Replica-exchange Wang–Landau simulations of lattice proteins for the understanding of the protein folding problem

by

Guangjie Shi

(Under the direction of David P. Landau)

Abstract

Protein folding is studied within the context of two coarse-grained lattice models that separate all amino acids into only a few types. The hydrophobic-polar (HP) model is a simplified lattice protein model for simulating protein folding and for understanding many biological problems of interest. In this work, an "improved" model, the semi-flexible H0P model, was proposed by introducing a new type of "neutral" monomer, "0", i.e., neither hydrophobic nor polar and also taking into consideration the stiffness of bonds connecting monomers. Even though both models are highly simplified protein models, finding the lowest energy conformations and determining the density of states are extremely difficult. We applied replica-exchange Wang–Landau sampling with appropriate trial moves for determining the density of states of multiple HP and H0P proteins, from which the thermodynamic properties such as specific heat can be calculated. Moreover, we developed a heuristic method for determining the ground state degeneracy of lattice proteins, based on multicanonical sampling. It is applied during comprehensive studies of single-site mutations in specific lattice proteins with different sequences. The effects in which we are interested include structural changes in ground states, changes of ground state energy, degeneracy, and thermodynamic

properties of the system. With respect to mutations, both extremely sensitive and insensitive positions in the protein sequence have been found. That is, ground state energies and degeneracies, as well as other thermodynamic and structural quantities may be either largely unaffected or may change significantly due to mutation. Moreover, comparison between the HP model and the semi-flexible H0P model have been performed based on two real proteins: Crambin and Ribonuclease A. We found that, compared with the HP model, the semi-flexible H0P model possesses significantly reduced ground state degeneracy, and rich folding signals as the proteins rearranging into native states from very compact structures at low temperatures. We calculated the free energy vs end-to-end distance as a function of temperature. The HP model shows a relatively shallow folding funnel and flat free energy minimum, reflecting the high degeneracy of the ground state. In contrast, the semi-flexible H0P model has a well developed, rough free energy funnel with a low degeneracy ground state. In both cases, folding funnels are asymmetric with temperature dependent shape.

INDEX WORDS: Monte Carlo simulations, Wang–Landau sampling, Replica-exchange Wang–Landau sampling, protein folding, protein folding funnel, hydrophobic-polar model, HP model, semi-flexible H0P model, protein mutation

Replica-exchange Wang–Landau simulations of lattice proteins for the understanding of the protein folding problem

by

Guangjie Shi

B.S., Xiamen University, 2011

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

2016

©2016

Guangjie Shi

All Rights Reserved

Replica-exchange Wang–Landau simulations of lattice proteins for the understanding of the protein folding problem

by

Guangjie Shi

Approved:

Major Professors:	David P. Landau
Committee:	Michael Bachmann Heinz-Bernd Schüttler Shan-Ho Tsai

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my adviser Prof. David P. Landau, for his invaluable guidance, support, and patience throughout my Ph.D studies. His passion for physics and insightful approach to research have been tremendous inspirations to me. I cannot image a better adviser and mentor.

I am grateful for other members in my advisory committee: Prof. Michael Bachmann, Prof. Heinz-Bernd Schüttler and Dr. Shan-Ho Tsai for carefully evaluating my dissertation and providing invaluable suggestions.

I would like to thank a number of collaborators: Dr. Thomas Vogel, Dr. Thomas Wüst and Dr. Ying Wai Li. They are great mentors and excellent scientists, who have provided me with tremendous, invaluable help.

I would like to express my appreciation to many members of the Center for Simulational Physics. I am grateful for a lot of beneficial discussions with Dr. Dilina Perera, Dr. Kai Qi, Mr. Tomas Koci, Mr. Lingyun Wu. I am indebted to the former administrative specialist Ms. Linda Lee and the current administrative specialist Ms. Stephanie Crowe for their cordial assistance. I also would like to thank systems administrators Mr. Mike Caplinger and Mr. Jeff Deroshia, whom I have constantly bothered with technical problems.

In addition, I would like to thank Dr. Ben Bergen and Dr. Christoph Junghans from Los Alamos National Laboratory, for selecting me to be one of participants of the 2015 Co-Design summer school.

Last but not least, I would like to thank my family for their endless support and unconditional love throughout all these years. Ever since I was a child, my parents have provided me with excellent education opportunities, and have respected every decision I made. Without them, I would not have been able to achieve this much.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The protein folding problem has been studied for more than 50 years, but much remains to be learned [1]. The attraction of understanding this problem comes from the fact that proteins are fundamental, molecular components which perform numerous biological functions in living organisms. For instance, protein enzymes catalyze many chemical reactions. In the cells, enzymes are very important for many metabolic processes in a way that they help catalyze the reaction in rates fast enough to sustain life. The biological function of a protein (e.g. enzyme) is strongly related to the native structure that it folds into. Therefore misfolded proteins are usually inactive, but in some cases they can have modified or even toxic functionality. Some examples of protein related diseases include Alzheimer's disease [2] and Parkinson's disease [3].Therefore, understanding the problem of protein folding is a primary step to recognize the principles of many biological processes. In a recent review article [1], Dill and colleagues concluded that the protein folding problem reduces to three main questions: (i) "How is the 3D native structure of a protein determined by the physicochemical properties that are encoded in its 1D amino acid sequence?" (ii) "How can proteins fold so fast?" (iii) "Can we devise a computer algorithm to predict a protein's native structure from its amino acid sequence?".

The goal of this dissertation is to contribute to the understanding of (i) and (ii) through means of advanced numerical simulations based on coarse-grained protein models. Instead of focusing on the atomic details of proteins, we apply advanced techniques from statistical physics on those models that only capture the most essential factors in protein folding. The simplicity of these models enables us the ability to carry out systematic studies and to understand the problems from the perspective of statistical physics.

To unveil the strength of the relationship between one-dimensional amino acid sequences and three-dimensional native structures (i.e. question (i)), one important question is how will simple changes on an amino acid sequence (i.e. sequence mutations) affect the structure that it folds into? Moreover, investigating the effect of mutations has more practical motivations since mutations have been known to be responsible for many diseases. For instance, Huntington's disease is caused by an autosomal dominant[1] mutation in either of an individual's two copies of a gene. Scientists used to think that any two naturally occurring proteins with a 40% or higher sequence identity would possess the same fold [4]. However, experiments discovered that proteins such as *Pfl6* and *Xfaso 1* with high sequence similarity end up with different folds [5]. Furthermore, Alexander *et al.* have successfully designed two proteins that share 88% of their sequence, but fold into totally different tertiary structures [6]. For this part of our work, the intent is to systematically examine how a simple mutation would affect the thermodynamic and structural properties, folding processes and ground state properties of a minimalistic lattice protein model: the HP model [7, 8] which classifies amino acids into only two types: hydrophobic and polar.

The time scale for a protein to fold into its functional state ranges from microseconds to seconds [9]. However, with the astronomical number of possible conformations, how could proteins fold so fast (i.e. question (ii))? The so-called "protein folding funnel" [10, 11] theory has been developed for answering this question. Within this theory, the states of a protein

[1]For genes on any chromosome other than sex chromosome, the alleles are autosomal dominant.

are treated as distributions of individual chain conformations, all of which tend to change the structures in ways that the free energy is minimized. At the same time, this process is subject to the influence of thermal fluctuation, i.e. instead of monotonically reducing the free energy, the protein might be directed to higher free energy states in some cases during this process. However, the folding funnel is always portrayed schematically as a relatively symmetric function of some unknown reaction coordinate about a unique minimum (the native state). In this part of our work, our goal is to uncover the protein folding funnels using minimalistic lattice protein models. However, the simplicity of the HP model yields large ground state degeneracies which stands in contrast to the generally unique native state of natural proteins. To ease this issue, we propose some modifications to the original HP model by introducing a hierarchical hydrophobicity among different monomers and also considering the natural rigidity of real proteins. This improved model, also known as the semi-flexible H0P lattice protein model ('0' stands for neutral in terms of hydrophobicity) shows much lower ground state degeneracy compared to its ancestor, the HP model, and also renders a more "realistic" folding funnel.

The arrangement of this dissertation is as follows: Chapter 2 gives a brief introduction to proteins and some related works. Chapter 3 describes the HP lattice protein model and its improved version, the semi-flexible H0P lattice protein model, as well as the structural and thermodynamic properties we investigated. Chapter 4 outlines the Monte Carlo techniques and trial moves we adopted in our work. Chapter 5 addresses the question of how mutations will affect the folding of lattice proteins. In Chapter 6, we first make brief comparisons between HP and semi-flexible H0P lattice protein models. It will be followed by the discussions of folding funnels for both lattice protein models and how these funnels help the understanding of folding processes. We will draw conclusions in Chapter 7.

Chapter 2

Brief Background on Proteins

2.1 What are proteins?

Proteins are large biomolecules consisting of one or more long polypeptides, each of which is a linear chain of amino acid residues. The importance of proteins lies in the fact that they are essential components which perform a vast array of biological functions in living organisms. As seen in Table. 2.1, there are twenty proteinogenic amino acids of the standard genetic code, serving as the building blocks of proteins [12].

All of the 20 amino acids can be described as four different components: a hydrogen atom, an amino group ($-NH_2$), a carboxyl group ($-COOH$) and a side chain ($-R$) attached to the central carbon atom (C_α). To form a linear backbone of the protein, the amino group of one amino acid is covalently bonded with the carboxyl group of another amino acid through a peptide bond. The side chain ($-R$), as the only feature that distinguishes between these amino acids, can have different properties: charged/uncharged polar or non-polar.

In terms of protein structures, there are four levels of complexity. Primary structure refers to the linear sequence of amino acids. A secondary structure is the arrangement of the amino acid sequence into helices and sheets, of which the most common types are the

alpha (α) helix and the beta (β) sheet. The secondary structures will be packed together to form a compact globule, namely a tertiary structure. The quaternary structure refers to the arrangement of multiple folded protein subunits.

In an aqueous environment, in order to minimize the free energy, hydrophobic amino acids tend to group together. Even though this behavior originates from the repulsive force by the water molecules, it has been treated as effective attraction between those hydrophobic amino acids. This hydrophobic interaction is believed to be the key driving force of protein folding and tertiary structure formation [13, 14]

Amino Acid	Abbreviation	One-letter code
Isoleucine	Ile	I
Valine	Val	V
Leucine	Leu	L
Phenylalanine	Phe	F
Cysteine	Cys	C
Methionine	Met	M
Alanine	Ala	A
Glycine	Gly	G
Threonine	Thr	T
Tryptophan	Trp	W
Serine	Ser	S
Tyrosine	Tyr	Y
Proline	Pro	P
Histidine	His	H
Glutamic acid	Glu	E
Glutamine	Gln	Q
Aspartic acid	Asp	D
Asparagine	Asn	N
Lysine	Lys	K
Arginine	Arg	R

Table 2.1: The twenty genetically encoded amino acids with their abbreviations and one-letter codes.

2.2 From pathway to folding funnel

The famous refolding experiment by Christian Anfinsen and colleagues [15, 16] postulated the hypothesis that, instead of the kinetic folding route, the native structure of a protein only depends on the conditions of solution and the intrinsic properties of amino acid sequences. Also, the native structure is thermodynamically stable. But Cyrus Levinthal made the argument, which is also known as the "Levinthal paradox", that the number of protein conformations is too large for a protein to find its native conformation by randomly searching ("needle in the haystack"). Thus he concluded the necessity of existing protein folding pathways that will guide proteins to fold into functional states [17]. This argument was framed in a way that the goal of achieving the global minimum and the goal of finding this minimum quickly are mutually exclusive; and it leads to the search of folding pathways of proteins as well as the intermediate states within these pathways [10].

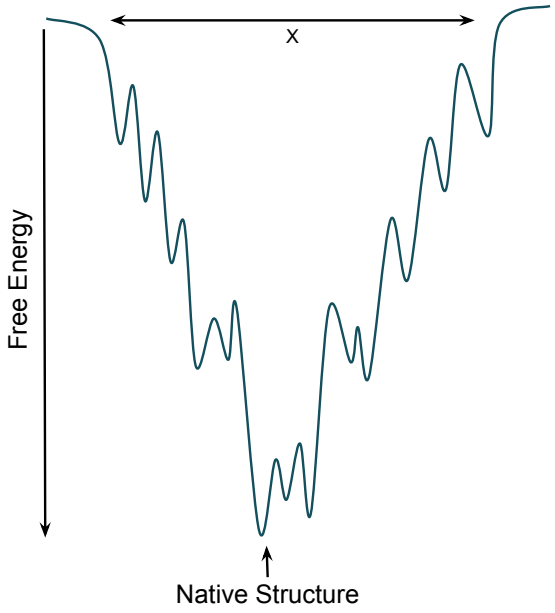


Figure 2.1: A sketch of a rugged protein folding funnel. The vertical axis represents the free energy of the protein; while the horizontal axis is an unknown reaction coordinate.

A more advanced view also known as "protein folding funnel" [10, 11] arose with the introduction of concepts from statistical physics. Instead of a single conformation, a protein state is actually a macroscopic state which is a distribution of individual chain conformations (or microscopic states). Therefore, two different macroscopic states could have substantial overlap of microscopic states. A protein folding funnel is just a funnel-shaped free energy landscape, i.e. many high-energy conformations and few low-energy conformations, of the protein with respect to the degrees of freedom, a sketch of which is shown in Fig. 2.1. In this case, protein folding should not be treated as a sequence of events that guides the protein changes from one conformation to another. Instead, it should be viewed as parallel events: upon folding, a protein tends to reduce its free energy through changing its conformations in different ways, but also constantly be directed to different states due to thermal fluctuation.

2.3 Tools in studying protein folding

2.3.1 Experimental approaches

For determining the protein structures at the atomic level, two major tools are X-ray crystallography and nuclear magnetic resonance spectroscopy.

X-ray crystallography [18] is a technique used for identifying the structure of a crystal at the atomic level. The resolution of this method usually ranges from one Angstrom (Å) to a few Angstroms. This is particularly useful when it comes to determining the tertiary structures of proteins, where the positions of atoms are necessary. However, since this method is only suitable for crystallized materials with repeating pattern, it might not work well when determining the details of folded protein structures which are usually globular lumps with irregular surfaces. Moreover, the preparation of protein crystals is time consuming, and the processes are easily affected by a number of factors: temperature, solvent pH and so on.

Nuclear magnetic resonance (NMR) spectroscopy [19] utilizes the nuclear magnetic moments inside some atoms, such as ^{1}H and ^{13}C, to learn the distances between pairs of atoms within the molecules. These distances can then be used for deriving the 3D structures. Even though NMR spectroscopy does not have resolution as high as x-ray crystallography, it has the advantage of resolving the proteins inside solution. Therefore, with this method, it is possible to get a structure closer to a specific physiological environment. The drawback of NMR spectroscopy, however, is that its applications are limited to small proteins with a few hundred residues.

2.3.2 Computational approaches

In terms of computational techniques, two major ones are molecular dynamics and Monte Carlo sampling.

Molecular Dynamics [20] is a deterministic method which predicts the positions of particles in the entire system by solving the equations of motion with a discretized time scale. The equations of motion for each particle in the system are determined by the forces experienced by them. In this way, this method simulates the evolution of the physical system throughout a limited time period. This provides the method with the advantages of observing the dynamical processes of protein folding. However, the quality of a molecular dynamics simulation relies on the empirical force field that it applied to model the interactions at the atomic level. Moreover, even with the computing power nowadays, it is still impossible to carry out molecular dynamics simulation at the realistic folding time range (from microseconds to seconds) of protein folding, using a fine time step (e.g. a femtosecond).

On the other hand, Monte Carlo sampling methods [21] belong to a family of stochastic methods. Instead of solving equations of motion for each particle, this type of method generates a sequence of states through trial moves. These trial moves are accepted or rejected according to a probability distribution. The physical properties of the system can be learned

through statistical analysis of these accepted configurations. Due to its stochastic nature, Monte Carlo compared to molecular dynamics can explore a much larger conformational phase space of the system, and thus provides the ability for investigating more general problems like "phase transition" (noticeable structural change or pseudophase transition) [22]. More details about Monte Carlo methods will be discussed in Ch. 4. Moreover, in the field of protein folding, Monte Carlo methods usually are applied with simplified protein models, which will be discussed in Ch. 3.

Chapter 3

Coarse-grained Lattice Protein Models

3.1 Overview of coarse-grained models

Coarse-grained models represent an amino acid residue of a protein by a spherical monomer, neglecting the atomic details of the amino acid. An amino acid sequence, then, is reduced to a chain of monomers that are connected through chemical bonds. Depending on the number of different types of monomers in the sequence, a coarse-grained model can be classified as a homopolymer, where only one type of monomer is considered, and a heteropolymer, which contains two or more different types of monomers. In both cases, to simulate a protein, the interactions between non-bonded monomers are chosen in reference to the realistic interactions between amino acids.

For off-lattice, coarse-grained protein models, a monomer chain forms different structures in a continuous three dimensional space. Different factors, such as bond length (i.e. the distance between two consecutive monomers), bond angle, torsional angle and so on have to be taken into consideration. While in the case of lattice protein models, the degrees of freedom

have been reduced even more. Instead of continuous space, now each monomer occupies a lattice site. Three dimensional lattices include simple cubic, face-centered cubic or body-centered cubic lattices. Moreover, bond length and bond angle in the case of lattice models are fixed, and, in the simplest case the interactions exist only between two neighboring, non-bonded monomers.

The advantages of employing lattice protein models mainly comes from the computational point of view. The calculations of energy, i.e. counting the number of non-bonded contacts or angles, are much easier on lattice compared with off-lattice models. For instance, on a simple cubic lattice, which is adopted in our work, there are only six sites ($\mathcal{O}(1)$ complexity) to be examined in order to find the number of non-bonded contacts for a given monomer. However, the same calculation will require examinations of the distance from N-1 monomers ($\mathcal{O}(N)$ complexity) to a given monomer, where N is the total number of monomers in a chain. Moreover, due to the discrete nature of the lattice model, the energy as well as the system coordinates can be represented by integer values, which allows us to take advantages of integer arithmetic. For example, for an off-lattice model, one usually has to decide the bin sizes of energy when employing advanced Monte Carlo methods, such as Wang–Landau sampling (Sec.4.3). This round-off error for floating-point numbers will introduce artifacts into the simulation.

3.2 Hydrophobic-polar (HP) lattice protein model

The Hydrophobic-polar lattice protein model [7, 8] classifies 20 amino acids into only two types: hydrophobic (H) and polar (P), based on the nature of their side chains. It characterizes the hydrophobic interaction, which is considered to be the driving force of protein folding, by an attractive interaction between only the non-bonded nearest neighbor hydrophobic monomers (also known as H-H contacts). The protein structure (one example is shown in

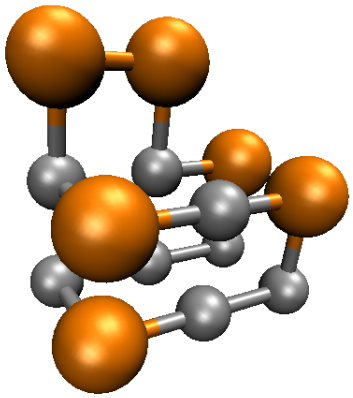


Figure 3.1: A three-dimensional structure of an HP lattice protein with 14 monomers, on a simple cubic lattice. Hydrophobic and polar monomers are colored in dark gray and orange, respectively.

Fig. 3.1) is mimicked by a self-avoiding walk on a rigid lattice. The Hamiltonian is given by

$$\mathcal{H} = -\epsilon_{\text{HH}} n_{\text{HH}} \,, \tag{3.1}$$

where n_{HH} is the number of non-bonded HH contacts. Hence, the ground state, i.e., lowest-energy state, of an HP protein is the state with a maximum number of n_{HH}. The HP model has served in the studies of many important topics over the years. Some examples include surface adsorption [23, 24, 25, 26, 27] and protein folding in membranes [28] or confined environments [29, 30].

Despite the simplicity of the HP model, finding the lowest energy structure of a given sequence is an NP-complete problem [31]. For long sequences (chain length $\gtrsim 30$), enumeration methods (see, e.g., [32, 33, 34]) are not accessible. However, different folding algorithms and Monte Carlo methods have been developed for approaching this problem. Examples include, but are not limited to, constraint-based approaches [35], chain-growth methods [36, 37], in particular the pruned-enriched Rosenbluth method (PERM) and its

variants [38, 39, 40, 41, 42, 43], sequential importance sampling [44], fragment regrowth Monte Carlo [45], multidomain sampler [46], genetic algorithms [47, 48], evolutionary Monte Carlo [49], and ant colony models [50]. Among those, the sampling method developed by Wang and Landau [51, 52, 53] has shown to be powerful and highly precise in simulating proteins and polymers. Moreover, a recently developed replica-exchange Wang–Landau [54, 55] sampling utilizes the power of parallel computing (see Chapter 4) and thus allows us to investigate systems with much larger sizes.

3.3 H0P lattice protein model

3.3.1 Motivation

The simplicity of the HP model yields large ground state degeneracies which stands in contrast to the generally unique native state of natural proteins. More advanced models such as the HPNX model [56, 57], which considers not only the hydrophobicity but also the charges of amino acids, have been shown to be effective in reducing degeneracy. From a different perspective, we propose some simple modifications to the original HP model, rendering the model more realistic without significantly increasing the difficulties of sampling.

In the original HP model, there are only two types of monomers: hydrophobic and polar. However, since different amino acids possess different levels of hydrophobicity [58], simplification into only two types might be insufficient to characterize the features of the hydrophobic interaction, which is believed to be the driving force of protein folding and tertiary structure formation. Therefore, we introduce a new level of "neutral" monomers, "0", which neither favors nor dislikes water as much as H or P monomers do. The mappings from amino acids to different types of monomers for the HP model [59] and the H0P model [60] are shown in Table 3.1.

Amino Acid	HP Model	H0P Model
I	H	H
V	H	H
L	H	H
F	H	H
C	H	H
M	H	H
A	H	0
G	P	0
T	P	0
W	H	0
S	P	0
Y	H	0
P	P	0
H	P	P
E	P	P
Q	P	P
D	P	P
N	P	P
K	P	P
R	P	P

Table 3.1: The mappings from amino acids to different types of monomer for the HP model and the H0P model. Details of the one-letter code of amino acids can be found in Table 2.1.

3.3.2 Description of the model

The H0P lattice protein model classifies 20 amino acids into three different types: hydrophobic (H), "neutral" (0) and polar (P). As shown in Table 3.1, this mapping is performed based on the hydrophobic index studied in ref [58]. We classify amino acids with hydrophobic index ≥ 1.9 as hydrophobic, those with hydrophobic index ≤ -3.2 as polar, and the rest of the amino acids are "neutral". The Hamiltonian of the H0P lattice protein model can be written as:

$$\mathcal{H} = -\epsilon_{\mathrm{HH}} n_{\mathrm{HH}} - \epsilon_{\mathrm{H0}} n_{\mathrm{H0}} - \epsilon_{00} n_{00} \,, \tag{3.2}$$

where n_{H0} and ϵ_{H0} are the number of non-bonded contacts between H and 0 monomers, and its effective interaction constant (n_{HH}, n_{00}, ϵ_{HH} and ϵ_{00} are defined likewise). One example of the H0P model is shown in Fig. 3.2, which includes a 2D structure with 9 monomers (4 Hs, 2 0s and 3 Ps).

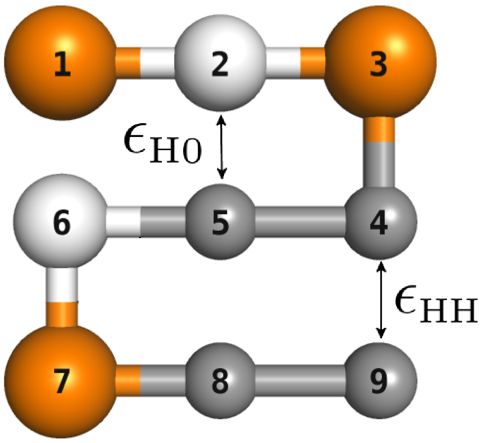


Figure 3.2: An example of an H0P model. Hydrophobic and "neutral" monomers are colored in dark gray and white, respectively, while polar monomers are colored in orange. The interaction between monomers 2 and 5 is ϵ_{H0}, and that between monomers 4 and 9 is ϵ_{HH}. In this particular 2-dimensional structure, $n_{\mathrm{HH}} = 2$, $n_{\mathrm{H0}} = 1$ and $n_{00} = 0$.

