Wang-Landau simulations of thermodynamic behavior in homopolymer systems

by

Daniel T. Seaton

(Under the direction of David P. Landau)

Abstract

This work focuses on the thermodynamic behavior of both flexible and semi-flexible homopolymer systems through the application of the Wang-Landau algorithm. Coarse-grained homopolymer systems are simple models used to understand complex phenomena in polymer and protein systems. The free energy landscape associated with these systems is similar to, yet less complex than those of real polymers and proteins, providing a tractable system for studying thermodynamic and conformational behavior. The results of these investigations can be divided into two separate sections, one studying the behavior of fully-flexible homopolymers for chain lengths up to $N = 561$, and the other mapping out similar conformational and transition behavior in semi-flexible homopolymers up to $N = 55$.

In the case of fully flexible homopolymers, results are generated using a one-dimensional implementation of the Wang-Landau algorithm. Investigations of transition phenomena are performed, with particular attention given to the infinite-chain behavior of the coil-globule and liquid-solid transitions. In addition, analysis of finite chains has revealed unique behavior in regard to solid-solid transitions occurring in the crystalline region of phase space. These features can be understood by considering the behavior of certain “magic number” chain lengths ($N = 13, 55, 147, 309, 561, \ldots$), for which perfect icosahedral geometries form. The
low temperature properties of other chain lengths can be correlated to this magic number behavior using packing considerations.

The conformational behavior of semi-flexible polymers is an extension of these results and methodology, with a key difference being the application of multi-dimensional Wang-Landau sampling. Using the energy and stiffness of the chain as sampling directions, three primary conformational regions occurring at low temperatures are described, namely, solid-globular, rod-like, and toroidal states. The existence of icosahedral ordering within the solid-globular region is of principal interest, and results showing Mackay and Anti-Mackay behavior are presented. The tendency for chains to fold into icosahedral geometries reduces as stiffness is increased, giving way to the less-ordered rod-like and toroidal structures at low temperatures. The overall conformational behavior is mapped out as a function of chain length and stiffness using various thermodynamic and structural quantities. Investigations of both flexible and semi-flexible homopolymers have been aided by the introduction of unique Monte Carlo trial moves, as well as algorithmic improvements, all of which are discussed in detail.

**INDEX WORDS:** Homopolymer, flexible, semi-flexible, Wang-Landau, Monte Carlo, conformational, coil-globule, liquid-solid, solid-solid, transitions, generalized ensemble
WANG-LANDAU SIMULATIONS OF THERMODYNAMIC BEHAVIOR
IN HOMOPOLYMER SYSTEMS

by

DANIEL T. SEATON

B.S., Auburn University, 2004

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010
Wang-Landau simulations of thermodynamic behavior in homopolymer systems

by

Daniel T. Seaton

Approved:

Major Professor: David P. Landau

Committee: Steven P. Lewis
Shan-Ho Tsai
Craig Wiegert

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2010
Acknowledgments

I gratefully thank my family and friends for their unwavering support throughout my doctoral studies. In particular, the love and support from Dr. Sarah H. Craven has allowed me to achieve goals beyond which I originally envisioned. Her presence in my life, along with Layla’s, is more appreciated than she will ever know. In addition, I would like to acknowledge the love, support, and “patience”, of my parents, who have remained a constant source of encouragement throughout my student endeavors. To all of my family, this work could not have been done without you!

I would like to thank my advisor Dr. David P. Landau, whose mentorship has helped me grow both as a scientist and as a person. My doctoral committee, Dr. Craig Wiegert, Dr. Shan-Ho Tsai, and Dr. Steven P. Lewis, has also been incredibly helpful throughout my graduate studies. Each has provided insight into issues ranging from computational methods to teaching approaches, and for this I am grateful. In addition, the advisement from Dr. Michael Bachmann has played a pivotal role in guiding the work included in this dissertation. A number of colleagues have also helped direct this work and my overall scientific demeanor, including Dr. Joao Plascak, Dr. Thomas Wüst, Dr. Stefan Schnabel, Dr. Steven J. Mitchell, and Dr. Claire Gervais. I must also thank a number of peers whose interactions have played a role in my graduate career: Sairam Tangirala, Dr. Wilson A. Smith, David L. Cotten, Ying-Wai Li, Dr. Steven P. Compton, and Dr. Scott J. Thompson.

I would also like to acknowledge various agencies that have provided funding throughout my doctoral studies, namely, The National Science Foundation, The National Institute of Health, and The Department of Physics and Astronomy of The University of Georgia.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACKNOWLEDGMENTS</strong></td>
<td>iv</td>
</tr>
<tr>
<td><strong>List of Figures</strong></td>
<td>viii</td>
</tr>
<tr>
<td><strong>List of Tables</strong></td>
<td>xi</td>
</tr>
<tr>
<td><strong>Chapter</strong></td>
<td></td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Background</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Introduction to polymer systems</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Generalized ensemble techniques</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Fully flexible homopolymers</td>
<td>10</td>
</tr>
<tr>
<td>2.4 Semi-flexible homopolymers</td>
<td>11</td>
</tr>
<tr>
<td>3 Model and predicted behavior</td>
<td>14</td>
</tr>
<tr>
<td>3.1 Single chain homopolymers</td>
<td>14</td>
</tr>
<tr>
<td>3.2 General Hamiltonian and interaction potentials</td>
<td>14</td>
</tr>
<tr>
<td>3.3 Perspective on interactions in fully flexible homopolymers</td>
<td>18</td>
</tr>
<tr>
<td>3.4 Thermodynamic behavior and transitions in fully flexible homopolymers</td>
<td>19</td>
</tr>
<tr>
<td>3.5 Connection to classical clusters</td>
<td>22</td>
</tr>
<tr>
<td>3.6 Perspective on interactions in semiflexible homopolymers</td>
<td>24</td>
</tr>
<tr>
<td>3.7 Thermodynamic behavior and transitions in semi-flexible homopolymers</td>
<td>25</td>
</tr>
</tbody>
</table>
4 Monte Carlo techniques ........................................ 29
   4.1 Overview of Monte Carlo techniques ..................... 29
   4.2 The partition function in Monte Carlo methods ......... 31
   4.3 Metropolis sampling ....................................... 33
   4.4 Estimates of thermodynamic behavior from Metropolis . 34
   4.5 Generalized ensemble methods ........................... 35
   4.6 Thermodynamic calculations from generalized ensemble
techniques ......................................................... 37

5 The Wang-Landau algorithm ..................................... 40
   5.1 Wang-Landau algorithm .................................... 40
   5.2 Modification factor and histogram ......................... 42
   5.3 Detailed balance and convergence ........................ 43
   5.4 Production run for structural quantities ................. 44
   5.5 Wang-Landau parameters in homopolymer simulations .. 45
   5.6 Homopolymer move set .................................... 46

6 Algorithmic considerations and analysis techniques ......... 56
   6.1 Overview .................................................... 56
   6.2 Binning resolution ....................................... 57
   6.3 Saturation of error ....................................... 61
   6.4 Energy range considerations ................................ 64
   6.5 Analyzing multiple independent simulations ............ 67
   6.6 Validating calculations ................................... 73
   6.7 Error bars in density of states and thermodynamic quan-
tities ............................................................... 75
   6.8 Conclusions ................................................ 76

7 Simulation results for flexible homopolymers ............... 79
7.1 One-dimensional Wang-Landau simulation parameters . . . 79
7.2 Heat capacity ................................................. 81
7.3 Structural quantities ........................................ 83
7.4 Transitions and clusters ..................................... 88
7.5 Conclusions ..................................................... 96

8 Simulation results for semi-flexible homopolymers ......... 99
  8.1 Wang-Landau parameters ................................. 100
  8.2 Specific heat ................................................. 103
  8.3 Structural quantities for $N = 30$ ......................... 116
  8.4 Mapping behavior in minimum energy configurations . . . 119
  8.5 Conclusions ..................................................... 128

9 Conclusions ......................................................... 131

Bibliography .......................................................... 146
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Diagram of homopolymer interactions in an $N = 5$ length chain</td>
<td>15</td>
</tr>
<tr>
<td>3.2</td>
<td>Bonded and nonbonded potentials versus the distance between two monomers</td>
<td>17</td>
</tr>
<tr>
<td>3.3</td>
<td>Conformational diagram of fully-flexible homopolymer chains</td>
<td>20</td>
</tr>
<tr>
<td>3.4</td>
<td>Depiction of Mackay and Anti-Mackay packing</td>
<td>23</td>
</tr>
<tr>
<td>3.5</td>
<td>Diagram of conformational behavior as it is related to stiffness (flexibility) and temperature in semi-flexible homopolymers</td>
<td>26</td>
</tr>
<tr>
<td>4.1</td>
<td>Traditional versus generalized ensemble simulations</td>
<td>32</td>
</tr>
<tr>
<td>5.1</td>
<td>Energy dependent displacement moves (step-size and efficiency)</td>
<td>50</td>
</tr>
<tr>
<td>6.1</td>
<td>Absolute error between reference specific heat $C_V^*$ and individual $C_V$ versus binning resolution $dE$ and individual temperatures $T$</td>
<td>59</td>
</tr>
<tr>
<td>6.2</td>
<td>Average density of states calculated for three different chain lengths, $N = 10$, 25, and 50</td>
<td>62</td>
</tr>
<tr>
<td>6.3</td>
<td>Saturation of error ($C_V$ and efficiency)</td>
<td>63</td>
</tr>
<tr>
<td>6.4</td>
<td>Specific heat for a chain length of $N = 50$, varying the minimum energy sampled</td>
<td>66</td>
</tr>
<tr>
<td>6.5</td>
<td>Comparison of two methods for averaging thermodynamic quantities from multiple simulations</td>
<td>70</td>
</tr>
<tr>
<td>6.6</td>
<td>Diagram for understanding energy ranges in analysis methods</td>
<td>72</td>
</tr>
<tr>
<td>6.7</td>
<td>Effects in canonical probabilities due to the minimum sampled energy for $N = 50$</td>
<td>74</td>
</tr>
<tr>
<td>6.8</td>
<td>Specific heat curves for $N = 50$ calculated with the advanced methods based on $P(E,T)$ from individual simulations</td>
<td>77</td>
</tr>
<tr>
<td>7.1</td>
<td>The relative density of states versus energy per monomer for a number of chain lengths</td>
<td>81</td>
</tr>
</tbody>
</table>
7.2 The heat capacity for fully flexible chain lengths 100, 200, 300, 400, and 500  
7.3 The squared radius of gyration versus energy per monomer for fully flexible 
chain lengths 50, 100, 200, 300, 400, and 500 ............................... 83 
7.4 Squared radius of gyration (top) and its derivative (bottom) versus tempera-
ture for fully flexible chain lengths 100, 200, 300, 400, and 500 ......... 85 
7.5 The core density (top) and its derivative (bottom) versus temperature for fully 
flexible chain lengths 100, 200, 300, 400, and 500 .......................... 87 
7.6 Transition temperatures ($T_{tr}$) from the specific heat (top), along with the cor-
responding specific heat ($C_V/N$) values (bottom), plotted versus chain length 
$N$ .................................................. 89 
7.7 Specific heat versus temperature for the magic set of fully flexible chain lengths 91 
7.8 Heat capacity (top) and the derivative of the squared radius of gyration 
(bottom) versus temperature for chain lengths between $13 \leq N \leq 55$ ..... 93 
7.9 Finite-size scaling of the liquid-solid and coil-globule transitions using transi-
tion temperatures provided by features in the specific heat ................. 95 
8.1 Specific heat as a function of temperature ($T$) and stiffness ($J_A$) for a chain 
length of $N = 15$ averaged over five independent simulations ............... 104 
8.2 Conformational diagram for $N = 15$ semi-flexible chain: peaks and shoulders 
found in the specific heat $C_V(J_A,T)/N$, with minimum energy configurations 
found at various $J_A$ ................................................. 106 
8.3 Color maps representing the 2D specific heat $C_V(J_A,T)/N$ for chain lengths 
$N = 20$ (top left), $N = 25$ (bottom left), $N = 30$ (top right), and the error 
associated with $N = 30$ (bottom right) using a different color scale ......... 108 
8.4 Conformational diagram for $N = 20$ semi-flexible chain: peaks and shoulders 
found in the specific heat $C_V(J_A,T)/N$, with minimum energy configurations 
found at various $J_A$ ................................................. 111
8.5 Conformational diagram for $N = 25$ semi-flexible chain: peaks and shoulders 
found in the specific heat $C_V(J_A, T)/N$, with minimum energy configurations 
found at various $J_A$ ................................................................. 112
8.6 Conformational diagram found analyzing specific heat of an $N = 30$ semi-
flexible chain ................................................................. 115
8.7 Color maps representing 2D structural quantities for $N = 30$, including radius 
of gyration $R_g^2$, end-to-end distance $R_{end}$, and the Khalatur parameters $K_1$ and 
$K_2$ ................................................................. 117
8.8 Diagram of conformational behavior for a variety of chain lengths as it is 
related to $J_A$, $E_{min}/N$, and structural fluctuations of MECs .................... 120
8.9 Minimum energy conformations found in the rod-like regions for the $N = 45$
length chain ................................................................. 122
8.10 Diagram of solid-globular states found in chain lengths $N = 30$, 35, 45, and 
55, for stiffness values of $J_A = 1.0$ and 4.0 ........................................ 125
8.11 Diagram of minimum energy configurations found in the toroidal region of 
states for an $N = 55$ semi-flexible chain ........................................ 127
LIST OF TABLES

7.1 Transitions founds in studies of classical clusters ........................................ 90
7.2 Transition temperature estimates for the coil-globule and liquid-solid transi-
tions in fully-flexible homopolymers ............................................................... 97
Chapter 1

Introduction

The work presented in this dissertation focuses on the simulation of coarse-grained polymer systems using state-of-the-art computational techniques. In particular, Monte Carlo simulation methods have been applied in the thermodynamic analysis of coarse-grained flexible and semi-flexible homopolymers. Homopolymer systems are simple models (containing identical monomers) used to understand polymer and protein phenomena at a general level. The free energy landscape associated with these systems is less complex than those of real polymers and proteins, providing a tractable system for studying various behaviors associated with these systems. The goal of this work is to further explore the broad class of problems associated with polymer and protein folding, while developing efficient simulation methods to describe thermodynamic behavior in these systems.

The Wang-Landau algorithm has proven to be exceptionally efficient in the simulation of coarse-grained homopolymers. The primary result of this generalized ensemble method is the density of states, which then leads to the partition function and subsequent thermodynamic properties. Considerable effort has been applied to the analysis of these quantities, using a number of statistical approaches. This has allowed for the full description of conformational behavior in both flexible and semi-flexible homopolymer systems.

The remainder of this dissertation is as follows: the second chapter provides a review of the fundamental concepts of polymers and their simulation. This includes a brief introduction to polymer systems, as well as an overview of computational techniques applied to these systems. In addition, a review of previous studies, particularly those affecting the trajectory of this work, is presented.
The third chapter presents the specific homopolymer model applied in this work. Specific interactions and their implementations are discussed in detail. The discussion includes an overview of the behavior expected in these systems, along with descriptions of studies that have helped shape this understanding.

The fourth chapter presents a broad overview of the Monte Carlo method, particularly as it is applied in statistical thermodynamics. This discussion starts from the Master Equation, highlighting topics such as detailed balance, which holds an important role in understanding methodologies used in subsequent chapters. In addition, the basics of Boltzmann sampling are introduced in the form of the Metropolis algorithm, and subsequently lead to discussions on generalized ensemble simulation techniques, particularly those generalized ensemble methods used to estimate the density of states.

The Wang-Landau algorithm, which is the primary method used to generate results in this dissertation, is presented in the fifth chapter. Contents include a detailed analysis of the algorithms, along with discussions of subtle tuning parameters used to control computational efficiency. Considerable discussion is given to how such parameters should be used in polymer simulations, and comparing these choices with the values presented in the original work on discrete systems. A detailed analysis of the Monte Carlo trial moves (or updates) applied to these polymer systems is also given, marking perhaps one of the most important methodological sections in this dissertation.

Chapter six marks the beginning of the results driven portion of this dissertation. This chapter details some of the most important algorithmic considerations discovered during the whole of this thesis work. These results offer topics that are not necessarily highlighted in the literature, and yet represent critical issues that researchers encounter when performing Wang-Landau simulations of polymer systems. In addition, the outline of this chapter is historical, meaning that results are presented in the order in which they were studied.

Chapter seven represents results of Wang-Landau simulations applied to fully-flexible homopolymers. These simulations were performed using a one-dimensional implementation
of the algorithm, whose parameters are discussed in detail. The results in this chapter can be split into two major categories: discussions of transition behavior in the infinite-chain limit, with particular comparison made with the results of other studies, as well as theoretical predictions; and analysis of finite-size effects, which can be directly compared with behavior found in classical cluster studies occurring for a broad range of chain lengths. The second of these categories represents a shift in the way homopolymers will be analyzed in the future, and concepts from this section play an important role in the analysis of semi-flexible systems.

Chapter eight is the final results driven chapter, and involves the application of Wang-Landau sampling to semi-flexible homopolymers. In order to effectively simulate these homopolymers, a two-dimensional implementation of the Wang-Landau algorithm is applied; algorithmic details are presented at the beginning of the chapter. Semi-flexible homopolymers differ by a single energy term from their fully-flexible counterparts, yet conformational behavior in semi-flexible systems is remarkably more complex. In order to highlight these complexities, conformational behavior is described using a number of quantities, including flexibility, energy, temperature, and chain length. Results are presented highlighting three major categories of minimum energy conformations, namely, solid-globular, rod-like, and toroidal states.

The final chapter is reserved for concluding remarks, further analysis of the data presented, and some indication of future directions for this work.
Chapter 2

Background

2.1 Introduction to polymer systems

Polymer systems represent a broad class of structures and materials, whose equally broad number of functions and applications represent challenges to researchers across a variety of disciplines. Materials scientists attempt to understand the properties of large collections of polymers, often referred to as melts, with the intent of improving material characteristics and sustainability [1]. Life scientists typically look at single chains in the form of proteins, trying to uncover function through its correlation with structure [2, 3]. Physical scientists often attempt to identify the general principles governing behavior of polymer systems, such that these principles can be used to understand specific systems and their applications [4].

A polymer is a large macromolecule made of repeating structural units, or monomers. Proteins represent an important example of a class of polymers referred to as heteropolymers, where sequences comprise a chain of amino acids for which there are twenty possibilities in most organisms [3]. Protein systems represent the biological machinery used to perform crucial tasks in cells, and additionally, play a key role in pharmaceutical applications [5, 6]. The sequence of amino acids is directly related to the overall protein structure, which in turn determines its specific function. Another polymer example comes in the form of polymer melts, where a large number of chains make up a macroscopic material. Polyethylene, is one such example of a homopolymer chain, made up of identical ethylene monomers, the fundamental building block of many modern plastics [1]. The behavior of individual chains affects the overall properties of the system, and in many cases must be considered when examining properties of a polymeric material.
The conformational behavior of single polymer chains remains a challenge to all areas of research. Protein folding provides one of the most tractable examples in this regard, as active research occurs in experiment, computer simulation, and theory. The main goals of these methodologies can be summarized as aiming to understand conformational behavior associated with function, as well as the transition behavior that affects this functionality. In the case of simple two-state folding proteins, the chain of amino acids has two conformational possibilities: folded (representing a functional chain) and unfolded (representing a dysfunctional chain). Researchers measure properties of the folded state in order to gain understanding between structure and function, while also studying the transition between folded and unfolded states. This becomes particularly relevant when conformational and transition behaviors are measured against physical parameters, such as temperature or chemical solvent. The functionality of all polymers, whether in the cell, a material, or in a pharmaceutical application, can be parameterized by various physically observable quantities, and hence greater understanding of the efficacy of such polymer systems can be gained.

All polymers have a wide set of conditions that affect their conformational and transition behavior. One of the most important factors affecting such behavior is the environment, with examples of external properties including temperature, chemical solvent, and constituent macromolecules. Of equal importance are the internal properties of polymer chains, such as sequence (and subsequent monomer type), length, and strength of interactions between monomers. These types of properties are essential in identifying generalities within a broad class of polymers, and all of the above quantities lend themselves to research in various disciplines.

2.1.1 Studying polymers - computer simulation

Many fields are dedicated to understanding conformational and transition behavior in polymer systems. Protein structure and folding again represents a cross-disciplinary area of research. Experimentalists are currently able to analyze protein structure using techniques
such as X-ray diffraction and nuclear magnetic resonance, but are still on the brink of being able to provide insight into the general behavior of all proteins [7]. Many of the resulting structures are stored in the Protein Data Bank [8], where further analysis of these structures can be performed by researchers looking for connections between structure and the sequence of amino acids [2]. Further searching for general behavior is performed in theoretical studies using simplified models that can be interpreted analytically [9, 10]. However, in most cases these models represent chains that are too small or over generalized. Computer simulation has provided an ideal bridge between experiment and theory, allowing for the understanding of a broad range of problems in a variety of systems [4, 11–13].

2.1.2 History of computer simulation

Computer simulation has become a widespread technique for studying a number of unique physical systems. These techniques vary as much as the physical systems, and each has a place in providing useful information to researchers. The work carried out in this dissertation relies solely on Monte Carlo methods, which represent a broad class of techniques relying on random numbers to produce their results. The dynamics of such methods are often well-suited for understanding many naturally stochastic processes, and thus hold a unique position in many areas of research, particularly in that of statistical thermodynamics [12, 13].

2.1.3 Perspective on Monte Carlo

The objective of many Monte Carlo techniques is to randomly generate as many statistically independent configurations of a physical system, while simultaneously measuring properties according to some distribution or ensemble. When considering Monte Carlo (MC) methods [12, 13], it is helpful to compare with another established method, Molecular Dynamics (MD) [11]. MD simulations evolve a given system in physical time by directly applying Newton’s Laws. However, MD is hindered by the accessible time scales over which a simulation can be performed (particularly in polymer and protein systems), and in many cases, is best suited
for studying specific properties over relatively short times. The power of MC methods come from the ability to rapidly generate states of a given system without (in most cases) regard to time-dependent properties. In addition, for the case of polymers and proteins, the time scales at which conformational and transition behavior occur are not always accessible in MD simulations. MC simulations provide a viable option for understanding transition and conformational behavior for a broad range of parameters.

Although the power of computer simulation has grown tremendously since the first application of MC methods in physics [14], protein and polymer systems have persisted in challenging these techniques. As was discussed previously, real polymer and protein systems have complex energy-landscapes, riddled with local minima [15], which are difficult to distinguish from the global energy minimum. In addition, the many transitions that occur in these systems present further difficulties, particularly those that represent subtle geometric changes at low temperatures. Computer simulation techniques are in turn forced to confront these issues through algorithm development. Traditional methods of simulation, such as the Metropolis method, have often been forced to avoid these issues by studying properties only at higher temperatures (or expanded states), analyzing features like the coil-globule transition, which does not require simulation of highly compact states.

2.1.4 Coarse-grained models of polymers

Even with significant gains in computing abilities, many studies of proteins and polymers still rely on simplified models to reduce complexity. Coarse-grained models of polymer systems [4] are used to understand complex phenomena in polymer and protein systems, where the free energy landscape associated with these models is similar to, yet less complex than, those of real polymers and proteins, providing a tractable system for studying thermodynamic and conformational behavior. Coarse-graining refers to the level of physical detail included in a given model, which can be further described as a simplification of monomers and their positions, as well as a simplification of the interactions between these monomers.
In the physical sciences, researchers tend to want to make more general statements regarding the behavior of a broad class of polymers, for example, how the hydrophobic effect influences folding in proteins. The coarse-graining mentioned above can be taken further by replacing the carbon-nitrogen protein backbone with generic monomers having no specific resemblance to physically realizable atoms. Simple models of heteropolymer chains using two generic monomers, (H)ydrophobic and (P)olar, have offered an insightful and simple approach for understanding the hydrophobic effect [16–18]. Furthermore, many of these models are implemented on a lattice, representing a further coarse-graining of polymer systems.

It is also noted that coarse graining can have many meanings and often depends on the desired information. In the life sciences, researchers want to know specific details regarding a biological protein, for example, at what temperature a protein unfolds and becomes dysfunctional. Such a case requires a more realistic representation of the individual monomers, with specific atomic interactions represented within the sequence of amino acids. In order to avoid computational inefficiencies created by all-atom models, a number of multi-step coarse-grained simulation processes have been introduced [3, 19, 20]. Initial coarse-graining in these methodologies removes atomic detail (typically side chains) of the various amino acids by simulating only their common carbon-nitrogen backbone. Once this reduced structure has been folded, and a class of equilibrium states identified, the previously-removed atomic detail is added back to the configuration, and the structure is further refined by the simulation. These multi-step simulation methods often require more procedures than described above, but this example provides an overview of how such processes are typically implemented.

2.2 Generalized ensemble techniques

In the past two decades, generalized ensemble techniques have been introduced to overcome many of the inefficiencies associated with traditional Monte Carlo methods [21–23]. These techniques differ from their traditional counterparts in that a traditional ensemble
used to simulate a particular system is often different for the ensemble from which results are desired. The advantage is that such techniques can be tuned to maximize efficiency in the sampling of conformational space, and still provide information regarding results in the ensemble of interest [21]. Examples of these methods include the Wang-Landau algorithm, multicanonical-sampling, parallel tempering, and simulated tempering. The first two methods use the “density of states” representation of the partition function in order to estimate properties of a system, while parallel and simulated tempering use temperature-dependent sampling criterion with additional methods of exchanging (or swapping) between various probability distributions $P(E, T_i)$. Although all of these methods (as well as others) have uses in various physical situations, the subset including Wang-Landau and multicanonical sampling are of particular interest to this work, representing algorithms aimed at estimating the density of states of a particular system. Techniques aimed at estimating the density of states have a direct relationship with the partition function, and hence with subsequent thermodynamic properties.

