are particular to the forcefield that is used with the AMBER molecular dynamics software package [65], although most of the information is generalizable for other forcefields that are commonly used, such as GROMOS [66], CHARMM [67], OPLS [68], and LAMMPS [69]. Additionally, we briefly discuss the necessity of developing new algorithms that have the goal of overcoming limitations in computational efficiency and conformational sampling; the details of these methods are addressed in later chapters of this work.

2.2 Modeling Biomolecules: Peptide and Polymer Model Systems

Factors that limit the feasibility of simulations of biological systems are the high level of detail and long timescale that are required to obtain accurate and biologically relevant results. One way to reduce the computational cost of a simulation is through simplification of the system under study, in terms of size and complexity. The use of model systems allows us to quickly gain insight into the use of new force field parameters or novel simulation algorithms before they are applied to larger, biologically relevant systems.

Useful peptide model systems provide known, stable examples of certain types of secondary structure. The work described in subsequent chapters focuses on the use of model peptides in the validation of new methods. In Chapter 3, the β-hairpin tryptophan zipper 2 [70], the α-helix K19 [71], and the trp-cage miniprotein [72] are used to test the self-guided Langevin algorithm. In Chapter 4, the alanine dipeptide model system, which is able upon solvation to fully sample the range of φ and ψ dihedral angles that is available to protein α-helix and β-strand motifs, is used to validate a new method involving replica exchange molecular dynamics.

Additionally, in Chapter 5, we use a simplified model consisting of 16 diblock copolymer arms bound to a rigid adamantane core in order to undertake simulations of nanogel star diblock copolymers. Although this model system is much smaller than any synthesized star polymer, we believe that it can serve as a useful model for the polymeric structure and solvent-nanoparticle interactions of more complex systems. This discussion of model systems for biomolecules brings us to an important point when discussing simulation methodology: all of our models, whether they are of the biomolecule, its interactions, or its environment, possess certain limitations. Exploring the shortcomings of these models allows them to be continually improved in accuracy and efficacy, and allows for a more complete description of the natural structures and phenomena that we aim to describe.

2.3 Modeling Interactions: The Molecular Dynamics Forcefield

Molecular dynamics (MD) simulations are grounded in the assumption that the motions of the atoms and molecules in a system can be modeled using classical mechanics. In theory, a quantum mechanical treatment of the system would be most accurate; however, accurate solutions to Schrödinger's equation are not feasible for any except the smallest systems. Using the Born-Oppenheimer approximation [73], which states that the nuclei remain fixed on the timescale of the motion of the electrons, the potential energy of the system can be described classically as a function of the nuclear positions, defining the Hamiltonian of the system as follows

$$H(\vec{p},\vec{r}) = K(\vec{p}) + U(\vec{r}) \tag{2.1}$$

where $K(\vec{p})$ is the kinetic energy and $U(\vec{r})$ is the potential energy. This Hamiltonian considers the molecule to be a collection of nuclei connected by bonds that are modeled as springs, and the molecule stretches, bends, and rotates about these bonds as a response to intermolecular and intramolecular forces. This potential function must be constructed empirically using the appropriate molecular data for each of its terms. Numerical minimization of this function allows for determination of favorable regions in the configuration space of the molecule.

The molecular mechanics Hamiltonian is the sum of the potentials of the contributing physical forces. In a protein, the forces are generally separated into bonded and nonbonded terms, which describe the local and long-range interactions within the molecule. Local terms describe potentials for lengths describing bond stretching, as well as angles describing bond bending and rotation. Nonlocal terms include a Lennard-Jones potential [74] to model repulsion at short interatomic separations and attraction at long distance, and a Coulombic potential [75] among the pairs of charged particles in the system. One functional form for a forcefield describing these interactions that is commonly used for biological macromolecules is written as follows [76]

$$\begin{aligned} U(\vec{r}^{\,N}) = {} & \sum_{bonds} \frac{k_i}{2}(\vec{r}_i - \vec{r}_{i,eq})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,eq})^2 + \sum_{dihedrals} \frac{V_n}{2}(1+\cos(n\omega-\gamma)) \\ & + \sum_{i<j}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \end{aligned} \tag{2.2}$$

where $U(\vec{r}^{\,N})$ denotes the potential energy as a function of the *N* atomic positions.

The first and second terms in Equation 2.2 model bond stretching and angle bending after Hooke’s law as a harmonic potential of bond lengths *r* and bond angles *θ* with force constant *k*. The empirical values in these terms are obtained from vibrational spectroscopy, as well as from solved X-ray crystal structures or quantum mechanical solutions to equilibrium structures of small molecules. The third term describes the potential as a function of the internal rotation of the molecule, which is dependent upon conformational barriers to rotation. The parameter V_n contributes to the barrier height, ω is the torsion angle, and γ is the phase factor. Empirical data obtained from NMR, IR, Raman, and microwave spectroscopy can be used for barrier height and periodicity estimation in compounds of low molecular weight. Torsional parameters in this term are obtained by using *ab initio* quantum-mechanical calculations combined with geometry optimizations.

The fourth term in Equation 2.2 is the Lennard-Jones 6-12 potential [74], which describes both the attractive long-distance dispersive interactions (or London forces) [77] and the short-range repulsions due to the Pauli exclusion principle [78]. Although the Buckingham potential [79] would provide a better fit over a broader range of separation distances, the Lennard-Jones potential is chosen for mathematical convenience. The Lennard-Jones potential includes two parameters: the molecular distance at which the energy is zero is described by the collision diameter σ, and the depth of the function's energy minimum is described by ε. These parameters can be empirically obtained by fitting to lattice energies and crystal structures, or by extrapolation from liquid simulations. The final term in Equation 2.2 is Coulomb's law, which describes the ionic interactions between partially or fully charged groups. In this equation, q_i describes the effective charge on atom i, r_{ij} is the distance between atoms i and j, and ε is the dielectric constant. Parameters for this term are obtained through *ab initio* quantum-mechanical calculations.

Evaluation of the nonbonded terms is computationally quite costly because the number of local terms grows linearly with the number of atoms, whereas the number of nonbonded terms grows quadratically. Because the non-bonded Coulomb forces change more slowly with distance than the bonded terms, algorithms using cutoff functions and Ewald summation [80,81,82] have been developed to reduce the computational complexity of the nonbonded calculation. The evaluation of the electrostatic terms is particularly costly when explicit representations of the solvent, which may involve hundreds of thousands of atoms, are included in the system. When structural water molecules and the local effects of solvation are not vital in the simulation of the protein, explicit water molecules may be replaced by a continuum representation, which saves computer time by reducing the computational complexity of the system. In Section 2.4, we discuss techniques for modeling the solvent using explicit all-atom, or implicit mean-field representations.

2.4 Modeling the Environment: Solvent Models

Due to the central role of water in the dynamic process of protein folding, accurate simulations depend upon accurate modeling of the protein’s surrounding environment. The most accurate water models for use with simulation are explicit water models, which model each water molecule individually. A few families of explicit water models have been developed [37,83], but the group of models most commonly used with the AMBER forcefield is known as the transferable intermolecular potential (TIP). In this model, each water molecule maintains a rigid geometry in which charges are placed on specific sites. In the TIP3P model [84], for example, partial positive charges are placed on the hydrogen atoms, which are balanced by a negative charge located on the oxygen atom. The TIP4P model [84,85] moves the negative charge off of the oxygen and onto a fourth site between the two hydrogen atoms in order to better reproduce thermodynamic and structural data obtained from experiment, such as the radial distribution function. Interaction between molecules is described using pairwise Coulombic and

Lennard-Jones expressions. Solvent models are parameterized to accurately reproduce bulk thermodynamic and kinetic properties of bulk water in order to accurately model the effect of placing a biomolecule in an aqueous solvent. The TIP4P-Ew solvent model [86], for example, is able to reproduce the experimental bulk density and enthalpy of vaporization of water, as well as structural properties such as x-ray scattering intensities, at a large range of temperatures. Despite the accuracy of these models, the number of particles that they require often leads to prohibitive computational expense, as discussed above.

An alternative to explicit modeling of water molecules is provided by replacing them with a dielectric continuum that approximates the properties of liquid water. The use of an implicit solvent model not only saves computational time and resources, but potentially gives improved sampling due to its lack of viscosity, which allows for a more complete search of conformational space to occur during the simulation. The potential energy function given in Equation 2.2 represents the energy of the molecule in vacuum. In order to calculate the total energy of the solvated system, the solvation free energy, which is the free energy of transferring the molecule from vacuum to solvent, must be added to the vacuum energy. The solvation free energy is assumed to be decomposable into electrostatic and nonelectrostatic components as follows

$$\Delta G_{solv} = \Delta G_{el} + \Delta G_{nonel} \tag{2.3}$$

where ΔG_{nonel} is the free energy of solvating a molecule with a partial charge of zero on every atom, and ΔG_{el} is the free energy of removing all charges from the molecule while in vacuum, then adding them back in the continuum solvent environment.

The most accurate model of electrostatic interactions in a dielectric medium is the Poisson-Boltzmann (PB) equation [87], which can in simple cases be used to exactly calculate solvation free energy of the solute and solvent. For systems of complex geometries, however, the PB equation cannot be solved exactly. The generalized Born (GB) method [88,89] has thereby been extensively implemented to approximate the physics of the PB equation, while decreasing its computational intensity. In this analytical approximation to the Poisson-Boltzmann equation, atoms are modeled as charged spheres with an internal dielectric that is lower than that of the environment. Each atom is assigned an effective radius such that the solvation free energy can be calculated using the Born formula. Different GB models differ in how they calculate the effective Born radius, which indicates the level of screening of atomic charge by the solvent. The Hawkins-Cramer-Truhlar (HCT) model [90] is known as a pairwise GB model, as it approximates the volume integral as a sum over the contribution of each atom. The Onufriev-Bashford-Case (OBC) GB model [91] is an improvement to the HCT model, which aims to correct the values of the effective Born radius for buried atoms, which were too low in the HCT model. Further improvements to these models, such as those given in the recent models known as GBNeck [92] and GBNeck2 [93] continue to increase the accuracy of the implicit solvent formulation.

As with many other approximations made in simulations, the use of an implicit solvent model represents a tradeoff between accuracy and tractability. Another approach to modeling the thermodynamic and kinetic effects of placing a protein in a solvent, which will be discussed below in Section 2.5, is to use a phenomenological model, such as Brownian or Langevin dynamics, which mimics the effect of solvent on the system.

2.5 Modeling Motion: Deterministic and Stochastic Models

The above sections of this work describe how the system of protein and explicit or implicit solvent are modeled in terms of the potential energy function. Once the system has been thus set up, propagation of the chosen equations of motion over a discrete time step yields the time-dependent motion of the system. Among the choices available for the equations of motion, this section will focus on deterministic dynamics as described by Newton's equations, and stochastic dynamics as described by the Langevin formalism.

2.5.1 Deterministic Dynamics: Newton's Equations of Motion

Under Newtonian dynamics, the equations of motion for a system of *N* atoms may be written as a first-order differential equation as follows:

$$\vec{F}(\vec{X}(t)) = M\dot{\vec{V}}(t) = -\nabla U(\vec{X}(t)) \quad . \qquad (2.4)$$

In this equation, $\vec{F}(\vec{X})$ is the systematic force as a function of the Cartesian vector $\vec{X}$ of the *N* atoms in the system. The force is equivalent to the diagonal mass matrix *M* multiplied by the derivative of the velocity vector $\vec{V}$, or to the negative gradient of the potential energy $U(\vec{X})$. This gradient may be expressed in terms of its components *i=1,…3N*, as follows

$$\nabla U(\vec{X})_i = \frac{\partial U(\vec{X})}{\partial \alpha_i} \quad , \qquad (2.5)$$

in which α_i denotes the *x*, *y*, or *z* component of an atom. Analytical solutions are only available for the simplest systems, requiring these equations to be integrated numerically. The result of the integration is a sequence of coordinate and velocity pairs at each chosen time step.

The choice of time step for the simulation is dictated by the timescale of the highest-frequency modes of motion that still impact molecular structure. Although many components of molecular motion occur on very short timescales, the collective motion of a biomolecule is highly cooperative, and local fluctuations have the ability to impact global structure. The frequencies of bond vibrations are on the scale of 10^{14} s^{-1}, which generally restricts the time step in an MD simulation to be on the order of 10^{-15} s, or 1 fs. A factor of two can be gained in the length of the time step if constrained dynamics is used to maintain the rigidity of bonds to hydrogen, as described below in Section 2.5.2. Typical simulations are run on the order of nanoseconds, although recent improvements in computing power, as well as the development of molecular mechanics programs such as NAMD, GROMACS, and Desmond for massively parallel computer architectures, are increasingly bringing simulations to the microsecond timescale [94,95,96].

The choice of starting structure or structures for the simulation depends largely upon the goal of the simulation, the size of the system to be studied, and the availability of experimental coordinates. In studies where the goal is to obtain statistical properties of an ensemble of biomolecules, a large number of starting structures in diverse conformations is customarily used. Additionally, the chaotic nature of individual trajectories obtained from MD simulations often results in more reliable data being obtained from averages over several trajectories, rather than taken from a single, long trajectory. Single trajectories, however, remain useful when examining the atomic details of protein or polymer conformational change and function.

In order to begin the simulation, an initial velocity vector is set in order to bring the total kinetic energy of the system to the expected value at the target temperature at which the simulation will be run. According to the classical equipartition theorem [97], each translational degree of freedom of a molecule in thermal equilibrium has the same average thermal energy. The average kinetic energy may thus be expressed as

$$\langle E_k \rangle = \frac{1}{2}\sum_{i=1}^{3N} m_i \vec{v}_i^{\,2} = (N_F k_B T)/2 \tag{2.6}$$

where N_F is the total number of translational degrees of freedom in the system. Velocity components are thereby assigned from a Gaussian distribution with a variance of *$(k_BT)/m$* and zero mean, as the following expression should hold at equilibrium:

$$\left\langle \vec{v}_{\alpha}^{2} \right\rangle = k_B T / m \quad . \qquad (2.7)$$

Using this relation, the instantaneous temperature is defined as

$$k_B T(t) = \sum_{i=1}^{N} \frac{m\vec{v}_{\alpha,i}^{2}(t)}{N_F} \qquad (2.8)$$

and the temperature can be adjusted during the simulation to match the desired temperature by scaling the velocities by $\sqrt{T/T(t)}$.

Before the production run is initiated, one or more rounds of equilibration are usually necessary in order to slowly bring the simulation to the desired temperature and to relax the protein and/or solvent structures to a state of low energy. During the equilibration, exchange between kinetic and potential energy occurs, and the system is considered to be equilibrated once there is convergence of these energy terms (this behavior describes that of the microcanonical ensemble; for a discussion of other ensembles, see Section 2.5.3 below). After the force on all particles is calculated according to Equation 2.4, the positions of the particles are propagated through time, resulting in a trajectory.

2.5.2 Integration Algorithms

The choice of integration algorithm is vital to ensuring the reliability of a molecular dynamics simulation. Criteria for a good integration algorithm include accuracy and stability for long time steps and the avoidance of long-term energy drift. One of the most widely-implemented group of integration algorithms is the Verlet family [98], including the leapfrog [99], velocity [100], and position Verlet [101] algorithms. These algorithms not only exhibit stability, but are also symplectic, which results in the bounding of energy fluctuations and preservation of the conservative property of the system.

In the Verlet propagation scheme for Newtonian dynamics, the equation of motion is written as in Equation 2.4. To discretize this problem, the continuous variables *X(t)* and *V(t)* are approximated by values at discrete time steps *nΔt* and written as X^n and V^n. The positions and velocities of the system are then recursively defined

$$V^{n+1} = V^n + \Delta t \tilde{F}^n \tag{2.9}$$

and the values of the positions from two previous steps are used to calculate the updated position as follows:

$$X^{n+1} = 2X^n - X^{n-1} + \Delta t^2 \tilde{F}^n \quad . \tag{2.10}$$

A Taylor expansion around *X(t)* is then used to solve for the updated position:

$$\begin{aligned} & X(t+\Delta t) + X(t-\Delta t) \\ & = X(t) + \Delta t V(t) + \frac{\Delta t^2}{2}\tilde{F}(x(t)) + \frac{\Delta t^3}{6}\ddot{V}(t) + O(\Delta t^4) \\ & + X(t) - \Delta t V(t) + \frac{\Delta t^2}{2}\tilde{F}(x(t)) - \frac{\Delta t^3}{6}\ddot{V}(t) + O(\Delta t^4) \\ & = 2X(t) + \Delta t^2 \tilde{F}(x(t)) + O(\Delta t^4) \end{aligned} \quad . \tag{2.11}$$

Velocities are not used in the algorithm, so they must be estimated using the positions, through subtraction of the Taylor expansion for *X(t-Δt)* from that for *X(t+Δt)* to obtain:

$$\frac{X(t+\Delta t) - X(t-\Delta t)}{2\Delta t} = V(t) + O(\Delta t^2) \tag{2.12}$$

Variants of the Verlet algorithm include the leapfrog scheme [99], the velocity Verlet [100], and the position Verlet [101]. Using the leapfrog scheme, the velocity is defined at half-timesteps $V^{n+1/2}$, while the positions are defined at whole timesteps X^n of *Δt*. A point in phase space is thereby transformed to the next as $\{V^{n-1/2}, X^{n}\} \Rightarrow \{V^{n+1/2}, X^{n+1}\}$, which may also be written as

$$\begin{aligned} V^{n+1/2} &= V^{n-1/2} + \Delta t \widetilde{F}^{n} \\ X^{n+1} &= X^{n} + \Delta t V^{n+1/2} \end{aligned} \quad . \qquad (2.13)$$

In order to increase the simulation timestep, constrained dynamics is often employed, in which the highest-frequency motions, such as vibration, are frozen through the addition of algebraic constants to the equations of motion. As an example, we have seen in Equation 2.2 that the bond length potential is approximated using the harmonic form

$$U_{bond} = \frac{k_i}{2}(r_i - r_{i,eq})^2 \qquad (2.14)$$

where *r* is the interatomic distance and *k* is the force constant. The bond length may be constrained as follows:

$$g_k = r_i^2 - r_{i,eq}{}^2 = 0 \quad . \qquad (2.15)$$

One form of constrained dynamics, known as the SHAKE algorithm [102], has been implemented with the leapfrog Verlet algorithm as follows:

$$\begin{aligned} V^{n+1/2} &= V^{n-1/2} - \Delta t \nabla U(X^{n}) - \Delta t g'(X^{n})^{T} \Lambda^{n} \\ X^{n+1} &= X^{n} + \Delta t V^{n+1/2} \\ g(X^{n+1}) &= 0 \end{aligned} \quad . \qquad (2.16)$$

where Λ is a vector of Lagrange multipliers. A symplectic variant of SHAKE known as RATTLE [103] is used with the velocity Verlet algorithm; although SHAKE is not symplectic, it produces results that are identical to those of RATTLE for the positions and results in velocities that are only slightly perturbed.

2.5.3: Other Ensembles

The quantities obtained in an MD simulation performed as described in the previous section are equivalent to quantities in the microcanonical (NVE) ensemble, as the volume and number of particles are fixed. Although the temperature fluctuates during the simulation, the total energy is a constant, as the Hamiltonian is a conserved quantity in the presence of conservative forces. Unless studying dynamic properties such as diffusion or relaxation phenomena, the microcanonical ensemble is not the most convenient for comparison with experimental data. For thermodynamic properties to be compared with experimental results, the simulation should be run in the canonical (NVT) or isothermal-isobaric (NPT) ensemble. To run simulations in these ensembles, techniques have been developed in which the system is considered to be in contact with a reservoir, and is no longer determined by a real Hamiltonian or held to satisfying the conservation laws of Newtonian dynamics.

The simplest method used for running constant-temperature MD simulation involves scaling the velocity vector at each timestep to fix the desired kinetic temperature T at the target value T_0, as described above. This approach, however, results in rapid energy transfer among the various degrees of freedom in the system, and often leads to the pumping of energy into low-frequency modes. A less drastic approach to running MD simulations at constant temperature is to use the Berendsen weak coupling thermostat [104]. In this approach, a diffusive process is mimicked by introducing a stochastic coefficient of friction which controls the relaxation rate of coupling to the heat bath. The modified equations of motion using the Berendsen thermostat are as follows:

$$\begin{aligned}&\dot{\vec{X}}(t)=\vec{V}(t)\\&M\dot{\vec{V}}(t)=-\nabla U(\vec{X}(t))-\gamma_t M\vec{V}(t)\\&\gamma_t=\frac{1}{2\tau}\left(1-\frac{T_0}{T}\right)\end{aligned} \quad , \tag{2.17}$$

where γ_t has units of inverse time, τ is an empirical constant of the decay time of the coupling to the heat bath, T_0 is the target temperature, and T is the instantaneous kinetic temperature. This implementation effectively scales the velocity vector by the factor

$$c_t=\sqrt{1-\frac{\Delta t}{\tau}\left(1-\frac{T_0}{T}\right)} \quad . \tag{2.18}$$

The use of the adjustable coupling parameter γ allows for careful control of the rate at which the target temperature is obtained. Weak coupling to the heat bath through the use of a small value of γ_t results in the scaling factor approaching unity, and the results approach those of the microcanonical ensemble. Strong coupling through the use of a large value of γ_t results in significant exchange of energy between the system and the thermal reservoir.

A similar approach is used for constant-pressure MD simulations, in which the Cartesian positions and volume are scaled as follows

$$X^{new}\leftarrow c_p X^{old} \tag{2.19}$$

$$c_p=\left[1-\frac{\beta\Delta t}{\tau}(P_0-P)\right]^{1/3} \tag{2.20}$$

where β is the isothermal compressibility, P_0 is the target pressure (or external pressure $P_{ex)}$, P is the instantaneous pressure, and τ controls the strength of the pressure coupling. This approach permits the volume of the system to fluctuate by uniformly changing the unit cell size as the internal pressure approaches the external pressure.

Although the weak coupling thermostat and barostat methods are convenient, they generally do not produce structures from the canonical or isothermal-isobaric ensembles. In order to generate the correct ensembles, more accurate extended-system methods introduce additional degrees of freedom to the system in order to mimic its environment. The Nosé-Hoover method [105,106], for example, produces the true canonical (NVT) by reducing the external heat bath to an additional degree of freedom in the system Hamiltonian. The thermal equilibrium process is controlled through choice of the friction coefficient as well as a fictitious mass. In order to obtain the correct isothermal-isobaric ensemble, an analogous method known as the Andersen barostat [107] was developed, in which additional degrees of freedom are used to describe a fictitious pressure piston which allows the volume of the system to fluctuate.

If stochastic, rather than deterministic, dynamics are used in the simulation, Langevin dynamics may be used to control the temperature of a system [108], and the end result is an approximation of the canonical ensemble. As with the extended-system methods discussed above, such as the Nosé-Hoover and Berendsen thermostats, the Langevin equations of motion introduce additional degrees of freedom [109]. Friction due to drag on the solute from the solvent, and random collisions between solvent and solute are added to the systematic force in order to represent the heat bath. Section 2.5.4 discusses the use of Langevin dynamics in simulations of biomolecules.

2.5.4 Stochastic Dynamics: The Langevin Equation

Langevin dynamics [110] aims to describe the motion of a particle subject to Brownian dynamics, which governs the motion of small particles immersed in a fluid. Early investigations into Brownian motion were made on small particles of colloidal size, such as pollen grains and dust particles. The theory, however, has been extended to collective properties of macroscopic systems, as the ensemble of random collisions undergone by the system is able to produce a systematic effect [111]. The Langevin equation approximates two degrees of freedom that are omitted in Newtonian dynamics: namely, random collision of the solute with the solvent, and the frictional drag on the solute as it moves through the solvent. Although the Langevin equations describe the behavior of a solute in a solvent, they do not constitute an implicit model of solvation, as there is no accounting for electrostatic screening or for the hydrophobic effect.

In the Langevin equation, a frictional term and a random force are added to the internal force as follows:

$$M\ddot{\vec{X}}(t) = -\nabla U(\vec{X}(t)) - \gamma M\dot{\vec{X}}(t) + \vec{R}(t) \tag{2.21}$$

where *U(x)* is the potential energy of the solute, $\vec{R}(t)$ is a random force which mimics molecular collisions of the solute with the solvent, and γ is the collision frequency of the solute. The random force is assumed to be independent of the positions, velocities, and forces of the particles on which it acts, and to obey a Gaussian distribution with zero mean:

$$\begin{aligned} &\left\langle R(t) \right\rangle = 0 \\ &\left\langle R(t)R(t')^T \right\rangle = 2\gamma k_B TM\delta(t-t') \end{aligned} \quad . \tag{2.22}$$

Increasing the collision frequency γ has the effect of damping the low-frequency vibrational modes of the protein molecule. Values of γ can be chosen for the system under study using hydrodynamic theory. Stokes' law for a particle of radius *a* and mass *m* describes how the frictional resistance of a spherical particle in solution varies linearly with its radius:

$$\gamma = 6\pi\eta a / m \tag{2.23}$$

where η is the solvent viscosity. Water, for example, has a value of γ equal to 54.9 ps^{-1} at room temperature. Alternatively, the Stokes-Einstein law may be used to choose a value of γ that reproduces experimental values of translational diffusion constants in the diffusive limit. Using this relation, in the diffusive limit the diffusion constant D_t is related to the collision frequency by

$$D_t = k_B T / \sum m\gamma \quad . \qquad (2.24)$$

The value of γ may also be chosen in order to accelerate configurational sampling in simulations, a point that will be revisited in Chapter 3, in which the effects of a novel method combining a self-guiding algorithm with Langevin dynamics on the resulting ensemble and sampling rate are examined.

2.6 Enhancing Sampling

Despite the advancements in computer power that have occurred since the simulation of proteins was first undertaken, it remains challenging to generate correct canonical ensembles of biomolecules simulated at biological temperature. Although simulations on the order of microseconds may now be routinely run [96], the approximations that are made in molecular mechanics force field functional form and parameterization, as well as limits in conformational sampling, limit the depth of new knowledge that may be gained from longer simulations [112]. In order to overcome these problems, many methods have been developed which aim to sample slow events such as large conformational changes and rare events such transitions between two states, to calculate reaction rates, and to increase the volume of conformational space that is available to the system under study. We briefly list here a few examples of these algorithms; the reader is directed to References [113,114] for comprehensive reviews.

Targeted MD [115], accelerated MD [116], and umbrella sampling [117] use restraining potentials in order to bias the system toward a specific state. Although these methods are useful for rapid, preliminary sampling of large-scale conformational changes, the biasing potential must be known in advance, and the resultant dynamics may be nonphysical. Approaches such as metadynamics [118], conformational flooding [119], and adaptively biased molecular dynamics [120] add potential energy terms to the system while the simulation is running which prevent it from revisiting areas of the landscape that have already been sampled. These methods do not require that an estimate of the energy landscape is provided at the outset of the simulation. Additionally, they may be used to calculate free energy as well as to enhance conformational sampling.

Transition path sampling [121] and nudged elastic band methods [122] aim to deduce reaction mechanisms when two stable states of the system are known. Transition path sampling generates an ensemble of probabilistically weighted trajectories between the two points after observing many transitions, while nudged elastic band performs constrained optimization to calculate a single minimum energy pathway between the two known states. Although these methods are very useful in the simulation of rare events, the ruggedness of the potential energy landscape introduces difficulty in locating saddle points, and requires the introduction of reaction coordinates, which are often difficult to determine before the simulation is run.

In order to calculate reaction rates and energies, many variants of Markov state models have been developed for application to protein folding simulations [123,124]. In assuming that a process is Markovian, one assumes that the distribution of future states of the process is only dependent upon the present state, not on the sequence of events which preceded it. In constructing a Markov model, state space is divided into discrete regions, and independent, short trajectories are run which visit each region. The observed transitions between states are then used to construct a matrix of transition probabilities, which may be used to describe the system at long timescales. A major advantage of Markov approaches is that the timescale of the simulations may be shorter than the longest relaxation time of the system. Difficulties arise, however, in the partitioning of state space, and the choice of reaction coordinate is vital to the success of the approach.

One common strategy of enhancing conformational sampling that has been incorporated into many simulation algorithms is to periodically afford the simulated molecule the opportunity to escape from energy minima and to thereby sample a much larger volume of phase space than by conventional methods. This strategy may be accomplished by enhancing low-frequency modes of motion in order to cross potential energy barriers, as in the self-guided Langevin dynamics algorithm [125] described in Chapter 3 of this work. In this formalism, the system is simulated using Langevin dynamics, but an additional force that is determined from averaging over previous time steps is added back into the system. Another strategy of enhancing sampling is an extension of parallel-tempering Monte Carlo to molecular dynamics, in which periodic attempts are made to exchange structures that are simulated at different temperatures. This method, known as replica exchange molecular dynamics [126], is outlined in more detail in Chapter 4 of this work. Although the trajectories obtained with these self-guided Langevin dynamics and replica-exchange molecular dynamics are not deterministic, their algorithms may be developed such that correct ensemble averages are obtained as functions of temperature.

3. Rigorous Evaluation of Thermodynamic Stability and Kinetic Rates of Peptide Folding Using Enhanced Sampling: Application to Self-Guided Langevin Dynamics

Abstract

A complete evaluation of any enhanced sampling method should rigorously address the effect of the method on the kinetic rate of folding and the distribution of structures in the resulting free energy landscape. Although self-guided Langevin dynamics (SGLD) simulations have been suggested as a powerful method for enhancing the conformational search efficiency of molecular dynamics (MD) simulations, the sensitivity to variation of this method’s two key parameters, as well as a quantitative description of the kinetic rates and thermodynamic ensembles resulting from the use of this method, have not yet been well explored.

In this work, 200 ns SGLD simulations of the β-hairpin tryptophan zipper 2 and the trp-cage miniprotein at 300 K were studied in comparison with standard 100 ns Langevin dynamics (LD) and replica exchange molecular dynamics (REMD) simulations at temperatures ranging from approximately 250 K to approximately 500 K. Twelve parameter sets were employed in the SGLD simulations for each system, with variations in the guiding factor and the averaging time used. Forty-eight trajectories were run for each temperature and averaging time in the LD and SGLD simulations, and for each set, native populations, first passage times, and rate constants were determined in order to assess the efficiency of SGLD versus simply raising the temperature of an LD simulation. Comparing the results from SGLD against a series of reference LD simulations at different temperatures, we explore whether the speedup obtained from a high-temperature LD simulation is accompanied by gains or losses in thermodynamic stability. This analysis gives a more complete and rigorous evaluation of the sampling method. The total simulation times for trpzip and trp-cage were 149.3 μs for each system, while the total

simulation time for K19 was 33.6 μs. The LD data produced in this study not only provides a point of comparison for the SGLD results presented here, but also provides thermodynamic and kinetic reference data for future methodological studies of these systems.

Our results indicate significant sensitivity; certain SGLD parameter sets are effective in accelerating the folding process while maintaining populations of native states of the peptides, whereas others are not. Additionally, certain parameter sets result in extreme distortion of the free energy landscape of the test system, requiring ensemble corrections to be used with this method. [127,128] Although trpzip2 and trp-cage are similar sizes, their differing topologies, folding rates, and mechanisms result in the requirement of different averaging times and guiding factors in order to fold efficiently and maintain a folded topology. In order to test the transferability of parameter sets between peptides of similar size but differing structural motifs, the most effective set of parameters from the SGLD simulations of trpzip2 was applied to the α-helix K19. The results obtained indicate that the parameters that are most successful in accelerating folding and maintaining populations of folded states of a β-hairpin are also able to accelerate the kinetics of an α-helix of comparable size, although the thermodynamic stability of the system is greatly decreased relative to the LD reference simulation. Careful choice of parameters must be made with SGLD in order to ensure the kinetic efficiency and thermodynamic stability of the system under study.

3.1 Introduction

One of the greatest obstacles facing molecular dynamics simulations of biopolymers is the size and complexity of the free energy landscape of folding. The presence of local minima on the energy landscape often causes large-scale conformational changes in proteins to occur on a timescale that renders brute force simulations of these rare events unfeasible, particularly when statistical characterization is necessary to obtain kinetic and thermodynamic information about the system rather than anecdotal observations. This limitation has motivated the development of methods that decrease the computational demands of long timescale simulations by improving their efficiency or by speeding slow events in the conformational search process [113,129,130].

One such method that has been developed to enhance the conformational search in a biomolecular simulation is the self-guided molecular dynamics method (SGMD) [129]. The method is known as "self-guided" because information obtained during the simulation is used during that same simulation to enhance the conformational sampling. In SGMD, the total force is defined as the sum of the interaction force $\vec{f}_i$ and the guiding force $\vec{g}_i$ as follows:

$$\vec{\dot{p}} = \vec{f}_i + \vec{g}_i \qquad . \tag{3.1}$$

From the definition of the local average of the property *P* at conformation *n* taken over *L* local conformations,

$$\left\langle P \right\rangle_L [n] = \frac{L-1}{L}\left\langle P \right\rangle_L [n-1] + \frac{1}{L}P[n] \tag{3.2}$$

the guiding force is calculated as the local average of the nonbonded forces

$$\vec{g}_i(t) = \lambda\left\langle \vec{f}_i(t) + \lambda\vec{g}_i(t-\delta t)\right\rangle_L = \left(1-\frac{\delta t}{t_L}\right)\vec{g}_i(t-\delta t) + \frac{\delta t}{t_L}\lambda(\vec{f}_i(t) + \vec{g}_i(t-\delta t)) \tag{3.3}$$

where λ is the guiding factor, δt is the time step, and t_L=$L\delta t$ is the time over which the local average is taken. SGMD simulations thus employ an equation of motion that uses the guiding force, calculated as a local average of the total instantaneous force, to accelerate the systematic motion defined by the local averaging time. Although this method has demonstrated efficiency in enhancing conformational search [130,131,132], it was found that the use of a guiding force may result in an altered conformational distribution, and that the inclusion of high-frequency bonded interactions in the guiding force calculation led to excessive noise. Additionally, this method was found to be insufficient in enhancing the conformational search in stochastic dynamics simulations [133].

These drawbacks to SGMD resulted in the development of self-guided Langevin dynamics (SGLD) [125], which uses the local average of momenta, rather than the forces, to calculate the guiding force as well as to enhance conformational sampling efficiency. As described in Chapter 2 of this work, Langevin dynamics has been utilized extensively in molecular simulations in order to mimic the random collisions of solute and solvent, and as a scheme for controlling the temperature [110,134]. Direct application of a guiding force based on momentum is problematic in a standard MD simulation, as it has the ability to cause an uneven distribution of kinetic energy through the system. In Langevin dynamics, each degree of freedom is independently coupled with a heat bath, which allows for the use of a momentum-based guiding force without compromising the energetic integrity of the system. The implementation of SGLD employed in this work follows that outlined in the original reference of Wu and Brooks [125].

For an *N* particle system, the equation of motion employed in an SGLD simulation is the sum of the Langevin equation of motion and a guiding force $\vec{g}_i$ as follows:

$$\dot{\vec{p}} = \vec{f}_i - \gamma_i \vec{p}_i + \vec{R}_i + \lambda_i \vec{g}_i \quad . \tag{3.4}$$

Here, γ_i is the collision frequency, $\vec{R}_i$ is a random force, and λ_i is the guiding factor for atom *i*. In an SGLD simulation, three parameters are used to define the self-guiding effect. The local averaging time, t_L, is the time period over which the guiding factor λ is calculated as an average of instantaneous forces. The local averaging time determines which of the slow motions are to be enhanced; because low-frequency motions of a protein occur on a long time period, it is intended that the guiding force is used to describe the systematic motion over a long time scale. It has been shown that the time derivative of the momentum may be dropped from Equation 3.4 for long-time dynamics [135,136], allowing the equation of motion to be rewritten as follows:

$$\gamma_i \dot{\vec{p}}_i = \vec{f}_i + \vec{R}_i + \lambda_i \vec{g}_i \quad . \tag{3.5}$$

As in Equation 3.3, the guiding force is calculated as a local average over *L* local conformations of the total force given in Equation 3.5

$$\vec{g}_i = \left\langle \vec{f}_i + \vec{R}_i + \lambda_i \vec{g}_i \right\rangle_L = \gamma_i \left\langle \vec{p}_i \right\rangle_L \quad . \tag{3.6}$$

The friction force on atom *i* is $\gamma_i \vec{p}_i$, indicating by Equation 3.6 that the guiding force $\vec{g}_i$ is the local average of the friction force. Through the guiding forces, extra energy is introduced into the system in the form

$$\sum_i^N \lambda_i \vec{g}_i \dot{\vec{r}}_i \tag{3.7}$$

where $\dot{\vec{r}}_i = \vec{p}_i / m$ is the velocity of atom *i*.

It is therefore necessary to add a constraint term ξ to the equation of motion to cancel the extra energy, as follows:

$$\dot{\vec{p}} = \vec{f}_i - \gamma_i \vec{p}_i + \vec{R}_i + \lambda_i \vec{g}_i - \zeta \vec{p}_i \tag{3.8}$$

where

$$\zeta = \frac{\sum_i^N \lambda_i \vec{g}_i \dot{\vec{r}}_i}{\sum_i^N \vec{p}_i \dot{\vec{r}}_i} \quad . \tag{3.9}$$

Having established the background and the equation of motion of the system, we now turn to a discussion of the simulation algorithm, which follows that of Brunger et al. [137] with the addition of Equation 3.9.

3.1.1 Simulation Algorithm

As the first step in the SGLD algorithm, forces (interaction, random, and guiding) are calculated at the initial time step. The interaction forces are those described by the force field, the random forces are assigned from a Gaussian distribution with zero mean, and the guiding forces are the local average of the friction forces calculated from simulation steps over the averaging time as follows:

$$\vec{g}_i(t) = \left(1-\frac{\delta t}{t_L}\right)\vec{g}_i(t-\delta t) + \frac{\delta t}{t_L}\gamma_i m_i \dot{\vec{r}}_i\left(t-\frac{\delta t}{2}\right) \qquad . \tag{3.10}$$

The value of the guiding force is initially set to zero. The scaling parameter χ_i is then determined through making an unconstrained half step in which the value of the constraint parameter is estimated from the unconstrained velocity and the value of the scaling parameter χ_i is calculated from the value of the constraint parameter:

$$\dot{\vec{r}}_i^{\,'}(t) = \dot{\vec{r}}_i\left(t-\frac{\delta t}{2}\right) + \frac{\delta t}{2m_i}\left(\vec{f}_i(t)+\lambda_i\vec{g}_i(t)+\vec{R}_i(t)\right) \tag{3.11}$$

$$\zeta = \frac{\sum_i^N \lambda_i \vec{g}_i(t)\dot{\vec{r}}_i^{\,'}(t)\left(1+\frac{\gamma_i\delta t}{2}\right)^{-1}}{\sum_i^N m_i \dot{\vec{r}}_i^{\,'2}(t)\left(1+\frac{\gamma_i\delta t}{2}\right)^{-2}} \tag{3.12}$$

$$\chi_i = \left(1+\frac{(\gamma_i+\zeta)\delta t}{2}\right)^{-1} \qquad . \tag{3.13}$$

The variable $\vec{\dot{r}}'(t)$ is the unconstrained velocity of atom *i* at time *t*. Finally, velocities are advanced to the next half-timestep

$$\vec{\dot{r}}_i\left(t+\frac{\delta t}{2}\right)=(2\chi_i-1)\vec{\dot{r}}_i\left(t-\frac{\delta t}{2}\right)+\chi_i\frac{\delta t}{m_i}(\vec{\dot{f}}(t)+\lambda_i\vec{\dot{g}}_i(t)+\vec{R}_i(t)) \tag{3.14}$$

and positions are advanced to the next time step:

$$\vec{\dot{r}}_i(t+\delta t)=\vec{\dot{r}}_i(t)+\vec{\dot{r}}\left(t+\frac{\delta t}{2}\right)\delta t \tag{3.15}$$

Coordinates are constrained using the algorithm of choice, such as SHAKE. Time steps are propagated by iterating from Equation 3.10, with a time value of *t*+*δt*.

Prior studies on SGMD [129,130,131,132,138] and SGLD [125,139,140,141,142] have compared the efficiency of these types of simulation to regular Langevin dynamics simulations for peptides and proteins of various conformations. Although reports have been made of the ability of SGLD to accelerate slow conformational changes, these studies have not quantified the effect of SGLD on these rates of transition, such as through analysis of first passage times and relaxation rates versus standard LD simulation. The choice of parameter sets in prior studies employing SGLD has been limited to a narrow range of averaging times and guiding factors. In their study of a 16-amino acid helical peptide, Wu and Brooks [125] tested an averaging time of 0.1 ps with guiding factors of 0.25 and 1.0. Study of structural relaxation of staphylococcal nuclease using SGLD [140,141] employed averaging times of 0.1 ps and 0.5 ps, with guiding factors of 0.25 and 1.0. All of these parameter sets have been anecdotally reported to be successful in accelerating slow conformational changes through analysis of a limited number of trajectories obtained from relatively short simulations.

It is the goal of the present study to systematically quantify the effect of SGLD on the thermodynamics and kinetics of folding using a statistically significant ensemble of long-time simulations and a broader range of parameter sets than has previously been explored. We hope to identify trends in successful, as well as unsuccessful, combinations of averaging times and guiding factors to determine which parameter sets have the ability to accelerate the kinetics of folding while maintaining reasonably accurate thermodynamic properties. We have also applied the most efficient parameter sets from our study of β-hairpin trpzip2 to α-helix K19 in order to test the transferability of parameters from one structural family of peptide to another of similar size but differing topology.

3.2 Methods

3.2.1 Model System: Trpzip2

The first model system chosen for study was the tryptophan zipper (trpzip). First developed by Starovasnik and coworkers [143], this β-hairpin structural motif is stabilized through cross-strand tryptophan pairs. Within the trpzip family of structures, trpzip2 (SWTWENGKWTWK, with a type I' β-turn at NG), has the most cooperative melting curve and highest stability (approximately 90% at 300K). Thermodynamic properties for this peptide have been determined by NMR and CD spectroscopy, and a family of structures was refined using restraints from NMR experiments [143] (PDB code 1LE1). In the simulations described below, the N-terminal of the peptide was acetlyated and the C-terminal was amidated, in accordance with experiment [143].

3.2.2 Model System: Trp-cage

The second peptide studied in this work is the trp-cage miniprotein, which was designed for use as a model system in protein folding studies [72]. Derived from the C-terminal fragments of the 39-residue exendin-4 peptide and made increasingly stable with the introduction of a solvent-exposed salt bridge and helical N-capping residues, trp-cage stably incorporates several elements of protein secondary structure. An α-helix extends from residues 2-9, a 3_{10} helix runs from residues 11-14, and a C-terminal polyproline II helix packs against the central tryptophan residue to form a highly stable hydrophobic core. The simulations in this study used the trp-cage TC5b amino acid sequence (NLYIQWLKDGGPSSGRPPPS), the structure of which was determined via NMR (PDB code 1L2Y) [72]. The high stability and fast folding rate of trp-cage make it an ideal model system for protein folding studies, and its thermodynamic and kinetic properties have been extensively studied via both experiment and simulation [144,145,146,147,148].

3.2.3 Model System: Helix K19

The third model system chosen for study was the peptide K19, which is an α-helix with sequence AcGGG-$(KAAAA)_3$-K-NH_2. Although the α-helical conformation has been experimentally shown to be unfavorable for short polyalanine peptides in water at room temperature [149,150,151], it is believed that helicity is favored by either the inclusion of polar side chains in the sequence [149,152,153], or the presence of positive charges at the C-terminus interacting with the helix macrodipole [71]. K19 has been synthesized, and CD and NMR have been used to characterize the fractional helicity of each residue of the peptide [154]. These experiments were complemented by computational studies in which the residue-specific fractional helicity of the peptide was determined, as well as its melting temperature. Additionally, the steps in helix formation of this peptide, as well as the residue contacts that maintain helical structure, were determined [154].