3.3.3 Semi-flexible H0P lattice protein model

Amino acid sequences are usually considered stiff chains with some persistence length. The idea of considering bond-stiffness energy has been involved in different studies previously, e.g., Thomas and Dill explored the relationship between helical propensities and conformations for globular proteins using HP model with locally helix interaction [61]; Bastolla and Grassberger [62] studied a lattice model of semi-flexible homopolymers with nearest neighbor attraction and energetic preference for straight joints between bonded monomers; Krawczyk etc. [63] have extensively investigated semi-flexible hydrogen-bonded and non-hydrogen bonded lattice polymers in both two and three dimensional lattice. As an exten-

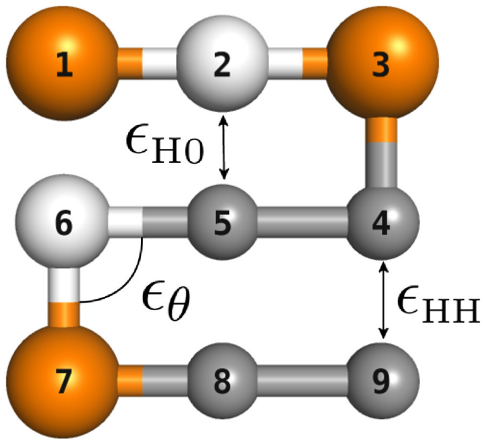


Figure 3.3: An example of a semi-flexible H0P model. Hydrophobic and "neutral" monomers are colored in dark gray and white, respectively, while polar monomers are colored in orange. The interaction between monomers 2 and 5 is ϵ_{H0}, and that between monomers 4 and 9 is ϵ_{HH}. The angle constituted by monomers 5, 6 and 7 contributes ϵ_θ energy. In this particular 2-dimensional structure, $n_{\mathrm{HH}} = 2$, $n_{\mathrm{H0}} = 1$ and $n_\theta = 4$.

tion to the H0P lattice protein model, the bond stiffness is taken into consideration in the semi-flexible H0P model, through an energetic term ϵ_θ for each angle exists in the protein structures. With n_θ representing the number of angles, the Hamiltonian of the semi-flexible

H0P model can be written as:

$$\mathcal{H} = -\epsilon_{\rm HH} n_{\rm HH} - \epsilon_{\rm H0} n_{\rm H0} - \epsilon_{00} n_{00} - \epsilon_\theta n_\theta \,. \tag{3.3}$$

One example of the semi-flexible H0P lattice protein model is shown in Fig. 3.3.

Crambin (46)	TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGAT CPGDYAN
HP3D46	PPHHPPHHHPPPHPHHPHPPPPPHHHHPHPPHHHHPPHP HPPPHHP
H0P3D46	OOHHOOHHOPOPHPHHPHOOOOPOHHOOOOOHHHHOOO OHOOPOOP
Ribonuclease A(124)	KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCK PVNTFVHESLADVQAVCSQKNVACKNGQTNCYQSYSTMSIT DCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDASV
HP3D124	PPPHHHPHPPPPHPPPPPHHPPPPHHPPHHPPPPHPPPPHP PHPPHHPPPHHPHPHHHPPPPHHHPPPPPPHHPPHPPHPHP PHPPPPPPPHPPHHHPPPPHPPPHHHHHPPPPHHPHPHPHPH
H0P3D124	PPOOOOPHPPPPHPOOOOOOOOOPOHPPHHPOPPHOPPPHP OHPOHHPPOHOPHPOHHOPPPHOHPPOPOPHOPOOOOHOHO PHPPOOOOPOOPHOOPOOPOPPPHHHOHPOPOOHOHPHPOOH

Table 3.2: Two real proteins: Crambin and Ribonuclease A converted into HP and H0P models, according to the mapping rule in Table 3.1

3.4 Mappings of two real proteins

We have chosen two real proteins: Crambin and Ribonuclease A, both of which have been converted into the HP model by Lattman et al. [59] following the mapping rule of HP model listed in Table 3.1. Applying the mapping rule of H0P model from the same table, we converted these two sequences into H0P models as well. Both original amino acid sequences and their converted results are shown in Table 3.2. As for Crambin which has 46 residues, the corresponding HP sequence (i.e. HP3D46) contains 22 H monomers and 24 P monomers; the H0P sequence (i.e. H0P3D46) has 15 H, 24 0 and 7 P monomers. A much longer protein Ribonuclease A is composed of 124 residues, out of which 47 and 77 are respectively converted into H and P monomers in the HP model (i.e. HP3D124). For the corresponding H0P sequence (i.e. H0P3D124), the numbers of H, 0 and P monomers are 29, 50 and 45.

3.5 Structural quantities

Structural quantities are essential in understanding the conformational changes during the folding process. We measure two commonly used quantities, radius of gyration (R_{g}) and end-to-end distance (R_{ee}):

$$R_{\mathrm{g}} = \left(\frac{1}{N} \sum_{i=1}^{N} (\vec{r}_i - \vec{r}_{\mathrm{cm}})^2 \right)^{1/2} \tag{3.4}$$

$$R_{\mathrm{ee}} = \left| \vec{r}_N - \vec{r}_1 \right|, \tag{3.5}$$

N is the number of monomers in the chain; $\vec{r}_i$ and $\vec{r}_{\mathrm{cm}}$ represent the positions of the ith monomer and the center of mass of the given configuration, respectively.

In addition, Wüst et al. [64] proposed another scalar structural observable, the tortuosity τ of the protein:

$$\tau = \left(\frac{1}{N-2} \sum_{i=1}^{N-2} (s_i - \overline{s})^2 \right)^{1/2} \tag{3.6}$$

where

$$s_i = \sum_{j=1}^{i} \vec{r}_{j,j+1} \times \vec{r}_{j,j+2}\,, \quad 1 \leq i \leq N-2\,, \tag{3.7}$$

and

$$\overline{s} = \frac{1}{N-2} \sum_{i=1}^{N-2} s_i. \tag{3.8}$$

Here $\vec{r}_{j,j+1}$ (or $\vec{r}_{j,j+2}$) denotes a vector pointing from monomer j to $j+1$ (or $j+2$). Unlike the radius of gyration and the end-to-end distance, which measure spatial extent only, τ is particularly sensitive to sequence-dependent internal topological features such as the breaking of HH contacts in compact denatured states upon folding to the ground state. For our purpose, it serves as a complementary structural quantity to better interpret features in the specific heat curves, for example. See also [64] for a discussion of this observable in the context of lattice polymers.

Chapter 4

The Methodology of Monte Carlo Simulation

4.1 A brief review of statistical physics and Monte Carlo techniques

4.1.1 Partition function and thermodynamic quantities

Monte Carlo techniques in statistical physics are primarily applied for the goal of estimating thermodynamic properties. For that purpose, it is essential to obtain the partition function Z, which is written as:

$$Z(T) = \sum_s \mathrm{e}^{-E_s/k_\mathrm{B}T} = \sum_E g(E)\,\mathrm{e}^{-E/k_\mathrm{B}T}\,. \tag{4.1}$$

The first term in this equation sums over all the possible microscopic states s with energy E_s. This energy is weighted by the Boltzmann factor $\mathrm{e}^{-E_s/k_\mathrm{B}T}$ (where k_B is the Boltzmann constant). The equivalent representation in the second term of this equation sums over

each energy level E, instead of configurations, which will be then weighted by the density of states ($g(E)$) and the Boltzmann factor. The density of states ($g(E)$) is the representation of the number of configurations at each energy level, and thus is independent of the canonical temperature. Therefore, the knowledge of the density of states gives access to all the thermodynamic properties of the system at different temperatures. Some quantities of particular interest in our work include the average energy $\langle E \rangle$ and the heat capacity C_V, which can be calculated as:

$$\langle E \rangle = Z^{-1} \sum_{E} E\, g(E)\, \mathrm{e}^{-E/k_\mathrm{B}T}, \tag{4.2}$$

$$C_V = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_\mathrm{B}T^2}. \tag{4.3}$$

Moreover, one could obtain the thermodynamics of a structural quantity Q (e.g. radius of gyration as mentioned in Ch. 3.5) through a similar manner with the knowledge of the two-dimensional density of states ($g(E, Q)$). In this case, the formulas of the partition function and averaged structural quantity $\langle Q \rangle$ are written as:

$$Z(T) = \sum_{E,Q} g(E, Q)\, \mathrm{e}^{-E/k_\mathrm{B}T}\,. \tag{4.4}$$

$$\langle Q \rangle = Z^{-1} \sum_{E,Q} Q\, g(E, Q)\, \mathrm{e}^{-E/k_\mathrm{B}T}, \tag{4.5}$$

4.1.2 Detailed balance condition

The applications of Monte Carlo methods in statistical physics can be described as the processes of randomly generating large amount of statistically independent configurations for a given physical model. Then the properties of interest of the physical system can be calculated by averaging over these configurations for a given ensemble of states. Even though

the manner of generating new configurations is model dependent, it is governed by the Master Equation:

$$\frac{dP_i(t)}{dt} = \sum_{i \neq j} \left(\omega_{j \to i} P_j(t) - \omega_{i \to j} P_i(t) \right), \tag{4.6}$$

where the $P_i(t)$ represents the probability of the system being in state i at "simulation time" t, and $\omega_{i \to j}$ is the transition rate from state i to state j; $P_j(t)$ and $\omega_{j \to i}$ are defined likewise. This equation leads to the detailed balance condition:

$$\omega_{j \to i} P_j(t) = \omega_{i \to j} P_i(t), \tag{4.7}$$

for equilibrium processes, i.e. when $\frac{dP_i(t)}{dt} = 0$. Since the probability $P_i(t)$ is constant with respect to time for equilibrium processes, from now on we will use P_i instead. We can further divide the transition rate $\omega_{i \to j}$ into the multiplication of two parts: the proposal probability $\omega^p_{i \to j}$, i.e. the probability of proposing an update, and acceptance probability $\omega^a_{i \to j}$, i.e. the probability of accepting this update:

$$\omega_{i \to j} = \omega^p_{i \to j} \cdot \omega^a_{i \to j}. \tag{4.8}$$

Therefore, the detailed balance condition in Eq. (4.7) then can be written as:

$$\frac{\omega^p_{i \to j} \cdot \omega^a_{i \to j}}{\omega^p_{j \to i} \cdot \omega^a_{j \to i}.} = \frac{P_j}{P_i}. \tag{4.9}$$

One solution to this equation is

$$\omega^a_{i \to j} = \min \left(\frac{P_j \cdot \omega^p_{j \to i}}{P_i \cdot \omega^p_{i \to j}}, 1 \right) \tag{4.10}$$

This form is particularly useful when we introduce the Monte Carlo trial moves that we have employed in our work in later sections of this chapter.

4.1.3 Metropolis sampling

Metropolis sampling [65] was the very first application of Monte Carlo techniques in statistical physics. As a special case of importance sampling, it generates an ensemble of conformations according to the Boltzmann distribution, at the temperature of interest. Then properties of the system can be calculated by averaging over these configurations. For a classical physical system at temperature T, the probability of it being in state i with energy E_i is given by

$$P_i(T) = Z^{-1} e^{-E_i/k_{\mathrm{B}}T}, \tag{4.11}$$

where k_{B} is the Boltzmann constant and Z is the partition function. Combining this probability distribution with Eq. (4.10), and simplifying the case by assuming a symmetric proposal probability, i.e., $\omega^p_{i\to j} = \omega^p_{j\to i}$, one can derive the acceptance criterion for a trial move in Metropolis sampling as:

$$\omega^a_{i\to j} = \min\left(e^{-(E_j - E_i)/k_B T}, 1\right) = \min\left(e^{-\Delta E/k_B T}, 1\right). \tag{4.12}$$

The "recipe" for Metropolis sampling can be described as:

1. Specify the temperature of simulation, and generate an initial state i with energy E_i.

2. Propose a trial move that changes the system from current state i to state j with energy E_j.

3. According to the acceptance criterion in Eq. (4.12), if $E_j < E_i$ the update will automatically be accepted. Otherwise, a random number $r \in [0, 1]$ will be generated. If $r < \omega^a_{i\to j}$, the trial move will be accepted. Upon acceptance, update the current state: $i \to j$ and $E_i \to E_j$.

4. Steps 2 and 3 will be repeated until a desired number of Monte Carlo steps have been

performed.

However, one unavoidable drawback of the Metropolis sampling method is that it is easily trapped in the metastable states during the simulation, especially at low temperature region. This is because, the probability of a system getting out of a metastable state drops off exponentially with the decrease of temperature, as it depends on the acceptance probability: $e^{-\Delta E/k_B T}$. More advanced Monte Carlo methods, such as Wang–Landau sampling [51, 52], have the advantage of overcoming this issue and will be introduced in later sections.

4.1.4 Remarks on random number generators

As Monte Carlo methods involve processes driven by random numbers [21], the quality of which is essential for the reliability of the simulation results. In our work, we have adopted the "Mersenne Twister" pseudo-random number generator [66], the implementation of which is provided in the GNU Scientific Library (GSL) as **gsl_rng_mt19937**. This pseudo-random number generator has an extremely long period, roughly 10^{6000}, and it passes numerous tests for statistical randomness, including the Diehard tests, which are a battery of statistical tests developed by George Marsaglia. Moreover, it has been in the comparison with another high quality pseudo-random number generator RANLUX [67] side by side on the simulation of the HP lattice model [68]. The comparison found that simulation results using both pseudo-random number generators agree with each other extremely well, and "Mersenne Twister" pseudo-random number generator is slightly better in terms of efficiency.

4.2 Monte Carlo trial moves for lattice models

Before giving a discussion of the Monte Carlo algorithms we employed, we give a brief description of some cleverly designed trial moves for studying lattice polymer models. Those

trial moves ideally can be combined with any types of Monte Carlo techniques. However, some combinations might just work better than others in terms of efficiency.

4.2.1 Local moves

Local moves refer to those trial moves that only modify a small part of a configuration, rendering a new configuration that is fairly similar to the old one. Some common examples include the end-flip (Fig. 4.1), kink-flip (Fig. 4.2) and crankshaft (Fig. 4.3). Even though local moves are relatively easy to implement with constant computational complexity, they share the same drawback of inducing long correlation times in the simulation.

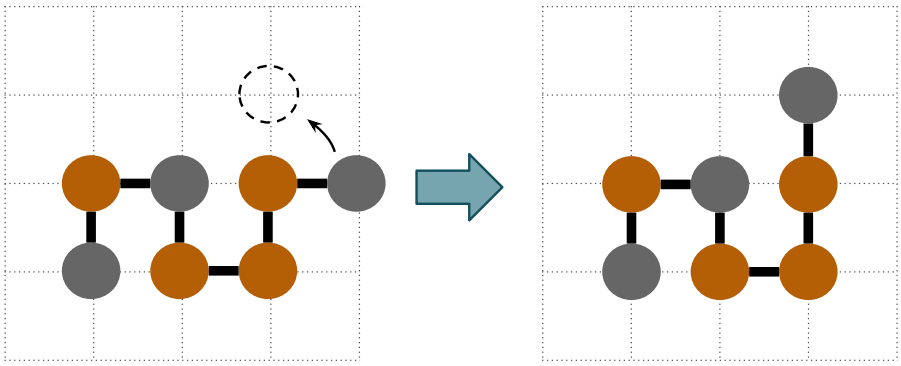


Figure 4.1: An end flip, showing the move of the black monomer at the end to the position outlined by the dashed circle.

4.2.2 Pivot move

Pivot move is a popular non-local trial move for simulating lattice polymers [69]. To perform a pivot move on a polymer chain with N monomers, one monomer (k) on the chain is picked at random as the pivot point, and a randomly picked symmetry operation is applied on the portion of chain comprising the monomers $k+1, \ldots, N$ (e.g. see Fig. 4.4). This move will be rejected if the resultant configuration is not self-avoiding. Upon acceptance, a chain portion of order N will be updated through a pivot move, and thus the chain conformation will be

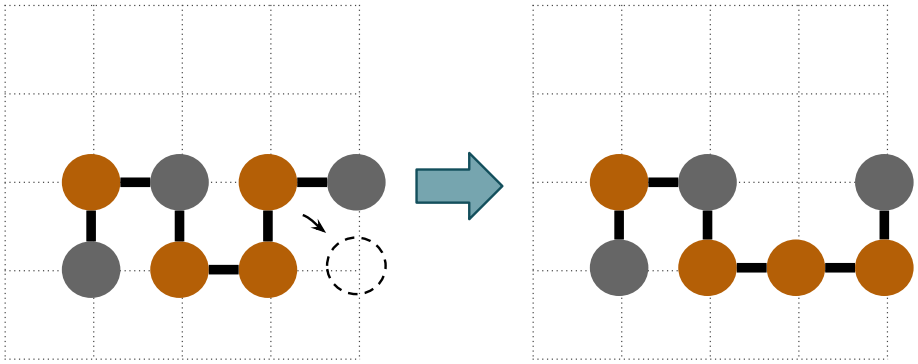


Figure 4.2: A kink flip, where a selected monomer is moved to its diagonally adjacent lattice site that is outlined by the dashed circle.

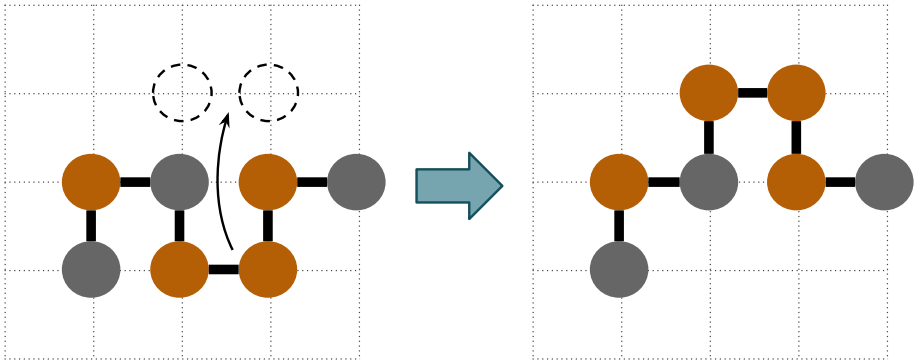


Figure 4.3: A crank shaft move, where two monomers are selected and moved to the positions outlined by the dashed circles.

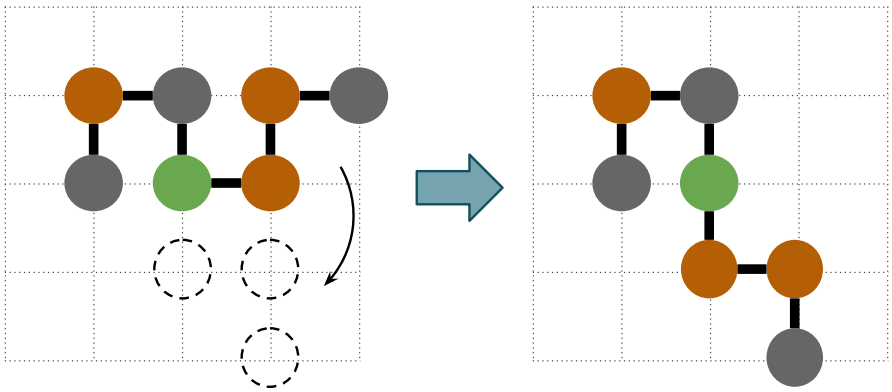


Figure 4.4: A pivot move, where the pivot point is colored in green and three monomers are moved to the positions outlined by the dashed circles.

drastically changed. However, the drawback of a pivot move is its high rejection rate. This is especially the case in the low temperature region, where the conformations tend to be compact. Therefore, a drastic change in conformation is likely to be rejected.

4.2.3 Pull move

Pull moves [70] combine the advantages of local and global moves. That is, depending on the choice of a monomer to start with, this type of moves can displace 1 or up to $N-1$ monomers of the entire chain. The reversibility and ergodicity of the pull moves that are described below have been mathematically proven. The former one refers to the property that for any move in the move set that changes the configuration, there exists another one in the same move set to restore the initial configuration. The latter one means that any configuration is reachable from any other valid configuration through a sequence of moves in the move set.

Details of the move

Assume a polymer chain consists of n monomers on a 2 dimensional square lattice, the basic steps of pull moves can be described as following (these steps can be easily generalized to a cubic lattice):

1. Randomly pick a monomer and mark it as $i\,(1 < i < n)$. See steps 4 and 5 for the cases where $i = 1$ or n. Identify the lattice site that is adjacent to $i-1$ and at the same time diagonally adjacent to i, and mark it as L. Therefore i, $i-1$ and L are three corners of a square lattice. The move can proceed only if the site L is not occupied, and the **fourth corner** is either monomer $i+1$ as in Fig. 4.5 or free (marked as C) as in Fig. 4.6 and Fig. 4.7. Otherwise, the whole procedure starts over with another monomer.

2. If the fourth corner is monomer $i+1$ as in Fig. 4.5, the whole move only consists of flipping monomer i to the site L.

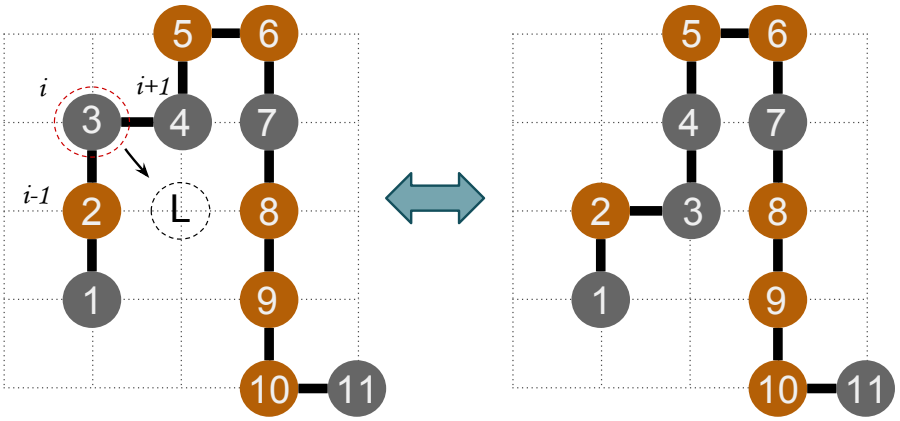


Figure 4.5: Pull move: single-monomer move, which is similar to the kink flip as in Fig. 4.1.

3. If the fourth corner is free and is marked as C in Fig. 4.6 and Fig. 4.7, we move monomer i and $i+1$ to site L and C respectively. If at this point, a valid configuration is obtained, then the move is finished, as in Fig. 4.6. Otherwise, the rest of chain starting from monomer $i+2$ will be pulled along the trajectory of the chain until a valid configuration is reached. In this case, the positions of multiple monomers will be updated, as in Fig. 4.7.

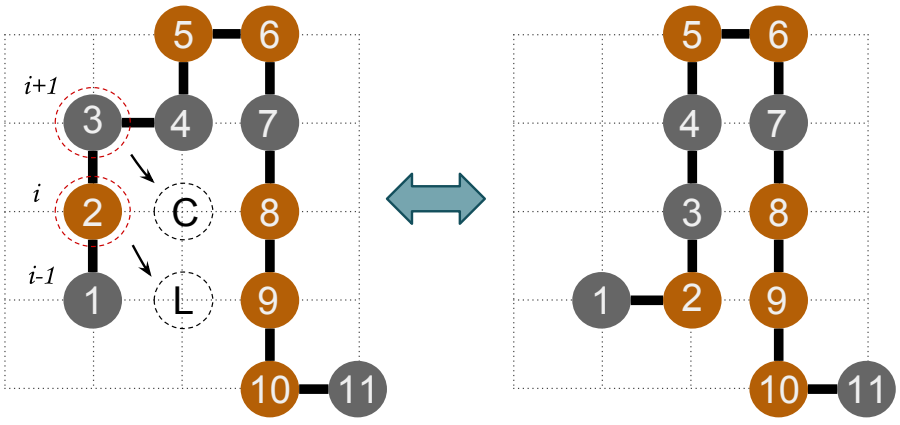


Figure 4.6: Pull move: two-monomers move, where two monomers are flipped to their diagonally adjacent lattice sites.

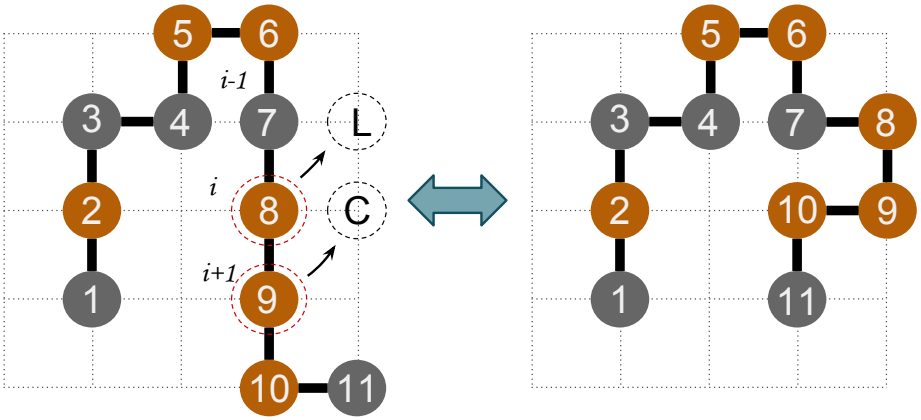


Figure 4.7: Pull move: internal multi-monomers move, where two monomers are flipped to their diagonally adjacent lattice sites and the rest of the chain is pulled to reach a valid configuration.

4. For the case where either end of the chain is chosen, i.e., $i = 1$ or n, the move will be performed slightly differently from above procedures. Assume $i = n$, that is the last monomer has been chosen. We consider any path of two free lattice sites with one of the locations adjacent to monomer n. Then monomers n and $n-1$ will be moved to these free locations, as in Fig. 4.8. Again, the rest of the chain will be pulled along the trajectory until a valid configuration is found.