Particularly, flexible and semi-flexible homopolymer chain models have benefited greatly from these methods, with studies being performed for a wide range of coarse grainings [24–26]. The nomenclature of “flexible” and “semi-flexible” has broad meaning in terms of coarse-graining, ranging from a number of models using generic monomers, to models with atomic detail meant to represent an experimentally realizable chain. Regardless, the classifications of “flexible” and “semi-flexible” are distinguished by a single bending energy term which regulates the persistence length of the chain (i.e., the average contour length along the chain). This energy term can be adjusted to vary the flexibility (or stiffness) of the chain, and each level of stiffness often has some biological relevance when comparing with the persistence length of real systems. The interactions in these systems are often the basis of more complicated models, and thus flexible and semi-flexible homopolymers are often used to study the most general properties of polymer systems.
2.3 Fully flexible homopolymers

The simulation of flexible homopolymers has been a topic of vigorous research over the past decade, with studies varying from investigations of single chains to the interaction of chains with surfaces or pores [27–30, 26]. In studies of single chains, the models have varied as much as the physical situations, with some being on-lattice [18, 25], some in the continuum [31–37], as well as bond fluctuation lattice models [38–40]. The level of interest is due in part to the number of interesting technological and biological applications, and in part to the availability of generalized ensemble simulation methods. These methods have allowed for the efficient sampling of all thermodynamic behavior, particularly at low temperatures where these systems are in their most collapsed states. Regardless of the model and method used, certain characteristic features remain consistent: As the temperature of an extended, single chain is decreased, a collapse transition from an unwound coil to a liquid globular state (coil-globule) occurs, followed by another transition from a liquid globule to a “solid”, compact globule (liquid-solid).

Recent studies have attempted to illuminate the behavior of these two transitions using Wang-Landau sampling. One set of studies using the bond-fluctuation model [38, 39], and another employing off-lattice simulations [34, 35], have analyzed these transitions for a number of different chain lengths. Noticeable discrepancies resulted in thermodynamic properties, and it was unclear whether these were due to differences between the models or in the details of the application of the Wang-Landau algorithm. As work began for this dissertation [41, 42], it was found that initial results were not in good agreement with those seen in the previously mentioned off-lattice case, particularly for the qualitative behavior of the liquid-solid transition.

In addition, lattice models using only nearest-neighbor interactions have further highlighted the complexities of this transition, showing that behavior does not develop smoothly with chain length [18, 25]. One might think simulating with long range interactions in the continuum would cause this behavior to decrease or disappear; however, subtle complexi-
ties remain near the liquid-solid boundary. Recent studies using multicanonical simulations [17, 32] compared classical Lennard-Jones clusters and their own continuous homopolymer equivalent, highlighting the behavior of chains near the liquid-solid transition. A number of unique, chain length dependent, transformations can occur in this region, making the finite-size analysis of the liquid-solid transition much more complicated than previously thought. These finite-size effects have an excellent correlation with the behavior of classical clusters, e.g., the “magic number” set \( (N = 13, 55, 147, 309, 561, ...) \), whose minimum energy states all have icosahedral geometries. This magic number behavior, along with other highly ordered geometries associated with particular chain lengths, has also been shown to occur in simulations of homopolymers [17, 32, 37]. In cluster simulations, the global minimum energy structures play an important role in transition behavior near the liquid-solid transition, and with the above correlations with homopolymers, these types of effects must be considered.

2.4 Semi-flexible homopolymers

Semi-flexible chains represent a class of systems whose overall flexibility is regulated by a single energy term that is dependent on the angle between bond vectors. Such an energy term directly affects the persistence length of a chain, and in turn creates unique behavior for a range of different flexibilities. This behavior has been studied using a number of different models and simulation techniques [43–47, 24, 48–50]. In addition, semi-flexible systems are often motivated by biological applications; for example, DNA and RNA are considered stiff chains with some measurable persistence length. Hence, many computational, as well as theoretical, studies are aimed at unveiling general properties of these biological systems using semi-flexible models.

The goals of semi-flexible homopolymer simulations can be split into two categories: the description of conformational behavior as it depends on stiffness \( J_A \) and temperature \( T \); and how particular conformations and their transitions behave in the infinite-chain limit. The behavior of these systems in the infinite limit has been a topic of great interest, particularly
due to the fact that these simulations are motivated by biomolecules like DNA, which usually exist in nature at much larger lengths than are practical in simulations. This motivation has overshadowed much of the analysis needed to describe conformational behavior in small systems, which would ultimately lead to a more complete understanding of behavior in larger chains. In addition, while semi-flexible homopolymers have been well-studied historically, the application of generalized ensemble techniques is notably lacking.

Some of the most recent work has explored a broad range of flexibilities for a few chains in the bond-fluctuation model. In these studies, the flexibility is regulated by an overall strength parameter (denoted $J_A$ in this dissertation) which, when varied, leads to a wide variety of conformational and transition behaviors. The principal conformations found include solid-globular, rod-like, and toroidal states. Sub-categories of these states have also been defined, such as disk-like globules and racket configurations. The overall conformational behavior of semi-flexible chains is remarkably more complex than for the fully-flexible homopolymer case. It is noted that the packing considerations previously discussed for flexible homopolymers have yet to be applied to the case of semi-flexible homopolymers. With the current knowledge of flexible homopolymers as defined by cluster behavior, it is only natural to apply this understanding to the study of semi-flexible systems.

Another motivation for studying semi-flexible systems comes from current experimental work being performed on DNA [51, 52], which has shown regions of high flexibility. These studies show that the bending stiffness is lower than expected from theory, particularly comparing against the worm-like chain model [9, 10]. DNA molecules are often rather large, with length scales reaching the macroscopic world. However, the condensed state of such DNA chains are remarkably compact structures, often measured on the nanoscale. With such a large amount of material condensed into remarkably small conformations, it is not surprising that bending stiffness would vary from theoretical predictions. In terms of computer simulation, the precise analysis of fully-flexible behavior, and its unique connection to clusters, has provided a basis with which to begin studying semi-flexible chains. At present, the method-
ology used to understand maximally condensed flexible homopolymers has not been applied to semi-flexible chains. It is fully expected that Mackay and Anti-Mackay behavior will be relevant in the solid-globular regions of conformational space, and therefore this type of behavior could provide insight into the general packing properties of DNA.
3.1 SINGLE CHAIN HOMOPOLYMERS

In computer simulation, single homopolymer chains represent a class of systems aimed at understanding general phenomena in protein and polymer systems. Here we model homopolymer systems as a bead-spring like model [4], whose inter-monomer interaction potentials are more advanced than those of the simple harmonic type. In addition, the interactions are purely classical; no quantum mechanical or electronic interactions are taken into account. Classical interactions are able to predict various properties in single homopolymers, as much of the interesting behavior in nature occurs at temperatures and length scales beyond the quantum regime. The following sections will explain the Hamiltonian of our system, as well as the individual interactions governing the behavior in this system.

3.2 GENERAL HAMILTONIAN AND INTERACTION POTENTIALS

A chain of $N$ identical monomers is defined in continuous space, with each monomer having bonded, nonbonded, and bond-angle interactions. The Hamiltonian is given by

$$H = J_L \sum_{i=1}^{N-1} U^{Bond}(l_i) + J_{NB} \sum_{i<j} U^{NB}(r_{ij}) + J_A \sum_{i=1}^{N-2} (1 - \cos \theta_i).$$  

The first term represents the bonded energy; the parameter $J_L$ regulates bond strength and is set to 2.0, with the sum ranging over all $N - 1$ bonds, and $U^{Bond}$ is the bonded potential as a function of bond length $l_i$ (i.e., the distance between two adjacent monomers). The second term represents the nonbonded energy; $J_{NB}$ regulates the magnitude of this energy and is
Figure 3.1: Diagram of homopolymer interactions in an $N = 5$ length chain. For clarity, bonded (red) and nonbonded (blue) interactions are drawn for only a single monomer. The bond-angle interactions (green) are drawn for a single bond-angle pair, but are present between all other bond-angle pairs.

set to 1.0, the sum is over all nonbonded $ij$ pairs, and $r_{ij}$ is the distance between each pair of monomers. The third term represents the bond angle interactions, i.e., the stiffness of the chain, where $J_A$ represents the strength of this interaction and is always positive, the sum is over all $N - 2$ bond angles, and $\theta_i$ represents the angle between two linear bond vectors; the minimum of this term occurs at $\theta_0 = 0^\circ$ (This angle comes from the dot product of two vectors whose directions are linear along the chain; other studies [53] consider the bond angle to be the supplement of the angle using in this work, in which case $\theta_0$ would be $180^\circ$).

Nonbonded monomers interact via a truncated-shifted Lennard-Jones potential given by

$$U^{NB}(r_{ij}) = \begin{cases} U^{LJ}(r_{ij}) - U^{LJ}(r_c) & \text{for } 0 < r_{ij} \leq r_c, \\ 0 & \text{otherwise,} \end{cases}$$

(3.2)
where

\[ U_{LJ}(r_{ij}) = \epsilon \left[ \left( \frac{\sigma}{r_{ij}} \right)^{12} - 2 \left( \frac{\sigma}{r_{ij}} \right)^{6} \right]. \]  \hspace{1cm} (3.3)

\( r_{ij} \) is the distance between a pair \((i,j)\) of nonbonded monomers, \(\epsilon\) is the potential well depth, and \(\sigma\) is the distance at which \(U_{LJ}\) has its minimum. Dimensionless units are used, where \(\epsilon = 1\) and \(\sigma = 2^{-1/6}r_0\). The equilibrium distance is \(r_0 = 1\) and the interaction cutoff distance \(r_c\) is \(3\sigma\).

Interactions between two bonded monomers consist of a combination of FENE (finite extensible nonlinear elastic) \([54, 55]\) and Lennard-Jones potentials

\[ U_B(l_i) = \begin{cases} 
U_{FENE}(l_i) + U_{LJ}(l_i), & 0 < l_i \leq R_0, \\
0, & \text{otherwise},
\end{cases} \]  \hspace{1cm} (3.4)

where the Lennard-Jones potential \(U_{LJ}(l_i)\) is of the same form as described above and the FENE potential \(U_{FENE}(l_i)\) is given by

\[ U_{FENE}(l_i) = -0.5kR_o^2 \ln\left[1 - \left(l_i/R_o\right)^2\right]. \]  \hspace{1cm} (3.5)

\(R_0\) is the finite extensibility, \(k\) is the stiffness constant, and \(l_i\) is the bond length. Dimensionless units are again implemented, where \(R_o = 1.2\) and \(k = 2\), and this potential diverges as \(l_i\) approaches \(R_0\). Analyzing the derivative of \(U_B(l_i)\) with the above values allows this potential to be set such that its minimum is zero and occurs at an equilibrium value \(l_0 = 1\), coinciding with the equilibrium position of the nonbonded potential. Note, the parameter \(\sigma\) found in the \(U_{LJ}(l_i)\) term in Eq. 3.4 is used to adjust the distance at which the minimum occurs and it has a different value than that applied in Eq. 3.3.

All parameters for \(U^{NB}\) and \(U^B\) were chosen such that the minima of both potentials occur at a dimensionless distance of 1. For comparison, these potentials are plotted in Fig. 3.2. Other recent studies of continuous homopolymers \([56, 34, 35, 31, 32]\) have used an alternate implementation of \(U^B\) compared to that described above. In these studies, the nonbonded
Figure 3.2: Bonded and nonbonded potentials versus the distance between two monomers. A harmonic potential is plotted as a reference for the bonded potential (FENE-Lennard-Jones combination). The nonbonded potential is plotted to a distance of only 1.5 for clarity (the actual cutoff occurs at 3).

Lennard-Jones potential is applied to bonded monomers, meaning that the models contain \( N - 1 \) more nonbonded interactions compared to the model presented here. Results are comparable, as the extra interactions are simply superimposed on an already dominant bonded interaction.

As an anecdote, the general form of the Hamiltonian often used to model “real” classical protein systems contains five energy terms: 1) bonded interactions, 2) van der Waals interactions (nonbonded), 3) bond angle interactions, 4) torsion angle interactions, and 5) electrostatic interactions. When considering all-atom models, each of these interactions must
be tuned so that they take into account the appropriate values for specific interacting atoms. Most simplified models only consider the first three terms, as the last two terms require specific atomic information, such as the presence of side chains and the charge values associated with amino acids. The work reflected in this document contains terms similar to the first three mentioned above; as computational techniques and hardware improve, it is expected that future research efforts will incorporate the remaining two terms.

3.3 Perspective on interactions in fully flexible homopolymers

Fully flexible homopolymers are realized by considering only the first two terms in Eq. 3.1, and in some cases the terminology “elastic-flexible” homopolymers has been used, where “elastic” denotes bonds with variable length. The effects of different potentials for bonded interactions have not been well-studied, and there is no clear consensus regarding which potentials are best suited for coarse-grained simulations. This is primarily because a broad range of interaction parameters barely influence the dominant physical processes. However, some evidence (unpublished comparisons with [31]) suggests that for small chain lengths, low temperature transitions can be modified, where minor changes to the Hamiltonian have greater influences over behavior. What is perhaps more important is the dynamics within a given simulation, where elastic bonds tend to allow for a quicker sampling of low energy states in systems with identical nonbonded interactions. Inelastic (or stiff bond) models are typically less efficient in reaching low energy configurations.

Simulation studies have placed more emphasis on the types of nonbonded interactions and their parameters. Nonbonded interactions range from the Lennard-Jones type in Eq. 3.3 to square-well potentials. The selection of these potentials is usually correlated to the coarse graining of each model, where lattice models are subjected to discrete interactions [18, 25, 44], while continuous models are able to use a variety of continuous interactions [26, 31, 34, 57]. It is noted that square-well potentials have also been applied with success in continuous homopolymer simulations [58, 33, 59, 60]. In much of the recent work on these systems, the
distance cutoff for nonbonded interactions has played a key role. In the case of the bond-fluctuation model, the distance cutoff has been shown to directly affect transition behavior, where the liquid-solid and coil-globule transitions actually merge for some parameter sets [40, 39, 38]. In similar work on a continuous model, such effects have been incorporated in the modeling of two-state protein folding [58, 60]. In the case of an elastic-flexible homopolymer model similar to our own, calculating energies with a larger cutoff distance ($r_c = 5\sigma$) can affect ground-state conformations for a number of chain lengths [31, 32].

3.4 Thermodynamic behavior and transitions in fully flexible homopolymers

The general behavior of the homopolymer system described above, with zero stiffness, is given in Fig. 3.3. Assuming a generic temperature scale, one expects two distinct transitions when lowering (or raising) temperature, 1) a “coil-globule” transition from an unwound, random coil to a liquid-like globule, and 2) a “liquid-solid” transition from a liquid globule to a solid, or crystallized, globule. These two transitions are quite general, and are observed in a number of distinct homopolymer models. Of particular importance to this work is the existence of “solid-solid” transitions occurring at temperatures below those of the liquid-solid transition. Solid-solid transitions represent transformations, typically of monomers existing on the outer-shell of a given configuration. In most cases, homopolymer chains have a highly ordered geometry associated with the minimum energy state, and for some cases, the rearrangement of monomers on the outer shell can lead to transition signals, which will be further understood in the upcoming section on classical clusters. Solid-solid transitions are highly dependent on the number of monomers $N$, which is why the dashed line representing this transition in Fig. 3.3 is not connected to the temperature axis.

The coil-globule transition is perhaps the most well-studied of the three transitions listed above, where both analytical [61, 43], and experimental [62, 63], research has been performed. A number of predictions have been offered for how this transition behaves in the infinite
Figure 3.3: Conformational diagram of fully-flexible homopolymer chains. As temperature is reduced (right to left), the chain goes through a coil-globule transition, representing a transformation from an extended, random coil, to a globular state. As temperature is reduced further, a liquid-solid transition occurs between a liquid globule (middle) and a solid-globule (left). At even lower temperatures, a solid-solid transition can occur for some chain lengths, where these states are typically representative of highly ordered, minimized energy states, and rearrangements of exterior atoms are responsible for transition signals in thermodynamic properties.

limit, with subtle differences existing [25]. In this dissertation, the coil-globule transition is analyzed for a number of chain lengths, with interest in understanding its role in individual systems, as well as how this behavior scales with chain length. We have chosen the following relationship in order to predict the Θ-point in the infinite-chain limit

\[ T(N) - T_\Theta = a_1/\sqrt{N} + a_2/N \],  \hspace{1cm} (3.6)\]

where \( T(N) \) is the length-dependent transition temperature and \( T_\Theta \) is the Θ temperature, or the transition temperature in the infinite-chain limit [64]. A small set of other predictions have been used in simulations of homopolymer systems [25], but with these considered, we have found very little difference in results. A significant reason for selecting this scaling
relation is due to our comparisons with other recent homopolymer studies employing similar models and simulation techniques [40, 39, 35, 34].

The liquid-solid transition has not been as well-studied in homopolymer systems, compared with the coil-globule transition. However, it is a fairly standard transition that has been well-studied in the cluster community [65–67], as well as other research areas. Unfortunately, there is not a great deal of work in the literature regarding the infinite-chain behavior for homopolymer systems. Recently a scaling prediction was suggested [40, 39], and we have applied this in our own work, where

$$T(N) - T_{ls} = a/N^{1/3}$$

where $T(N)$ is the transition temperature for each chain length and $T_{ls}$ is the liquid-solid transition temperature for an infinite-chain. Infinite-chain estimates have been reported in homopolymer studies [39, 38, 35, 34] for the liquid-solid transition, but studies of chain length dependencies [25, 31, 32] have shown that more care must be taken when analyzing this transition. In particular, the chain lengths simulated in this study are riddled with finite-size effects, many of which make the prediction of an infinite-chain limit unreliable. This calculation is performed later in this dissertation, particularly as a means of comparing with other studies analyzing the liquid-solid transition. This type of analysis becomes even more important when considering the solid-solid transitions in conjunction with the liquid-solid transitions, which will be a large portion of Chapter 8.

Solid-solid transitions are seemingly unknown to a large part of the polymer community. These transitions are well-known in cluster communities, but have been slow to become a rigorous feature analyzed in homopolymer simulations. Solid-solid transitions are due to packing considerations on the exterior of a globular state. Typically, a core-shell structure develops, where the interior atoms form a stable core with a given geometry (e.g., icosahedral) and the remaining monomers arrange themselves on the surface of the core. Understanding the core formation can be described in terms of the behavior of classical clusters, and will be discussed in detail in the following section. The various packing arrangements of monomers
on the surface of this core can also be explained in terms of classical clusters. Initially, monomers simply pack in an energetically favorable manner on the surface of the core, and as temperature is reduced, the optimal packing according to the core is realized. Solid-solid transitions are highly dependent on chain length, and are not present in all lengths. The finite-size effects associated with these transitions make it nearly impossible to describe an infinite-chain transition temperature. Furthermore, it is also unclear how well these transitions will correlate with other models, particularly those not using Lennard-Jones potentials for their nonbonded interactions.

3.5 Connection to classical clusters

One of the most prominent features discussed in this dissertation is the relation of results for homopolymer simulations, with those found in the atomic cluster community. Lennard-Jones clusters have been studied in detail using a variety of techniques aimed at both thermodynamics and ground-state determination. In most cases, correlations between clusters and homopolymers can be considered directly, with a prominent feature being that the minimum energy conformations in both systems are often one-to-one. The reason for this connection has much to do with packing considerations of atoms (or monomers), and in many respects the choice of a Lennard-Jones interaction potential. Being that there are many connections, one might think much of the work has been done, and is understood for these types of systems. However, phases exist for homopolymers that are not seen in cluster systems, such as random-coil states, and modern studies providing new insights for cluster systems have recently been published for sizes $N = 32$ [68] and 309 [69], implying some features of these systems are still not fully understood.

Atomic cluster studies are typically interested in understanding the minimum energy states associated with the solid region of conformational space. These states usually represent conformations with specific geometries, which are highly dependent on the number of atoms present. Icosahedral geometries are particularly important in studies using Lennard-Jones
Figure 3.4: Depiction of Mackay and Anti-Mackay packing, along with three Mackay states, an $N = 13$ icosahedral core (left), an $N = 40$ incomplete icosahedron with Mackay overlayer (middle), and Mackay’s $N = 55$ icosahedron.
this \( N = 40 \) length chain represents an example of a system that goes through a solid-solid
transition in order to reach the final Mackay state. The term Mackay packing comes from
the original analytical work of Mackay on the packing of atoms in icosahedral geometries,
with the \( N^* = 55 \) case now known as the Mackay icosahedron [71]. There are a number of
system sizes that do not follow this general packing behavior, but these tend to be special
cases that are well-understood using standard geometric formalisms. In general, any state
that represents an incomplete icosahedral geometry is termed a Mackay state, and any other
state containing one or more icosahedral cores, but a non-icosahedral overlayer, is termed as
Anti-Mackay.

Schnabel et al. have made a vigorous analysis of the connection between flexible
homopolymers and clusters [32, 31], showcasing the direct correlations, as well as the
subtle differences, for particular chain lengths. To our knowledge, this represents the wave
front of a future community response to understand the crystallization behavior in a broad
class of homopolymers. Until this point, studies have been more interested in understanding
behavior in the infinite-chain limit. Behavior of these collapsed states, along with their
complex finite-size effects, has given researchers a broad class of new problems to study,
which typically occur in system sizes that are completely accessible to modern computing
techniques. In addition, the introduction of stiffness terms into the Hamiltonian creates
completely new set of conformations that cannot necessarily be described by cluster models.

3.6 Perspective on interactions in semiflexible homopolymers

Semiflexible systems differ from the above mentioned flexible case by a stiffness term intro-
duced to the Hamiltonian, which in turn regulates the persistence length (or stiffness) of the
chain. A number of variations exist for this energy term, but are usually all of the harmonic
type, differing only in implementation. The potential used in this work is modeled after
that used by Vogel et al. [72], but could have easily been taken from other studies on this
subject. The implementation applied in this dissertation has a connection to models used
in theoretical studies like the worm-like chain [10], where the bond angle is defined as the angle between two consecutive bond vectors. It is noted that the bond angle is defined as zero for a perfectly straight chain, where some studies define the bond angle as 180 degrees for a straight chain. Either measurement is readily available using a simple transformation.

In many studies, a stiffness is often set to mimic a particular persistence length of a given polymer, focusing on specific behavior as it is related to some relevant biomolecule. Other work varies the stiffness in order to describe a complete phase diagram as a function of the interaction strength. Such a phase diagram is rich with conformations such as globules, rods, disks, and toroids. The ordering into more macroscopically understandable geometries is caused by the bond angle interaction’s tendency to cause “looped” structures. This is drastically different compared to their flexible counterparts, which often have monomers packed into unique, higher-order geometric shapes, such as icosahedra, tetrahedra, or dodecahedra. The existence of so many unique conformations implies that a number of independent transformations occur for a given chain length.

3.7 Thermodynamic behavior and transitions in semi-flexible homopolymers

As mentioned previously, a number of simulation studies have focused on the overall conformational behavior of semi-flexible systems. These simulations have been performed at various levels of coarse graining, using lattice, bond-fluctuation, and continuous models. The majority of these studies focus on equilibrium configurations for a broad range of temperatures $T$ and stiffness values $J_A$. The body of work produced by these studies has found three principal stable configurations, namely, solid-globular, rod-like, and toroidal states. Sub-categories of these states have also been defined, such as the disk-like globule, and racket configurations, but it is noted that these sub-categories often fit quite well-into the broader categorization of principal configurations.

Fig. 3.5 provides a diagram of the expected conformational behavior of semi-flexible homopolymers as they depend on $T$ and $J_A$. This depiction relies heavily on a number of
Figure 3.5: Diagram of conformational behavior as it is related to stiffness (flexibility) and temperature in semi-flexible homopolymers. Three classes of minimum energy conformations occur for low temperatures: 1) solid-globular, 2) rod-like, and 3) toroidal states. In the case of low stiffness, behavior is similar to flexible homopolymers, where liquid-solid and coil-globule transitions occur as temperature is increased. However, a dividing line exists between solid-globules and rod-like states such that the liquid-solid and coil-globule transitions are suppressed with increasing stiffness. Scales for temperature and stiffness are generic, representing rough estimates of these quantities.
works simulating the bond-fluctuation model [45, 46, 24], but is also taken from work on other models [44, 47, 50]. In addition, certain aspects of what is known for fully flexible homopolymers are included. Starting from the low temperature, low stiffness region, a solid-globular region exists. This region exhibits similar behavior to the case of the fully-flexible homopolymers, where an increase in temperature leads to first a liquid-solid transition, and a further increase to a coil-globule transition. For semi-flexible homopolymers, it is expected that higher order geometries similar to the flexible case exist in the solid-globular region, and it is noted that these features have not been studied with the same rigor as in the flexible case. However, much work needs to be done in classifying such geometries within this region.

The liquid region in Fig. 3.5 still holds many of the same properties found in the flexible case, but is expected to disappear as stiffness increases, giving way to abrupt transformations from random coils and straight chains, to more ordered states such as rods-like and toroidal configurations. The nature of these transformations has been studied and are typically of the rounded-first-order type. The first-order classification also reveals that for flexibilities producing rod-like and toroidal structures, a second-order coil-globule transformation no longer exists. Rather, these transitions become of the order to disorder type.