3.2.4 Langevin Dynamics Simulation

For trpzip2, 48 independent 100 ns LD trajectories were obtained at each of the following temperatures: 300 K, 325 K, 350 K, 363 K, and 375 K. Version 10 of the AMBER molecular dynamics package was used for these simulations [65]. A version of the ff99 force field with modified backbone parameters to reduce α-helical bias [76] was employed. All nonbonded interactions were evaluated at each time step and SHAKE was used to constrain all bond lengths. The time step used was 2 fs, and systems were maintained at constant temperature. The collision frequency used was 1.0 ps^{-1}. All simulations used the Generalized Born (GB) implicit solvent model [155] with GB^{HCT} implementation [90] in AMBER. This protocol was chosen to match that used in a previous study, in which it was effective in quantitatively reproducing the structure of trpzip2, as well as its temperature-dependent stability [156]. Forty-eight random starting structures were selected from a trajectory obtained from a standard high-temperature LD simulation at 400 K.

For trp-cage, 48 independent 100 ns LD trajectories were obtained at each of the following temperatures: 250 K, 285 K, 300 K, 320 K, and 355 K. All simulations used the Generalized Born (GB) implicit solvent model with GB^{OBC} implementation [157] in AMBER. All other simulation parameters were identical to those described above for trpzip2. This protocol was chosen to match that used in a previous unpublished study which quantitatively reproduced the structure and temperature-dependent stability of trp-cage. Forty-eight random initial structures were selected from a REMD simulation at 400 K. For the helix K19, 48 independent 100 ns Langevin dynamics trajectories were obtained at each 280 K and 300 K. Once the LD simulations were initiated, all nonbonded interactions were evaluated at each time step and SHAKE was used to constrain bonds to hydrogen atoms. The time step used was 2 fs, and systems were maintained at constant temperature. The collision frequency used was 1.0 ps^{-1}. All simulations used the Generalized Born (GB) implicit solvent model [155] with GB^{OBC} implementation [89,91]. The ff99SB forcefield [76] was employed. This protocol has previously been reported to be effective in the reproduction of experimental determination of the structure and folding process of K19, as well as its melting temperature [154]. Extended starting structures of K19 were built using the leap module included in the AMBER 10 package.

3.2.5 Replica Exchange Molecular Dynamics Simulation

In the REMD simulation of trpzip2, fourteen replicas covering a temperature range of 251.7-554.7 K were used. Exchanges between neighboring replicas were attempted at intervals of 1 ps. The REMD simulation was run to 100,000 exchange attempts, for a total of 100 ns per replica. Other parameters were as described above for the LD simulations of trpzip2. In the REMD simulation of trp-cage, fourteen replicas covering a temperature range of 251.5-540.2 K were used. Exchanges between neighboring replicas were attempted at intervals of 1 ps. The REMD simulation was run to 80,000 exchange attempts, for a total of 80 ns per replica. Other parameters were as described above for the LD simulations of trp-cage.

3.2.6 Self-Guided Langevin Dynamics Simulation

For each of the 12 SGLD parameter sets used in simulation of trpzip2 and trp-cage, 48 independent trajectories of 200 ns were generated at 300 K. The sander module of AMBER [65], which includes an SGLD implementation, was employed. Initial structures and parameters were as described above for the LD simulations, with the addition of averaging time and self-guiding force parameters. Four averaging times (0.2 ps, 1.0 ps, 2.0 ps, and 10.0 ps) were employed. For each averaging time, three self-guiding factors were used: 1.0, 5.0, and 10.0 (see

Table 3.1 for nomenclature of parameter sets employed in SGLD simulations of trpzip2 and trp-cage). For each system, 48 trajectories were obtained for each of the 12 parameter sets.

For the helical system K19, LD simulations were run at 280 K and 300 K, and three SGLD parameter sets were tested: an averaging time of 2.0 ps was used with self-guiding factors of 1.0, 5.0, and 10.0. These parameter combinations correspond to sets SGLD 3, SGLD 3a, and SGLD 3b, as indicated in Table 3.2. Due to the necessity of a small guiding factor to produce correctly folded states (discussed below), two additional parameter sets with guiding factors of 0.25 and 0.5 (SGLD 3c and SGLD 3d, respectively) were tested. All other parameters were as discussed above for the Langevin dynamics simulations of K19. The same starting structures were used for the SGLD simulations as in the LD simulations. Forty-eight independent trajectories of 100 ns in length were obtained for each temperature of LD, and forty-eight independent trajectories of 200 ns in length were obtained for each SGLD parameter set.

Table 3.1: Parameter sets used in self-guided Langevin dynamics simulations of trpzip2 and trp-cage.

Parameter Set	Averaging Time (ps)	Guiding Factor
SGLD 1	0.2	1.0
SGLD 1a	0.2	5.0
SGLD 1b	0.2	10.0
SGLD 2	1.0	1.0
SGLD 2a	1.0	5.0
SGLD 2b	1.0	10.0
SGLD 3	2.0	1.0
SGLD 3a	2.0	5.0
SGLD 3b	2.0	10.0
SGLD 3c	2.0	20.0
SGLD 4	10.0	1.0
SGLD 4a	10.0	5.0
SGLD 4b	10.0	10.0

Table 3.2: Parameter sets used in self-guided Langevin dynamics simulations of K19.

Parameter Set	**Averaging Time (ps)**	**Guiding Factor**
SGLD 3	2.0	1.0
SGLD 3a	2.0	5.0
SGLD 3b	2.0	10.0
SGLD 3c	2.0	0.25
SGLD 3d	2.0	0.5

3.2.7 Analysis

The trajectories obtained from the LD, REMD, and SGLD simulations were analyzed using the AMBER ptraj module [65]. For trpzip2, the root-mean-square deviation (RMSD) of the backbone atoms of residues 2-11 of trpzip2 from the experimentally determined native structure (model 1 of PDB code 1LE1) [143] was calculated. Terminal residues were omitted to remove the effects of fluctuations. An RMSD cutoff of 1.7 Å was used to determine native structures on the basis of the free-energy profile along RMSD where the native minimum reached up to 1.7 Å (data not shown). For trp-cage, the RMSD of the backbone atoms of residues 3-18 from the native structure as determined by NMR (model 1 of PDB code 1L2Y) [72] was calculated. An RMSD cutoff of 2.5 Å was used to define the native state of trp-cage. The choices of RMSD values used in this analysis for trpzip2 and trpcage have previously been shown to be effective in producing a description of the folding that agrees with experiment [144,156,158].

Using these cutoffs, the time evolution of the average fraction of native content was determined for each parameter set for each system. Error bounds on the precision of this measurement were calculated by treating each set of 48 trajectories as two independent sets of 24 trajectories. For both systems, the stability of each SGLD parameter set relative to LD was quantified by calculating the value of ΔG using the ratio of the average population of native structure in the SGLD to the average population of native structure from the REMD trajectory at 300K. For trpzip2 and trp-cage, melting curves were generated by calculating, across each of the 48 LD trajectories and for each of the REMD temperature trajectories, the average population of native structure for each simulated temperature. The fraction of native population as a function of time was calculated by averaging the native population at each time point across the 48 LD runs using the criterion of RMSD to the native structure.

To determine the first passage time of folding of trpzip2, the first time step in each of the 48 trajectories at which the instantaneous backbone RMSD for residues 2-11 fell below 1.7 Å was extracted. For trp-cage, the RMSD cutoff of 2.5 Å on residues 3-18 was used. For both systems, the distribution of the first passage times of folding for the 48 runs was fit to single and double-exponential equations in order to quantify the relaxation time. In order to determine unfolding rates of trpzip2, first passage times out of the folded basin were extracted when the backbone RMSD of residues 2-11 rose above 3.0 Å, subsequent to the first folding event. For trp-cage, a cutoff of 6.0 Å for residues 3-18 was used to determine the rate of unfolding, subsequent to the first folding event. The use of these cutoffs ensures that structures have fully unfolded, and are not fluctuating around the native conformation.

In order to quantify the rate and diversity of structural sampling, cluster analysis was performed on the backbone of residues 2-11 of trpzip2 with a cutoff of 1.7 Å using MOIL-View [159]. All 48 trajectories for each temperature or SGLD parameter set were combined, and clusters were formed with the bottom-up approach. Using this algorithm, each structure is initially assigned to its own cluster. RMSD values between all pairs of clusters are calculated, and the cluster pair with the lowest RMSD is merged into a single cluster. This procedure is repeated until the remaining pair of clusters has an RMSD that is less than that of the similarity cutoff. The population of each cluster is then determined.

Multi-dimensional population histograms were used to obtain the free energy as a function of the radius of gyration and the backbone RMSD to the native structure. The free energy values obtained by this analysis were calculated relative to the most populated histogram bin. Potentials of mean force were then plotted in order to visualize the areas of the conformational landscape that were explored in each of the simulations.

For the helix K19, DSSP [160] (as implemented in the ptraj module of AMBER) was used to quantify the fractional helicity of each residue in the peptide. The average helicity across the non-terminal residues of the peptide was calculated, and the error in the precision of this measurement was determined by treating the set of 48 trajectories as two independent sets of 24 trajectories. The average fractional helicity across all 48 trajectories was calculated at each time step and subsequently plotted as a function of time.

3.3. Results and Discussion

3.3.1 Langevin Dynamics Simulations of Trpzip2 and Trp-cage

Independent 100 ns Langevin dynamics simulations were performed on trpzip2 at each of the following temperatures: 300 K, 325 K, 350 K, 363 K, and 375 K. Forty-eight trajectories were obtained for each of the five temperatures, for a total of 24 μs simulation time. The thermodynamic stability, first passage time, folding rate constant, and rate of structural sampling were determined for the set of trajectories at each temperature. Because 100 ns is too short a simulation time to converge the folding thermodynamics at low temperature using LD (Figure 3.1), the thermodynamic stability of the SGLD and LD simulations was compared to converged data from a replica exchange simulation of trpzip2 that employed the same parameters as those used in the LD simulations (Table 3.3). The data obtained from the REMD simulations thus serves as a reference against which the thermodynamic stability of the SGLD simulations may be compared, while data from the LD runs serve as the benchmark in determining the kinetic efficiency of the SGLD simulations.

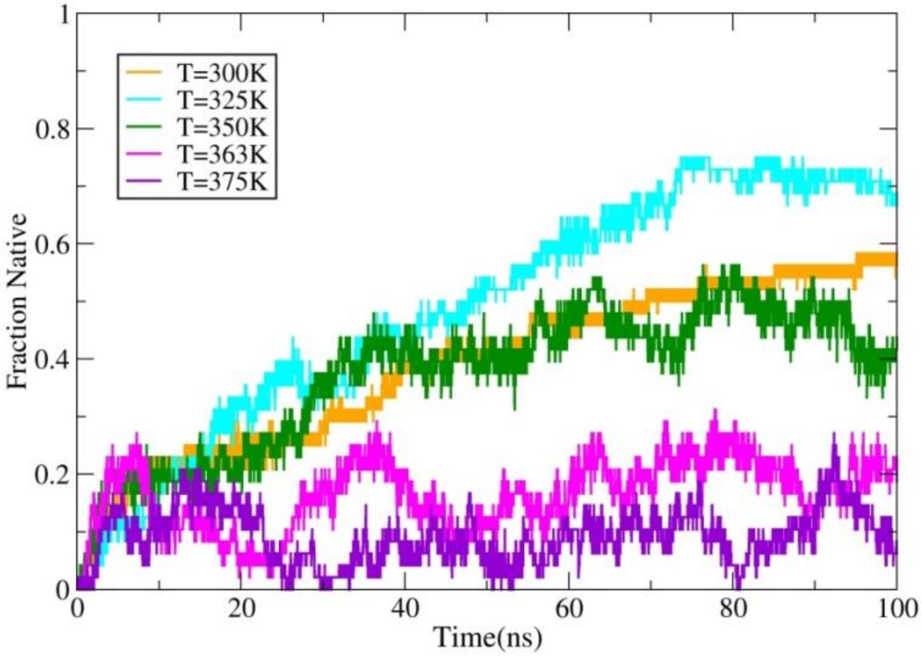

Figure 3.1**:** Time-dependent average fraction native of the 48 trajectories for 100 ns LD simulations of trpzip2 at temperatures ranging from 300-375 K.

Table 3.3: Thermodynamic and kinetic data from REMD and LD simulations of trpzip2. Column 2 lists the fraction of native structure obtained for each temperature trajectory at the end of the 100 ns reference REMD simulation. For the 100 ns LD simulations, column 4 lists the average fraction of the 48 simulations that are in the native state during the 100 ns simulation time, with their associated error bounds. Column 5 is the value of ΔG for each of the LD simulations versus the REMD simulation at 300K. Columns 6 and 7 are the relaxation times of folding and unfolding obtained from a single-exponential fit of first passage times of folding and for escape from the native basin, respectively. Column 8 lists the number of clusters found by each set of trajectories at the end of the 100 ns LD simulation.

Parameter Set	**Fraction Folded**	**Parameter Set**	**Fraction Folded**	ΔG (kcal/mol)	**Folding t_{relax} (ns)**	**Unfolding t_{relax} (ns)**	**Number of Clusters**
REMD 300K	99%	LD 300K	55±3%	0.4	75.4	n/a	395
REMD 327.5K	87%	LD 325K	71±4%	0.2	31.4	37.6	482
REMD 350K	55%	LD 350K	46±2%	0.5	16.2	16.8	792
REMD 360.5K	34%	LD 363K	22±1%	1.1	10.6	3.9	1192
REMD 373K	17%	LD 375K	10±1%	1.7	11.3	2.3	1386

First passage times for folding were fit to a single-exponential equation in order to estimate the relaxation times of folding, and Table 3.3 lists the relaxation times and pre-exponential factors obtained from the fit. The relaxation time of folding is seen to decrease by a factor of approximately 7 as the temperature increases from 300 K-363 K. The relaxation time increases slightly from 363 K to 375 K; this non-Arrhenius behavior of the system at higher temperatures is likely due to the entropic barrier to folding caused by an increase in the number of states available to the system [161]. Table 3.3 also lists the relaxation time of unfolding out of the native basin, which is seen to decrease with temperature, from 37.6 ns at 325 K to 2.3 ns at 375 K. A relaxation time of unfolding for the 100 ns LD simulation at 300 K was not calculated because trajectories that do fold at that temperature tend to become trapped, and do not exhibit a subsequent unfolding transition; at 300 K, the relaxation time for unfolding is longer than the length of the simulation. At temperatures of 363 K and 375 K, the relaxation time of unfolding is faster than either of the two phases of folding, which results in relatively low populations of folded structures.

The extent and rate of sampling of structures was determined using cluster analysis, which has been established as an effective method of quantifying the convergence of conformational diversity within a trajectory [162]. By determining the total number of clusters sampled over the course of each simulation, we may compare the extent to which each of the simulations sampled diverse structures on the energy landscape. Table 3.3 lists the number of clusters sampled in each of the LD simulations, and Figure 3.2 shows the rate at which new clusters were found for each of the LD trajectories of trpzip2. The number of clusters is seen to increase as the temperature of the simulation increases, from 395 clusters at 300 K to 1386 clusters at 375 K. Sampling above the melting point of the peptide appears to increase the number of conformations that are accessible to the peptide during the timescale of the simulation.

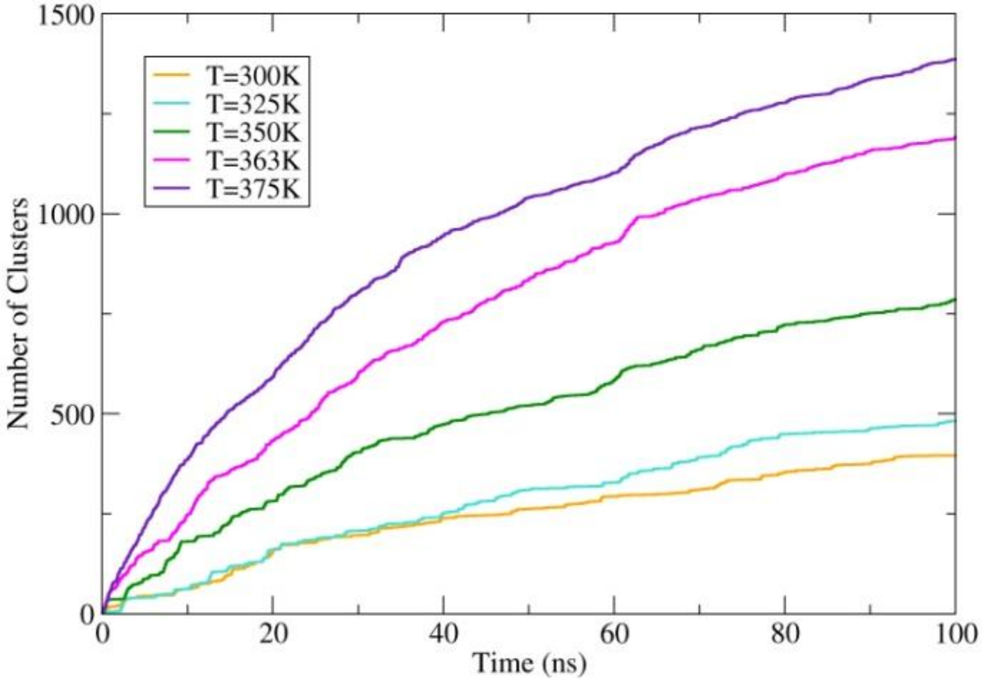

Figure 3.2**:** Number of clusters identified versus time for 100 ns LD simulations of trpzip2.

As with trpzip2, reference data for the trp-cage SGLD simulations was obtained using standard LD and REMD simulations. The data obtained from the LD runs served as the benchmark in determining the kinetic efficiency of the SGLD simulations, while that obtained from the REMD simulations served as a benchmark against which the thermodynamic stability of the SGLD simulations could be compared. Independent 100 ns Langevin dynamics simulations were performed on trp-cage at each of the following temperatures: 250 K, 285 K, 300 K, 320 K, and 355 K. Forty-eight trajectories were obtained for each of the five temperatures, for a total of 24 μs simulation time. As with the trpzip2 runs, the thermodynamic stability of the SGLD and LD simulations of trp-cage was compared to converged data from REMD simulations (Table 3.4) due to the inability of the 100 ns LD simulations to produce accurate ensembles (see Figure 3.3). Table 3.4 also contains the kinetic data for trp-cage, which is compared against the standard LD simulations. The relaxation time of folding is seen to decrease by a factor of approximately 10 as the temperature is increased from 250 K to 285 K. Increasing the temperature from 285 K to 355 K results in a further decrease in the folding relaxation time by a factor of approximately three. The relaxation time of unfolding was not obtained for the LD simulations at 250 K or 285 K, as trajectories at those temperatures which underwent folding at some point in the 100 ns of simulation time did not subsequently unfold. The relaxation time of unfolding at the higher temperatures is seen to decrease by a factor of approximately 40 as the temperature is increased from 300 K to 355 K. The relaxation time of unfolding is approximately seven times faster than that of folding, which contributes to the lower fraction of native population at high temperature.

Table 3.4: Thermodynamic and kinetic data from REMD and LD simulations of trp-cage. Column 2 lists the average fraction of native structure obtained for each temperature trajectory over the 100 ns reference REMD simulation. For the 100 ns LD simulations, column 4 lists the average fraction of the 48 simulations that are in the native state during the 100 ns simulation time, with their associated error bounds. Column 5 is the value of ΔG for each of the LD simulations versus the REMD simulation at 300K. Columns 6 and 7 are the relaxation times of folding and unfolding obtained from a single-exponential fit of first passage times of folding and for escape from the native basin, respectively. Column 8 lists the number of clusters found by each set of trajectories at the end of the 100 ns LD simulation.

Parameter Set	Fraction Folded	Parameter Set	Fraction Folded	ΔG (kcal/mol)	Folding t_{relax} (ns)	Unfolding t_{relax} (ns)	Number of Clusters
REMD 251.5K	85%	LD 250K	23±2%	0.4	129.1	n/a	216
REMD 282.9K	73%	LD 285K	58±3%	0.01	13.4	n/a	299
REMD 300K	60%	LD 300K	38±4%	0.3	13.9	81.2	622
REMD 318.2K	44%	LD 320K	46±5%	0.2	6.0	12.8	1068
REMD 357.9K	9.7%	LD 355K	11±3%	1.3	4.3	2.0	3156

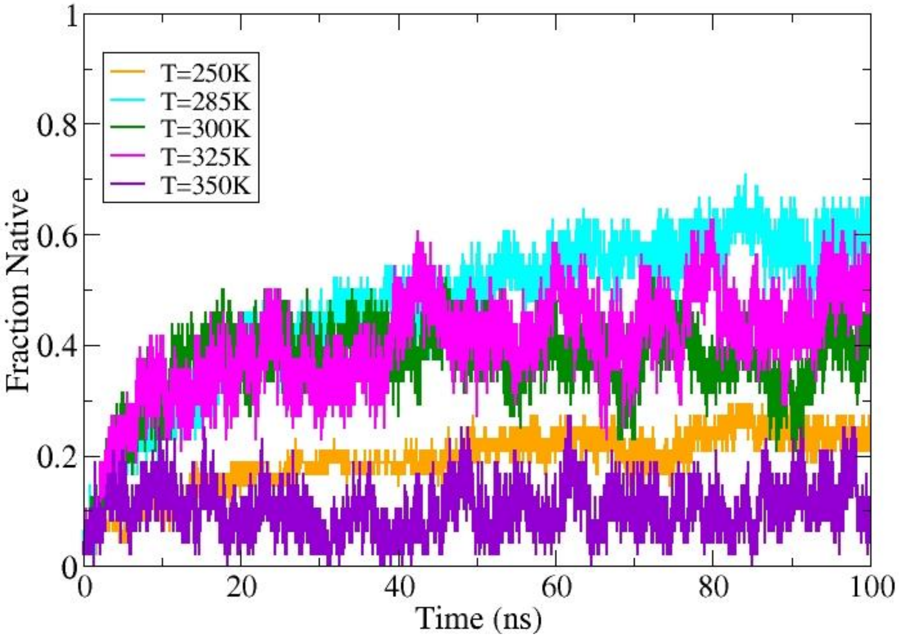

Figure 3.3: Time-dependent average fraction native of the 48 trajectories for 100 ns LD simulations of trp-cage at temperatures ranging from 250-355 K.

By determining the total number of clusters sampled over the course of each simulation, as well as the rate at which new clusters were sampled, we may compare the efficiency with which each of the simulations sampled diverse structures on the energy landscape. Table 3.4 lists the number of clusters sampled in each of the LD simulations of trp-cage, and Figure 3.4 shows the rate at which new clusters were sampled. The number of clusters sampled, as well as the rate of sampling of new clusters, are both seen to increase with temperature. The LD simulations at 250 K and 285 K sample the smallest number of clusters, with 216 and 299 clusters found, respectively. Trajectories at 300 K sample 622 clusters, those at 320 K also sample 1068 clusters, and those at 355 K sample 3156 clusters. As seen with trpzip2, sampling above the melting point of the peptide appears to increase the number of conformations that are accessible to the peptide during the time scale of the simulation.

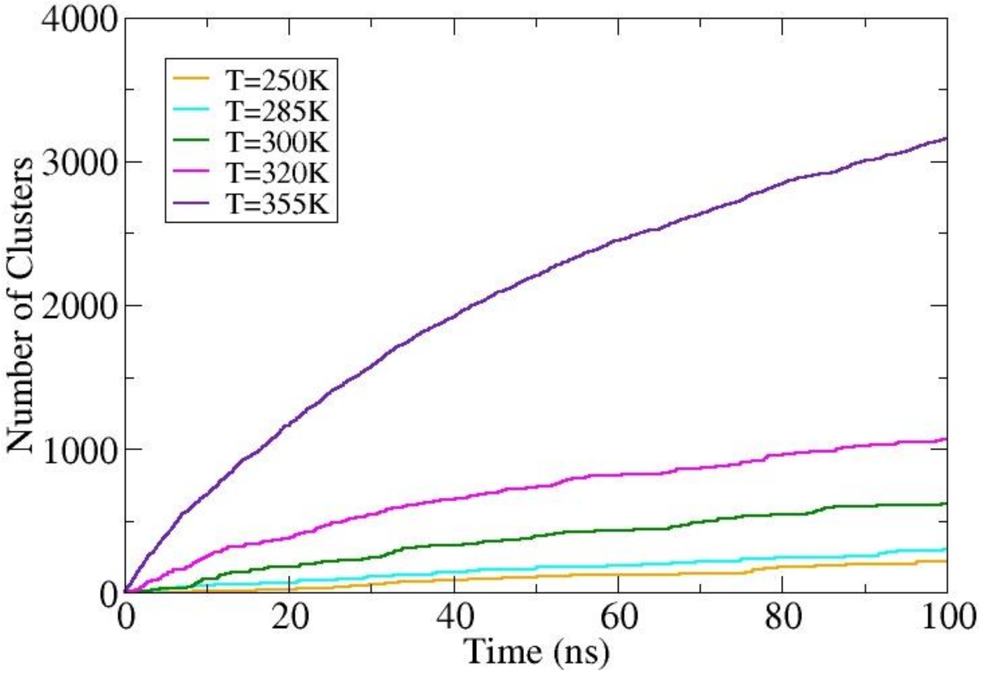

Figure 3.4: Number of clusters identified versus time for 100 ns LD simulations of trp-cage.

3.3.2 Self-guided Langevin Dynamics Simulations of Trpzip2 and Trp-cage

The standard REMD and LD simulations described in the previous section serve as a benchmark against which we can estimate the effective thermodynamic temperature of an SGLD simulation and judge the efficiency of performing an SGLD simulation versus performing a standard LD or REMD simulation at high temperature. In this section, we describe the results obtained from these SGLD simulations. We generated 48 independent 200 ns SGLD simulations of trpzip2 and trp-cage at 300 K for each of the thirteen sets of averaging times and guiding factors listed in Table 3.1, for a total of 124.8 μs simulation time per peptide. For each parameter set, we determined the thermodynamic stability and free energy surface, first passage time distribution, relaxation times of folding and unfolding, and time dependence of structural sampling. Through comparison of these results with those of the reference REMD and LD simulations presented above, we may identify which, if any, of the SGLD parameter sets are effective in accelerating the kinetics of folding while maintaining reasonably accurate thermodynamics.

Table 3.5 includes, for each SGLD parameter set, the averaging time, guiding factor, average fraction of trajectories in the folded state during the last 40 ns of the 200 ns simulation of trpzip2, and the fraction of trajectories that reach the native state at least once over the course of the simulation. Table 3.5 also includes the relaxation times of folding and unfolding obtained from the single-exponential fits of the first passage time data, and a comparison of these relaxation times against the reference LD simulation at 300 K.

Table 3.5: Thermodynamic and kinetic data from 200 ns SGLD simulations of trpzip2. Column 4 lists the average fraction of native structure obtained for each SGLD parameter set, with their associated error bounds. Column 5 is the value of ΔG for each of the SGLD parameter sets relative to the REMD reference trajectory at 300K. Column 6 lists the relaxation time of folding obtained from the single-exponential fit of first passage times, and column 7 lists the relaxation time for escape from the native basin. Column 8 is the value of the relaxation time of folding for each LD parameter set divided by that for the SGLD simulation at 300K (speedup), and column 9 gives the number of clusters found by each parameter set at the end of the 200 ns simulation time.

Parameter Set	Averaging Time (ps)	Guiding Factor	Fraction Folded	ΔG (kcal/ mol)	Folding t_{relax} (ns)	Unfolding t_{relax} (ns)	SGLD Speedup	Number of Clusters
SGLD 1	0.2	1.0	81±6%	0.1	14.9	61.8	5.1	471
SGLD 2	1.0	1.0	82±3%	0.1	39.9	69.4	1.9	364
SGLD 3	2.0	1.0	76±7%	0.2	26.9	46.3	2.8	374
SGLD 4	10.0	1.0	82±5%	0.1	33.6	89.9	2.2	333
SGLD 1a	0.2	5.0	1±0%	n/a	12.6	0.5	6.0	3340
SGLD 2a	1.0	5.0	70±3%	0.2	13.7	43.4	5.5	851
SGLD 3a	2.0	5.0	81±2%	0.1	23.6	67.9	3.2	486
SGLD 4a	10.0	5.0	82±2%	0.1	32.0	>200	2.4	389
SGLD 1b	0.2	10.0	0±0%	n/a	4.1	0.2	18.4	4324
SGLD 2b	1.0	10.0	36±2%	0.6	10.4	19.4	7.3	1764
SGLD 3b	2.0	10.0	84±6%	0.1	13.5	62.0	5.6	728
SGLD 4b	10.0	10.0	77±4%	0.2	34.2	44.7	2.2	429
SGLD 3c	2.0	20.0	77±4%	0.2	14.0	41.0	5.4	1386

Examining Table 3.5, it is evident that with all tested averaging times, the use of a guiding factor of 1.0 (parameter sets SGLD 1, SGLD 2, SGLD 3, and SGLD 4) with trpzip2 results in relatively large populations of native structure, ranging from approximately 76% to approximately 82%. These populations are significantly larger than that obtained using LD at 300 K, which achieves a population of approximately 55% but is unconverged. The fraction of folded structures is less than that obtained from the converged REMD trajectory at 300 K, which achieves a population of 99% native structure. A comparison of the relaxation times of folding and unfolding for each of these parameter sets to those of the reference LD simulation at 300 K allows for a determination of the kinetically efficient SGLD parameter sets. All trajectories with a guiding factor of 1.0 fold more quickly than the LD simulation. Parameter sets SGLD 1, SGLD 2, SGLD 3, and SGLD 4 exhibit folding rates that are 5.1, 1.9, 2.8, and 2.2 times faster than the LD simulation at 300 K, respectively. These results indicate that this relatively small guiding factor is successful at both accelerating folding and maintaining populations of native structures when the averaging time ranges from 0.2 ps to 10.0 ps. Among these parameter sets, SGLD 1 has the largest kinetic rate acceleration versus the LD reference simulation (5.1 times that of LD at 300 K) and also has a relatively long relaxation time of unfolding, which results in its large fraction of native population.

With a larger guiding factor of 5.0 (SGLD 1a, 2a, 3a, and 4a), the populations of native structures, as well as the relaxation times of folding and unfolding, are seen to decrease as the averaging time increases from 0.2 ps to 10.0 ps. With the exception of parameter set SGLD 1a, the fraction of folded structures obtained by these parameter sets is approximately equal to those obtained by the set with a guiding factor of 1.0 (SGLD 1, 2, 3, and 4), while the folding rates are slightly faster than these sets. Parameter sets SGLD 3a and 4a have the largest populations of native structure (approximately 81%), however SGLD 3a has the fastest relaxation time of folding. The slow relaxation time of unfolding for SGLD 3a likely contributes to the relatively large population of native structure that is obtained using this parameter set.

Increasing the guiding factor to 10.0, the populations of folded structures range from 0% to approximately 84% as the averaging time is increased from 0.2 ps to 2.0 ps. Increasing the averaging time from 2.0 ps to 10.0 ps leads to a decrease in folded population to approximately 77%. Parameter set SGLD 3b is the most kinetically efficient and thermodynamically stable, achieving a fraction of native structure of approximately 84%, which is comparable to the REMD simulation at 327.5 K. The folding first passage time of this parameter set is 13.5 ns, which is comparable to that of an LD simulation run between 350K and 363K. Parameter set SGLD 4b also exhibits thermodynamic stability, but is less kinetically efficient than SGLD 1, SGLD 3, SGLD 3a, or SGLD 3b.

From these results, we may conclude that for trpzip2, the application of SGLD is most successful at maintaining thermodynamic stability and accelerating folding when a relatively low guiding factor of 1.0 and short averaging time of 0.2 ps is used (SGLD 1). Additionally, an averaging time of 2.0 ps used with all tested guiding factors (SGLD 3, SGLD 3a, SGLD 3b) results in populations of native structure that are the most stable and the fastest to fold of all the SGLD parameter sets that were tested. In order to test whether the trends of accelerated and stable folding would continue as the guiding factor was increased to 20.0 with an averaging time of 2.0 ps, simulation of an additional parameter set (SGLD 3c) was also performed. Although the first passage time of folding remained approximately constant compared to that of SGLD 3b, the fraction of native structures decreased to approximately 77% due to faster rates of unfolding.

It is notable that averaging times of 1.0 and 2.0 ps give rather different results for both the thermodynamic and kinetic properties when quantities obtained with the same guiding factor are compared. These differences underscore the sensitivity of the method to the combination of averaging time and guiding factor used. It is also significant that averaging times larger than 1.0 ps have not been used in prior SGLD studies, but our work indicates that an averaging time of 2.0 ps is beneficial in enhancing the kinetic rate of trpzip2, as well as maintaining its thermodynamic stability.

In order to quantify the extent to which each SGLD parameter set sampled diverse conformations during the 100 ns simulation, the number of structural clusters and the rate at which new clusters were sampled were determined. Table 3.5 lists the number of clusters sampled in each of the SGLD simulations. For each averaging time, the number of distinct clusters found during the simulation increases as the guiding factor increases, which is a sensible result given that increasing the guiding factor is predicted to result in increased sampling [125]. The use of a short averaging time results in a larger number of clusters, and for a single guiding factor, the number of clusters decreases as the averaging time increases. Increasing the averaging time is predicted to enhance slower motions of the peptide [125], so it is sensible that the extent of sampling would decrease as the enhanced motions become slower.

Figure 3.5 shows the rate at which new clusters were identified by each of the SGLD parameter sets, and includes the data from the LD run at 300 K as a reference point. Trajectories with small guiding factors and long averaging times are among the slowest to sample new structures, while those with large guiding factors and short averaging times generally sample at the fastest rate. With the exception of parameter set SGLD 4, which samples structures at a rate that is approximately equal to that of the LD simulation at 300 K, the use of SGLD was able to accelerate the rate of sampling of trpzip2 with all tested combinations of guiding factors and averaging times relative to the reference 300 K LD simulation. Parameter sets SGLD 3a and SGLD 1 sampled new structures at a rate that is approximately equal to that of the LD simulations run at 325 K. SGLD 3b and SGLD 2a sampled new structures at a rate that is approximately equal to that of the LD simulations at 350 K. Use of parameter sets SGLD 1b, SGLD 1a, or SGLD 2b resulted in sampling rates that were higher than those of any of the LD reference simulations, including those run at 375 K.

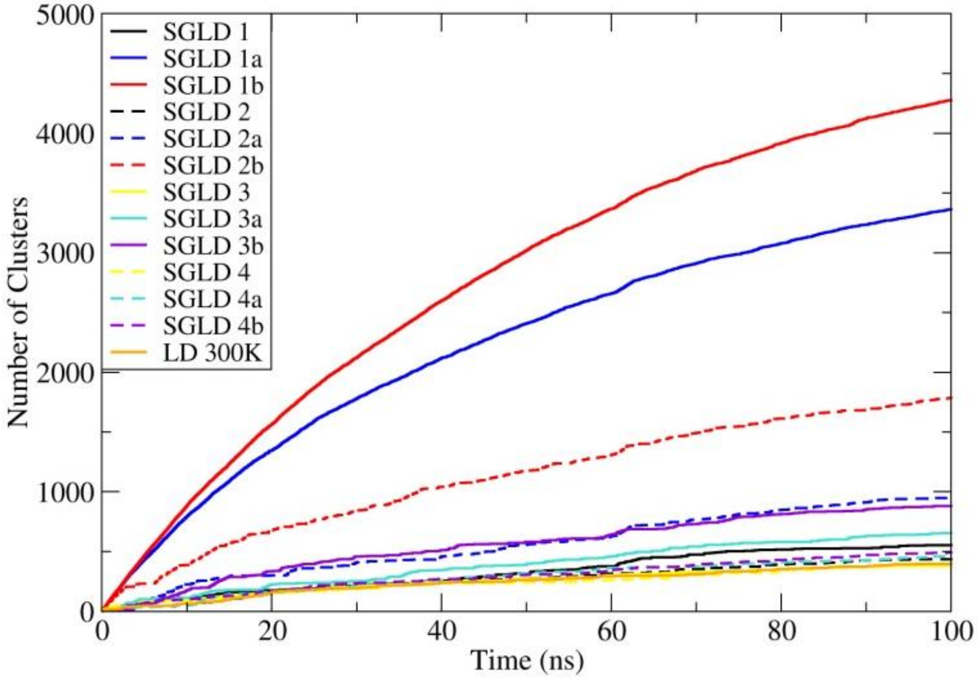

Figure 3.5: Number of clusters vs. time for 100 ns SGLD simulations of trpzip2.

Having quantified the effect of the use of SGLD on the rate of sampling, we now examine whether accelerated folding occurs at the cost of obtaining correct thermodynamic ensembles. In order to examine the effect of the use of SGLD on the energy surface of trpzip2, we have plotted the free energy as a function of two order parameters: the radius of gyration of the hydrophobic cluster of four tryptophan residues, and the RMSD of the backbone residues to the native structure. PMFs were also obtained for each of the REMD temperature trajectories, and comparisons may therefore be made between the SGLD ensembles and the well-converged REMD ensembles. As an example, parameter sets SGLD 1, SGLD 1a, and SGLD 1b are shown in Figure 3.6. As discussed above, trpzip2 was seen to rapidly fold to a large population (approximately 81%) of native structure using SGLD 1. The PMF obtained for the SGLD 1 trajectories resembles that of the REMD surface from the replica run at 327 K, with a compact minimum centered on a radius of gyration of approximately 6.5 Å and a backbone RMSD of 0.75 Å. The PMF obtained using SGLD 1 contains a higher proportion of extended structures than those in the REMD simulation, although this population of extended structures is small. The use of parameter sets SGLD 1a and SGLD 1b resulted in zero population of folded structures. The surfaces obtained from these parameter sets are comparable to that of the REMD trajectory at 465 K, with its loss of distinguishable minima. It is apparent from the surfaces that the ensembles obtained from running SGLD simulations at 300 K with a short averaging time and large guiding factor are comparable to those obtained from running LD simulations at temperatures that are much higher than 300 K.

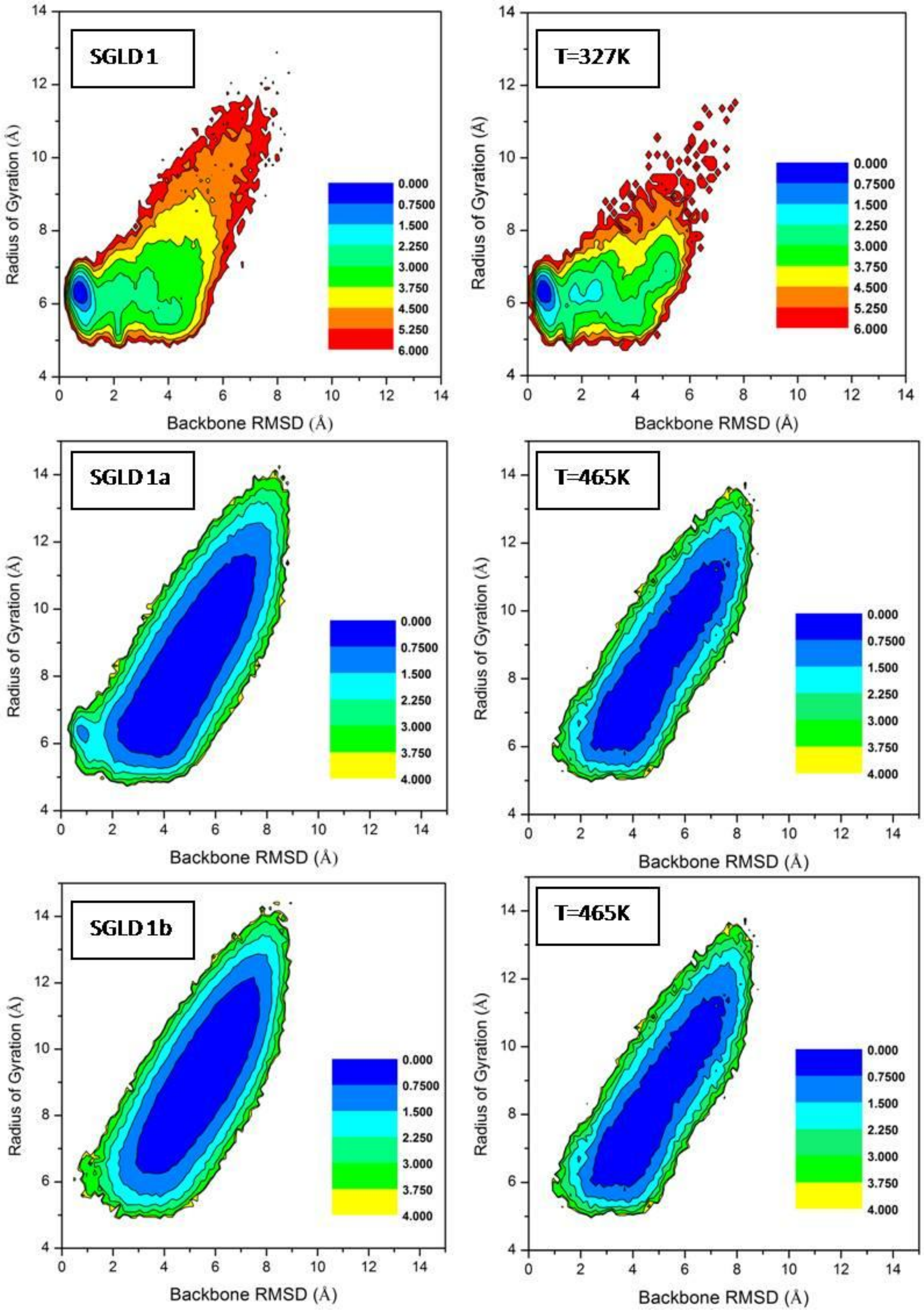

Figure 3.6: Comparison of free energy surfaces obtained for SGLD simulations of trpzip2 with averaging time of 0.2 ps and guiding factors of (a) 1.0, (c) 5.0, and (e) 10.0 with REMD simulations at (b) 327 K, (d) 465 K, and (f) 465 K. Energies are in units of kcal/mol.

Figure 3.7 is a comparison of the free energy surfaces obtained for simulations SGLD 3, SGLD 3a, and SGLD 3b with the comparable surface from the REMD simulations, that of the 327 K temperature trajectory. Each of these SGLD parameter sets attained stable thermodynamic populations at a faster rate than the LD simulation at 300 K. The PMFs for these parameter sets each have a compact minimum centered on a radius of gyration of approximately 6.5 Å and a backbone RMSD of 0.75 Å, which suggests the convergence of the ensemble to the native state. Surfaces obtained from REMD at 327.5K, as well as those from SGLD with a guiding factor of 1.0, exhibit two distinct ensembles of misfolded structures: one with a radius of gyration of approximately 6.0 Å and backbone RMSD ranging from approximately 2-3 Å, and a second with a radius of gyration of approximately 7.0 Å and a backbone RMSD of approximately 5.0 Å. Increasing the guiding factor removes the latter population of misfolded structures, although it appears that the former population of misfolded structures increases, and structures become more extended. The use of parameter sets SGLD 3, SGLD 3a, and SGLD 3b with trpzip2 resulted in rates of folding that were approximately equal to that of the standard LD simulation at 327 K. From these surfaces, we may conclude that the populations of folded structures that were obtained in these simulations also resembled those obtained from the LD simulation at 327 K. These parameter sets are more kinetically efficient than the standard LD simulation at 300 K, but are able to maintain a stable population of folded structures resembling that found in a relatively low-temperature LD simulation.