5. However, the pull move as shown in Fig. 4.9 is not reversible and thus is forbidden in our simulation.

Consideration about detailed balance

As seen in Eq. (4.10), in order to obey detailed balance, one needs to carefully consider the acceptance criterion for this move set due to the possible non-symmetric proposal probability:

$$\omega^a_{i\to j} = \min\left(\frac{P_j}{P_i}\frac{n_{j\to i}/n_j}{n_{i\to j}/n_i}, 1\right), \tag{4.13}$$

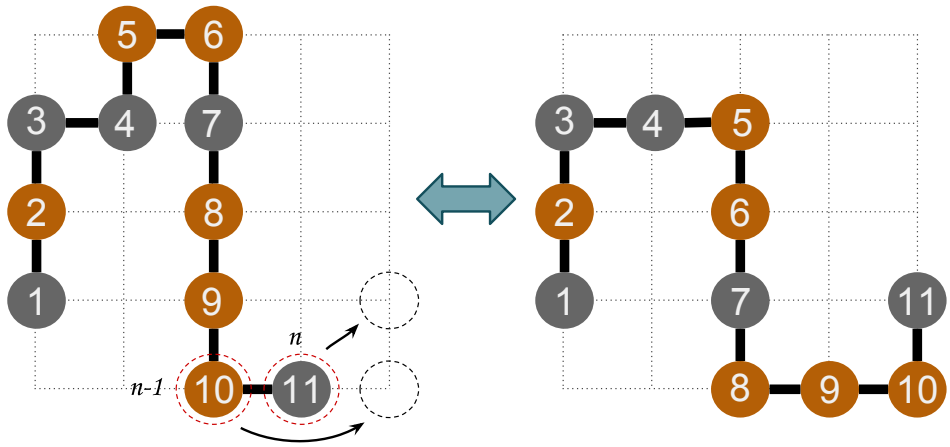


Figure 4.8: Pull move: chain-terminal move, where the two monomers at the end are moved to the positions that are outlined by dashed circles, and the rest of the chain is pulled to reach a valid configuration.

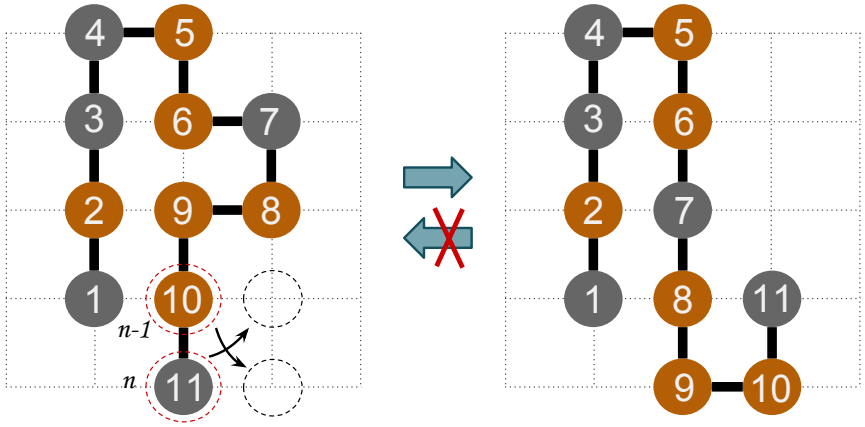


Figure 4.9: Pull move: chain-terminal move forming a "hook". This move is non-reversible and thus is forbidden.

where n_i denotes the number of possible pull moves which can be performed on state i, and $n_{i\to j}$ represents the number of pull moves that change the system from state i to state j; n_j and $n_{j\to i}$ have the similar meanings. Considering the reversibility of pull moves, we know that $n_{i\to j} = n_{j\to i}$, and thus Eq. (4.13) can be simplified as:

$$\omega^a_{i\to j} = \min\left(\frac{P_j}{P_i}\frac{n_i}{n_j}, 1\right). \tag{4.14}$$

Therefore, to perform a pull move in our simulation, the number of possible pull moves (n_i) that can be performed on the current state i is counted first. Out of all these possible moves, one is picked at random and changes the system to state j. Then the number of possible pull moves (n_j) that can be performed on state j is counted as well. In this way the acceptance probability can be calculated.

4.2.4 Bond-rebridging move

The essential idea behind the bond-rebridging move [71] is that, instead of the displacement of monomers, it reconstructs the bonds connecting these monomers such that the resultant conformation, if valid, is very different from the original one. This type of move is particularly useful for proposing the change for dense conformations. As a bond-rebridging move obeys detailed balance itself, it is unnecessary to correct the acceptance criterion as in the case of pull moves.

Details of the move

Assume a polymer chain consists of n monomers on a 2 dimensional square lattice, the basic steps of bond-rebridging moves can be described as following (these steps can be easily generalized to a cubic lattice):

1. Two consecutive monomers i and $i+1$ ($1 < i, i+1 < n$) are picked at random. Denote

the unit vector from i to $i+1$ as $\mathbf{v}$. See step 6 for the case where an end monomer is picked.

2. Randomly pick one out of two (four for a simple cubic lattice) neighboring unit vectors of $\mathbf{v}$, and denote it as $\mathbf{u}$. If the lattice sites on two ends of $\mathbf{u}$ are occupied by two consecutive monomers, proceed to the next step. Otherwise, the whole process starts over again.

3. For two monomers connected by $\mathbf{u}$, denote the one neighboring to monomer i as j and another one as k, as in Fig. 4.10. The move will be performed differently depending on the relationship between j and k.

4. If $j-k=1$, the two vectors $\mathbf{v}$ and $\mathbf{u}$ are anti-parallel, as seen in Fig. 4.10, then:

   (1) Cut the links between i and $i+1$, between j and k. By connecting i and j, $i+1$ and k, the configuration become a segment and a loop.

   (2) Randomly pick two consecutive monomers i' and $i'+1$ on the loop to form a vector $\mathbf{v}'$. Look for a vector $\mathbf{u}'$ on the segment that neighbors to $\mathbf{v}'$. If such a vector could be found, the "cut-and-join" procedures in step (1) will be performed. Otherwise, the whole process starts over again.

   (3) Renumber the monomers and also reassign the monomer types to restore the sequence of the chain.

5. If $j-k=-1$, the two vectors $\mathbf{v}$ and $\mathbf{u}$ are parallel, as seen in Fig. 4.11(a), then:

   (1) Cut the links between i and $i+1$, between j and k. Connect i and j, $i+1$ and k.

   (2) Renumber the monomers and also reassign the monomer types to restore the sequence of the chain.

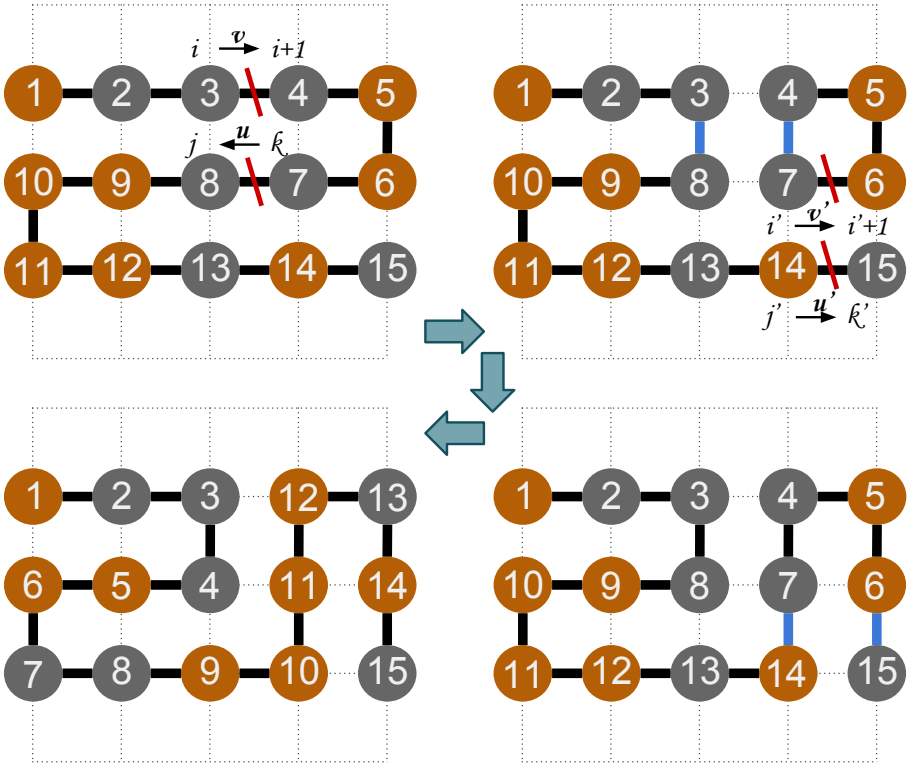


Figure 4.10: Bond-rebridging move: one pair of anti-parallel bonds is cut to form a segment and a hoop; another pair of bonds between these two parts is found and cut; reconnect these two parts and reorder the sequence.

6. If either of the end monomers is picked to perform a bond-rebridging move, denote it as i, as in Fig. 4.11(b) and perform following steps:

   (1) Look for neighboring sites that are occupied by monomers but not directly connected to i, and randomly pick one from them as j. If none of these sites are found, the whole process starts over again. Otherwise, link monomers i and j, and cut the bond between j and its neighbors so that a valid configuration is reached.

   (2) Renumber the monomers and also reassign the monomer types to restore the sequence of the chain.

4.2.5 Combination of different moves

The combination of pull moves and bond-rebridging moves works amazingly well with (replica-exchange) Wang–Landau and sampling for both the determination of the minimum energy state and the estimation of the density of states for the models used here [72, 64, 60, 73]. In our work, these two moves, combined with pivot moves are called with different probabilities: 75% pull moves, 23% bond-rebridging moves and 2% pivot moves.

4.3 Replica-exchange Wang–Landau (REWL) sampling

4.3.1 Wang–Landau (WL) algorithm

Wang–Landau (WL) sampling [51, 52] is an iterative Monte Carlo algorithm to estimate the density of states $g(E)$, by ideally performing a random walk in energy space. In Wang–Landau sampling, the probability of finding state i with energy E_i is the reciprocal of the density of states (i.e., $1/g(E_i)$). Therefore, with the assumption of symmetric proposal

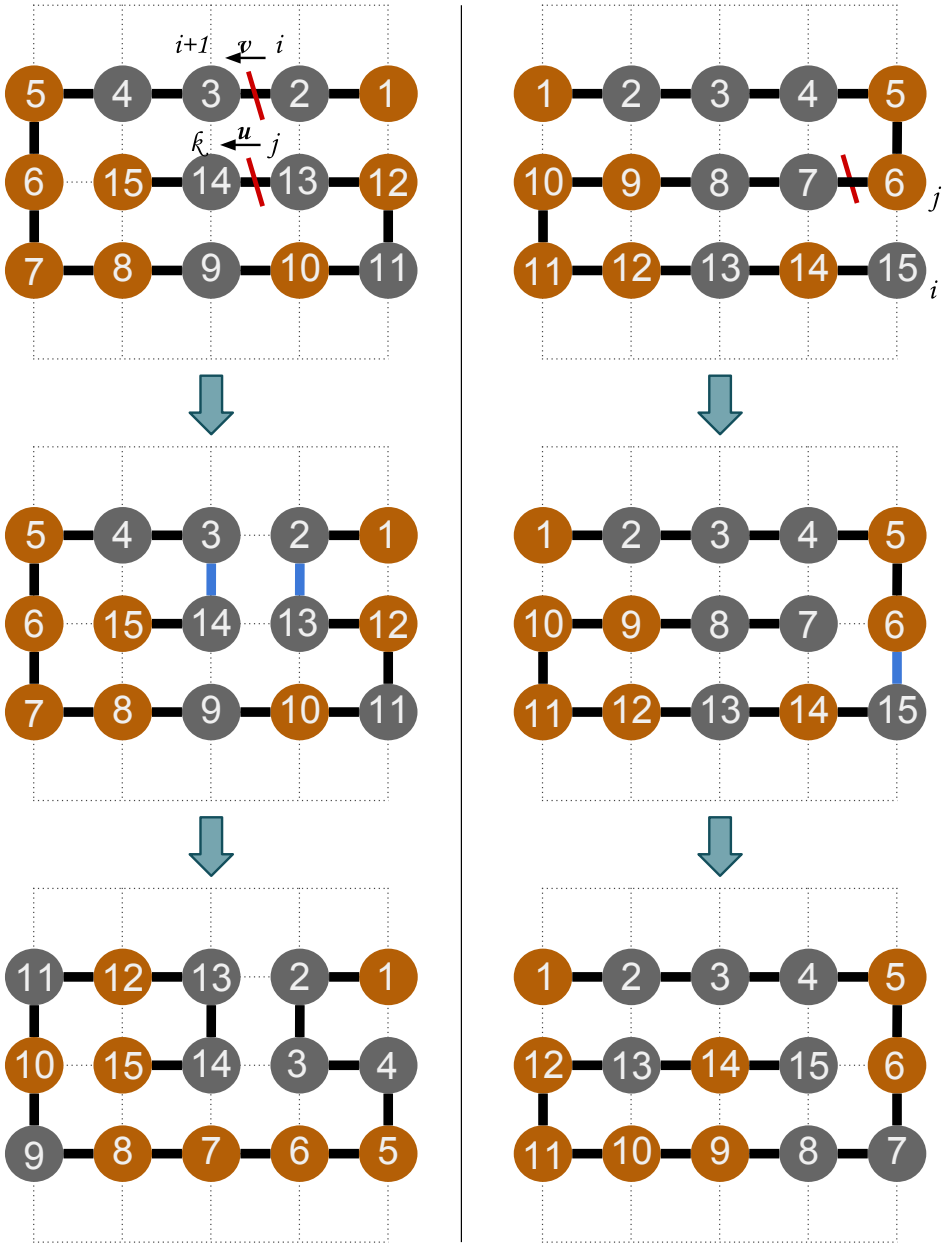


Figure 4.11: Bond-rebridging move: (a) one pair of parallel bonds is found and cut; reconnect these two parts and reorder the sequence; (b) chain-terminal move, where only one bond is cut and another one is rejoined; the sequence is reordered in the end.

probability (i.e. $\omega^p_{j\to i} = \omega^p_{i\to j}$), the acceptance criterion in Eq. (4.10) can be easily derived for Wang–Landau sampling:

$$P(i \to j) \equiv \omega^a_{i\to j} = \min\left(\frac{g(E_i)}{g(E_j)}, 1\right). \tag{4.15}$$

One could notice the most advantageous part of this formalism is that it is independent of the canonical temperature and thus the Boltzmann factor. This advantage allows Wang–Landau sampling to avoid many drawbacks of temperature dependent Monte Carlo techniques, such as Metropolis sampling as described in Ch. 4.1.3.

Details of the algorithm

The algorithm starts with a pre-defined energy range $E \in [E_{\min}, E_{\max}]$; an empty histogram $H(E) = 0, \forall E \in [E_{\min}, E_{\max}]$; an estimator for density $g(E)$ of states with initial guess (e.g. $g(E) = 1, \forall E \in [E_{\min}, E_{\max}]$); a modification factor f with an initial value as f_{init} (e.g. $\ln f_{\text{init}} = 1$) and a final value as f_{final} (e.g. $\ln f_{\text{final}} = 10^{-6}$ or 10^{-8}). With an initial configuration i and energy E_i, the following procedures will be repeated until the criterion $f \leq f_{\text{final}}$ is satisfied:

1. Propose a trial move that changes the system from state i to j.
2. Accept this move with probability $P(i \to j) = \min\left(\frac{g(E_i)}{g(E_j)}, 1\right)$.
3. Upon acceptance, update the histogram: $H(E_j) \to H(E_j) + 1$ and density of states: $g(E_j) \to g(E_j) \times f$. If the trial move was rejected, $H(E_i)$ and $g(E_i)$ are updated instead in the same way.
4. Check the flatness criterion of the histogram. In our work, we chose the criterion as the minimum value of the histogram is larger than the 80% of the averaged histogram

values. If this flatness criterion is satisfied, reset the histogram $H(E) = 0, \forall E \in [E_{\min}, E_{\max}]$ and update the modification factor: $f \to \sqrt{f}$.

As the density of states ($g(E)$) are constantly updated throughout the simulation, the detailed balance condition as in Eq. (4.7) is not satisfied for Wang–Landau sampling during its early iterations. However, the modification factor approaches unity by the end of the simulation, and thus the detailed balance is restored.

4.3.2 Details of the parallelization

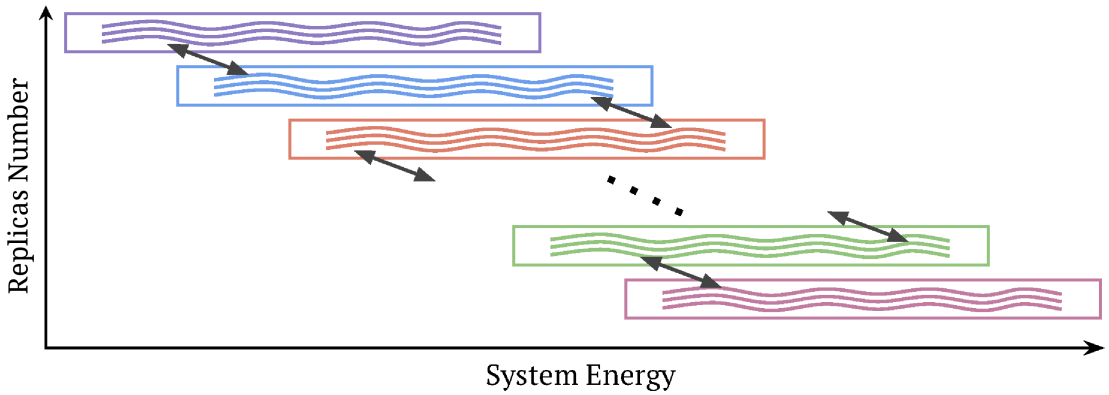


Figure 4.12: Illustration of the general framework of replica-exchange Wang–Landau sampling. The system energy range is split into multiple, overlapping energy windows which are colored differently. In each window, the wavy lines represent multiple independent processes (or walkers) running serial Wang–Landau sampling. At fixed time intervals, the replica exchange procedure will be performed between two neighboring energy windows, as indicated by two-way black arrows.

Replica-exchange Wang–Landau (REWL) sampling [54, 55] is a generic, parallel computing framework for WL sampling. It fully exploits and combines the power of WL sampling and replica-exchange Monte Carlo. As seen in Fig. 4.12, the energy range of the whole system of interest is split into multiple, overlapping energy windows, in each of which one or more processes (or walkers) are employed to run serial Wang–Landau sampling independently. Each of these walkers carries its own estimator for the density of states $g(E)$ and its

own histogram $H(E)$ and is required to satisfy the flatness criterion individually. But all the walkers inside an energy window share the same modification factor and stopping criterion. At fixed time intervals the replica exchange will be performed between neighboring windows (details will be discussed later in this section). The percentage of the overlap region between two neighboring windows depends on the system. However, excessive overlaps, e.g. at the extreme case 100%, will make the framework inefficient, while small overlaps would result in small acceptance rates for replica exchange. In our work, we have adopted percentages range from 60% to 80%.

Flatness criterion and stopping criterion

Recall that in serial Wang–Landau sampling, the simulation proceeds into the next iteration with reduced modification factor if the flatness criterion is satisfied, and the whole simulation terminates if the modification factor is smaller than a predefined threshold. In replica-exchange Wang-Landau sampling, different energy windows proceed into the next iteration independently. Only if all the walkers in an energy window satisfy the flatness criterion, the modification factor of this energy window will be reduced, and the estimators of density of states will be averaged over these walkers before proceeding to the next iteration. The whole simulation will be terminated only if all energy windows satisfied the stopping criterion, i.e. the modification (f) is smaller than a predefined threshold (as discussed in ref. [64], for short sequences, the error saturated after $\ln f \leq 10^{-6}$; while for longer sequences a more stringent criterion should be used, e.g. $\ln f \leq 10^{-8}$). In the end of the simulation, each energy window will result in one piece of density of states that is averaged over all the walkers inside.

Replica-exchange procedures

Assume there are $1 \ldots N$ energy windows, and there are n_i walkers running WL sampling inside the range of window i. At fixed time intervals the replica exchange will be performed

as following:

1. For two neighboring energy windows: i and j, randomly pick one walker from each window and denote them as w_i and w_j respectively.

2. Denote the present conformation of w_i as X and that of w_j as Y. Propose the replica exchange with acceptance probability as:

$$P_{acc} = \min\left(\frac{g_i[E(X)]}{g_i[E(Y)]}\frac{g_j[E(Y)]}{g_j[E(X)]}, 1\right), \tag{4.16}$$

   where $g_i[E(X)]$ is the current estimator for the density of states of walker i at the energy of conformation X; $g_i[E(Y)]$, $g_j[E(X)]$ and $g_j[E(Y)]$ are defined likewise..

The replica exchange allows each replica to travel through different energy windows of the system and helps resolving configurational trapping, if any, of some random walkers. However, this procedure should not be performed at a high frequency, since the message exchange process is usually costly, especially between two different computing nodes. The interval we have adopted in our work is in the range of $1,000 \sim 10,000$ Monte Carlo steps.

4.3.3 Concatenation of density of states pieces

At the end of REWL sampling, there will be one piece of density of states from each energy window. In order to obtain the final density of states for the whole energy range, in this section we will describe the method we adopted for merging these pieces together.

One-dimensional density of states

In the case of a one-dimensional density of states, assume $g_i(E)$ and $g_j(E)$ are two pieces from neighboring energy windows i and j respectively. We first calculate the inverse of microcanonical temperatures through $\beta(E) = d\ln[g(E)]/dE$, and denote them as $\beta_i(E)$ and $\beta_j(E)$ respectively. The difference can then be calculated as $\Delta\beta = |\beta_i(E) - \beta_j(E)|$. The point where the value of $\Delta\beta$ is the smallest is chosen as the joining point E_{join}. Therefore, we use E_{join} and $g(E_{join})$ as reference values to rescale the density of states, for instance, $g_j(E) \to g_j(E)g_i(E_{join})/g_j(E_{join})$. In this way the final density of states that combines $g_i(E)$ and $g_j(E)$ can be obtained:

$$g(E) = \begin{cases} g_i(E), \text{ for } E < E_{join} \\ g_j(E), \text{ for } E \geq E_{join} \end{cases} \tag{4.17}$$

In this manner, the potential discontinuities in the final density of states due to improper merging can be avoided [54, 55, 74, 75]. One example of the merging process is shown in Fig. 4.13, in which there are five pieces of density of states obtained through a replica-exchange Wang–Landau sampling run for the H0P3D46 with $\epsilon_{\text{HH}} = 4$, $\epsilon_{\text{H0}} = 2$, $\epsilon_{00} = 1$, and zero for the rest of the interactions. In this example, we apply five-point stencil method [76] for evaluating numerical derivatives of the logarithm of density of states.

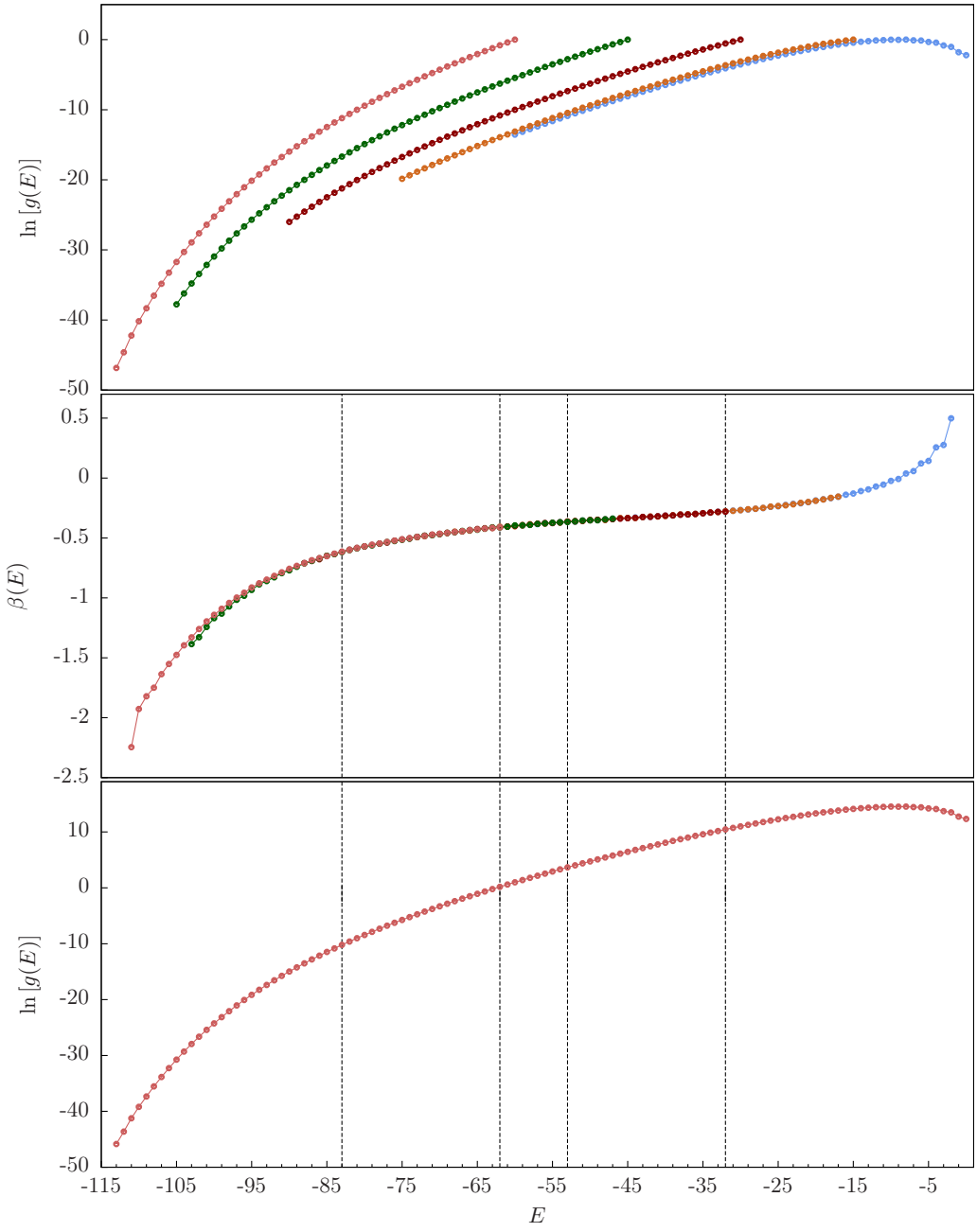


Figure 4.13: Joining five pieces of $g(E)$ of the H0P3D46 model with $\epsilon_{\text{HH}} = 4$, $\epsilon_{\text{H0}} = 2$, $\epsilon_{00} = 1$, and zero for the rest of the interactions. (Top) Raw density of states pieces from a replica-exchange Wang–Landau sampling run, in which the system energy is divided into five energy windows and each is simulated by one random walker. (Middle) Derivatives of the logarithm of each piece of density of states. The points where the derivatives coincide the best are marked by dotted lines. (Bottom) Final, merged density of states.