The nature of the transitions discussed above have been studied in a variety of works. Some pressing issues involve how these transitions and their subsequent conformational regions behave in the infinite limit. Unfortunately, in order to study these effects, large systems must be simulated. One recent study has gone as high as \( N = 50000 \) for a simplified model on the lattice [50]. Currently, continuous models of the type described in this chapter have no real chance of reaching these lengths, and therefore, no comments are made regarding the nature of the transitions described in Fig. 3.5. It is not to say that such lengths cannot be reached in the future, as certain phenomena could certainly be simulated for chains in the range of \( N = 1000 \). However, in light of the small systems studied in this dissertation, no discussion is given regarding the behavior of transitions in the infinite limit.
In addition to the above discussion, the solid-globular region presents an opportunity to apply knowledge from cluster behavior to semi-flexible homopolymers. This dissertation will provide evidence of higher order geometries existing in the solid-region of the semi-flexible conformation diagram given in Fig 3.5. In particular, the Mackay / Anti-Mackay symmetry described for fully flexible homopolymers (along with other geometries) can exist for a significant range of stiffness values within the solid-globular region. How these geometries vary with parameters such as flexibility and chain length are of principal interest. To the knowledge of this author, studies aimed at understanding the crystalline behavior of semi-flexible homopolymers in terms of Mackay and Anti-Mackay geometries have not been performed systematically. The reason for the lack of attention has been two-fold: the first being the inability to effectively simulate in the low $T$ compact globular region, and the second is that many of the most interesting questions remaining for researchers studying semi-flexible systems pertain to large chains, with the intent of approaching the infinite limit. With the newly developed methods of studying flexible homopolymers in place, it is only natural that such investigations take place for semi-flexible systems.
Chapter 4

Monte Carlo techniques

4.1 Overview of Monte Carlo techniques

Monte Carlo techniques involve processes driven by random numbers [12, 13]. In statistical physics applications, their function can be generalized to the idea of randomly creating as many statistically independent configurations as possible for a given physical model. This allows one to average over these configurations for a given ensemble of states, and in turn, calculate various properties of a given system. The generation of new configurations, which is often a model dependent property, is governed by the master equation

\[
\frac{dP_n(t)}{dt} = - \sum_{n \neq m} \left[ \omega_{n \rightarrow m} P_n(t) - \omega_{m \rightarrow n} P_m(t) \right] \tag{4.1}
\]

where \(P_n(t)\) and \(P_m(t)\) represent the probabilities of being in states \(n\) and \(m\) at some time \(t\), and \(\omega_{n \rightarrow m}\) and \(\omega_{m \rightarrow n}\) are the transition probabilities between these states. For equilibrium processes \(dP_n(t)/dt = 0\), leading to

\[
\omega_{n \rightarrow m} P_n(t) = \omega_{m \rightarrow n} P_m(t) \tag{4.2}
\]

which is the detailed balance condition and must be satisfied in all equilibrium Monte Carlo techniques. The time \(t\) is not “physical” time, but represents a “simulation” time which is often set to unity, playing a subtle role in the dynamics of a simulation. Hence, further discussions will concern generic probabilities \(P_n\) and \(P_m\), and later these probabilities will be represented as a function of some physical observable.

A variety of Monte Carlo techniques exist for sampling physical systems, and these techniques differ in their implementation of Eq. 4.2. However, the majority follow a general
process when simulating a given system. This process involves measuring a physical quantity \( A \) of some system, proposing a change to the system, and then measuring the change to quantity \( A \). A key component of any Monte Carlo simulation is how changes (or updates) are made to the system. The transition probability \( \omega_{n \rightarrow m} \) represents the probability that a system currently in state \( n \) will change to state \( m \), and these rates are often determined by the measure of some quantity such as \( A \). The goal of these techniques is to generate as many independent configurations as possible within a given ensemble (or distribution) of states, and the design of efficient updates (or moves) is a crucial component of any simulation. Updates vary as much as the physical system, and \( \omega_{n \rightarrow m} \) must be considered when devising these updates. This quantity can be expressed in more detail as

\[
\omega_{n \rightarrow m} = W^p_{n \rightarrow m} \cdot W^a_{n \rightarrow m},
\]

where \( W^p_{n \rightarrow m} \) is the probability for proposing an update and \( W^a_{n \rightarrow m} \) is the probability of accepting this update. These two components of the overall transition probability rely heavily on the type of change being proposed, and will be considered in more detail when specifics are discussed regarding the model used in this work. Understanding the transition probability in terms of these components changes Eq. 4.2 to

\[
\frac{W^p_{n \rightarrow m} \cdot W^a_{n \rightarrow m}}{W^p_{m \rightarrow n} \cdot W^a_{m \rightarrow n}} = \frac{P_m}{P_n},
\]

which will play an important role when discussing the “move-set” used in this work. In many cases, symmetric proposal probabilities exist \( W^p_{n \rightarrow m} = W^p_{m \rightarrow n} \), simplifying Eq. 4.4 into an expression based primarily on acceptance rates

\[
\frac{W^a_{n \rightarrow m}}{W^a_{m \rightarrow n}} = \frac{P_m}{P_n}.
\]

This format allows for a more direct analysis of the detailed balance equation (Eq. 4.5), for which a solution is given by

\[
W^a_{n \rightarrow m} = \min \left( \frac{P_m}{P_n}, 1 \right).
\]
This solution represents the rate at which changes are accepted to a system for a proposed update or move, and again, depends highly on the type of algorithmic implementation of a given Monte Carlo technique.

Of equal importance are the moves when $W_{n \rightarrow m}^p \neq W_{m \rightarrow n}^p$. These types of moves are often implemented in order to improve efficiency in a simulation, and detailed balance must be taken care of in this implementation. The proposal probabilities for a particular move can be included in the general solution of detailed balance

$$W_{n \rightarrow m}^a = \min \left( \frac{P_m \cdot W_{m \rightarrow n}^p}{P_n \cdot W_{n \rightarrow m}^p}, 1 \right). \quad (4.7)$$

Now one is left to understand these proposal probabilities and simply include them in the Monte Carlo method. These types of probabilities will be better understood when considering various move sets used in this work.

The above text represents necessary dialogue for applying such techniques to the simulation of physical systems, and this dialogue is primarily meant to help facilitate further, more detailed, discussion of applications in this work.

4.2 THE PARTITION FUNCTION IN MONTE CARLO METHODS

Monte Carlo methods in statistical physics are primarily used to estimate thermodynamic properties of a given system. In almost all cases, these properties can be related to the partition function $Z$, whose canonical form is written as

$$Z(T) = \sum_i e^{-E_i/k_BT} \equiv \sum_E g(E) e^{-E/k_BT}, \quad (4.8)$$

where the first term represents a sum over all possible configurations $i$ with individual energies $E_i$, which are weighted by the Boltzmann factor $e^{-E_i/k_BT}$, where $k_B$ is Boltzmann’s constant, and $T$ is the canonical temperature. The second term is an equivalent representation, only the sum over all configuration is replaced by the sum over all energy state $E$ weighted by the density of states $g(E)$, and again the Boltzmann factor. The density of states
Figure 4.1: Traditional versus generalized ensemble simulations. Density of states for $N = 50$ length chain (black) and two probability distributions $P(E,T)$ for $T = 0.1$ (red) and $T = 1.0$ (green). Each curve has been divided by its maximum, so that all curves are easily visible, and only part of the data for $g(E)$ is plotted. This is meant to be a generic example, hence, no error bars are drawn.

is simply a representation of the number of states with a particular energy. The partition function is perhaps the most important quantity in statistical mechanics in that it allows access to all thermodynamic properties of a physical system.

These two expressions also represent a rough analogy between traditional and generalized ensemble simulation methods. The first term relates to a more traditional formalism, where $T$ plays a crucial role in the dynamics of a system (which is why at low $T$ and near phase transitions, these methods can have problems sampling effectively). The second term can be related to a class of algorithms that no longer sample with $T$, but are only concerned with producing a random walk in $E$ in order to generate $g(E)$, hence circumventing many issues associated with traditional simulation methods.
This section is not meant to be a refresher on statistical mechanics, but is rather meant to aid the reader in understanding certain formalisms of Monte Carlo techniques applied to statistical physics. In the remainder of this chapter, two such methods are discussed: Metropolis-importance and Wang-Landau algorithms. Before introducing details of these techniques, a visual representation is offered in Fig. 4.1, where two probability distributions $P(E, T)$ (whose full definition will be given shortly) for a fully-flexible $N = 50$ length chain are plotted for $T = 0.1$ and $1.0$ ($k_B$ is set to unity), along with the density of states for the same system, all normalized by their maxima for clarity. The individual $P(E, T)$ are the distributions within which the Metropolis method would be confined to sample, as $T$ has a direct role in controlling the dynamics of the simulation (where the width of the $T = 0.1$ distribution gives an idea of why sampling can be constrained at low $T$ or for complicated distributions). The density of states $g(E)$ (black curve) was produced by the Wang-Landau algorithm, in which the simulation’s goal is to sample all energies, and provide a function that can reproduce thermodynamics at all temperatures. This comparison and visualization are meant only to motivate the following sections, where each method is discussed in detail. It is noted that although generalized ensemble techniques are often much more powerful than traditional techniques, the efficacy of these approaches is often more dependent on the problem to which they are being applied.

4.3 Metropolis sampling

As an example of Monte Carlo sampling in Statistical Mechanics [12, 21], the Metropolis method [14] is introduced, which is the first Monte Carlo technique applied in physics. This importance sampling technique generates Markov chains of states through Monte Carlo updates satisfying Eq. 4.2. Consider some classical physical system at temperature $T$. From statistical mechanics, the probability of this system being in state $n$ with energy $E_n$ is written as

$$P(E_n, T) = e^{-E_n/k_B T} / Z ,$$

(4.9)
where $k_B$ is Boltzmann’s constant, and $Z$ is the aforementioned partition function. In most cases, a proper estimate of $Z$ is not available, and subsequently, $P_n$ is unknown. The Markov process satisfies this issue since each newly generated state relies on the previous state, meaning probabilities $P_n$ and $P_m$ have a common, canceling factor in $Z$. When applying these probabilities to Eq. 4.6, the acceptance rate is

$$W_{n\rightarrow m}^a = \min\left(e^{-\Delta E/k_B T}, 1\right),$$

where $\Delta E = E_m - E_n$. In the above equation, the min function bases its selection on $\Delta E$, with particular importance on whether energy is reduced or enlarged. In the standard Metropolis approach, if $E_m < E_n$, the probability of accepting the change is 1, meaning changes lowering the system’s energy are always selected. This implies that when possible (based on the types of updates applied), equilibrium is reached at a faster rate than would be sampled by a simple random process. In the case that $E_m > E_n$, the exponential factor is selected, and its value compared with a random number $r \in (0, 1)$. If $r < e^{-\Delta E/k_B T}$, the move is accepted. Since many systems provide difficulties in sampling equilibrium states, the number of moves raising energy often outnumber those that reduce energy.

The above discussions have neglected the types of updates used to change a given system, and how their design may affect the acceptance probability. These will be discussed in detail regarding our homopolymer model in the next chapter.

4.4 ESTIMATES OF THERMODYNAMIC BEHAVIOR FROM METROPOLIS

The Metropolis algorithm proceeds as a Markov process, where new states are generated and accepted according to the probabilities given in the above section. In order to keep track of various properties of a system simulated with Metropolis, one must periodically measure properties of the system, and keep track of their average value through

$$\langle A \rangle = \frac{1}{S} \sum_i^S A(x_i),$$

(4.11)
where $A$ is some thermodynamic property, $x$ are the instantaneous coordinates of the system, $i$ represents this instantaneous Monte Carlo time, and $S$ is the total number of measurements. There are any number of properties that can be measured in this way, such as the average energy, specific heat (from fluctuations in the average energy), and structural quantities that rely directly on $x_i$. This approach is quite straightforward to implement, but is faced with subtle issues regarding the accuracy of the results. For example, in many polymer systems, the $x_i$ between subsequent states are quite correlated. In this case, gaps between measurements should be placed to remove correlations that would otherwise provide inaccurate results. This issue can be realized in the initial stages of a simulation, where a model often has some observable relaxation time between the initial and equilibrium states. Furthermore, often in temperature dependent simulations, models are trapped in metastable states near the equilibrium state (particularly at low temperatures), so this simple averaging of properties may lead to statistically accurate results, being physically incorrect due to systematic error. These types of issues (as well as others) are not unknown, and in fact have been studied [12] since the introduction of the Metropolis method.

4.5 **Generalized ensemble methods**

Many physical systems are characterized by a complex energy landscape, riddled with local minima and free-energy barriers. These minima and barriers can represent the difference between two separate phases, as in the case of a first-order transition, or can represent simple conformational transformations between two mutually similar states, often leading to small thermodynamic signals. In more complicated systems, such as proteins and polymers, these types of phenomena occur at a variety of temperatures and physical parameters. Traditional Monte Carlo methods are often unable to fully describe behavior in these systems, as their ability to move between local minima is stalled by the nature of their algorithms. This can be seen by considering the Metropolis algorithm, whose acceptance probability is based on the Boltzmann factor and change in energy $e^{-\Delta E/k_BT}$. This implies that at very low
$T$, the likelihood of making changes that approach equilibrium configurations (minimum energy) drops off exponentially, meaning simulations often are stuck in metastable states corresponding to minima outside of equilibrium. This can be further realized by considering the probability distribution $P(E,T)$ in Eq. 4.9, whose width decreases as $T$ decreases.

As stated in Chapter 2, generalized ensemble techniques [21–23] have been introduced to overcome many of the inefficiencies associated with traditional Monte Carlo methods. These techniques are not constrained to a single “Boltzmann” distribution, allowing them to avoid the inefficiencies associated with Boltzmann-weighted simulations. There are a number of useful generalized ensemble techniques, but this work focuses on the subset whose primary goal is the estimate of the density of states [73, 74], in particular, the Wang-Landau algorithm [75, 76].

Generalized ensemble techniques involving the density of states $g(E)$ for a system with energy $E$ find a connection to statistical mechanics through the partition function $Z$, which is written as a function of $g(E)$ in Eq. 4.8. This representation allows for the estimation of $g(E)$ to be the sole focus of the simulation. The Wang-Landau algorithm [75, 76, 73], as well as multicanonical sampling [77, 74], is based on the observation that a random walk in energy, with an acceptance probability proportional to the inverse density of states, will produce a flat histogram of energy visits [76]. This feature means that flat-histogram methods more freely visit local minima and are better able to bypass energy barriers compared to temperature dependent methods. The acceptance rate for these types of simulations can be written as

$$W_{1 \rightarrow 2}^a = \min \left( \frac{g(E_1)}{g(E_2)}, 1 \right),$$

representing a class of algorithms whose dynamics rely directly on $g(E)$ \footnote{Multicanonical sampling uses a weight function which is inversely proportional to $g(E)$.}. The acceptance probability proceeds by comparing two density of states values, and always accepting moves when $g(E_2) \leq g(E_1)$. This insures that the probability of moving to “under-sampled” (or low
probability) states is 1, where typically these states correspond to low energy (or low temperature) regions of \( g(E) \), which are typically too difficult to sample using purely canonical approaches. When \( g(E_2) > g(E_1) \), the ratio of these values is compared to a random number \( r \), and the transition (or move) is accepted if \( r < g(E_1)/g(E_2) \). The above text describes the basic features of a density of states based, generalized ensemble technique. However, further algorithmic considerations needed for these methods will be discussed in the context of Wang-Landau simulations in the following chapters.

The above procedure offers the general context for performing a generalized ensemble simulation aimed at estimating the density of states. There are a number of algorithmic constraints that are necessary, and will be discussed in the following chapters. However, the general idea is that these simulations are no longer restricted to the calculation of single quantities for a single preselected temperature, but are instead able to generate a distribution from which the partition function can be estimated, and subsequently, all thermodynamic quantities.

### 4.6 Thermodynamic Calculations from Generalized Ensemble Techniques

Once an estimate of the density of states is available, calculating thermodynamic properties is straightforward, where again, quantities of importance in this dissertation can all be related to \( Z(T) \). We start by considering the average energy as a function of temperature

\[
\langle E \rangle_T = \sum_E E \ g(E) e^{-E/k_B T} / Z(T),
\]

where \( T \) is the canonical temperature, \( e^{-E/k_B T} \) is the Boltzmann factor, and \( E \) represents the overall energy of a given system. Average energy as a function of \( T \) gives indications of changes in the system, and can be even more useful when particular terms in the Hamiltonian Eq. 3.1 are isolated, e.g., \( \langle E_{\text{Bond}} \rangle \). However, of more importance in identifying transitions in these systems is the specific heat \( C_V(T)/N \), defined as

\[
\frac{C_V(T)}{N} = \frac{\langle E^2 \rangle_T - \langle E \rangle_T^2}{NT^2},
\]

(4.14)
where \( \langle E^2 \rangle_T \) can be calculated in the same manner as in Eq. 4.13. Since this quantity can be considered a measure of the fluctuation in energy, transition signals are much stronger and more easily identifiable than in energy alone. All of the above quantities depend on \( g(E) \) and can be calculated for any temperature. There are subtleties related to the validity of these calculations, and these will be discussed throughout this dissertation.

The \( \langle E(T) \rangle \) and \( C_V \) are the primary thermodynamic quantities calculated in all simulations. However, other meaningful quantities are readily available. The free energy \( F(T) \) can be calculated by \( F(T) = -k_B T \ln Z \), and the entropy can be calculated as \( S(T) = (\langle E \rangle_T - F(T))/T \). The behavior we are currently interested in does not require analysis of the free energy or entropy, therefore, discussion will be limited. It is also mentioned that an offset is typically used in all of these calculations in order to avoid overflow and underflow issues caused by the orders of magnitude differences occurring in some \( g(E) \). This factor, usually written as \( \lambda \), is discussed in detail in one of the early publications on Wang-Landau sampling [73].

Structural quantities are another important measure needed to understand the behavior of polymer systems and are typically estimated through a production run procedure once an estimate of \( g(E) \) is available. Details regarding production runs can be found in Chapter 5. In order to calculate the canonical form of a structural quantity \( Q \), their average values are estimated as a function of energy during each production run. These estimates are obtained by storing numerical values of a structural quantity for many individual configurations found within the same energy bin. Various structural quantities \( \overline{Q}(E) \) can be estimated by keeping track of individual histograms as a function of energy, \( H(E) \), leading to

\[
\overline{Q}(E) = \frac{1}{H(E)} \sum_E Q(E, X),
\]

where, \( X \) represents the coordinates for a specific configuration. These averages can then be used to calculate the canonical form using

\[
\langle Q \rangle_T = \frac{1}{Z(T)} \sum_E g(E)\overline{Q}(E)e^{-E/k_BT}
\]

(4.16)
which is not so different to the standard approach of calculating the average energy. This approach has been used in other studies as well [40, 39]. A number of structural quantities are employed in this work, such as the radius of gyration, end-to-end distance, core density, and several others. Quantities will be defined individually along side their results.

It is noted that error estimation issues associated with these thermodynamic and structural calculations are discussed in an upcoming chapter.
Chapter 5

The Wang-Landau algorithm

5.1 Wang-Landau algorithm

The Wang-Landau algorithm was introduced in 2001 as a new approach to the “flat-histogram” (or extended ensemble) class of sampling techniques [75, 76, 73]. It provides an iterative procedure for estimating the density of states $g(E)$, whose sampling criterion is the inverse of the current estimate of $g(E)$, where $E$ is the continuous or discrete energy of some physical system. The power of this method comes from its ability to efficiently sample physical quantities over a broad range of values, and what is perhaps more important, is that its ease of use makes it more accessible than similar techniques. The algorithm’s effectiveness can be measured by the number of articles employing this method in the last decade (highlights of these articles [78–82, 18, 83]).

Consider a generic physical system with energy values $E_i$ and corresponding density of states values $g(E_i)$, in which all values of this function are initially unknown. Monte Carlo updates or moves are proposed to the system and accepted through

$$W_{1 \rightarrow 2}^a = \min \left( \frac{g(E_1)}{g(E_2)}, 1 \right).$$  \hspace{1cm} (5.1)

which shows that the probability of making a move is based on the inverse of current estimate of the density of states, which represents a solution to the detailed-balance condition in Eq. 4.2. The most convenient part of this formalism is that the Boltzmann factor is not included in this solution, meaning the technique is a temperature independent method, and

\[\text{It is noted that the detailed balance condition is not satisfied in the early portions of the simulation, and it is not until the late stages that this condition is satisfied, which will be discussed in further detail in a following section.}\]
is therefore not subjected to many of the inefficiencies related to some temperature-based sampling methods.

The algorithm proceeds as follows. First, all values of $g(E_i)$ are set to 1.0, a generic initial guess that allows us to begin the random walk through energy space. Next, an MC move is used to change the system from initial energy $E_1$ to a new energy $E_2$, and this move accepted or rejected according to Eq. 5.1. The density of states is then updated by a modification factor $f$ (which will be discussed shortly) and a histogram of energy values $H(E_i)$ is also updated to keep track of what energies have been visited. These function are updated by

$$g(E_i) = g'(E_i) * f$$
$$H(E_i) = H'(E_i) + 1$$

(5.2)

where if the move is accepted, these functions are updated at $E_2$, and if they are rejected, these functions are updated at $E_1$. The initial value of $f_0$ is typically $f_0 = e^1$. In order to estimate $g(E_i)$ correctly, $f$ must be reduced to a small value ($f \approx 1.0$), and the histogram is used to regulate this reduction by ensuring that all energies have been sampled appropriately ($f$ and $H(E_i)$ will be discussed in detail). Once $g(E_i)$ and $H(E_i)$ have been updated for a single MC move, the process is repeated for a fixed value of $f$. The number of repetitions is generally related to the system in question, e.g., total number of $E_i$. After this fixed value of moves, the histogram is checked and if appropriate sampling has taken place, all $H(E_i)$ are reset to zero, and $f$ is reduced. If the histogram lacks appropriate sampling, the process continues with the same histogram and value of $f$. This process is repeated until $f$ is sufficiently small ($f \approx 1.0$), meaning that large changes are no longer being made to $g(E_i)$ during the random walk, and the estimate is close to the actual value.

Another important point associated with this procedure is that typically $g(E_i)$ is stored in terms of the natural logarithm $\ln[g(E_i)]$ in order to deal with functions that can easily contain many orders of magnitude (in some cases thousands) differences between highest and lowest $E_i$ values. The sampling procedure described above remains the same, only the density of states is updated by $\ln[g(E_i)] = \ln[g(E_i)] + \ln[f]$. Although the values are stored
in terms of the natural logarithm, the ratio in Eq. 5.1 must be in terms of the ‘real’ values. In addition, it is noted that the density of states is known only as a relative quantity, i.e., the estimate provided by the Wang-Landau algorithm is actually \( g(E) + C \), where \( C \) is a constant offset. In calculations of secondary properties such as specific heat, this constant offset cancels and becomes irrelevant.

5.2 Modification factor and histogram

The modification factor \( f \) and the procedure for how it is reduced plays an important role in the dynamics of the random walk in energy space. The original, and still widely used, approach to reducing \( f \) is given by the following

\[
f_{i+1} = \sqrt{f_i},
\]  

(5.3)

which in terms of the natural logarithm is equivalent to \( \ln[f_{i+1}] = \ln[f_i]/2 \). The initial value of the modification factor \( f_0 \) is typically set to \( e^1 \) (or \( \ln[f_0] = 1.0 \)) and reduced to a value \( \approx 1.0 \) (or typically \( \ln[f] = 10^{-8} \)). These parameters can be different for various physical systems, and the values applied in simulations of homopolymer systems will be discussed in detail.

The histogram \( H(E) \) is used to regulate the reduction of \( f \). Traditionally, the flatness \( p \) of the histogram was used to determine the appropriate level of sampling, where \( p \) is the minimum value of the histogram divided by its overall average. This insures that all energy states are sampled sufficiently before \( f \) is reduced to values that would no longer be able to modify \( g(E) \). Other procedures have been introduced for determining the rate at which \( f \) is reduced, such as simply checking to make sure that each \( H(E_i) \) has been sampled a set number of times, without regard to overall flatness [84, 85]. Another method involves the round-trip time, which is a measure of the number of moves needed to move between the high and low energy boundaries of a given system [78]. One of the more promising modifications to the original framework uses the MC time (number of moves) to reduce \( f \), and is hence
referred to as the $1/t$ method [86–88]. This procedure has worked well for simple systems, but is still being proven as a method applicable to proteins and polymers [89].

Deciding which procedure to apply in the reduction of $f$ is entirely dependent on the system being simulated, as well as the desired information. In this work, the traditional flatness approach has been used in simulations over energy (1D) of systems up to $N = 561$, while in multi-dimensional simulations (2D), the simple check of sampling each bin a certain number of times is used. In the 2D simulations, a flatness approach becomes inefficient due to both the size and complexity of the 2D density of states. These types of issues will be discussed in detail in the results section of this dissertation. Attempts to implement the $1/t$ method have also been made, but preliminary results on its efficiency were unsatisfactory\(^2\), and were therefore not pursued.

5.3 Detailed balance and convergence

It is noted that the procedure described above does not satisfy the detailed balance condition (Eq. 4.2) until the end of the simulation. Since the probability of being in a given state is proportional to $1/g(E)$, and $g(E)$ is subject to change while $f$ is large, these values are not stable enough to precisely satisfy this condition. However, as $f$ is reduced, and the estimate of $g(E)$ approaches the true value, the system fulfills Eq. 4.2, which has been proven in various studies [90–92]. A more important topic deals with the convergence of the estimate of $g(E)$ towards the real value. In many cases, saturation of error can occur leading to systematic errors in estimations of physical quantities, or to unnecessarily large simulation times, where corrections to $g(E)$ are no longer being made. This saturation of error has been studied, but in many cases this systematic error is still much smaller than the statistical error. In this regard, it is often appropriate to produce a large set of individual runs, with the intent of averaging these results. If the set is large enough, one can often cease the reduction of $f$ at

\(^2\)The inefficiencies found with the $1/t$ method are most likely due to the implementation in a continuous model. The original works for this method were all implemented in discrete systems.
larger values ($10^{-6}$ or $10^{-7}$) than recommended by the original work ($10^{-8}$). Certain methods have been proposed to address the saturation of error issue, but have not shown sufficient generality to any physical system [86–88]. In addition, it is noted that these issues are still of interest to researchers, but regardless, the convergence of the Wang-Landau algorithm has been shown [90–92].