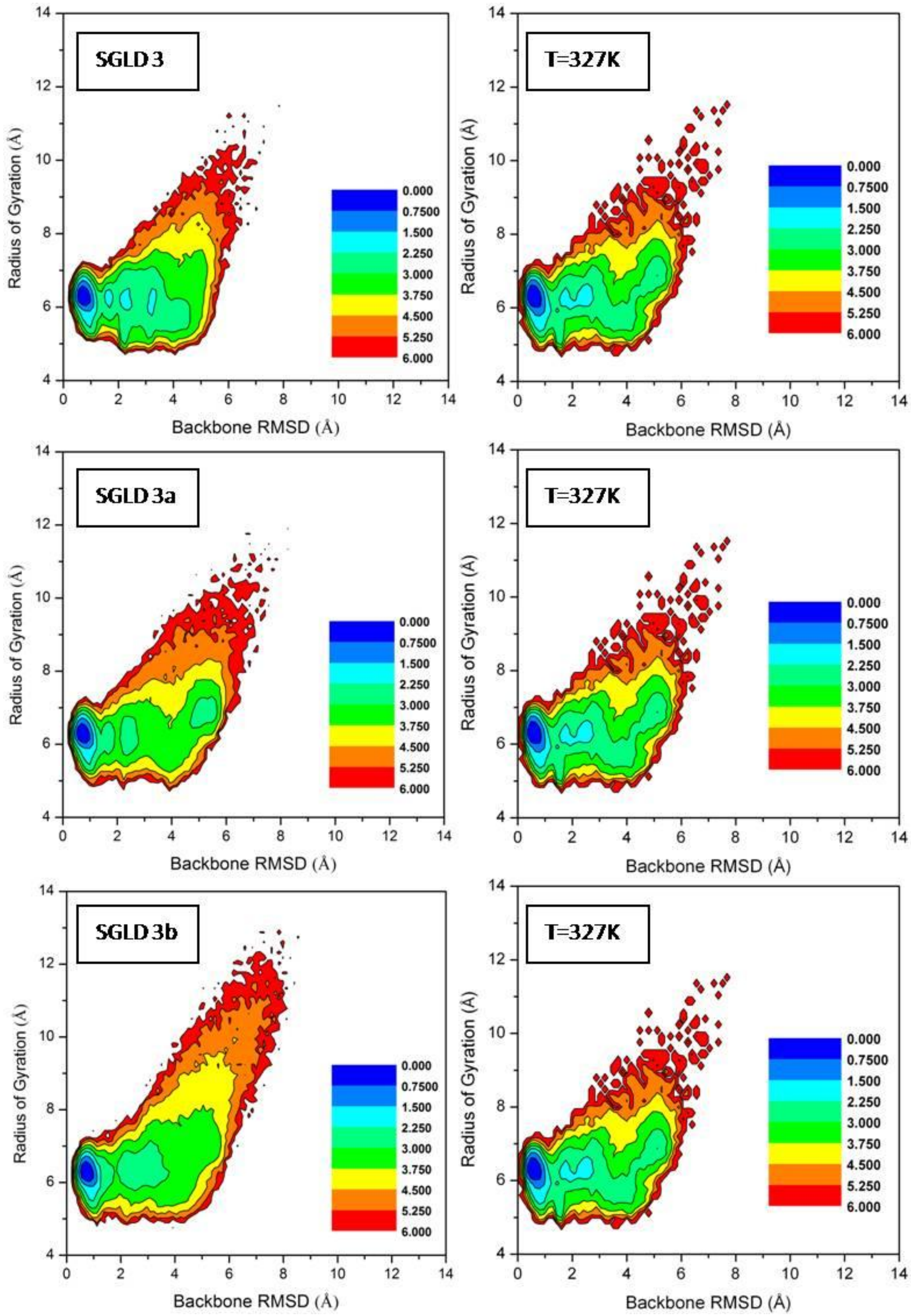

Figure 3.7**:** Comparison of free energy surfaces obtained for SGLD simulations of trpzip2 with averaging time of 2.0 ps and guiding factors of (a) 1.0, (c) 5.0, and (e) 10.0 with REMD simulations at (b) 327 K, (d) 327 K, and (f) 327 K. Energies are in units of kcal/mol.

Table 3.6 includes the thermodynamic and kinetic data obtained in SGLD simulations of trp-cage. Reviewing the results, it is evident that the largest populations of native structure are obtained when the lowest guiding factor of 1.0 is used with averaging times of 1.0 ps, 2.0 ps, and 10.0 ps (SGLD 2, SGLD 3, and SGLD 4). The use of these parameter sets results in populations of native structure (44%, 40%, and 48%) that are approximately equal to that obtained from the REMD trajectory at 318.2K (44%). Additionally, all three of these parameter sets have relaxation times of folding (6.4 ns, 9.7 ns, and 7.1 ns) that are faster than the reference LD simulation at 300 K, which has a relaxation time of folding of 13.9 ns. Relative to the LD simulation at 300 K, the use of these parameter sets result in speedup factors of 1.9, 2.8, and 2.2, respectively.

With the larger guiding factors of 5.0 and 10.0, an averaging time of 10.0 ps (SGLD 4a and SGLD 4b) is necessary in order to obtain populations of folded structures that are approximately 30%, but these populations are lower than those obtained with a guiding factor of 1.0, which are approximately 44%. The use of guiding factors of 5.0 and 10.0 with averaging times ranging between 0.2 ps and 10.0 ps is unsuccessful in SGLD simulations of trp-cage; although the use of these guiding factors exhibit an acceleration of the folding rate that is equal to 1.8 and 1.4 times those of the LD simulation of at 300 K, these simulations do not produce stable structures that are maintained over the course of the simulation. These results are in contrast to those obtained with trpzip2, which exhibited stable, accelerated folding using guiding factors of 5.0 and 10.0 with an averaging time of 2.0 ps (SGLD 3a and SGLD 3b).

Table 3.6: Thermodynamic and kinetic data from 200 ns SGLD simulations of trp-cage. Column 4 lists the average fraction of native structure obtained for each SGLD parameter set, with their associated error bounds. Column 5 is the value of ΔG for each of the SGLD parameter sets versus the REMD reference trajectory at 300 K. Column 6 lists the relaxation time of folding obtained from the single-exponential fit of first passage times, and column 7 lists the relaxation time for escape from the native basin. Column 8 is the value of the relaxation time of folding for each LD parameter set divided by that for the SGLD simulation at 300 K (speedup), and column 9 gives the number of clusters found by each parameter set at the end of the 200 ns simulation time.

Parameter Set	**Averaging Time (ps)**	**Guiding Factor**	**Fraction Folded**	**ΔG (kcal/mol)**	**Folding t_{relax} (ns)**	**Unfolding t_{relax} (ns)**	**SGLD Speedup**	**Number of Clusters**
SGLD 1	0.2	1.0	15±1%	0.8	3.3	2.9	4.2	2479
SGLD 2	1.0	1.0	44±2	0.2	6.4	14.0	2.2	1059
SGLD 3	2.0	1.0	40±3%	0.3	9.7	14.0	1.4	990
SGLD 4	10.0	1.0	48±3%	0.1	7.1	30.1	2.0	588
SGLD 1a	0.2	5.0	0±0%	4.4	35.9	0.1	0.4	6014
SGLD 2a	1.0	5.0	17±1%	0.8	3.3	0.5	4.2	2766
SGLD 3a	2.0	5.0	7.9±1%	1.2	3.5	1.0	4.0	3660
SGLD 4a	10.0	5.0	30±7%	0.4	7.8	4.1	1.8	1542
SGLD 1b	0.2	10.0	0±0%	6.1	41.4	0.1	0.3	5842
SGLD 2b	1.0	10.0	9.6±2%	1.1	12.5	0.2	1.1	3693
SGLD 3b	2.0	10.0	1.6±0%	2.2	6.0	0.4	2.3	4996
SGLD 4b	10.0	10.0	30±7%	0.7	9.7	2.0	1.4	2441
SGLD 3c	2.0	20.0	0.3±0%	3.2	15.5	0.1	0.9	5049

The number of structural clusters identified for each SGLD parameter set is given in Table 3.6. As seen in the results for trpzip2, for each averaging time, the number of distinct clusters found during the simulation increases as the guiding factor increases. For each guiding factor, the use of a short averaging time results in a larger number of clusters. With the exceptions of parameter sets SGLD 3a and SGLD 3b, the number of clusters decreases as the averaging time increases. Figure 3.8 shows the rate at which new clusters of trp-cage structures were identified by each of the SGLD parameter sets, with the LD simulation run at 300 K shown as reference. Trajectories with small guiding factors and long averaging times were among the slowest to sample new structures, while those with large guiding factors and short averaging times generally sampled at the fastest rate. With the exception of parameter set SGLD 4, the use of SGLD with all tested combinations of guiding factors and averaging times was able to accelerate the rate of sampling of trpcage relative to that of the LD simulation at 300 K. The LD simulations run at 325 K sampled new structures at a rate that is approximately equal to that of SGLD 4. The LD simulation at 350 K samples at a rate that is between those of SGLD 2a and SGLD 3b. Parameter sets SGLD 1a, SGLD 1b, SGLD 3c, SGLD 3b, SGLD 3a, and SGLD 2b sample new structures more efficiently than the simulations run at 350 K using LD.

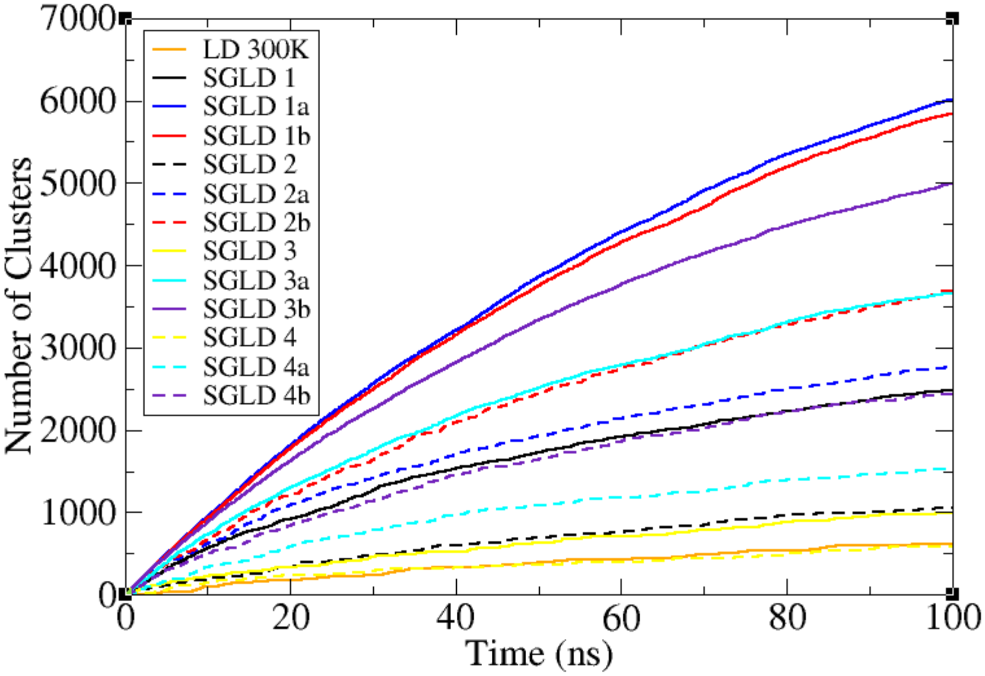

Figure 3.8: Number of clusters vs. time for 100 ns SGLD simulations of trp-cage.

PMFs comparing the ensembles obtained in the SGLD simulations to the well-converged SGLD ensembles are shown in Figures 3.9 and 3.10. As discussed above, SGLD 1 exhibited a native population of approximately 15% after 200 ns of simulation, while neither SGLD 1a nor SGLD 1b were able to maintain any native population. The PMF obtained for the SGLD 1 trajectories resembles that of the REMD surface from the replica run at 357.9 K, with a broad minimum centered on a radius of gyration of approximately 8 Å and a backbone RMSD of 3.5 Å. The surfaces obtained with SGLD 1a and SGLD 1b are comparable to that of the REMD trajectory at 452.8 K, which exhibits an extremely broad minimum with backbone RMSD ranging from approximately 5-9 Å and a radius of gyration that ranges from approximately 9-13 Å.

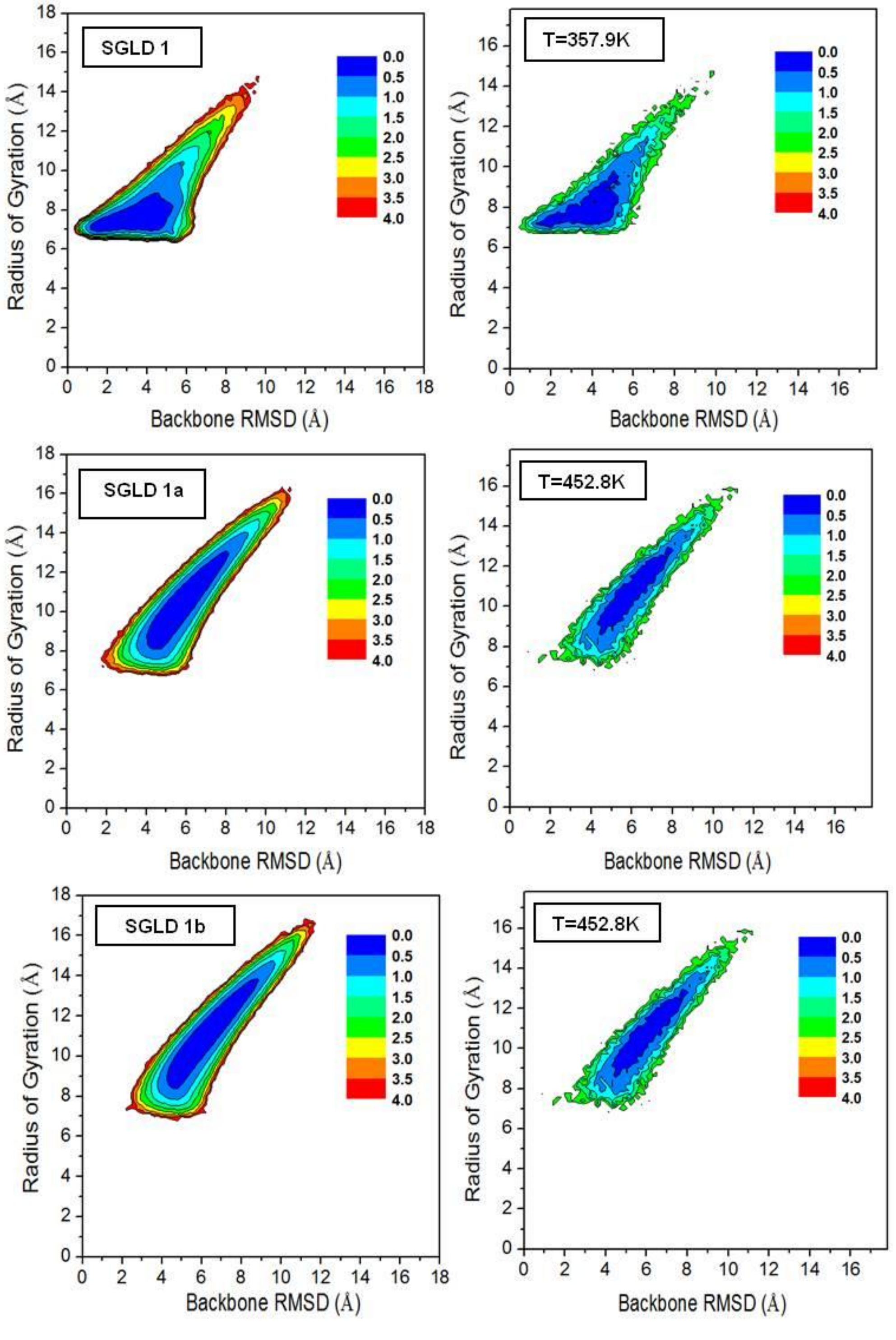

Figure 3.9**:** Comparison of free energy surfaces obtained for SGLD simulations of trp-cage with averaging time of 0.2 ps and guiding factors of (a) 1.0, (c) 5.0, and (e) 10.0 with REMD simulations at (b) 357.9 K, (d) 452.8 K, and (f) 452.8 K. Energies are in units of kcal/mol.

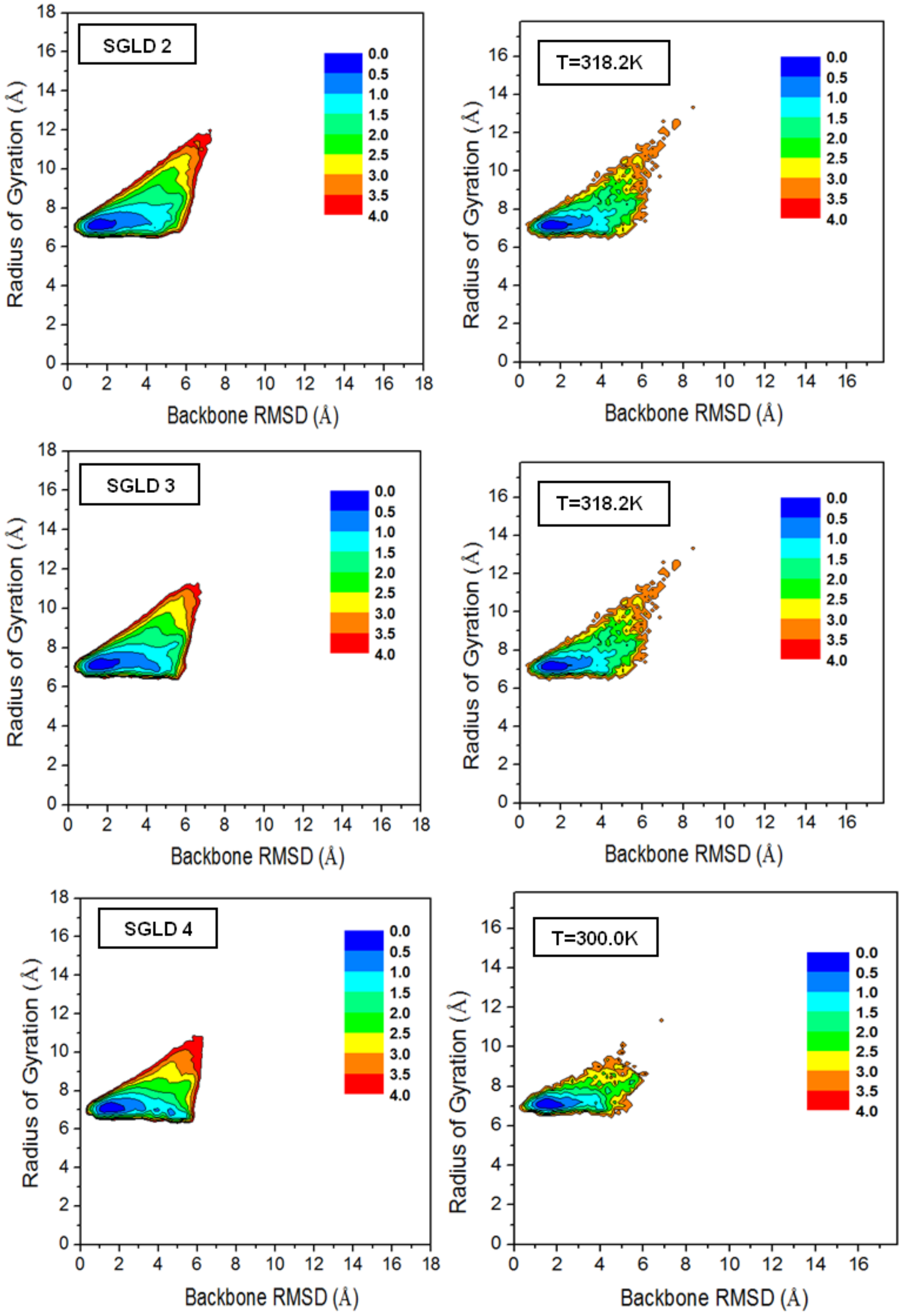

Figure 3.10**:** Comparison of free energy surfaces obtained for SGLD simulations of trp-cage with guiding factor of 1.0 and averaging times of (a) 1.0 ps, (c) 2.0 ps, and (e) 10.0 ps with REMD simulations at (b) 318.2 K, (d) 318.2 K, and (f) 300 K. Energies are in units of kcal/mol.

Figure 3.10 is a comparison of the free energy surfaces obtained for simulations SGLD 2, SGLD 3, and SGLD 4 of trp-cage with comparable surfaces from the REMD simulations. The PMFs for parameter sets SGLD 2 and SGLD 3 each have a compact minimum centered on a radius of gyration of approximately 7 Å and a backbone RMSD of 1.75 Å, which suggests that the ensemble has populated the native state, and resembles the surface obtained for the REMD simulation run at 318.2 K. The surface obtained for SGLD 4 comprises structures that are slightly more compact that those of SGLD 2 and SGLD 3, and resembles the REMD trajectory run at 300 K, with a minimum centered on a radius of gyration of approximately 7 Å and a backbone RMSD of 1.5 Å.

3.3.3 Discussion: Comparison of SGLD with Trpzip2 and Trp-cage

Looking at the results for trpzip2 and trp-cage together, we may generalize that peptides of similar size but different topologies require different combinations of guiding factors and averaging times in order to optimize folding rates and stability. Stable, accelerated folding of trpzip2 required the use of parameter sets with a guiding factor of 1.0 and an averaging time of 0.2 ps, or an averaging time of 2.0 ps with guiding factors of 1.0, 5,0, and 10.0. The use of parameter sets SGLD 1, SGLD 3, SGLD 3a, and SGLD 3b resulted in ensembles of structures that exhibited thermodynamic stability comparable to REMD simulations run at approximately 327.5 K, with kinetic rates of folding that were approximately 3-5 times faster than LD simulations run at 300 K. For trp-cage, the simulations run with a guiding factor of 1.0 and averaging times of 1.0 ps, 2.0 ps, and 10.0 ps (SGLD 2, SGLD 3, and SGLD 4) were the most thermodynamically stable and kinetically efficient. The use of these parameters resulted in ensembles of structures with native populations that were approximately equal to that of the REMD simulation run at 318.2 K, with relaxation times of folding that were approximately 5-10 times as fast as that obtained using LD at 300 K. Even using the most successful of the SGLD parameter sets that were tested, neither trpzip nor trpcage was able to achieve the fraction of native content that was observed in the REMD ensembles at 300K.

3.3.4 Alpha-helix K19: Langevin Dynamics and Self-guided Langevin Dynamics Simulations

The results obtained from our SGLD simulations indicate that certain parameter sets are reasonably effective in accelerating the folding of trpzip2 and trp-cage while maintaining populations of folded states, whereas others are not, and some are less effective than using LD and result in unpredictable and extreme distortions in the free energy landscape. In order to further test the transferability of parameter sets between peptides of similar size but differing structures, three of the most effective sets of parameters from the SGLD simulations of trpzip2 (SGLD 3, SGLD 3a, and SGLD 3b) were applied to the α-helix K19. These SGLD parameters were chosen because the SGLD simulations of trpzip2 using these parameters were more successful at accelerating folding while maintaining stable populations of native structures relative to the converged REMD ensemble. In addition, SGLD 3 was successful in its application to trp-cage and thus exhibited transferability between two peptide systems of differing topologies. The averaging time for these parameter sets was 2.0 ps, and the guiding factors were 1.0, 5.0, and 10.0, respectively (SGLD 3, SGLD 3a, and SGLD 3b). In addition, two sets of simulations were run using guiding factors less than 1.0 when it became apparent that the folded ensembles of K19 were most stable with a relatively small guiding factor. Parameter set SGLD 3c had an averaging time of 2.0 ps and a guiding factor of 0.25, while parameter set SGLD 3d had an averaging time of 2.0 ps and a guiding factor of 0.5. Forty-eight trajectories of 100 ns each were obtained for each of the five parameter sets, for a total of 24 μs simulation time. Two standard Langevin dynamics simulations serving as benchmarks were performed at 280 K and 300 K, with forty-eight 100 ns trajectories obtained at each temperature, for a total of 9.6 µs simulation time.

For all SGLD and LD runs, the average fractional helicity of each residue over all 48 trajectories was calculated using DSSP [160]. The acetyl and amide groups at the termini of the peptide contained no helical content and are therefore not included in the analysis. The resulting values of the percent helical content per residue are compared in Figure 3.11, and the average values of the fractional helicity per residue across residues 4-16 of the peptide are given in Table 3.7. This analysis follows that used in a previous computational study to determine the temperature dependence of the helical propensity of residues in K19 [154].

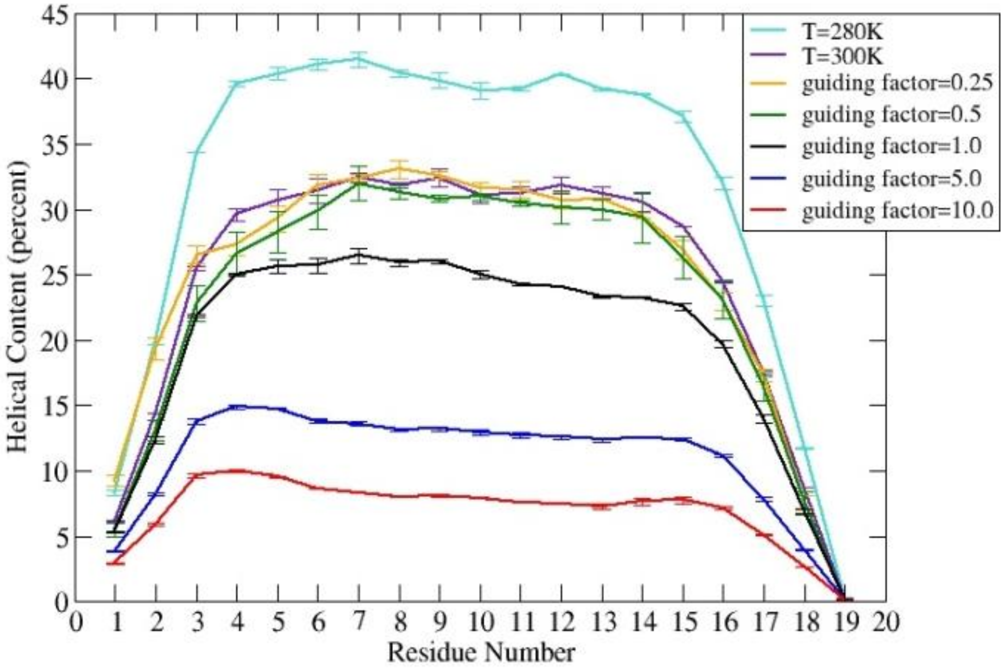

Figure 3.11: Helical content per residue of the peptide K19 obtained from LD simulations at 280 K and 300 K and SGLD simulations at 300 K with an averaging time of 2.0 ps.

Table 3.7: Thermodynamic and kinetic data from 100 ns LD and SGLD simulations of K19. Column 4 lists the values of the average helicity across residues 4-16 of the peptide, with their associated error bounds. Column 5 is the value of ΔG for all simulations versus the LD reference trajectory at 300 K. Column 6 lists the relative percent decrease in helicity for each parameter set versus the LD simulation at 300 K. Column 7 lists the folding time of the peptide, estimated from the fractional helicity vs. time.

Parameter Set	**Averaging Time (ps)**	**Guiding Factor**	**Average Helicity**	**ΔG (kcal/mol)**	**Relative Decrease in Helicity**	**Folding Rate (ns)**
LD 280K	n/a	n/a	39.3±0.3%	-0.2	n/a	5.0
LD 300K	n/a	n/a	30.7±0.3%	0	n/a	3.0
SGLD 3	2.0	1.0	24.6±0.2%	0.1	20%	3.0
SGLD 3a	2.0	5.0	14.7±0.1%	0.4	48%	1.0
SGLD 3b	2.0	10.0	8.9±0.1%	0.7	71%	0.5
SGLD 3c	2.0	0.25	30.3±0.6%	0.01	1%	2.8
SGLD 3d	2.0	0.5	29.1±1.0%	0.03	29%	2.5

All of the simulations exhibit a profile with very little helicity at the C-terminus, a rapid increase in helicity between residues 2 and 4, a plateau region of approximately constant helicity from residues 4-16, and a rapid decrease in helicity between residues 16 and 19. The LD simulation at 280 K exhibits the highest fractional helicity of the simulations that were compared, with a maximum helical content of 41.4% at the alanine residue in position 7. The LD simulation at 300 K has a maximum fractional helicity of 31.4% at alanine 7. The termini exhibit a difference of approximately 5% between the helicities of their residues in the 280 K and 300 K simulations, while residues 4-16 have an average absolute difference of approximately 10% between the two simulations (39.3% versus 30.7%) and a relative difference of 25%. The absolute difference of approximately 10% in this region is supported by previous work [154] in which the helical propensity of residues in K19 versus temperature was calculated, and an absolute difference of approximately 15% was found between the helicities of all of the residues at 275 K and 300 K.

Introduction of a guiding factor less than 1.0 maintains the helical content present in the LD reference simulations at 300 K, which is 30.7%. The addition of a guiding factor greater than 1.0 results in a uniform decrease in helical content across all residues of the peptide, although the trend across the sequence is maintained. The percent decreases in helicity for each SGLD parameter set relative to the LD control simulation at 300 K are given in Table 3.7. Although the decreases in helicity appear to be large, these differences in native population correspond to relatively small differences in thermodynamic stability, as evidenced by the values of ΔG given in Table 3.7. Average helical content at each time step was calculated over all of the 48 trajectories in order to determine the time needed for the ensembles to reach their converged distribution from the linear starting structures.

Folding rates were estimated from the time required for the fractional helicity vs. time to equilibrate, and are given in Table 3.7. As shown in Figure 3.12, equilibration was reached within 10 ns for all parameter sets. At 300 K, approximately 3.0 ns were needed for the starting structures to reach a fractional helicity of approximately 27%. The use of SGLD resulted in folding rates that are equal to, or up to six times faster than, those of the LD simulations at 300 K. With the exception of parameter set SGLD 3, increasing the guiding factor lead to a decrease in the folding rate. Parameter sets SGLD 3, SGLD 3c, and SGLD 3d exhibited folding rates that were approximately equal to that of the LD simulation at 300 K. SGLD 3a and SGLD 3b exhibited folding rates that constitute a three-fold and six-fold increase, respectively, versus the LD simulation at 300 K, but the resulting ensembles exhibited the largest degree of thermodynamic destabilization relative to the LD simulation. The limited improvement in the folding rates obtained using SGLD with K19 is likely due to the fact that the system folds rapidly without assistance. One direction for future study may therefore be the use of SGLD with a system such as K19 in explicit solvent. As folding proceeds more slowly in the presence of explicit solvent, the effect of SGLD on the kinetics of the system will be clearer.

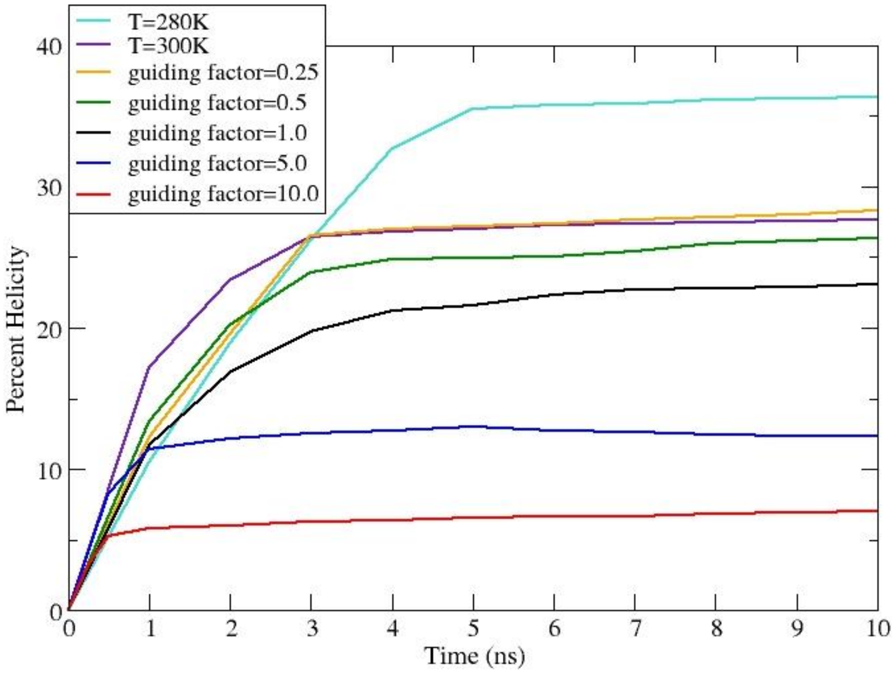

Figure 3.12. Fractional helicity vs. time obtained from first 10 ns of 100 ns LD and SGLD simulations of helix K19.

3.4 Conclusions

The results obtained in this study indicate that simulations using self-guided Langevin dynamics are very sensitive to the combination of averaging time and guiding factor that is chosen. Despite the enhancement in folding rate and thermodynamic stability that may be attained using self-guided Langevin dynamics, the optimal combination of parameters is not apparent before the simulation is run. As shown in the examples above, incorrect choice of parameters may lead to a slowing of the rate of folding, a destabilization of the folded state of the system, or even the complete loss of recognizable features on the free energy landscape. Additionally, our systematic testing of identical parameter sets for the β-hairpin trpzip2 and the trp-cage miniprotein, as well as our attempt to transfer successful parameter sets from the beta-hairpin trpzip2 to the α-helical system K19, indicates that the choice of optimal parameters may need to be determined on a system-by-system basis if the systems are of similar size but different topology.

Despite these drawbacks, self-guided Langevin dynamics remains a powerful and sensitive method by which to enhance sampling in simulations of biological systems. In cases where an optimal combination of parameters has been determined through preliminary testing, a simulation employing SGLD can be more efficient than a Langevin dynamics simulation in its exploration of the conformational space available to the protein. This effect would likely be more prominent in large systems, where the large number of degrees of freedom requires longer simulation time for the observation of conformational changes. Additionally, the cluster analysis undertaken in this study indicates that SGLD simulations with large guiding factors and short averaging times may be employed to efficiently produce alternate structures that may be used for decoy screening, or as a reservoir for reservoir replica exchange molecular dynamics simulations if native as well as non-native structures are generated [156]. The potential utility of self-guided Langevin dynamics therefore warrants its continued study, particularly in its application to larger biomolecular systems, and systems in explicit solvent.

Chapter 4: Non-Boltzmann Reservoir Replica Exchange Molecular Dynamics with User-defined Weights

Abstract

Many algorithms have been developed to enhance the conformational sampling of biomolecules that is achieved in Monte Carlo and molecular dynamics simulations. One method that has proven to be very efficacious in enhancing sampling is the replica-exchange molecular dynamics algorithm (REMD), which achieves a random walk in temperature space in order to surmount conformational barriers in the energy landscape. Variants of this technique have been developed over the years in order to increase the efficiency of REMD simulations of biomolecules. In particular, approaches have been developed in which a structural reservoir is used to decouple the high-temperature search for structures from the exchanges and annealing which occur at lower temperatures. It has been shown that the contents of this reservoir need not comprise a Boltzmann-weighted ensemble; any ensemble of structures may be used as long as its probability distribution is known. Expanding on this method, we have developed an algorithm to further enhance the efficiency of reservoir REMD through the inclusion of a weight factor that relates the relative probabilities of the highest-temperature replica structure and the structure in the reservoir under exchange. In Chapter 4, we outline attempts to apply this method to the model system alanine dipeptide, and discuss the results obtained using a coarse-grained model that considers only the potential energy of the dipeptide as a function of its dihedral angles and does not consider its atomistic degrees of freedom.

4.1 Introduction

As discussed in Section 2.6 of this work, conformational sampling remains one of the most significant current limitations in the simulation of biomolecules. The existence of a rough underlying energy landscape in which local energy minima are separated by large barriers causes structures to become trapped in local minima, and prevents Boltzmann-weighted sampling from occurring within a timescale that is computationally feasible. In order for a system to exhibit ergodicity, it must be able to reach any point in phase space from any initial state. The existence of large energy barriers on the landscape, however, often leads to quasi-ergodocity, by which the system has an extremely small probability of exiting a local minimum [162]. A simulation may appear converged with respect to certain order parameters, but a simulation of the same system initiated from a different state will reveal that the entirety of conformation space was not explored.

This problem is particularly prevalent at the relatively low temperatures (~300 K) often used to simulate the conditions at which biological systems function, as the tendency for trapping generally increases as the thermal energy of the system decreases. One method that has been developed to address the problem of accurate sampling at realistic temperatures is known as replica-exchange molecular dynamics (REMD) [126,163,164,165]. In an REMD simulation, a number of MD simulations of a system, each at a different temperature, are simultaneously run. Each independent simulation is known as a “replica,” and the temperatures at which they are run range from biologically relevant, experimentally accessible temperatures, such as 280-300 K, to higher temperatures, such as 600 K, at which the system is expected to have enough thermal energy to easily overcome potential energy barriers. At a predetermined number of time steps, each replica attempts to exchange its structure with the replica that is adjacent to it in temperature. Whether a move is accepted or not is determined based on a Metropolis-type criterion [33], which considers the probability of sampling the alternative structure at the current temperature. Canonical ensemble properties are thus preserved through the construction of the transition probability. Further details on this method may be found in Section 4.2 below. Due to the increase in computational efficiency that it affords, REMD has seen extensive application in protein folding studies of peptides and small proteins [148,166,167,168,169,170,171].

In spite of the advantages that it confers in sampling, difficulty remains in implementing REMD for large systems; as an extensive algorithm, its computational cost increases with the size of the system [172,173,174,175]. The number of replicas required in an REMD simulation increases with the square root of the number of degrees of freedom in the system, which limits the size of the systems under study as well as the length of the simulations that can be obtained using this method. Another issue which limits the efficiency of REMD is that although the high-temperature replicas allow the system to surmount energy barriers to sample conformational space, they do not necessarily confer any advantage in locating the native state. The increase in conformational entropy at high temperature often leads to weak, or non-Arrhenius, dependence of the folding rate on temperature [161,176,177]. If a high-temperature replica does sample the native state, the exchange criterion will likely require that this structure be exchanged down to lower temperature, requiring that the search begin again at high temperature. In order to obtain a correctly weighted ensemble, the high-temperature replica may need to sample the native state several times before convergence is obtained, which may require a large amount of computational time.

In order to reduce the search time required by the highest-temperature reservoir to sample many folding transitions, several algorithms have been developed which decouple the high-temperature conformational search from the low-temperature simulations [178,179,180]. The "J-walking" method [181] was developed in order to couple a single low-temperature Monte Carlo (MC) simulation to a collection of structures that had been generated at higher temperature. This algorithm allows the low-temperature walker to occasionally jump directly to the distribution at the higher temperature, thereby surmounting any energy barriers that may have separated the two structures. Tandem conformational searches by the low-temperature and high-temperature systems were attempted with this method; however, it was found that correlation between the tandem walkers, as well as increased computational overhead, limited the utility of the method. Instead, it was found that the generation of conformations prior to any jumps yielded accurate results without any additional computational cost. Other methods subsequently extended this algorithm, including exchanges between structures of different resolution [182,183] and from finite reservoirs of structures [184].

Inspired by J-walking, Okur et al. [158] developed a replica-exchange molecular dynamics simulation scheme known as reservoir replica-exchange molecular dynamics (R-REMD) in which a high-temperature Boltzmann-weighted collection of structures is independently obtained using MD and subsequently used as a reservoir for the lower-temperature exchange simulations (details in Section 4.2). This method was applied to the β-hairpin trpzip2 and the 3-stranded antiparallel β-sheet dPdP model systems in implicit solvent with a reservoir of 10,000 structures obtained at 400 K. This relatively high temperature was chosen because it is above the experimental melting temperature of each peptide, and therefore high enough to allow for extensive sampling of conformation space. The use of this high temperature requires exchanges from the reservoir to significantly transform this ensemble in order to obtain accurate populations at lower temperatures. For both peptides, this technique was shown to exhibit increased speed of convergence and an accurate thermal stability profile when compared against standard REMD, as exchanges are made into a correctly-weighted ensemble that has already been generated and does not require any additional computational overhead.

The nature of the ensemble that is used as the reservoir, particularly the question of whether it needs to be Boltzmann-weighted in order for the structures to be passed to lower temperatures with correct probability, was the subject of a subsequent study by Roitberg et al. [156]. If the purpose of R-REMD is to increase the efficiency of the simulation, then the reservoir-generating step must not be overly time-consuming. Because is it often difficult to generate a reservoir of Boltzmann-weighted structures, a variant of R-REMD was developed that uses a reservoir in which the structures are not Boltzmann-weighted, demonstrating that an arbitrary distribution of structures may be included in the reservoir given that its probability distribution is known. In this formalism, trpzip2 structures were selected for inclusion in the non-Boltzmann reservoir by performing a cluster analysis based on structural similarity on the Boltzmann-weighted reservoir of 10,000 structures. The representative structure of each of the 700 resultant clusters was extracted for inclusion in the non-Boltzmann reservoir, resulting in an ensemble in which each member had a probability of 1/700 of being selected for an exchange to lower temperature. Whereas the Boltzmann-weighted reservoir that had originally been used contained approximately 3% native structures, the non-Boltzmann reservoir contained only a single native structure, with a probability of 1/700. The use of this “flat” reservoir required the re-derivation of the Metropolis criterion, as described in Reference [156] and derived below in Section 4.2.

The results of using the equally-weighted non-Boltzmann reservoir indicated that accelerated convergence and an accurate thermal stability profile can be reached with an arbitrary reservoir as long as the correct Metropolis criterion is used. Melting curves obtained from simulations with this reservoir were virtually identical to those obtained using a standard REMD simulation without the use of a reservoir. Additionally, convergence speedup of a factor on the order of 10 was obtained using R-REMD with either the Boltzmann-weighted or non-Boltzmann-weighted reservoir compared against using REMD without a reservoir. These results indicate that similar acceleration of convergence was observed from using either the Boltzmann-weighted or non-Boltzmann-weighted reservoir. However, if the time required for reservoir generation is included in calculating the total simulation time, we expect that the use of the non-Boltzmann reservoir decreases the total simulation time.

In this chapter, we outline a method that uses the R-REMD exchange formalism in combination with a non-Boltzmann-weighted reservoir containing structures that have been assigned to clusters using a structural similarity metric. Although the method of R-REMD with evenly distributed weights described above was effective, the use of only a single conformation from each structural cluster may lead to incomplete sampling with larger systems, as the reservoir in this case is populated only by structures that are located toward the bottom of their respective conformational basin. By clustering the reservoir structures and including the complete ensemble in the reservoir during the simulation, we increase the diversity of structures that may be sampled. In order to drive the acceptance of structures that in standard REMD would have a low probability of acceptance, we use the relative probabilities of the reservoir structure and the highest-temperature replica structure in order to scale the value of the Metropolis exchange criterion.