Two-dimensional density of states

We are not only interested in the energetic quantities of the system, but also structural quantities, which are usually helpful in understanding the structural changes of the system. For this reason, two-dimensional densities of states are necessary in order to obtain the thermodynamics of these structural quantities. The replica-exchange Wang–Landau sampling we described in Ch. 4.3.2 can be easily extended for performing two-dimensional random walks on the space of the system energy E and a structural quantity Q to obtain the joint density of states $g(E, Q)$. Moreover, in the next section, we also will introduce an alternative method for obtaining $g(E, Q)$, if more than one structural property is of interest.

Assume $g_i(E, Q)$ and $g_j(E, Q)$ are two pieces of the joint density of states from window i and j, the normal vectors of surfaces $\ln[g_i(E, Q)]$ and $\ln[g_j(E, Q)]$ can be calculated:

$$\hat{n}_i(E, Q) = (\frac{\partial \ln[g_i(E, Q)]}{\partial E}, \frac{\partial \ln[g_i(E, Q)]}{\partial Q}, -1)$$

and

$$\hat{n}_j(E, Q) = (\frac{\partial \ln[g_j(E, Q)]}{\partial E}, \frac{\partial \ln[g_j(E, Q)]}{\partial Q}, -1)\,.$$

The joining point (E_{join}, Q_{join}) is determined as the one that maximizes the value of $\hat{n}_i \cdot \hat{n}_j$. With that, the $g_j(E, Q)$ will be rescaled: $g_j(E, Q) \to g_j(E, Q) g_i(E_{join}, Q_{join}) / g_j(E_{join}, Q_{join})$ and the final density of states can be obtained as:

$$g(E, Q) = \begin{cases} g_i(E, Q), \text{ for } (E, Q) \text{ in window } i \\ g_j(E, Q), \text{ for } (E, Q) \text{ in window } j \\ \frac{1}{2}\left[g_i(E, Q) + g_j(E, Q)\right], \text{ for overlapped region} \end{cases} \tag{4.18}$$

4.4 Replica-exchange Multicanonical (MUCA) Sampling

For cases where structural properties of a system are of interest, one could perform multi-dimensional random walkers using the replica-exchange Wang–Landau sampling framework described in the previous section. However, it becomes time-consuming if more than one structural quantity is of interest. Therefore, an alternative method we adopted is to apply multicanonical sampling [77, 78] with the same replica-exchange framework. The procedures of replica-exchange multicanonical sampling is exactly the same as REWL except the following three points:

1. During the simulation, the density of states $g(E)$ obtained from previous REWL is held fixed without updating and the reciprocal of which will be used as the weight for accepting or rejecting trial moves. A trial move is accepted with a probability that is the same in Eq. (4.15). The proposal of replica exchange shares the same formula in Eq. (4.16) as well.

2. Each random walker maintains multiple two-dimensional histograms for structural quantities. Let Q represent a structural quantity of interest and $H_i(E, Q)$ is the two-dimensional histogram from random walker i. After a trial move is performed and judged for acceptance, the value of Q will be evaluated for the current conformation, and the histogram will be updated: $H_i(E, Q) \rightarrow H_i(E, Q) + 1$. At the end of the simulation, the joint density of states $g_i(E, Q)$ is then obtained through: $g_i(E, Q) = g_i(E)H_i(E, Q)$. A final density of states $g(E, Q)$ then can be obtained by merging these pieces together as described in Ch. 4.3.3.

3. The simulation is terminated when a predefined number of Monte Carlo steps have been performed.

4.5 Ground state degeneracy estimation

Despite the simplicity of the HP model, finding the lowest energy structure of a given sequence is an NP-complete problem [31]. For long sequences (chain length $\gtrsim 30$), enumeration methods (see, e.g., [32, 33, 34]) are not accessible. Moreover, due to different symmetry operations on a simple cubic lattice, identifying a unique lattice protein structure is costly. All of these render the determination of the ground state degeneracy of a lattice protein extremely challenging. In this section, we propose a method for estimating the ground state degeneracy and the results are compared with other methods, including exact enumeration, for short chains.

4.5.1 Sequence of directions

For a given lattice protein conformation on the simple cubic lattice, the "path" from the first monomer through the end can be uniquely recorded as a sequence of directions. We define two sets of values, $\vec{B}_1, \vec{B}_2, \ldots, \vec{B}_{N-1}$ and $D_1, D_2, \ldots, D_{N-1}$, for the $N-1$ bonds that connect consecutive monomers in a sequence of length N. The first set, $\vec{B}_k$, is determined by the difference of coordinates between monomer k and $k+1$. Therefore $\vec{B}_k$ will be assigned one of values from $\{+\vec{X}, -\vec{X}, +\vec{Y}, -\vec{Y}, +\vec{Z}, -\vec{Z}\}$, where $+\vec{X}$ denotes positive X-axis direction, etc. The second set is the sequence of direction (SoD), which contains five elements: *Forward*(F), *Left*(L), *Right*(R), *Up*(U) and *Down*(D) (see [79, 80] for similar representations). The procedure of calculating the sequence of directions can be described as follows:

1. Along the polymer chain, pick the first three bonds $(1, i, j)$ that are all perpendicular to each other, i.e., such that $\vec{B}_1 \perp \vec{B}_i \perp \vec{B}_j \perp \vec{B}_1$.

2. Then $\vec{B}_1$, $\vec{B}_i$ and $\vec{B}_j$ define a new coordinate system (i.e., new directions $+\vec{X}$, $+\vec{Y}$, and $+\vec{Z}$) in which we calculate the remaining bonds.

3. The first bond (D_1) is, by definition, the *Forward* direction, while the next non-forward bond (D_i) is defined as *Left*. The other directions are then determined in step 4.

4. Assign F to $D_2 \dots D_{i-1}$ and calculate D_h, $i < h < N$:

$$
D_h = \begin{cases}
F, & \text{IF} & \vec{B}_h = \vec{B}_{h-1} \\
L, & \text{ELIF} & \vec{B}_h = \vec{B}_{h-k} \\
 & \text{AND} & o(\vec{B}_{h-1}, \vec{B}_h) = s(\vec{B}_{h-1}, \vec{B}_h) \\
R, & \text{ELIF} & \vec{B}_h = \vec{B}_{h-k} \\
 & \text{AND} & o(\vec{B}_{h-1}, \vec{B}_h) \neq s(\vec{B}_{h-1}, \vec{B}_h) \\
U, & \text{ELIF} & o(\vec{B}_{h-k}, \vec{B}_{h-1}) = s(\vec{B}_{h-1}, \vec{B}_h) \\
 & \text{AND} & s(\vec{B}_h, +) = 1 \\
U, & \text{ELIF} & o(\vec{B}_{h-k}, \vec{B}_{h-1}) \neq s(\vec{B}_{h-1}, \vec{B}_h) \\
 & \text{AND} & s(\vec{B}_h, -) = 1 \\
D, & \text{ELIF} & o(\vec{B}_{h-k}, \vec{B}_{h-1}) = s(\vec{B}_{h-1}, \vec{B}_h) \\
 & \text{AND} & s(\vec{B}_h, -) = 1 \\
D, & \text{ELIF} & o(\vec{B}_{h-k}, \vec{B}_{h-1}) \neq s(\vec{B}_{h-1}, \vec{B}_h) \\
 & \text{AND} & s(\vec{B}_h, +) = 1
\end{cases}
$$

where $\vec{B}_{h-k}$ is the closest preceding element that satisfies $\vec{B}_{h-k} \neq \vec{B}_{h-1}$ and $o(\vec{B}_m, \vec{B}_n)$,

$s(\vec{B}_m, \vec{B}_n)$ are functions defined below:

$$o(\vec{B}_m, \vec{B}_n) = \begin{cases} 1\,, & (\left|\vec{B}_m\right|, \left|\vec{B}_n\right|) \in \{(X,Y),(Y,Z),(Z,X)\} \\ 0\,, & \text{otherwise} \end{cases}$$

$$s(\vec{B}_m, \vec{B}_n) = \begin{cases} 1\,, & \vec{B}_m,\, \vec{B}_n \text{ have the same sign} \\ 0\,, & \text{otherwise} \end{cases}$$

By this procedure we uniquely assign a SoD to each conformation and vice versa, taking the symmetries into account. That is, conformations are equal (modulo symmetry transformation of the cubic lattice) *iff* their SoD are identical.

4.5.2 Details of the algorithm

To obtain all the ground state structures of a given polymer sequence we perform a (replica-exchange) multicanonical sampling on the whole energy space. During that process, if the putative ground state energy, i.e., the lowest energy found during the (replica-exchange) Wang–Landau run, is met, we calculate the direction sequence of this state and compare it to those of previously found ground state structures, which we store in a tree structure container with a branching factor of at most five (the number of elements in an SoD). In this tree data structure, a direction sequence is uniquely represented by a path of length $N-1$ from the root node to a leaf node. Hence, the complexity of verifying a new found direction sequence is $\mathcal{O}(N)$ [81]. If the actual ground state is already present in that container, we just proceed. Otherwise, the new structure will be added to the database and the counter of degeneracy increased, as seen in Fig. 4.14. The simulation ends when a predefined number of MC steps have been performed with no change in ground-state degeneracy over a long period of time (e.g. 10% of the total MC steps). Note that even though we will use this method mainly to estimate ground state degeneracies, it is of course applicable to any other

energy level just as well.

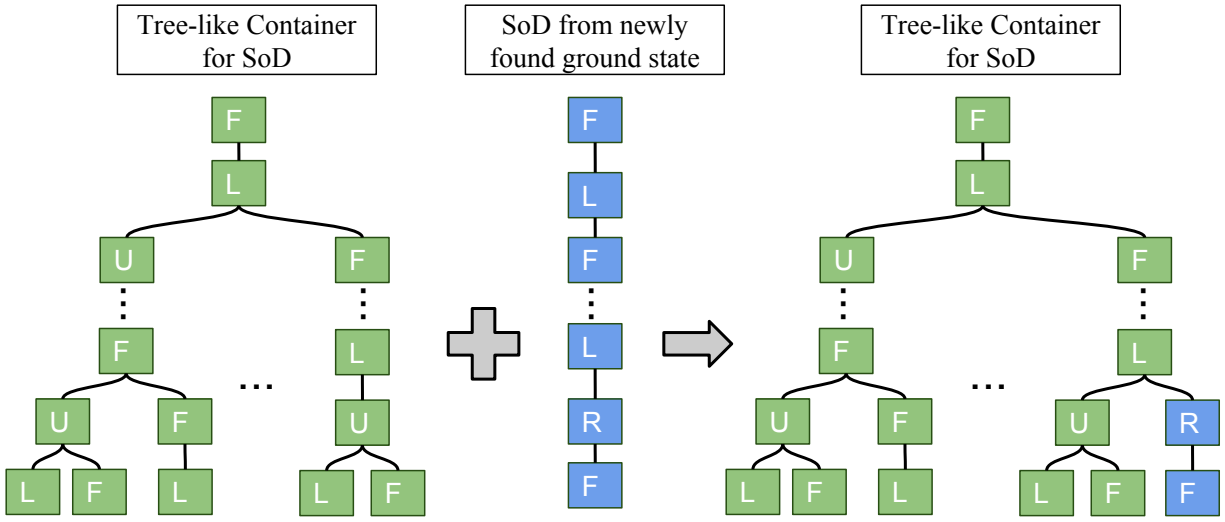


Figure 4.14: A schematic drawing for showing the procedure of inserting a SoD (sequence of direction) of a newly found ground state structure into a tree-like container.

4.5.3 Comparison with other methods

As a verification of our method, we chose 4 prominent HP sequences with length of 14, all of which have been studied using an exact enumeration method [32], and performed a multicanonical scan counting the absolute density of states. That is, we estimate the degeneracy of all energy levels analogously to the ground state sampling described above. The results are shown in Table 4.1, where the numbers of unique structures at each energy level are given. By identifying the dimension of each structure and considering different symmetries ($1\mathrm{D} \times 6; 2\mathrm{D} \times 24; 3\mathrm{D} \times 48$), we calculated the densities of states and found them to be exactly the same compared to enumeration results [32]. Each of these simulations took less than 4×10^9 Monte Carlo steps for $g(E_0)$ to converge. Even for a sequence with only 14 monomers, the total number of structures is close to 10^9. For longer sequences, the problem becomes tremendously more difficult and is only accessible through our method.

SeqID: 14.1 (*HPHPHHPHPHHPPH*)				
E	All	n_{2D}	n_{3D}	$g(E)$
−8	1	0	1	48
−7	262	0	262	12 576
−6	3 380	5	3 375	162 120
−5	28 163	84	28 079	1 349 808
−4	176 076	713	175 363	8 434 536
−3	754 422	3 809	7 50 613	36 120 840
−2	2 466 457	14 059	2 452 398	118 052 520
−1	6 533 719	38 605	6 495 114	312 691 992
0	9 758 750	52 912	9 705 837	467 150 070
SUM	19 721 230			943 974 510

SeqID: 14.2 (*HHPPHPHPHHPHPH*)				
E	All	n_{2D}	n_{3D}	$g(E)$
−8	2	0	2	96
−7	220	0	220	10 560
−6	2 929	4	2 925	140 496
−5	22 738	68	22 670	1 089 792
−4	139 052	561	138 491	6 661 032
−3	625 336	3 014	622 322	29 943 792
−2	2 102 592	10 872	2 091 720	100 663 488
−1	5 710 617	31 935	5 678 682	273 343 176
0	11 117 744	63 733	11 054 010	532 122 078
SUM	19 721 230			943 974 510

SeqID: 14.3 (*HHPHPHPPHPHPHH*)				
E	All	n_{2D}	n_{3D}	$g(E)$
−8	2	0	2	96
−7	200	1	199	9 576
−6	2 631	2	2 629	126 240
−5	21 987	68	21 919	1 053 744
−4	125 858	510	125 348	6 028 944
−3	591 753	3 110	588 643	28 329 504
−2	2 286 507	13 296	2 273 211	109 433 232
−1	6 392 045	37 796	6 354 249	305 911 056
0	10 300 247	55 404	10 244 842	493 082 118
SUM	19 721 230			943 974 510

SeqID: 14.4 (*HHPHPPHPHPHHPH*)				
E	All	n_{2D}	n_{3D}	$g(E)$
−8	4	0	4	192
−7	232	0	232	11 136
−6	3 348	7	3 341	160 536
−5	26 267	74	26 193	1 259 040
−4	163 540	757	162 783	7 831 752
−3	801 505	4 370	797 135	38 367 360
−2	2 702 687	15 734	2 686 953	129 351 360
−1	6 575 905	39 087	6 536 818	314 705 352
0	9 447 742	50 158	9 397 583	452 287 782
SUM	19 721 230			943 974 510

Table 4.1: Our Monte Carlo results for absolute densities of states for four 14mers. Columns from left to right: energy level, total number of structures of all dimensions, 2D structures (n_{2D}), 3D structures (n_{3D}) and $g(E) = 6\,n_{1D} + 24\,n_{2D} + 48\,n_{3D}$. Each sequence has only 1 1D-structure (with $E = 0$) which is not shown. Our results are identical to results from exact enumeration.

[1] values of $g^{\mathrm{L}}(E_0)$ can be referred to Yue and Dill [79].
[2] values of $g^{\mathrm{L}}(E_0)$ can be referred to Yue and Dill [82].
[3] values of $g^{\mathrm{L}}(E_0)$ can be referred to Bachmann and Janke [83]

	SeqID	E_0	$g^{\mathrm{L}}(E_0)$	$g(E_0)$
[1]	HP3D27.1	−16	36691	51537
	HP3D27.2	−15	297	297
	HP3D27.3	−16	25554	25554
	HP3D31	−28	1114	1114
[2]	HP3D42	−34	4	4
	HP3D67	−56	3	3
[3]	HP3D48.1	−32	$(5.2 \pm 0.8) \times 10^6$	$(10.3 \pm 0.4) \times 10^6$
	HP3D48.2	−34	$(1.7 \pm 0.8) \times 10^4$	$(2.84 \pm 0.02) \times 10^4$
	HP3D48.3	−34	$(6.6 \pm 2.8) \times 10^3$	5.09×10^3
	HP3D48.4	−33	$(6.0 \pm 1.3) \times 10^4$	$(4.97 \pm 0.16) \times 10^4$
	HP3D48.5	−32	$(1.2 \pm 0.3) \times 10^6$	$(1.94 \pm 0.04) \times 10^6$
	HP3D48.6	−32	$(9.6 \pm 1.9) \times 10^4$	$(1.84 \pm 0.02) \times 10^6$
	HP3D48.7	−32	$(5.8 \pm 2.1) \times 10^4$	$(10.8 \pm 0.1) \times 10^4$
	HP3D48.8	−31	$(2.2 \pm 0.7) \times 10^7$	$(1.59 \pm 0.03) \times 10^7$
	HP3D48.9	−34	$(1.4 \pm 0.5) \times 10^3$	2.614×10^3
	HP3D48.10	−33	$(1.9 \pm 0.9) \times 10^5$	$(5.53 \pm 0.14) \times 10^5$

Table 4.2: Estimated ground state degeneracy of some widely studied HP proteins. For each of them we listed the ground state energy E_0, the ground state degeneracy $g^{\mathrm{L}}(E_0)$ found in earlier studies and $g(E_0)$ estimated with our method. Converged sequences do not have statistical errors; otherwise error bars were obtained from multiple extrapolation fits (see text).

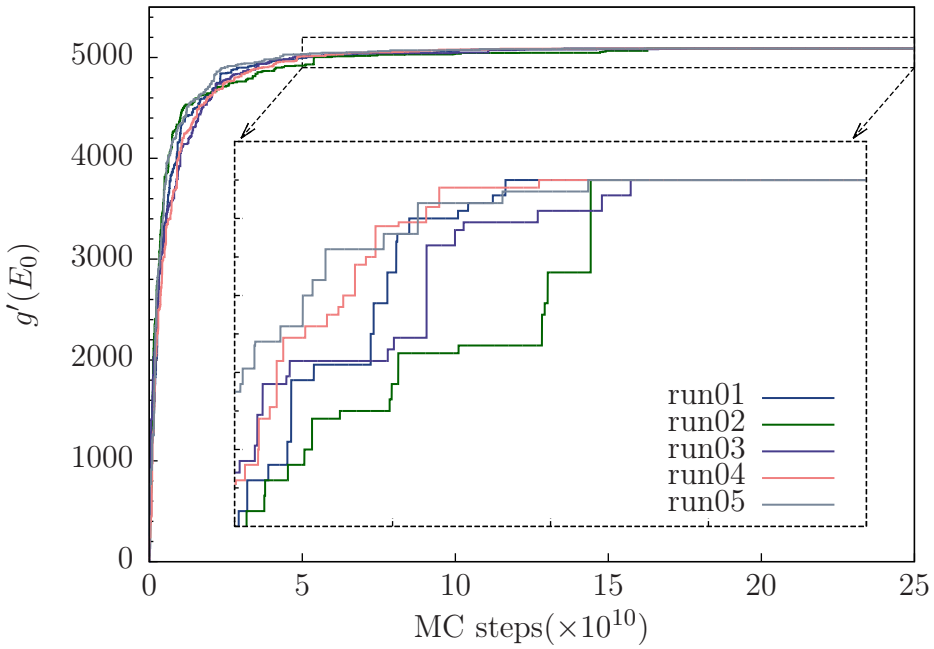


Figure 4.15: Number of different ground states found over time for HP sequence HP3D48.3 as seen in Table 4.2. $g'(E_0)$ is the actual estimator of GS degeneracy.

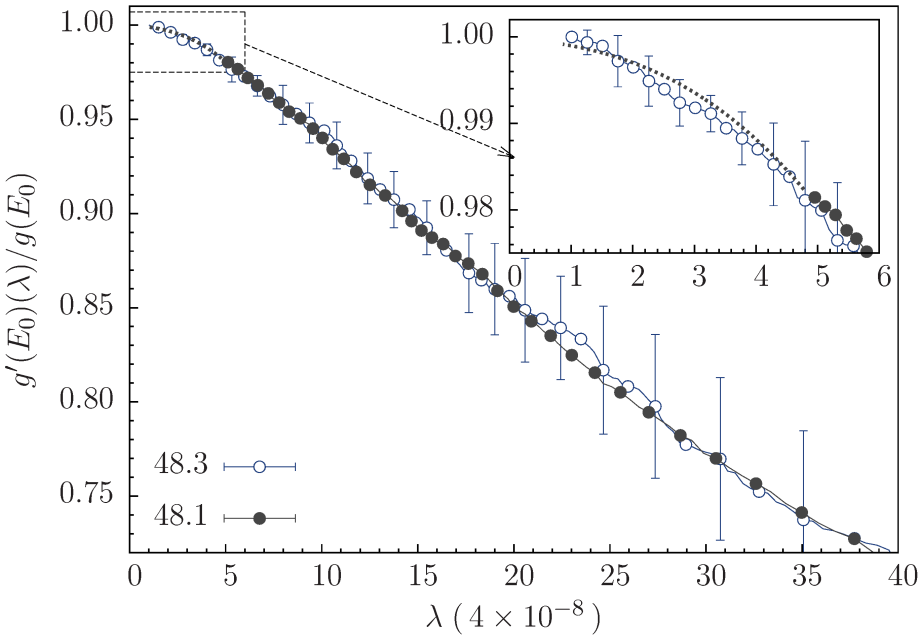


Figure 4.16: Number of different ground states found over time for HP sequences HP3D48.1 fitted to the corresponding curve of protein HP3D48.3. $g'(E_0)(\lambda)$ is the actual estimator of GS degeneracy and $\lambda = 1/(10000$ MC Steps) is the inverse Monte Carlo time.

After this proof of concept, some other widely studied HP sequences [79, 82, 83] have been chosen for testing our scheme. We carried out simulations for estimating the ground state degeneracy for each of these sequences, and listed the results in Table 4.2. For short sequences (e.g. HP3D27.2, HP3D27.3 and HP3D31) or sequences with low ground state degeneracy (e.g. HP3D42 and HP3D67), the results of our simulation agree with other studies perfectly. Due to the high ground state degeneracy, it is extremely challenging to reach all ground states for the 48mers [83] in finite simulation time (we ran up to 2.5×10^{11} MC steps). However, there are two sequences (HP3D48.3 and HP3D48.9) for which the ground state degeneracy stayed stable for a long time as seen in Fig. 4.15, and we thus believe we have converged to the true value $g(E_0)$. By normalizing the number of ground states and Monte Carlo time, we find that these two sequences share the same convergence behavior. The assumption of a fundamental convergence pattern provides a means to extrapolate the true ground state degeneracy for other long sequences. Hence, we fitted the normalized number of visited, different ground states vs. time for other sequences to the known curve for HP3D48.3 and extrapolated their ground state degeneracy. As an example, in Fig. 4.16 we show the corresponding fit for sequence HP3D48.1. The extrapolated part is marked by the dashed line in the figure. However, instead of a unique fit, there are multiple choices which fit equally well. Therefore, by doing 20 different fits, the average value as well as an error bar could be calculated. Even though not every curve fits as well as Fig. 4.16, it provides a better estimation of the true value $g(E_0)$. We note that our procedure yields rather different values than those obtained earlier using an approach where $g(E_0)$ is obtained from an implicit estimate of the partition function [83]. However, since our estimates are obtained from explicit enumeration of ground states with very high statistics, we believe that our procedure provides more reliable results.

Chapter 5

Effect of Mutations on Protein Folding

5.1 Mutations on lattice protein models

The Hydrophobic-polar lattice protein model has been found to have similar mutational properties compared to real proteins [84, 85]. A study on 2D HP-proteins with chain lengths ≤ 30 shows a single-mutation-induced fold switching [86], which has also been discovered in experimental studies [87, 88, 89, 90, 6, 91, 92]. Different from previous studies, which investigated relatively small systems on 2-dimensional simple cubic lattice, the main focus of this Chapter is the study of the effect of single-site substitution mutations on multiple longer HP sequences on 3-dimensional simple cubic lattice as a way of systematically approaching some of the questions introduced before, such as the sequence-structure relationship, on a very fundamental level.

A single-site mutation (SSM) of a HP protein consists of a 'flip' of one monomer from its original type to the other. For example, HP**H**HP becomes HP**P**HP under SSM on the third monomer. To understand possible effects of SSM on HP proteins, we choose two sequences

that were designed to study the origins of tertiary structures in proteins [82]:

$$\begin{aligned} \text{HP3D42}: &\quad \mathrm{P(H(HP)_2)_2HP_2H_3(PH)_2\ (HP)_2H_3P_2H((PH)_2H)_2P} \\ \text{HP3D67}: &\quad \mathrm{(PH(PH_2)_2PHP_2H_3PP)_3PH(PH_2)_2PHP_2H_3P} \\ &\quad \equiv \text{xPxPxPx, where x=}\mathrm{PH(PH_2)_2PHP_2H_3P} \end{aligned} \tag{5.1}$$

There are symmetries present in these two sequences: HP3D42 reads the same forward and backwards, and HP3D67 is composed of four identical pieces, each of which is represented by x in Eq. 5.1. Two xs are connected through a P monomer. The ground state structures of these two lattice proteins mimic construction of α/β-barrels and the β-helix, respectively (see Fig. 5.4 and 5.3). Note that the ground state degeneracy is extremely small (4 for the 42mer and 3 for the 67mer, as seen in Table 4.2). We systematically performed the SSM on each monomer of both HP chains and thus created 42 and 67 mutated sequences, respectively. We denote HP3D42sk as the mutated sequence generated by applying SSM on kth monomer of HP3D42 (and analogously for HP3D67). We performed simulations of each of these mutated sequences independently as described earlier (Ch. 4). Results are shown and discussed in the following sections.