5.4 Production run for structural quantities

Once an estimate of $g(E)$ is available, it is sometimes desirable to estimate quantities not directly available from this function of the energy $E$. Production runs are one such way of estimating such quantities, e.g., the radius of gyration $R_g^2 = \sum_i N (\vec{r}_i - \vec{r}_{cm})^2 / N$, which is a measure of the overall size of a polymer chain. The production run is performed after a final estimation of the density of states $g_f(E)$ has been provided by a single Wang-Landau simulation. This $g_f(E)$ is then used to perform a random walk in $E$, while simultaneously keeping track of various quantities as functions of $E$, such as $R_g^2(E)$. Once all energy bins have been sampled sufficiently, an average value of each quantity is available, e.g., $\overline{R_g^2}(E)$. This estimation in many cases can provide insight into general behavior, but more often, the canonical form $\langle R_g^2 \rangle_T$ is desired. This can be realized through the formalism presented in Eq. 4.16. Therefore, parameters associated with the production run must be tuned such that the energy dependent estimate of various quantities is as accurate as possible.

In this work, the production run has been implemented in two separate ways, both of which are discussed below. The first method simply uses the Wang-Landau algorithm, generating a random walk in energy space over which structural averages are stored as a function of energy. An issue that arises during this procedure involves the decision of what value of the modification factor should be used. This work has experimented with using different value and comparing the results. The intuitive approach is to simply reduce $f$ one iteration past the final value used to provide the estimate of $g(E)$, which would allow for an even more precise $g(E)$. However, these final values typically take the longest of all iterations
and in some cases the number of MC trial moves needed for the production run equals the number needed to generate the final $g(E)$! Tests were made using a larger $f \approx 10^{-3}$, which in terms of visiting all energies, usually represents the most dynamic region of the simulation. Considerations of the balance between CPU time and improving the estimate of $g(E)$ were taken into account when applying these methods, and in most cases, CPU time outweighed the need for a better estimate of $g(E)$. Very little difference was found between these two approaches in the production run.

The second method used to produce structural data is a multicanonical production run. Multicanonical sampling is a similar approach to the Wang-Landau algorithm, where details can be found elsewhere [77, 74]. The benefits of multicanonical sampling is that this method avoids issues involving saturation of error (which will be discussed in detail in a subsequent chapter). The important message here is that if the current estimate of $g(E)$ is reasonably accurate, this method will reduce error with some proportionality of the simulation time. In some studies [18, 17, 57], this has been the method of choice for production runs, as it is somewhat “guaranteed” to reduce error in $g(E)$ (although some technical details can cause this method to have sampling problems). This work has experimented with this style of production run, choosing methods that are best suited for the desired results.

5.5 Wang-Landau parameters in homopolymer simulations

The exact details of how the Wang-Landau algorithm is implemented can vary according to the desired results of a simulation. The studies represented in these documents fall into two categories: 1D sampling aimed at simulating the liquid-solid and coil-globule transitions for very large chains (up to $N = 561$) and 2D sampling aimed at uncovering conformation behavior as a function of stiffness ($J_A$). These categories have slight variations in their common parameters and will be discussed in detail in the corresponding results sections.

However, certain features regarding multi-dimensional Wang-Landau sampling should be explained. One-dimensional (1D) simulations typically use the energy $E$ as the sampling
direction, where sampling direction refers to the quantity over which a random walk is pursued. When considering multi-dimensional Wang-Landau sampling, a second quantity is placed orthogonal to the primary direction such that a grid of values is available to estimate the density of states, e.g., in this work that secondary direction is flexibility \( J_A \), leading to a density of states function \( g(E, J_A) \). The multi-dimensional density of states does not change the algorithm presented above, but rather allows the simulation to move more freely through computational space. Although, this freedom can in some cases come at the expense of CPU time. Issues with this method will be discussed in chapter 8, along with results of this method applied to semi-flexible homopolymers. In addition, multi-dimensional Wang-Landau sampling has been applied with success in other studies \([84, 93, 94]\), including simulations of proteins and polymers \([95, 96, 27]\).

5.6 Homopolymer move set

In this work, a number of Monte Carlo (MC) moves are employed in order to rapidly change configurations in the homopolymer model described in Chapter 3. These moves represent changes in energy (or stiffness) as described in the “general” MC approach in Chapter 4. We employ the following types of moves in various combinations depending on the study: 1) single-monomer displacement, 2) end cut-and-join, 3) reptation, 4) crank-shaft, 5) random-pivot, and 6) stiffness change (which changes the flexibility \( J_A \), not monomer or bond positions). Some of these moves have been implemented as energy dependent moves, which has led to dramatic increases in performance. In initial work, only moves 1, 2, 4, and 5 were used, with the remaining moves added at later times to improve efficiency.

The general procedure for making a MC move is as follows: 1) calculate the initial energy \( E_i \) of the configuration (if not already stored from a previous move), 2) make a change to the system (the trial move or update), and 3) calculate the final energy \( E_f \) following this change. In the following sections, the acceptance of the move is based on the standard Wang-Landau criterion, but in principle, the moves described below could be implemented
for any Monte Carlo technique. The original description of detailed balance and transition probabilities discussed in the beginning of this chapter, as well as in Chapter 4, will be used to describe trial moves used in this work, where Eq. 4.4 will be of particular importance. An introduction to MC trial moves applied to polymer systems can be found here [4].

5.6.1 Single monomer displacement

Displacement moves take a single monomer and displace it by some random amount. In general, these moves are responsible for the majority of changes made to the system, and in this work, were the most critical of all the moves when considering the entire phase space of each homopolymer chain. The selection of a monomer is done randomly, such that all monomers have an equal probability of being selected. Once selected, the monomer is displaced by a random amount along each of the standard $xyz$ coordinate axes.

The random amount in which the monomer is displaced is a subtle issue. Typically, these types of moves rely on a maximum displacement value $D_{\text{max}}$ multiplied by a random number (between 0 and 1), such that each step length falls between 0.0 and $D_{\text{max}}$. Traditional generalized ensemble simulations require that this value be able to satisfy the need for large moves at high energies (unwound, random-coil chains), and small moves at low energies (dense, solid-like chains). Initial work relied on value of $D_{\text{max}} = 0.2$, which was a suitable combination of small and large step sizes. It is noted however, that in this case a large portion of the moves at low energy are rejected due to the compact nature of these configurations. The benefit of this simple implementation of a single value of $D_{\text{max}}$ is that detailed balance is easily satisfied in Eq. 4.5, as the proposal probabilities ($W^n_{p}^{m-n}$ and $W^n_{p}^{m-n}$) are symmetric. However, there is a need for step sizes that vary based on the current conformation, allowing compact conformations to make small changes to monomer positions, while expanded conformation make larger changes.
This move has become even more effective with the introduction of energy dependent step lengths. Generalized ensemble simulations aimed at estimating the density of states typically rely on energy as a sampling parameter, making it a suitable choice in which to create a function $D(E)$ representing the maximum value in which a monomer can be displaced. Such a function would allow a variable step size that could be directly related to the properties of each configuration. However, a scheme for generating this function must be introduced.

There are any number of ways to select or generate an appropriate $D(E)$ function, and in this work, the method proposed by Schnabel et al. [97] was selected, which relies solely on how the energy changes from move to move. Consider a single displacement move from $E_i$ to $E_f$, where initially, $D(E)$ is set to 0.1 for all energy values. As the move is made, the following criterion is used to regulate $D(E)$:

$$D(E_i) = \begin{cases} 
D(E_i) - \epsilon, & \text{if } E_f > E_i \\
D(E_i) + 2\epsilon, & \text{if } E_f \leq E_i
\end{cases}, \quad 0 < \epsilon \ll 1,$$

(5.4)

where $\epsilon$ is a modifier that adjusts the function $D(E)$. In general, the criterion is designed to create an approximately 66.6% acceptance probability. In figure 5.1, $D(E)$ is plotted for a chain length $N = 13$, showing the values of the displacement step lengths as a function of energy. This function allows for small moves at energies with compact, closely packed, monomers, while also allowing for larger step sizes at high energies where the chain is typically in a random-coil state. A noticeable plateau is seen at high energies, where the maximum displacement is set such that this step length does not become unphysical, or more importantly, does not become so large that it upsets the bond potential with each move. A similar restriction should be placed on low energy values of $D(E)$, where in this work a value of 0.0001 is typically set (although this is a subtle issue that depends on bin resolution, the minimum energy sampled, and the desired information). This energy dependent move dramatically improves the efficiency of the simulation of both flexible and semiflexible

---

3This section and its notation relies heavily on the pioneering work of Schnabel et al. [97]
homopolymers, and is applicable in a wide range of other physical systems. However, certain constraints must be discussed regarding its implementation, and satisfying the detailed balance condition.

Consider the selection of the \( m^{th} \) monomer, where the overall conformation is currently at some high energy \( E_h \), and a well-defined value of \( D(E_h) \) exists. Suppose further that the displacement of the \( m^{th} \) monomer changes the overall energy to some lower energy \( E_l \), where \( D(E_l) \) also exists, but is much smaller ( \( D(E_l) \ll D(E_h) \) ). This would mean that an additional displacement of the \( m^{th} \) monomer may not allow it to return to its original position. This is a critical violation of the detailed balance condition, which must be checked within each move. Therefore, if a trial move has a new position that makes moving back to the previous location unattainable, this move “must” be rejected. As a more straightforward example, consider a configuration with energy \( E_1 \) with a single monomer whose position is at the origin. The current value of \( D(E_1) \) allows this monomer to be moved along the x-axis to position \((1,0,0)\), hence changing the energy to \( E_2 \) and the displacement length to \( D(E_2) \). In order for detailed balance to be fulfilled, \( D(E_2) \) must be at least equal to 1.0, otherwise, it is impossible for the monomer to return to its original location. As stated previously, in order for detailed-balance to be satisfied, displacement moves must have some overlap, i.e., must satisfy the condition \( |R_1^{(k)} - R_2^{(k)}| \leq D(E_n) \), where \( R_1^{(k)} \) is the initial position of monomer \( k \) and \( R_2^{(k)} \) is the final position.

Regardless of the blatant violation above, moves that satisfy this condition still violate detailed balance. This is because the proposal probabilities are not symmetric due to the function \( D(E) \). When \( D(E) \) is a constant value, probabilities are symmetric \((W_{n\rightarrow m}^{p})\), and there is no need to correct for this information. In the case where \( D(E) \) varies as a function of energy, this violation can be corrected in the solution to Eq. 4.7 through the inclusion of the proposal probabilities, where \( W_{n\rightarrow m}^{p} = 1/D^3(E_n) \), which is the inverse volume of the cube representing all possible locations to which the monomer can be displaced. Including
Figure 5.1: Energy dependent displacement moves (step-size and efficiency). (Top) Displacement step-size as a function of energy $D(E)$ for two values of epsilon. This function regulates the maximum amount by which a monomer can be displaced, where the black curve represents a standard implementation, and the red line is the function generated using the energy dependent scheme described in the text. $\epsilon = 0.0001$ is the amount by which this function is modified during the simulation. (Bottom) The number of Monte Carlo sweeps is plotted as a function of the modification factor $\ln[f]$, showing that the energy dependent displacement moves are indeed more efficient than standard implementations.
this in to Eq. 4.7 gives

\[ W_{n \rightarrow m} = \min \left( \frac{P_m \cdot D^3(E_n)}{P_n \cdot D^3(E_m)}, 1 \right) . \quad (5.5) \]

Both detailed balance conditions written above can be compressed into a single formalism for the acceptance criterion such that

\[ W_{n \rightarrow m}^a = \begin{cases} 
\min \left( \frac{P_m \cdot D^3(E_n)}{P_n \cdot D^3(E_m)}, 1 \right), & \text{if } |R_n^{(k)} - R_m^{(k)}| \leq D(E_n) \\
0, & \text{otherwise} \end{cases} \quad (5.6) \]

This energy dependent approach to Monte Carlo trial moves is rather new to the community currently working on simple polymer systems. However, with the dramatic effects in sampling efficiency and minimum-energy state identification presented here, as well as those from Schnabel et al. [97], it is expected that this approach will become broadly used in the near future. Brief results for an \( N = 13 \) fully-flexible homopolymer are presented showing the effectiveness of this technique compared to standard displacement moves. The upper plot in Fig. 5.1 shows \( D(E) \) for two values of \( \epsilon \). The black curve represents a standard implementation with \( \epsilon = 0.0 \), meaning there is no energy dependence, and the red line is the function generated using \( \epsilon = 0.0001 \). Even from this rather simple function indications of thermodynamic behavior can be observed, where a small dip indicative of the liquid-solid transition can be observed at \( E/N \approx -1.75 \). The lower plot represents the number of Monte Carlo sweeps, averaged for 10 runs per curve, as a function of the modification factor \( \ln[f] \), showing that the energy dependent displacement moves are indeed more efficient than standard implementations. In addition, they provide a remarkably fast approach in the early stages, indicating that the moves locate minimum energy states faster than standard methods. It is also noted that the progression of MC time appears to be much smoother.

Another important note is that this move should be turned off once the function has had sufficient time to equilibrate. This is so that in the late stages of the Wang-Landau algorithm, where the histogram should become flat and detail balanced is no longer violated, major changes based on moves are not visible in the estimates of \( g(E) \). A scheme designed to reduce \( \epsilon \) over time could be introduced, such that \( D(E) \) becomes a “very” smooth function.
However, this is not absolutely necessary, as its purpose in improving efficiency is quite apparent without this type of step.

5.6.2 End cut-and-join

End cut-and-join moves are a type of bond breaking, or bond rearrangement, move applied in this work. Although there are three types of cut-and-join moves [98], only the “end cut-and-join” type is employed. This is due to the fact that these moves lose efficiency when flexibility is varied, therefore, time was not spent implementing the other, more complicated, cut-and-join moves. This procedure begins by randomly choosing one end of the polymer $n_{\text{end}}$, and then attempting to make a new bond with a non-neighboring monomer $n_i$. If a monomer is within an appropriate range to satisfy the bond potential, a new bond is drawn. Once this new bond is in place, one of $n_i$’s two initial bonds must be destroyed, and in the case of “end cut-and-join” moves, the bond closest in bond number to the selected end must be broken. This prevents loops from forming in the chain. Now that these bonds have been removed, the monomers along the chain are renamed such that a linear order is returned.

The efficacy of these moves is apparent when proposing trial moves for highly compact configurations. This is important because it gives a much needed boost in efficiency of sampling low energy configurations. In particular, many of the ground state configurations associated with flexible homopolymers, e.g., magic number icosahedrons, have one end of the chain in the center and the other on the outer shell. The dynamics of a simulation using only displacement moves does not always find this bond arrangement when searching for minimum energy states. Bond rearrangement moves allow for the sampling of these condensed states, in particular the search for bond arrangements that best satisfy energy constraints. However, it must also be understood that the probability of making one of these moves can be very low when implemented in a purely random fashion, where the proposal probability for the end $i$ is a function of the total number of possible exchange monomers ($n^{em}$), i.e., $W_{n \rightarrow m}^p = 1/(n^{em}_m)$. This is primarily because there are typically at most 12 neighboring
monomers (in some cases slightly more) in the most stable minimum energy configurations. This means that as \( N \) increases, \( n^e m \) remains constant, and the probability of randomly selecting one of the possible monomer pairs diminishes rapidly, especially since 12 is viewed as a rough maximum for the number of monomer pairs.

To further improve the efficiency of these moves, a search for neighboring monomers with which to make an end cut-and-join move has been implemented. At the beginning of this move, one end of the chain is selected at random and a list is created of monomers within a distance of \( r < 1.2 \) (which is the bond length maximum). This list is now used to randomly select a monomer with which to perform a cut-and-join move. This list provides a more efficient approach for making this move, but detailed balance must still be considered. In many cases the number of monomer pairs differs between two sequential moves, e.g., a move that takes an end of the chain from the core to the surface. In order to correct for this, the proposal probabilities must be included in the acceptance criterion, meaning that the general solution to the detailed balance equation becomes

\[
W^{a}_{n \rightarrow m} = \min \left( \frac{P_{m} \cdot n^{e m}_{m}}{P_{n} \cdot n^{e m}_{n}}, 1 \right). \tag{5.7}
\]

### 5.6.3 Reptation

Reptation, or “slithering-snake”, moves involve breaking a bond at one end of the polymer chain, and then reattaching this bond and monomer to the opposite end. In this work, one of the two ends of the polymer chain is selected at random, and then detached and placed at the opposite end at a random direction. The original design of these moves was to promote more drastic changes in compact configurations, particularly when an end is buried in a collapsed configuration. However, at extremely low energies these moves are typically only accepted when one of the ends exists on the surface of a compact configuration, meaning that the acceptance probability remains low in this energy region. On the other hand, these moves tend to promote changes at mid and high energies, helping rapidly rearrange configurations that would take longer using only moves based on particle dynamics, such as displacement.
moves. The implementation of this move fulfills detailed balance since it is a purely random process, meaning there are no special considerations to try and improve their efficiency.

5.6.4 Crank-shaft

Crank-shaft moves belong to a class of rotation moves which involve rotating segments of the polymer chain. Initially, two individual monomers are selected at random on the chain and the segment between is marked for rotation. In order for a rotation to be worth while, at least one monomer must be rotated. Similarly, a rotation of the chain between two ends should not be performed, as this would result in an identical state. Once two suitable monomers have been selected, the segment between is rotated around an axis drawn between these two monomers. The two monomers and the segment between are translated to the origin before rotation, and then translated back to their original positions after rotation. The rotation angle is selected randomly between $-\pi$ and $\pi$.

Rotational moves played a large role in test simulations when this work started. However, the focus of this work shifted more and more toward collapsed states where these moves can play an almost non-existent role in changing configurations. Therefore, in the most up-to-date work, these moves are not used.

5.6.5 Random pivot

Random pivot moves are another rotation move used to move multi-monomer segments to new locations. A random monomer is chosen initially, and then the segment below or above the monomer is selected randomly. These monomer are then marked for rotation around a randomly generated axis. All monomers belonging to the randomly selected segment are translated to the origin, rotated, and then translated back to their initial coordinates. The rotation angle is selected randomly between $-\pi$ and $\pi$. 
Again, a disclaimer is offered regarding these rotational moves. Since the focus of this work has shifted to collapsed states, these moves are no longer included due to their inefficiencies at high densities.

5.6.6 $J_A$ Change

In multi-dimensional Wang-Landau sampling, multiple quantities are used to perform a random walk in conformational space. In this work, a two-dimensional method is implemented with directions being energy $E$ and stiffness $J_A$, meaning an update is required to change $J_A$. This update simply changes an initial value of $J_A$ by some random amount between 0.0 and $\epsilon dJ_A$, where $dJ_A$ is the binning resolution in the stiffness direction, and $\epsilon$ is some numerical constant used to vary the maximum change to $J_A$ as a function of $dJ_A$. Some testing was done to try and optimize this move, where $\epsilon$ was varied between 1.0 and 4.0, but no dramatic differences in efficiency were found in preliminary testing. This move generally has a high acceptance probability due to the small step size, and parameters set for this move were selected such that they are more efficient for low energy states. In the future, this move should be marked for improving efficiency, or at least a greater attempt to understand how the dynamics of the simulation are affected by different variations of parameters.
Chapter 6

Algorithmic considerations and analysis techniques

This chapter represents a timeline of some of the most important algorithmic and analysis developments produced by this work, and the presentation of these results flows directly along this timeline. This work represents a time when the author was performing introductory analysis of the Wang-Landau algorithm, as well as techniques for estimating the error produced by multiple independent simulations. Although results are presented for only the flexible case, techniques and understanding garnered here have been applied in all aspects of work related to Wang-Landau simulations. This chapter will begin with preliminary simulation results based on testing fundamental parameters associated with the algorithm, and ending with the description of advanced analysis techniques used to calculate thermodynamic quantities from multiple, independent simulations.

6.1 Overview

As discussed in the introductory chapters in this dissertation, the Wang-Landau algorithm is a relatively new method, whose power allows for the generation of large (and often complex) data sets. With any new method, advanced analysis techniques of simulation results must be built up along side the algorithm. This chapter represents work aimed at understanding how various parameters associated with the Wang-Landau algorithm affect results, while also providing sophisticated analysis techniques needed to generate accurate results with error bars. The power of this method to efficiently generate the density of states for a given system

---

allows for the calculation of advanced thermodynamic quantities, but, one must take great care when interpreting these results and their statistical properties.

As previously mentioned, this chapter represents a timeline, beginning with the introductory analysis of the algorithm and flexible homopolymers, and ending with the development of sophisticated analysis techniques. Therefore, questions may arise regarding why certain analysis techniques were not implemented for results appearing at the beginning of this chapter. The answer here is that these techniques were not in place when these results were generated. Regardless, all results presented here are of high quality (error bars presented), and are the results of multiple independent simulations (unless specifically stated). The sections presented in this chapter represent important highlights of studying algorithmic and analysis based calculations, all of which affect how results are attained and presented in the following chapters on flexible and semi-flexible homopolymers.

This chapter proceeds as follows: 1) results are presented for a few of the important parameters directly associated with the Wang-Landau algorithm, 2) description of analysis techniques used to calculate average properties from multiple independent simulations, 3) how to validate calculations of thermodynamic quantities using information from the aforementioned analysis techniques. In addition, this chapter will rely heavily on density of states based calculations of the heat capacity $C_V(T)$ or specific heat $C_V(T)/N$, as well as basic statistical analysis such as averaging ($\bar{A} = \sum_{i}^{M} A_i$, where $M$ is the number of samples) and error-bar generation using standard error of the mean ($\sigma/M$, where $\sigma$ is the sample standard deviation). These calculations are discussed in greater context in Chapter 4.

6.2 Binning resolution

A subtle factor in any extended ensemble simulation pertains to the storage of density of states. Simulations of systems with continuous energies require that the energy range be divided up into a set of discrete intervals (or bins) with some width $dE$. The partition
function can be approximated in terms of such a discrete set

\[ Z(T) = \int_{E} g(E)e^{-E/k_bT} dE \approx \sum_{i} g(E_i)e^{-E_i/k_bT} \]  

(6.1)

where the partition function has been transformed from a continuous representation, to a sum over all binned energy states with resolution \( dE \), \( E_i = E_0 + i \cdot dE \). This discretization does not change the continuous nature of \( E \), but is only used to regulate which values have been visited, and hence, store information regarding \( g(E) \). A subtle consequence of this adjustment is the choice of \( dE \) values.

In this work, a variety of \( dE \) values have been used, which most often depends on the chain length and desired information regarding the system. Published results for fully flexible homopolymers [26] have used a range from 0.1 → 2.0 (more recent results have used values as low as 0.01), with the smallest values typically corresponding to the smallest chain lengths \( (N \approx 10) \) and the largest values to the largest chain lengths \( (N \approx 561) \). One might think that this resolution would greatly affect the accuracy of calculations of thermodynamic quantities, i.e., higher resolutions produce better results because more information is stored. However, results presented here show a broad range of \( dE \) capable of producing results with similar statistical error.

In order to test correlations in error and resolution, a systematic analysis was performed of heat capacity curves calculated from \( g(E) \) with different \( dE \) values. Calculations of heat capacity from the density of states can be found in Chapter 4. Results for an \( N = 25 \) length chain are presented here, where a reference \( C_V^*(T) \) was generated by averaging the results from 100 independent simulations using standard averaging techniques. These 100 \( g(E) \) all used a resolution of \( dE = 0.01 \), such that the reference incorporated as much information as possible. This reference was then compared to sets of \( k \) independent simulations (typically 20), with each set representing some variation of \( dE \), which were then used to produce a number of heat capacity curves, i.e., \( C_V(T, dE)_k \). The absolute error between \( C_V(T, dE) \) and
the reference $C^*_V(T)$ was then measured

$$
\epsilon_{C_V}(T, dE) = \frac{|C^*_V(T) - C_V(T, dE)|}{C^*_V(T)},
$$

(6.2)

and the average $\overline{\epsilon_{C_V(T)}}$ was taken over all $k$ runs for each temperature. In the following results, this error is discussed as a function of $dE$ and temperature.

The error associated with $dE$ is not as great as one would expect. In Fig. 6.1 a graph of “error” vs $dE$, corresponding to the above discussion, is shown for $N = 25$, and it is seen that a plateau exists where error remains roughly the same for a broad range of $dE$ values. The highest temperature, $T = 2.0$, reveals only a slight increase in $\epsilon_{C_V}(T, dE)$ as $dE$ is increased. This is not surprising, as simulations are typically much better at sampling energies associated with this type of temperature. As $T$ is decreased, all of these curves shift...
toward larger errors, which has more to do with the difficulties of sampling low energy states than error associated with $dE$. A striking feature of the lowest temperature value $T = 0.075$ is the dramatic increase in error for large resolutions ($dE > 0.1$). This introduces another issue with binning resolution in that small values of $dE$ are required in order to sample near the ground state (low $T$ values correspond to low energies). Connections between binning resolution and sampling minimum energy states will be discussed through the remainder of this dissertation.

An important note is that implementing a large number of bins can lead to excessively large CPU times; the nature of the algorithm is to visit all binned energies a significant number of times in each simulation. Extremely small values of $dE$ produce algorithmic constraints, where many more moves are required to fill the large number of bins, e.g., for $N = 100$, $dE = 0.01$ requires $\approx 70000$ bins, while $dE = 1.0$ requires $\approx 700$ bins. Larger values of $dE$ allow the algorithm to fill each bin more quickly, but error must be considered (as has been shown above). Another important consideration is how $f$ is reduced. If the condition is to only visit each bin a small number of times, then it is reasonable to have a large number of bins. While if using a traditional “flatness” approach, this criterion can be held back by only a small percentage of bins (or even a single bin).