This method, which we have termed “non-Boltzmann reservoir replica exchange molecular dynamics with user-defined weights,” was tested on the model system alanine dipeptide in implicit solvent, with limited success. Discovering why this method yielded inaccurate results required that simpler coarse-grained simulations were undertaken in which only the potential energies of the alanine dipeptide conformations were used, and their atomistic degrees of freedom were ignored. These energies were then used to subject the system to a random walk in an energy space in which the partition function, and therefore the relative probability of each state, was *a priori* known. This simulation methodology allowed for greater control in testing different structural metrics for use with the clustering algorithm in order to elucidate its influence on the resultant ensembles. The following section of this work outlines the formulation of standard replica exchange molecular dynamics, reservoir replica exchange molecular dynamics with both Boltzmann-weighted and non-Boltzmann-weighted reservoirs, and non-Boltzmann reservoir replica exchange molecular dynamics with user-defined weights.

4.2 Derivation of Equations

We consider a system of N atoms of mass m_k *(k=1,…,N)* with coordinate and momentum vectors $q \equiv \{\vec{q}_1,\ldots,\vec{q}_N\}$ and $p \equiv \{\vec{p}_1,\ldots,\vec{p}_N\}$, respectively. The Hamiltonian of this system is the sum of its kinetic energy *K(p)* and its potential energy *E(q)* as follows:

$$H(q,p) = K(p) + E(q) \tag{4.1}$$

where

$$K(p) = \sum_{k=1}^{N} \frac{\vec{p}_k^{\,2}}{2m_k} \quad . \tag{4.2}$$

In a standard replica-exchange molecular dynamics simulation, the system comprises *M* independent copies, also known as replicas, of the original system at *M* different temperatures T_m *(m=1,…M)*. Only one replica is present at each temperature, and we label the replicas as *i* *(i=1,…M)*. In the single replica *i* at temperature T_m, the state *X* is specified by the *M* sets of coordinates and momenta of the *N* atoms in replica *i* at temperature T_m as follows:

$$X = \{x_1^{[i(1)]},\ldots,x_M^{[i(M)]}\} = \{x_{m(1)}^{[1]},\ldots,x_{m(M)}^{[M]}\} \tag{4.3}$$

where

$$x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m \quad . \tag{4.4}$$

At intervals during the simulation, we attempt to exchange the structures presented by two replicas that are adjacent in temperature. Considering the exchange of states *i* and *j*, which are at temperatures of T_m and T_n, respectively, the exchange of state *X* to state *X'* may be expressed as

$$X = \left\{\ldots, x_m^{[i]},\ldots,x_n^{[j]},\ldots\right\} \rightarrow X' = \left\{\ldots, x_m^{[j]},\ldots,x_n^{[i]},\ldots\right\} \quad . \tag{4.5}$$

In the canonical ensemble, the equilibrium probability of each state at temperature T_m is given by the Boltzmann factor *W* as follows:

$$W(p^{[i]},q^{[i]},T_m) = \exp\left\{-\frac{1}{k_B T_m} H(q^{[i]}, p^{[i]})\right\} \tag{4.6}$$

where k_B is the Boltzmann constant. Because the replicas in the system do not interact, the weight factor of state *X* in this ensemble is the product of the Boltzmann factors of each of the replicas:

$$W(X) = \exp\left\{-\sum_{i=1}^{M}\frac{1}{k_B T_m} x_m^{[i]}\right\} \quad . \tag{4.7}$$

Imposing the condition of detailed balance on the transition probability for the exchange of states *i* and *j* ensures that the exchange process converges towards an equilibrium distribution:

$$W(X)w(X \rightarrow X') = W(X')w(X' \rightarrow X) \quad . \tag{4.8}$$

Decoupling of the coordinates and momenta in the Hamiltonian as outlined in Reference [126] allows the Boltzmann factor to be written as

$$W(q^{[i]},T_m) = \exp\left\{-\frac{1}{k_B T_m} E(q^{[i]})\right\} \quad , \tag{4.9}$$

and substituting the Boltzmann factor for the weight of each conformation, we may express Equation 4.8 for an exchange between states at temperatures T_m and T_n as follows:

$$\begin{aligned}&\exp\left\{-\frac{1}{k_B T_m} E(q^{[i]}) - \frac{1}{k_B T_n} E(q^{[j]})\right\} \cdot w(X \to X') = \\ &\exp\left\{-\frac{1}{k_B T_m} E(q^{[j]}) - \frac{1}{k_B T_n} E(q^{[i]})\right\} \cdot w(X' \to X)\end{aligned} \quad . \tag{4.10}$$

Through rearrangement of Equation 4.10, we obtain

$$\frac{w(X \to X')}{w(X' \to X)} = \exp\left\{(\beta_n - \beta_m)(E(q^{[i]}) - E(q^{[j]}))\right\} \tag{4.11}$$

where $\beta_n \equiv \frac{1}{k_B T_n}$.

Exchanges between replicas must therefore obey the following exchange criterion

$$\rho = \min\left(1, \exp\left\{(\beta_n - \beta_m)(E(q^{[i]}) - E(q^{[j]}))\right\}\right) \quad , \tag{4.12}$$

known as the Metropolis criterion [33], to ensure that exchanges drive each replica to Boltzmann weighting.

During a standard REMD simulation, replicas at different temperatures are independently run using MD. At a predetermined number of steps, an exchange is attempted between replicas that are adjacent in temperature, with the probability of success determined from Equation (4.12). If the exchange is accepted, the temperatures of the replicas are swapped through velocity rescaling as follows:

$$\begin{aligned} p^{[i]\prime} &\equiv \sqrt{\frac{T_n}{T_m}}\, p^{[i]} \\ p^{[j]\prime} &\equiv \sqrt{\frac{T_m}{T_n}}\, p^{[j]} \end{aligned} \qquad (4.13)$$

and these structures then continue in their MD trajectories at their new temperatures. If the exchange is not accepted, the replicas remain at their current temperatures and continue in their MD trajectories.

The central idea of reservoir REMD is to generate a collection of high-temperature structures using MD before the REMD simulation is run, and to then attempt to use these static structures for exchange with the highest-temperature replica. Simultaneously, the lower-temperature replicas are attempting to exchange structures among each other as in standard REMD. When performing REMD simulations with a pre-generated reservoir of Boltzmann-weighted structures, the same set of exchange equations are used to either accept or reject the randomly selected reservoir structure as in Equation (4.12). For exchanges between the highest-temperature replica and the reservoir, the temperatures that are used in this equation are the temperatures of the high-temperature reservoir and the temperature at which the reservoir structures were generated. For exchanges between neighboring replicas, the temperatures of the two replicas are used, as in standard REMD. If an exchange between the highest-temperature replica and the reservoir is accepted, the coordinates and velocities of the reservoir structure are sent to the highest-temperature replica. Formally, the coordinates and velocities from the highest-temperature replica should be added to the reservoir; however, for computational convenience, they are discarded, as we assume that the reservoir already comprises a complete representation of the ensemble. By the same reasoning, a reservoir structure that is accepted for exchange from the high-temperature replica is left in the reservoir and is not discarded.

When a non-Boltzmann-weighted reservoir is used with R-REMD, a new exchange criterion needs to be derived in order for the exchanges to drive the ensemble to Boltzmann weighting. In the study of Roitberg et al. [156], it was proven that any ensemble could be used for the reservoir; the collection of structures does not need to be Boltzmann weighted as long as its probability distribution is known. Considering *i* replicas associated with *M* temperatures, we explicitly write the exchange probability between two replicas as follows:

$$\begin{aligned}W\left(X_l^l,\ldots,X_j^s,X_k^t,\ldots,X_M^M \rightarrow X_l^l,\ldots,X_j^t,X_k^s,\ldots,X_M^M\right)w\left(X_l^l,\ldots,X_j^s,X_k^t,\ldots,X_M^M\right)=\\ W\left(X_l^l,\ldots,X_k^s,X_j^t,\ldots,X_M^M \rightarrow X_l^l,\ldots,X_j^s,X_k^t,\ldots,X_M^M\right)w\left(X_l^l,\ldots,X_k^s,X_j^t,\ldots,X_M^M\right)\end{aligned} \tag{4.14}$$

which may be rewritten as the product of populations as

$$\frac{W\left(X_j^S,X_k^t \rightarrow X_j^t,X_k^S\right)}{W\left(X_k^S,X_j^t \rightarrow X_j^S,X_k^t\right)}=\frac{w\left(X_k^s\right)w\left(X_j^t\right)}{w\left(X_j^s\right)w\left(X_k^t\right)} \quad . \tag{4.15}$$

Inserting the Boltzmann populations into Equation (4.15) yields the exchange criteria equation from standard REMD:

$$\frac{W\left(X_j^S,X_k^t \rightarrow X_j^t,X_k^S\right)}{W\left(X_k^S,X_j^t \rightarrow X_j^S,X_k^t\right)}=\exp\left\{\left(\beta_t-\beta_s\right)\left(E\left(q^{[k]}\right)-E\left(q^{[j]}\right)\right)\right\} \quad . \tag{4.16}$$

In order to create the reservoir used in [156], a Boltzmann-weighted reservoir was clustered using a structural similarity metric, and the representative structure from each of the resultant 700 clusters was included in the reservoir. In this case, the probability of selecting structure *i* from the reservoir of size *M* is *1/M*. Exchanges between replicas that do not involve the reservoir obey the Metropolis criterion of Equation (4.16), while exchanges between a replica and the reservoir obey the following criterion:

$$\frac{W\left(X_j^R, X_k^t \rightarrow X_j^t, X_k^R\right)}{W\left(X_k^R, X_j^t \rightarrow X_j^R, X_k^t\right)} = \frac{w\left(X_k^R\right)w\left(X_j^t\right)}{w\left(X_j^R\right)w\left(X_k^t\right)} = \frac{(1/M)\exp\left\{-\beta_t E\left(q^{[j]}\right)\right\}}{(1/M)\exp\left\{-\beta_t E\left(q^{[i]}\right)\right\}} =$$
$$\exp\left\{-\beta_t \left(E\left(q^{[j]}\right) - E\left(q^{[k]}\right)\right)\right\}$$

(4.17)

Because the probability distribution of the reservoir is flat, the temperature of the reservoir may be assumed to be infinite $(\beta_R = 0)$, and only the temperature of the replica exchanging with the reservoir is used. The temperature of the reservoir itself is not needed in the expression for the exchange criterion.

In our new implementation of R-REMD, we would like to maintain the use of a non-Boltzmann reservoir that has been clustered by structural similarity; however, instead of using only the representative structure of each cluster for exchange, we would like to include all of the structures in each cluster in order to increase the structural heterogeneity of the reservoir structures available for exchange, and to accelerate convergence, particularly in application to Hamiltonian replica exchange MD in which different forcefields are used. In order to implement this idea, we first tried weighting the value of the Metropolis acceptance probability by a limiting probability that compares the relative populations of the clusters to which the highest-temperature replica structure and reservoir structure under exchange belong. Although the highest-temperature replica structure is not included in the reservoir, before exchange, this structure is compared against the full set of reservoir structures and assigned to a cluster as if it were part of the reservoir ensemble. We then compare the energies of the highest-temperature replica structure and the randomly chosen reservoir structure, and use these energies to calculate the Metropolis acceptance probability.

The probability distribution of each structure in the MD replica may now be written as $w\left(X_j^t\right)=\frac{N_{MD}}{N_{tot}}$, where N_{MD} is the population of the cluster to which structure *j* would belong if it were in the reservoir, and N_{tot} is the total number of structures that have been clustered (in this implementation, as in the work of Okur et al. [158], the size of the reservoir is limited to 10,000 structures; this number may be increased in the future). The probability distribution of the randomly chosen structure from the reservoir is $w\left(X_k^R\right)=\frac{N_{rsv}}{N_{tot}}$, where N_{rsv} is the population of the cluster in the reservoir to which structure *j* belongs, and N_{tot} is the total number of structures. We may now rewrite the Metropolis acceptance criterion as follows:

$$\begin{aligned}&\frac{W\left(X_j^R, X_k \rightarrow X_j^t, X_k^R\right)}{W\left(X_k^R, X_j^t \rightarrow X_j^R, X_k^t\right)}=\frac{w\left(X_k^R\right)w\left(X_j^t\right)}{w\left(X_j^R\right)w\left(X_k^t\right)}=\frac{(N_{MD}/N_{tot})\exp\{-\beta_{MD}E_{rsv}\}}{(N_{rsv}/N_{tot})\exp\{-\beta_{MD}E_{MD}\}}= \\ &\rho=\min\left\{1,\frac{N_{MD}}{N_{rsv}}\exp\{-(\beta_{MD})(E_{rsv}-E_{MD})\}\right\}\end{aligned} \tag{4.18}$$

As outlined below in Section 4.4, test simulations run using this exchange criterion were not successful. Even after systematic changes were made to the reservoir contents in order to discern the influence of reservoir ensemble on the output ensemble, we were unable to determine the source of error in our methodology. We then modified Equation (4.18) to place the decision to accept or reject the structure after the determination of the Metropolis criterion. Thus, rather than scaling the Metropolis criterion itself, we scale how often a structure is accepted or rejected. This six-case formalism was used for subsequent test simulations. As in regular R-REMD, a random structure is chosen from the reservoir for exchange with the structure undergoing MD in the highest-temperature replica. The Metropolis criterion is then calculated using the potential energies of these two structures according to Equation (4.17), and the decision on whether to accept or reject the reservoir structure is based not only on the value of the Metropolis criterion, but also on the value of the weighting factor. According to the formalism of Equation 4.12, if the value of the Metropolis criterion is greater than the chosen random number, this structure would be accepted. However, in this new formalism, the value of the weight factor is now used to determine the frequency at which the structure will be accepted.

We begin with the situation where the value of the Metropolis criterion and random number suggest a favorable exchange between the reservoir structure and the structure in the highest-temperature replica. If $\frac{N_{MD}}{N_{rsv}} = 1$, no correction is necessary to the acceptance frequency; because these two structures are in the same cluster, scaling is unnecessary, and we accept the structure. If $\frac{N_{MD}}{N_{rsv}} > 1$, this indicates that the reservoir structure is underrepresented relative to the high-temperature replica structure. Again, no correction to the acceptance frequency is needed; we accept the structure. If $\frac{N_{MD}}{N_{rsv}} < 1$, the reservoir structure is overrepresented relative to the highest-temperature replica structure. Although the value of the Metropolis criterion would require that we accept the structure based on its energy, the ratio of the weight factors indicates that we should scale back the acceptance frequency of this structure by a factor of $\frac{N_{MD}}{N_{rsv}}$. We therefore generate a second random number against which to compare the value of the weight factor. If the value of the weight factor is greater than this second random number then we accept the exchange, and if the weight factor is less than or equal to the random number, we reject the structure.

We now consider the situation in which the value of the Metropolis criterion and random number suggest that the exchange of structures is rejected. If $\frac{N_{MD}}{N_{rsv}} = 1$, both structures are in the same cluster and no correction is needed to the acceptance frequency. If $\frac{N_{MD}}{N_{rsv}} < 1$, the reservoir structure is overrepresented relative to the structure of the highest-temperature replica; no correction to the acceptance frequenecy is needed, and we reject this structure. If $\frac{N_{MD}}{N_{rsv}} > 1$, the reservoir structure is underrepresented relative to the MD structure. Although the value of the Metropolis criterion would require that we reject the structure based on its energy, the ratio of the weight factors indicates that we should scale up the acceptance frequency of this structure by a factor of $\frac{N_{MD}}{N_{rsv}}$. As before, we generate a second random number against which to compare the value of the weight factor, and if the value of the weight factor is less than the random number, the structure is rejected as the weight factor is not large enough to drive the acceptance.

Tests using both the two-case and six-case exchange criteria were run on the model system alanine dipeptide in implicit solvent. The results of these simulations were unpredictable, and no correlation could be made between the population of the reservoir ensemble and the ensemble that was obtained after the simulation was run. Our inability to identify a source of systematic error indicated that a simpler system was required in order to accurately test this methodology. Following the initial tests, code was written to test this method using only the potential energies of the alanine dipeptide molecule to achieve a random walk in potential energy space. Using only the energies of the peptide allowed inaccuracies in the algorithm to be more easily diagnosed, as we were able to both analytically write the partition function and divide the energy grid into different structural clusters for testing purposes. Results from this coarse-grained simulation are outlined in Section 4.4.

4.3 Model System: Alanine Dipeptide

Terminally blocked alanine dipeptide (AdP), with the amino acid sequence Ace-Ala-Nme (Figure 4.1), is commonly used as a model system for larger nonglycine/nonproline protein backbones in computational studies of conformational sampling. Although this peptide is extremely simple, it is able upon solvation to fully sample the range of φ and ψ dihedral angles that is available to protein alpha helix and beta strand motifs. Additionally, the peptide contains two amide peptide bonds that are capable of hydrogen bonding to one another, as well as to polar solvent molecules. The simplicity of the molecule has allowed its conformational and energetic landscapes to be fully quantified experimentally, as well as theoretically through high-level quantum calculations, as well as Monte Carlo or molecular dynamics approaches. The series of alanine peptides is often used for the parameterization of dihedral angles in molecular mechanics forcefields.

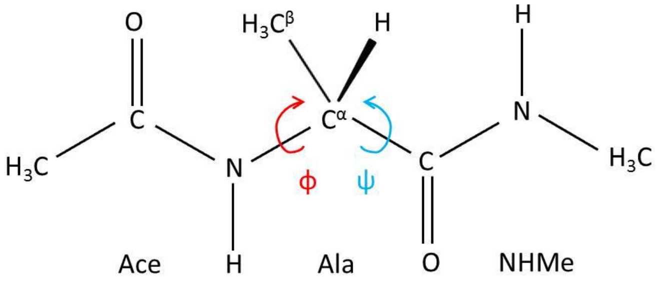

Figure 4.1: Terminally blocked alanine peptide (Ace-Ala-Nme).

4.4 Methods and Results

In this Section, we describe efforts to undertake simulations of alanine dipeptide using a non-Boltzmann-weighted reservoir. We describe the steps taken in the construction of this reservoir, and show results obtained using the exchange formula outlined in Equation 4.18. Because this scheme was unsuccessful, we outline subsequent attempts to uncover the cause of the error in our algorithm. To this end, we created a non-Boltzmann-weighted reservoir with systematic structural variation. When tests using this reservoir were not successful, we wrote code to mimic the random walk of alanine dipeptide in φ/ψ space, using only the potential energy of the structure in order to reduce the complexity of the problem. We attempted four different clustering schemes with this reduced code in order to examine the effect of clustering algorithm on the resulting ensemble; these results are presented below.

In order to obtain a converged trajectory of alanine dipeptide for subsequent construction of a Boltzmann-weighted reservoir, the molecule was built using the Leap module, and a single Langevin dynamics trajectory was run at 400K for 1 μs. Version 10 of the AMBER molecular dynamics package was used for these simulations. A version of the ff99 forcefield with modified backbone parameters to reduce α-helical bias [76] was employed. All nonbonded interactions were evaluated at each MD time step and SHAKE was used to constrain all bonds to hydrogen. The time step used was 2 fs, and the collision frequency used was 1.0 ps^{-1}. All simulations used the Generalized Born (GB) implicit solvent model [155] with GB^{OBC} implementation [157] in AMBER. In order to assess the convergence of the trajectory, a two-dimensional histogram of the free energy dependence on the dihedral angles was plotted (Figure 4.2).

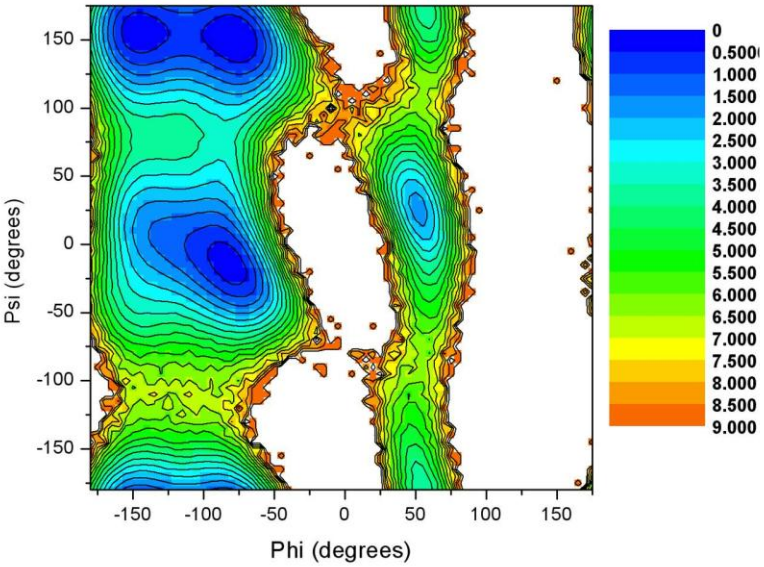

Figure 4.2: Free energy landscape of alanine dipeptide trajectory run for 1μs at 400 K. Energy units are in kcal/mol.

In order to create reservoirs for subsequent R-REMD simulations, single structures were extracted at each 0.2 ns of the microsecond-long trajectory, yielding a total of 5,000 structures. In order to assure that this ensemble of structures was Boltzmann-weighted, its structural content was compared to the 400 K, microsecond-long trajectory from which it was extracted (Table 4.1). The distribution of structures in the ensemble agrees well with a Boltzmann-weighted ensemble of structures [185]. This ensemble of structures was then clustered by structural similarity, with a cutoff of 0.5 Å on the heavy atoms, resulting in 13 distinct structural clusters. The structural content of each cluster was determined and is given in Table 4.2.

Ensemble	**α**	**β**	**P^{II}**	**α^{L}**
1μs MD trajectory at 400 K, all frames	31.9±0.3%	23.7±0.05%	28.4±0.09%	2.4±0.4%
5,000 frames extracted from 1μs MD trajectory	32.4±0.8%	23.0±0.5%	28.9±0.4%	2.6±0.9%

Table 4.1: Structural content of the 1μs MD trajectory at 400K, all frames and the ensemble of structures extracted from that trajectory.

Cluster Number	**Population**	**Percentage of Reservoir**	**Structural Content**
1	1423	28.5%	91% P^{II}
2	948	19.0%	97% α
3	185	3.7%	63% β
4	426	8.5%	87% α
5	1128	22.6%	93% β
6	349	7.0%	72% α
7	173	3.5%	20% β
8	36	0.7%	No discernible content
9	119	2.4%	92% P^{II}
10	93	1.9%	82% α
11	86	1.7%	6% β
12	24	0.5%	83% P^{II}
13	10	0.2%	No discernible content

Table 4.2: Populations and structural content of each of the 13 clusters obtained from clustering the 5,000 structures extracted from the 1 μs alanine dipeptide trajectory at 400 K using a 0.5 Å cutoff on the heavy atoms.

Reservoir replica-exchange molecular dynamics simulations were then run using the 5,000 Boltzmann-weighted structures as the reservoir. The simulation was run assuming Boltzmann-weighting of the reservoir (acceptance probability in accordance with Equation 4.16) in order to provide a point of comparison for future R-REMD runs with a non-Boltzmann reservoir. In this R-REMD simulation of alanine dipeptide, four replicas covering a temperature range of 275-375 K were used. Exchanges between neighboring replicas were attempted at intervals of 1 ps. The REMD simulation was run to 500,000 exchange attempts, for a total of

500 ns per replica. Other parameters were as described above for the LD simulations of AdP. Table 4.3 gives the populations at end of the R-REMD run with the Boltzmann-weighted reservoir of 5,000 structures; these populations agree well with those given in Table 4.1, indicating that the simulation is converged and gives accurate results.

Temperature	α	β	P^{II}	α^{L}
275 K	35.0%	24.7%	30.4%	2.2%
300 K	34.3%	24.4%	30.3%	2.5%
325 K	33.9%	23.7%	28.9%	3.0%
350 K	34.8%	23.3%	28.2%	3.1%

Table 4.3: Populations of structural families from control R-REMD simulation using reservoir of 5,000 Boltzmann-weighted structures.

In order to test the procedure of using R-REMD with a non-Boltzmann-weighted reservoir with user-defined weights, we created a non-Boltzmann-weighted reservoir from the 5,000 Boltzmann-weighted structures by systematically increasing the populations of certain structures in the reservoir. In accordance with this new method, we expect the weight factor to correct for the inhomogeneity of the reservoir. Because the structural contents of each cluster are known (Table 4.2), we can independently increase the populations of different types of structures to create reservoirs containing very different ensembles. We may then compare the reservoir ensemble with the ensemble resulting from the simulation in order to detect any systematic structural undersampling or oversampling.

To this end, we created three non-Boltzmann-weighted (NBW) ensembles through independently increasing the populations of cluster 1, cluster 2, and cluster 5, such that each of these clusters comprised approximately 54% of the NBW reservoir (Table 4.4). These clusters were chosen because they each contain predominantly a single type of secondary structure, so that we may examine the effect on the resultant ensemble of independently increasing the population of a single type of structure and thereby have control over the contents of the reservoir. Table 4.4 outlines the contents of each of the three new reservoirs.

Cluster Number	**Population in BW Reservoir**	**Structural Content**	**Percentage of BW Reservoir**	**NBW Reservoir Population Enhancement**	**Total Number of Structures in Cluster**	**Number of Members in NBW Reservoir**	**Percent of Reservoir**
`1	1423	91% P^{II}	28.5%	x2	4269	7846	54.4%
2	948	97% α	19.0%	x4	4740	8792	53.9%
5	1128	93% β	22.6%	x3	4512	8384	53.8%

Table 4.4: Construction of non-Boltzmann-weighted (NBW) reservoirs from the original Boltzmann-weighted (BW) reservoir of 5,000 frames.

Following the construction of these new reservoirs, we ran non-Boltzmann R-REMD using the cluster weights in order to examine the effect on the output ensemble. Exchanges between the highest-temperature reservoir proceeded according to Equation 4.18, while those between any two replicas proceeded according to Equation 4.16. The results in Table 4.5 are averaged over the four replica trajectories for each system at 275 K, 300 K, 325 K, and 350 K, and should be compared against Table 4.1 above in order to discern their difference from Boltzmann weighting.

Reservoir	**α**	**β**	**P^{II}**	**α^L**
Increased Cluster 1 (91% P^{II})	26.6%	26.8%	35.6%	2.0%
Increased Cluster 2 (97% α)	34.3%	19.5%	23.4%	10.6%
Increased Cluster 5 (93% β)	36.7%	18.3%	22.3%	10.5%

Table 4.5: Average structural content over replica trajectory structures at 275 K, 300 K, 325 K, and 350 K for user-defined non-Boltzmann R-REMD runs using non-Boltzmann-weighted reservoirs with increased populations of specific types of structures.

From these results, it is clear that the reweighting scheme in Equation 4.18 does not adequately correct for reservoir overpopulation of structures belonging to particular clusters. Additionally, the error in the results is not predictable based on the type of structure that is overrepresented in the reservoir, as we hoped it would be when designing the NBW reservoirs. Comparing the results in Table 4.5 to those in Table 4.1, it is evident that increasing the population of polyproline-II helical structures in the reservoir leads to oversampling of polyproline-II structures, undersampling of α-helical structures, slight undersampling of β-sheet structures, and very slight undersampling of the left-handed α-helical structures. Independently increasing the populations of α-helical or β-sheet structures resulted in ensembles with nearly identical populations of structures, with oversampling of α-helical structures, undersampling of β-sheet and polyproline-II structures, and a marked oversampling of left-handed α-helical structures. Careful examination of exchange and acceptance rates of each cluster did not reveal any trends that were helpful in determining the sampling trends that are indicated in Table 4.5.

We thus determined that it was necessary to reduce the complexity of the problem by constraining the alanine dipeptide structures to discrete values of φ and ψ dihedral angles for which we were able to determine the potential energy, and subsequently writing code that would simulate a random walk in the peptide's dihedral angle space. Knowing the energy of each state allows for exact calculation of the partition function, which in turn allows for determination of the correct probability of each state on the energy landscape and a knowledge of areas that are being incorrectly treated by the exchange algorithm.

In order to determine the energy of alanine dipeptide as a function of its dihedral angles, a script was written in the Leap module of AMBER which generated structures corresponding to every 5 degrees in the φ and ψ angles of the Ramachandran plot. The result of running this script was a grid of 5,184 alanine dipeptide structures for which the potential energies are known. Code was then written to simulate a random walk between neighboring states on this grid with exchange probability calculated according to Equation 4.12. Exchanges with the reservoir were simulated by allowing a periodic random move to a non-neighboring grid point, with the probability calculated according to Equation 4.18. The code was constructed so that this grid could be divided into structural regions in order to mimic the clustering of structures by structural similarity that was performed in the unsuccessful tests runs. By systematically changing the regions of clustering, we hoped to gain insight into the effect of the choice of structural clustering metric on the output ensemble. The potential energy landscape of alanine dipeptide is shown in Figure 4.3. A range of tests was performed with this random walk toymodel, including variations in the clustering metrics and the placement of the weight factor.

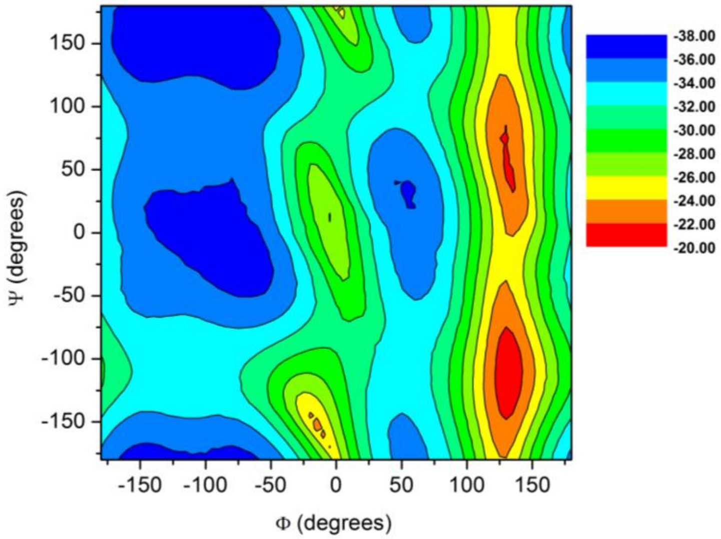

Figure 4.3: Potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space. Energy is in units of kcal/mol.

The first test was to perform a random walk on the potential energy grid with Metropolis-criterion-based exchanges attempted only between neighboring states. No random moves on the energy grid were included, and the exchange probability was calculated according to Equation

4.12. As in REMD exchanges between two temperatures, this equation takes into account the energies and temperatures of the two states that are being considered for exchange. We are not using a reservoir of structures in this test. The random walk was simulated for a total of 100 million steps, and three different temperatures were run: 300 K, 375 K, and 450 K. The results, shown below in Figure 4.4, show the error for each of the 5,184 states in dihedral angle space. This landscape should be compared against Figure 4.3 in order to qualitatively understand which structural areas of the landscape are correctly or incorrectly sampled in this test. Quantitative comparison of Figure 4.4 with Figure 4.3 is not possible, as Figure 4.4 shows the sampling errors in terms of populations, whereas Figure 4.3 shows the landscape in terms of energies. Error was calculated as the difference between the observed probability of that state and the expected probability calculated based on the partition function. The scale of error in Figure 4.4 is in hundredths of a percent; all colored areas on the graph have errors that are bounded by 25% in either undersampling or oversampling. Areas with negative values are oversampled, and those with positive values are undersampled. Regions with error larger than 25% in undersampling or oversampling are shown in white; in Figure 4.4, these are the barrier areas of relatively high energy that have a low probability of being sampled during the simulation. As expected, we see that the sampling improves as the temperature increases; regions that are unexplored and have high error at low temperature, particularly at the barrier regions, see this error diminish as the temperature increases.

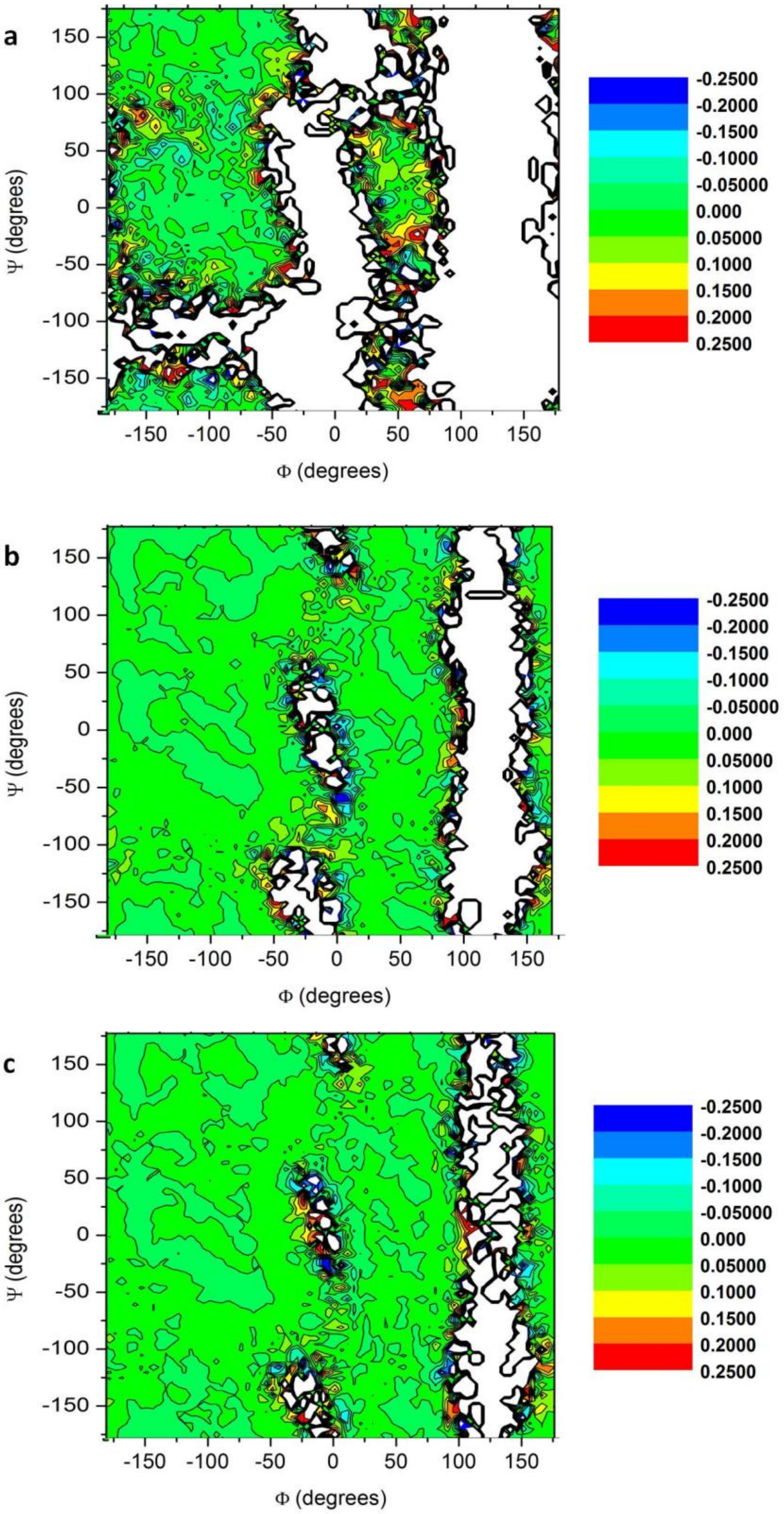

Figure 4.4: Superimposition of population error for nearest-neighbor random walk onto potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space for (a) 300 K, (b) 375 K, (c) 450 K. Errors are in hundredths of a percent.

Following this nearest-neighbor random walk, we then ran the same simulation with the addition of random moves to enable increased exploration of the landscape and to mimic the random exchange attempts of the highest-temperature replica with the reservoir. Nearest-neighbor exchanges were attempted as before, but every tenth move, exchanges were permitted to occur via random exploration or 'jumps' on the energy landscape. Using these jumps, we expect to sample high-energy areas of the landscape that would be inaccessible if only a random walk between adjacent states is performed. Two sets of simulations were run in order to set up a control against which later tests could be compared. In the first test, acceptance of both nearest-neighbor exchange moves and jump moves was determined according to Equation 4.12; no weights were used in the calculation of the exchange probability. In the second test, acceptance of nearest-neighbor moves was determined according to Equation 4.12, while Equation 4.18, with N_{MD} and N_{rsv} both set equal to 1, was used to determine the acceptance of the jump moves. The purpose of the second procedure is to test that the weight calculation has been correctly integrated into the simulation code, as correct calculation of the weight factors will give a result that is equal to that obtained using Equation 4.12.

The results from these two tests were identical, indicating that the algorithm written with the inclusion of Equation 4.18 for jump moves was correctly written and implemented. Results from only the second test are shown in Figure 4.5. When compared with Figure 4.4, these results indicate that sampling is much improved, particularly at lower temperature; the inclusion of a random jump move allows a greater area of the energy landscape to be explored than in the case of only nearest-neighbor moves. As before, the sampling error is seen to decrease as the temperature increases. Sampling errors are low; all states shown in color Figure 4.5 have undersampling or oversampling errors within 25%, and the majority of the landscape exhibits sampling errors that are less than 5%. Regions shown in white have errors that are larger than 25% in undersampling or oversampling.

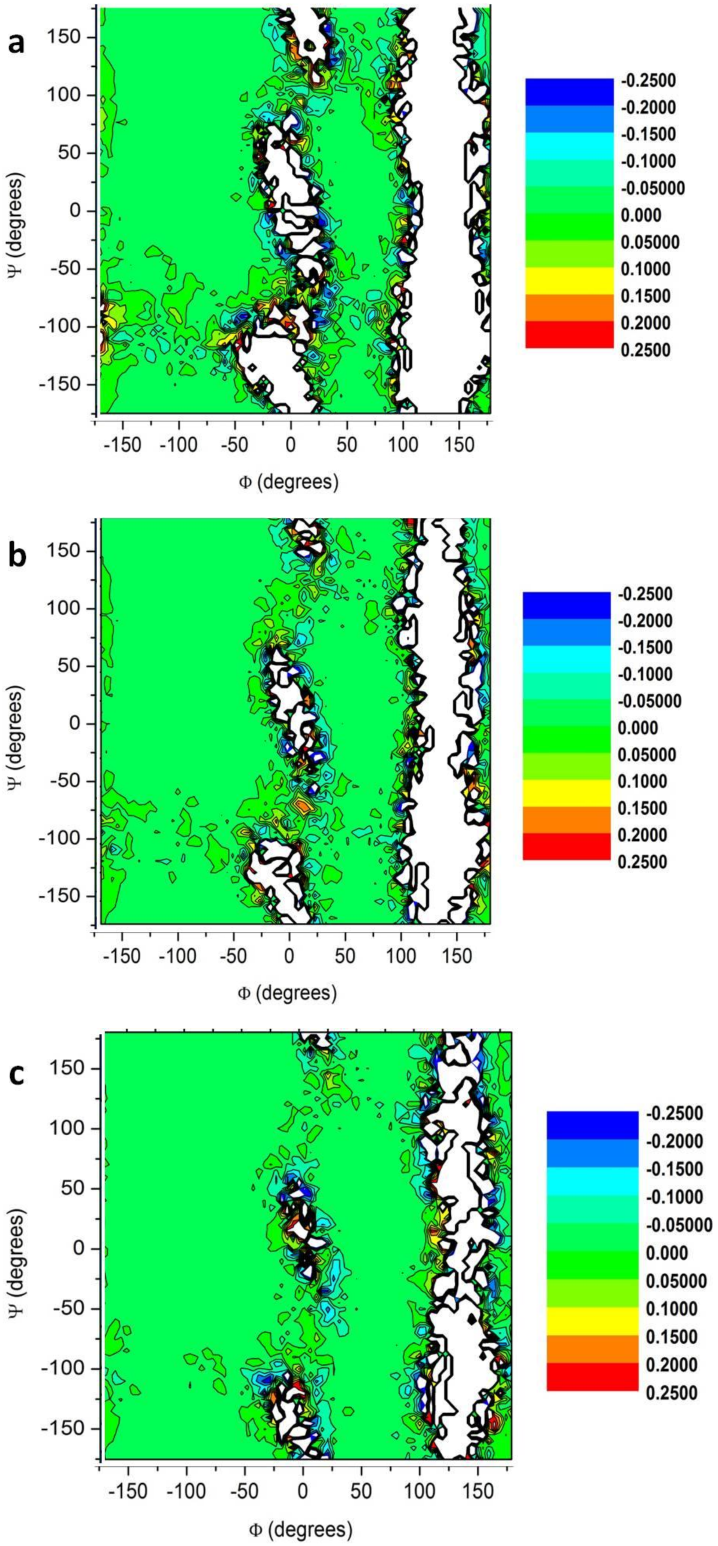

Figure 4.5: Superimposition of population error for nearest-neighbor random walk with random landscape exploration onto potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space for (a) 300 K, (b) 375 K, (c) 450 K. Errors are in hundredths of a percent.

In order to investigate the effect of using Equation 4.18 for exchanges, the energy landscape was divided into areas of structural similarity in order to mimic the structural clustering of the reservoir and replica structures that was used to weight the exchanges in the simulation of alanine dipeptide. Using the potential energy landscape as a guide, clusters were created by visually dividing the peptide's dihedral angle space into discrete structural regions. For each cluster, delineated in Figure 4.6 by bold black lines, the population and average energy were calculated. The average energies of each cluster were not used in the calculation of the Metropolis criterion, but are rather shown here to give a sense of the energies of the structures contained within each cluster, and to give an idea of whether certain energetic areas of the landscape were excluded from exchanges. Figure 4.6 shows two clustering schemes that were used.

The first scheme did not distinguish between the β and P^{II} basins, and separates both of the higher-energy barrier regions at $-50° < \varphi < 25°$and $75° < \varphi < 180°$ from the basin regions at -$180° < \varphi < -50°$ and $75° < \varphi < 180°$. The second scheme divides the β and P^{II} basins and places the highest-energy barrier regions at $75° < \varphi < 180°$ in their own clusters, but includes structures from the barrier region at $-50° < \varphi < 25°$ in the clusters containing structures from the adjacent basins at $-180° < \varphi < -50°$. Using these differing clustering schemes, we hoped to determine the effect of the choice of clustering metric on the ensemble that is output after the simulation. In each of these two tests, 10 nearest-neighbor moves with acceptance probability determined according to Equation 4.12 were followed by a random jump on the grid which used Equation 4.18, with N_{MD} and N_{rsv} equal to the cluster weights, to determine the acceptance of the jump moves.