5.2 Ground state degeneracies and structures

We are first interested in how the SSM affect the ground states (GS) of HP proteins in terms of their energy, actual conformations and degeneracy. For each of the 109 mutated sequences and two unmutated ones, we performed multiple, independent Wang–Landau sampling runs, followed by multiple MUCA runs up to 1×10^{11} MC steps for estimating the density of states and ground state degeneracies. In Figs. 5.1 and 5.2 we plot the ground state energy and degeneracy for all SSM of HP3D42 and HP3D67, respectively. For both sequences we find that the effect of SSM can vary significantly, depending on the monomer position. About

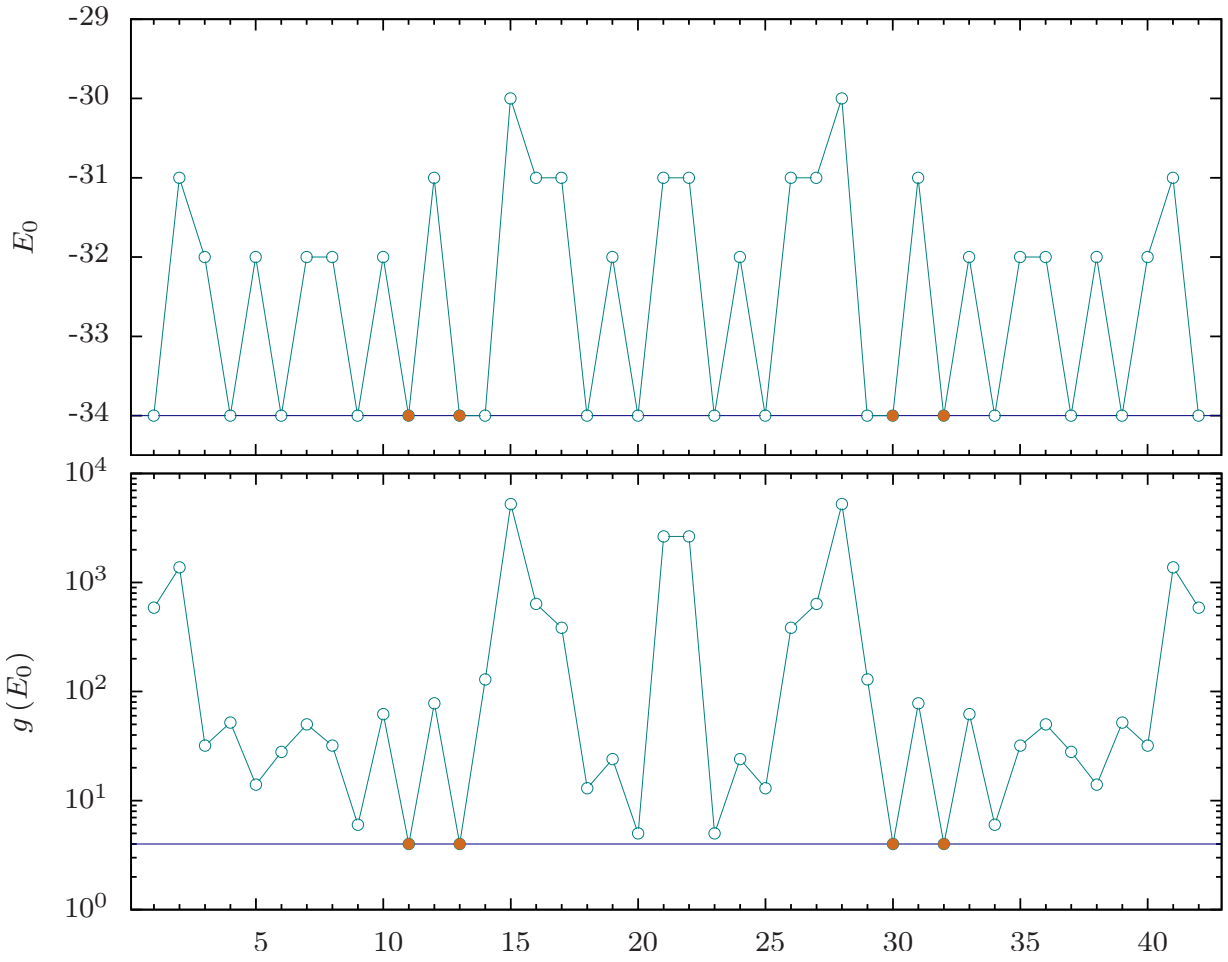


Figure 5.1: Ground state energy E_0 (top panel) and ground state degeneracy $g(E_0)$ (bottom panel) of mutated HP3D42. The X-axis value indicates the position which has been affected by the single-site mutation. Properties of the original, unmutated sequence are marked by horizontal lines. Positions where both ground state energy and ground state degeneracy are unchanged under the mutation are colored in orange.

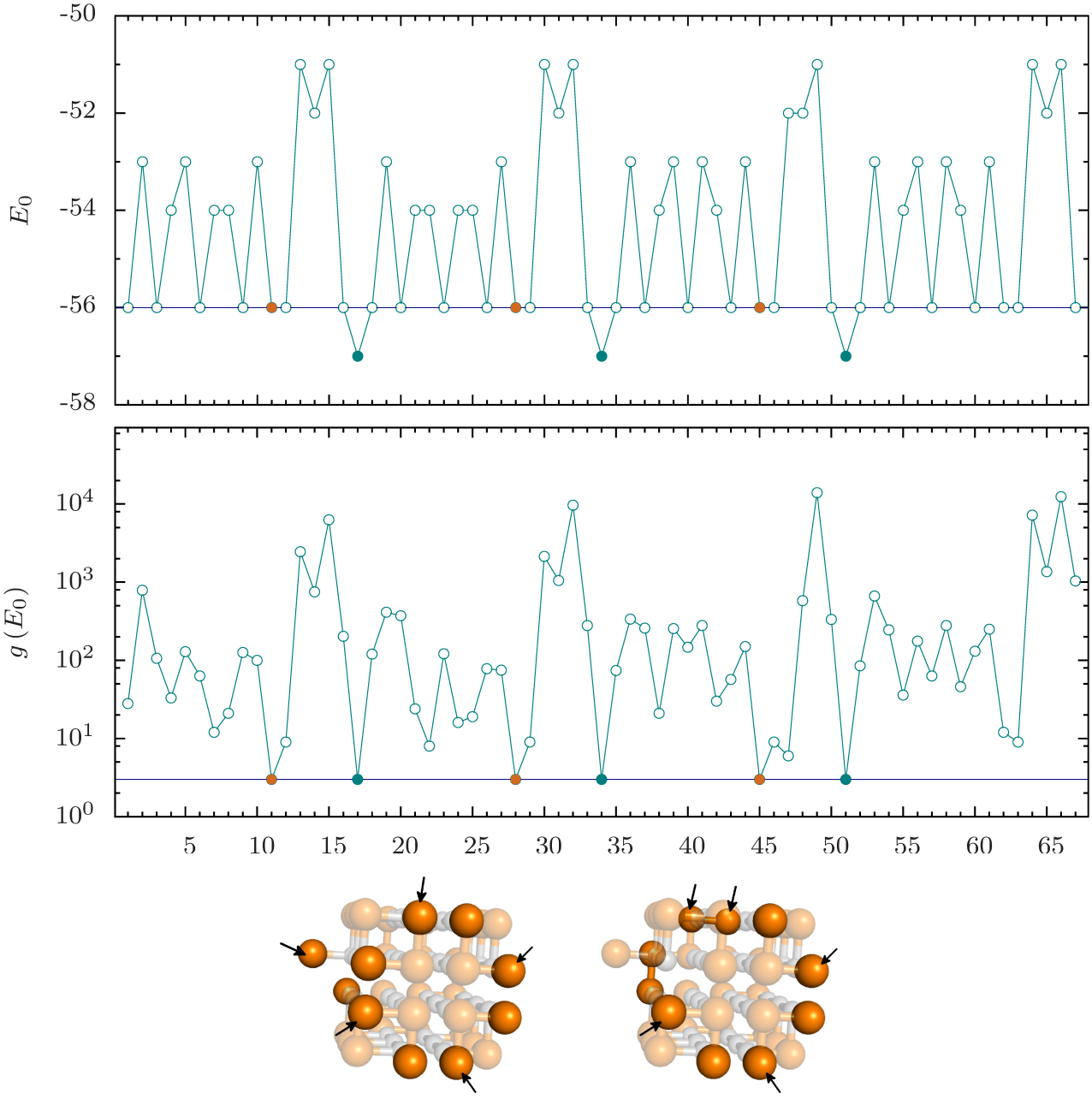
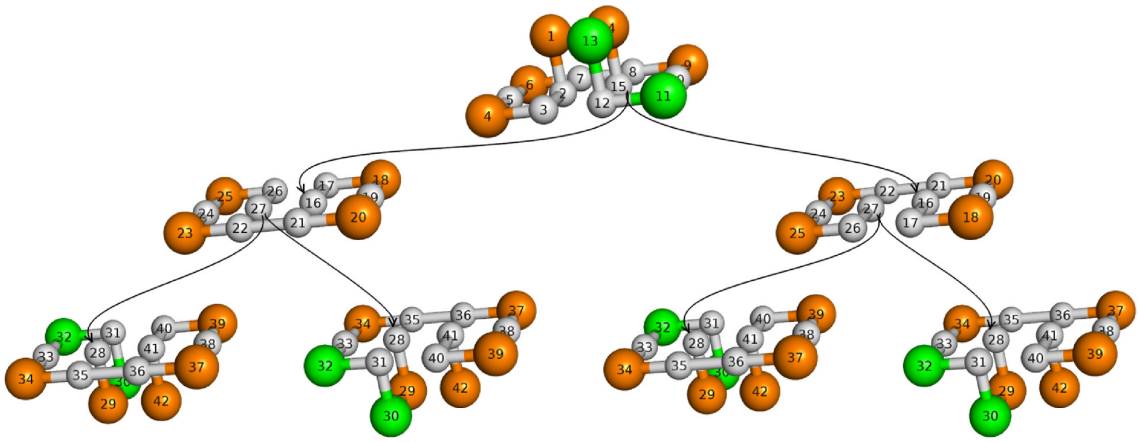


Figure 5.2: Ground state energy E_0 and ground state degeneracy $g(E_0)$ of mutated HP3D67 (cp. Fig. 5.1). Each of the bottom pictures shows the result of two overlapping ground state structures of HP3D67. Overlapped monomers are shown in shadowed color. Monomers pointed to by arrows belong to the same structure, while the rest belong to different structures. Positions where both ground state energy and ground state degeneracy are unchanged under the mutation are colored in orange. Cases where the ground state energy lowered by 1 while the ground state degeneracy kept unchanged are colored in green.

half of the mutated sequences retain their ground state energy (GSE), while others changed significantly. For example, mutations on the 15th monomer of HP3D42 or the 13th of HP3D67 change the GSE from $E = -34$ to -30 and from $E = -56$ to -51, respectively. Moreover, three mutated sequences (HP3D67s17, HP3D67s34 and HP3D67s51, using the notation defined at the end of Sec. 5.1) even have lower GSE ($E = -57$) compared to the original sequence. Interestingly, none of the mutated sequences has a GSE of $E = -33$ or -55, which corresponds to the first excited states of the unmutated HP3D42 and HP3D67, respectively.

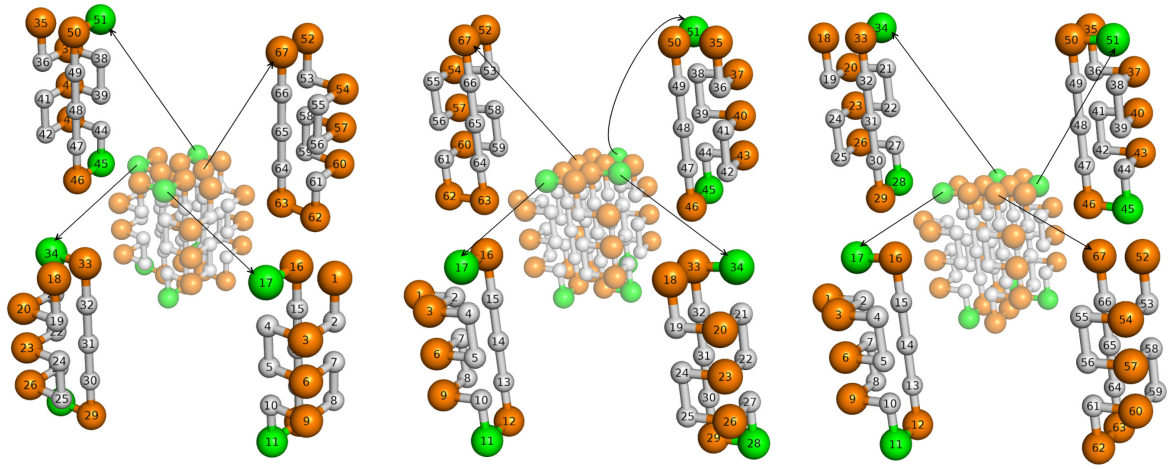
Regarding the ground state degeneracy (GSD), most of the mutated sequences show dramatically larger values than the unmutated ones. However, by comparing their ground state structures, we found that 88 out of 109 (36 for the 42mer and 52 for the 67mer) mutated sequences retain the ground state structures of the original sequence.

For HP3D67 we identified six P monomers (at positions 11, 17, 28, 34, 45 and 51 that are colored either in orange or green; cp. Fig. 5.2) which are "immune" against SSM in the sense that the ground state degeneracy and the actual ground state structures stay exactly the same except for the substituted site. For three of them (17, 34, and 51) SSM even results in lower ground state energy. Furthermore, we saw that the ground states remain unaffected under multiple site mutations, i.e., mutating up to all three sites simultaneously. We also find sequences where a single-site mutation does not affect the ground state energy but its degeneracy. In these cases we observe that SSM lowers the thermal stability of ground states by notably increasing the degeneracy of the first excited states, for example. Through examining the 3D ground state structures of the 67mer, we found that these "immune" positions are at the joints of lattice helices and lattice strands (cp. Fig. 5.4), while those extremely sensitive sites (e.g. 15, 32, 49) are usually located at the lattice strands (cp. Fig. 5.4). We note that the symmetries observed in Fig. 5.1 reflect the symmetry in the HP sequence of HP3D42 as expected, providing further evidence for the validity of our method.



(a) (b) (c) (d)

Figure 5.3: Four different ground state structures for the 42mer. Each ground state structure is sliced into three layers. The top layer is shared by all the ground state structures, while there are two different structures for middle and for bottom layers respectively which result in 4 ground state structures in total. Arrows in the figure pointed to the bonded monomer in the next layer. Hydrophobic monomers are colored in light grey, while polar monomers are either orange or green. Green colored monomers are those for which the ground state degeneracy remains the same under single-site mutation.



(a) (b) (c)

Figure 5.4: Three different ground state structures for the 67mer. Residues 2-10 form a lattice helix, and residues 12-16 form a lattice strand. Hydrophobic monomers are colored in light grey, while polar monomers are either orange or green. Green colored monomers are those for which the ground state degeneracy remains the same under single-site mutation. And these green monomers are also at the joints of lattice strands and lattice helices.

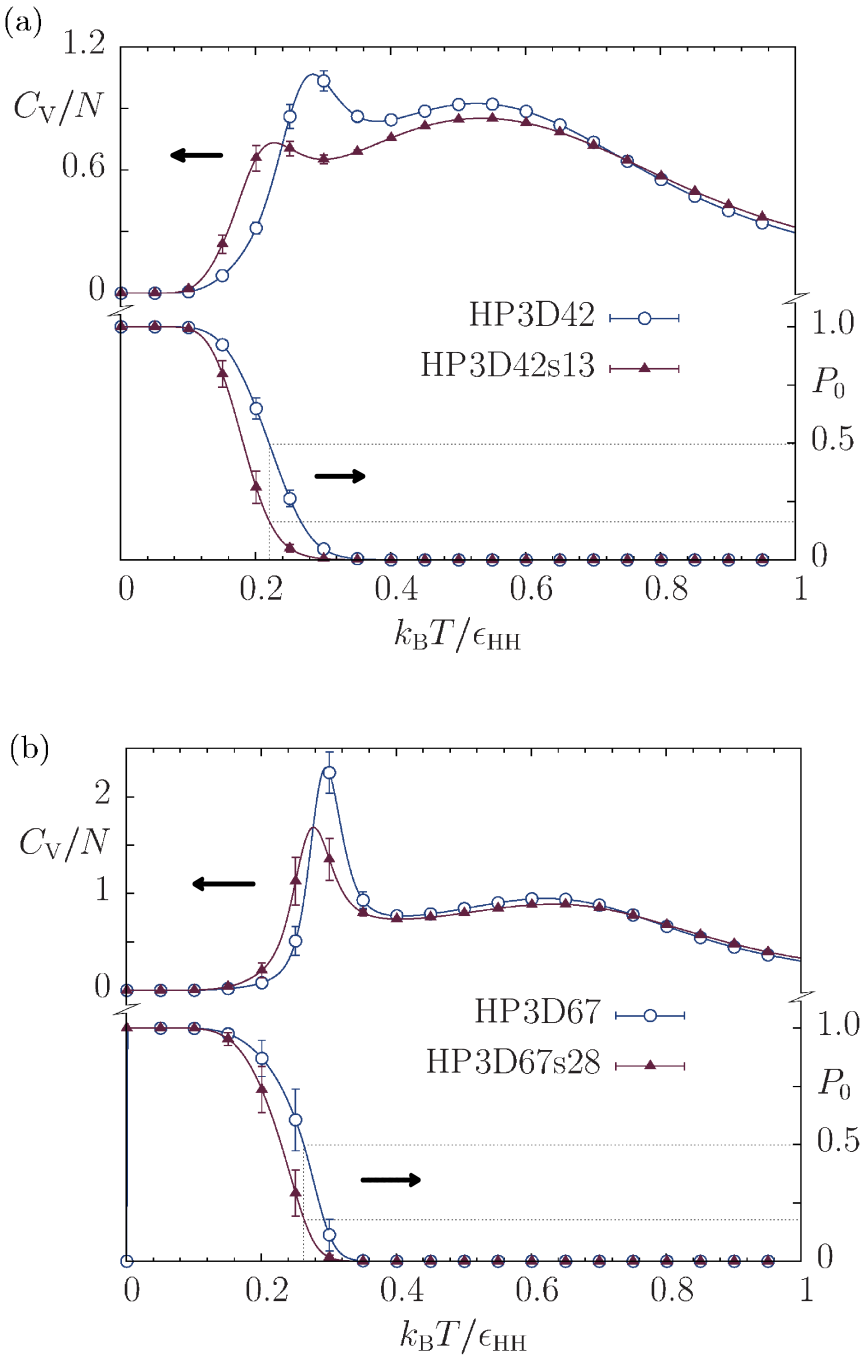


Figure 5.5: Examples: Thermal stability of ground state (a) Comparison between HP3D42s13 and HP3D42 based on specific heat (top curves, left scales) and ground state population (bottom curves, right scales) . (b) Comparison between HP3D67s28 and HP3D67 based on specific heat and ground state population ($P_0(T)$). In both figures, error bars smaller than data points are not shown.

One of the important properties we investigated is the ground state population $P_0(T)$ [86] which is defined as:

$$P_0(T) = \frac{g(E_0)\,\mathrm{e}^{-E_0/k_\mathrm{B}T}}{Z(T)}, \tag{5.2}$$

where E_0 is the ground state energy. In Fig. 5.5, we plot the specific heat $C_\mathrm{V}(T)/N$ (Eq. 4.3) and the ground state population $P_0(T)$ of two sequences for which both the ground state energy and degeneracy does not change compared to the corresponding unmutated sequence. We see a shift of the ground state population P_0 to lower temperatures. For example, at the temperature where 50% of the conformations in the canonical distributions of the unmutated sequences correspond to ground states, this percentage drops to 17%–18% for both mutated sequences (see grid-lines in Fig. 5.5). Thus, the ground state population is much more sensitive to temperature increase compared to the original protein. Looking at the specific heat, we also note a shift of the low-energy peak which corresponds to the formation of the compact hydrophobic core, i.e., this core breaks apart at lower temperatures as an effect of the mutation. Both observations show that the mutations lower the thermal stability of the ground states of the investigated HP proteins.

5.3 Thermodynamics of energetic and structural properties

Finally, we investigate in more detail how single-site mutations can affect the thermal behavior of HP proteins. Such knowledge could help unveil the effect of mutations on the folding process, for example. The quantities we are interested in here are the specific heat, the end-to-end distance, the tortuosity (τ) and the radius of gyration (R_g), as defined Sec. 4.1.1 and Sec. 3.5.

By examining all mutated sequences of HP3D42 and (most of) HP3D67, we have discovered cases where all quantities revealed very similar behavior compared to the unmutated sequences (see Fig. 5.6 (a) and Fig. 5.7 (a)). This comparison strongly indicates that those mutations do not affect the folding behavior significantly. More than 50% of all single-site mutations fall into this class, in which sequences contain the original ground state structures and add, if at all, only a small number to the ground state degeneracy.

On the other hand, there are instances where mutations significantly affect thermal quantities. As shown in Fig. 5.6 (b) and Fig. 5.7 (b), mutated sequences can present different behaviors in various ways. Typically, the left peak of heat capacity, which corresponds to the hydrophobic core formation (see, for example, Ref. [83] for a more detailed discussion of this transition), becomes lower or fades into a shoulder, along with raised or lowered τ and R_{ee}. Significant change of R_{g} has not been observed in all of our cases, which implies that this quantity is, not surprisingly for this model, quite stable under single-site mutation. Under this type of effects, mutated sequences also show up with sharp increase in ground state degeneracy and might lose the original ground state structures.

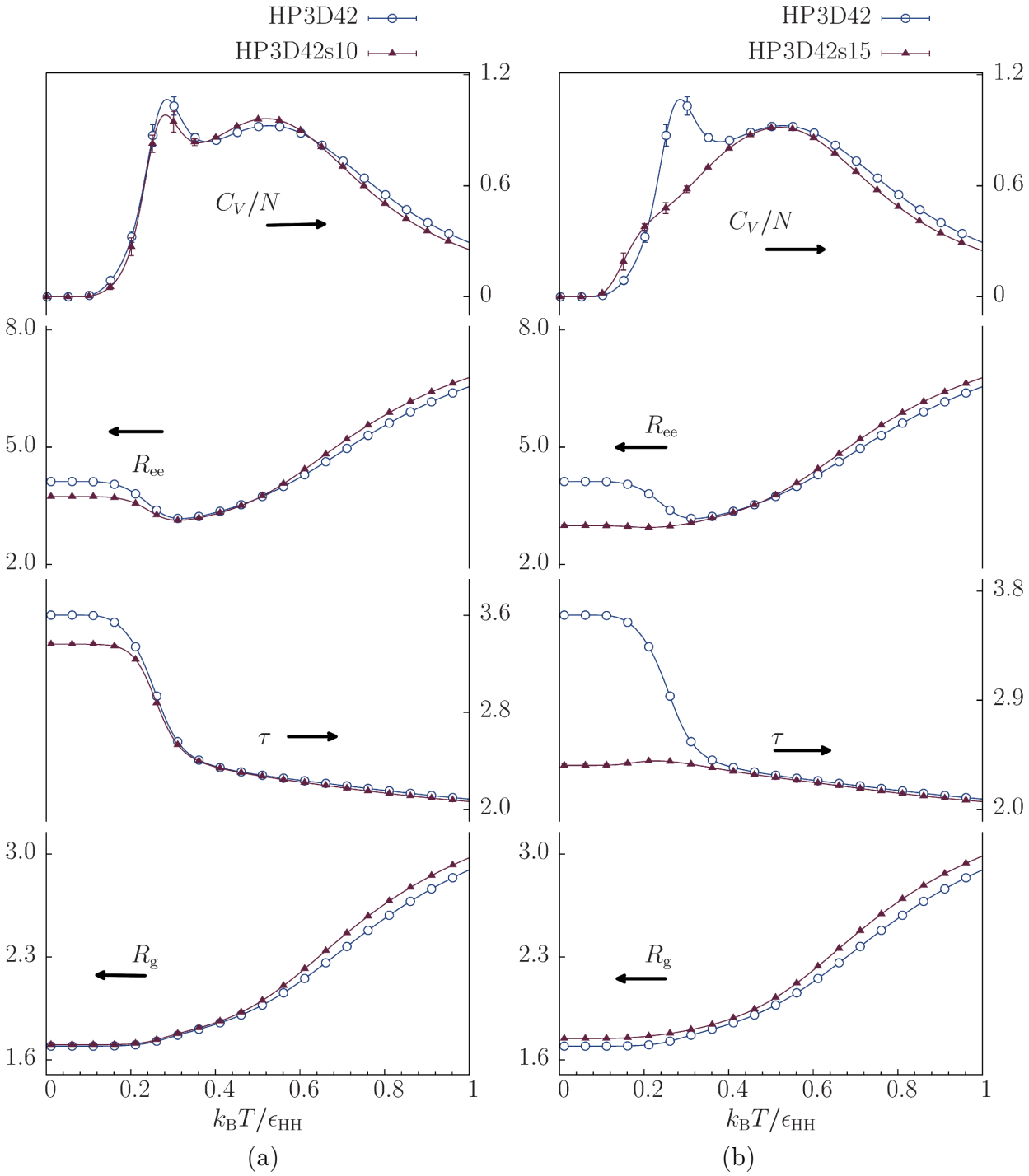


Figure 5.6: Effect of mutation on folding behavior for HP3D42: specific heat (C_V/N), end-to-end distance (R_{ee}), tortuosity (τ) and radius of gyration (R_g). (a) and (b) show respectively the cases where the mutation does not affect the folding behavior and where the thermodynamic quantities are changed under mutation. In all figures above, error bars smaller than data points are not shown.