As mentioned in relation to Fig. 6.1, one must also consider how effectively minimum energy states are sampled. In many cases, a very small resolution ($dE = 0.01$) is required in order to reach these states, e.g., for $N = 13$ assuming a set energy range, simulations with $dE = 0.01$ are often faster (particularly when reducing $f$ without a histogram flatness) than simulations using $dE = 0.1$, because the algorithm is forced to sample configurations near the minimum energy states. In the case of large $dE$, the algorithm can be satisfied if only states at one boundary of the bin width are sampled, which is often the case at low energies. However, as chain length is increased, one must be careful that the number of energy bins is not growing too large, primarily for algorithmic reasons discussed in the previous paragraph. The number of moves required to fill each bin can become quite large, very quickly.
Binning resolution is a very subtle issue in these simulations, as well as any generalized ensemble simulation. One must consider the type of information desired and what $dE$ will be most effective at producing this desired results. If interested in transition phenomena occurring at relatively high temperatures (liquid-solid and coil-globule), then larger $dE$ values are able to provide accurate results. If interested in producing minimum energy states, then a smaller $dE$ should be employed, and one must understand how to modify the simulation in order to maintain an efficient algorithm. The conclusion of this work is that a wide range of $dE$ exists from which precise results can be produced, and as long as one is considerate of the above effects, large errors will not occur.

6.3 Saturation of error

Simulations of protein-like systems face an important issue in the choice of the final value of the modification factor $f$. The original Wang-Landau study on discrete spin models [75] used a final value of $\ln[f] = 10^{-8}$; but in most recent studies on polymer models, a final modification factor of $\ln[f] = 10^{-6}$ was used. The aim of this section is to better understand why $10^{-6}$ is an appropriate choice for polymer and protein systems.

In order to resolve whether the selection of $\ln[f_{\text{min}}] = 10^{-6}$ made in some studies was warranted, 50 independent runs were made for each of the following chain lengths: $N = 10$, 25, and 50. In each run, $g(E)$ was recorded for each value of the modification factor. The final value of $\ln[f]$ was reduced to at least $\ln[f_{\text{min}}] = 10^{-8}$, and the final values of $\ln[g(E)]$ were then averaged together. In Fig. 6.2, the average values of $\ln[g(E)]$ are plotted as a function of energy for each of the chain lengths, where the energy ranges have been set such that the maximum and minimum boundaries have been extended far enough to insure that important thermodynamic features are reproduced correctly\(^1\). Using these average values of $\ln[g(E)]$, a reference value of the specific heat, $C^{\text{ref}}_V$, is calculated for each chain length.

\(^1\)The energy range for $N = 50$ does not extend to points very near the minimum energy. These results were taken at a time when minimum energy configurations were not a major concern.
Figure 6.2: Average density of states calculated for three different chain lengths, \( N = 10, 25, \) and 50. Each point is the average of 50 independent runs and the error bars are much smaller than the symbol size. These curves are used to create reference thermodynamic quantities, which are in turn used to study the efficacy of individual simulations.

reference is then compared to specific heat curves calculated from the \( \ln[g(E)] \) values at each iteration of \( \ln[f] \). When comparing the reference with specific heat curves calculated at a particular value of \( \ln[f] \), the absolute differences, \( |C_V^{\text{ref}}(T) - C_V(T)| \), are then averaged and stored for each value of \( \ln[f] \). These differences are then averaged for all 50 independent runs and error bars are calculated.

The top of Fig. 6.3 shows the absolute difference between \( C_V^{\text{ref}} \) and the calculated \( C_V \) for each value of \( \ln[f] \), and the bottom of Fig. 6.3 shows the average number of MC sweeps versus \( \ln[f] \) (where both plots are generated from the same 50 run set for the individual chain lengths). The top plot shows how the error decreases during the course of a simulation. For the \( N = 10 \) system length, the curve decreases with \( \ln[f] \), but then levels off at \( \ln[f] \approx 10^{-6} \). The simulations for this system size were continued to a final value of the modification factor of \( \ln[f] = 10^{-10} \), and it is clear that any extra time spent running a single simulation
Figure 6.3: Saturation of error \( (C_V \text{ and efficiency}) \). Top: average absolute difference in the specific heat between the reference curve \( C_V^{ref} \) and single iterations \( C_V \) of \( \ln[g(E)] \) versus modification factor. Bottom: average number of MC sweeps versus modification factor, where these data come from the same independent runs used in the top plot. A total of 50 independent runs were performed for each chain length. Flatness in all these simulations is \( p = 0.8 \).
past \( \ln[f] \approx 10^{-6} \) does not necessarily improve the final result (at least for our choice of flatness criterion \( p = 0.8 \)). The other two chain lengths plotted were simulated to a value of \( \ln[f] = 10^{-8} \). Again, the average deviation levels off at \( \ln[f] \approx 10^{-6} \). However, the bottom plot reveals that a huge price is paid in terms of the number of MC sweeps needed to iterate \( \ln[f] \) from \( 10^{-6} \) to \( 10^{-8} \). For \( N = 25 \), the number of MC sweeps executed after \( \ln[f] = 10^{-6} \) is twice the number executed before \( \ln[f] = 10^{-6} \). For \( N = 50 \), a similar behavior is observed, only the number of sweeps needed to reach \( \ln[f] = 10^{-8} \) is nearly 3 times larger than the number needed to reach \( \ln[f] = 10^{-6} \). Thus, when considering both efficiency and the error, \( \ln[f_{\text{min}}] = 10^{-6} \) is found to be a very reasonable choice.

This study is meant to be only a reference in understanding the behavior of \( f \), and its final value. In the next chapter, certain results, which will be discussed in detail, required \( \ln f = 10^{-7} \). The issue of saturation of error has been studied in the past few years [99, 87, 88, 90–92], with various groups providing new techniques for addressing this issue. In this author’s opinion, no method has been introduced which has general application in polymer and protein simulations. In addition, in this work the statistical error has been significantly larger than the saturated error in a single simulation, meaning there is a greater need for the averaging of multiple, independent simulations. This is the approach taken in the remainder of this document.

6.4 Energy range considerations

Often in simulations of systems with continuous energies it is not clear where the energy boundaries occur. For the interactions described in this study, a unique minimum energy \( E_{\text{min}} \) typically occurs for different variations of the systems (\( N \) or \( J_A \)), while a maximum energy \( E_{\text{max}} \) exists as a more generic stopping point which is more consistent for different variations of the same system. \( E_{\text{min}} \) depends highly on the ground-state configurations produced by the simulation, while \( E_{\text{max}} \) is often set to account for energies whose properties are more
similar as a function of chain length. The above values are typically system dependent, and certain checks can be used to make sure they are appropriate.

One such check is to monitor behavior of various thermodynamic properties. Specific heat is one such quantity, where noticeable differences in behavior can be observed when certain portions of the energy range have not been sampled. For example, if low energies bins are not sampled during a particular simulation, a drop off to zero is observed, where the equipartition theorem is clearly violated\(^2\). Fig. 6.4 shows the low temperature region of the specific heat \(C_V/N\) for a chain length of \(N = 50\), using a number of different energy ranges for the estimation of \(g(E)\). Each curve has a different minimum energy value in the density of states, i.e., the random walk \(g(E)\) is pushed closer and closer to the ground state. The maximum energy sampled in each curve is the same, \(E_{\text{max}}/N = 3.0\). For clarity, the low \(T\) values drop off is shown for each chain, representing the low energy region where \(E_{\text{min}}/N\) is very important. For \(E_{\text{min}}/N = -3.0\), a small shoulder just below \(T = 0.5\) can be observed, that then drops off to zero as \(T\) is lowered further. For \(E_{\text{min}}/N = -3.3\), the small shoulder has become a small peak at lower \(T\) (\(\approx 0.3\)). The next curve is for \(E_{\text{min}}/N = -3.6\) and a dramatic peak at \(T \approx 0.3\) is clearly seen, and a shoulder that drops off at lower \(T\). The final curve is for \(E_{\text{min}}/N = -3.75\), where the resultant peak representing the liquid-solid transition is within the error bars of the previous curve, and the low \(T\) portion extends further with a nonzero value. Note that for the \(E_{\text{min}}/N = -3.75\) curve, the specific heat approaches a constant value at low \(T\) as would be expected from the predictions of the classical Equipartition theorem. This value can be extrapolated and compared to this prediction by considering the degrees of freedom for each monomer. Although no rigorous comparison is made, it is noted that this provides a physical check on simulation accuracy at low temperature. These curves show the importance of sampling range selection on the final determination of various properties. Based upon these findings, it is believed that an overly restricted range of energy

\(^2\)The equipartition theorem can provide meaningful results at low \(T\) for classical models, and the homopolymer model studied here is no exception. Although detailed comparisons are not made in this dissertation, it is noted that this is a tool that can be used to check simulation accuracy at low \(T\).
Figure 6.4: Specific heat for a chain length of $N = 50$, varying the minimum energy sampled. Each curve is produced using the standard average of 15 independent simulations, and error bars were found through a jackknife analysis.

for the random walk in earlier works [34, 35] is a major factor in the differences between their results and those presented here. This type of analysis becomes even more important in future discussion of solid-solid transitions, which typically occur at temperatures well-below the liquid-solid transition presented in Fig. 6.4.

The goal of all simulations should be to produce results that incorporate the physical phenomena one wants to observe. When trying to understand behavior at all temperatures, one must push the energies to values that will generate appropriate results for these temperatures. As seen in Fig. 6.4, a $E_{\text{min}}/N = -3.75$ produces reasonable behavior at low $T$ compared to higher $E_{\text{min}}/N$ according to the equipartition theorem. Understanding where these minimum energies should occur for all chain lengths was a big concern in the early stages of this work. It is noted that the correlation with Lennard-Jones clusters, which was
found at a much later date than when these data in Fig. 6.4 were produced, made determining these minimum energies a matter of looking them up in a table and making a small transformation. However, these results show the importance of checking and interpreting the behavior of thermodynamic calculations using results from generalized ensemble simulations. In particular, if this type of analysis had not been performed, the full form of the liquid-solid transition may have remained hidden as was seen in other studies [34, 35]. At the end of this chapter, more sophisticated analysis methods are presented that automatically account for the validity of a calculation based on the canonical probability.

6.5 Analyzing multiple independent simulations

In Chapter 4, the basic analysis of generalized ensemble techniques involving the density of states $g(E)$ were introduced. The algorithmic estimation of $g(E)$ allows for the calculation of the partition function $Z$, from which all thermodynamic quantities can be calculated. This is also true for structural quantities, although energy dependent averages of these quantities must be generated in order to weight them with $Z$. Generally, the error in any subsequent quantity can be traced back to the original estimation of $g(E)$. In order to provide quantitative estimates of error in this work, multiple independent simulations are considered. The following section discusses how multiple density of states functions should be handled for error estimation, and additionally, how error is estimated in subsequent thermodynamic calculations.

All data presented in this dissertation (unless specifically stated) are the result of multiple independent simulations, which are averaged together and errors estimated as function of these multiple runs. In order to average independent $g(E)$ values, and subsequently find the error, certain features must first be explained. As stated in Chapter 5, $g(E)$ is a relative quantity whose values are only known in relation to their neighbors, where $g(E) + C$ is the algorithm’s estimate and there is no simple way of estimating the constant offset $C$. A typical average, along with an estimation of the standard error of the mean, is not a viable
option due to this unknown offset, but also because the density of state is typically stored as the \( \ln[g(E)] \), and \( f(x) \neq f(\bar{x}) \), or \( \ln[\ln(g(E))] \neq \ln(\ln(g(E))) \) (This simple relation will also be applied in understanding averaging issues in thermodynamic quantities). The overall goal of the analysis of \( g(E) \) is to provide an average over multiple runs, along with an error estimate for this averaged value.

Error in the density of states, in principle, should be connected to the modification factor, or MC time, used in simulating a given system. In standard implementations of the Wang-Landau algorithm, saturation of error may occur, leaving significant error in \( g(E) \). In that same regard, the ability to simulate larger and more complex systems has increased the general size and scope of \( g(E) \), where in many cases this function spans 1000’s of orders of magnitude. Handling curves which span this type of numerical landscape becomes not only an issue of statistical analysis, but also of computational analysis. Overflow and underflow often make the computational analysis of these curves difficult. Therefore, in practice the \( \ln[g(E)] \) is stored, and for some approaches, this function must be transformed into “real” space (removing the natural logarithm) in order to be averaged.

With regard to the above discussion, there are three major issues facing the averaging and error analysis of these simulations: 1) avoiding the need to know \( C \) when averaging multiple \( g(E) \). 2) managing underflow and overflow issues when \( g(E) \) spans many orders of magnitude, and 3) providing error estimates (and validity analysis) of secondary quantities calculated from the primary \( g(E) \) results. Each of these issues has been addressed, and the following sections provide detail into the methods implemented in this work. This portion directly affects all calculations beyond this point, but in addition provides a means for understanding the accuracy associated with an estimate of \( g(E) \), and subsequent thermodynamic properties, for any system.
6.5.1 Averaging the density of states from multiple simulations

In order to resolve the first issue involving the unknown offset, all $g(E)$ functions are normalized by the area under their curves. This normalization must be done in “real” space, meaning the natural logarithm must be removed from the stored value. In doing this, overflow and underflow become an issue (particularly, for large chain lengths), where $g(E)$ may span many orders of magnitude. If overflow and underflow are not issues (small chains), a simple average of these normalized curves can be taken

$$\bar{g}(E) = \frac{1}{S} \sum_{k} S_{k} g_{k}(E) \quad (6.3)$$

where $S$ is the number of individual simulations, and it is understood that these curves are summed for each value of $E$. This averaging allows us to pass $\bar{g}(E)$ into calculations of thermodynamic and structural properties, further avoiding the issue where $\bar{f}(x) \neq f(\bar{x})$. It is noted that the overflow / underflow issue does not occur if one is only concerned with calculating thermodynamics from a single run in these homopolymer simulations. However, because $g(E)$ must first be averaged to provide estimates of thermodynamic quantities, this issue cannot be avoided, and the following section will be dedicated to this issue.

As an example of the unacceptable averaging of individual thermodynamic curves (secondary quantities), Fig. 6.5 shows the mishaps involved with the relation $\overline{C_{V}[g(E)]} \neq C_{V}[\overline{g(E)}]$. These data were generated from 12 independent simulations of an $N = 300$ length chain, with flatness $p = 0.4$ and final modification factor $\ln[f] = 10^{-6}$. The black curve (single peak) was generated using the correct formalism $C_{V}[\overline{g(E)}]$, i.e., using the average of the primary simulation results to generate $C_{V}$, while the red curve (multiple peaks) was generated using the incorrect formalism $\overline{C_{V}[g(E)]}$, i.e., first calculating the secondary resulting $C_{V}$ for each individual simulation, and then averaging these results. The incorrect averaging of individual $C_{V}(T)$ curves produces a highly erratic curve, with multiple peaks. This is an important issue in that small systems typically do not see this kind of difference when performing both analysis techniques. This is due to the fact that it is easier to produce high
Using average $g(E)$ to find $C_V$ 

Average of individual $C_V$

Figure 6.5: Comparison of two methods for averaging thermodynamic quantities from multiple simulations: 1) individual $C_V$ are first calculated for each run, and then averaged to give $\overline{C_V[g(E)]}$ (red curve), and 2) individual $g(E)$ are first averaged, and then the secondary quantity $C_V[\overline{g(E)}]$ (black curve) is calculated.

quality results for these smaller systems, and strong peaks do not affect the averaged results as they do here. Simply averaging “secondary” thermodynamic quantities from individual simulations is incorrect, often leading to incorrect results. It is essential to average $g(E)$ first, and then use $\overline{g(E)}$ to calculate subsequent thermodynamic quantities. More information regarding error in this issue can be found here [13].

6.5.2 AVERAGING WHEN RELATIVE DIFFERENCE OF THE DENSITY OF STATES CAUSES UNDERFLOW OR OVERFLOW

In the case where the relative differences in individual $g(E)$ become too large, one must find a way to average these values while avoiding overflow and underflow. In this work, a method has been applied that uses canonical probabilities $P(E,T)$ to estimate a range of energies
which are relevant to estimates of thermodynamic properties for single temperatures $T$. The canonical probability is given by 

$$P(E, T) = g(E) e^{-E/k_B T} / Z(T),$$  \hspace{1cm} (6.4)$$

where these curves represent a distribution of possible states for a given temperature over the entire energy range. This formalism can be seen in the calculation of a thermodynamic property $A$ as a function of $T$ by

$$\langle A(T) \rangle = \frac{\sum_E A \ P(E, T)}{\sum_E P(E, T)} = \frac{\sum_E A \ g(E) e^{-E/k_B T}}{\sum_E g(E) e^{-E/k_B T}}.$$  \hspace{1cm} (6.5)$$

The majority of $P(E, T)$ have distributions with non-zero portions that peak for a range of energies, and tails that rapidly approach zero as $E$ is increased or decreased away from the maximum peak distribution$^3$. Therefore, a range of relevant energies $E^*$ can be extracted from $P(E, T)$. This set of energies covers the width of a particular $P(E, T)$ over which calculations have some impact on the final result. For example, the tail of most $P(E, T)$ are much smaller (essentially zero) than double precision of a 32-bit system (in practice, $\approx 10^{-14}$), which is in turn still much smaller than the statistical error in these simulations. With this in mind, all calculations can be made over the energy range \{\(E^* \mid P(E, T)/P_{\text{max}}(E, T) \geq 10^{-16}\}\}. This cuts the overall energy range into a $P(E^*, T)$ relevant portion, which become manageable regarding overflow and underflow issues.

With the above in mind, the calculation of thermodynamic quantities must proceed as calculations for individual temperatures, based on an average of multiple density of states for a specified energy range, $\bar{g}(E^*)$. This method proceeds as follows for $k$ individual simulations: 1) a temperature $T_i$ is selected, 2) $P_k(E, T_i)$ are calculated, 3) energies pertaining to $P_k(E, T)/P_k^{\text{max}}(E, T) \geq 10^{-16}$ are stored as an energy range $E^*_k$, 4) a final energy range $E^*$ is selected for all $k$ simulations to insure overlap of these energies, 5) $g_k(E^*)$ are normalized (in “real” space) by their individual areas and then averaged to give $\bar{g}(E^*)$, and 6) $\bar{g}(E^*)$ is then used in the calculation of some thermodynamic property $\langle A(T) \rangle$ as seen in Eq. 6.5.

$^3$There are also cases of bimodal distributions, where double peaks occur, and these must be taken into account in any analysis program.
Figure 6.6: Diagram for understanding energy ranges in analysis methods. Density of states for \( N = 50 \) length chain (black) and two probability distributions \( P(E, T) \) for \( T = 0.1 \) (red) and \( T = 1.0 \) (green). Each curve has been divided by its maximum, so that all curves are easily visible.

As a visual representation in the case of an \( N = 50 \) length chain, Fig. 6.6 gives probability distributions for \( T = 1.0 \) and 0.1, each normalized by their maximums, and then superimposed over the density of states divided by its maximum. The \( E^* \) ranges discussed above can be inferred from this figure and are labeled with double-ended arrows, where as expected, for low temperatures these ranges become small (as well as the distributions). This is important for the low \( T \) case as this is where the relative differences in \( g(E) \) are the largest and most likely to cause underflow or overflow issues. One can imagine how this behavior translates to “extremely” large chains, e.g., \( N = 1000 \) (which have not been simulated in this dissertation). At some point, even this method will be unable to cope with the many orders of magnitude differences these size systems would produce in the density of states.
However, as will be shown in Chapter 7, this method has been able to perform analysis on chain lengths as large as $N = 561$, which is quite large for the type of model in question.

6.6 Validating calculations

Now that there is an understanding of how multiple independent simulations are averaged, the validation of this averaging and the resultant data may also be incorporated. This is an important step in any generalized ensemble technique aimed at producing the density of states. In principle, thermodynamic quantities can be calculated at any temperature, even if technically no information exists for the energies associated with this distribution. Methods utilized here again rely heavily on the $P(E, T)$ from each simulation, with the understanding that all of these functions will be used to calculate quantities such as the specific heat.

Utilizing the previously discussed Fig. 6.6, imagine if the probability distribution were calculated for a temperature lower than $T = 0.1$ (red curve). This function would quickly become unrecognizable as a complete probability distribution, meaning the distribution would extend below the lower energy bound, and energy bins associated with the distribution would no longer exist. In order to motivate this further, we plot various $P(E, T)$ for a chain length of $N = 50$ in this low temperature region. The highest temperatures shown are $T = 0.2$ and $T = 0.1$, which represent complete distributions. The idea of a complete distribution simply means that both tails of the distribution smoothly trail to zero. The next curve is for $T = 0.083$, which is the lowest “valid” (still complete) temperature found according to the method discussed in the following paragraphs. Of more importance is the slightly lower temperature $T = 0.06$ which ends abruptly where the peak of the distribution should exist. In essence, this type of curve means that only half of the relevant energies are accounted for in an analysis of this temperature. As stated above, the calculation of thermodynamic properties would continue for such a distribution unless some check is in place.
Figure 6.7: Effects in canonical probabilities due to the minimum sampled energy for \( N = 50 \). The analysis method described in this section, which uses properties of \( P(E, T) \) in its calculations, found the lowest “correct” temperature to be \( T = 0.083 \). Temperatures above this represent \( P(E, T) \) with a complete distribution that can be used in calculations of thermodynamic quantities. Temperature below \( (T = 0.06) \) represent temperatures which should be excluded, which can be seen in the figure, where \( P(E, T = 0.06) \) shows an incomplete distribution. The purpose of having advanced analysis techniques is to avoid making calculations based on incorrect, or incomplete, distributions.
In order to avoid issues of the nature discussed above, a method recently introduced by Parsons and Williams [34] is incorporated into our overall analysis technique. When calculating thermodynamics at a given temperature, each distribution \( P(E, T) \) is checked for “completeness”, i.e., \( P(E, T)/P_{\text{max}}(E, T) < 0.001 \) on both sides of the distribution. This value can be adjusted to enforce a stricter measure of completeness (decrease 0.001), however in this study, it was found that 0.001 provides a balanced assessment of completeness. It is again mentioned that an extra check must be incorporated when dealing with bimodal distributions, where often the region between the two peaks can produce \( P(E, T)/P_{\text{max}}(E, T) < 0.001 \), meaning that the analysis method might only analyze a single peak.

6.7 Error bars in density of states and thermodynamic quantities

Discussions so far have relied solely on the average density of states, and how it produces thermodynamic quantities for this average. A major question still remains regarding the estimation of error for multiple independent simulations. In the initial stages of this work, error bars were calculated using the standard error of the mean of some averaged quantity \( x \), written as \( SE_{\bar{x}} = s/\sqrt{N} \), where \( s \) is the standard deviation, defined as

\[
s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}.
\] (6.6)

This error analysis would still have meaning for the primary results of these simulations, namely the density of states, but again considering the inequality \( f(x) \neq f(\bar{x}) \), some other method for estimating error must be employed. This can be understood from the fact that an average density of states would only give a single specific heat curve, therefore, there is no real data set with which to generate a “standard” analysis of error.

Error in this method can be found using a jackknife analysis [13], which the previous citation recommends as the standard for error bar calculations. This is due primarily to its ability to account for bias, but also because in unbiased situations, the jackknife and standard
error estimates agree. An estimate of the standard deviation can be found using jackknife estimators for a secondary quantity $f$. These estimators are defined as $\bar{f} = \frac{1}{N} \sum_{i=1}^{N} f_i$, where $f_i^J = f(x_i)$ and $x_i = \frac{1}{N-1} \sum_{k \neq i} x_k$, leading to the standard deviation estimate of

$$s_J = \sqrt{\frac{N - 1}{N} \sum_{i=1}^{N} (f_i^J - \bar{f})^2}. \quad (6.7)$$

Further details regarding this method, along with a discussion of bias and other issues regarding statistical analysis in Monte Carlo simulations, can be found here [13].

6.8 Conclusions

The chapter has shown that algorithmic considerations are highly dependent on the nature of the investigation. The analysis techniques proposed here, however, are applicable to all generalized ensemble simulations. As in any computer simulation technique, one must understand how error propagates in the system being studied. Generalized ensemble techniques are some of the most powerful simulation methods available, but as has been shown above, the huge data sets accompanying these techniques are not error free. Researchers must take into account systematic and statistical error that can occur.

As a final implementation of the analysis techniques discussed above, the specific heat for the data shown previously in Fig. 6.4 is recalculated and plotted for comparison in Fig. 6.8. This correctly analyzed data shows three key features: 1) the importance of selecting an appropriate energy range over which to simulate, 2) the drop off at low $T$, which dissatisfied the equipartition theorem, has been replaced by a single-validated temperature, and 3) subtle differences in peak height can be seen, which are due to the averaging of secondary quantities issues previously mentioned, as well as to the “completeness” check of individual probability distributions.

In addition, it is noted that this author has been made aware of other techniques aimed at handling underflow and overflow issues that can greatly improve the efficiency associated with calculating and averaging thermodynamic properties. Unfortunately, the author has not
Figure 6.8: Specific heat curves for $N = 50$ calculated with the advanced methods based on $P(E, T)$ from individual simulations. Again, this shows the importance of selecting an appropriate energy range for your simulations. It also shows the power of this analysis technique to tell one when they have made valid calculations, i.e., a safer way of interpreting features in a curve. Curves are calculated from identical data sets used in Fig. 6.4 and a jackknife analysis is used to calculate error bars.
had a chance to directly compare these methods with the analysis techniques described in this chapter, therefore, no definitive statements can be made regarding efficiency. However, it is mentioned that one of these methods seems quite promising, as it involves using the “GNU Scientific Library” [100], which allows for the storage of any number with arbitrary precision, meaning underflow and overflow are no longer issues. Although all methods proposed in this chapter on analyzing thermodynamic data are correct, such a standard computational library could provide dramatic increases in computational efficiency of this analysis.
The following results represent one-dimensional Wang-Landau simulations of single, fully-flexible \((J_A = 0)\), homopolymer chains ranging in length from \(N = 10\) to 561. Simulations are aimed at understanding three general features: transition behaviors and their scalings with respect to \(N\); comparing conformational and transition behavior with results from studies on classical clusters; and how the results in this dissertation compare with previous studies using similar models and methodologies. All thermodynamic data presented beyond this point are results of multiple, independent simulations, unless otherwise specified.