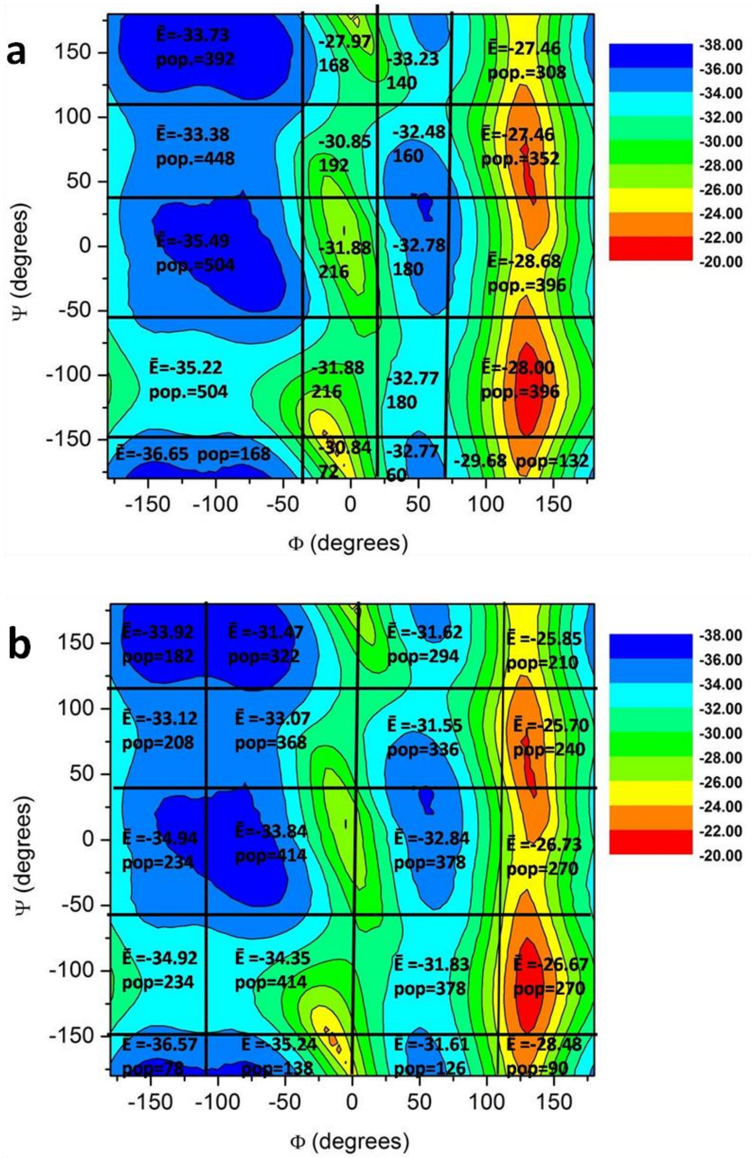

Figure 4.6: Potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space showing division of landscape into areas of structural similarity. Figure (a) considers the β and P^{II} basins to be continuous, whereas Figure (b) divides those structural basins into two discrete clusters. For each such area, the average energy Ē and population of the cluster is given. Energy is in units of kcal/mol.

As is evident from Figure 4.7, the addition of the cluster weights to the acceptance probability calculation lead to serious errors in the sampling. In Figure 4.7, colored regions outline areas where the error in undersampling and oversampling ranges from -100% to +100%. Areas shown in white have sampling errors that are larger than 100%. For all temperatures, the error in undersampling appears to be largest in the region defined by -180°< φ < -25° and -50°< ψ<50°, which exhibits error ranging from approximately 40%-70%. This region corresponds to the cluster containing the α-helical basin, which is the most populated cluster in the scheme presented in Figure 4.6a. At low temperature, sampling errors are generally within an error of 25% in the rest of the region defined by -180°< φ < -25°, although the areas of 150°< ψ<180° and -125°< ψ<-100°do exhibit large oversampling errors.

As expected, the sampling rate of the higher-energy areas of the landscape does increase with temperature; however, the lower-energy areas of the landscape do not exhibit more accurate sampling as the temperature is increased. As temperature increases, the undersampling of the α-helical basin increases. Undersampling also becomes markedly worse in the transition region separating the α-helical basin from the β/P^{II} basin at 35°< ψ<110°. Two areas that are oversampled at lower temperatures, those centered on (φ = -100°, ψ = -100°) and (φ = -75°, ψ = 150°), show a decrease in oversampling with temperature.

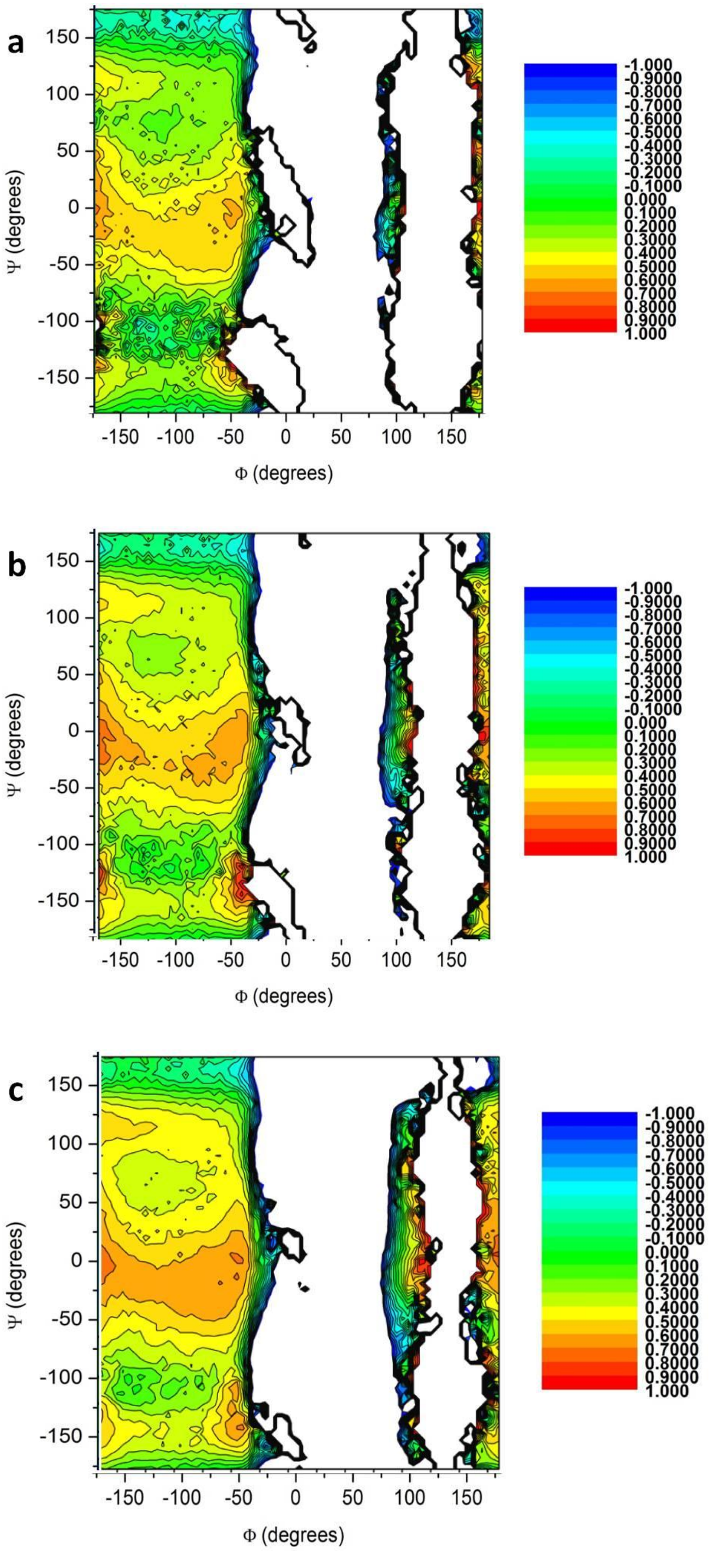

Figure 4.7: Superimposition of population error for nearest-neighbor random walk with random landscape exploration using cluster weights onto potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space for (a) 300 K, (b) 375 K, (c) 450 K. Errors are in hundredths of a percent.

Figure 4.8 shows the population errors after the clustering scheme is used in which the α-helical and β/P^{II} basins are divided, as shown in Figure 4.6b. In Figure 4.8, the error in undersampling and oversampling ranges from -100% to +100% is shown in color; areas in white exhibited errors that are greater than 100%. In comparison with Figure 4.7, sampling is now exhibited in the regions of the left-handed α-helical basin (centered on φ = 50°). Examining the two cluster schemes shown in Figure 4.6, we note that structures from the higher-energy barrier region centered on φ = 0° are contained in discrete clusters in the clustering scheme in Figure 4.6a, whereas the clustering scheme in Figure 4.6b incorporates these higher-energy structures into the clusters containing the lower-energy P^{II} and α-helical basins.

Despite this global improvement in the area of the landscape that is sampled using this clustering scheme, the rates of sampling remain incorrect. A high rate of undersampling (40-50%) is again exhibited in the region of the α-helical basin. In the sampling scheme shown in Figure 4.6b, this basin is divided into two clusters, and only one of those clusters is undersampled; the other exhibits sampling that is within a smaller error bound (~25%). The left-handed α-helical basin is also seen to exhibit significant undersampling. Oversampling is exhibited in the area of the β-sheet cluster, as well as in the transition region between the α-helical and β-sheet clusters centered on (φ = -150°, ψ = -125°). Both oversampling and undersampling errors are seen to increase as the temperature increases.

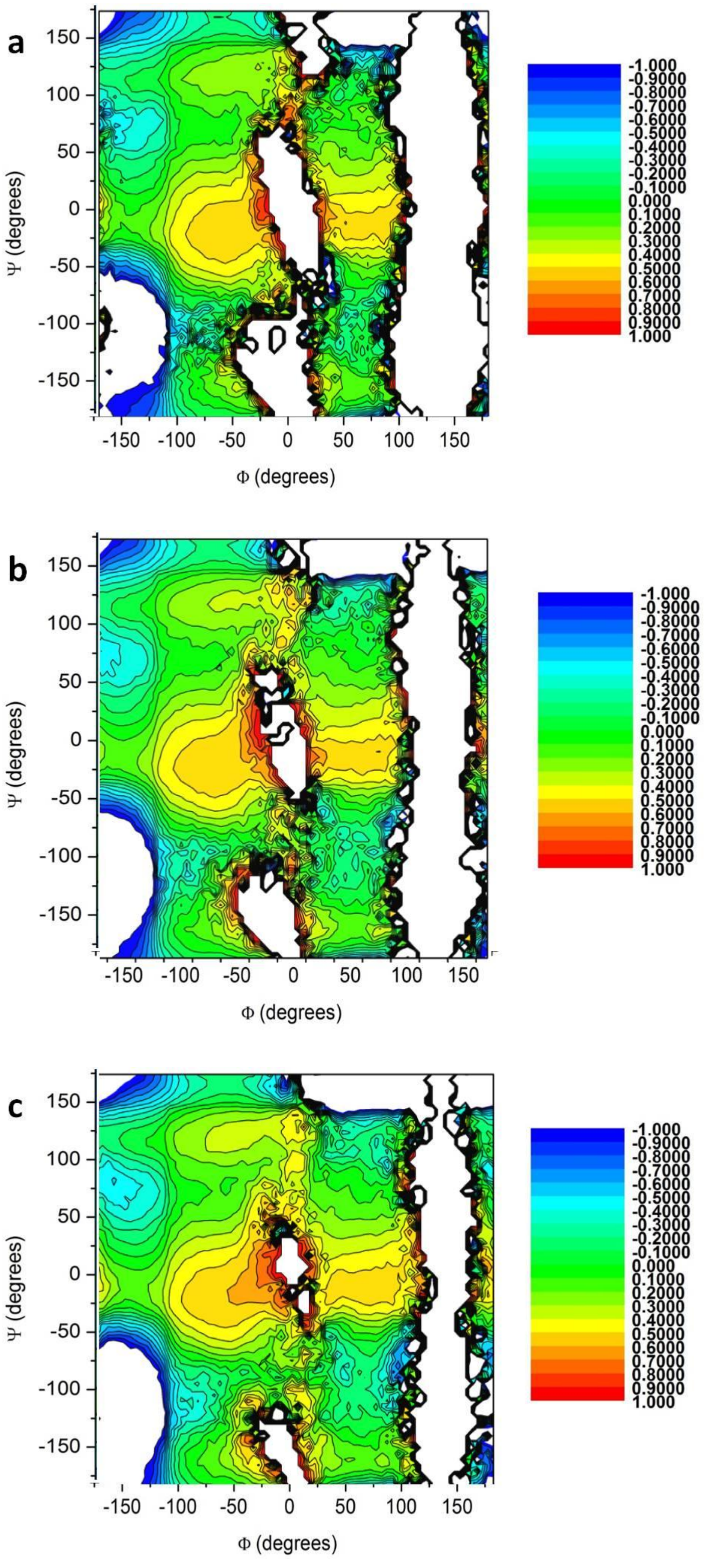

Figure 4.8: Superimposition of population error for nearest-neighbor random walk with random landscape exploration using cluster weights onto potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space for (a) 300 K, (b) 375 K, (c) 450 K. Errors are in hundredths of a percent.

It is remarkable that such a small change in the clustering scheme results in such a large change in the rates of sampling of certain clusters. Comparing the clustering schemes presented in Figure 4.6 and their results in Figures 4.7 and 4.8, it is evident that separating the high-energy barrier regions into their own low-population clusters did not drive sampling of these structures. According to Equation 4.18, structures belonging to clusters with small populations should be favored in the exchange of structures. From the results in Figures 4.7 and 4.8, it appears that the α-helical basin was undersampled in both schemes. Because its cluster had the largest population in both schemes, we may conclude that the scaling factor in Equation 4.18 works to scale back these structures, although it appears to be scaling these structures too often. Similarly, in the second clustering scheme, the relatively large population of the left-handed α-helical basin is undersampled, its population having been decreased through application of Equation 4.18. The structures in these basins should be energetically quite favorable to accept during an exchange, so it is surprising that they are undersampled using both of these clustering schemes.

It is notable that the second clustering scheme, in which each cluster had a larger range of energies among its structures, was more successful in overall sampling of the landscape than the first scheme. In the clusters obtained using the algorithm described above for the atomistic treatment of alanine dipeptide, in which a reservoir was created using a high-temperature reservoir and structures were grouped and assigned to clusters using the metric of backbone RMSD, there was a large energetic heterogeneity within each cluster. In order to test whether the energetic homogeneity of the clusters as defined in Figure 4.6 was the root cause of the sampling errors exhibited in Figures 4.7 and 4.8, the structures on the energy landscape were clustered by structural similarity with a cutoff of 0.5 Ǻ on the heavy atoms. This metric is the same as that described above in the protocol to create the Boltzmann-weighted and non-Boltzmann-weighted reservoirs. These cluster populations were then used with Equation 4.18 during random jump exchange attempts on the potential energy landscape shown in Figure 4.3. We also decided to test both the formalism of Equation 4.18, as well as the the six-case formalism outlined above, in which the value of the cluster scaling factor is used to scale the rate of acceptance rather than the Boltzmann factor, in order to determine the effect on the sampling.

Results from the simulations using the cluster weights determined through backbone RMSD are presented in Figure 4.9a for the two-case exchange criterion using Equation 4.18, and in Figure 4.9b for the six-case exchange criterion to scale the rate of acceptances as outlined above. Colored areas exhibit error in undersampling and oversampling ranging between -100% and +100%; areas in white have larger error. Results from simulations employing both of these exchange schemes are extremely similar. In both cases, oversampling with an error of approximately 25-50% occurs at the regions of transition that are adjacent to the potential energy minima of the β/P^{II} basin, the α-helical basin, and the basin with left-handed α-helical structures. The basin regions have errors ranging from approximately 25% oversampling to approximately 25% undersampling. Neither of these exchange algorithms appeared to yield a Boltzmann-weighted ensemble.

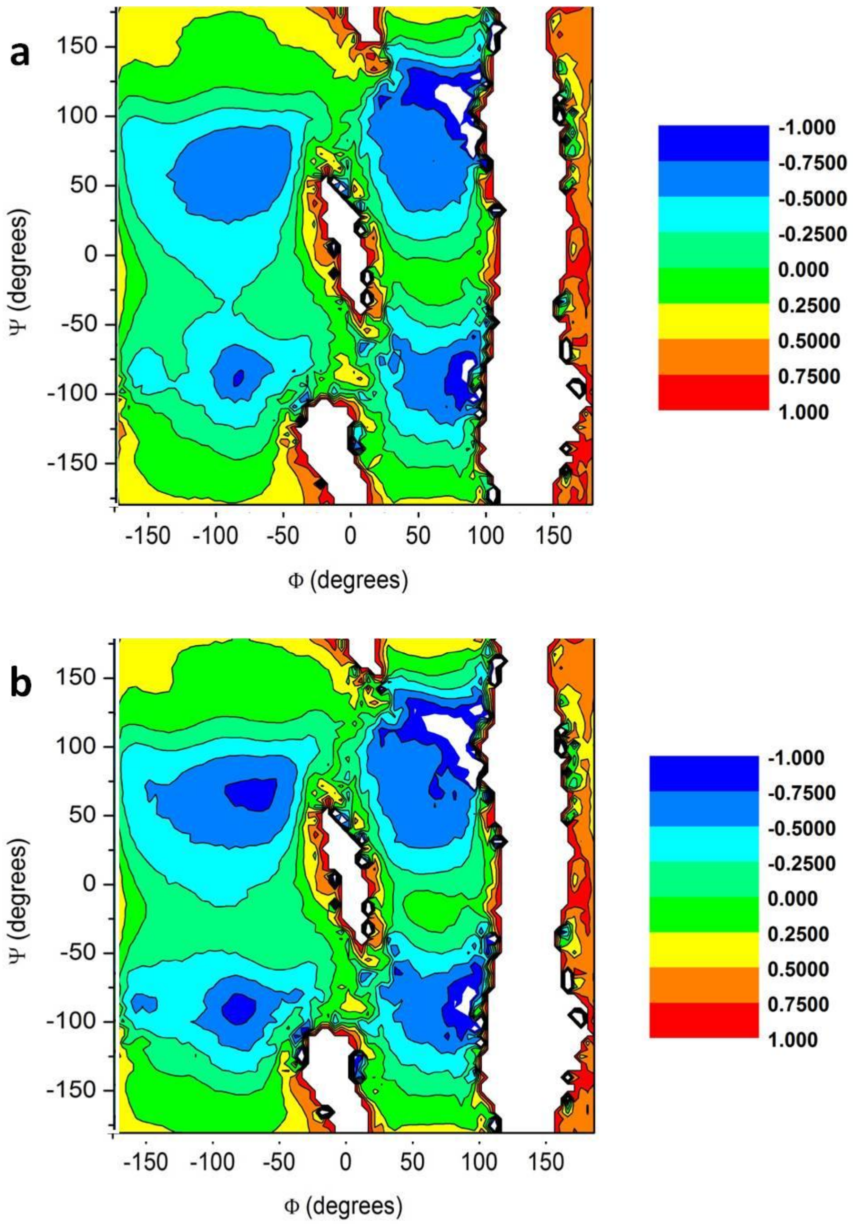

Figure 4.9: Superimposition of population error for nearest-neighbor random walk with random landscape exploration using cluster weights onto potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space for (a) 2-case exchange criterion, (b) siz-case exchange criterion. Errors are in hundredths of a percent. T=300K.

Thus far in our tests, we have been using the populations of the structures under exchange in order to weight their acceptance probability. We recall Equation 4.18, in which the probability distributions of the structures between which the exchange is being attempted are used to weight their Boltzmann factors. Strictly, these probability distributions should be some function of the potential energy. We therefore attempted a final simulation scheme in which the probability densities of the potential energies of the structures under exchange are used to weight their Boltzmann factors, as follows:

$$\frac{W\left(X_j^R, X_k \rightarrow X_j^t, X_k^R\right)}{W\left(X_k^R, X_j^t \rightarrow X_j^R, X_k^t\right)} = \frac{w\left(X_k^R\right)w\left(X_j^t\right)}{w\left(X_j^R\right)w\left(X_k^t\right)} = \frac{E_{MD}}{E_{rsv}} \exp\{-(\beta_{MD} - \beta_{rsv})(E_{rsv} - E_{MD})\} \quad . \qquad (4.19)$$

In this equation, E_{MD} and E_{rsv} are the potential energy densities of the structure from the highest-temperature replica and the structure from the reservoir, respectively. Potential energy densities were determined by grouping all of the energies on the potential energy landscape of the alanine dipeptide molecule by a specific energy cutoff.

The result of the simulation using the potential energies to weight the Boltzmann factor is shown in Figure 4.10. In this simulation, an energy cutoff of 1.5 kcal/mol was used to define the energy clusters; other values were used for the energy cutoff, but 1.5 kcal/mol was seen to yield the most accurate results. Errors in undersampling and oversampling ranged from -25% to 25%. Areas shown in white had larger errors in sampling. This simulation was run at 300 K. From the results below, it is evident that weighting the Boltzmann factors by the potential energy is the most successful of the schemes attempted thus far. The greatest errors in oversampling are observed in the transition regions centered on (φ = -100°, ψ = -125°), (φ = 60°, ψ = -125°), and (φ = 60°, ψ = 110°), which comprise structures that are higher in energy than the structures in the adjacent energy basins. Oversampling is also observed in regions bordering peaks of high energy. Oversampling in these regions indicates that the algorithm is attempting to increase the population of high-energy structures in the annealed ensemble. The regions of the α-helical and left-handed α-helical basins are also oversampled by approximately 5-10%. The region of the β/P^{II} well exhibits undersampling ranging from 0-25%.

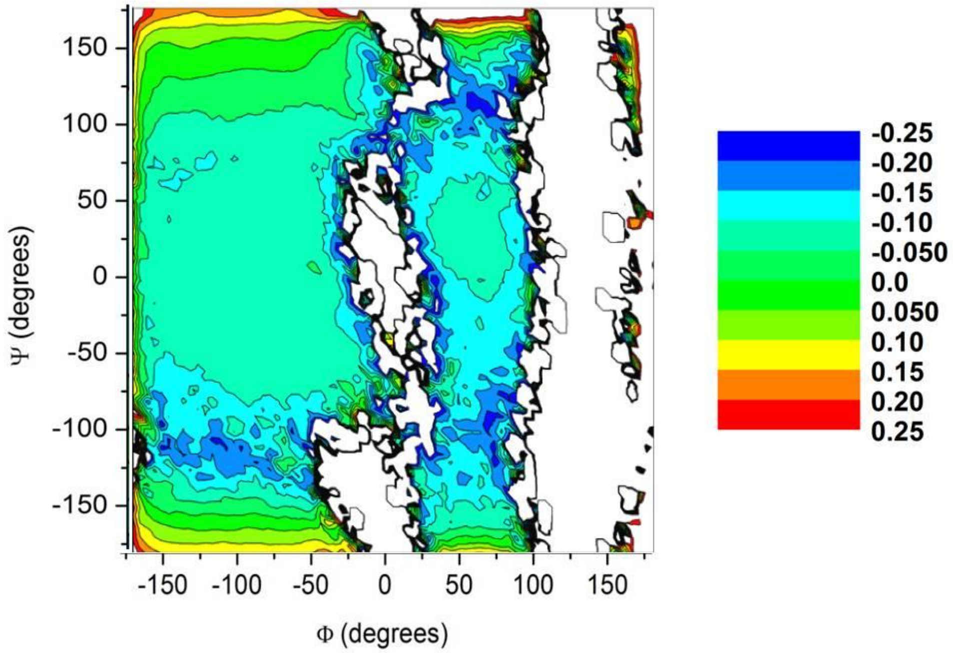

Figure 4.10: Superimposition of population error for nearest-neighbor random walk with random landscape exploration using cluster weights onto potential energy landscape of alanine dipeptide obtained through construction of states in phi/psi space. Structures were clustered by the value of their potential energy with a cutoff of 1.5 kcal/mol. Errors are in hundredths of a percent. T=300K.

4.5 Summary and Conclusions

Accurate conformational sampling remains one of the challenges facing the field of biomolecular simulation. Through development of a variant of replica-exchange molecular dynamics, we have aimed to provide an algorithm with enhanced sampling and increased computational efficiency. This algorithm uses a structural reservoir to decouple the high-temperature search of conformational space, which is often a bottleneck in REMD simulations, from the annealing exchanges to lower temperatures. This reservoir does not need to contain an ensemble that is Boltzmann-weighted, which further decreases the computational demand of the algorithm.

In this chapter, we have outlined attempts to use structural similarity clustering as a metric by which to drive the exchanges from the non-Boltzmann-weighted reservoir to produce a Boltzmann-weighted ensemble. In the first cases of testing the algorithm, alanine dipeptide was used as the model system for simulations. The algorithm did not yield successful results, and even after the systematic construction of three non-Boltzmann-weighted reservoirs of known content, errors were not able to be remedied. We therefore constructed a simpler system which placed the potential energies of alanine dipeptide on a grid at every 5° in the peptide's φ/ψ space. Simulations run using this system considered random-walk exchanges between neighboring grid points in order to mimic the MD simulation that occurs in each replica in an REMD simulation. Periodic 'jumps' by the random walker to a random point on the landscape were used to mimic the selection of a reservoir structure. Three different structural clustering schemes were used in order to test this method, but none were successful. A greater accuracy of results was observed when the probability densities of the energies of the two structures was employed as the weighting factor, rather than their relative probabilities based on their structure. The success of the initial tests of this energy clustering scheme indicate that this may be the correct avenue to pursue in the future with regards to this algorithm.

5. Molecular Dynamics Simulations of Star Polymeric Molecules with Diblock Arms, a Comparative Study

Abstract

The utilization of nanoparticles has become widespread in the field of medicine, where their application extends from the cellular to the organ levels. The advancement of targeted drug delivery systems has occurred alongside of advances in polymer chemistry that have increased the synthetic possibilities of polymeric materials. These materials may be used to increase the solubility of drug molecules, target the drug molecules to a particular type of cell within the body, facilitate their transport to that target, and control their rate of release. Targeted drug delivery requires that the delivery platform be compatible with the drug, customizable with respect to the biological target to be reached, nontoxic, and biodegradable. The success of polymeric nanoparticles in this application is evidenced by the clinical trials that are currently underway for polymeric drug delivery platforms. Improvements in the synthetic potential of polymers, along with the increase in numbers of identified drug-like molecules, have lead to a proliferation of combinatorial possibilities for pairings of drugs and their polymeric carriers. On one hand, these advancements give therapies the potential to be tailored to a specific patient or type of disorder; while on the other, they necessitate a rational methodology for the design and optimization of such therapies.

Computational methods have the potential to reduce the time and expense that are required by the material design and optimization that occur in the laboratory. In this work, we have performed all-atom explicit solvent molecular dynamics simulations of three different star polymeric systems in water, each star molecule consisting of 16 diblock copolymer arms bound to a small adamantane core. The arms of each system consist of a relatively hydrophobic block

(either polylactide, polyvalerolactone, or polyethylene), and an outer hydrophilic block of polyethylene oxide (PEO). These models exhibit unusual structure very close to the core that is clearly an artifact of our model, but which we believe becomes bulk-like at relatively short distances from this core. We report on a number of temperature-dependent thermodynamic (structural and energetic) properties as well as kinetic properties. Our observations suggest that under physiological conditions, the hydrophobic regions of these systems may be solid and glassy, with only rare and shallow penetration by water, and that a sharp boundary exists between the hydrophobic cores and either the PEO or water. The PEO in these models is seen to be fully water-solvated at low temperatures but tends to phase separate from water as the temperature is increased, reminiscent of a lower critical solution temperature exhibited by PEO-water mixtures. Water penetration concentration and depth are strongly composition- and temperature- dependent, with greater water penetration for the most ester-rich star polymer, polylactide. It is our hope that the results of this study may be extended and used to determine the utility of these materials for drug delivery applications.

Acknowledgements

The material presented in this chapter contains direct excerpts from the manuscript by William C. Swope, Amber C. Carr, Amanda J. Parker, Joseph Sly, Robert D. Miller, and Julia E. Rice, published in the Journal of Chemical Theory and Computation in 2012, volume 8, number 10, pages 3733-3749. Co-author Dr. William Swope listed in the publication directed and supervised the research that forms the basis for this chapter of the dissertation. The manuscript was written by Dr. William Swope with suggestions and revisions from Amber Carr. Simulations were performed by Dr. William Swope. Data analysis was performed by Amber Carr and Amanda Parker.

5.1 Introduction

There is growing interest in the use of biocompatible polymeric nanoparticles for drug delivery. The hope is that such materials can be engineered to absorb therapeutic (drug) molecules before their delivery into the body and then to release them in a controlled or programmed manner under physiological conditions. Moreover, such nanoparticles could be functionalized on their exteriors to adhere to the membranes of cells of particular tissue types, or even to cells in a particular disease state. Such targeted delivery systems would result in much more effective therapies than can be achieved by normal means, with a consequent lowering of dosage and reduction of potential side effects.

Polymer chemists have shown amazing ingenuity [186,187,188] in producing polymers for these types of applications using complex sequences of monomeric units (e.g., diblock, triblock, and/or random copolymers) with various types of chemical functionality, a range of topologies, and various types and amounts of covalent and noncovalent cross-linking. Candidate nanoparticles under consideration for drug delivery include micelle and vesicle assemblies made from polymers and polymer blends [189,190], as well as unimolecular systems with dendritic [191] and star [188] topologies and nanogel star polymers [186,187], which have polymeric arms emanating from a nanogel core. Molecular systems with star, nanogel star, and dendritic topologies show promise over molecular assemblies because, being entirely covalently bonded, they are much more likely to be structurally stable over the range of environmental conditions seen in a living organism. However, all of these systems may be useful in different contexts or for different applications.

Of particular interest in this area is the design of general purpose vehicular nanoparticles where, with a relatively small range of chemical or topological variation, one would be able to engineer delivery of a potentially large number of different types of cargo molecules to provide precise control of their release rate and/or target delivery site, or to transport multiple drug types simultaneously. To enable these types of functional variation using, for example, a star polymer topology, the polymer chemist has the freedom to change the number and length of the arms, and each arm may itself be a diblock or triblock polymer with different segment lengths. Furthermore, with nanogel star polymers, the arms of such a molecule do not need to be identical in length and composition. Through variation of these features, a polymer chemist can, in principle, design star molecules with multiple compartments tailored for different types of cargo.

We are reminded that star polymer topologies exist in much broader classes of materials than just those with small molecule, dendritic or nanogel core junctions, such as in the case of polymer-coated gold nanoparticles [192,193]. Even micelles, though not chemically bound, can have polymer-solvent interactions that bear a strong resemblance to those of the star polymers. Star polymers can be synthesized and modeled with much smaller molecular systems, yet can serve as useful and characterizable models for the polymeric structure and solvent-nanoparticle interactions of these more complex systems.

This work concerns the study of an important class of polymeric nanoparticle, namely, the star polymer topology where each arm of the star is itself a diblock copolymer. In general, each arm of these molecules consists of a relatively hydrophobic region positioned close to the interior of the star and a relatively hydrophilic region on the exterior that serves to make the molecule water-soluble and prevent aggregation at finite concentrations.

Because of their use in a number of commercial application areas, star polymers have been studied extensively from both a theoretical and experimental perspective. Comprehensive reviews have been developed by Grest et al. [194] as well as by Likos [195]. Both of these reviews also provide numerous references to a large body of Monte Carlo and molecular dynamics simulation studies, most of which have employed coarse grained (e.g., beaded string) models with intramolecular interactions designed to model good and poor solvents in an implicit way. These simulations have been very useful in helping to interpret experimental results and to validate theoretical scaling laws that describe the relationship, for example, between the radius of gyration and chain number and length in various qualities of solvent. On the other hand, the literature of all-atom and explicit solvent simulations of star polymers is relatively sparse, particularly of star polymers with diblock arms. Ganazzoli et al. [196] used Monte Carlo techniques and a mean field type of approach to determine how the arms of a star polymer might behave in a generic poor solvent. Chang et al. [197] studied heteroarm copolymers, with some arms purely hydrophobic and some hydrophilic. Lee and Larson [139,198] used coarse-grained molecular dynamics to model star polymers with a range of arm numbers and lengths with relatively long polyethylene oxide (PEO) arms bound to different sizes of dendrimeric cores.

The work most relevant to this study is that of Huynh et al. [63]. They reported on a set of all-atom explicit solvent simulations of six-arm star polymers that explored the effect of varying the lengths of the hydrophobic and hydrophilic segments of each diblock arm made from polycaprolactone (PCL, a polyester with five methylene groups between ester groups) and PEO. Each of the six arms was attached to one of the terminal carbon atoms of diethylether. Using the OPLS force field [68,199] (much like the work reported here) and the SPC [200] water model, they studied 13 arm length variants, each with long simulations (200 ns) at 300 K, and were able to establish several important scaling relationships. In their simulations, they observed that the hydrophobic material is densely packed and excludes both water and the hydrophilic material (i.e., strongly segregated, or phase separated), with a well-defined boundary between the hydrophobic material and water. Also, the PEO is highly mobile and adopts disordered conformations, and it is well solvated. If the PEO segments are short, the densely packed hydrophobic region is somewhat solvent-exposed. However, as the length of the PEO segments is increased, the fractional coverage of the hydrophobic core by PEO increases, providing "protection" from water, leading them to suggest that sufficiently long PEO segments might inhibit aggregation in solutions of star copolymers at higher concentrations.

Whereas the work of Huynh et al. focused on the varying the polymer chain length for a single set of hydrophobic and hydrophilic materials at one temperature, the main goal of this work is to investigate the effect on diblock star polymer structure, stability, and kinetics of changes in the hydrophobic region over a range of temperatures. This work examines three different polymers for the construction of the hydrophobic region of each diblock arm: polylactic acid (PLA), polyvalerolactone (PVL) and polyethylene (PE). These differ in the amount of ester versus alkane content, with PLA being the most rich in ester functional groups and the least rich in alkane; and PE being the least rich in ester (having none), and most rich in alkane. The alkane component provides flexibility as well as hydrophobicity due to its nonpolar nature. Ester groups, in contrast, have an effective charge distribution that produces relatively strong electrostatic interactions, and the hydrophobicity of ester-rich polymers is due to their stronger attraction to other ester groups than to water, causing them to "phase separate" from water, as the ester-rich condensate is more stable than water-solvated conformations. We note that PVL is very similar to the polycaprolactone (PCL) studied by Huynh et al.[63], having four methylene groups between ester groups, one less than PCL.

The star polymers studied in this work have arms that are bound to a small adamantane junction. Star polymers with an adamantane junction have actually been synthesized, as reported by Huang et al. [63]. The star polymers they prepared had four arms, compared with our 16, and theirs consisted of different polymeric materials, including styrenes and methacrylates. The four arms on their star polymers were not copolymers, and they were much longer than the ones we have simulated. We wish to emphasize at the outset, however, that our study is not meant to be about any particular adamantane-based star polymer. We are using this type of junction simply to generate a model that can be determined to exhibit bulk-like behavior at a quantifiable distance from its center, and will therefore be useful to help understand much larger star polymers of a similar composition.

This study makes use of fixed charge force fields, where a single charge model is used regardless of the environment in which an atom is situated. This is clearly an approximation since, from the physics of the situation, one would expect electronic polarization for a molecule that depends in degree upon its environment. Even simple reaction field theory [63], for example, predicts that the surrounding solvent causes an enhancement of the polarization of a molecule that depends on the dielectric constant of the solvent. One might reasonably expect, therefore, that the appropriate charge model to use for an ester surrounded by water would be different than for one surrounded by alkane moieties or other ester groups. Therefore, understanding the structure and energetics of diblock star polymers might imply a need for, at least, polarizable force fields [201,202,203,204,205,206]. However, the development and validation of these force fields is currently an evolving field and their computational cost is still significantly greater than that of fixed charge force fields. Therefore, fixed charge force fields are still a good starting point for the study these types of molecular systems, pending the availability, validation and improved performance of more sophisticated treatments. Additionally, as outlined in the chapters above, extensive and successful work has been done in the application of fixed-charge force fields to biopolymers.

Since the use of fixed charge force fields raises questions about the quality of the balance of the intra- and inter-component interactions among the three types of components of these systems (hydrophobic polymeric material, hydrophilic polymeric material, and water), this work examines the star polymer behavior as a function of temperature. Behavior in the simulations that is seen to persist over a wide temperature range under physiological conditions is more likely to be predictive than behavior that is very sensitive to temperature near physiological conditions. Also, studying the temperature-dependent behavior of star polymer structure and kinetics will provide a point of comparison with similar changes in bulk polymeric material properties that occur at glass transition and melting temperatures.

This chapter is structured as follows: Section 2 describes our methods, including a description of the molecular systems and the force fields and methods used to prepare, equilibrate and simulate them, as well as the types of analyses performed to determine structural, thermodynamic, and kinetic properties. Section 3 presents results, and Section 4 is a discussion of these results as well as predictions to be experimentally tested. Finally, Section 5 presents general conclusions, and also discusses further possible implications of this work for the design of star diblock copolymers for drug delivery.

5.2 Methods

5.2.1 Molecular Systems

Each of the star polymer systems was prepared by connecting 16 linear diblock copolymer arms to an adamantane junction. The carbon atoms of adamantane, $C_{10}H_{16}$, have a rigid 10-atom diamond-lattice structure (Figure 5.1a). The 16 sites that are hydrogen atoms in adamantane were used as the connection sites for the hydrophobic part of each diblock arm. For each star system, the hydrophobic part of each arm was then connected to a short chain of six polyethylene oxide units. The first star polymeric system consisted of 16 arms each with sixteen monomeric units of L-lactic acid (LA=-C(HCH$_3$)-CO-O-), a linker methylene unit (-CH$_2$-), then six units of ethylene oxide (EO=-CH$_2$-O-CH$_2$-), the last of which was terminated with a hydrogen atom to form a terminal methyl group. This system can also be described as A[LA$_{16}$-CH$_2$-EO$_6$-H]$_{16}$. The methylene group was included to link the PLA appropriately to the PEO. Each adamantane carbon is bonded to the stereocenter carbon of the first lactic acid unit in either one or two arms. This system will later be referred to as the polylactic acid (PLA) star polymer. The second type of star polymer system was constructed using eight units of delta-valerolactone (VL=- CH$_2$-CH$_2$-O-CO-CH$_2$-CH$_2$-) for the hydrophobic part of the arm, followed, as before, with six units of ethylene oxide terminated with a methyl group. This system is described as A[VL$_8$-EO$_6$-H]$_{16}$, and will be referred to as the polyvalerolactone (PVL) star polymer. The third type of star polymer system was constructed using 12 units of ethylene (E=-CH$_2$-CH$_2$-; i.e., an alkane chain of 24 carbon atoms) for the hydrophobic part of the arm, followed with six units of ethylene oxide terminated with a methyl group. This system is described as A[E$_{12}$-EO$_6$-H]$_{16}$, and will be referred to as the polyethylene (PE) star polymer.

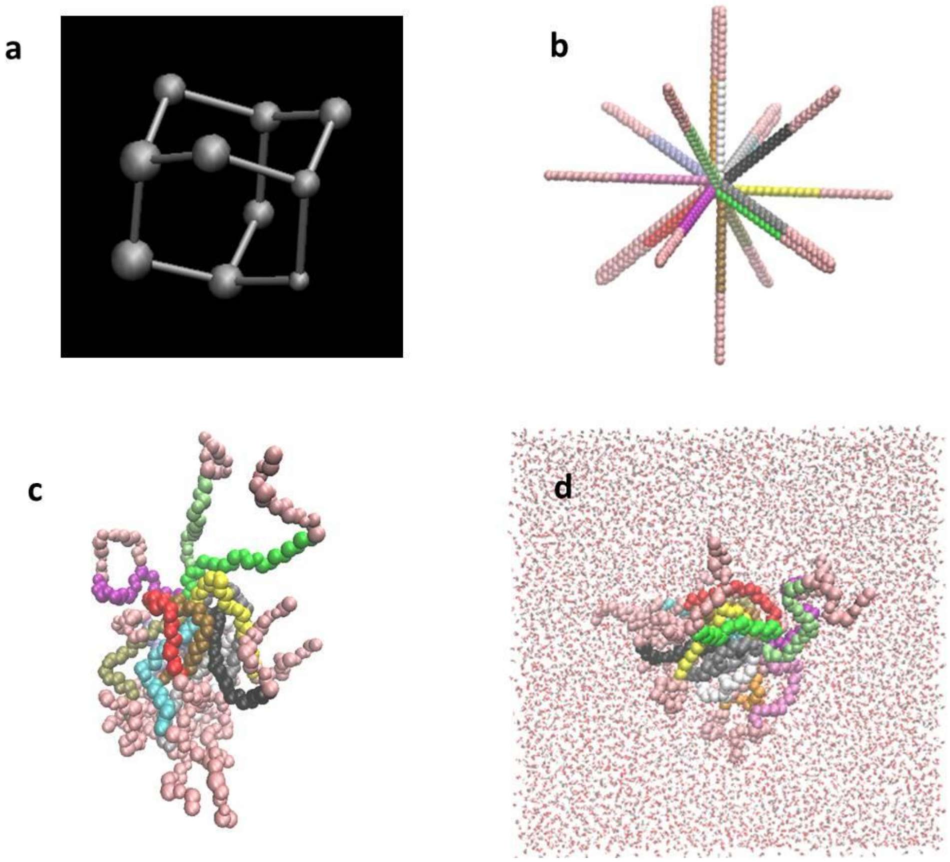

Figure 5.1: Schematic representations of star polymer construction for this study. a) The adamantane junction showing the frame of 10 carbon atoms, four of which support the attachment of one arm and six of which support two arms. b) A fully extended 16-arm star polymer with arms attached to the adamantane junction; hydrophobic portions of the diblock arms are in different colors, the hydrophilic portions are all colored light brown. c) A representative “open” conformation produced after a small amount of simulation in the vacuum phase. d) A solvated structure. These figures are not drawn to the same scale; see the text for relative sizes.

As mentioned above, these systems are identical in composition except for the hydrophobic regions, and the variation was designed to span range of polar vs. nonpolar character, ester vs. alkane content, and torsional flexibility. The lengths of the hydrophobic segments in each of the three star polymers were chosen to yield approximately equal arm lengths. The molecular systems were built with arms in a fully extended state (Figure 5.1b) and their length is noted in Table 5.1. The amount of this length due to the PEO part of each arm is approximately 21 Å.

Table 5.1: Summary of star polymer systems studied. Core volume is used to derive scaling factors that allow comparison among systems (see text).

Name	System	Approx. Extended Arm Length, Å	Number Atoms	Number Water	Core Volume, $Å^3$
PLA	$A[LA_{16}\text{-}CH_2\text{-}EO_6\text{-}H]_{16}$	67	3050	45377	5860
PVL	$A[VL_8\text{- }EO_6\text{-}H]_{16}$	78	2618	45702	4650
PE	$A[E_{12}\text{-}EO_6\text{-}H]_{16}$	50	1850	46115	2432

5.2.2 Force Field

The force fields used for these simulations were mainly the OPLS-AA (all atom) force field [68] but with a number of exceptions. For adamantane, parameters similar to those for cyclohexane were used, but with the improved parameters of Price et al. [199] for the alkane torsion angle energy expressions. For the linkage of the adamantane to the hydrophobic chains and for polyethylene segments, standard alkane parameters were used with the same improved torsion expressions. OPLS-AA parameters for esters (the polylactic acid and polyvalerolactone star polymers) were obtained from the same reference [199].

Although most of the OPLS-AA force field parameters for esters were readily available, polylactic acid, an alpha-polyester, required some parameters that had not been published. These relate to the torsional expressions for sites along the backbone of the PLA polymer. Parameters for the CT-C-OS-CT torsion exist [199], but not for C-OS-CT-C or OS-CT-C-OS. In OPLS-AA notation, C represents a carbonyl carbon; OS, an alkoxy oxygen in an ester; and CT, a generic alkane-like carbon. For the missing torsional parameters we substituted those for C-OS-CT-CT and CT-CT-C-OS torsions, respectively, since the middle bond in each case has the same character and multiplicity. Because of its commercial importance, recent efforts [207] have attempted to develop a better OPLS-like parameter set for polylactic acid simulations. These efforts have included extensive fitting to bulk properties of PLA, such as glass transition temperatures, and resulted in functional forms that are more complex than the ones used in this work. It is not clear whether such a potential would improve the model accuracy in the context of a star polymer in water.