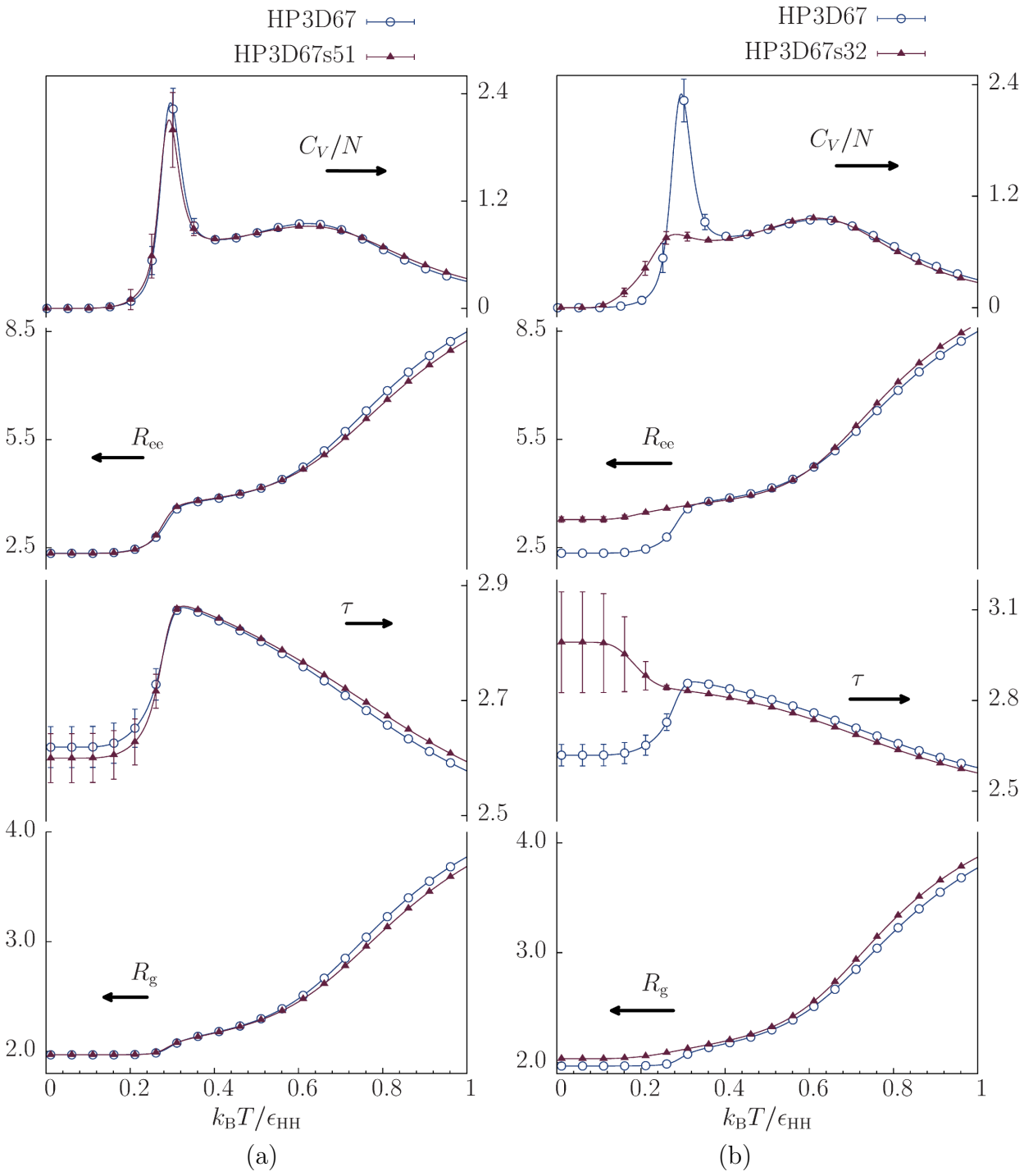


Figure 5.7: Effect of mutation on folding behavior for HP3D67: specific heat (C_V/N), end-to-end distance (R_{ee}), tortuosity (τ) and radius of gyration (R_g). (a) and (b) show respectively the cases where the mutation does not affect the folding behavior and where the thermodynamic quantities are changed under mutation. In all figures above, error bars smaller than data points are not shown.

Chapter 6

Lattice protein folding funnels

6.1 Introduction

As briefly discussed in Ch. 2.2, the resolution of Levinthal's paradox concerning the ability of proteins to fold rapidly postulates the existence of a rough, "folding funnel" in free energy space that "guides" the protein to its lowest free energy, native state. The funnel is always portrayed schematically as a relatively symmetric function of some unknown reaction coordinate about a unique minimum (the native state), as shown in Fig. 2.1. To calculate the free energy landscape with the reaction coordinates of choice, the knowledge of the density of states is essential. Let Q represents a structural quantity which can be used as a reaction coordinate. The joint density of states $g(E, Q)$ can be obtained through Monte Carlo methods, e.g. REWL, that are described in Ch. 4. The free energy can then be formulated as a function of Q and the temperature:

$$F\left(T, Q\right) = -k_{\mathrm{B}}\, T\, \ln Z(T, Q), \tag{6.1}$$

where $Z(T,Q)$ is the partition function based on temperature and end-to-end distance:

$$Z\left(T,Q\right)=\sum_{E}g(E,Q)\,e^{-E/k_{\mathrm{B}}T}. \tag{6.2}$$

In the rest of this Chapter, we will first make brief comparison between the HP model and the (semi-flexible) H0P model mappings for two real proteins as described in Ch. 3.4, based on the thermodynamic and structural behaviors. Then we will discuss some appropriate choices of reaction coordinates that might be used for describing protein folding funnels, which will be followed by the description of folding funnels for both the HP model and the semi-flexible H0P model.

6.2 The thermodynamic and structural behaviors

6.2.1 Crambin: HP3D46 vs H0P3D46

Crambin is a hydrophobic protein [93] with 46 residues. It has been converted into an HP sequence (denoted as HP3D46) by Lattman et al. [59] and also an H0P sequence (denoted as H0P3D46) as described in Ch. 3.4. It is interesting to note that the mapping onto an HP model yields a slightly polar protein whereas the H0P model mapping produces significantly more H-mers than P-mers as seen in Table. 3.2! We are interested in comparing the HP3D46 ($\epsilon_{\mathrm{HH}}=1$), with H0P3D46 on two different sets of coupling constants: H0P3D46 ($\epsilon_{\mathrm{HH}}=2$, $\epsilon_{\mathrm{H0}}=1$) and H0P3D46 ($\epsilon_{\mathrm{HH}}=4$, $\epsilon_{\mathrm{H0}}=2$, $\epsilon_{00}=1$). The specific values we have chosen for the coupling constants are based on the idea that the strength of H-H interactions should be stronger than H-0 interactions, which should be stronger than 0-0 interactions.

We performed multiple, independent Wang–Landau/replica-exchange Wang–Landau runs for determining the density of states of each model protein to high precision. As shown in Fig. 6.1 and Fig. 6.2, we present respectively the densities of states as well as specific heat

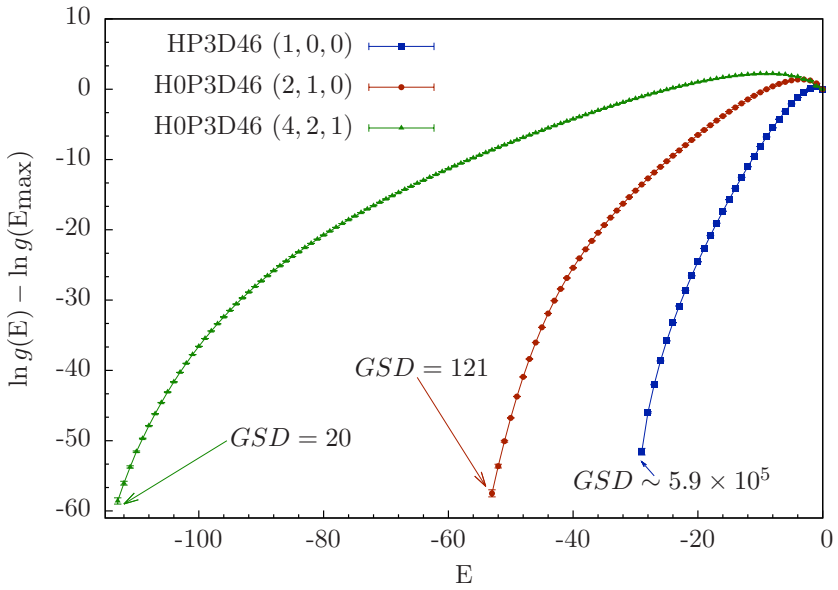


Figure 6.1: Densities of states (g(E)) of HP3D46 and H0P3D46 lattice models for Crambin. The ground state degeneracy (GSD) of each model is shown in the figure. Error bars smaller than data points are not shown; tuples in the legend indicate the values of (ϵ_{HH}, ϵ_{H0}, ϵ_{00}).

of all three model proteins. Based on the density of states, we then performed multiple, independent multicanonical/replica-exchange multicanonical sampling runs for estimating the ground state degeneracy and structural quantities. In Figure 6.1, along with densities of states, we labeled the ground state degeneracies (GSD) at the position of the lowest energy for each model. For each model, the GSD was obtained with 1×10^{12} MC steps, using the method described in Ch. 4.5. As seen in Fig. 6.1, the GSD has decreased significantly from HP3D46 to H0P3D46 (up to 4 orders of magnitude) due to the addition of the "neutral" type of monomer and the H-0 and the 0-0 interactions.

The specific heat (C_V/N) as shown in Fig. 6.2 indicate that different from HP3D46, both H0P3D46 proteins show clear two-step folding process signals as observed in all-atom simulations [94]. The coil-globule collapse transition occurs at the same temperature (around $T = 0.6$) for all three lattice models, but the rearrangement at lower temperature (around

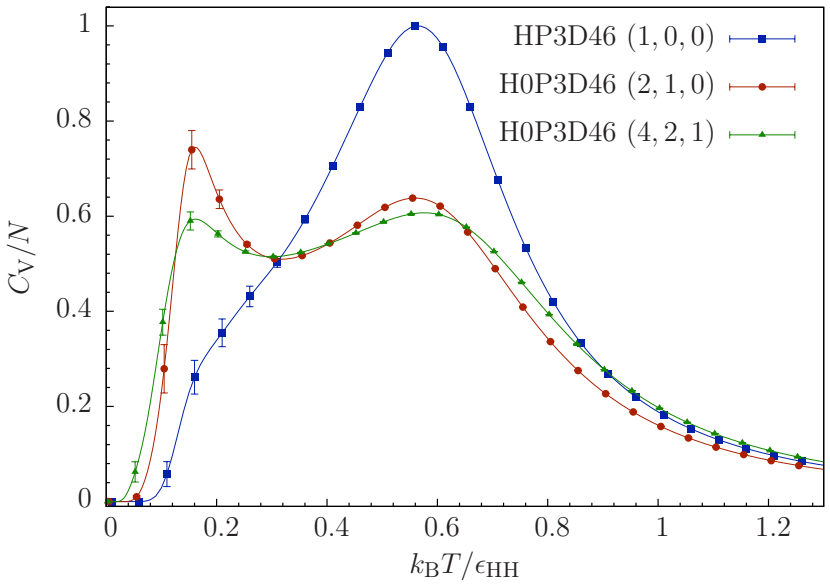


Figure 6.2: Specific heat (C_V/N) of HP3D46 and H0P3D46 lattice models for Crambin. Error bars smaller than data points are not shown; tuples in the legend indicate the values of (ϵ_{HH}, ϵ_{H0}, ϵ_{00}).

$T = 0.2$) clearly depends upon the details of the model. In the following, this can be further explained by investigating the behavior of the radius of gyration (R_g) for each monomer type and the number of non-bonded contacts.

In Fig. 6.3, we present the radius of gyration and the number of contacts for HP3D46. For the coil-globule collapse transition ($T \approx 0.6$), R_g, R_g(H) and R_g(P) all go through a sharp decrease, as lowering the temperature, which can be confirmed by looking at the sharp peaks of dR_g/dT, $dR_g(\mathrm{H})/dT$ and $dR_g(\mathrm{P})/dT$ around this temperature. As a result of this collapsing, all different types of contacts (n_{HH}, n_{HP}, n_{PP}) are increasing as seen in Fig. 6.3. As the temperature is lowered $T : 0.6 \rightarrow 0.2$, n_{HH} and n_{PP} continue to increase until saturated at $T \approx 0.2$; as the opposite, the decrease of n_{PP} is observed, which becomes stable around $T \approx 0.2$ as well. At the same time, the radius of gyration continues to decrease with a much smaller rate compared with previous processes.

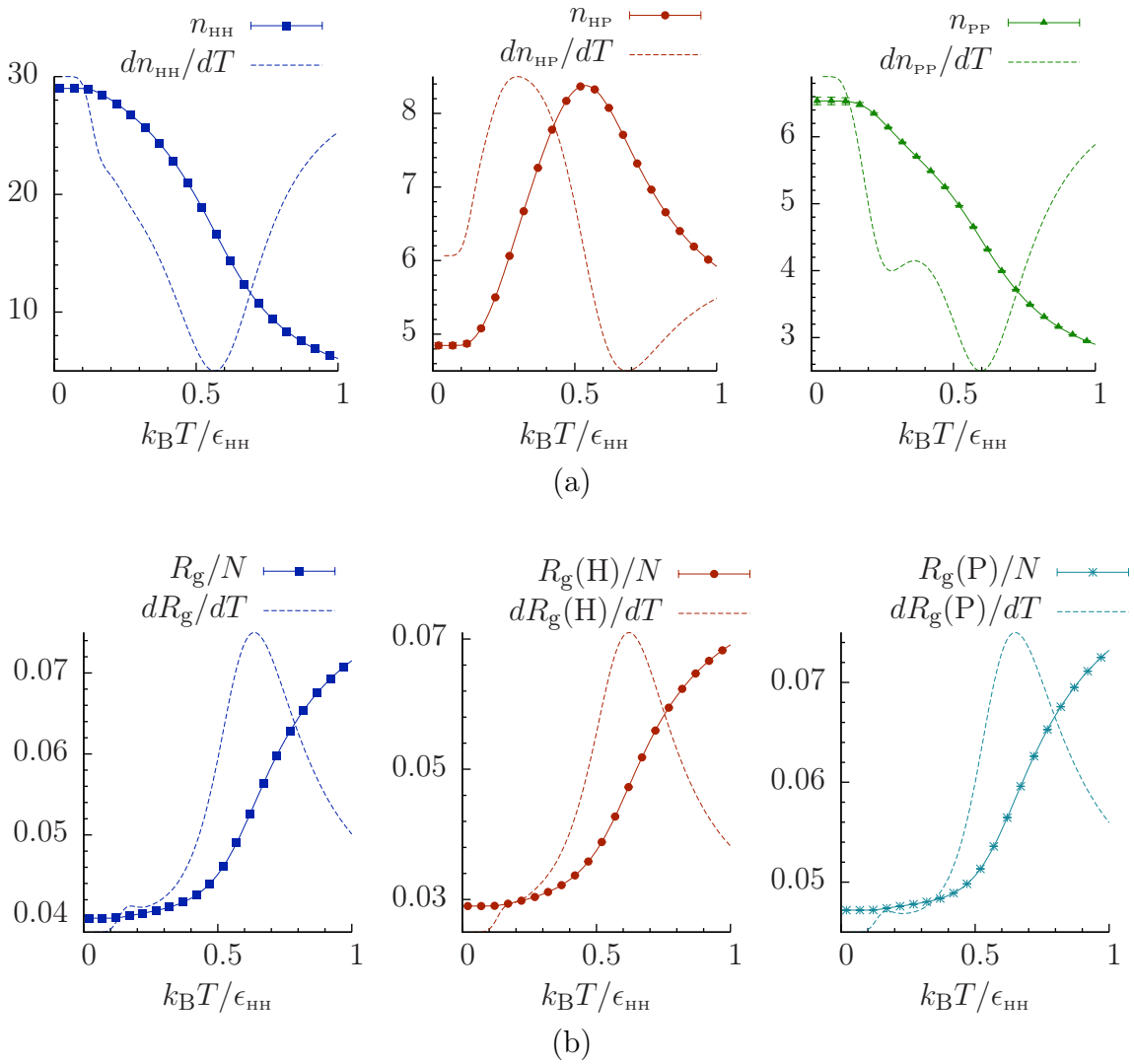


Figure 6.3: For HP3D46 with $\epsilon_{\rm HH} = 1$ and zero for the rest of the interactions: (a) number of non-bonded contacts for different monomer types: $n_{\rm HH}$, $n_{\rm HP}$ and $n_{\rm PP}$; (b) radius of gyration for whole chain ($R_{\rm g}$), and that for each monomer type: $R_{\rm g}({\rm H})$ and $R_{\rm g}({\rm P})$. In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

In Fig. 6.4 and Fig. 6.5, we respectively show the radius of gyration and the number of contacts for H0P3D46 with only H-H and H-0 interactions contributing to the Hamiltonian. Even though the behaviors of these properties are very similar to those of HP3D46 at $T \approx 0.6$, differences are observed at lower temperatures. For example, with lowering the temperature, a drastic increase of n_{H0} (Fig. 6.4) is observed at $T \approx 0.2$, which contributes to the pronounced peak in the specific heat (Fig. 6.2) at the same temperature. At the same time, as the protein rearranges into native structures, instead of continuously decreasing, $R_{\mathrm{g}}(0)$ bounces up after reaching the minimum.

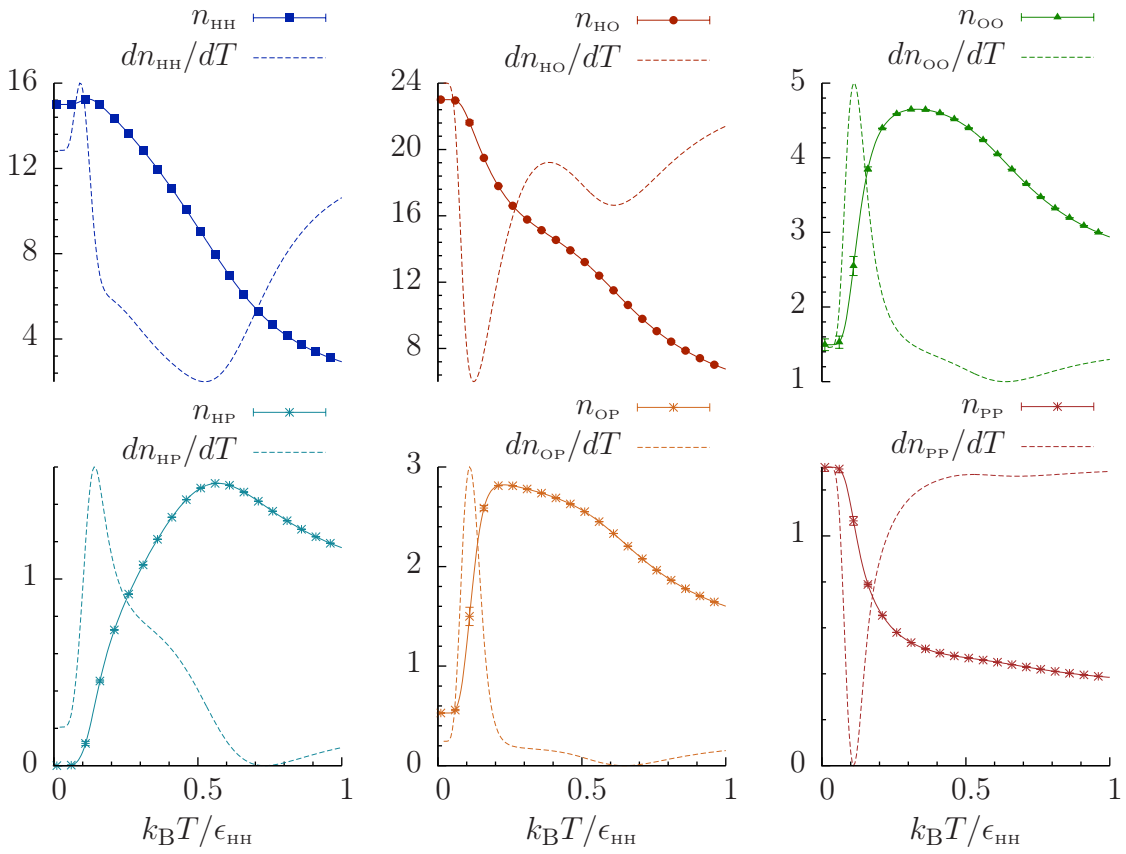


Figure 6.4: For H0P3D46 with $\epsilon_{\mathrm{HH}} = 2$, $\epsilon_{\mathrm{H0}} = 1$ and zero for the rest of interactions: the number of non-bonded contacts for different monomer types: n_{HH}, n_{H0} n_{00}, n_{HP} $n_{0\mathrm{H}}$ and n_{PP}. In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

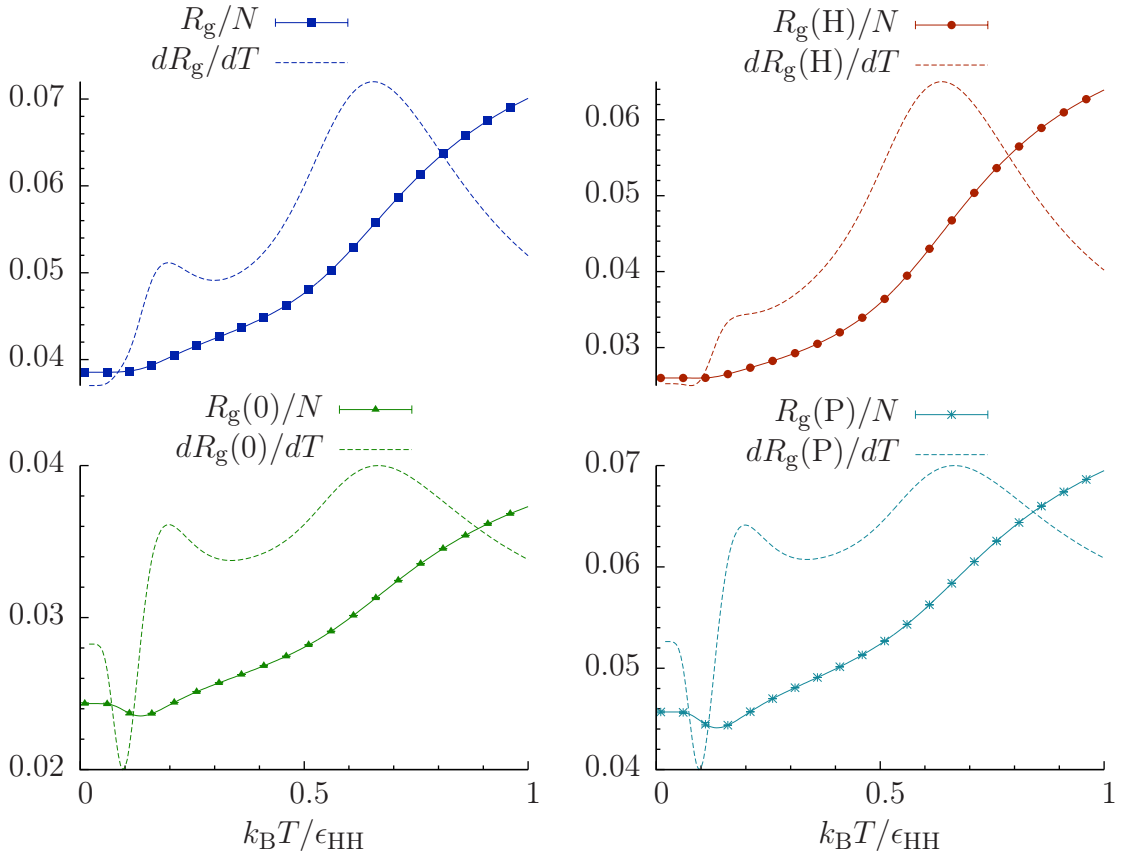


Figure 6.5: For H0P3D46 with $\epsilon_{HH} = 2$, $\epsilon_{H0} = 1$ and zero for the rest of interactions: radius of gyration for whole chain (R_g), and that for each monomer type: R_g(H), R_g(0) and R_g(P). In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

Likewise, for H0P3D46 with H-H, H-0 and 0-0 interactions contributing to the Hamiltonian, the number of contacts and the radius of gyration are shown respectively in Fig. 6.6 and Fig. 6.7. Different from the previous case, the sharp peak in specific heat of this protein at $T \approx 0.2$ is due to the drastically increase of both n_{H0} and n_{00} as seen in Fig. 6.6. Moreover, as this protein rearranges into native structures, drastically decreased $R_{\mathrm{g}}(0)$ and $R_{\mathrm{g}}(\mathrm{P})$ are observed as seen in Fig. 6.7.