7.1 One-dimensional Wang-Landau Simulation Parameters

Simulation parameters associated with the following results were set such that the coil-globule and liquid-solid transitions could be easily studied. The investigation of these systems started with only these features in mind, and the recognition of solid-solid transitions became an exciting byproduct of these results. With this said, much of this work was performed with a move set and simulation parameters not fully tuned for studying minimum energy states and solid-solid transitions, which is why these features aren’t studied in depth for chain lengths above \(N = 55\).

The Wang-Landau algorithm and its significant parameters are described in Chapter 5. The data presented in this chapter are a result of the following: the modification factor is reduced to \(\ln[f] \rightarrow 10^{-6}\) and a flatness of \(p = 0.4\) is required for the histogram, where \(p\)
is again the minimum histogram entry divided by its average. The reduction of $f$ has been chosen to as $\ln[f] \rightarrow 10^{-6}$ based on other studies of similar systems [39, 38, 34, 33, 101] and also through testing of the effects of varying this parameter (see Chapter 6 or [42]). Recent studies [102, 18] have shown that $\ln[f] \rightarrow 10^{-6}$ can be insufficient, meaning that the modification factor must be reduced further (or in a different way altogether [103, 104]). For the majority of chains presented here, very little improvement was found in results when reducing the modification factor below $10^{-6}$, with the exceptions being the largest chains, $N = 500$ and 561, where fluctuations in thermodynamic properties were improved by reducing $\ln[f] \rightarrow 10^{-7}$. However, these two chains pose the greatest challenges to these simulation techniques and general conclusions regarding $\ln[f]$ should not be taken from these cases alone. The flatness criterion was based on preliminary testing and offers a balance between the most stringent case ($p \geq 0.8$) and cases where little or no flatness is used at all ($p \leq 0.1$). In Fig. 7.1, the logarithm of the density of states is plotted versus energy for chain lengths ranging from $N = 50$ to $N = 500$. Each curve is the result of a single Wang-Landau simulation and the maximum of each curve is normalized to zero for clarity. The most noticeable aspects of Fig. 7.1 are the orders of magnitude differences in relative height between the highest and lowest energy states. In addition, the minimum energy $E_{\text{min}}/N$ is set for individual chains, and analysis (as presented in Chapter 6) is performed so that all relevant thermodynamic properties are included.

Once an estimate of $g(E)$ is available, thermodynamic and structural quantities are calculated using a resolution of $dT = 0.001$ for temperatures between $0.001 \leq T \leq 5.0$. The analysis techniques presented in Chapter 6 provide an applicable temperature range that mainly depends on $E_{\text{min}}/N$. In this study, the lowest applicable temperature is for the smallest length ($N = 10$) and is $T = 0.036$. This lower limit increases with $N$, and is $T = 0.511$ for the largest chain ($N = 561$). Care is taken to incorporate as much of the low temperature behavior as is possible within the limits of CPU time. Thermodynamic properties not only are difficult to determine at these temperatures, but also exhibit rich phenomena in
Figure 7.1: The relative density of states versus energy per monomer for a number of chain lengths. The maximum of each curve is set to zero for plotting purposes. Values for the highest and lowest energy states for the longest chain ($N = 500$) differ by $\approx 1000$ orders of magnitude. (Each curve represents a single run and the majority of symbols were omitted for clarity.)

homopolymers, heteropolymers, and clusters [18, 102, 31, 32, 70, 105]. The error bars of all quantities are calculated using a jackknife analysis [13] as shown in Chapter 6.

7.2 Heat Capacity

In Fig. 7.2, the heat capacity is plotted versus temperature for chain lengths between $100 \leq N \leq 500$ (heat capacity was chosen for visual clarity). Two distinct features are observed, the most discernible being sharp (weakly rounded first-order) peaks at low temperatures ($\approx 0.3$ to $0.6$) representing the liquid-solid transition, and the other being shoulders at higher temperatures ($\approx 1.75$ to $2.75$) indicating the coil-globule transition. The two transitions are well-separated, with no indication that they would merge in the infinite-chain limit. (In view of the arguments [33, 38] about when such a merging could occur, it is not expected for
Two noticeable features appear: large peaks at low $T$ representing the liquid-solid transition, and shoulders at high $T$ representing the coil-globule transition. Each curve is the average of at least 20 independent simulations, where the error bars are calculated using a jackknife analysis ($dT = 0.001$).

The region between these two transitions is also of interest since one study claims that a liquid-liquid transition occurs [34, 35], while another has hypothesized only that “interesting behavior” may occur [38]. The results show no evidence of a third independent transition, and analysis of structural quantities supports this conclusion. The heat capacity, together with visualizations of the homopolymer, describes the general trends discussed previously in Chapter 3, i.e., as $T$ decreases, a single chain undergoes a coil-globule transition, followed by a liquid-solid transition. However, it will be shown that certain chain lengths exhibit non-monotonic finite-size effects in the liquid-solid temperature regime. Estimation of infinite-chain behavior will be discussed in detail later in this chapter.
Figure 7.3: The squared radius of gyration versus energy per monomer for fully flexible chain lengths 50, 100, 200, 300, 400, and 500. All curves are monotonic at both low and high $E/N$. (The majority of symbols were omitted for clarity and error bars are smaller than the symbol size.)

7.3 Structural quantities

Production runs of the first type mentioned in Chapter 5 have been applied to all fully-flexible homopolymer data. The aim of these runs is to generate results for a wide variety of structural quantities, all of which aid in the description of conformational and transition behavior. During the production run, the average values of these structural quantities are calculated as a function of the same energy scale used to define $g(E)$. Because of inefficiencies in sampling at low values of $f$ [42, 103], a saved copy of $g(E)$ from a previous iteration (around $\ln[f] \approx 10^{-3}$) is sampled with normal Wang-Landau transition rates. While resampling this iteration, the average values of structural quantities for each bin visited are calculated, making these calculations whether the new trial state is accepted or rejected, and requiring
that each bin be visited at least $10^6$ times. These structural averages are then applied to the final $g(E)$, which is unaffected by this procedure.

In Fig. 7.3, the average squared radius of gyration is plotted versus energy per monomer. These data show the general collapse behavior of this system, where at high $E/N$ the chain is unwound, and at low $E/N$ the chain becomes more compact. These curves are all monotonic, which is most discernible at high $E/N$, where the value of $R_g^2(E)$ increases systematically with $N$. To check the accuracy of our work, this technique was applied to the final iteration of $g(E)$ instead of to a previous iteration ($\ln[f] \approx 10^{-3}$), with the only differences being that using the final iteration required longer simulation times. In addition, many more structural quantities have been calculated, but from this point, will only be shown as functions of temperature.

The canonical form of such structural quantities is of more importance than the energy-dependent form presented above. The transformation between these two forms was presented in Chapter 4, and in Fig. 7.4, the squared radius of gyration per monomer is plotted versus temperature, along with its derivative with respect to $T$, for the same chain lengths considered previously. The radius of gyration behaves monotonically at both low and high temperatures, showing the relation between overall size and chain length. The derivatives with respect to temperature provide more illuminating results. Temperatures between $\approx 0.3$ to $0.6$ show five sharp, individual peaks, each representing the liquid-solid transition for a particular chain length. The temperatures at which these peaks occur correspond with those seen in Fig. 7.2. Clear indications of the coil-globule transition are the most noticeable feature, indicated by the large, broad peaks between temperatures of $\approx 1.75$ to $2.75$. The similarity between the $N = 400$ and $N = 500$ curves suggests that data are reaching chain lengths in the asymptotic regions. The liquid-solid peaks remain well-separated, but are still less distinguishable compared to the coil-globule peaks. The radius of gyration gives an excellent description of the coil-globule transition, as does the end-to-end distance (which is not
Figure 7.4: Squared radius of gyration (top) and its derivative (bottom) versus temperature for fully flexible chain lengths 100, 200, 300, 400, and 500. Features that were seen in the heat capacity are also present in this figure. The derivative indicates the coil-globule transition by large, broad peaks at high $T$. The liquid-solid transition is also represented at low $T$, but the peaks are much smaller in height and less distinguishable. (The majority of symbols were omitted for clarity and error bars are smaller than the symbol size, $dT = 0.001$.)
shown because the two quantities give nearly identical results). However, other quantities are needed to better characterize the liquid-solid transition.

One such quantity is the core density, which is plotted in Fig. 7.5, along with the negative of its derivative with respect to $T$, for the chain lengths previously considered. This quantity counts the number of monomers within a spherical shell whose origin is at the center of mass and radius is the maximum interaction distance of the nonbonded potential, $3\sigma$. For chain lengths $\geq 200$, large peaks similar to those seen in the heat capacity form in the derivative at $T \approx 0.3$ to 0.6, giving clear indications of the liquid-solid transition. Since this quantity simply counts the number of monomers within an interaction sphere around the central monomer, the $N = 100$ length chain has only a small peak, due to the fact that nearly all monomers are already within the interaction sphere at the temperature where the liquid-solid transition occurs. The largest chains show very similar behavior in the core density since they are simply reaching the limit of the number of monomers that can be packed into this single interaction sphere. However, the derivatives still describe the liquid-solid transition quite well, where large peaks can be seen steadily increasing in height and temperature as $N$ increases. Furthermore, there are indications of the coil-globule transition at high $T$, but they are not as discernible as those seen in Fig. 7.4.

The results for the chain lengths discussed so far offer a qualitative description of the collapse behavior in flexible homopolymers and can be compared directly to other studies of homopolymers with similar length. Results from the bond-fluctuation model [38–40] show qualitative agreement with the behavior of the model presented here, while results from off-lattice models remain mixed [34, 35, 31, 32]. The comparison with off-lattice cases is especially important, particularly since our models are so similar. Quantitative comparisons will be made later in this section.
Figure 7.5: The core density (top) and its derivative (bottom) versus temperature for fully flexible chain lengths 100, 200, 300, 400, and 500. Features that were present in previous figures are also found here, namely, peaks at low $T$ indicative of the liquid-solid transition and broad peaks (similar to the shoulder seen in the specific heat) at high $T$ indicative of the coil-globule transition. (The majority of symbols were omitted for clarity and error bars are smaller than the symbol size, $dT = 0.001$.)
7.4 Transitions and Clusters

All of the quantities discussed above offer insight into the nature of the coil-globule and liquid-solid transitions. Analyzing these quantities gives the location of transition temperatures $T_{tr}(N)$ for each chain length, which then leads to estimations of infinite-chain transition temperatures. However, a more detailed resolution in $N$ reveals complex finite-size effects associated with the liquid-solid transition. These effects must be addressed before estimates of the infinite-chain transition temperatures can be calculated. In Fig. 7.6, transition temperatures $T_{tr}(N)$ (top) determined from the specific heat, along with the subsequent peak values $C_V(N)$ (bottom), are considered for a number of chain lengths between $10 \leq N \leq 561$. For such small systems in the canonical ensemble, no sharp phase transition can occur, and only a rounded transition region appears. “Transition temperature” is defined as the temperature at which the specific heat maximum occurs, marking the center of the region. Three distinct trends in peak behavior can be found in these data: the coil-globule transition, and two transitions found in the liquid-solid regime. The coil-globule transition is represented by filled circles, the liquid-solid transition is represented by filled and unfilled squares, and the solid-solid transition (not reported in many previous studies [38, 40, 34, 35] with which this work has been compared) appears for chain lengths $30 \leq N \leq 54$ and is represented by unfilled triangles. The discussion begins by considering the coil-globule transition, which is found by analyzing the shoulder seen at high temperatures in the specific heat (see Fig. 7.2). In Fig. 7.6, each chain length has an estimate of $T_{tr}$ and fluctuations are small relative to neighboring chain lengths. The largest fluctuations are seen for the largest chain lengths, where the shoulders representative of this transition can become increasingly difficult to analyze and error becomes more relevant. The coil-globule transition has been analyzed in detail in other studies [25, 38, 40, 34, 35] of homopolymers and the results presented here compare qualitatively well with those works. However, the quantitative description of the infinite-chain still differs according to this model (which will be discussed shortly).
Figure 7.6: Transition temperatures \( (T_{tr}) \) from the specific heat (top), along with the corresponding specific heat \( (C_V/N) \) values (bottom), plotted versus chain length \( N \). The coil-globule (circles), the liquid-solid (squares), and solid-solid (triangles) transitions are all represented. The coil-globule and liquid-solid transitions are present in nearly all chain lengths, while the solid-solid transition appears between \( 30 \leq N \leq 54 \). The small gap between \( 42 \leq N \leq 54 \) is a result of difficulties in distinguishing the liquid-solid and solid-solid transitions. This plot has striking similarities compared to recent studies of classical Lennard-Jones clusters [106, 105, 70]. (Error bars are on the order of or smaller than the symbols.)
Table 7.1: Table of transformations found near the liquid-solid transition in studies of classical Lennard-Jones clusters [106, 105, 70] for various cluster sizes. These transformations are typically identified using peaks in the specific heat. A recent study of homopolymers [31, 32] with Lennard-Jones interactions revealed striking similarities.

<table>
<thead>
<tr>
<th>cluster size</th>
<th>ground state</th>
<th>peaks in $C_V$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13, 55, 147, 309</td>
<td>complete icosahedral</td>
<td>1</td>
</tr>
<tr>
<td>14 ≤ $N$ ≤ 30</td>
<td>anti-Mackay</td>
<td>1</td>
</tr>
<tr>
<td>31 ≤ $N$ ≤ 54</td>
<td>Mackay</td>
<td>2</td>
</tr>
<tr>
<td>56 ≤ $N$ ≤ 81</td>
<td>anti-Mackay</td>
<td>1</td>
</tr>
<tr>
<td>82 ≤ $N$ ≤ 146</td>
<td>Mackay</td>
<td>2</td>
</tr>
</tbody>
</table>

exceptions

| 38, 75 – 77 | non-icosahedral | 2 |
| 98, 102 – 104 | non-icosahedral | 2 |

The most striking feature seen in Fig. 7.6 is the low temperature behavior in the “liquid-solid” region. The belief regarding various homopolymer models has been that these systems undergo only two distinct transformations, namely, the coil-globule transition and a single liquid-solid (crystallization) transition. What was found by Schnabel et al. [31, 32], and verified by results presented here, is that for certain chain lengths a number of unique structural transformations, not limited to a single crystallization, can be found at low temperatures. The data plotted in Fig. 7.6 for low temperature transitions ($T \leq 0.6$) show remarkable similarities to the behavior of Lennard-Jones clusters [105, 106, 70, 107], and Schnabel et al. [31, 32] have confirmed that there is a near one-to-one correspondence between these two systems. To fully understand the behavior of the model presented here, general features of classical Lennard-Jones clusters must be considered.

In Table 7.1, features are listed which have been reported in recent cluster studies [106, 105, 70] regarding the ground state and transition behavior of cluster sizes up to $N = 309$ (these features are also present in [31, 32] with only a few exceptions). As discussed in Chapter 3, the magic number set $N = 13, 55, 147, 309$, and 561 all have icosahedral geometries, and
Figure 7.7: The specific heat versus temperature for the magic set of fully flexible chain lengths, where minimum energy configurations for each chain correspond to an icosahedral geometry. These chain lengths offer two generic features: a shoulder in $C_V/N$ indicative of the coil-globule transition (high temperatures) and a first-order like peak indicative of a liquid-solid transition (low temperatures). They also offer a basis for understanding the behavior of chain lengths between two magic lengths.

smaller magic number configurations remain as cores of the longer chains. The specific heat versus temperature from our own homopolymer simulations of the magic set lengths can be found in Fig. 7.7. These chains offer the previously expected transition behavior, namely, a single liquid-solid transition and a coil-globule transition. However, they also account for the plethora of unexpected ground state and transformation behaviors occurring for chain lengths between these numbers. The general descriptions of anti-Mackay and Mackay arrangements [108, 69, 106] are used to classify the ground state configurations of clusters. In Chapter 3, Fig. 3.4 provides a depiction of the anti-Mackay and Mackay overlayers as represented on one face of an $N = 13$ icosahedral core. These terms describe packing in the outer shell of a cluster, where typically an icosahedral core (based on the magic set) forms and additional monomers begin arranging on its surface. In the Mackay case, enough monomers are present
that the next layer of a magic number icosahedron forms. Anti-Mackay packing can form a variety of geometries where monomers locate the most energetically favorable place to attach to the core’s surface. In clusters, sizes between $14 \leq N \leq 30$ have anti-Mackay ground states and a single peak associated with melting is visible in the specific heat. At $N = 31$, the ground state behavior changes to the Mackay type and a peak forms well-below the melting transition. Again, this new peak corresponds to a transformation in the outer shell of the cluster and is representative of a change from anti-Mackay to Mackay states (aM/M). This transition has been well-studied in cluster systems [106, 105, 68], but has not been considered in great detail in homopolymer studies until recently [31, 32].

In Fig. 7.6, a very similar behavior is shown as compared to the cluster studies [106, 105, 70] discussed above, as well as the recent homopolymer study [31, 32]. In particular, chain lengths between $30 \leq N \leq 42$ exemplify the finite-size effects seen in these studies. Chains below this range have a single peak representative of the liquid-solid transition (unfilled squares), the most prominent being the magic length $N = 13$, whose ground state is a perfect icosahedron. At $N = 30$, a small peak forms (open triangles) and it continues to grow as $N$ increases, eventually merging with the liquid-solid transition. This is contrary to both recent cluster studies and Schnabel et al. where in those studies the small peak indicative of a transformation in the outer shell of the homopolymer appears for $N = 31$ rather than $N = 30$ (which is attributed to the minor model differences discussed in Chapter 3). In Fig. 7.8 we provide more detailed evidence of this behavior by plotting $C_V/N$ and $d(R_g^2)/dT$ as functions of temperature. In our work, the solid-solid (aM/M) transition first appears for $N = 30$ at $T = 0.1$ in both quantities. Chain lengths $30 \leq N \leq 34$ offer the most discernible indication of the solid-solid (aM/M) transition and the liquid-solid transition existing in a single chain. As $N$ increases, the solid-solid transition becomes more defined, while the liquid-solid transition becomes broader, appearing to eventually overcome the solid-solid transition near $N = 42$. A shoulder still exists for many larger lengths in the $42 < N < 55$ range, but this is not easily analyzed within our current framework. The merger of these
Figure 7.8: Heat capacity (top) and the derivative of the squared radius of gyration (bottom) versus temperature for chain lengths between $13 \leq N \leq 55$. These chain lengths offer unique behavior in the liquid-solid regime, where for some chains two peaks can be seen in the heat capacity and a peak and shoulder in the radius of gyration. Each curve is shifted by an arbitrary amount for visual clarity. ($dT = 0.001$)
two transitions is an interesting, but also challenging topic. Cluster simulations [106, 105], as well as the recent homopolymer study [31, 32], have given details regarding behavior in this region and results presented here are in good agreement. Cluster studies have shown these two transitions to merge just below the $N = 55$ Mackay icosahedron [71], where the Mackay ground states become much more favorable compared to the anti-Mackay states as they approach $N = 55$. What these chain lengths indicate, along with the studies discussed above, is that the interplay between Mackay and anti-Mackay states can lead to significant thermodynamic behavior.

An important issue brought about by these effects is whether or not these types of transformations occur in larger chain lengths. The range of sizes between the two magic number $N = 55$ and 147 have shown similar Mackay and anti-Mackay structure in the outer shell of particular cluster sizes (as represented in Table 7.1) and Schnabel et al. [31, 32] have found similar behavior in this range (with a few exceptions). This becomes more interesting for the next magic number range $147 < N < 309$, where in Fig. 7.6 one might expect these effects to have become insignificant because of core-surface arguments. However, cluster studies still show signs of multiple transitions at low temperatures. In fact, one study [69] has shown two-peak behavior within the magic number $N = 309$. This implies that these effects may continue well-past the range of chain lengths considered in most recent studies. This could all be further complicated by transformations in larger clusters $> 1690$, where systems prefer decahedral geometries over icosahedral [31, 32, 109]. The results shown in Fig. 7.6 and Fig. 7.8 reveal that transition behavior in homopolymers is not as ideal as has been presented in previous studies. The coil-globule transition is still rather consistent, but the liquid-solid transition is far more susceptible to finite-size effects. Real polymer chains are much more complicated than the present model, and optimal packing near the liquid-solid transition will subsequently depend on chemical structure.

In Fig. 7.9 the predictions discussed in the methods section are considered, where Eq. 3.6 and 3.7 are applied to results (seen in Fig. 7.6) from the analysis of the specific heat and the
radius of gyration. The coil-globule transition behaves as is expected, and the fit is seen to coincide with the data for all chain lengths. In Tab. 7.2 fitting results are presented for the coil-globule transition using both specific heat and radius of gyration data, where the infinite chain prediction of the $\Theta$-temperature is $T_\Theta = 3.156 \pm 0.007$ for $C_V/N$ and $T_\Theta = 3.176 \pm 0.004$ for $R_g^2$. This fit was applied to chain lengths between $10 \leq N \leq 561$ and the results agree quite well with the prediction in Eq. 3.6. There are some fluctuations, especially at large $N$, but these fluctuations do not fall outside the bounds of statistical error. In comparing these fitting results with those seen in the work from Parsons and Williams, it is found that the estimate of $T_\Theta$ is significantly larger in results presented here. This is most likely due to this studies sampling of a broader energy range, where $E_{\text{max}}/N = 3.0$ rather than zero. It is also noted that when the shortest chains are excluded from the fitting procedure, the estimates
for $T_\Theta$ increase slightly, but systematically. In all cases the “best fit” for the $R_g^2$ data yield a slightly larger value of $T_\Theta$ than is obtained from the specific heat. Thus, the error estimates for the values in Tab. 7.2 may be underestimated due to possible, residual finite-size effects for the smallest chain lengths.

The analysis using Eq. 3.7 to describe the liquid-solid transition data is also shown in Fig. 7.9; however, finite size effects previously discussed must be considered. The variations with size below $N = 100$ makes it hard to include these results for extrapolation to the infinite-chain limit. In Fig. 7.6 the $N = 55$ specific heat peak is uncharacteristically high compared to the other chain lengths, as well as other magic numbers. This large peak is not novel and can be seen in many similar analyses of clusters [106, 105, 70]. Understanding this peak may take more analysis of the liquid-solid and solid-solid transitions, and in particular, how these two merge before the $N = 55$ chain length. Because of the finite-size effects discussed above, analysis is restricted to $100 \leq N \leq 561$, when applying the prediction of Eq. 3.7. The data for $N \geq 100$ appear to be rather smooth and magic number spikes decrease as $N$ increases (see points $N = 147$, $N = 309$, and 561). Results for this fit are presented in Tab. 7.2, where $T_{ls} = 0.790 \pm 0.007$ using the $C_V/N$ data and $T_{ls} = 0.79 \pm 0.01$ using the $R_g^2$ data. These results still offer perspective into previous issues in homopolymer studies. The first is that the coil-globule and liquid-solid transition remain separated in the infinite-chain limit. The second is that our estimate of the liquid-solid transition is significantly lower than that given by Parsons and Williams, providing further evidence that sampling a restricted energy range will yield imprecise results. However, this analysis is further complicated by finite-size effects that have been shown to occur for chain lengths greater than $N = 55$ and which are currently beyond this study’s ability to characterize.

### Conclusions

Results describing the general behavior of single flexible homopolymer chains have been obtained for a wide range of chain lengths. The application of the Wang-Landau algorithm
Table 7.2: Table of transition temperatures for both the liquid-solid ($T_{ls}$) and the coil-globule ($T_\theta$) transitions. These transitions have been analyzed using both specific heat and radius of gyration data.

<table>
<thead>
<tr>
<th>Fit function</th>
<th>Data</th>
<th>$T(N \to \infty)$</th>
<th>$a_1, a_2$</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T(N) - T_\theta = a_1/\sqrt{N} + a_2/N$</td>
<td>$C_V/N$</td>
<td>$3.156 \pm 0.007$</td>
<td>$-11.90, 13.98$</td>
<td>$10 \to 561$</td>
</tr>
<tr>
<td></td>
<td>$R_g^2$</td>
<td>$3.176 \pm 0.004$</td>
<td>$-11.56, 12.77$</td>
<td>$10 \to 500$</td>
</tr>
<tr>
<td>$T(N) - T_{ls} = a_1/N^{1/3}$</td>
<td>$C_V/N$</td>
<td>$0.790 \pm 0.007$</td>
<td>$-1.94$</td>
<td>$100 \to 561$</td>
</tr>
<tr>
<td></td>
<td>$R_g^2$</td>
<td>$0.79 \pm 0.01$</td>
<td>$-1.9$</td>
<td>$100 \to 500$</td>
</tr>
</tbody>
</table>

has proven to be an efficient and straightforward method for studying thermodynamic and structural properties of homopolymers. Nonetheless, the few parameters involved require a careful adjustment in order to yield an optimal balance between simulation accuracy and computational effort, e.g., the density of states must be obtained for a sufficiently broad range of energies. Unlike earlier studies of continuous models [34, 35], only two transitions are identified for similar chains, and as has been shown in studies of a bond-fluctuation model [38], these two transitions remain distinct in the limit of infinite size. We also see a significantly higher $T_\theta$ compared to these off-lattice studies [34, 35], most likely due to the broader sampling of energy. The liquid-solid transition is more difficult to analyze because of finite-size effects occurring for chain lengths $30 \leq N \leq 54$, where additional solid-solid transitions associated with transformations in the outer shell of a collapsed chain are visible in thermodynamic and structural quantities. Such effects have been seen in a very recent study of Lennard-Jones homopolymers [31, 32], with indications of such behavior in longer chain lengths. However, for $N \geq 100$ variations in behavior with chain length become smoother and we were able to estimate the infinite-chain transition temperature. Our value of $T_{ls}$ is
significantly lower than that found in previous off-lattice studies, which is again attributed
to energy range effects. Results also show that size dependent effects are prominent and do
not behave in a simple, monotonic manner as chain length increases. These finite-size effects
must be considered in future studies of single homopolymer chains.
Using two-dimensional (2D) Wang-Landau sampling, the density of states of a single homopolymer chain is obtained as a function of energy $E$ and flexibility $J_A$ for selected lengths. Conformational diagrams obtained from analysis of $g(E, J_A)$ reveal complex conformational behavior, showing a number of unique transitions, all of which are accompanied by a large number of finite-size effects. In addition, minimum energy configurations are also stored during these simulations, providing a useful tool for further analysis of conformational behavior. Although there are many complications, an effort is made to understand the behavior as function of chain length, such that qualitative predictions can be made for larger systems.