Because of its importance for numerous applications ranging from use as a polymeric solvent in batteries to improving the solubility of pharmaceuticals, a considerable amount of effort [208,209,210,211,212,213,214] has been spent to develop and improve force field parameters for polyethylene oxide (PEO), also known as polyethylene glycol (PEG). These efforts usually start with quantum chemical studies of a simple commercially available dimer of ethylene oxide, 1,2-dimethoxyethane (DME), for which a great deal of reliable experimental data [215,216] exist that are useful for guiding and validating force field efforts. It has turned out to be particularly difficult for a force field model to yield the correct conformer populations for bulk liquid DME, available from the analysis of Raman spectra [215], where the gauche conformation of the central O-C-C-O torsion appears to be unusually stable relative to, say, that of butane, a phenomenon known as the "gauche effect." The work of Anderson and Wilson [209] sought to produce a DME force field that improved upon both OPLS-AA and the potential of Smith, et al. [210,211], with respect to these conformer populations. They employed higher quality quantum calculations than had been used in previous work to produce torsional energy maps for the three torsion angles involving the six heavy atoms of DME. Then, using the OPLS-AA charges, Lennard-Jones, bond and angle parameters, they refitted only the C-C-O-C and the O-C-C-O torsional parameter to best reproduce the quantum potential energy surface. This resulted in a substantial improvement in the conformer populations in bulk DME over previous implementations.

One of our concerns with the resulting Anderson-Wilson DME force field, however, was the possibility that it was insufficiently polarized to represent PEO in an aqueous environment. Earlier work [217,218] has shown that charges based on a solvent-polarized wave function can produce better solvation results. Using an approach very similar to that of Anderson and Wilson, we produced a potential for the star polymer simulations that we felt might represent the polarization of DME in water. Briefly, we generated an optimized structure for DME in the ttt conformation using an MP2 level of theory, but with a polarizable continuum [219] model (PCM) to represent an aqueous environment. The resulting charge density was used to evaluate the electrostatic potential (ESP) around the molecule, which was then fitted using point charges at the nuclear sites. After symmetrization, these charges were 0.003/0.068 for C/H of the terminal methyl groups, 0.142/0.049 for the C/H of the methylene group, and -0.447 for the oxygen. (The unit of charge is the proton). This is in contrast to the charges 0.110/0.030, 0.140/0.030 and -0.400, respectively, used in OPLS-AA and in the Anderson-Wilson potential. Then, using these charges along with the OPLS-AA parameters for the bond and Lennard-Jones expressions, the C-C-O-C and the O-C-C-O torsional expressions were fitted to best reproduce a gas phase torsional energy map of a quality (MP2) similar to that used by Anderson and Wilson.

The results are shown in Table 5.2, where one can see that the torsional fitting parameters are very sensitive to the charge model, even though in OPLS the 1-4 electrostatic interactions are scaled by a factor of 0.5. Conformational analysis using the resulting potential for liquid NpT simulations of bulk DME at 298 K and 1 atm is shown in Table 5.3, along with other published results. One can see that the new model, polarized appropriately for solvation by water, still does fairly well at obtaining the conformer populations in bulk DME, slightly overpopulating the tgt conformer, underpopulating the tgg’ conformer (like the other force fields listed), and doing slightly worse with population of the less prevalent ttt conformer.

Table 5.2: Parameters for the torsional energy expressions involving backbone atoms in DME. Values, in units of kcal/mol, are used in the expression $E(\phi)=(V_1/2)(1+\cos(\phi))+(V_2/2)(1-\cos(2\phi))+(V_3/2)(1+\cos(3\phi))+(V_4/2)(1-\cos(4\phi))$. DMEFF refers to the work of Anderson and Wilson (28); IBM refers to the current work.

Torsion	Force field	V_1	V_2	V_3	V_4
O-C-C-O	DMEFF	2.8198	-2.5606	0.8216	-0.9203
C-O-C-C	DMEFF	1.6678	-0.5653	-0.0033	-0.2931
O-C-C-O	IBM	1.6224	-2.4022	0.2672	-0.1864
C-O-C-C	IBM	0.2770	-0.0086	0.2630	0.0178

Table 5.3: Populations of conformers of DME in bulk liquid given by different force fields and by experiment. OPLS-AA and DMEFF data is from Anderson and Wilson [209]; the column labeled SJY is from Smith, Jaffe, and Yoon [212]; Raman data are from Goutev et al. [215].

Conformation	OPLS-AA	SJY	DMEFF	IBM	Raman
ttt	13.5	18	15.4	5	12
tgt	50.3	45	51	55	42
ttg	5.2	9	5.1	2	4
tgg	13	8	7.8	10	9
tgg'	14.6	17	18.4	24	33
ggg	1.4		0.4	1	
ggg'	1.4		1.2	2	
gg'g	0.1		0.3	0	
gtg	0.3		0.2	0	
gtg'	0.2		0.2	0	
Total	100	97	100	99	100

The charges actually used on the PEO segments of the star polymers were 0.142/0.049 for the methylene group and -0.48 for the oxygen, the final oxygen charge being more negative by 0.08 than used in OPLS-AA. Admittedly, the change in charge model relative to OPLS-AA is rather small. The improvement of conformer populations is quite likely due to the fitting of torsional energy parameters to high quality quantum chemical results. Further testing of this potential should be done to assess its applicability in other contexts.

The water model used was TIP4P-Ew, developed [86] as a variant of TIP4P [84], and tuned for use in the context of Ewald treatments of the long range electrostatic interactions.

Since these data were used in the parameterization of the model, the TIP4P-Ew potential accurately reproduces pure liquid water density and heat of vaporization data over a broad temperature range. However, the model also reproduces structural properties such as the radial distribution function and kinetic properties such as the self-diffusion coefficient over a broad temperature range, and these were not used in the parameterization. In addition to being developed for use in the context of Ewald models, the fitting of the TIP4P-Ew parameters to experimental heat of vaporization data took account of the difference in energy between an unpolarized gas phase water molecule and one polarized to the extent implied by the fixed charges of the model, also known as the electronic polarization cost. Consideration of polarization cost in the fitting was also done in the development of TIP4P/2005 [85], but it was not done in the development of TIP4P [84]. Treating the polarization cost has lead to substantial improvement in the accuracy of the structural, thermodynamic, and kinetic properties of both of these models [85,86].

5.2.3 Simulations

Structural models were prepared for all three star polymer molecules using locally developed software. The models are simply the atomic Cartesian coordinates for each atomic site of the molecules constructed with each polymeric arm in a fully extended state (see Figure 5.1). These molecular structures have end-to-end distances as large as 160 Å (see Table 5.1). Force field parameters were assigned, also using locally developed software, and input files were prepared for the LAMMPS simulation package [69]. Short structural optimizations were performed followed by short simulations on these molecules without solvent, i.e., in the "gas phase", during which the molecules were allowed to partially collapse into more compact but still somewhat open structures. These partially collapsed structures had end-to-end distances ranging from 48 Å (PVL) to 58 Å (PLA).

A cubic simulation cell of 46,656 TIP4P-Ew water molecules was prepared and equilibrated using locally developed software and a protocol previously described [220] with a control temperature of 300 K and an external pressure of 1 atm. A set of particle coordinates was obtained from the resulting simulation (cell edge length 111.9453 Å) representing an instantaneous density of 0.9949 g/cm^3, in excellent agreement with experimental values (at a temperature of 298 K and a pressure of 1 atm, the mean density of water using the TIP4P-Ew water model [86] is 0.9954 g/cm^3; the corresponding experimental value is 0.99716 g/cm^3).

The starting conformation for the simulations of solvated star polymers was made as follows: (1) The coordinates of the star polymer sites were translated so that the center of geometry was at the center of the cubic simulation cell from the water equilibration simulation. (2) For each water molecule, the smallest distance from any of its three sites to any of the sites on the star polymer was computed. (3) Using these distances, water molecules were removed from the simulation, starting with the one closest to the star polymer and then the one next closest and so on, until the total mass of removed water molecules first exceeded the total mass of the star polymer. Between 541 (PE system) and 1279 (PLA system) water molecules were removed by this procedure. This process produced a set of three systems, one for each star polymer. By construction, each of these systems had the same volume and very nearly the same mass, and hence the same mass density, as that of water. With the closest water molecules removed in this way, dynamical simulations could be started without any additional preparation (see Figure 5.1).

Most of the production simulations were performed using a version of the LAMMPS software [69] dated July 7, 2009. The LAMMPS software performs and scales well on massively parallel computers for molecular systems of the size and type studied here. Molecular dynamics simulations were performed on an IBM BlueGene/L supercomputer and most of the analysis of the resulting trajectory data was performed on a cluster of IBM AIX workstations.

All equilibration and production simulations were performed using the NVT ensemble, with thermal control implemented in LAMMPS using a Nosé-Hoover extended Lagrangian procedure, with a fictitious mass set so as to establish a fluctuation period [221] of approximately 100 fs in the thermostat variable, known as the thermostat damping factor in LAMMPS. The dynamical integration scheme was velocity-Verlet [100] with a time step size of 1 fs. All bond lengths involving hydrogen, as well as the H-O-H angle for the TIP4P-Ew water, were constrained using a SHAKE procedure [103] to guarantee that bond length constraints were satisfied to a tolerance of 10^{-5} Å. Lennard-Jones interactions and direct space electrostatic interactions were truncated at 9 Å. A tail correction for the part of the Lennard-Jones potential beyond this cutoff was included in the energy and pressure computation. Electrostatic interactions were evaluated with a particle-particle-particle mesh (PPPM) procedure [99] with an accuracy parameter (10^{-5}) that resulted in a 3D grid of 120-by-120-by-120. In accordance with the OPLS potential, neither Coulomb nor Lennard-Jones interactions are evaluated for particle pairs that are 1-2 and 1-3 interactions, and both of these types of interactions are scaled by a factor of 0.5 for 1-4 interactions. Geometric combining rules were used to establish the Lennard-Jones parameters.

Thermal equilibration was performed for each of these systems at four different temperatures: 300 K, 350 K, 400 K and 450 K. Some of these equilibration simulations were as long as 50 ns. During the thermal equilibration phase for each solvated star polymer system, the star polymer molecules collapsed slightly more, and for each system, the edge length of the simulation cell was greater than twice the linear dimension of the star polymer molecule. Under the minimum image convention, this guarantees that the closest image of each site in the star polymer to any other site is from the same image, thereby preventing the apparent direct interaction between copies of the star polymer molecule in different periodic images. Production runs for these three solvated molecular systems at the four temperatures were at least 20 ns.

5.2.4 Analysis

During the production phase of the simulations, solute and solvent coordinates were saved to disk for analysis at intervals of 10 ps, resulting in at least 2,000 sets of coordinates for each of the three molecules and at each of the four temperatures. Since in most cases these represented highly correlated data, detailed analysis was performed only on one-fourth of this data, on coordinates spaced at intervals of 40 ps. The analysis consisted of the calculation of several structural observables as well as a Voronoi analysis. The structural observables included such things as maximum end-to-end distance, spherically averaged mass density, and molecular shape descriptors [222] derived from the eigenvalues of the gyration tensor such as radius of gyration, asphericity and anisotropy. Notation and formulae related to geometric shape descriptors do not appear to be standardized. We follow Theodorou and Suter [222] where the elements of the gyration tensor are given by $S_{\alpha,\beta} = (1/N)\sum_i r_{i,\alpha} r_{i,\beta}$ where the sum is over the *N* sites $\mathbf{r}_i$ and α and β refer to Cartesian components (*x, y or z*); $\mathbf{r}_i$ is measured relative to the center of geometry where $\sum_i r_{i,\alpha} = 0$. The eigenvalues of *S* are denoted λ_x, λ_y, and λ_z. The radius of gyration is defined as $(R_g)^2 = \lambda_x + \lambda_y + \lambda_z$ and when ordered, $\lambda_x < \lambda_y < \lambda_z$. Asphericity is defined as $b = \lambda_z - 0.5(\lambda_x + \lambda_y)$. Acylindricity is defined as $c = \lambda_y - \lambda_x$. Shape anisotropy is $\kappa^2 = (b^2 + 0.75c^2)/(R_g)^4$. We report κ^2 and call it anisotropy. Huynh, et al. [223] report this as well but call it asphericity.

In order to characterize intramolecular chain structure and dynamics without the effect of overall star polymer molecular rotation, a molecule-centered reference frame was defined with an origin and orientation determined by the coordinates of the sites of the relatively rigid adamantane core. For each coordinate set, the atomic site coordinates as well as the coordinates of an orientational unit vector associated with each monomeric unit of the star polymer were measured with respect to the molecule-centered reference frame.

Voronoi analyses were also performed for each set of coordinates. A Voronoi analysis [224,225] constructs a set of polyhedra, one polyhedron around each atomic site of the system (all water sites plus all star polymer sites). These polyhedra collectively fill all space of the simulation cell, and each one encloses a volume of space that is closer to its associated site than to any other site. Voronoi polyhedra faces that are shared by two polyhedra consist of points that are equidistant to the two sites associated with the polyhedra. Similarly, points on a polyhedron edge are equidistant to three such sites, and each polyhedron vertex is equidistant to four. Voronoi analysis uses only the Cartesian coordinates of the atomic sites in the simulation, without any knowledge of the molecular identity, chemical nature, or bond connectivity of the material. However, all of this information manifests itself in the Voronoi analysis through the resulting distribution in the number of faces, the facial shapes and areas, polyhedral volumes, etc. For our purposes, we partitioned the sites into different classes and used the Voronoi polyhedra associated with each class to compute a volume for the class, and the interfacial surface area shared between pairs of classes. The interface between two classes consists of the union of all

Voronoi faces that are shared by two polyhedra where one polyhedron is associated with one class and the other polyhedron is associated with the other class. For example, by establishing three classes that we associate with water, hydrophobic, and hydrophilic sites, we can compute the total volume occupied by each class and the interfacial surface area between the hydrophobic and hydrophilic material, as well as between the water and the hydrophilic material. This technique provides a somewhat more general alternative to the solvent accessible surface area metric [226] that is often used.

An important issue for the use of star polymeric materials for drug delivery relates to the amount of water in the interior of the polymer, specifically the hydrophobic region. Water content potentially affects the release rate of the drug as well as the rate and mode of degradation of the polymer itself. Important metrics therefore include the thermal and compositional dependence of the amount of water, its penetration depth, its diffusion within the star polymer, and the exchange rate of water into and out of (e.g., water lifetime) the hydrophobic regions. The difficulties of measuring this in simulations begins with how one should define what molecules are actually in the hydrophobic region, since there is really not a well-defined interface between components when examined at a molecular level.

The issue of water penetration was explored also using Voronoi analysis, since it provides a useful way of determining if two molecules are neighbors: two sites are Voronoi neighbors if their Voronoi polyhedra share a face. A cluster analysis was performed on the water molecule sites in each coordinate set using the neighbor list developed by the Voronoi analysis. Two water sites were defined as being in the same water cluster if they were Voronoi neighbors. This clustering procedure partitions the water molecules into one or more sets, each of which is a contiguous water cluster, with the largest, of course, representing the bulk solvent. Other clusters of water molecules, if they exist, are, by construction, not neighbors of any of the bulk water molecules, and, so, are surrounded only by star polymer sites. We designate such clusters as interior clusters, as distinguished from the bulk cluster. With this approach, information can be collected about the number, size and shape distribution of such interior water clusters, the nature of their environment (e.g., the amount of the water cluster's surface area that is in contact with hydrophobic vs. hydrophilic regions of the star polymer), and their evolution and lifetime within the star polymer. By computing the closest distance from each oxygen site in an interior water molecule to the oxygen sites in the bulk water, one can get a sense of the penetration depth into the star polymer of the interior absorbed water.

In general and at most sampling rates, observables computed from molecular dynamics simulation data are highly correlated. In several instances where statistical uncertainty is reported, correlation in the data was taken into account by computing the fluctuation autocorrelation function for the observable of interest. The correlation time was taken as the area under this function. Since this function itself is subject to uncertainty and becomes more noisy with greater lag times, it is integrated only up to the point where it first goes negative, giving an estimate of the correlation time, τ_c. The uncertainty (standard deviation of the mean) in an observable is then given as the root-mean-square deviation in that observable times the square root of $2\tau_c/T$, where T is the simulation time over which the observable is averaged and $T/2\tau_c$ is the effective number of uncorrelated samples. There is always some danger in computing such correlation times and resulting uncertainty estimates, since for insufficiently long simulations the true magnitude and temporal variation of the fluctuation of a signal is not observed and, hence, both the standard deviation and the correlation time are underestimated, resulting in an artificially small uncertainty which makes the result appear to be more statistically significant than it really is. Uncertainties in the numbers of rare events observed, such as instances of water penetrating into the hydrophobic regions of the star polymers, were assumed to be the square root of the number of observations, assuming these are governed by Poisson statistics.

5.3 Results

Renderings of representative configurations of the three molecular systems at the lowest (300 K) and highest (450 K) temperatures studied are shown in Figures 5.2-5.4. For the PLA and PVL molecules, the images do not convey much difference between the high and low temperature, nor even between the PLA and PVL star polymers. The PE system, on the other hand, shows a considerable amount of ordered structure, even at high temperature. Animated models (videos) of these molecules suggest that the PEO regions are very mobile, but the hydrophobic regions are very rigid (crystalline or glassy) on the 20 ns time scale of these simulations.

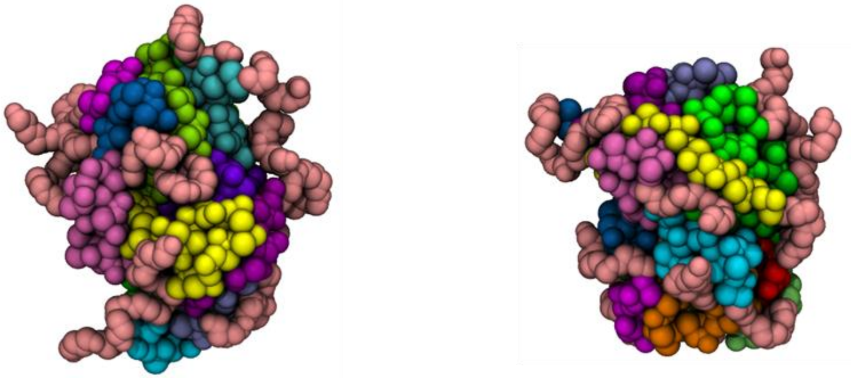

Figure 5.2: Renderings of representative configurations of the PLA star polymer at 300 K (left) and at 450 K (right). The hydrophobic region of each of the 16 arms is shown in a different color. The hydrophilic (PEO) terminal region of each arm is colored light brown.

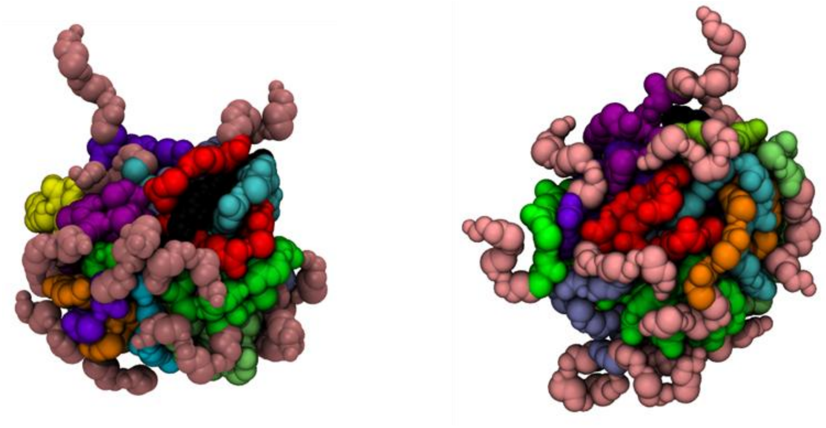

Figure 5.3: Renderings of representative configurations of the PVL star polymer at 300 K (left) and at 450 K (right), colored as in Figure 5.1. The hydrophilic (PEO) terminal region of each arm is colored light brown.

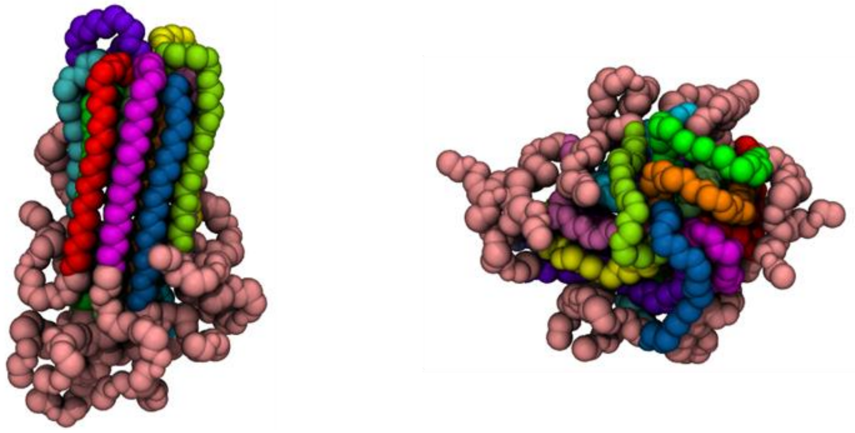

Figure 5.4: Renderings of representative configurations of the PE star polymer at 300 K (left) and at 450 K (right), colored as in Figure 5.2.

Radius of gyration (R_g) data are shown in Figure 5.5. For each molecule and at each temperature, the average radius of gyration was computed using the entire star polymer (filled symbols and solid lines) as well as without consideration of the hydrophilic part, in order to characterize the spatial extent of the hydrophobic core. From the figure, one can see that there is very little temperature dependence except for the PE system, which shows a significant drop in R_g between 350 K and 400 K, suggesting a transition to a more compact state. Figure 5.6 shows the thermal dependence of the magnitude of the fluctuations in R_g. The general increase with temperature indicates a gradual increase in the compressibility. It is notable that even though there is a discontinuous change in R_g for the PE star, there is no such behavior in the magnitude of the fluctuation in this quantity. Figure 5.7 shows the thermal dependence of the anisotropy. This dimensionless metric can range from zero (spherical) to unity (long rods). The change in anisotropy for the PE star in going from 350 K to 400 K indicates a change from an elongated to a more spherical shape.

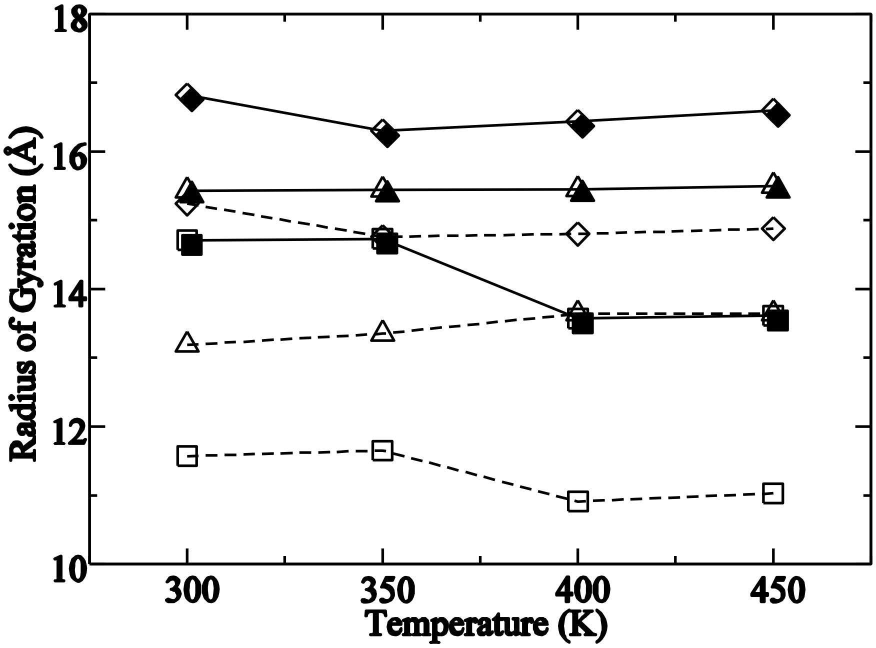

Figure 5.5: Radius of gyration, computed from the gyration tensor, for each star polymer at each of four temperatures. Symbols represent PLA (diamonds), PVL (triangles), and PE (squares). Solid symbols and solid lines represent the radius of gyration for the entire star polymer; open symbols and dashed lines represent that of just the hydrophobic material. Uncertainty estimates ± two standard deviations are not shown but are approximately the size of the symbols, usually 0.1Å or less.

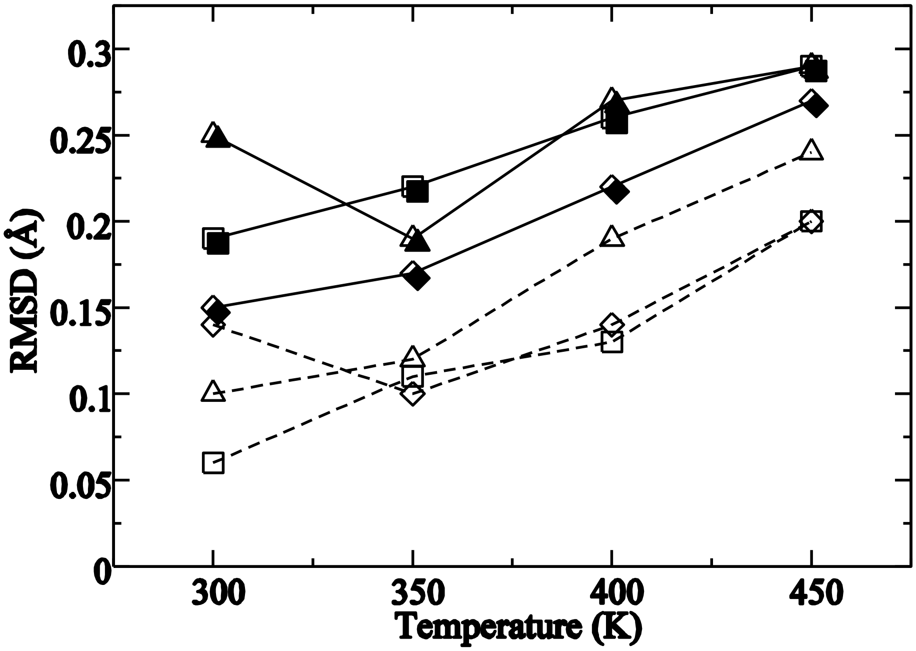

Figure 5.6: Root mean square deviation in R_g. Symbol and line type notation is the same as in Figure 5.5.

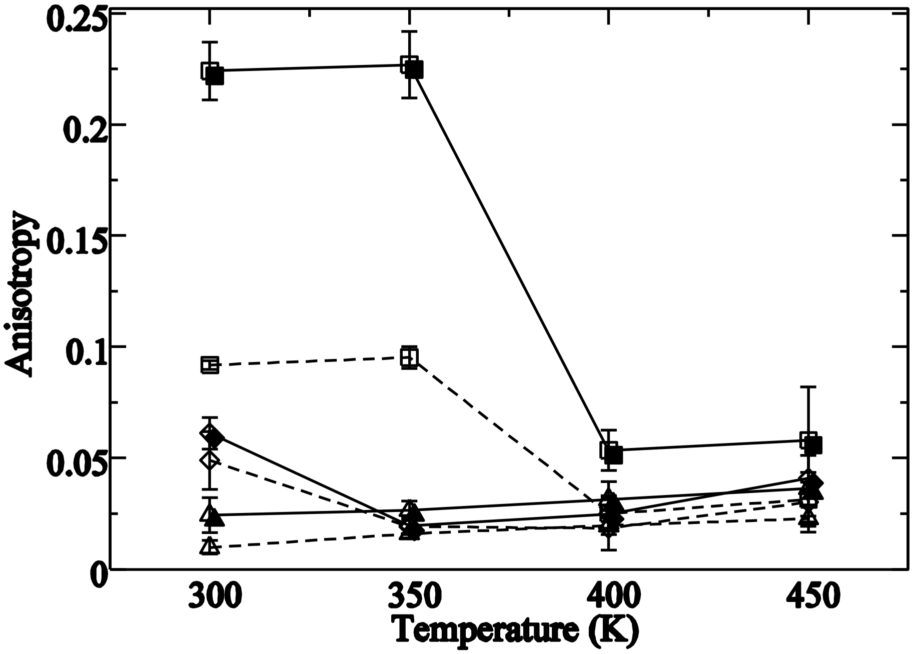

Figure 5.7: Anisotropy. Uncertainty estimates are ±2 standard deviations.

The behavior of some of these star polymers at the lowest temperature shows some anomaly in R_g (PLA), its fluctuation (PVL and the hydrophobic core of PLA), and the anisotropy (PLA). This may be an indication that at 300 K the relaxation times and correlation times may be so long for PLA and PVL that thorough sampling is difficult to achieve over the 20 ns period of the production simulations. Other observables suggest this as well. The uncertainty estimates are larger at these temperatures, but might still be underestimated. The question thus arises as to whether these simulations were sufficiently long for adequate sampling, particularly given that those in the preceding study of Huynh et al. were at least 200 ns per star polymer system at 300 K. Although we experienced difficulty at that temperature with sampling some of the observables for the PLA and PVL star polymers over the course of the 20 ns production simulation, Huynh et al. noted that most of their observables were stable after about 15 ns of sampling. We feel that at our higher temperatures (350 K, 400 K and 450 K), we are able to sample adequately due to the shorter correlation times at these temperatures.

The spherically averaged mass density at 350 K, as a function of distance from the center of mass of the adamantane, is shown for the three molecules in Figures 5.8-5.10. These figures illustrate the contribution to the total mass density from various components: adamantane, the hydrophobic material, the hydrophilic material, and water. These figures indicate that close to the adamantane the material is highly structured. In fact, having 16 polymeric arms attached to an adamantane core produces a significant amount of local strain in the model that persists along each arm until the density decreases enough for more favorable and random chain conformations to be adopted. These graphs show very little thermal dependence (data not shown for other temperatures), except for a gradual smearing of some of the features of the hydrophobic material. One can see that the hydrophilic material adopts a broad featureless distribution in each case. The figures also appear to suggest that there are rather diffuse boundaries between materials, with considerable interpenetration of water and hydrophilic material into the hydrophobic material. However, further analysis indicates this is an illusion created by the spherical averaging process when the hydrophobic core is highly nonspherical (PE) or has a rough surface with many grooves and valleys (PLA and PVL) that allow water to get relatively close to the adamantane core without actually penetrating into the hydrophobic core material itself.

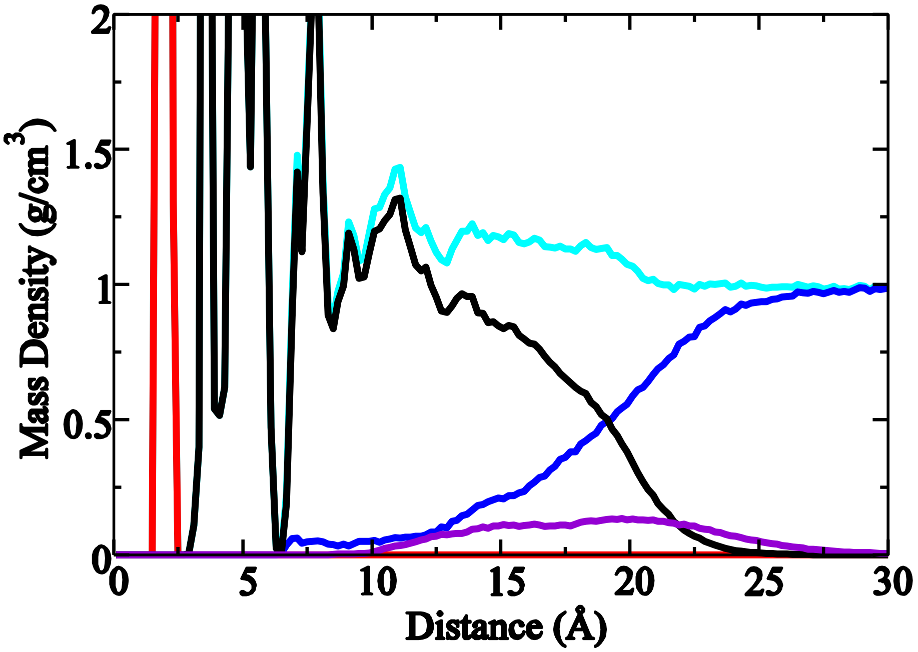

Figure 5.8: Orientationally averaged mass density for the PLA star polymer at 350 K as a function of distance from the center of mass of the adamantane. The curves represent contributions to the total mass density (cyan) from the adamantane (red), from the hydrophobic material (black), from the hydrophilic PEO (purple), and from water (blue).

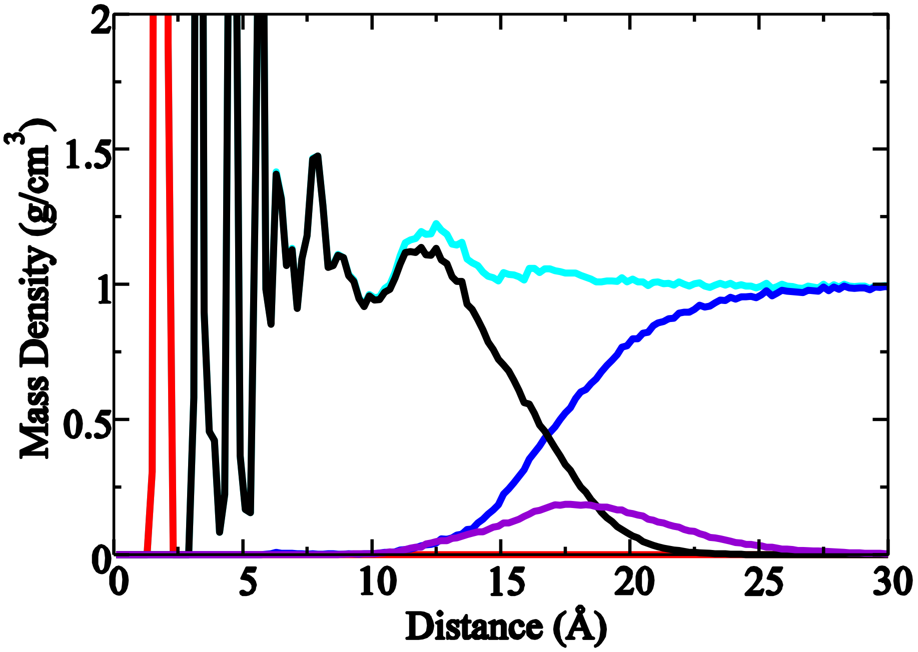

Figure 5.9: Orientationally averaged mass density for the PVL star polymer at 350 K. Coloring scheme is the same as for Figure 5.8.

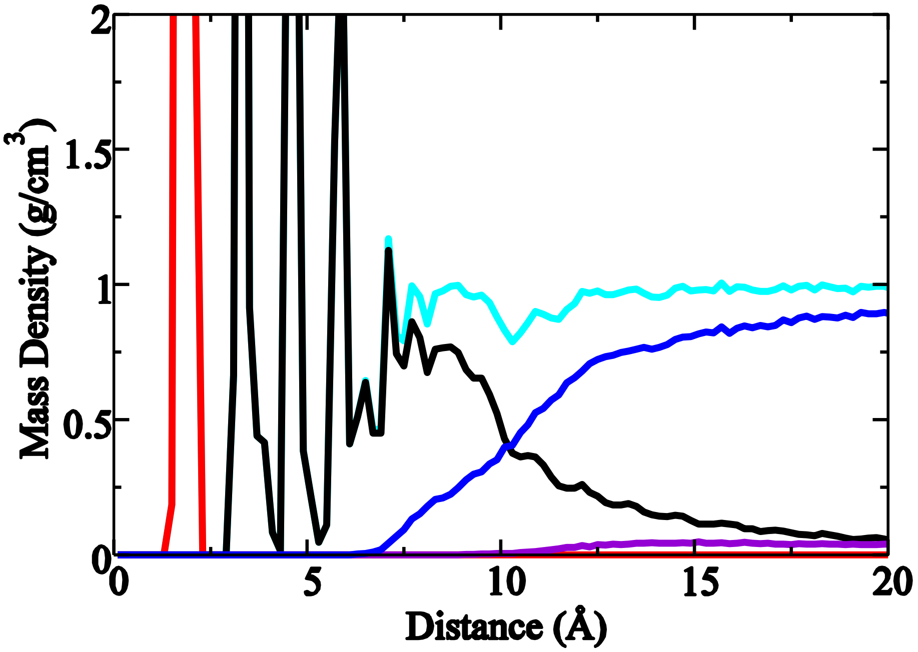

Figure 5.10: Orientationally averaged mass density for the PE star polymer at 350 K. Coloring scheme is the same as for Figure 5.8. In comparing this figure with Figures 5.8 and 5.9, one should note that the range of the x-axis is different.

The mass density plot for the PE polymer (Figure 5.10) exhibits a feature not seen in the other two polymers: there is a much longer and more slowly decaying curve for the hydrophobic material. Whereas the mass density due to hydrophobic material falls from 0.5 g/cm^3 to 0.1 g/cm^3 over a distance of 2.9 Å for the PLA and 3.2 Å for the PVL, this decrease occurs over a distance of 6.4 Å for the PE polymer. This is a manifestation of the crystalline and cylindrical nature of this polymer at 350 K, as indicated visually (Figure 5.4) and through other structural observables.

The structure in the mass density graphs can be resolved by monomeric unit, and this is shown for PLA at 350 K in Figure 5.11. The first monomeric units of LA along each of the 16 arms produce three peaks in this graph (shown in black) near a distance of about 5 Å; the second units of LA along the 16 arms produce peaks near 7 Å (shown in red), and so on. As one moves out along the chains, the distributions become less structured and broader. Using this graph, one can determine which monomeric units are responsible for the peaks in the total mass density. For example, the peak at 10.7 Å in black in Figure 5.8 (PLA) is due to the positioning of the third LA units along the arms (shown in blue in Figure 5.11). The peak at 11.7 Å is due to the fourth (shown in purple) and fifth (shown in magenta), and the soft peak at 13.3 Å is due to the fifth (shown in magenta) and sixth. LA units farther than the fourth or fifth, and out to the last (16th), are behaving similarly and probably produce a density and chain packing at distances larger than about 12.5 Å that is more representative of what might be seen in even larger star polymers. In fact, even the fourth unit (shown in purple) contributes significant mass density at distances closer to the adamantane than the third (shown in blue), suggesting an ability to pack more closely. The fourth and fifth units along the chains are, therefore, transitional between the highly structured and more random units. Similar analysis identifies the transitional hydrophobic units for the other star polymers as the second for PVL and the third and fourth for PE (the PVL units are much larger and floppier than the PLA and PE units).

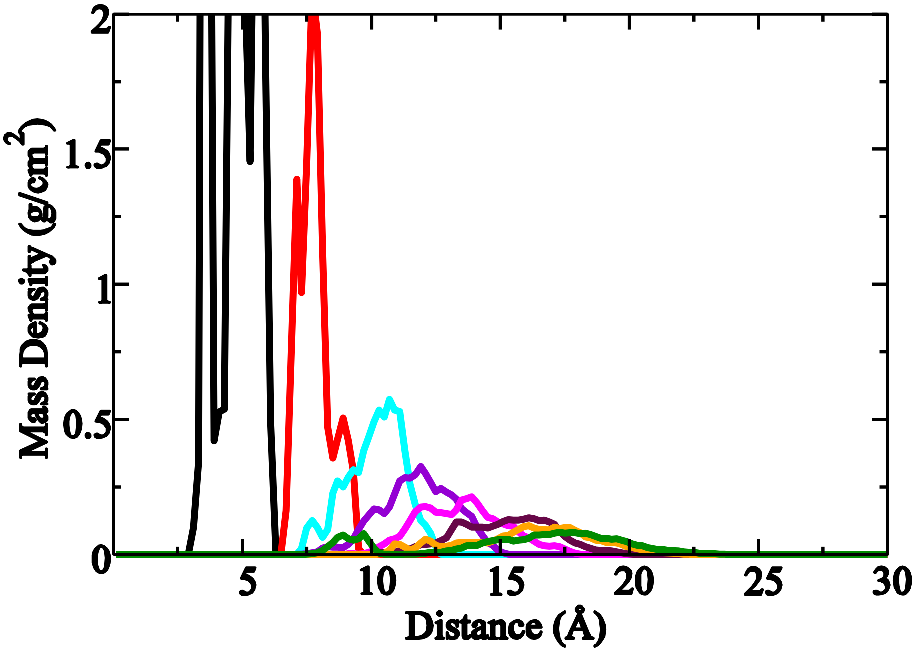

Figure 5.11: Orientationally averaged mass density contribution from hydrophobic material in the PLA star polymer at 350 K resolved by contribution from different monomeric units. Each color corresponds to the mass density contributed by a different set of 16 lactic acid monomer units that are all at the same position along the arm as measured from the adamantane connection. The black curve represents mass density from the 16 lactic acid units that are directly connected to the adamantane.

Orientational autocorrelation functions for individual repeat units for each star polymer at 350 K are shown in Figure 5.12. These were computed as follows. First, a local orientational unit vector (**u**) was defined for each monomeric unit on each arm of each star polymer. These vectors were directed between specific pairs of atomic sites on each monomeric unit. For the PLA and PVL, the vectors were directed between an alkoxy oxygen site and the first carbon site immediately opposite the nearest carbonyl group. For PE and PEO, the vectors were directed between pairs of adjacent (bonded) carbon sites. Only three vectors were selected within the six-unit PEO part of the chain. Second, for each saved set of coordinates, these vectors were measured and projected from the lab frame onto the molecule-centered reference frame. Third, the time evolution of these vectors in the molecule frame was determined at 40 ps resolution over the 20 ns of the production simulations, and then an autocorrelation function (<**u**(0)•**u**(t)>) was computed for each monomeric unit. Next, groups of 16 of these functions that correspond to monomeric units at the same position along the star polymer arms were averaged. There is, therefore, a set of curves (19 for PLA, 11 for PVL, and 15 for PE) for each star polymer that shows different rates of decay, corresponding to the rate of loss of orientational memory for monomer units at various positions along each arm. Curves that correspond to monomeric units close to the adamantane do not decay at all, and the curves that correspond to the most distant PEO group decay very rapidly. (In Figure 5.12, approximately half of these curves are shown).

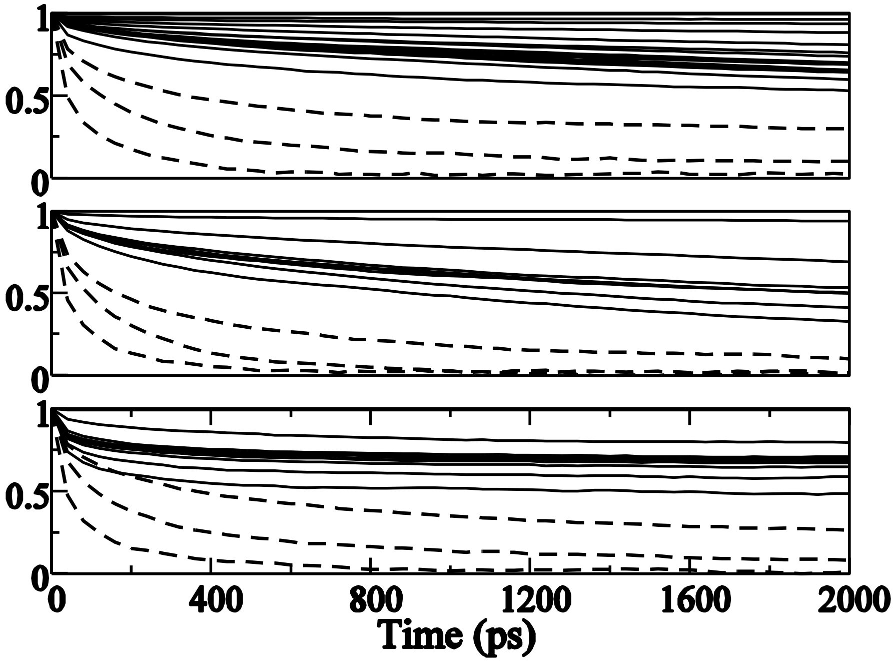

Figure 5.12: Orientational autocorrelation functions for monomeric units at 350 K for PLA (top), PVL (middle), and PE (bottom) star polymers. Each curve represents an average of the 16 autocorrelation functions that correspond to monomeric units at the same distance along each arm. Dot-dashed lines correspond to the transitional repeat units near the adamantane within the hydrophobic material. Solid lines correspond to various other hydrophobic monomeric units. Dashed lines correspond to hydrophilic PEO units. Correlation functions decay more and more quickly as one moves farther out along each chain away from the adamantane. For PLA (top) the lines refer to repeat units 2, 4*, 5*, 6, 8, 10, 12, 14, 16. For PVL (middle) the lines refer to repeat units 1, 2*, 4, 6, 8. For PE (bottom) the lines refer to repeat units 2, 3*, 4*, 6, 8, 10, 12. Asterisks indicate the repeat units to be transitional.