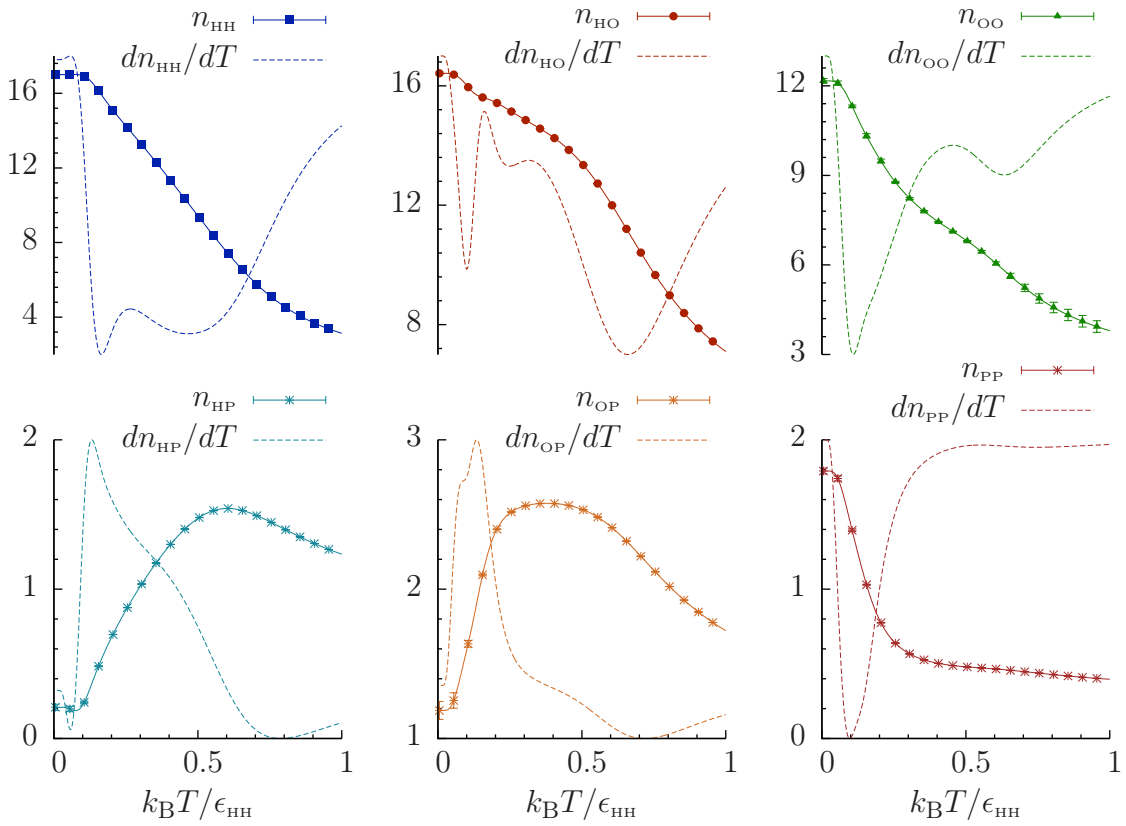


Figure 6.6: For H0P3D46 with $\epsilon_{\mathrm{HH}} = 4$, $\epsilon_{\mathrm{H0}} = 2$, $\epsilon_{00} = 1$ and zero for the rest of interactions: the number of non-bonded contacts for different monomer types: n_{HH}, n_{H0} n_{00}, n_{HP} n_{0H} and n_{PP}. In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

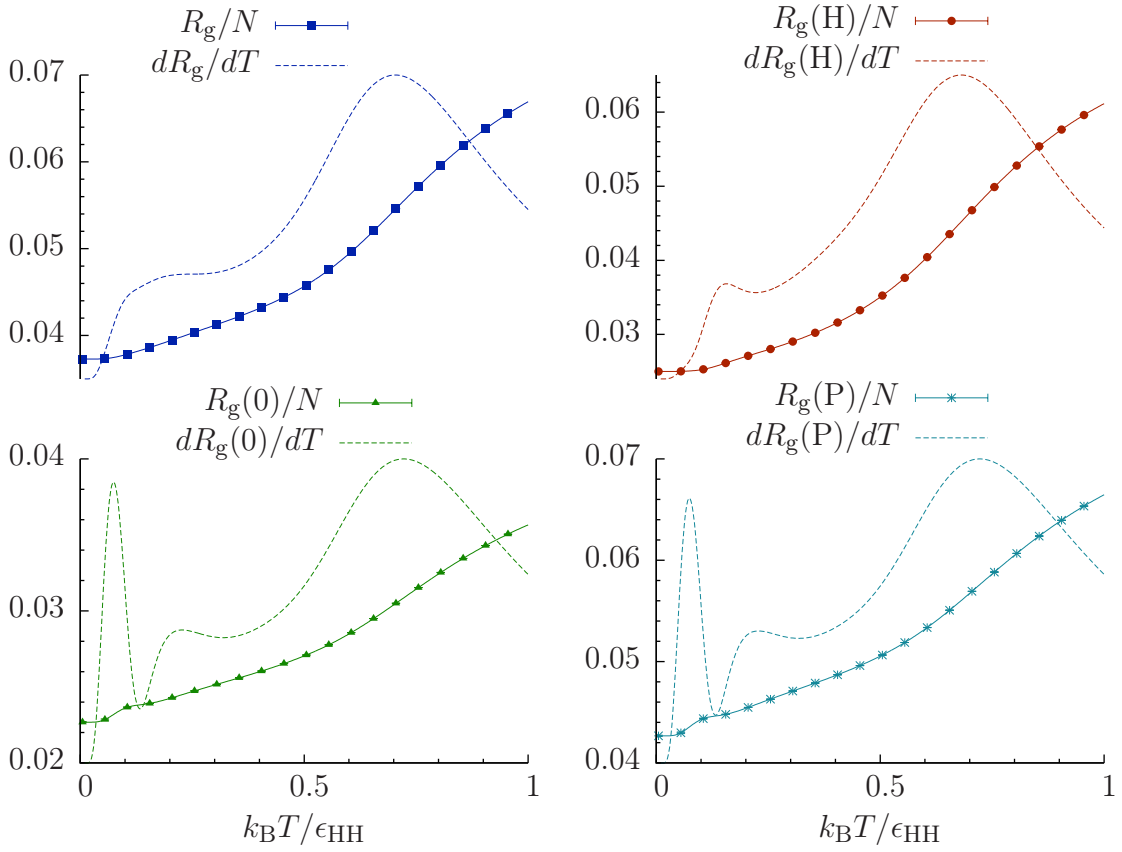


Figure 6.7: For H0P3D46 with $\epsilon_{HH} = 4$, $\epsilon_{H0} = 2$, $\epsilon_{00} = 1$ and zero for the rest of interactions: radius of gyration for whole chain (R_g), and that for each monomer type: R_g(H), R_g(0) and R_g(P). In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

6.2.2 Ribonuclease A: HP3D124 vs H0P3D124

Ribonuclease A has been converted into a HP sequence (denoted as HP3D124) by Lattman et al. [59] and also an H0P sequence (denoted as H0P3D124) as described in Ch. 3.4. For HP3D124, the coupling constant $\epsilon_{\text{HH}} = 1$ and zero for the rest; for H0P3D124 the $\epsilon_{\text{HH}} = 4$, $\epsilon_{\text{H0}} = 2$, $\epsilon_\theta = -1$ and zero for the rest of the interactions. The specific values we have chosen for the coupling constants are based on the idea that the strength of H-H interaction should be stronger than H-0 interaction, and the stiffness of the bond should be contributing the least to the Hamiltonian.

For a protein with length > 100, obtaining the density of states covering the whole energy space is already extremely difficult even with Wang–Landau sampling as discussed in ref. [64]. We performed multiple, independent replica-exchange Wang–Landau runs for determining the density of states of each model protein to high precision. These are followed by multiple, independent replica-exchange multicanonical sampling runs for estimating the structural quantities.

As seen in Fig. 6.8, the specific heat and end-to-end distance (as introduced in Ch. 3.5) for HP3D124 both show a clear protein collapse "transition" near $T \approx 0.5$ followed by a very slight "bump" at quite low T. Typical protein configurations in Fig. 6.8 show this folding process including one of the degenerate ground states. We also present the radius of gyration and the number of contacts for HP3D124 in Fig. 6.9. The behaviors of these two properties are very similar to those of HP3D46. That is, at $T \approx 0.5$, the protein already collapses into very compact structures, which are indicated by R_{g} in Fig. 6.9 (b). Moreover, through the method described in Ch. 4.5 with 1×10^{13} MC steps, we estimated the ground state degeneracy for HP3D124 to be $\sim 1.4 \times 10^6$, i.e. far above the "desired" unique native state.

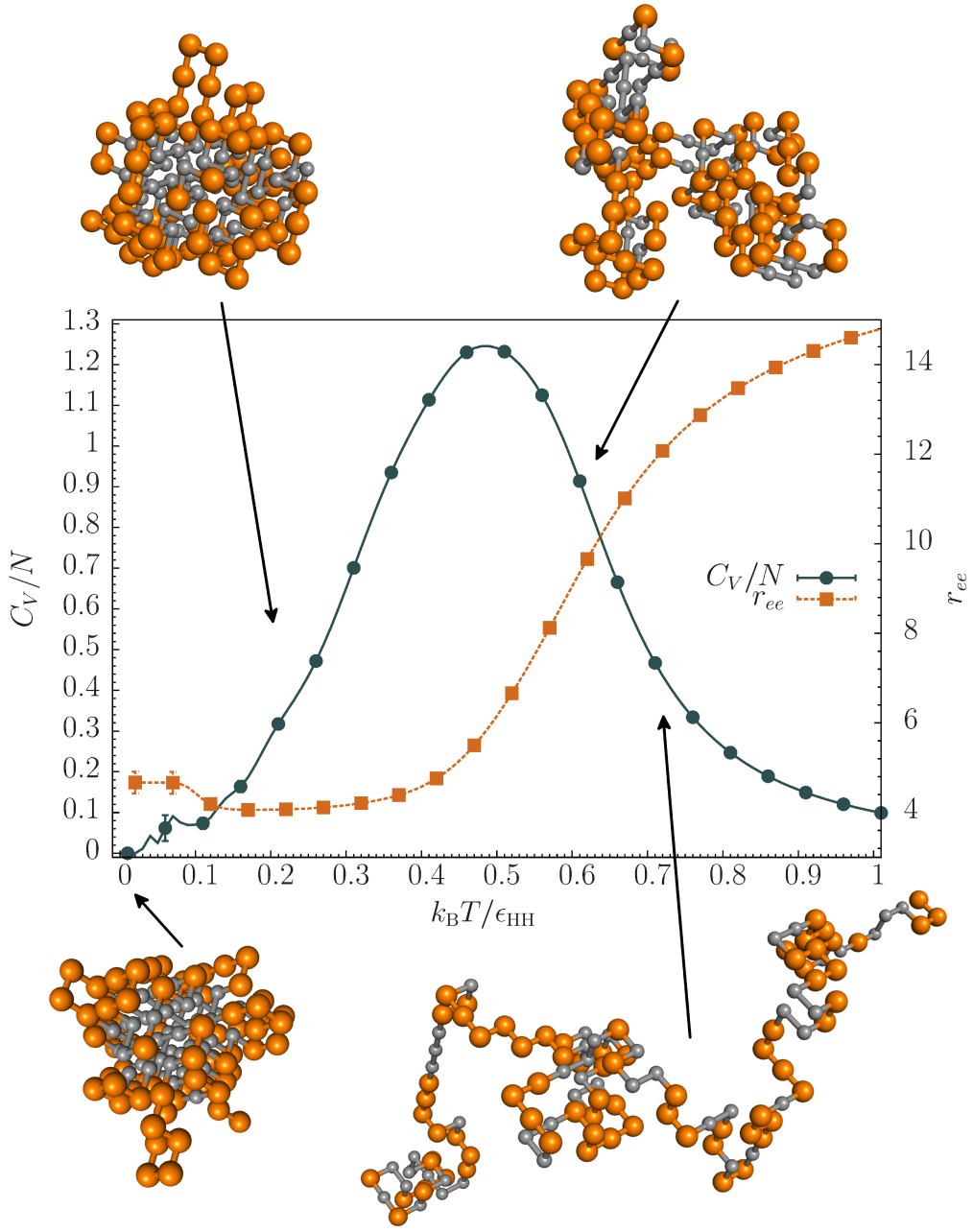


Figure 6.8: The specific heat (C_V/N) and end-to-end distance vs temperatures for the HP3D124 lattice protein model, i.e., $\epsilon_{\rm HH} = 1$ and zero for the rest of the interactions. Typical configurations are shown at the indicated temperatures: Hydrophobic monomers are colored dark gray while polar monomers are colored orange. Error bars smaller than the data points are not shown.

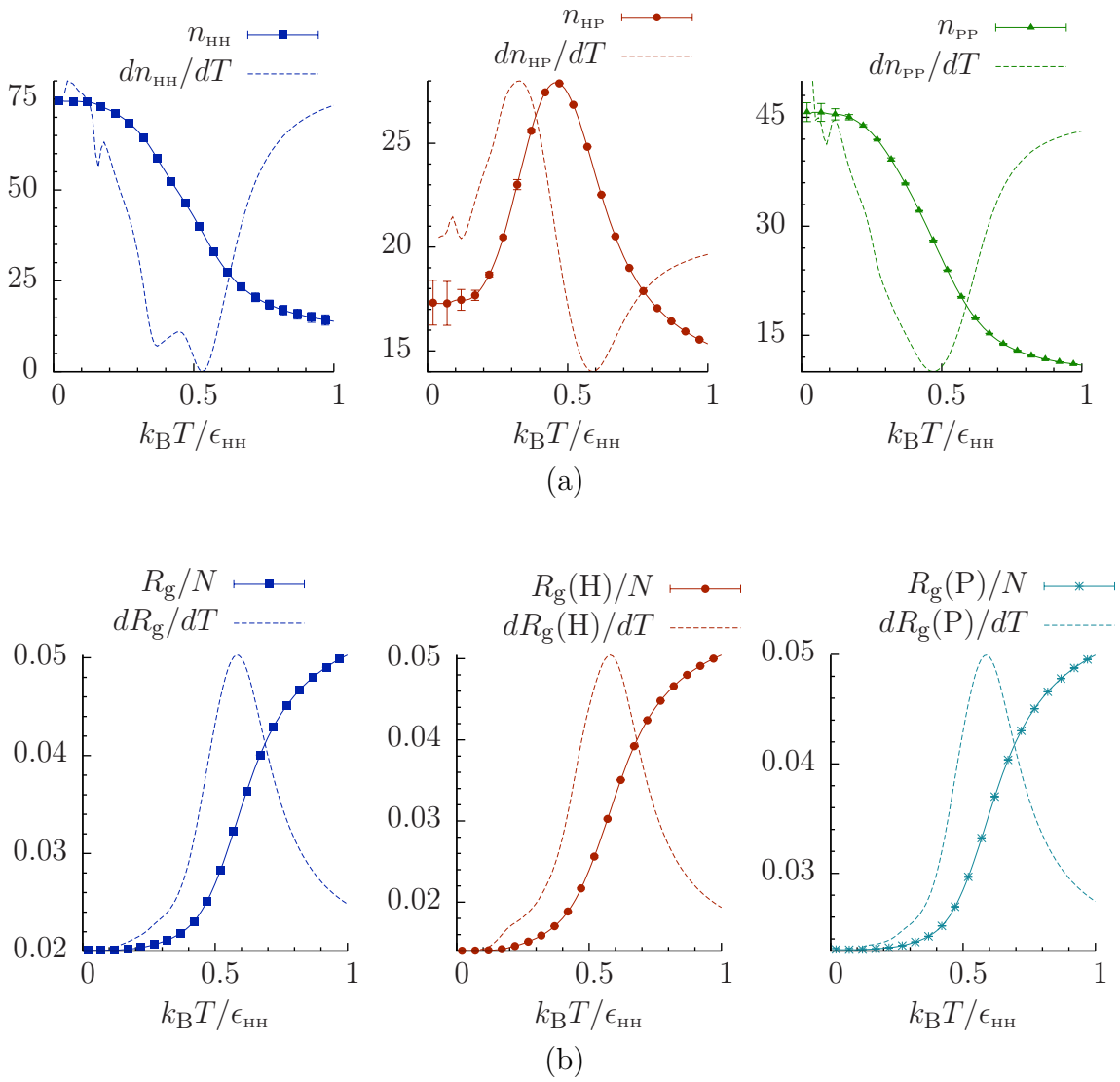


Figure 6.9: For HP3D124 with $\epsilon_{\mathrm{HH}} = 1$ and zero for the rest of the interactions: (a) number of non-bonded contacts for different monomer types: n_{HH}, n_{HP} and n_{PP}; (b) radius of gyration for whole chain (R_{g}), and that for each monomer type: $R_{\mathrm{g}}(\mathrm{H})$ and $R_{\mathrm{g}}(\mathrm{P})$. In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

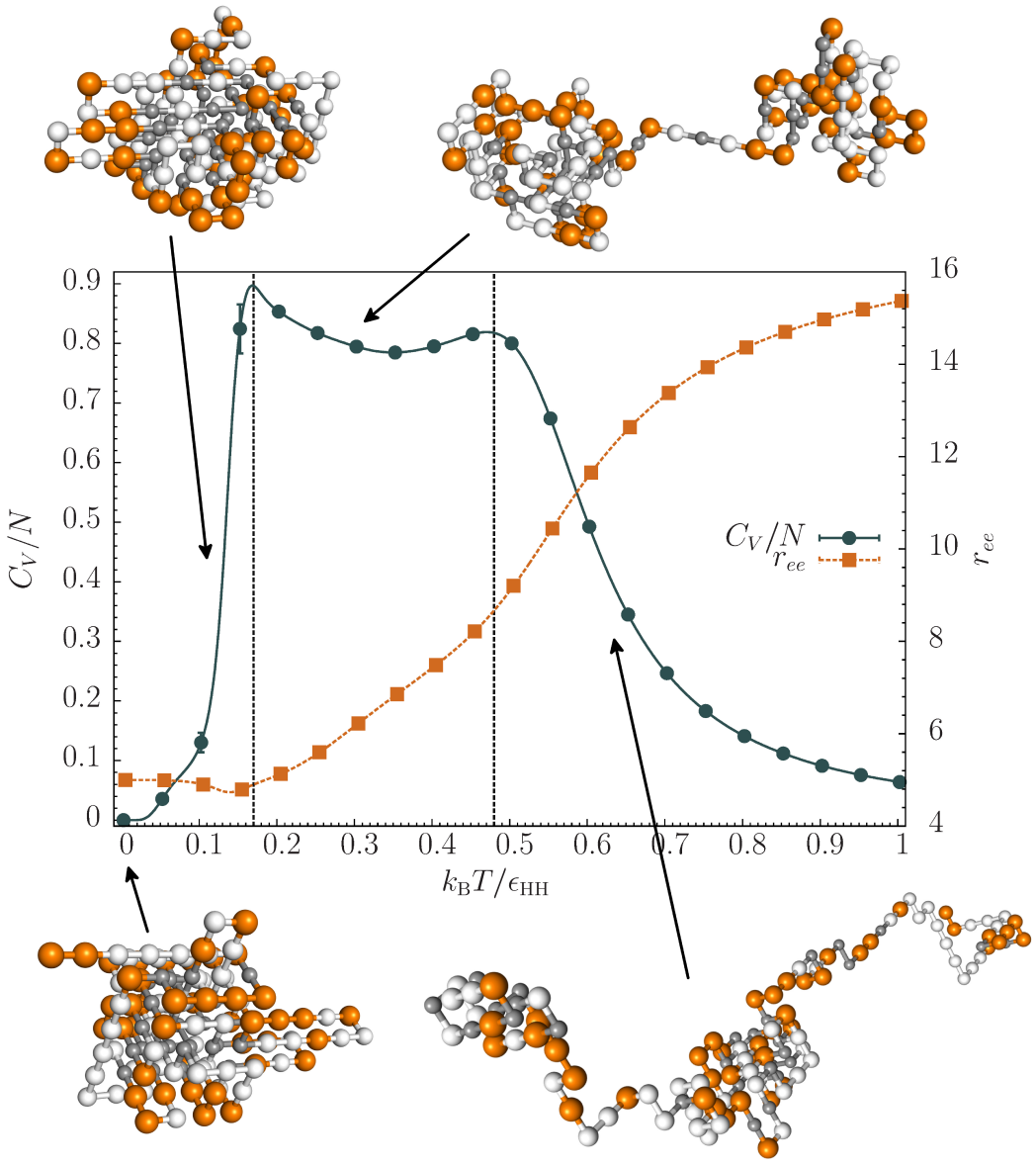
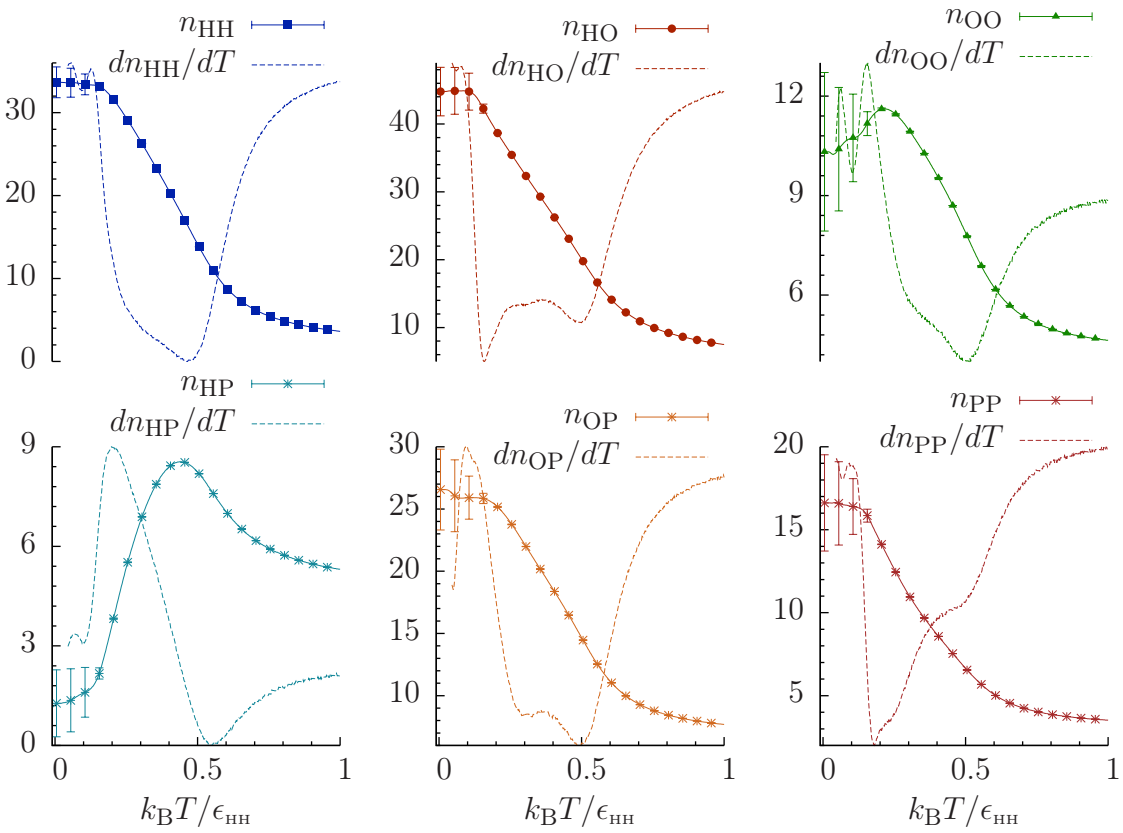
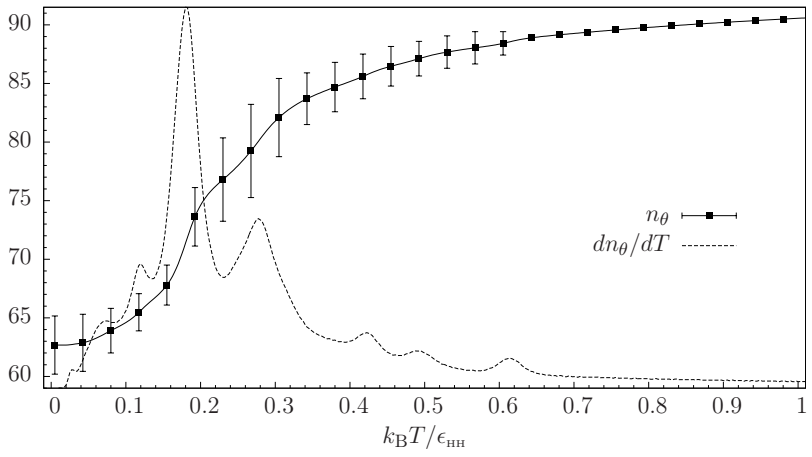


Figure 6.10: The specific heat (C_V/N) and end-to-end distance vs temperatures for the H0P3D124 lattice protein model, i.e., $\epsilon_{\mathrm{HH}} = 4$, $\epsilon_{\mathrm{H0}} = 2$, $\epsilon_\theta = -1$ and zero for the rest of interactions. For structures shown, H- and '0'-mers are colored dark gray and white, respectively, while P-mers are colored orange. Error bars smaller than the data points are not shown.