The minimum energy state behavior of semi-flexible chains has been studied in a number of works over recent years [43–47, 53, 24, 48–50], where particular emphasis has been put on the existence of structures with high persistence length, e.g., disks, toroids, and rods [44–46, 24, 48]. Low persistence length states have been labeled simply as globular in these studies, with no real investigation of the order in these states. The primary focus of these studies has been on how these structures are affected by temperature $T$ and flexibility $J_A$, and rough conformational diagrams have been presented for a number of chain lengths. These studies have typically been simulated with more traditional Monte Carlo methods, usually involving temperature dependent methods, meaning some features of the conformational behavior may have been overlooked.

With an implementation of the 2D Wang-Landau algorithm, the conformational behavior is mapped out for a number of chain-lengths ranging from $N = 10$ to 55. The goals of this
investigation can be summarized as follows: 1) to understand basic transition phenomena and their dependence on $T$ and $J_A$, 2) to investigate the minimum energy states associated with a broad range of $J_A$, namely, the maximally collapsed globule ($J_A = 0.0$) to the completely straight chain, and 3) to show the existence of Mackay / Anti-Mackay behavior in semiflexible systems. The latter is perhaps the most exciting of these goals, where previously this type of behavior has been overlooked or was impossible to investigate with techniques employed.

8.1 Wang-Landau parameters

In order to effectively study the conformational behavior of single homopolymer chains, two-dimensional Wang-Landau sampling has been employed using energy $E$ and flexibility $J_A$ as the sampling directions. These directions allow for the estimation of canonical ensemble properties for all values of flexibility, and further estimation of transition and minimum energy state behavior as a function of temperature $T$. Such a simulation requires a 2D grid of $E$ and $J_A$ bins, which is ultimately used to keep track of a 2D density of states $g(E, J_A)$. Effectively handling these functions requires certain modifications to the parameters associated with the Wang-Landau algorithm. In addition, little is known about minimum energy states (as well as maximum energy cutoffs) of all chains with a non-zero flexibility, where classes of these states have been identified, but with no real catalogue of values. Therefore, an adaptive method for sampling new states has been implemented.

The Wang-Landau algorithm proceeds as discussed in Chapter 5, only here, the 2D density of states $g(E, J_A)$ is placed into the acceptance probability given in Eq. 5.1. A principal issue with these simulations is that the range of $E$ and $J_A$ needed to cover the conformational space associated with $g(E, J_A)$ becomes quite large for $N$ considered to be moderate lengths in 1D simulations of the flexible-case. With this increased coverage, $g(E, J_A)$ becomes quite
large with respect to the total number of bins, e.g., $N = 55$ requires $\approx 2 \times 10^6$ bins\(^1\) to fully describe the conformational behavior, where the binning resolutions are typically $dE = 0.1$ and $dJ_A = 0.1$. A major obstacle associated with simulations of chain lengths whose density of states require such large numbers of data points is the reduction of the modification factor $f$. The standard approach of using a flatness can be highly inefficient, where one or two bins, particularly near minimum energies at a variety of $J_A$ values, can keep $f$ from being reduced, even when a large percentage of the DOS has already satisfied the flatness criterion. In order to allow $f$ to be modified in situations like this, an alternative criterion for reducing $f$ is used, which to the best knowledge of this author was first introduced by Shell et al. [84, 85]. This method checks that all bins have been sampled a certain number of times before reducing $f$. This condition is further relaxed in this work in that a small number of bins are left unsampled between iterations. The alternative method for reducing $f$ works quite well in protein and polymer systems, because typically these systems must fold and unfold a large number of times in order to sample all states (whether 1D or 2D). This folding dynamic often leads to evenly sampled histograms, even without a very strict flatness criterion.

Of equal importance to the above discussion on reducing the modification factor is the existence of a large number of unique minimum energy states. When using a standard flatness criterion, a single simulation might find a highly improbable state for some values of $E$ and $J_A$, particularly near minimum energy values. This single bin representing this state may keep the modification factor from being reduced, while all other bins have been sampled a large number of times. It is also mentioned that this is not merely a thought experiment, but instead is from actual observations of test and production simulations. The minimum energy states are highly unique when considered as a function of $J_A$, and in many cases can not be reached by moving from neighboring $J_A$ values. Instead, the chain must be unfolded and then refolded into these unique states, which even for random processes with a highly

\(^1\)In 1D simulations of $N = 55$, highly accurate results can be generated when considering only thousands of bins.
efficient move set can be quite challenging. Not only does this type of physical situation require a different way of handling the modification factor, but it also requires that some of these states be marked for exclusion when considering the need for a smooth function that can be used to calculate thermodynamics for number of temperatures.

The final difference between this 2D implementation, and the 1D implementation used in the study of flexible homopolymers, is an exploration feature that allows for new states to be added in the early stages of each simulation. This method is particularly important to the exploration of minimum energy states. In the case of fully-flexible homopolymers, an exploration feature is not necessary because the correlation with classical clusters allows for minimum energies to simply be looked up in tables. However, semi-flexible chains have no direct comparison, and therefore, the minimum energy is unknown for each value of $J_A$. In order to avoid preliminary simulations aimed at defining the minimum energy boundaries for each value of $J_A$, the simulation is allowed to explore new energies during the first 6 iterations of the Wang-Landau algorithm, i.e., $\ln[f] \approx 0.01$. After these early iterations, bins which have been sampled are marked, and no new bins are allowed to be added to the current estimation of $g(E, J_A)$. In addition, once these states have been marked and a clear minimum energy boundary has been defined, a small percentage of the lowest energy states are removed for each value of $J_A$. This step corresponds to the issues described above regarding the multitude of minimum energy states occurring in semi-flexible homopolymers. In these simulations, the percentage of states removed for each $J_A$ value typically corresponds to $\approx 0.1\%$ of the total explored energy range. The states removed do not dramatically affect thermodynamics, and such effects are checked during the analysis of thermodynamic data. The most dramatic effect regarding these low energy states actually has to do with CPU time, where order of magnitude increases can occur when attempting to include all minimum energy states in a simulation.

Results are generated using a 2D implementation of the Wang-Landau algorithm. Differences with the original implementation include: 1) reducing the modification factor when
each bin has been sampled a certain number of times (typically 50), 2) an exploration feature which is included in the first few iterations (typically 6), 3) a resizing of $g(E, J_A)$ once the exploration feature is turned off (always includes at least 99.8% of the explored density of states), 4) a small percentage of bins are left unsampled between iterations (differs during the exploration and non-exploration portions of each simulation, but is typically a small percentage of the number of $J_A$ values).

8.2 Specific heat

The two-dimensional density of states $g(E, J_A)$ provides access to the specific heat, which is calculated by using the analysis techniques provided in Chapter 6. Essentially, each value of $J_A$ provides a one-dimensional density of states that can be averaged and subsequently transformed into the canonical form. As is the general case for these types of simulations, the specific heat provides a standard starting point for understanding conformational and transition behavior. These functions can be both revealing, and remarkably complex, even for the relatively small systems studied here. In order to remove some of this complexity, results are presented for a small chain length, and then systematically increased. In this section, the two-dimensional specific heat functions are presented for $N = 15, 20, 25,$ and 30, along with highlights of expected conformational behavior as predicted by the knowledge of flexible homopolymers and the analysis of minimum energy configurations (MECs) found in each simulation. All results presented here are the average results of multiple independent simulations (typically, between 5 and 10), unless stated otherwise.

8.2.1 Understanding behavior in a small system, Specific heat of an N=15 length chain

First, the $N = 15$ length chain is analyzed so that a knowledge base is formed for a simple system. Fig. 8.1 shows the average (5 independent simulations) 2D specific heat as a function of $J_A$ and $T$, where this 3D function is plotted as a 2D color map to remove some of the
Figure 8.1: Specific heat as a function of temperature ($T$) and stiffness ($J_A$) for a chain length of $N = 15$ averaged over five independent simulations. Various transformations can be seen as indicated by peaks in this function, where peaks correspond to the brightest color (yellow). The error in this plot will be presented in the broader context of the largest simulated chain length later in this chapter.

complexity. The color key to the right of the plot represents the height of $C_V(J_A, T)/N$, while the x-axis corresponds to $J_A$ and the y-axis to $T$. It is immediately recognized that these systems are characterized by a more complex conformational space compared to fully flexible homopolymers. In order to classify transitions in terms of their conformational transformations, the MECs (minimum energy configuration) are also stored during each simulation. At least in terms of the low temperature regions, this allows for conclusions to be made regarding certain regions in the conformational space. Error in this plot is discussed in a broader context as it relates to the $N = 30$ length chain.
In order to further simplify the 2D specific heat, analysis of peak and shoulder positions is performed, and these are plotted in Fig. 8.2, allowing for a clearer depiction of regions in conformational space. These extrema are found by observing the behavior of the first and second derivative of $C_V(J_A, T)/N$, and classifying them accordingly. In this plot, circles represent peaks or maxima defined by the first derivative crossing zero, while the squares represent significant shoulders or “second-order” like transitions defined by peaks in the second-derivative. The peaks (circles) can be likened to “first-order” like transition such as the liquid-solid or solid-solid transitions, while the shoulders (squares) often belong to second-order like transitions similar in nature to the coil-globule transitions. However, it is repeated that these are not “real” transitions in the thermodynamic sense, but rather strong indications of conformational changes. All of the points in Fig. 8.2 can be traced out and interpreted as conformational boundaries, creating regions of unique conformational behavior.

In Fig. 8.2, conformational regions are labeled A through F for the $N = 15$ length chain. Region A represents the random-coil states seen in the fully flexible homopolymer case. The region $A^*$ is similar to A, but as $J_A$ increases, these states become straight chains whose behavior is more predictable conformational behavior than the $J_A = 0.0$ case of random coils. Region B is inferred to be the liquid-globular region, again using knowledge from the flexible case. The remaining regions represent conformations classified from MEC data: C is a solid-globular region in which icosahedral Mackay and Anti-Mackay states can be found, $D_3$ and $D_2$ provide solid-rod like configurations made up of $m$ parallel segments (of course still connected), $D_3$ being three and $D_2$ being two packed segments. Understanding how these conformations are affected by temperature, hence more effectively defining these boundaries, is done using structural quantities later in this chapter. Another important feature are the peaks occurring for low $J_A$ and low $T$ in region C. Some peaks are due to transformations as simple as bond rearrangements with no real geometric change of monomer positions, while others correspond to significant changes in geometrical order, such as Mackay to Anti-
Figure 8.2: Conformational diagram for $N = 15$ semi-flexible chain: peaks and shoulders found in the specific heat $C_V(J_A,T)/N$, with minimum energy configurations found at various $J_A$. Letter in this diagram represent conformational regions, separated by first-order like (grey circles) or second-order like (red squares) transformations, where these are defined as A - random-coil, B - liquid, C - solid-globular, (blue monomers surrounding a red central monomer represent an icosahedral core), D - rods with segment number. Dashed lines represent guides for the eye and are not any type of extrapolation of data.
Mackay. These details will be analyzed in a broader context at the end of this chapter, and for now this region holds a generic title of solid-globules.

One critique of these data is that some points seem too irregular, or not connected to a particular boundary. Unfortunately, there is currently no analysis scheme for differentiating between noise and real transitions, therefore, errant points do occasionally exist. However, some features should not be mistaken for errant data, such as the breaks between lines near regions A, A*, and B. For example, it is expected (and will be shown for larger sizes) that the coil-globule transition line extending from approximately $J_A = 0.0$ to 2.0 would be smooth and eventually connect with the boundary between regions A* and D. But for such small systems, a change in conformation that removes even just a few interactions can drastically affect thermodynamic behavior. This implies that the second-order (red squares) transition occurring between region A and B is still the coil-globule transition, even with this break at $J_A = 1.2$ and $T = 0.7$. Understanding these types of breaks for such small systems can be realized by observing the number of interacting pairs. In the remainder of this work, the liquid-globular region is particularly hard to analyze, as it becomes a melting pot of so many unique conformations.

The analysis of the specific heat in such a small system provides understanding that will be extended to larger systems. Although behavior becomes more complex as $N$ increases, many features remain, or change only slightly. The following sections will rely heavily on the understanding of this small chain.

8.2.2 Behavior in larger systems, Specific heat for $N=20$, 25, 30

In Fig. 8.3, two-dimensional specific heat $C_V(J_A, T)/N$ color maps are plotted for $N = 20$ (top left), 25 (bottom left), and 30 (top right), along with the error associated with the $N = 30$ chain (bottom right, gray-blue color scale). Each color map represents the averaged results of at least five simulations. This compact format allows the visualization of the overall curvature of each system, before peaks are analyzed and plotted separately. The
Figure 8.3: Color maps representing the 2D specific heat $C_V(J_A, T)/N$ for chain lengths $N = 20$ (top left), $N = 25$ (bottom left), $N = 30$ (top right), and the error associated with $N = 30$ (bottom right) using a different color scale. All of these plots give insight into the increasing complexity of thermodynamic behavior with $N$. The error for $N = 30$ also shows that accurate results can be generated for these systems.
most prominent feature of this collection of plots is the increasing complexity with $N$, where a number of distinct transition lines exist in the low temperature regions. The error plotted for the $N = 30$ system of Fig. 8.3 (bottom right) is representative of all data sets, and since $N = 30$ has the largest errors associated with $C_V(J_A,T)/N$, the errors are not plotted for the smaller chains (including the previously discussed $N = 15$ length chain). It is noted that the region with the highest error occurs in the solid-globular region, which is typical of all chain lengths. In this region, conformations are of highest densities and persistence length while still maintaining solid-globular form, and in addition, this region of error represents the border of solid-globule and rod-like bundles. These conditions make sampling states in this region particularly difficult. The ability to plot all interesting features, along with their error, is a principal challenge in reporting results in these types of 2D simulations.

In the following, each chain length from Fig. 8.3 is analyzed in detail as was done for the case of $N = 15$. It is noted that an attempt was made to put all of these data sets into a single figure, so that scaling relations could be checked. However, no straight-forward method for scaling could be made, demonstrating the dominant influence of finite-size effects. Hence, plots are again presented for individual chains.

8.2.3 $N = 20$

In Fig. 8.4, extrema from the specific heat for the $N = 20$ length chain are plotted, allowing for the clear representation of different conformational regions. Regions A and $A^*$ again represent unwound states, random coils and straight chains, respectively, with $B$ again being the liquid phase. Region C again represent solid-globular conformations, where Mackay and Anti-Mackay states are abundant, particularly, a double-icosahedral geometry which can be directly associated with MEC of an $N = 19$ fully flexible homopolymer. This Anti-Mackay state is found throughout the MECs of a number of chains, and is perhaps one of the more efficient geometries due to its similar nature to rod-like configurations. This can be further clarified by noticing that the solid-globule to solid-rod transformation in $J_A$ occurs at a much
higher value of $J_A$ compared to $N = 15$. The rod-like states are again represented by $D_3$ and $D_2$. It is noted that a twist, promoting the possibility of helix-like behavior, is now apparent for the $D_3$ MECs. This twist does remain in many of the larger systems, and will be analyzed for these larger systems.

8.2.4 $N = 25$

Fig. 8.5 shows extrema from the specific heat for the $N = 25$ length chain. Regions A, A*, and B, are again comparable to the previously studied chain lengths, representing random coils, straight chains, and liquid-like conformations, respectively. However, it is observed that the MEC regions have undergone some changes. Region C again represent solid-globular conformations (blue and red monomers represent icosahedral cores), but with many more complexities as $T$ is increased to the liquid region B. This can be somewhat understood by the Anti-Mackay MEC depicted at the bottom of the figure, where three icosahedral cores make up a single configuration. Further analysis of region C shows a strong line of peaks starting from $\approx J_A = 0.8, T = 0.15$, and extending down to $\approx J_A = 3.0, T = 0.05$. Through analysis of the MECs in region C, this line represents an Anti-Mackay state containing three icosahedral cores, where as $J_A$ is increased beyond $\approx 3.0$, the MEC becomes a solid-globule with no dominant geometrical order.

Another important feature of Fig. 8.5 is the existence of rod-bundle states of higher order, namely the $D_4$ region. This gives the first impression that rod-like states will exist with larger segment numbers as $N$ increases. Rod-like states with three $D_3$ and two $D_2$ segments are again an important part of the conformational diagram, and it is noted that a twist remains in both the $D_4$ and $D_3$ MECs. This twist does remain in many of the larger systems, and will be shown for these sizes later in this chapter.

In addition to the above features, Fig. 8.5 provides first insight into a new class of conformations, referred to here as toroidal states, and represented by E in the conformational diagrams. Some may argue that the hair-pin state depicted at the bottom of this figure is
Figure 8.4: Conformational diagram for $N = 20$ semi-flexible chain: peaks and shoulders found in the specific heat $C_V(J_A, T)/N$, with minimum energy configurations found at various $J_A$. The letters in the $J_A$ versus $T$ diagram represent C - solid-globules (blue and red monomers represent an icosahedral core), $D_m$ - rod-bundles with $m$ number of segments. Regions A and A* are again unwound states, and B is the liquid region, where complexities arise at the boundaries of globular and rod-like liquids. Dashed lines represent guides for the eye and are not any type of extrapolation of data.
Figure 8.5: Conformational diagram for $N = 25$ semi-flexible chain: peaks and shoulders found in the specific heat $C_V(J_A, T)/N$, with minimum energy configurations found at various $J_A$. The letters found in the $J_A$ versus $T$ diagram again represent C - solid-globules (blue and red monomers represent icosahedral cores), D - rods with number of segments, and E - first representation of disks toroids in the form of a hairpin conformation. Regions A and A* are again unwound states, where a MEC for $J_A = 14.0$ is drawn to the right of the diagram. B is again the liquid state, but for this length considerable transformation can be seen due to the mixing of globular and rod-like states. Dashed lines represent guides for the eye and are not any type of extrapolation of data.
closer to rod-like configurations, than to a torus, but this behavior has a clear relationship with the states found in larger systems, and is therefore categorized with these states. In particular, one can imagine that as chain-length grows, states may exist where the circular region at the top of the hair-pin becomes significantly large, making a stronger argument for its categorization. Regardless of the classification, the existence of this configuration for such a simple system is quite motivating. These types of structures are found in real protein systems [110], and in addition, similar configurations are meaningful structures in short, single-stranded RNA [111].

8.2.5 $N = 30$

As the final data set for this section, Fig. 8.6 shows extrema from the specific heat for the $N = 30$ length chain. Regions A, A*, and B, are again comparable to the previously studied chain lengths, representing random coils, straight chains, and liquid-like conformations, respectively. It is noted that region B has become somewhat more well-defined compared to the $N = 25$ case. However, the regions leading to this phase with increasing $T$ have become significantly complicated. Region C is riddled with peaks and shoulders with no clear recognizable boundaries as could be inferred for the $N = 25$ (change between Anti-Mackay states and solids with unspecific geometries). These peaks and shoulders were initially correlated with possible error in the simulation, but Fig. 8.3 shows no tendency toward large statistical errors. Interestingly, there is a compelling explanation that can be found by analyzing MECS in this region. At the bottom of this figure, a unique conformation containing a “triangular-bipyramid” is drawn for a stiffness of $J_A = 1.0$. To this author’s knowledge, such a configuration has not been found in such small homopolymer systems, and energetic competition between the standard Mackay / Anti-Mackay and other highly ordered (and symmetric) states could cause serious disruptions in thermodynamic behavior. This geometrical competition is a compelling example of the uniqueness of the MECS found in these systems, and because of the complexity, these types of issues will be further analyzed later.
in this chapter. Such arguments lead to a better understanding of the error for \( N = 30 \) shown in Fig. 8.3 being largest in region C.

In addition to the interesting physical behavior described above, new features again occur in the rod-like region for this chain length. Fig. 8.6 depicts the existence of rod-bundle states of even higher order compared to \( N = 25 \), namely the \( D_5 \) region. This \( D_5 \) region is smaller than its counterparts, but does have some clear boundaries that extend to the liquid-phase (B). An interesting part of the rod-like behavior in this system is that all \( D_m \) regions appear to have clear boundaries that terminate near the liquid-phase. (This was hard to determine in the \( N = 25 \) case, due to the difficulty of distinguishing between rod and toroidal regions.) \( D_3 \) appears to terminate at the \( D_4 \) boundary, but more interestingly, has a undefined region between it and \( A^* \). Behavior in this unlabeled region can be further understood by observing that no \( D_2 \) can be found in the analysis of MECs. One might expect that this small (unlabeled) region is simply the \( D_2 \) region, and toroidal states have overcome the MECs for these values of flexibility. However, further analysis is needed in order to properly distinguish this region from the surrounding \( D_3, A^*, \) and toroidal regions, \( E^*_{H} \) and \( E_{L} \).

The toroidal region for \( N = 30 \) can be broken into two categories, \( E_{L} \) representing pure toroidal like geometries (simple spools, or rings), while \( E^*_{H} \) represents hairpin states. An interesting feature of these two regions is that their boundaries appear to merge as \( T \) decreases and \( J_A \) increases. This is backed up by analysis of MEC states, where hairpin states were found in a small percentage of the total number of runs, and the minimum energy hairpin states found by these simulations were often slightly higher in energy than the toroidal configurations. This would mean that in these simulations the merger of these boundaries had not occurred, and hence the hairpin states were still visible.
Figure 8.6: Peaks from the specific heat $C_V(J_A,T)/N$ for $N = 30$, with minimum energy configurations found at various $J_A$. The letters placed in the $J_A$ versus $T$ diagram represent C - solid-globules (blue and red monomers represent an icosahedral core), $D_m$ - rods with $m$ number of segments, $E_L$ - toroid and disks, and $E_H^*$ - hair-pins and other elongated conformations with circular symmetry. Regions A and $A^*$ are again unwound states, and B is again the liquid state, which has several visible transformations making this region hard to identify. Dashed lines represent guides for the eye and are not any type of extrapolation of data.
8.3 Structural Quantities for $N = 30$

In this section, various structural quantities are used to further interpret the specific heat data presented in the previous sections. Results presented here are restricted to $N = 30$, a system that contains all of the interesting characteristics found in the smaller systems, along with challenges of describing conformational regions which can not be done with MECs alone. Structural quantities are calculated through the application of a multicanonical production run to the $g(E)$ functions for $N = 30$, which are the same data used to calculate $C_V(J_A, T)/N$ in Fig. 8.3.

The following results will show data for four structural quantities: squared-radius of gyration $\langle R^2 \rangle_{J_A,T}$, end-to-end distance $\langle R_{\text{end}} \rangle_{J_A,T}$, first Khalatur parameter $\langle K_1 \rangle_{J_A,T}$, and second Khalatur parameter $\langle K_2 \rangle_{J_A,T}$. The end-to-end distance is perhaps the most straightforward of these parameters, whose definition is simply the distance between the end monomers. The other three quantities require a more rigorous definition, where they are estimated using the eigenvalues of the gyration tensor, $S_{mn} = \frac{1}{N} \sum_{i=1}^{N} r^i_m r^i_n$, where $m, n$ are 1, 2, or 3, representing Cartesian coordinates of the center of mass positions, i.e., $\sum_{i=1}^{N} r^i = 0$. The gyration tensor is a real-symmetric $3 \times 3$ matrix that can be diagonalized, with the subsequent eigenvalues representing the principal moments $\lambda_1^2 \leq \lambda_2^2 \leq \lambda_3^2$. In this work, the calculation of these eigenvalues is performed using the GNU Scientific Library [100]. The following definitions of structural quantities can be written in terms of these eigenvalues: squared radius of gyration $R_g^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$, first Khalatur parameter $K_1 = (\lambda_1^2 + \lambda_2^2)/(\lambda_2^2 + \lambda_3^2)$, and second Khalatur parameter $K_2 = (\lambda_1^2 + \lambda_3^2)/(\lambda_2^2 + \lambda_3^2)$. The Khalatur parameters represent a means for understanding broader geometrical properties, where a perfect sphere is characterized by $K_1 = K_2 = 1$, a perfect rod by $K_1 = 0$, $K_2 = 1$, and a perfect disk by $K_1 = K_2 = 1/2$. These values represent the ideal case, and it is noted that for the small systems studied here, the concern is often that these criterion are fulfilled within a certain percentage of the ideal case.