The curves in Figure 5.12 illustrate several interesting points. First, the rate of decay in the correlation functions is nearly monotonic as one moves along each chain from the units closest to the adamantane outward. Second, for each molecule there is a group of very slowly decaying curves that includes the transitional units (identified above, and shown in Figure 5.12 with dot-dashed lines), and then a band of decaying curves that are closely spaced and sometimes overlapping that correspond to a more homogeneous temporal behavior of hydrophobic material outside of this transitional region. Third, there are one or two units in the hydrophobic region nearest the PEO that have much faster decay than these, presumably because they are being "pulled" about by the more rapidly reorienting and solvated PEO units. Finally, the PEO units (dashed lines in Figure 5.12) have the fastest decay of orientational correlation, with the slowest one of those corresponding to the ones "tethered" to the hydrophobic units. These time correlation functions were computed at each of the four temperatures (not shown) and all graphs share these general features, but with correlation times becoming shorter with increasing temperature.

The curves in Figure 5.12 also indicate that at 350 K the hydrophobic material is not reorienting on a 2 ns time scale for the PLA and PE, and the PVL core is behaving only slightly more fluid-like, suggesting that these cores are more like a solid (crystalline for PE, disordered and glassy for PLA and PVL) than a liquid. In fact, the middle band of PE correlation times (bottom of Figure 5.12) shows no decay at all, and the structure appears to be rigid and crystalline (Figure 5.4). Even though they do not decay much on a 2 ns timescale, the hydrophobic PVL correlation functions show faster decay than those of either PLA or PE, indicating that PVL could be somewhat more fluid-like than either the PLA or the PE polymers. This may be due to the alkane regions of the hydrophobic segments offering more flexibility than what is available in the PLA, and the ester regions preventing the chain registration and alignment of the alkane segments that is seen to stabilize the crystalline structures of the PE.

The Voronoi analysis was performed on coordinate sets at 40 ps temporal resolution, and interfacial surface areas between water, hydrophobic, and hydrophilic materials were measured and averaged for each star polymer at each temperature. Figure 5.13 shows the interfacial area exposed by the hydrophobic material, which is the sum of the hydrophobic-water and hydrophobic-hydrophilic interfacial areas. Except for the PLA star at 300 K, which might be exhibiting insufficient sampling as discussed above, these show a gradual increase with temperature probably related to the expansion seen in the fluctuations of the radius of gyration. Figure 5.14 shows the part of this that is due to the interface between the hydrophobic material and water. These generally show an increase with temperature, except for PLA at 300 K, and a noticeable drop between 350 K and 400 K for the PE star, where it changed from an elongated structure to a more globular one.

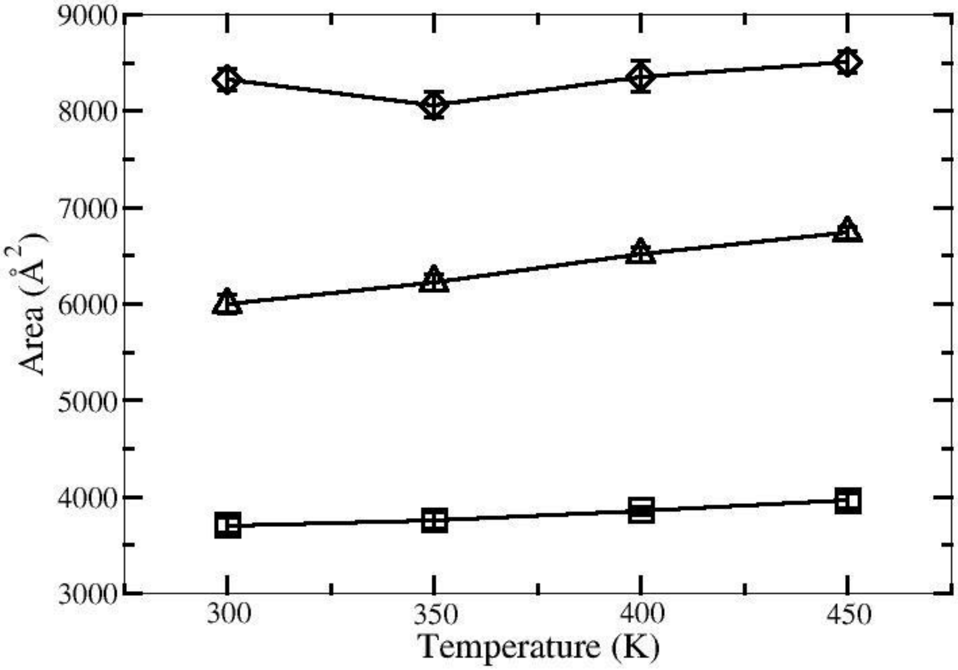

Figure 5.13: Total interfacial area of hydrophobic material (sum of hydrophobic-hydrophilic and hydrophobic-water interfacial areas) for the PLA (diamond), PVL (triangle), and PE (square) star polymers. Uncertainty estimates are ±2 standard deviations.

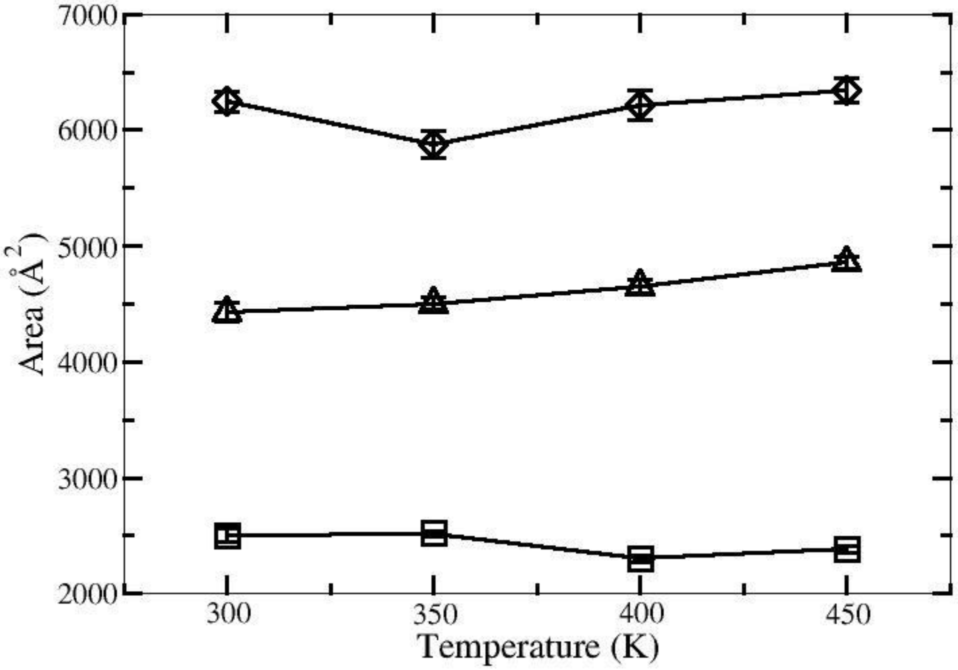

Figure 5.14: Interfacial area between hydrophobic material and water for the PLA (diamond), PVL (triangle), and PE (square) star polymers. Uncertainty estimates are ±2 standard deviations.

It would make sense for some aspect of the PE star hydrophobic surface area to change during this structural change. A decrease in the hydrophobic contact area with water (shown in Figure 5.14) makes sense, but the fact that the total surface area (Figure .13) did not change very much indicates that that the hydrophilic-hydrophobic interfacial area increased to compensate. Images of the PE star (Figure 5.4) suggest why this happened. The elongated cylindrical PE structures seen at low temperature are very crystalline, organized into a cylindrical shape with all of the PEO at one end. At higher temperatures where this structure gives way to a more globular shape, there is more opportunity for the hydrophilic PEO parts of the chains to come into contact with the hydrophobic parts, thereby simultaneously decreasing the water contact (Figure 5.14) and increasing the PEO contact (Figure 5.16) after the collapse. The trend in hydrophobic surface areas seen in Figures 5.13 and 5.14, PLA>PVL>PE at all temperatures, is simply a reflection of the fact that these star polymers are of somewhat different sizes, simply because of the number and size of the repeat units in their hydrophobic regions (see Table 5.1).

Figure 5.15 shows the total PEO interfacial area, which is the sum of the contact area with water and with the hydrophobic material. The increase in PEO total surface area for the PE star in going from 350 K to 400 K can be understood as the dense PEO structure at the end of the cylindrical structures seen at low temperature (Figure 5.4) is broken up at higher temperatures, allowing more PEO to be exposed to water and to the hydrophobic PE in these more globular conformations. However, other features of Figure 5.15 are rather surprising in that there is a general and significant decrease in PEO surface area with increasing temperature for each polymer, even for the PE on each side of the transition. This decrease suggests that PEO may be aggregating with itself, decreasing contact with water and perhaps with hydrophobic material, and that this is somehow more pronounced at higher temperatures.

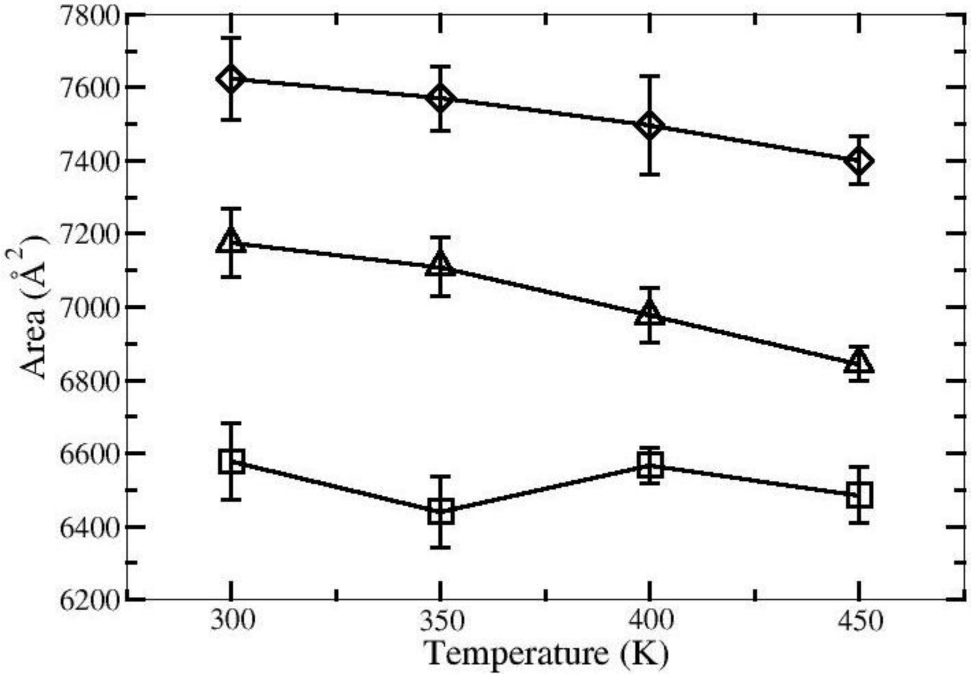

Figure 5.15: Total interfacial area of hydrophilic material (sum of hydrophobic-hydrophilic and hydrophilic-water interfacial areas) for the PLA (diamond), PVL (triangle) and PE (square) star polymers. Uncertainty estimates are ±2 standard deviations.

Figures 5.16 and 5.17 show the PEO-hydrophobic and PEO-water interfacial areas, respectively. Contact between PEO and hydrophobic material actually increases or is relatively flat with temperature, but the contact with water decreases by even more, indicating a preference for PEO to attempt to phase separate from water at higher temperatures. This rather surprising observation may be consistent with the fact that PEO-water mixtures exhibit a lower critical solution temperature (LCST) [227] wherein a PEO-water mixture can change from a single phase (miscibility) to a two-phase system with increasing temperature. The possibility of this phenomenon being exhibited in star polymers, where it might be tunable by modifications of chain length, number of chains or the chemical nature and size of the hydrophobic region, is worthy of further experimental and theoretical investigation. Figures 5.16 and 5.17 indicate that this tendency of PEO to self-associate is in competition with its tendency to associate with the hydrophobic material.

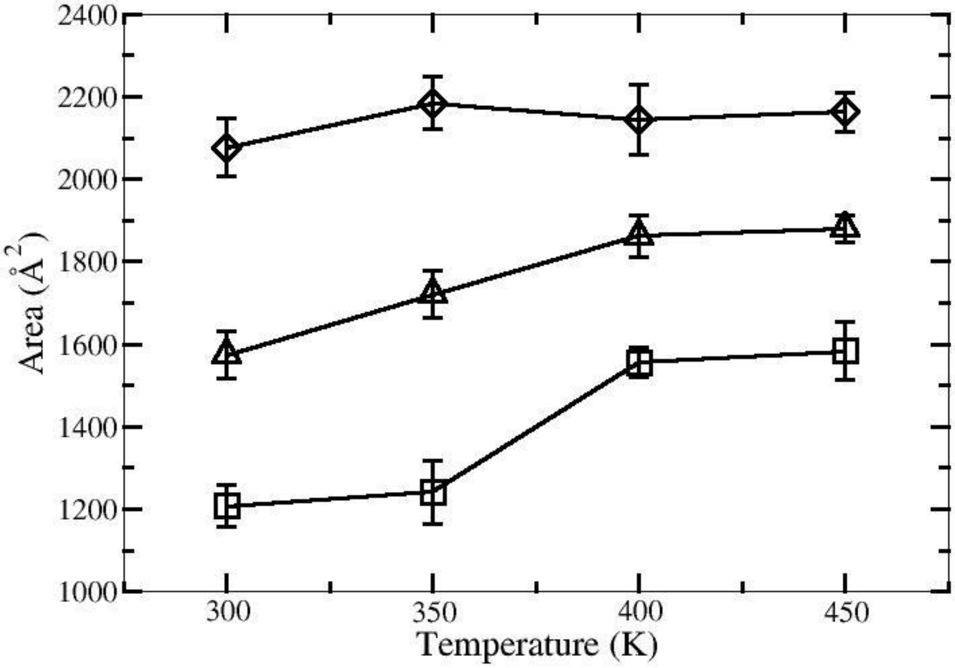

Figure 5.16: Interfacial area between hydrophilic (PEO) and hydrophobic material for the PLA (diamond), PVL (triangle), and PE (square) star polymers. Uncertainty estimates are ±2 standard deviations.

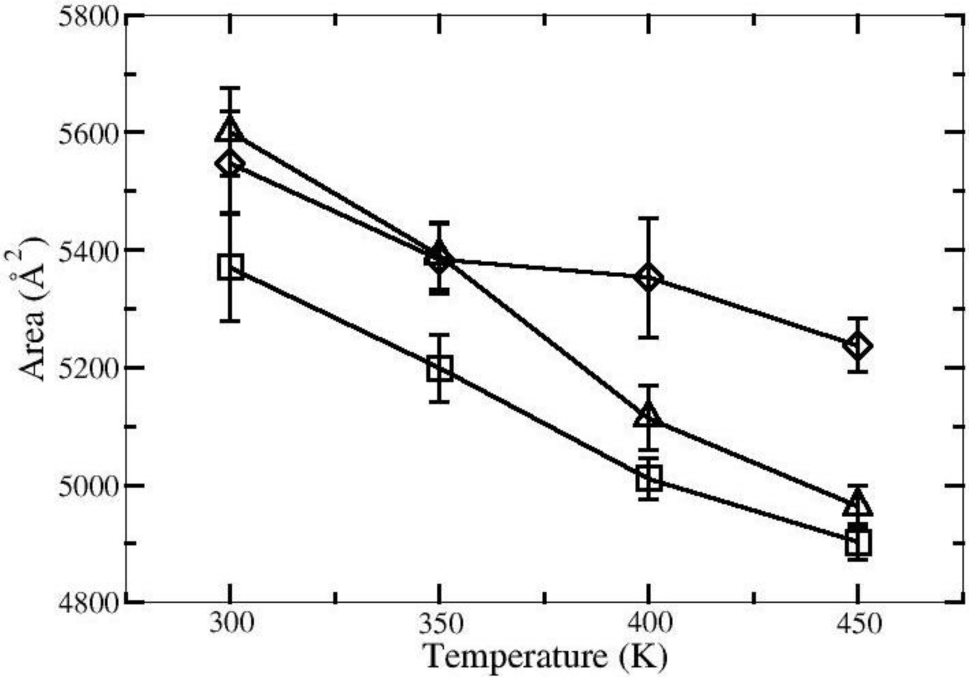

Figure 5.17: Interfacial area between hydrophilic (PEO) material and water for the PLA (diamond), PVL (triangle), and PE (square) star polymers. Uncertainty estimates are ±2 standard deviations.

Also apparent from Figure 5.16 is the decrease at all temperatures in the tendency of PEO to cover the hydrophobic core as one goes from an ester-rich hydrophobic material (PLA) to a pure alkane hydrophobic material (PE), with PVL in between. Recall that the length of the PEO region of each arm is the same across all three polymers, so differences between PLA, PVL, and PE star polymers in the PEO interfacial areas with water and with hydrophobic material seen in Figures 5.15-5.17 are not due simply to differences in the sizes of the hydrophobic regions as is the case in Figures 5.13 and .14. Finally, regarding the numerical values of the PEO-hydrophobic interfacial surface areas in Figure 5.16, it should be recognized that these include the areas from the region of connection where the PEO segment of each arm connects to the hydrophobic segment of that arm. The contribution from these junctions varies considerably with chain orientation, but 35 Å^2 is a reasonable approximation, suggesting that there is about 560 (16*35) Å^2 of the interfacial PEO-hydrophobic area that should not be considered to be in contact in the usual sense of a PEO chain laying against the hydrophobic material. With this number subtracted from each of the values in Figure 5.16, and the result compared with the values in Figure 5.17, one sees that there is about 2.5 times more contact between PEO and water than between PEO and the hydrophobic region of the PLA star. For the PVL star polymer this factor is about 3, and for the PE star it goes from over 4 at low temperatures to about 3 at high temperatures.

Apparent from comparison of Figures 5.13 and 5.14 is that hydrophobic cores in these models are exposed to a great deal of water. From 60% to 75% of the hydrophobic surface area is in contact with water, and this increases to 70% to 80% if account is made for the approximately 560 Å^2 of junction regions that cannot be solvated. This may be due to the fact that the PEO regions of the chains are relatively short, with only six PEO units. However, from comparison of Figures 5.15-5.17, as discussed above, one can see that given the choice between contact with water or with hydrophobic material, the PEO in our model has overwhelming preference for water contact. A study of individual conformations reveals that the maximum contact area between one of the PEO segments and the hydrophobic region is approximately 234Å^2, including the junction region. Therefore, a hypothetical fully coated hydrophobic region would produce about 3744 (16*234) Å^2 of PEO-hydrophobic interfacial area with PEO segments of the length studied here. Figure 5.16 shows significantly less than this amount at all temperatures and for all molecules. Similarly, the maximum interfacial area between an extended PEO segment and water is approximately 450 Å^2. So, if all PEO segments were maximally solvated, the aggregate PEO-water interfacial area would be approximately 7200 Å^2. Figure 5.17 shows that for each molecule and at all temperatures, the PEO-water interfacial area is a significant fraction of this "available" 7200 Å^2. Some of the deficit can be explained by PEO-PEO contact, including interchain contacts and intrachain contacts observed when a PEO segment adopts hairpin and coil conformations.

Voronoi analysis allowed for the identification of water clusters that permeated into the interior of the star polymers. The average number of interior water molecules per configuration for each star and at each temperature is shown in Figure 5.18. Water penetration generally increases with temperature, but for the PLA showed a very large increase (almost six-fold) in going from 350 K to 400 K. PLA is the most ester-rich of the star polymers studied, and this might explain its increased tendency to absorb water. However, even for this polymer and at the highest temperatures, there are only about 2.5 water molecules in a typical star polymer, so water penetration is rather rare.

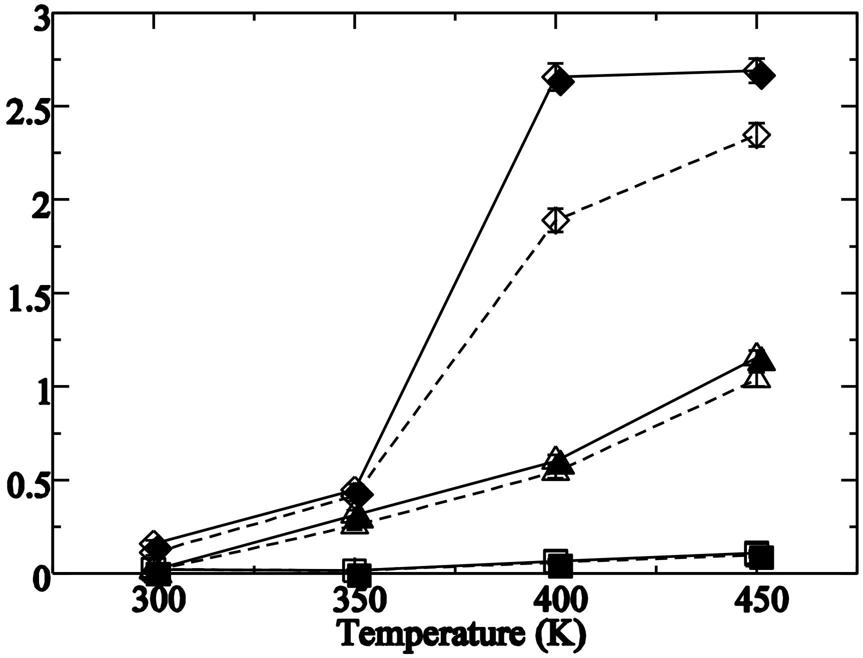

Figure 5.18: Average number of interior water molecules and water molecule clusters per configuration in the PLA (diamond), PVL (triangle), and PE (square) star polymers. The number of molecules is indicated with the solid lines and filled symbols; the number of clusters is indicated with dashed lines and open symbols. Uncertainty estimates are ±1 standard deviation.

The trend in the amount of interior water shown in Figure 5.18, PLA>PVL>PE, could be due to differences in the water-accessible volumes of these star polymers, given that these polymers are of different sizes (see Table 5.1). To account for this, we normalized the values of Figure 5.18 by the water-accessible volume for each star polymer. The Voronoi volumes of the hydrophobic regions of each star polymer were observed to be relatively temperature-independent (24,400 Å^3 for the PLA star; 19,200 Å^3 for the PVL; 10,400 Å^3 for the PE). This volume includes the adamantane component and all of the hydrophobic material of each star polymer. However, the deepest part of this volume, consisting of the adamantane and the first few hydrophobic repeat units of each arm, is inaccessible to water due to the high density and steric crowding. Therefore, the volume of hydrophobic material accessible to water in each case is the total hydrophobic volume, minus a core volume (see Table 5.1) that includes the adamantane and the first hydrophobic repeat units of each arm out to the transitional units discussed earlier. The core volume includes adamantane plus the first four lactic acid units in the PLA star polymer, the first two valerolactone units of the PVL star, and the first three ethylene (six carbons) of the PE star. Scaling the water content data of each star polymer in Figure 5.18 by the water-accessible volume to account for these size differences does not significantly change the trend, but makes the PLA and PVL appear a bit more similar, so that the water concentration follows the trend PLA~PVL>PE. An alternative to this, normalizing the water content by the total hydrophobic surface area, produced similar results.

For each interior water molecule, its depth into the polymer was computed as the distance from the oxygen site of that interior water to the closest oxygen of a water molecule in the bulk. By our definition of an interior water molecule, there must be at least one intervening Voronoi polyhedron from a star polymer site separating these oxygen sites, so there is a minimum bulk to interior water distance observed of about 4 Å by this metric. These depth profiles are shown in Figure 5.19 for each star polymer at the four temperatures studied. It can be seen that these curves are very noisy at low temperatures since so few interior water molecules were observed (<0.5 per frame). As the temperature increases, the curves get smoother and penetration is deeper into the interiors of the star polymers. The penetration depth follows the trend PLA> PVL> PE, suggesting again that ester-rich environments may be somewhat more favorable to water than alkane environments, however one must also keep in mind that the sizes of these star polymer differ, and some of the penetration depth trend could be explained by that difference.

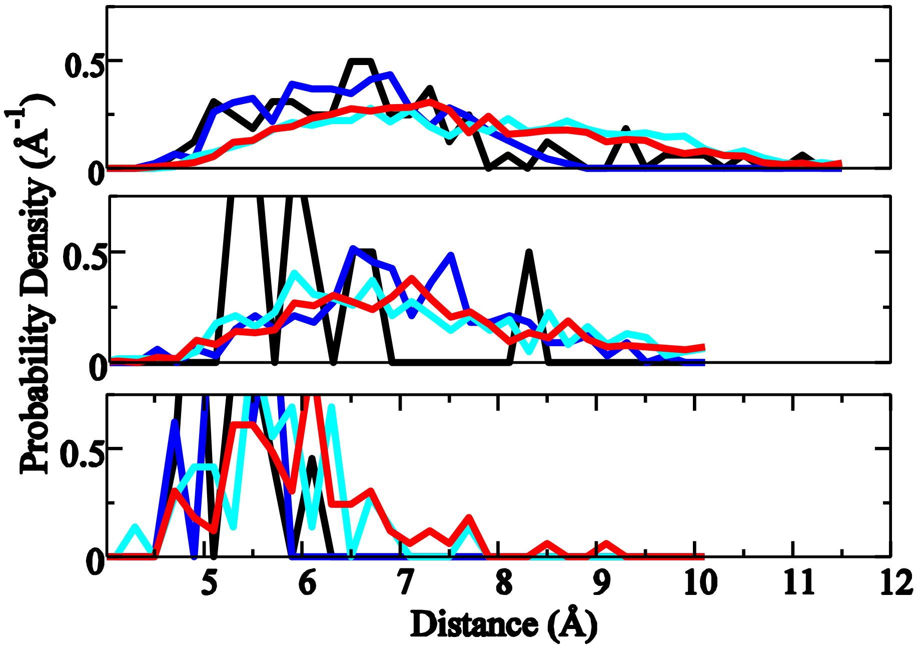

Figure 5.19: Depth profiles for water molecules penetrating into the interior of star polymers. Top panel is for PLA, middle for PVL, and bottom for PE star polymers. In each case, the lines represent the probability density by depth for 300 K (black), 350 K (blue), 400 K (cyan), and 450 K (red).

The number and duration of water penetration events were also noted. We defined a water penetration event as having a beginning time, the time of first observation of the water molecule in the interior (immediately before which it was not in the interior), some number of consecutive observations, and an ending time (immediately after which that water was not an interior water). If an interior water molecule was observed in only one coordinate frame, the event was given a lifetime of 40 ps, the sampling period for coordinate sets used in the Voronoi analysis. If the same water molecule was observed to be interior at two successive times the event was given a lifetime of 80 ps, and so on. This does not, of course, account for the possibility that water molecules might have come and gone and returned between sampling periods. The number of water penetration events observed per picoseconds is shown in Figure 5.20, where, again, one can see that the number water penetration events follows the trend PLA>

PVL>PE. Normalized as above to account for differences in the volumes of the star polymers, the trend is only slightly different (PLA>PVL>>PE). Histograms giving the fraction of events as a function of their lifetimes are shown in Figure 5.21, where it can be seen that by far most penetration events last for only one sampling.

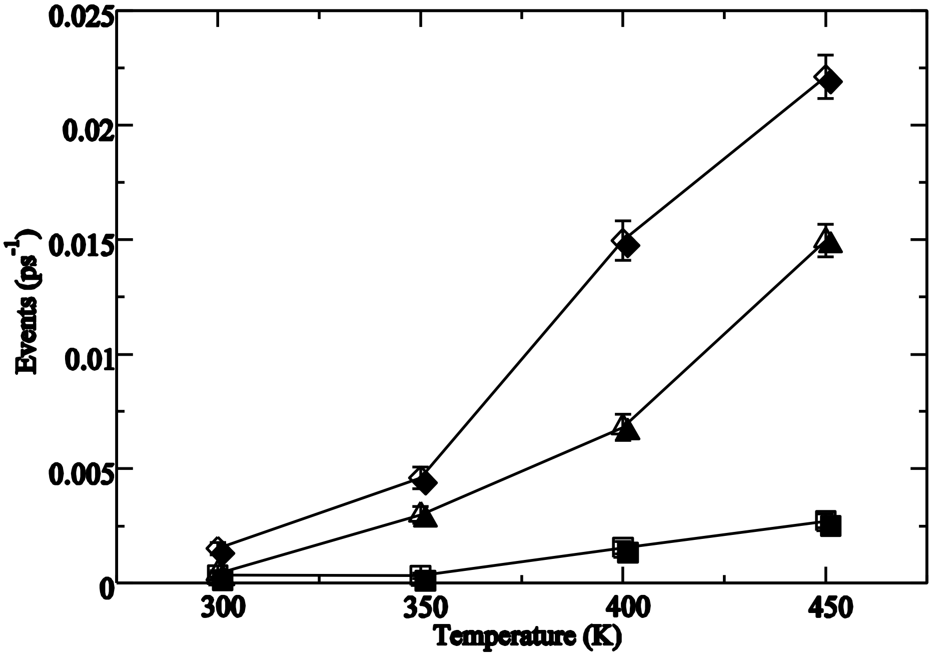

Figure 5.20: Number of water penetration events observed per picosecond for PLA (diamond), PVL (triangle), and PE (square) star polymers. Uncertainty estimates are ±1 standard deviations.

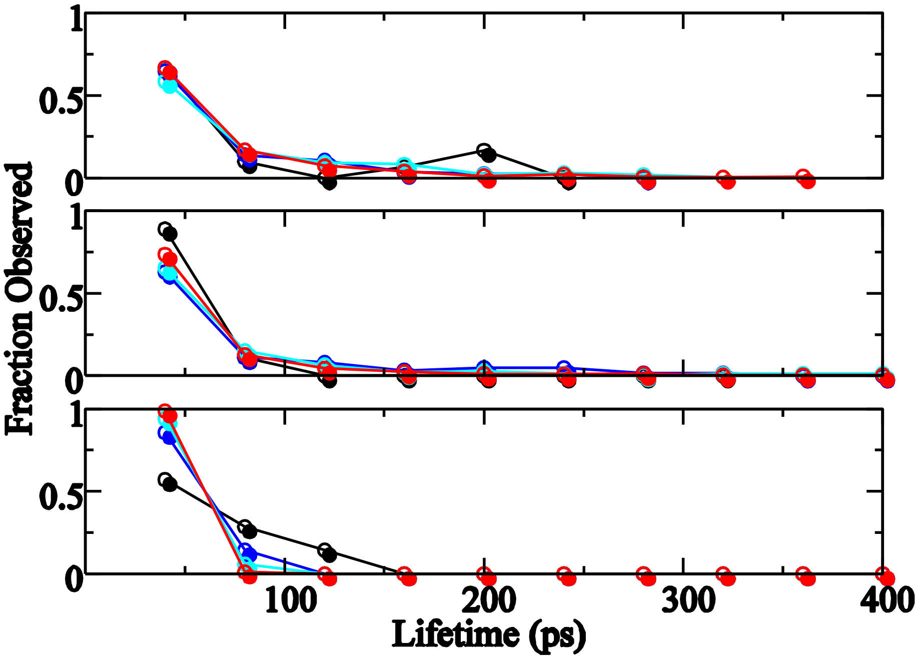

Figure 5.21: Water penetration lifetime distributions. The top panel is for PLA, middle for PVL, and bottom for PE star polymers. In each case, the data represent the fraction of water penetration events lasting various amounts of time for 300 K (black), 350 K (blue), 400 K (cyan), and 450 K (red).

There were rare but notable instances of water molecules existing in clusters, for example, as water dimers or trimers. The average number of water clusters per configuration is shown in Figure 5.18 along with the average number of water molecules observed. When nearly all observed water occurred as monomers, as for the PE star, these curves lay on top of each other. The size distribution for these clusters is shown in the bar charts in Figure 5.22, where it can be seen that most interior water molecules were monomers, and that water trimers were extremely rare.

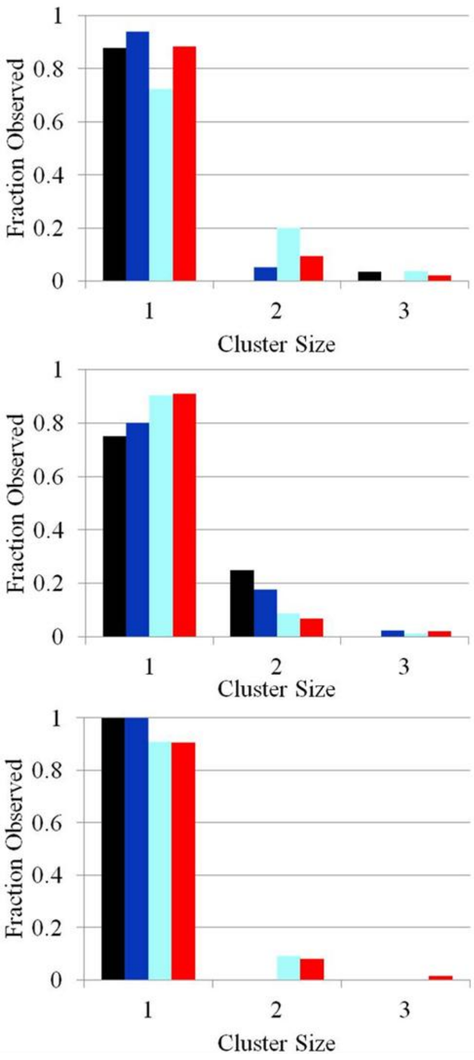

Figure 5.22: Interior water cluster size histogram. The top panel is for PLA, middle for PVL, and bottom for PE star polymers. In each case, the data represent the fraction of interior water clusters that were water monomers, water dimers or water trimers for 300 K (black), 350 K (blue), 400 K (cyan), and 450 K (red).

5.4 Discussion

This simulation study precedes a series of small angle neutron scattering (SANS) and backscattering spectroscopy (BASIS) experiments to be performed on very similar star copolymer molecules. These experiments are meant to probe the structure, organization, water content and kinetics of these kinds of star polymers over a range of temperatures and should produce results that can be compared against this study. The predictions are outlined below.

5.4.1 Hydrophobic Cores

The hydrophobic cores in these molecules are strongly phase-separated from the other material, both from the hydrophilic segments of the strands and from water. Moreover, the hydrophobic regions in each case are relatively rigid, with the PLA and PVL polymers showing glassy (i.e., disordered and slow) behavior, and the PE showing crystalline behavior until rather high temperatures. The slow kinetics manifests itself in a difficulty to sample adequately some of the observables for the PLA and PVL star polymers at the lowest temperature studied (300 K) on the 20 ns timescale of the production simulations. Slow kinetic behavior in the hydrophobic core is also apparent from the reorientational correlation functions.

The highly ordered cylindrical structure of the PE star polymer that persists even at 350 K is quite striking given that the 16 arms are more or less symmetrically placed around the adamantane in a way that should destabilize this kind of structure. We note that the structures formed at low temperatures by the polyethylene in the PE star are reminiscent of the structures formed in aqueous solvent by attached alkane chains on the surface of gold nanoparticles [193].

Further evidence of the solid-like behavior of these molecules is that the surfaces of the hydrophobic regions are highly irregular showing asymmetry, pits, and grooves. The solid behavior persists in these materials until temperatures reach 400 K or higher, at which we see much shorter correlation times for structural changes in the hydrophobic region. Of the two glassy polymers, the PLA polymer seems to be more rigid than the PVL, possibly due to some combination of a higher density of ester groups in the PLA, or more flexibility from the larger alkane segment between ester groups in the PVL. This would suggest that one could produce more liquid-like hydrophobic regions in star polymers by placing longer alkane segments between the ester groups. There may be a limit to this since as one approaches very large segments of alkane, one might begin to see the ordered and very solid structures seen in the PE star polymer.

The fact that the rigidity and structure of the hydrophobic regions persist over such a range of temperatures suggests that this is not likely to be an artifact of the force field and model. We note that the glass transition temperatures for bulk long chain polymers of the same material as our hydrophobic cores suggest these regions could possibly be glassy in the context of a star polymer, but it is not clear whether or not glass transition temperatures of bulk materials apply for the case of a star polymer, with arms that are short, tethered to common connection points, and in contact with water. In fact, our original expectation was that the hydrophobic regions would be much more fluid-like.

5.4.2 Hydrophilic (PEO) Region

The hydrophilic (PEO) regions of all of the star polymers are very disordered and dynamic, exhibiting significant motion on 100 ps timescales even at the lowest temperatures (300 K) studied. This behavior is apparent in the reorientational correlation functions, where one can see the striking behavioral contrast with the hydrophobic regions to which they are connected. The hydrophilic regions are highly solvated with water, and the PEO segments in these simulations appear to prefer contact with water to contact with the hydrophobic core. However, the thermal behavior of the PEO is most interesting and unusual. There is a slight increase in PEO-hydrophobic core interfacial area, but a large decrease in PEO-water interfacial area with increasing temperatures, suggesting a tendency of the PEO to self-aggregate with increasing temperature, reminiscent of a lower critical solution temperature (LCST) seen in PEO-water mixtures where one can observe a mixture go from a single phase (miscibility) to two phases with increasing temperature. We believe that further experimental and simulation studies should be done to investigate this effect.

We note that the hydrophilic (PEO) regions of the diblock arms in our star polymers are very short compared with those in some of the recently synthesized and simulated [223] star polymer systems. Because of this, there is a limit to how large a fraction of the hydrophobic region can be covered by the PEO in our models. The recent work by Huynh et al. [223] has suggested that longer PEO segments would have a greater tendency to protect the hydrophobic region than the relatively short chains of this study. We have not varied the PEO segment length in this study. However, the tendency for association of PEO with the hydrophobic core in our model appears to be so weak that the chain length may not adversely affect our results for these types of hydrophobic materials. Clearly, there are entropic forces at play in this matter as well, but we speculate that the balance of interactions exhibited by the force field among water, PEO, and hydrophobic material is more likely to affect these observations than the PEO chain length, since stronger or weaker interactions could tip the balance in various directions. Our water model was TIP4P-Ew, whereas the work of Huynh et al. employed the SPC water model, which has a smaller dipole moment. Also, we used a more solvent-polarized model for the PEO than Huynh et al., resulting in a larger partial charge on the PEO oxygen sites. Both of these aspects could serve to increase the solubility of our PEO. Subsequent work might assess the sensitivity to the water and/or star polymer force field of the relative tendency of PEO to associate with itself, with water, or with the hydrophobic material. However, whether there is sensitivity due to chain length, force field, or temperature, this hints that the degree of protection of the hydrophobic core offered by the hydrophilic region, or its thermal dependence, could be engineered by minor changes in chain length and composition.

The trend in affinity of PEO for the hydrophobic core seen in the interfacial surface area (Figure 5.16; PLA>PVL>PE) suggests that PEO has greater affinity for ester-rich hydrophobic material than for alkane-rich material, which might not be surprising. We note, however, that some of the trend in Figure 5.16 is also due to differences in the volumes of the hydrophobic regions, but the conclusions remain the same when this is accounted for.

Similar observations have been made in a different context by other workers. Yang et al. [192] reported on simulations of PE chains and PEO chains attached to gold nanoparticles in an aqueous environment where they found solvation and closer penetration by water near PEO-coated surfaces, and greater water exclusion from the PE-coated gold nanoparticles.

5.4.3 Interior Water

The orientationally averaged mass density seems to suggest that water penetrates rather far into the hydrophobic region of the polymer. However, the Voronoi analysis suggests that this is an illusion created by orientational averaging over the misshapen and rough hydrophobic surface. In fact, there is a very well defined water-hydrophobic region interface. These observations are consistent with those of Huynh et al. [223].

In spite of the phase separation between hydrophobic and hydrophilic material, there is a very small amount of water that transiently enters the hydrophobic region, with a probability trend of PLA>PVL>>PE, showing a decrease with decreasing ester and increasing alkane content. Water penetration increases with temperature and, so, appears to be thermally activated. However, part of this effect may be due to a change in phase as the hydrophobic core becomes more fluid-like at the higher temperatures. The increase in water content with temperature is much greater for PLA. Some of the differences in water uptake seen among star polymers in our simulations can be explained by different sizes and surface areas of the hydrophobic regions of these systems, but the trend remains the same when allowance is made for these differences.

Analysis of the temporal behavior of water entry into the hydrophobic regions reveals that such events are rather rare, very short lived, not very deep into the interior (even for our admittedly small polymers), and predominantly involve single water molecules rather than dimers or trimers of water. Because small molecule esters are relatively soluble in water, our expectations were that there would be much more water diffusing into the ester-rich interiors of these star polymers. Apparently, although ester groups interact favorably with water, the interactions among this hydrophobic material are stronger than the water-ester interactions and lead to the expulsion of water. Finally, we feel that there is enough bulk water contact with the hydrophobic material and enough penetration of water that it could help facilitate a slow degradation of the star polymer as well as enable drug release. Based on the differences among the star polymers of this study we feel that such attributes might be controllable with changes in polymer composition and topology.

5.4.4 Caveats

In most respects, our results are remarkably consistent with the observations of Huynh et al. [223] in their study of 13 different six-arm star polymers based on differing lengths of hydrophobic polycaprolactone (PCL) and hydrophilic PEO segments. A direct numerical comparison of our results with theirs is not possible because of a number of differences in the two studies: the numbers of arms, the segment lengths, the connection mechanism, and different metrics for the computation of water contact and interfacial surface area. We note that their studies showed that with increasing length of PEO segments relative to that of the PCL segment, the fraction of the total hydrophobic surface area that is protected from solvent increased from about 60% (short PEO segments) to about 90% (long PEO segments). We did not explore the effects of changing the length of the PEO segment, but our PEO segments appear to be significantly more soluble in water than theirs, resulting in less protection of the hydrophobic region. We would predict that a PCL-based star polymer, with five methylene groups between ester groups, should behave similarly to our PVL star polymer, with four methylene groups.