\footnote{This analysis relies heavily on work found here [34, 24].}
Figure 8.7: Color maps representing 2D structural quantities for $N = 30$, including radius of gyration $R_g^2$, end-to-end distance $R_{\text{end}}$, and the Khalatur parameters $K_1$ and $K_2$. Regions are labeled for $R_g^2$ whose meanings are the same as found in Fig. 8.6, $C$ - solid-globules, $D_m$ rod-like bundles, $E_L$ - toroids, $E_H$ - hairpins. Since these regions are the same for each color map, labels are left off the others for clarity. The error for these data is of the same order as that presented for the $N = 30$ length chain in Fig. 8.3.
Figure 8.7 contains four color maps representing the two-dimensional canonical representation of the previously discussed structural quantities for $N = 30$: (top left) radius of gyration $\langle R_g^2 \rangle_{J_A,T}$, (bottom left) end-to-end distance $\langle R_{\text{end}} \rangle_{J_A,T}$, first Khalatur parameter $\langle K_1 \rangle_{J_A,T}$, and second Khalatur parameter $\langle K_2 \rangle_{J_A,T}$. These data are used to make a final distinction of the regions previously defined by MECs only in Fig. 8.6. These data reinforce those findings at very low $T$, but also shows how these states extend to higher $T$. The first plot of $\langle R_g^2 \rangle_{J_A,T}$ (top left) gives an overall view of the behavior as it is related to size. The most compressed states are solid-globules occurring for low values of $J_A$ (region C - dark purple). As stiffness is increased, new zones are formed, each having distinguishable values of $\langle R_g^2 \rangle_{J_A,T}$, namely D - rod-like bundles (light purple), and E - toroidal states (light purple, but hard to distinguish from rods). However, this quantity only indicates an overall change in size, and therefore, gives no substantial indication as to what types of changes are occurring, e.g., changes in rod-like bundles and toroidal states. The easily calculable end-to-end distance $\langle R_{\text{end}} \rangle_{J_A,T}$ (bottom left) gives a good indication of changes in rod-like bundles, where different folds of these rods correlate well with where the two ends will exist, e.g., three rod bundles have two folding points and the ends at the extrema of the overall bundle, four rod bundles have three folding points and the ends can only exist at one extrema of the overall bundle.

The remaining data in Fig. 8.7 are for the two Khalatur parameters $\langle K_1 \rangle_{J_A,T}$ (top right) and $\langle K_1 \rangle_{J_A,T}$ (bottom right). Clear fluctuations in these parameters are apparent, but in order to gain a full understanding, their values must be compared in order to understand the difference between ideal spherical ($K_1 = K_2 = 1$), ideal rod-like ($K_1 = 0, K_2 = 1$), and ideal disk-like ($K_1 = K_2 = 1/2$) states. Again, it is emphasized that these values represent the ideal case and the small sizes here only approximate these conditions. First, the color scales of each plot show that in region C approximately spherical states exist, where $K_1$ and $K_2$ are both $\approx 0.9$ which represents that the chain attempts to form highly symmetric, compressed states. More importantly, the expected rod like conformations have $K_1 \leq 0.25,$
while $K_2 \geq 0.95$. The ideal rod condition is best satisfied for $D_3$, and as the number of rod-like segments increases, this condition is less satisfied. The results reveal that rod-like states indeed occur, but differentiating between region C and higher order $D_m$ regions can be difficult.

Of further importance is the ability for the Khalatur parameters to distinguish between regions $E_L$ (toroid) and $E_H$ (hairpin). The condition for an ideal disk is again $K_1 = K_2 = 1/2$ and for region $E_L$ these parameters are $K_1 \approx 0.4$ and $K_2 \approx 0.6$, showing a clear indication of a toroidal state. This was expected from observations of MECs in this region, but the reader is reminded that a small number of hairpin states were also found in this region, and an assumption was made as to the existence of these states in region $E_H$. The Khalatur parameters for region $E_H$ are $K_1 \approx 0.0$ and $K_2 \approx 0.95$, meaning states in this region are closer to rod-like states. With the combination of MEC information and these results, it is more definitively stated that this region represents hairpin states. A remaining question regarding this region is whether or not $D_2$ states can exist along with these hairpins. Currently, these structural quantities give no direct indication of a $D_2$. An attempt has been made using the total number of interacting pairs, which should be slightly different for these states, but results were inconclusive. Studies are planned that will address this issue directly.

8.4 Mapping behavior in minimum energy configurations

In this section, a classification of the minimum energy configuration behavior (MEC) is presented by plotting the minimum energy $E_{min}/N$ versus $J_A$ for a number of chain lengths $N$. Figure 8.8 represents a diagram of conformational behavior for a variety of chain lengths, including various transformations occurring at various values of $J_A$. An attempt to correlate these transformations among different chain lengths leads to a broader understanding of MEC behavior as it depends on $N$. Rod-like bundles, previously labeled by $D_m$, are particularly emphasized.
Figure 8.8: Diagram of conformational behavior for a variety of chain lengths as it is related to $J_A$, $E_{\text{min}}/N$, and structural fluctuations of MECs. Solid lines (color) represent the average $E_{\text{min}}/N$ values found in multiple simulations for a given $N$. Symbols of the same color appearing on these lines represent fluctuations in MECs (e.g., change from solid to a rod, or change within rods, $D_3$ to $D_4$), which are calculated by measuring fluctuations in the radius of gyration of these states as a function of $J_A$. Particularly, rod-like bundles have been distinguished, and similar transformations connected using small dashed lines, where regions are labeled using previously employed $D_m$ labeling. $D_2$ through $D_5$ are consistent with previous discussions, but this pattern is upset for $N \geq 35$, where a jump to $D_7$ (hcp-packing) occurs, with only a single instance of $D_6$ (red pattern) occurring for $N = 45$. Various states in the solid-globule and toroidal regions were not distinguished, primarily because the radius of gyration is not an appropriate structural quantity for distinguishing between states in these classes. In addition, open symbols represent data where this fluctuation analysis begins to become less accurate due to the limited number of MEC data collected for chains $N \geq 40$. Error in $E_{\text{min}}$ is smaller than the line-width, and for the symbols indicating fluctuations, error in $J_A$ is on the order of the symbol size. Dashed lines with arrowheads are visual indicators of possible behavior for $N > 55$, and are not any formal extrapolation.
The first aspect discussed in Figure 8.8 are the solid lines (color) indicated in the legend, which are the averaged $E_{\text{min}}/N$ values found in multiple simulations for a given $N$. As stated previously, these simulations are free to explore at any energy, and it is quite impressive that multiple independent simulations find the same results with such accuracy. The error in $E_{\text{min}}/N$ is smaller than the line-width, and $\mathcal{J}_A$ has a resolution of 0.1 for all $N$.

Another feature in Fig. 8.8 corresponds to the symbols of the same color appearing on the $E_{\text{min}}/N$ lines mentioned in the previous paragraph. These symbols are transformations in MECs (e.g., change from solid to a rod, or change within rods, $D_3$ to $D_4$), which are calculated by measuring fluctuations in the radius of gyration of these states as a function of $\mathcal{J}_A$. MECs are stored for each value of $\mathcal{J}_A$ in all simulations, and the configurations can be quickly analyzed using some external program to search for transformations as $\mathcal{J}_A$ is varied. Signals indicating a transformation are usually quite abrupt and can be measured in a number of ways. In Figure 8.8, rod-like bundles have been distinguished, where conformational regions are labeled using previously employed $D_m$ labeling. These regions are separated by small dashed line that connect transformations existing in multiple chains, e.g., all points representing changes from $D_3$ to $D_4$ conformations are connected. Once boundaries between these $D_m$ regions were drawn, these regions were shaded for clarity. $D_2$ through $D_5$ are consistent with previous discussions chains $N \leq 30$, but this pattern is upset for $N \geq 35$, where a jump to $D_7$ (hexagonal-close packed) occurs, with only a small patch of $D_6$ states (red shading - see legend) occurring for $N = 45$. This jump is not surprising in such a small system prone to finite-size effects, and in addition, assumptions about packing efficiency in such states can easily be made to justify such configurations. In Fig. 8.9, rod-like conformations of variable segment length are presented for $N = 45$, where these visualizations represent typical rod-like states found in conformational regions $D_7$, $D_6$, and $D_5$. The hexagonal-close packed (hcp) geometry found in the $D_7$ region is of particular interest, where a twist that has existed in all rod-like bundles is no longer visible. In addition, this type of hcp configuration
Figure 8.9: Minimum energy conformations found in the rod-like regions for the $N = 45$ length chain. These states represent typical conformations found in conformational regions $D_7$, $D_6$, and $D_5$. $D_7$ is of particular interest, representing a hexagonal-close packed structure.

appears to be particularly efficient for the $N = 55$ length, where it occurs for the majority of $J_A$ values in the rod-like region of this chain length.

The final features in Fig. 8.8 correspond to various states in the solid-globule and toroidal regions. Specific behaviors were not distinguished in these regions, primarily because the radius of gyration is not an optimal structural quantity for distinguishing between states in these classes. In the solid-globular region, geometry specific structural quantities must be introduced in order to distinguish between subtle changes in packing. In order for this to be effective, a broader classification scheme must be introduced, such that occurrences of unique geometries, such as the triangular-bipyramid in $N = 25$, can be explained along with more standard cases, such as Mackay / Anti-Mackay behavior.

In addition, open symbols represent data where this fluctuation analysis begins to become less accurate due to the limited number of MEC data collected for $N \geq 40$ chain lengths\(^3\). Currently, only one configuration is saved for each $J_A$ value, and properties of the confor-

\(^3\)It is noted that a complete thermodynamic and structural analysis of chains $N \geq 40$ has not been performed as was done for $N = 30$ previously in this chapter. This was also a contributing factor in choosing open symbols for these transitions.
mations are then averaged between five to ten simulations. For larger chains, it was more
difficult to hone in on the transformations as a function of $J_A$, and slightly larger errors
occurred. This identified a need for future simulations to save more MEC configurations
for each $J_A$ value. One such implementation would be to save these configurations for each
iteration of the Wang-Landau method, multiplying the total number of MECs by a factor of
$\approx 20$. These concerns are driven more by predictions of what kind of data will be needed for
chains larger than $N = 55$. The MEC data that is available and plotted with open symbols
in Fig. 8.8 has error in the position of these transformations (which would appear as x-axis
error bar for $J_A$) on the order of the symbol size.

Finally, dashed lines with arrowheads are visual indicators of possible behavior for $N > 55$, and are not any formal extrapolation. They rather reveal the intuition of the author as to
what will be found as $N$ is increased further. One interpretation of these data is that perhaps
the toroidal region eventually takes over all minimum energy states, however, the literature
for similar models [44, 24, 50] suggests rod-like configurations will play an important role in
system sizes comparable to those studied in the fully-flexible case ($N \leq 561$).

8.4.1 SOLID-GLOBULAR STATES AND ICOSAHEDRAL ORDERING

The solid-globular region represents collapsed states ranging from $J_A = 0.0$ to $\approx 5.0$. Con-
formations within this region can take on many geometries, where geometries identified in
this work include icosahedral (Mackay / Anti-Mackay), tetrahedral, hexagonal close packing,
and others. The classification of these geometries is a difficult task facing future simulations
of these systems, as this behavior is highly dependent on $N$. Furthermore, the analysis per-
formed here concentrates only on MEC data ($T \to 0$), meaning for higher $T$ in this region,
“solid-solid” like transitions could exist that make dramatic changes to geometrical order.
Regardless of these concerns, the knowledge of Mackay and Anti-Mackay packing in flexible
homopolymers allows us to make some generalizations regarding the behavior in semi-flexible
systems. It is noted that this type of packing has been realized in all chain lengths simulated in this work.

To more fully understand behavior in the solid-globular region, MECs are analyzed by searching for the existence of icosahedral cores. As discussed for the flexible homopolymer case, and icosahedral core is made up of 13 monomers, where one monomer resides at the center of the core and 12 neighboring monomers make up a shell. In order for this collection of monomers to be icosahedral, these 12 monomers must arrange accordingly to produce 30 unique interactions (not including those to the central monomer), all of which represent the edges of an icosahedron. In order to measure for these cores, first monomers possessing 12 neighbors are labeled, where two monomers are considered neighbors if the distance between them is $d_i < 1.2$ (reminding that the equilibrium distance between two bonded monomers is $d_0 = 1.0$, and the bond potential diverges for $d_0 + 0.2$). Once these “potential” central monomers are labeled, all neighboring interactions are counted within each respective 12 monomer shell, again, with the condition $d_i < 1.2$. This distance represents the maximum bond length, and it is noted that many of the cores found here are “approximate”. In order to effectively identify these cores, one must leave room for the distances to fluctuate. Finally, if the total number of interactions within the outer shell (edges) equals 30, this group of monomers is labeled as an icosahedral core, and re-colored such that a central monomer is red, and the surrounding shell is blue.

Figure 8.10 presents semi-flexible homopolymer conformations for stiffness values $J_A = 1.0$ and $J_A = 4.0$, for a number of chain lengths $N$. Single conformations are presented for $J_A = 1.0$, while three different visualizations are shown for the $J_A = 4.0$ case: 1) atomic representation of monomers, 2) tubular representation of monomers and bonds, and 3) contact maps showing helical and anti-helical behavior. The first representation allows for the close comparison with behavior in flexible homopolymers, while the second and third make a better comparison with structures typically identified in semi-flexible systems. Looping structure are common in these systems, but are often not analyzed with the intent of identi-
Figure 8.10: Diagram of solid-globular states found in chain lengths $N = 30, 35, 45, \text{ and } 55$, for stiffness values of $J_A = 1.0$ and 4.0. These states give a rough estimation of behavior in the solid-globular region, showing the existence of Mackay / Anti-Mackay states, as well as fcc-like states for $N = 30$. IC is the number of icosahedral-cores in each configuration (blue monomers surrounding a red central monomer represent an icosahedral core). The $J_A = 4.0$ columns represent three ways of visualizing a single configuration: 1) atomic monomers, 2) tubular representation of bond and monomer, and 3) a contact map plotting interacting monomers. The * implies that these states only have multiple icosahedral cores, and that the outer layer (silver monomers) has not been checked rigorously. For $N = 30$, all monomers making up a non-icosahedral cell are colored red.
fying higher-order geometries. These three visualizations show that even configurations with multiple windings (helical-like behavior) can still find these energetically favorable states. 

$N = 55$ is perhaps the most prominent feature of this figure, where the Mackay’s icosahedron is the MEC for all $J_A$ values in the solid-globular region. Although, $N = 30$ is also an interesting case where tetrahedral cores exist, $J_A = 1.0$ leads to a triangular-bipyramid core, and $J_A = 4.0$ produces a single tetrahedral core (half of a triangular-bipyramid). Data for $N = 35$ and $N = 45$ point toward a consistent behavior in that low values of $J_A$ promote Mackay states identical to those in the flexible case, while higher $J_A$ values produce Anti-Mackay states made up of multiple icosahedral cores (typically between 2 and 4). The most impressive feature of this figure is that highly wound states are able to find geometries associated with classical-clusters and fully-flexible homopolymers.

In addition, contact maps corresponding to conformations of high stiffness (tubular conformations) are plotted in Fig. 8.10, which show detail regarding helical and anti-helical sub-structures. The term helical is used in the most general sense, simply highlighting the coiled (or looped) sub-structure within each conformation. Each point in a contact map represents the monomer number of two interacting monomers, e.g., the line of positive slope extending from the origin to $N, N$ in each plot corresponds to bonded interactions. Underneath these bonded interaction points, lines of positive slope represent helical sub-structure, and lines of negative slope represent anti-helical sub-structure. Contact maps provide a useful tool in identifying highly wound structures, and here they show that the packing within these solid-globules is very different compared to the fully-flexible case.

### 8.4.2 Toroidal states

As was seen previously in Fig. 8.8, the toroidal region makes up a significant portion of the conformational space. As indicated by the symbols in this figure, a number of conformational changes takes place within the toroidal region. The more interesting transitions are those seen in the study of $N = 30$, where at high $J_A$ there seems to be a small region where
toroids \((E_L)\) and hairpins \((E_H)\) compete. As was seen in the case of \(N = 30\), these two states exist at high \(J_A\), but are separated by a boundary that splits off from the \(D_3\) boundary. Preliminary results for \(N > 30\) have also shown this type of behavior, and similar boundary between these states is expected. Unfortunately, this means that MEC data for hairpins is not available, and that configurations will have to be stored separately for this conformational region. However, the transformations indicated by symbols in Fig. 8.8 mean that significant changes are occurring in the \(E_L\) toroidal states. Unfortunately, a detailed thermodynamic and structural analysis of these changes has not been performed, but the available MEC data has been analyzed.

In Fig. 8.11, toroidal conformations (region E) for an \(N = 55\) length chain are presented. The primary observation is that as stiffness increases, conformations within the toroidal region become less and less wound, meaning the inner radius of the toroid becomes larger. In addition, the states drawn in Fig. 8.11 represent MECs that exist between transition points found on the \(N = 55\) minimum energy line plotted in Fig. 8.8.

As stated above, a thermodynamic and structural analysis of these states will be necessary in order to see if the transformations occurring in the MEC data have some more well-defined transition properties. Toroidal transformations occur more dramatically for chain
lengths $N > 35$; therefore, it is difficult to know what conformations will occur. In addition, previous studies of the toroidal region in semi-flexible systems have not focused on these properties for the chain lengths considered here, making a prediction of behavior even more difficult.

8.5 Conclusions

Results describing the conformational behavior of semi-flexible homopolymers have been presented. Wang-Landau sampling of the multi-dimensional density of states $g(E, J_A)$ provides an efficient way of describing the conformational and transition behavior of a number of chain lengths. Using an array of analysis techniques, further understanding of conformational regions has been gained. Semi-flexible chains are significantly more complex compared to their fully-flexible counterparts studied in Chapter 7. Regardless of this complexity, the data presented here provide a tool set that makes understanding this behavior attainable.

The conformational results presented agree quite well with studies of similar semi-flexible systems using the bond-fluctuation model [45, 46, 24]. These works have shown a tendency for semi-flexible systems to collapse into three categories of low-temperature conformations, namely, solid-globular, rod-like, and toroidal. These studies have focused on a few chain lengths, some quite close to the results presented in this dissertation ($N = 40$ and 80, as well as some that are quite larger ($N = 128$ and 256). The results presented in this chapter not only agree with this earlier work, but also present a greater level of detail within each conformational region. It is expected that this work will provide greater insight into the geometric packing occurring in various conformational regions, and in addition, should provide a basis from which to make predictions for larger systems where packing details have not been studied.

One of the most impressive features of these results is the existence of high-order geometrical packing in the solid-globular region. Although a more thorough analysis of minimum energy configurations is needed, along with analysis of low-temperature transitions between
unique geometries, this work provides a starting point from which future investigations may start. Regardless, the discovery of icosahedral Mackay and Anti-Mackay states for such broad values of stiffness within the solid-globular region should lead to changes in the way future investigations study these collapsed states.

8.5.1 CPU time in multi-dimensional Wang-Landau sampling

A major issue facing 2D sampling in these systems are the large CPU times. In the case of the $N = 55$ chain length, single simulations are unable to finish in under 30 days of continuous processing on a modern CPU. Initial work on these systems showed dramatic increases in CPU time, even when moving from chain lengths $N = 20$ to $N = 22$. It seems that conformational complexity and the sampling of minimum energy states can cause dramatic increases in CPU time. More analysis of the reasons behind these large simulation times is needed.

One technique for trying to reduce CPU time is to sample less of the energies near the ground-state of each $J_A$ value. This corresponds well with the analysis techniques presented in Chapter 6, as one can check the validity of low temperature calculations, directly assessing the quality of data in this region. However, initial results show that CPU time still increases dramatically, and it simply might not be possible to simulate all conformational space for chains much larger than $N = 55$ within a reasonable amount of CPU time.

A possible better choice for reducing CPU time in 2D Wang-Landau sampling could be breaking the density of states into different sections, and running individual simulations for each section[112]. However, the complexities of behavior for semi-flexible systems does not offer a straight-forward method for dividing the density of states into sections. This author feels that the best approach might be to study individual conformational regions separately, e.g., simulating stiffness values that incorporate only the solid-globular region. This would allow moves to be tuned to states typically of this region, and will also give researchers
the chance to focus on a particular type of behavior, as opposed to trying to study all
conformations existing for a single chain.

8.5.2 2D vs 1D Wang-Landau sampling in semi-flexible homopolymers

A key component of future work is to analyze conformational behavior in semi-flexible sys-
tems with results from one-dimensional simulations. Preliminary evidence has shown that
the two-dimensional simulations performed in this dissertation reveal features indistinguish-
able in one-dimensional simulations with similar move sets. This is particularly true for low
temperature states found in the rod-like and toroidal conformational regions. The dynamics
of two-dimensional simulations, with sampling directions of energy and flexibility, allow for
the sampling of minimum energy states whose conformations are nearly unattainable with
the move set implemented here. This is true even with the many improvements to this
move set described in Chapter 5. However, appropriate results have yet to be fully gath-
ered in comparing one-dimensional and two-dimensional methods of simulating semi-flexible
homopolymers, but this author expects this to be a major component of future work using
multi-dimensional Wang-Landau sampling.
Chapter 9

Conclusions

The work presented in this dissertation has focused on the simulation of coarse-grained polymer systems using state-of-the-art computational techniques. In particular, the Wang-Landau algorithm has been applied in the thermodynamic analysis of coarse-grained flexible and semi-flexible homopolymers. These homopolymers have provided a tractable system from which general results were obtained regarding conformational and transition behavior, and additionally, the efficient simulation of these coarse-grained models has provided methodologies that can be applied to a broad range of physical systems. The understanding gained in these studies will provide general insight in the broader problems associated with folding in polymer and protein systems.

Results describing the general behavior of single flexible homopolymer chains have been obtained for a wide range of chain lengths. A one-dimensional implementation of the Wang-Landau algorithm has been particularly useful in describing the coil-globule and liquid-solid transitions for chain lengths $N \leq 561$. In addition, solid-solid transitions have been described for chain lengths $N \leq 55$. Of particular importance is the comparison with other researchers studying a variety of different homopolymer models. These comparisons have shown a number of inconsistencies that are now understood through the work in this dissertation. One of the most crucial issues involves the large number of finite-size effects occurring in flexible homopolymers. These effects can cause significant error in the analysis of transition temperature aimed at understanding behavior in the infinite-chain limit, and must be considered when performing such analyses. Another critical issue described in this dissertation is the need to validate results from density of states based thermodynamic calculations.
These issues can produce systematic error not described by standard statistical analysis, and therefore, the analysis methodologies described within this dissertation become a valuable tool in understanding results from density of states based calculations.

Future work performed on flexible homopolymers should focus on two goals: development of methods aimed at fully describing finite-size effects in these systems, which often lead to unique behavior such as solid-solid transitions, and development of improved techniques that can simulate chain lengths $N > 561$. The first goal has much to do with the implementation of the Wang-Landau algorithm, which in this case was tuned more to understanding the liquid-solid and coil-globule transitions in the infinite limit. Many of the methods used to study semi-flexible homopolymers in this dissertation would be well-suited in meeting this goal in a one-dimensional implementation of the Wang-Landau algorithm, particularly the method used to reduce the modification factor. The second goal involves improving simulation methods so that larger and larger systems can be simulated. In particular, cluster studies [109] describe a crossover from icosahedral to decahedral geometries for chain lengths $> 1690$, and considering the correlation between clusters and homopolymers, realizing such a crossover in homopolymers would be a significant accomplishment. This becomes even more significant when considering collaborations with experimentalists working on cluster or polymer systems, as such a crossover occurring for larger systems could be more obtainable than the system sizes presented in this dissertation.

The thermodynamic behavior of semi-flexible homopolymer systems is more complicated compared to flexible homopolymer systems. A variety of unique minimum energy states are available when adjusting the flexibility of semi-flexible homopolymers, namely, solid-globular, rod-like, and toroidal states. The behavior of these states has been described in terms of the flexibility, energy, temperature, and chain length, leading to a more well-defined understanding of behavior in small chain lengths. One of the most impressive features of these results is the identification of icosahedral Mackay and Anti-Mackay states for such broad values of stiffness within the solid-globular region. It is expected that knowledge of this
packing behavior will lead to changes in the way future investigations study these collapsed states.

As studies advance in understanding conformational and transition behavior in semi-flexible systems, greater focus should be placed on describing each of the three conformational regions described above. The solid-globular region can be studied in a manner similar to fully-flexible homopolymers, where a first step has been given in terms of icosahedral Mackay and Anti-Mackay behavior. Even within the results presented here for $N \leq 55$, significant understanding of possible solid-solid transitions, or transformations from icosahedral to other geometries, are left to be studied. The rod-like conformational regions require similar studies in understanding the most efficient form of packing of rod segments. In addition, understanding how the boundary between rod-like and toroidal states behaves in the infinite-chain limit is of equal importance, and will require simulation of large chain lengths. The toroidal region is also of interest, where this dissertation has shown that conformational changes occur as flexibility is varied. Understanding these changes should lead to further classification of conformations in this region.

As stated in the concluding remarks of Chapter 8, a large amount of computing time is needed to describe all of the interesting conformational behavior in $N \leq 55$ chain lengths. The methods introduced in Chapter 8 allow completion of a small number of simulations for $N = 45$ length chains in approximately 30 CPU days. However, other simulations within that same set of simulation have taken much longer times, or remain incomplete. Methods can be developed to make this CPU time more consistent and shorter, as was shown in Chapter 8 when removing some of the bins near the minimum energy states. Other possible methods have been presented and future work should take these into consideration. It is fully expected that significant improvements can be made to allow more efficient sampling of multi-dimensional density of states for semi-flexible homopolymers. This becomes particularly important when considering that some of the minimum energy configurations found in
these multi-dimensional simulations may not be attainable using standard one-dimensional Wang-Landau sampling.

In addition, it is the hope of this author that collaborations can be made with experimentalists studying biomolecules such as DNA or RNA, with the goal being to help drive these scientists to verify unique packing in solid-globular biomolecules. Such experiments already exist for genetic material in viral capsids [113, 114], which represent a suitable example of the solid-globular state. It is not to say that packing in such systems will “absolutely” be of the icosahedral Mackay and Anti-Mackay type, as it is understood that packing in DNA will certainly depend on a number of factors, possibly requiring more realistic simulation models. However, the simulation work presented in this dissertation should provide techniques for studying other models that may more closely resemble the real packing in these biomolecules.

The results presented in this dissertation have furthered the overall knowledge of thermodynamic behavior in flexible and semi-flexible homopolymers. The Wang-Landau algorithm has been exceptionally efficient in simulating the coarse-grained models used to mimic these homopolymers, particularly in describing conformational and transition behavior through a number of physical properties. Although a significant amount of results have been presented here, a number of unanswered questions remain. Results presented in this dissertation will act as a basis from which future studies can be performed.