Because of the internal strain caused by connecting 16 diblock polymer arms to adamantane, we have observed that as one moves from monomer unit to monomer unit along each of the arms that the first couple of units are structurally and kinetically constrained until one reaches some transitional units, after which the material begins to behave in a way that is probably more representative of larger star polymers. These transitional units can be identified from analysis of the orientationally averaged mass distribution function and the reorientational correlation functions, both resolved by monomeric unit. The transitional units might change somewhat with temperature, moving closer to the core with increasing temperature. Some of our results may be affected by the small size of these star polymer models since the rigid part of the hydrophobic core might provide a template that artificially stabilizes anomalous structures and adversely affects kinetics and sampling. Subsequent work should be performed with longer arms and/or a more extended or realistic core than adamantane to validate or challenge these results. In general, details of our conclusions may depend somewhat on force field parameters, but we feel the general trends with temperature and with composition of the hydrophobic core are realistic.

One might ask whether the simulations were sufficiently long for adequate sampling, especially given that those of Huynh et al. were at least 200 ns per star polymer system. Their study was done on systems at 300 K, and we note that at that temperature we had difficulty with sampling some of the observables for the PLA and PVL star polymers using 20 ns production simulations. However, they noted that although their simulations exceeded 200 ns, most observables were stable after about 15 ns of sampling. Moreover, we feel that at our higher temperatures (350 K, 400 K and 450 K) we were able to sample adequately due to the shorter correlation times at these temperatures. Except for a few of the observables measured at 300 K, all of our results follow reasonable systematic trends with temperature, and the uncertainty estimates appear to be realistic. That is, simulations at higher temperatures allow us to estimate the temperatures for which we were not able to sample adequately, and to help establish an effective glass transition temperature for the polymer where the relaxation times begin to exceed the production simulation time. In fact, this is part of the basis for our suggestion that these polymers are glassy at 300 K and 350 K.

We note that the lamellar structure of the PE at 300 K is similar to the extended structure of this polymer before equilibration and production are performed, and that the lamellar structure might therefore be seen as an artifact of incomplete sampling at low temperature. Due to the persistence of order in the PE structure even at 350 K, we argue that the low-temperature structure is not artifactual. In order to test this hypothesis, future work might include starting a simulation of PE at 300 K using a starting structure from the high-temperature simulation at 400 K or 450 K.

Our study employed constant volume simulations to mimic canonical (NVT) ensembles, rather than temperature- and pressure- controlled simulations to mimic isobaric-isothermal (NpT) ensembles. Due to the method of preparation, the density in all of these simulations was designed to be appropriate near 300 K, and consequently slightly too large for the higher temperatures. However, our goal in performing the higher temperature simulations was to assess the stability of our results rather than to represent accurately these higher temperatures. In any case, the higher two temperatures of our study are known to be above the vaporization temperature of the TIP4P-Ew water model. Moreover, the thermal expansion coefficient of water is such that the error in the density is still rather small, provided one remains in the liquid state.

Pressure denaturation is a phenomenon known in protein science where proteins can be unfolded by subjecting them to increased pressure. The exact mechanism of this effect has been debated, but one suggestion is that increasing the pressure increases the chemical potential of water molecules in the bulk water (i.e., surrounding the protein) and drives them into the interior of the protein, causing disruption of intraprotein hydrogen bonding. If such an effect were operative here, the increased pressure at our higher temperatures might drive water into the hydrophobic interior, and/or cause disruption of PEO-PEO interactions. The Voronoi analysis does indicate a slight increase in water content with temperature that could be pressure-induced, but the water content is so small, this is probably not significant if operating at all. The effect of increased pressure on the high-temperature aggregation of the PEO might cause the aggregation phenomenon to shift in our simulations to slightly higher temperatures, and this might be worthy of investigation.

Finally, we note that the thermal control mechanism used in this study could, in principle, affect temporal observables such as correlation times, rates of water penetration events, and water absorption lifetimes. However, we do not expect this to affect the trends we have observed with respect to composition and temperature.

5.5 Conclusions

We have performed molecular dynamics simulations at four different temperatures on three different star polymers, each with 16 linear diblock copolymer arms bonded to a small adamantane core. Across the three star polymer types, there is a difference in the degree of ester versus alkane content in the hydrophobic component of each arm, including one rich in ester content (PLA), one with a mix of ester and alkane content (PVL), and one with pure alkane content (PE). Whereas earlier simulation studies have explored star polymer behavior at a single temperature and investigated the effect of variations in chain number, length, and composition, but for a given type of hydrophilic and hydrophobic material, we have looked at thermal effects and considered three different types of hydrophobic material.

In all situations, there is a pronounced phase separation of the alkane (PE) and ester (PLA, PVL) hydrophobic material from the rest, producing a phase with virtually no water, and with no mixing with the hydrophilic (PEO) material. The hydrophilic material mixes very well with the water at low temperatures, but exhibits signs of phase separation itself at higher temperatures, reminiscent of a lower critical solution temperature (LCST) effect also seen in PEO-water mixtures. At higher temperatures the PEO material condenses and increases its contact with the hydrophobic material.

Structural (density profiles) and kinetic analysis (orientational correlation functions) indicate that the hydrophobic material is solid-like and either very viscous/glassy (ester-based hydrophobic material, PLA and PVL) or structured (alkane material, PE). The phase separation and solidity of the hydrophobic material in these systems renders them rather impermeable to water, and water entry events are rare, short-lived, and shallow in spite of the fact that the ester-based material has a high density of hydrophilic functional groups.

With respect to the use of these kinds of star polymers as transporters of hydrophobic drug molecules, we feel that the above observations imply that drug molecule cargos would have a difficult time being absorbed into them and are more likely to be adsorbed onto the hydrophobic surface of these materials, i.e., at either a water-hydrophobic material interface, or at a hydrophilic-hydrophobic material interface, rather than being encapsulated in their interiors. If this is the case, we predict drug loading to be proportional to the surface area rather than the volume of the star polymer for star polymers of these compositions.

As evidenced by various shape descriptors and surface area measurements, the surfaces of the hydrophobic regions of our star polymers are generally irregular and misshapen, with grooves and pits. For star polymers with larger hydrophobic regions, there might be a greater tendency to adopt more spherical shapes, but depending on the composition, it is also possible that the hydrophobic surface could have a very rough or fractal nature with a surface area that increases more rapidly than as $(\text{volume})^{2/3}$. This, of course, would affect drug loading as well, if it occurs at interfaces. We feel shape, surface area, rigidity, and their effect on cargo loading will be an important area for experimental characterization.

We believe that the tendency for PEO segments to self-aggregate with increasing temperature deserves further experimental and theoretical investigation and could offer a means to control the behavior and function of these types of polymers. In the context of the use of star polymers for drug delivery, an example might be the use of hydrophilic chains engineered in composition and density to increase contact with water when the temperature is lowered slightly, allowing greater exposure of the hydrophobic core and controlled release of cargo stimulated by temperature drop.

“Nature is alive and is talking to us. This is not a metaphor.”

--Terence McKenna

6. Conclusions

Major advances have been made in the field of biomolecular simulation since the discipline came to life 50 years ago. Improvements in computational resources have progressed alongside advances in our theoretical understanding of the physics underlying biological systems. These strides have increased not only the temporal extent of our simulations, but also their accuracy in reproducing and predicting the thermodynamic and kinetic properties of biomolecules. The next 50 years of computational biology hold as much promise as the first, as computational techniques are increasingly applied to some of the greatest challenges currently facing the fields of biology and medicine.

Although biomolecular simulation has thus far achieved great success, limitations remain in the ability of this methodology to accurately model biomolecular structure and dynamics, and to solve problems efficiently in terms of the computational time and resources required. These limitations are rooted in the complexity of the systems under study; in order for many of the problems probed by simulation to be computationally tractable, we must build models of these systems which are reduced in complexity when compared with those found in Nature. The force fields that are used in modeling biological systems serve to approximate the real physics, and are in many cases chosen for computational tractability rather than accuracy. While these force fields have certain shortcomings, their use by a large community ensures that errors are quickly discovered, and ongoing work by many research groups ensures that force fields are continually improved. An additional shortcoming in our simulation arises due to the inability of simulations to completely sample the rugged free energy landscape on which biomolecules exist, as the presence of many local minima limits the extent of conformational sampling that may be achieved. The development of enhanced sampling techniques aims to bypass this limitation by manipulating the formulation of the system's energy function and the equations of motion.

This dissertation has presented two algorithms that aim to enhance the conformational sampling of peptides in stochastic and molecular dynamics simulations while retaining accurate ensemble properties. In Chapter 3, the application of self-guided Langevin dynamics (SGLD) to the folding of three different peptides was explored. The two parameters used with SGLD, the guiding factor and the averaging time, were systematically varied in order to determine their effect on the kinetic rates and thermodynamic stability of the ensembles obtained. In Chapter 4, a variant of replica-exchange molecular dynamics (REMD) was presented which aims to enhance sampling of proteins while increasing the computational efficiency of the simulation.

Chapter 5 of this work outlined the application of molecular dynamics simulation techniques to three polymeric nanoparticle model systems for their potential use in targeted drug delivery applications. Model systems possessing reduced complexity when compared to those fabricated in the laboratory were employed in this case in order to make simulation computationally tractable. As outlined in Chapter 5, we believe that these model systems yield results that are applicable to the laboratory system, and also generalizable to other star polymeric systems.

In building models, it is our responsibility to continually assess their limitations, as the improvement of our models is a vital part of the model-building process. Inaccuracies in the force fields and solvent models used in biomolecular simulation, for example, result in pathological problems in the simulation structures when compared with experimental results. When discovered, however, these errors are reported in the literature by a community of users, and the models are continually improved in order to account for any deficiencies. As we look to the future, much work remains to be done to increase the accuracy of our models and to ensure that our simulations accurately reproduce natural phenomena.

References

1. Anderson P (1972) More is different. Science 177: 393-396.

2. Davies PCW (2004) Emergent biological principles and the computational properties of the universe: Explaining it or explaining it away. Complexity 10: 11-15.

3. Schrodinger E (1944) What is Life? Cambridge, England: The University Press.

4. Laughlin RB, Pines D, Schmalian J, Stojković BP, Wolynes P (2000) The middle way. Proceedings of the National Academy of Sciences 97: 32-37.

5. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. Proceedings of the National Academy of Sciences 37: 251-256.

6. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature 171: 737-738.

7. Franklin RE, Gosling RG (1953) Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. Nature 172: 156-157.

8. Wilkins MHF, Stokes AR, Wilson HR (1953) Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. Nature 171: 738-740.

9. Caspersson T, Schultz, J. (1939) Pentose nucleotides in the cytoplasm of growing tissues. Nature 143: 602-603.

10. Pauling L, Corey RB, Branson HR (1951) The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proceedings of the National Academy of Sciences 37: 205-211.

11. Prusiner SB (1997) Prion diseases and the BSE crisis. Science 278: 245-251.

12. Lansbury PT (1999) Evolution of amyloid: What normal protein folding may tell us about fibrillogenesis and disease. Proceedings of the National Academy of Sciences 96: 3342-3344.

13. Perutz MF (1999) Glutamine repeats and neurodegenerative diseases: molecular aspects. Trends in Biochemical Sciences 24: 58-63.

14. Flory PJ (1953) Principles of Polymer Chemistry. New York: Cornell University Press.

15. de Gennes P-G (1990) Introduction to Polymer Dynamics. Cambridge, England: Cambridge University Press.

16. Bryngelson JD, Wolynes PG (1990) A simple statistical field theory of heteropolymer collapse with application to protein folding. Biopolymers 30: 177-188.

17. Dill KA (1985) Theory for the folding and stability of globular proteins. Biochemistry 24: 1501-1509.

18. McCammon JA (1984) Protein dynamics. Reports on Progress in Physics 47: 1-46.

19. Fischer E (1906) Untersuchungen über Aminosäuren, Polypeptide und Proteïne. Berichte der deutschen chemischen Gesellschaft 39: 530-610.

20. Sanger F, Thompson, EOP. (1953) The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochemical Journal 53: 353-366.

21. Sanger F, Thompson, EOP. (1953) The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochemical Journal 53: 366-374.

22. Dill KA (1990) Dominant forces in protein folding. Biochemistry 29: 7133-7155.

23. Horwich AL, Fenton WA, Chapman E, Farr GW (2007) Two families of chaperonin: physiology and mechanism. Annual Review of Cell and Developmental Biology 23: 115-145.

24. Anfinsen CB, Haber E, Sela M, White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proceedings of the National Academy of Sciences 47: 1309-1314.

25. Levinthal C (1969) How to fold graciously. In: DeBrunner J, editor. Mossbauer Spectroscopy in Biological Systems: Proceedigs of a meeting held at Allerton House, Monticello, Illinois: University of Illinois Press. pp. 22-24.

26. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins: Structure, Function, and Bioinformatics 21: 167-195.

27. Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proceedings of the National Academy of Sciences 89: 8721-8725.

28. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, et al. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181: 662-666.

29. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, et al. (1960) Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5-[angst]. resolution, obtained by x-ray analysis. Nature 185: 416-422.

30. Wüthrich K (1969) High-resolution proton nuclear magnetic resonance spectroscopy of cytochrome c. Proceedings of the National Academy of Sciences 63: 1071-1078.

31. Bartlett AI, Radford SE (2009) An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. Nature Structural and Molecular Biology 16: 582-588.

32. Bernal JD (1964) The Bakerian Lecture, 1962. The Structure of Liquids. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 280: 299-322.

33. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21: 1087-1092.

34. Metropolis N, Ulam S (1949) The Monte Carlo method. Journal of the American Statistical Association 44: 335-341.

35. Alder BJ, Wainwright TE (1957) Phase transition for a hard sphere system. The Journal of Chemical Physics 27: 1208-1209.

36. Rahman A (1964) Correlations in the motion of atoms in liquid argon. Physical Review 136: A405-A411.

37. Stillinger FH, Rahman A (1974) Improved simulation of liquid water by molecular dynamics. The Journal of Chemical Physics 60: 1545-1557.

38. Lifson S, Warshel A (1968) Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. The Journal of Chemical Physics 49: 5116-5129.

39. Levitt M, Lifson S (1969) Refinement of protein conformations using a macromolecular energy minimization procedure. Journal of Molecular Biology 46: 269-279.

40. Karplus M (2003) Molecular dynamics of biological macromolecules: A brief history and perspective. Biopolymers 68: 350-358.

41. Levitt M (2001) The birth of computational structural biology. Nature Structural and Molecular Biology 8: 392-393.

42. Levitt M, Warshel A (1975) Computer simulation of protein folding. Nature 253: 694-698.

43. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267: 585-590.

44. Allen F, Almasi G, Andreoni W, Beece D, Berne BJ, et al. (2001) Blue Gene: A vision for protein science using a petaflop supercomputer. IBM Systems Journal 40: 310-327.

45. Swope WC (2001) Deep computing for the life sciences. IBM Systems Journal 40: 248-262.

46. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annual Review of Biophysics 37: 289-316.

47. Hansson T, Oostenbrink C, van Gunsteren W (2002) Molecular dynamics simulations. Current Opinion in Structural Biology 12: 190-196.

48. van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, et al. (2006) Biomolecular modeling: goals, problems, perspectives. Angewandte Chemie International Edition 45: 4064-4092.

49. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nature Structural and Molecular Biology 9: 646-652.

50. Liu J, Lee H, Allen C (2006) Formulation of drugs in block copolymer micelles: drug loading and release. Current Pharmaceutical Design 12: 4685-4701.

51. Kipp JE (2004) The role of solid nanoparticle technology in the parenteral delivery of poorly water-soluble drugs. International Journal of Pharmaceutics 284: 109-122.

52. Drummond DC, Meyer O, Hong K, Kirpotin DB, Papahadjopoulos D (1999) Optimizing liposomes for delivery of chemotherapeutic agents to solid tumors. Pharmacological Reviews 51: 691-744.

53. Zhang L, Gu FX, Chan JM, Wang AZ, Langer RS, et al. (2007) Nanoparticles in medicine: therapeutic applications and developments. Clinical Pharmacology and Therapeutics 83: 761-769.

54. Cabral C (2010) Multifunctional nanoassemblies of block copolymers for future cancer therapy. Science and Technology of Advanced Materials 11: 1-9.

55. Huynh L, Grant J, Leroux J-C, Delmas P, Allen C (2008) Predicting the solubility of the anti-cancer agent docetaxel in small molecule excipients using computational methods. Pharmaceutical Research 25: 147-157.

56. Tyrrell ZL, Shen Y, Radosz M (2010) Fabrication of micellar nanoparticles for drug delivery through the self-assembly of block copolymers. Progress in Polymer Science 35: 1128-1143.

57. Huynh L, Neale C, Pomès R, Allen C (2012) Computational approaches to the rational design of nanoemulsions, polymeric micelles, and dendrimers for drug delivery. Nanomedicine: Nanotechnology, Biology and Medicine 8: 20-36.

58. Licciardi M, Giammona G, Du J, Armes SP, Tang Y, et al. (2006) New folate-functionalized biocompatible block copolymer micelles as potential anti-cancer drug delivery systems. Polymer 47: 2946-2955.

59. Kabanov AV, Batrakova EV, Alakhov VY (2002) Pluronic® block copolymers as novel polymer therapeutics for drug and gene delivery. Journal of Controlled Release 82: 189-212.

60. Zhang X, Wu D, Chu C-C (2004) Synthesis and characterization of partially biodegradable, temperature and pH sensitive Dex–MA/PNIPAAm hydrogels. Biomaterials 25: 4719-4730.

61. Zeng F, Lee H, Allen C (2006) Epidermal growth factor-conjugated poly(ethylene glycol)-block- poly(δ-valerolactone) copolymer micelles for targeted delivery of chemotherapeutics. Bioconjugate Chemistry 17: 399-409.

62. Huynh L, Neale C, Pomès R, Allen C (2011) Computational approaches to the rational design of nanoemulsions, polymeric micelles, and dendrimers for drug delivery. Nanomedicine: Nanotechnology, Biology and Medicine.

63. Huynh L, Neale C, Pomes R, Allen C (2010) Systematic design of unimolecular star copolymer micelles using molecular dynamics simulations. Soft Matter 6: 5491-5501.

64. Cui Y (2011) Using molecular simulations to probe pharmaceutical materials. Journal of Pharmaceutical Sciences 100: 2000-2019.

65. Case DA, Darden TA, Cheatham TE, 3rd, Simmerling CL, Wang J, et al. (2008) Amber 10. University of California, San Francisco.

66. Christen M, Hünenberger PH, Bakowies D, Baron R, Bürgi R, et al. (2005) The GROMOS software for biomolecular simulation: GROMOS05. Journal of Computational Chemistry 26: 1719-1751.

67. Mackerell AD, Feig M, Brooks CL (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. Journal of Computational Chemistry 25: 1400-1415.

68. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. Journal of the American Chemical Society 118: 11225-11236.

69. Plimpton S (1995) Fast parallel algorithms for short-range molecular dynamics. Journal of Computational Physics 117: 1-19.

70. Cochran AG, Skelton NJ, Starovasnik MA (2001) Tryptophan zippers: Stable, monomeric β-hairpins. Proceedings of the National Academy of Sciences 98: 5578-5583.

71. Schneider JP, DeGrado WF (1998) The design of efficient alpha-helical C-capping auxiliaries. Journal of the American Chemical Society 120: 2764-2767.

72. Neidigh JW, Fesinmeyer RM, Andersen NH (2002) Designing a 20-residue protein. Nature Structural and Molecular Biology 9: 425-430.

73. Born M, Oppenheimer JR (1927) On the quantum theory of molecules. Annalen Physik 84: 457.

74. Jones JE (1924) On the determination of molecular fields. II. From the equation of state of a gas. Proceedings of the Royal Society of London Series A 106: 463-477.

75. Coulomb CAd (1785) Premier memoire sur l'electricite et le magnetisme. Histoire de l'Academie Royale des Sciences: 569-577.

76. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65: 712-725.

77. London F (1937) The general theory of molecular forces. Transactions of the Faraday Society 33: 8-26.

78. Pauli W (1925) Uben den zusammenhang des abschlusses der elektronengruppen im atom mit der komplexstruktur der spektren. Zeitschrift fur Physik 31: 765-783.

79. Buckingham RA (1938) The classical equation of state of gaseous helium, neon and argon. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 168: 264-283.

80. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. The Journal of Chemical Physics 98: 10089-10092.

81. York D, Yang W (1994) The fast Fourier Poisson method for calculating Ewald sums. The Journal of Chemical Physics 101: 3298-3300.

82. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, et al. (1995) A smooth particle mesh Ewald method. The Journal of Chemical Physics 103: 8577-8593.

83. Berendsen HJC, Postma, J.P.M., van Gunsteren, W.F., Hermans, J., editor (1981) Interaction models for water in relation to protein hydration. Dordrecht: D. Reidel Publishing. 331 p.

84. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 79: 926-935.

85. Abascal JLF, Vega C (2005) A general purpose model for the condensed phases of water: TIP4P/2005. The Journal of Chemical Physics 123: 234505.

86. Horn HW, Swope WC, Pitera JW, Madura JD, Dick TJ, et al. (2004) Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. The Journal of Chemical Physics 120: 9665-9678.

87. Fogolari F, Brigo A, Molinari H (2002) The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology. Journal of Molecular Recognition 15: 377-392.

88. Onufriev A, Case DA, Bashford D (2002) Effective Born radii in the generalized Born approximation: The importance of being perfect. Journal of Computational Chemistry 23: 1297-1304.

89. Onufriev A, Bashford D, Case DA (2000) Modification of the generalized Born model suitable for macromolecules. Journal of Physical Chemistry B 104: 3712-3720.

90. Hawkins GD, Cramer CJ, Truhlar DG (1996) Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. Journal of Physical Chemistry 100: 19824-19839.

91. Onufriev AB, D.; Case, D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized Born model. Proteins: Structure, Function, and Bioinformatics 55: 383-394.

92. Mongan J, Simmerling C, McCammon JA, Case DA, Onufriev A (2006) Generalized Born model with a simple, robust molecular volume correction. Journal of Chemical Theory and Computation 3: 156-169.

93. Nguyen H, Roe DR, Simmerling C (2013) Improved generalized Born solvent model parameters for protein simulations. Journal of Chemical Theory and Computation 9: 2020-2034.

94. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, et al. (2012) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. Journal of Chemical Theory and Computation 8: 1542-1555.

95. Shaw D, Dror R, Salmon J, Grossman JP, Mackenzie K, et al. Millisecond-scale molecular dynamics simulations on Anton; 2009; Portland, Oregon. ACM. pp. 1-11.

96. Pierce LCT, Salomon-Ferrer R, Augusto F. de Oliveira C, McCammon JA, Walker RC (2012) Routine access to millisecond time scale events with accelerated molecular dynamics. Journal of Chemical Theory and Computation 8: 2997-3002.

97. Kundt A, Warburg, Emil (1876) Uber die specifische warme des quecksilbergases (On the specific heat of mercury gases). Annalen der Physik 157: 353-369.

98. Verlet L (1967) Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. Physical Review 159: 98-103.

99. Hockney RW, Eastwood, J.W. (1989) Computer Simulation Using Particles. New York: Taylor and Francis.

100. Swope W, Andersen H, Berens P, Wilson K (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. The Journal of Chemical Physics 76: 637-649.

101. Tuckerman M, Berne BJ, Martyna GJ (1992) Reversible multiple time scale molecular dynamics. The Journal of Chemical Physics 97: 1990-2001.

102. Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics 23: 327-341.

103. Andersen H (1983) Rattle: A “velocity” version of the SHAKE algorithm for molecular dynamics calculations. Journal of Computational Physics 52: 24-34.

104. Berendsen HJC, Postma JPM, Gunsteren WFv, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics 81: 3684-3690.

105. Nose S (2002) A molecular dynamics method for simulations in the canonical ensemble. Molecular Physics 100: 191-198.

106. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Physical Review A 31: 1695-1697.

107. Andersen HC (1980) Molecular dynamics simulations at constant pressure and/or temperature. The Journal of Chemical Physics 72: 2384-2393.

108. Grest GS, Kremer K (1986) Molecular dynamics simulation for polymers in the presence of a heat bath. Physical Review A 33: 3628-3631.

109. Lemons DS, Gythiel A (1997) Paul Langevin's 1908 paper ``On the Theory of Brownian Motion" Comptes Rendus de l'Academie des Sciences (Paris) 146, 530--533 (1908)]. American Journal of Physics 65: 1079-1081.

110. van Gunsteren WF, Berendsen HJC (1982) Algorithms for Brownian dynamics. Molecular Physics 45: 637-647.

111. Kubo R (1966) The fluctuation-dissipation theorem. Reports on Progress in Physics 29: 255.

112. Piana S, Lindorff-Larsen K, Shaw David E (2011) How robust are protein folding simulations with respect to force field parameterization? Biophysical Journal 100: L47-L49.

113. Christen M, van Gunsteren WF (2007) On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. Journal of Computational Chemistry 29: 157-166.

114. Lei H, Duan Y (2007) Improved sampling methods for molecular simulation. Current Opinion in Structural Biology 17: 187-191.

115. Schlitter J, Engels M, Krüger P (1994) Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. Journal of Molecular Graphics 12: 84-89.

116. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. The Journal of Chemical Physics 120: 11919-11929.

117. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics 23: 187-199.

118. Laio A, Parrinello M (2002) Escaping free-energy minima. Proceedings of the National Academy of Sciences 99: 12562-12566.

119. Grubmüller H (1995) Predicting slow structural transitions in macromolecular systems: Conformational flooding. Physical Review E 52: 2893-2906.

120. Darve E, Pohorille A (2001) Calculating free energies using average force. The Journal of Chemical Physics 115: 9169-9183.

121. Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. Annual Review of Physical Chemistry 53: 291-318.

122. Jónsson H, Mills G, Jacobsen KW (1998) Nudged elastic band method for finding minimum energy paths of transition. Classical and quantum dynamics in condensed phase simulations: World Scientific. pp. 385-404.

123. Chodera J, Swope W, Pitera J, Dill K (2006) Long-time protein folding dynamics from short-time molecular dynamics simulations. Multiscale Modeling & Simulation 5: 1214-1226.

124. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. The Journal of Chemical Physics 126: 155101.

125. Wu X, Brooks BR (2003) Self-guided Langevin dynamics simulation method. Chemical Physics Letters 381: 512-518.

126. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters 314: 141-151.

127. Wu X, Brooks BR (2011) Toward canonical ensemble distribution from self-guided Langevin dynamics simulation. The Journal of Chemical Physics 134: 134108.

128. Wu X, Brooks BR (2011) Force-momentum-based self-guided Langevin dynamics: A rapid sampling method that approaches the canonical ensemble. The Journal of Chemical Physics 135: 204101.

129. Wu X, Wang S (1998) Self-guided molecular dynamics simulation for efficient conformational search. Journal of Physical Chemistry B 102: 7238-7250.

130. Wu X, Wang S (2000) Folding studies of a linear pentamer peptide adopting a reverse turn conformation in aqueous solution through molecular dynamics simulation. Journal of Physical Chemistry B 104: 8023-8034.

131. Wu X, Wang S (2001) Helix folding of an alanine-based peptide in explicit water. Journal of Physical Chemistry B 105: 2227-2235.

132. Wu X, Wang S, Brooks BR (2002) Direct observation of the folding and unfolding of a beta-hairpin in explicit water through computer simulation. Journal of the American Chemical Society 124: 5282-5283.

133. Lahiri A, Nilsson L, Laaksonen A (2001) Exploring the idea of self-guided dynamics. The Journal of Chemical Physics 114: 5993-5999.

134. Pastor RW, Brooks BR, Szabo A (1988) An analysis of the accuracy of Langevin and molecular dynamics algorithms. Molecular Physics 65: 1409-1419.

135. Tildesley D, Allen MP (1987) Computer Simulations of Liquids. Oxford: Clarendon Press.
136. Lax M (1966) Classical noise IV: Langevin methods. Reviews of Modern Physics 38: 541-566.

137. Brünger A, Brooks CL, Karplus M (1984) Stochastic boundary conditions for molecular dynamics simulations of ST2 water. Chemical Physics Letters 105: 495-500.

138. Wu X, Wang S (1999) Enhancing systematic motion in molecular dynamics simulation. The Journal of Chemical Physics 110: 9401-9410.

139. Olson MA, Chaudhury S, Lee MS (2011) Comparison between self-guided Langevin dynamics and molecular dynamics simulations for structure refinement of protein loop conformations. Journal of Computational Chemistry 32: 3014-3022.

140. Damjanovic A, Wu X, Garcia-Moreno E. B, Brooks BR (2008) Backbone relaxation coupled to the ionization of internal groups in proteins: A self-guided Langevin dynamics study. Biophysical Journal 95: 4091-4101.

141. Damjanovic A, Miller BT, Wenaus TJ, Maksimovic P, Garcia-Moreno E. B, et al. (2008) Open science grid study of the coupling between conformation and water content in the interior of a protein. Journal of Chemical Information and Modeling 48: 2021-2029.

142. Lee MS, Olson MA (2010) Protein folding simulations combining self-guided Langevin dynamics and temperature-based replica exchange. Journal of Chemical Theory and Computation 6: 2477-2487.

143. Cochran AG, Skelton NJ, Starovasnik MA (2001) Tryptophan zippers: Stable, monomeric beta-hairpins. Proceedings of the National Academy of Sciences of the United States of America 98: 5578-5583.

144. Simmerling C, Strockbine B, Roitberg AE (2002) All-atom structure prediction and folding simulations of a stable protein. Journal of the American Chemical Society 124: 11258-11259.

145. Qiu L, Pabit SA, Roitberg AE, Hagen SJ (2002) Smaller and faster: The 20-residue trp-cage protein folds in 4 μs. Journal of the American Chemical Society 124: 12952-12953.
146. Chowdhury S, Lee MC, Xiong G, Duan Y (2003) Ab initio folding simulation of the trp-cage mini-protein approaches NMR resolution. Journal of Molecular Biology 327: 711-717.

147. Snow CD, Zagrovic B, Pande VS (2002) The trp-cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. Journal of the American Chemical Society 124: 14548-14549.

148. Pitera JW, Swope W (2003) Understanding folding and design: Replica-exchange simulations of ``Trp-cage" miniproteins. Proceedings of the National Academy of Sciences of the United States of America 100: 7587-7592.

149. Scheraga HA, Vila JA, Ripoll DR (2002) Helix-coil transitions re-visited. Biophysical Chemistry 101-102: 255-265.

150. Miller JS, Kennedy RJ, Kemp DS (2001) Short, solubilized polyalanines are conformational chameleons: Exceptionally helical if N- and C-capped with helix stabilizers, weakly to moderately helical if capped with rigid spacers. Biochemistry 40: 305-309.

151. Lin JC, Barua B, Andersen NH (2004) The helical alanine controversy: An (Ala)6 insertion dramatically increases helicity. Journal of the American Chemical Society 126: 13679-13684.

152. Nymeyer H, Garcia AE (2003) Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized Born approximation with explicit solvent. Proceedings of the National Academy of Sciences of the United States of America 100: 13934-13939.

153. Garcia AE, Sanbonmatsu KY (2002) Alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds. Proceedings of the National Academy of Sciences of the United States of America 99: 2782-2787.

154. Song K, Stewart JM, Fesinmeyer RM, Andersen NH, Simmerling C (2008) Structural insights for designed alanine-rich helices: comparing NMR helicity measures and conformational ensembles from molecular dynamics simulation. Biopolymers 89: 747-760.

155. Still WC, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. Journal of the American Chemical Society 112: 6127-6129.

156. Roitberg AE, Okur A, Simmerling C (2007) Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. Journal of Physical Chemistry B 111: 2415-2418.

157. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized Born model. Proteins: Structure, Function, and Bioinformatics 55: 383-394.

158. Okur A, Roe DR, Cui G, Hornak V, Simmerling C (2007) Improving convergence of replica-exchange simulations through coupling to a high-temperature structure reservoir. Journal of Chemical Theory and Computation 3: 557-568.

159. Simmerling C, Elber R, Zhang J, editors (1995) MOIL_View - A Program for Visualization of Structure and Dynamics of Biomolecules and STO - A Program for Computing Stochastic Paths. Netherlands: Kluwer Academic Publishers. 241-265 p.

160. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.

161. Zheng W, Andrec M, Gallicchio E, Levy RM (2007) Simulating replica exchange simulations of protein folding with a kinetic network model. Proceedings of the National Academy of Sciences 104: 15340-15345.

162. Smith LJD, X.; van Gunsteren, W.F. (2002) Assessing equilibration and convergence in biomolecular simulations. Proteins: Structure, Function, and Genetics 48: 487-496.

163. Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chemical Physics Letters 281: 140-150.

164. Swendsen RH, Wang J-S (1986) Replica Monte Carlo simulation of spin-glasses. Physical Review Letters 57: 2607.

165. Tesi MC, van Rensburg EJJ, Orlandini E, Whittington SG (1996) Monte Carlo study of the interacting self-avoiding walk model in three dimensions. Journal of Statistical Physics 82: 155-181.

166. Feig M, Karanicolas J, Brooks CL (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. Journal of Molecular Graphics and Modelling 22: 377-395.

167. Garcia AE, Sabonmatsu KY (2001) Exploring the energy landscape of a beta-hairpin in explicit solvent. Proteins: Structure, Function, and Genetics 42: 345-354.

168. Kinnear BS, Jarrold MF, Hansmann UHE (2004) All-atom generalized-ensemble simulations of small proteins. Journal of Molecular Graphics and Modelling 22: 397-403.

169. Roe DR, Hornak V, Simmerling C (2005) Folding cooperativity in a three-stranded beta-sheet model. Journal of Molecular Biology 352: 370-381.

170. Zhou R, Berne BJ, Germain R (2001) The free energy landscape for β-hairpin folding in explicit water. Proceedings of the National Academy of Sciences 98: 14931-14936.

171. Sugita Y, Kitao A, Okamoto Y (2000) Multidimensional replica-exchange method for free-energy calculations. The Journal of Chemical Physics 113: 6042-6051.

172. Cheng X, Cui G, Hornak V, Simmerling C (2005) Modified replica exchange simulation methods for local structure refinement. Journal of Physical Chemistry B 109: 8220-8230.

173. Fukunishi H, Watanabe O, Takada S (2002) On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. The Journal of Chemical Physics 116: 9058-9067.

174. Kofke DA (2002) On the acceptance probability of replica-exchange Monte Carlo trials. The Journal of Chemical Physics 117: 6911-6914.

175. Rathore N, Chopra M, Pablo JJ (2005) Optimal allocation of replicas in parallel tempering simulations. The Journal of Chemical Physics 122: 024111.

176. Ferrara P, Apostolakis J, Caflisch A (2000) Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. The Journal of Physical Chemistry B 104: 5000-5010.

177. Cavalli A, Ferrara P, Caflisch A (2002) Weak temperature dependence of the free energy surface and folding pathways of structured peptides. Proteins: Structure, Function, and Bioinformatics 47: 305-314.

178. Zhou R, Berne BJ (1997) Smart walking: A new method for Boltzmann sampling of protein conformations. The Journal of Chemical Physics 107: 9185-9196.

179. Andricioaei I, Straub JE, Voter AF (2001) Smart darting Monte Carlo. The Journal of Chemical Physics 114: 6994-7000.

180. Brown S, Head-Gordon T (2003) Cool walking: A new Markov chain Monte Carlo sampling method. Journal of Computational Chemistry 24: 68-76.

181. Frantz DD, Freeman DL, Doll JD (1990) Reducing quasi-ergodic behavior in Monte Carlo simulations by J-walking: Applications to atomic clusters. The Journal of Chemical Physics 93: 2769-2784.

182. Lyman E, Ytreberg FM, Zuckerman DM (2006) Resolution exchange simulation. Physical Review Letters 96: 028105.

183. Lyman E, Zuckerman DM (2006) Resolution exchange simulation with incremental coarsening. Journal of Chemical Theory and Computation 2: 656-666.

184. Li H, Li G, Berg BA, Yang W (2006) Finite reservoir replica exchange to enhance canonical sampling in rugged energy surfaces. The Journal of Chemical Physics 125: 144902.

185. Okur A, Wickstrom L, Layten M, Geney R, Song K, et al. (2006) Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. Journal of Chemical Theory and Computation 2: 420-433.

186. Lee VY, Havenstrite K, Tjio M, McNeil M, Blau HM, et al. (2011) Nanogel star polymer architectures: A nanoparticle platform for modular programmable macromolecular self-assembly, intercellular transport, and dual-mode cargo delivery. Advanced Materials 23: 4509-4515.

187. Appel EA, Lee VY, Nguyen TT, McNeil M, Nederberg F, et al. (2012) Toward biodegradable nanogel star polymers via organocatalytic ROP. Chemical Communications 48: 6163-6165.

188. Heise A, Hedrick JL, Frank CW, Miller RD (1999) Starlike block copolymers with amphiphilic arms as models for unimolecular micelles. Journal of the American Chemical Society 121: 8647-8648.

189. Srinivas G, Pitera JW (2008) Soft patchy nanoparticles from solution-phase self-assembly of binary diblock copolymers. Nano Letters 8: 611-618.

190. Ebrahim Attia AB, Ong ZY, Hedrick JL, Lee PP, Ee PLR, et al. (2011) Mixed micelles self-assembled from block copolymers for drug delivery. Current Opinion in Colloid and Interface Science 16: 182-194.

191. Hedrick JL, Trollsås M, Hawker CJ, Atthoff B, Claesson H, et al. (1998) Dendrimer-like star block and amphiphilic copolymers by combination of ring opening and atom transfer radical polymerization. Macromolecules 31: 8691-8705.

192. Yang A-C, Weng C-I, Chen T-C (2011) Behavior of water molecules near monolayer-protected clusters with different terminal segments of ligand. The Journal of Chemical Physics 135: 034101.

193. Lane JMD, Grest GS (2010) Spontaneous asymmetry of coated spherical nanoparticles in solution and at liquid-vapor interfaces. Physical Review Letters 104: 235501.

194. Grest GS, Fetters LJ, Huang JS, Richter D (2007) Star polymers: experiment, theory, and simulation. Advances in Chemical Physics: John Wiley & Sons, Inc. pp. 67-163.

195. Likos CN (2001) Effective interactions in soft condensed matter physics. Physics Reports 348: 267-439.

196. Ganazzoli F, Kuznetsov YA, Timoshenko EG (2001) Conformations of amphiphilic diblock star copolymers. Macromolecular Theory and Simulations 10: 325-338.

197. Chang Y, Chen W-C, Sheng Y-J, Jiang S, Tsao H-K (2005) Intramolecular Janus segregation of a heteroarm star copolymer. Macromolecules 38: 6201-6209.

198. Lee H, Larson RG (2009) Molecular dynamics study of the structure and interparticle interactions of polyethylene glycol-conjugated PAMAM dendrimers. The Journal of Physical Chemistry B 113: 13202-13207.

199. Price MLP, Ostrovsky D, Jorgensen WL (2001) Gas-phase and liquid-state properties of esters, nitriles, and nitro compounds with the OPLS-AA force field. Journal of Computational Chemistry 22: 1340-1352.

200. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. Intermolecular Forces: 331-342.

201. Borodin O, Smith GD (2003) Development of quantum chemistry-based force fields for poly(ethylene oxide) with many-body polarization interactions. The Journal of Physical Chemistry B 107: 6801-6812.

202. Ren P, Wu C, Ponder JW (2011) Polarizable atomic multipole water model for molecular mechanics simulation. Journal of Chemical Theory and Computation 7: 3143-3161.

203. Lamoureux G, MacKerell A, Roux B (2003) A simple polarizable model of water based on classical Drude oscillators. The Journal of Chemical Physics 119: 5185-5197.

204. Yu H, Hansson T, Gunsteren WF (2003) Development of a simple, self-consistent polarizable model for liquid water. The Journal of Chemical Physics 118: 221-234.

205. Rick SW, Stuart SJ, Bader JS, Berne BJ (1995) Fluctuating charge force fields for aqueous solutions. Journal of Molecular Liquids 65–66: 31-40.

206. Patel S, Mackerell AD, Brooks CL (2004) CHARMM fluctuating charge force field for proteins: II Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. Journal of Computational Chemistry 25: 1504-1514.

207. McAliley JH, Bruce DA (2011) Development of force field parameters for molecular simulation of polylactide. Journal of Chemical Theory and Computation 7: 3756-3767.

208. Danz R (1979) Theory of electric polarization Vol. 11. Dielectrics in time-dependent fields. Acta Polymerica 30: 65-65.

209. Anderson PM, Wilson MR (2005) Developing a force field for simulation of poly(ethylene oxide) based upon ab initio calculations of 1,2-dimethoxyethane. Molecular Physics 103: 89-97.

210. Jaffe RL, Smith GD, Yoon DY (1993) Conformation of 1,2-dimethoxyethane from ab initio electronic structure calculations. The Journal of Physical Chemistry 97: 12745-12751.

211. Smith GD, Jaffe RL, Yoon DY (1993) Force field for simulations of 1,2-dimethoxyethane and poly(oxyethylene) based upon ab initio electronic structure calculations on model molecules. The Journal of Physical Chemistry 97: 12752-12759.

212. Smith GD, Jaffe RL, Yoon DY (1995) Conformations of 1,2-dimethoxyethane in the gas and liquid phases from molecular dynamics simulations. Journal of the American Chemical Society 117: 530-531.

213. Borodin O, Smith GD, Bandyopadhyaya R, Byutner O (2003) Molecular dynamics study of the influence of solid interfaces on poly(ethylene oxide) structure and dynamics. Macromolecules 36: 7873-7883.

214. Borodin O, Douglas R, Smith GD, Trouw F, Petrucci S (2003) MD simulations and experimental study of structure, dynamics, and thermodynamics of poly(ethylene oxide) and its oligomers. The Journal of Physical Chemistry B 107: 6813-6823.

215. Goutev N, Ohno K, Matsuura H (2000) Raman spectroscopic study on the conformation of 1,2-dimethoxyethane in the liquid phase and in aqueous solutions. The Journal of Physical Chemistry A 104: 9226-9232.

216. Kříž J, Dybal J (2011) Hydration modes of an amphiphilic molecule 2: NMR, FTIR and theoretical study of the interactions in the system water–1,2-dimethoxyethane. Chemical Physics 382: 104-112.

217. Swope WC, Horn HW, Rice JE (2010) Accounting for polarization cost when using fixed charge force fields. I. Method for computing energy. The Journal of Physical Chemistry B 114: 8621-8630.

218. Swope WC, Horn HW, Rice JE (2010) Accounting for polarization cost when using fixed charge force fields. II. Method and application for computing effect of polarization cost on free energy of hydration. The Journal of Physical Chemistry B 114: 8631-8645.

219. Tomasi J, Mennucci B, Cammi R (2005) Quantum mechanical continuum solvation models. Chemical Reviews 105: 2999-3094.

220. Hornak V, Simmerling C (2004) Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations. Journal of Molecular Graphics and Modelling 22: 405-413.

221. Martyna GJ, Tobias DJ, Klein ML (1994) Constant pressure molecular dynamics algorithms. The Journal of Chemical Physics 101: 4177-4189.

222. Theodorou DN, Suter UW (1985) Shape of unperturbed linear polymers: polypropylene. Macromolecules 18: 1206-1214.

223. Huynh L, Neale C, Pomes R, Allen C (2010) Systematic design of unimolecular star copolymer micelles using molecular dynamics simulations. Soft Matter 6.

224. Brostow W, Dussault J-P, Fox BL (1978) Construction of Voronoi polyhedra. Journal of Computational Physics 29: 81-92.

225. Finney JL (1979) A procedure for the construction of Voronoi polyhedra. Journal of Computational Physics 32: 137-143.

226. Connolly M (1983) Solvent-accessible surfaces of proteins and nucleic acids. Science 221: 709-713.

227. Dormidontova EE (2004) Influence of end groups on phase behavior and properties of PEO in aqueous solutions. Macromolecules 37: 7747-7761.