For H0P3D124, as seen in Fig. 6.10, the lattice protein collapses from a random coil at a temperature $T \approx 0.5$. Different from the HP3D124 which is already very compact at this point as discussed before, the R_{g} of H0P3D124 still needs to go through a substantial decrease when $T : 0.5 \rightarrow 0.2$, as seen in Fig. 6.12. Moreover, when T is lowered to 0.2, H0P124, in contrast to HP3D124, exhibits a second peak in the specific heat which corresponds to the sharply increased n_{H0} (as seen in Fig. 6.11 (a)) and drastically decreased n_{θ} (as seen in Fig. 6.11 (b)). After studying large amount of the protein configurations at some temperatures between $T = 0.5$ and $T = 0.2$, we present some typical ones in Fig. 6.13. In this temperature region, instead of forming a highly compact hydrophobic core, most protein conformations are moderately collapsed, and some even obtain an extended, small tail. These along with typical configurations in Fig. 6.10 indicate that the first "transition" is only partial and the protein collapses completely only at the lower temperature. At very low temperature there is only a slight shoulder in the specific heat but the end-to-end distance (Fig. 6.10) still increases slightly. Using the method described in Ch. 4.5 with 1×10^{13} MC steps, we estimated the ground state degeneracy for H0P3D124 was greatly reduced and only 343 inequivalent ground states were found, i.e., a reduction of more than 4 orders of magnitude as compared to the HP model! This characteristic is much closer to what is expected for a real protein.



(a)



(b)

Figure 6.11: For H0P3D124 with $\epsilon_{\mathrm{HH}} = 4$, $\epsilon_{\mathrm{H0}} = 2$, $\epsilon_\theta = -1$ and zero for the rest of interactions: (a) the number of non-bonded contacts for different monomer types: n_{HH}, n_{H0} n_{00}, n_{HP} $n_{0\mathrm{H}}$ and n_{PP}; (b) the number of angles with respect to temperature. In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

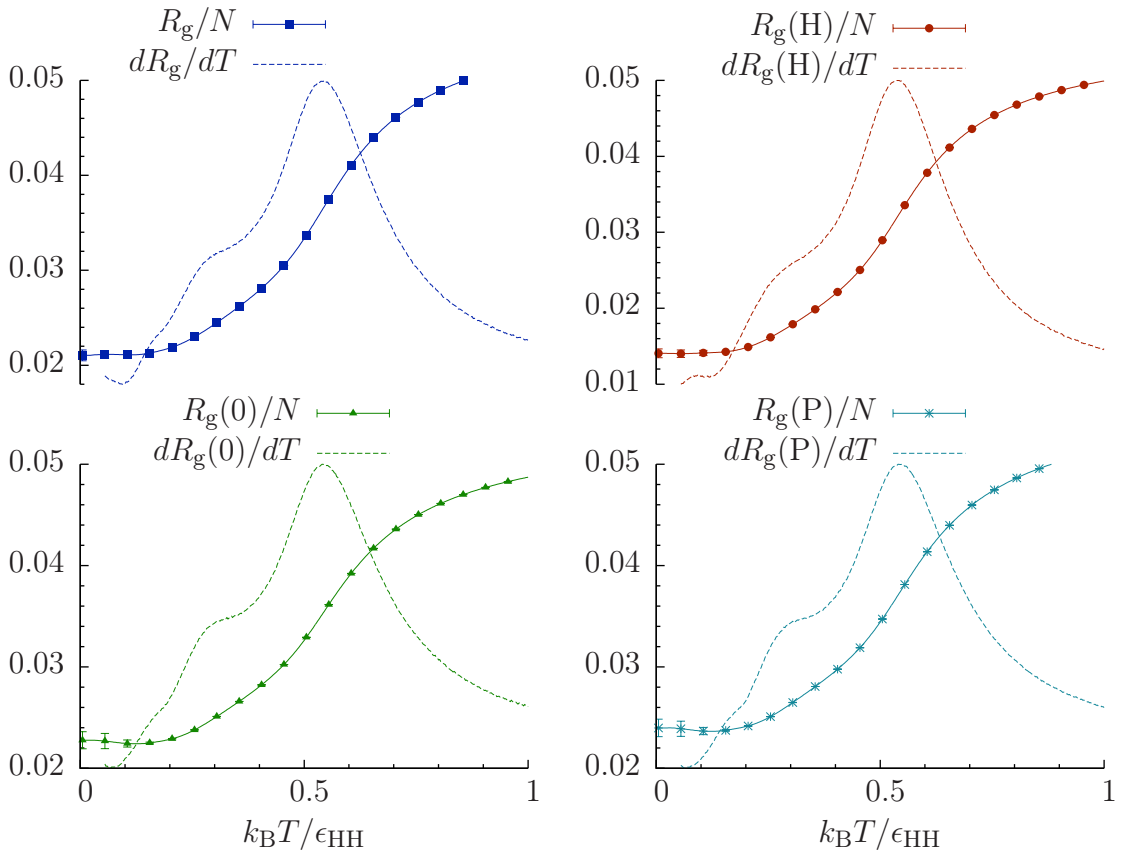


Figure 6.12: For H0P3D124 with $\epsilon_{\mathrm{HH}} = 4$, $\epsilon_{\mathrm{H0}} = 2$, $\epsilon_\theta = -1$ and zero for the rest of interactions: radius of gyration for whole chain (R_g), and that for each monomer type: R_g(H), R_g(0) and R_g(P). In each figure, the y-axis indicates the value of the solid line, the derivative of which with respect to the temperature is represented by the corresponding broken line. In all figures, error bars smaller than data points are not shown.

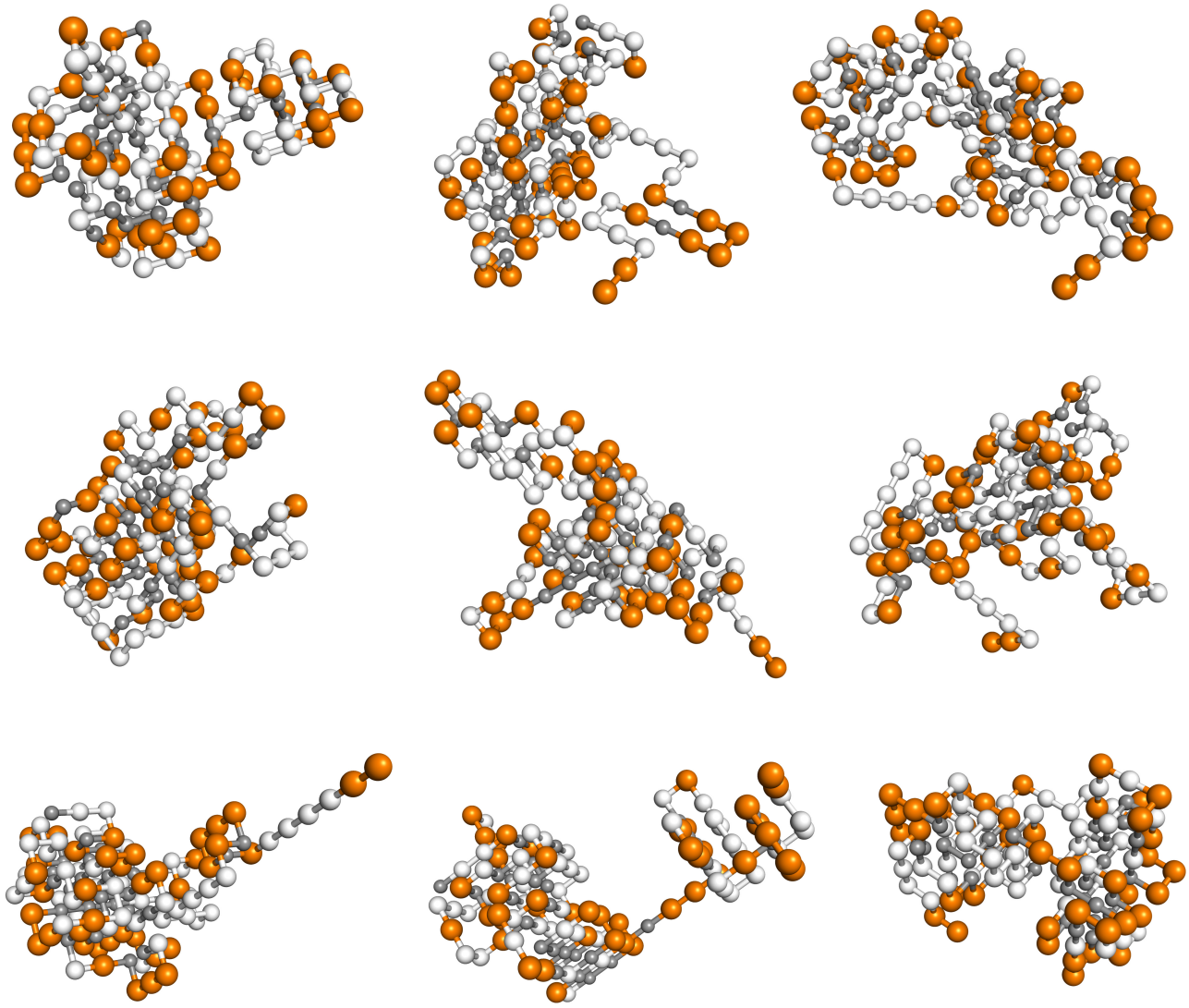


Figure 6.13: For H0P3D124 with $\epsilon_{\mathrm{HH}} = 4$, $\epsilon_{\mathrm{H0}} = 2$, $\epsilon_{\theta} = -1$ and zero for the rest of interactions: nine typical configurations at $T = 0.3$. For structures shown, H- and '0'-mers are colored dark gray and white, respectively, while P-mers are colored orange.

6.3 Folding funnels in Ribonuclease A: choice of reaction coordinates

In order to observe the free energy landscape in terms of a folding funnel, the reaction coordinate has to be chosen carefully. We have examined some properties, such as radius of gyration ($R_{\rm g}$), tortuosity (τ), number of angles (n_θ) and end-to-end distance ($R_{\rm ee}$) as the candidates of reaction coordinates.

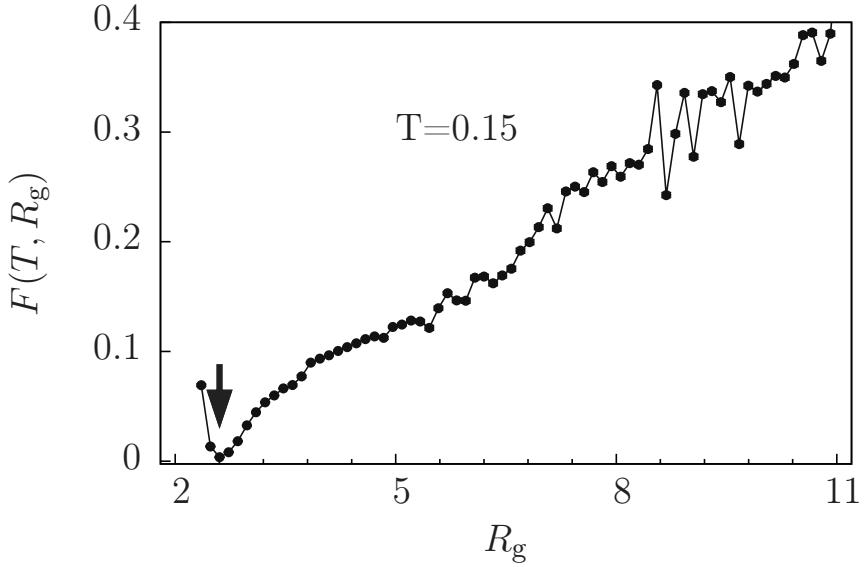


Figure 6.14: Normalized free energy vs radius of gyration ($R_{\rm g}$) at $T = 0.15$ for the H0P3D124 lattice protein ($\epsilon_{\rm HH} = 4$, $\epsilon_{\rm H0} = 2$ and $\epsilon_\theta = -1$). Black, filled arrow indicates the lowest free energy at this temperature. Error bars smaller than the data points are not shown.

The radius of gyration ($R_{\rm g}$) as a way of measuring the compactness of the protein structure is very useful for characterizing the coil-globule collapse transition. However, at low temperatures, the protein structure is usually very compact, and the compactness does not change as much. Therefore, $R_{\rm g}$ is not very useful for representing the processes of rearranging protein structures into native states. As shown in Fig. 6.14, we plot the free energy of H0P3D124 vs $R_{\rm g}$ at $T = 0.15$, which is within the temperature region where the structure of the protein is rearranged into native state. We can see that the free energy shows

a rugged shape at the positions far away from the free energy minimum, while presenting pretty smooth shape around the free energy minimum.

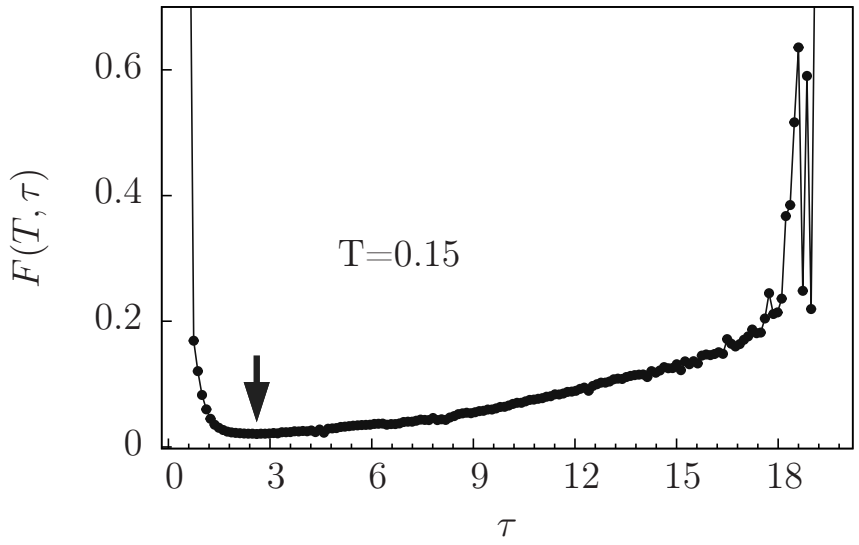


Figure 6.15: Normalized free energy vs tortuosity (τ) at $T = 0.15$ for the H0P3D124 lattice protein ($\epsilon_{\text{HH}} = 4$, $\epsilon_{\text{H0}} = 2$ and $\epsilon_\theta = -1$). Black, filled arrow indicates the lowest free energy at this temperature. Error bars smaller than the data points are not shown.

The tortuosity (τ), proposed by Wüst et al. [64], is particularly sensitive to sequence-dependent internal topological features such as the breaking of HH contacts in compact denatured states upon folding to the ground state. As shown in Fig. 6.15, we plot the free energy of H0P3D124 vs τ at $T = 0.15$. A smooth and relatively flat bottom is observed for this quantity. Similarly the free energy of H0P3D124 vs the number of angles (n_θ) at $T = 0.15$ is shown in Fig. 6.16. From this figure, we found the shape of free energy curve is, even though rugged, against our understanding of folding funnel, as in Fig. 2.1.

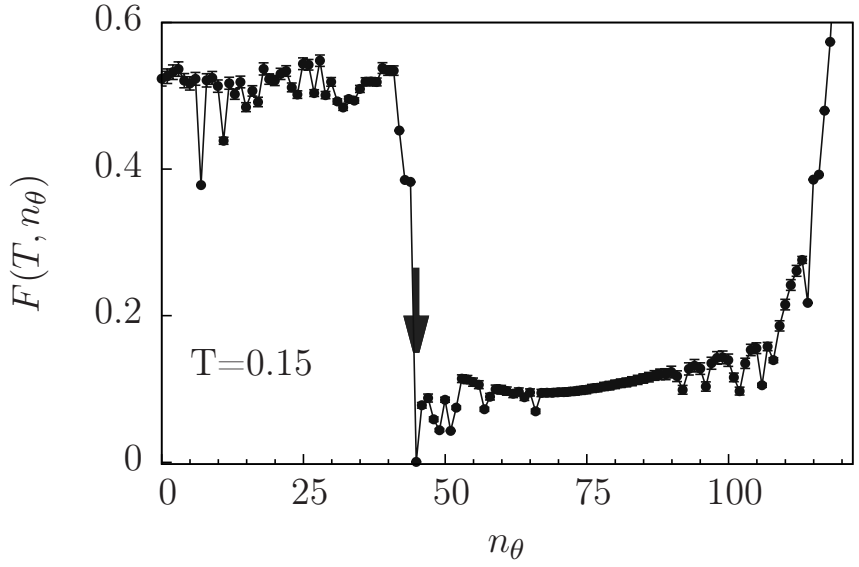


Figure 6.16: Normalized free energy vs number of angles (n_θ) at $T = 0.15$ for the H0P3D124 lattice protein ($\epsilon_{\text{HH}} = 4$, $\epsilon_{\text{H0}} = 2$ and $\epsilon_\theta = -1$). Black, filled arrow indicates the lowest free energy at this temperature. Error bars smaller than the data points are not shown.

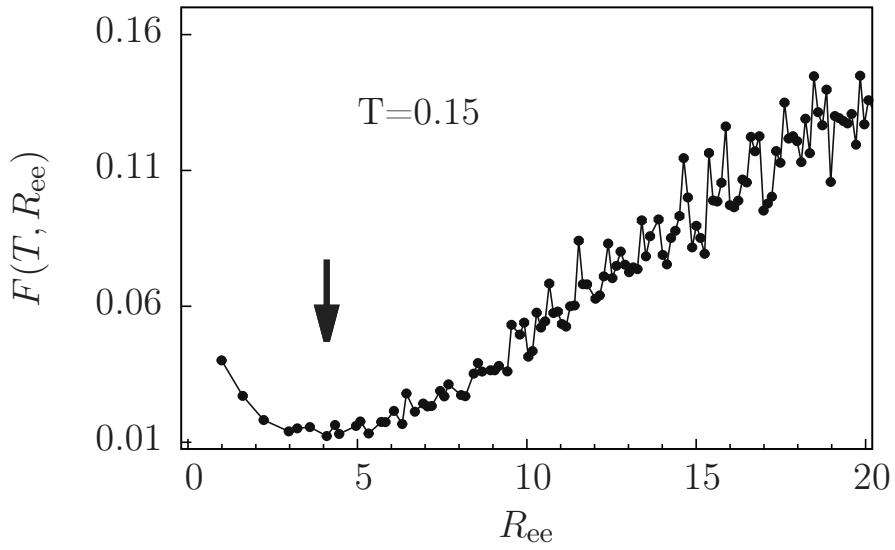


Figure 6.17: Normalized free energy vs end-to-end distance (R_{ee}) at $T = 0.15$ for the H0P3D124 lattice protein ($\epsilon_{\text{HH}} = 4$, $\epsilon_{\text{H0}} = 2$ and $\epsilon_\theta = -1$). Black, filled arrow indicates the lowest free energy at this temperature. Error bars smaller than the data points are not shown.

Finally, we found that the end-to-end distance (R_{ee}), even though it is one of simplest structural quantities, turns out to be a good choice as the reaction coordinate. As shown in Fig. 6.17, we plot the free energy of H0P3D124 vs R_{ee} at $T = 0.15$, which presents a rugged, funnel-like free energy curve. In the following two sections, we will present respectively the folding funnel of HP3D124 and H0P3D124 lattice proteins, using the end-to-end distance as the reaction coordinate.

6.4 Folding funnels in Ribonuclease A: the HP model

The free energy vs end-to-end distance at various temperatures is calculated according to Eq. 6.1, and results are shown in Fig. 6.18. The free energy curves contain many local maxima and minima at all temperatures. These variations in free energy are significant since statistical errors in the results are smaller than the size of symbols. The lowest free energy state is indicated by a filled, black arrow, while the mean end-to-end distance is marked by an orange arrow. At high temperature the behavior shows a shallow, "symmetric" but quite rough landscape. Upon lowering the temperature, we found that the free energy forms a clear, funnel-like structure that is skewed toward the region with low end-to-end distance values. Schematic portrayals of the protein folding funnel always present a static structure that simply guides the protein towards a fixed minimum as the temperature is lowered. Instead, we find that the lowest free energy position shifts with the change of temperature, indicating a dynamic, instead of static, nature of the folding funnel. At lower temperatures, the free energy landscape becomes relatively flat near the minimum and oddly shaped for large end-to-end distance. The relative smoothness means that the system can easily move between states, i.e., small changes in end-to-end distance do not result in significant differences in the free energy. When $T < 0.2$, the point where the free energy is lowest coincides with the mean end-to-end distance.

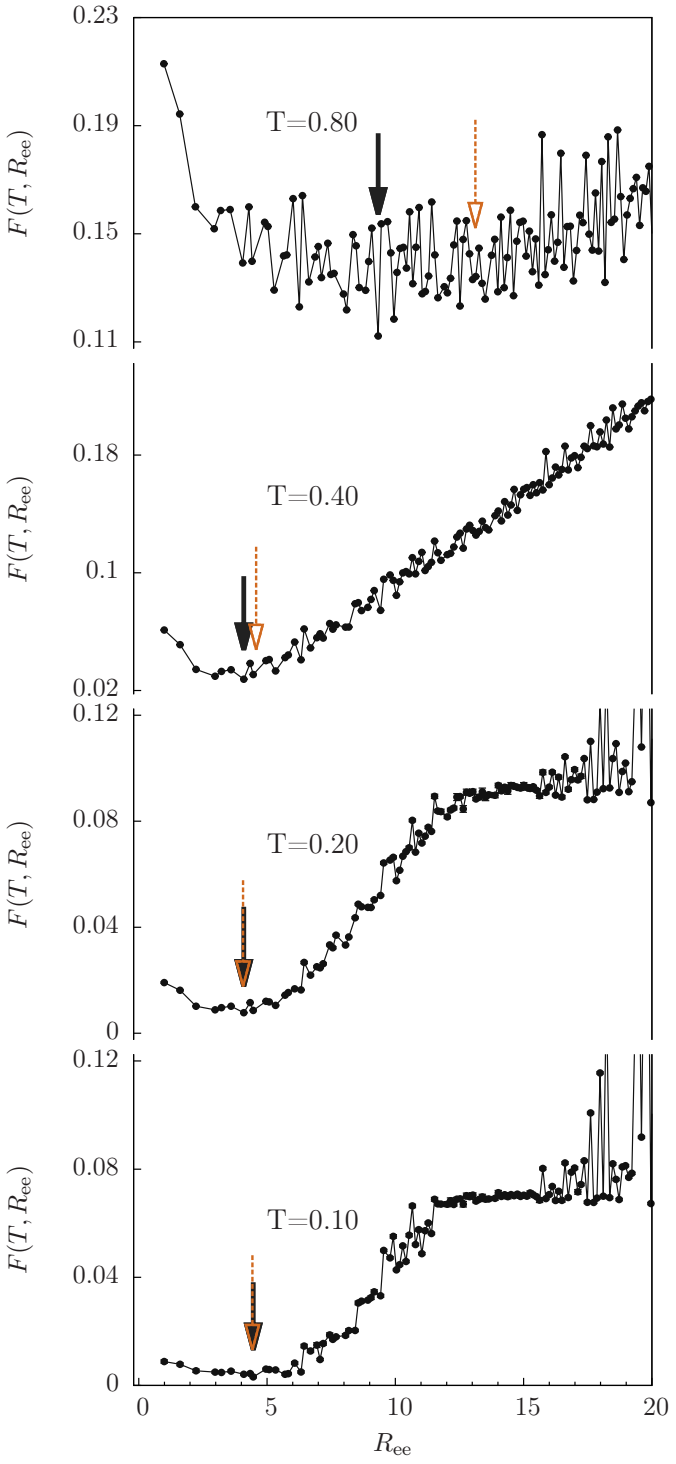


Figure 6.18: Normalized free energy vs end-to-end distance at four different temperatures for the HP3D124 lattice protein ($\epsilon_{\text{HH}} = 1$). Black, filled arrows indicate the lowest free energy at each temperature, orange arrows point to the mean end-to-end distance at the temperatures. Error bars are smaller than the data points.

6.5 Folding funnels in Ribonuclease A: the H0P model

The free energy vs end-to-end distance at various temperatures for H0P3D124 is shown in Fig. 6.19. The state with lowest free energy is indicated by a filled black arrow, while the mean end-to-end distance is marked by an orange arrow. At $T = 0.8$, we found a fairly shallow but rugged free energy landscape, which is similar to that for HP3D124. With decreasing temperature the free energy skews clearly toward the region with low end-to-end distance values, but at low T the funnel remains rough, even near the bottom! However, the shift of the lowest free energy position with temperature, indicates a dynamic, instead of the usually depicted static, rugged folding funnel.

Although both lattice protein models posses complex, funnel-like free energy landscapes, clear differences exist between them. For HP3D124 (Fig. 6.18) at lower temperatures, the free energy curve is oddly shaped and relatively flat near the bottom whereas the entire funnel remains rugged for H0P3D124. When $T < 0.2$, the position of the free energy minimum for HP3D124 coincides with the mean end-to-end distance, but for H0P3D124, even at low T, the free energy landscape remains rough near the minimum. For example, the free energy barrier preventing escape from the 2nd lowest state (R_{ee} between 5 and 6) is approximately $7k_{\mathrm{B}}T$, and this state is not even immediately adjacent to the lowest free energy state. Moreover, the lowest free energy is clearly separated from the averaged end-to-end distance and the protein can easily become trapped in a local minimum. Whereas the specific heat shows only two major events, the mean end-to-end distance changes often with temperature. This indicates that folding occurs through a series of small rearrangements that give rise to two major configurational changes. For both models the density of states, g(E), is smooth, even as the energy approaches its minimum. As a consequence, schematic representations of the funnel with a width given by a multivalued function of the entropy (see e.g. Wolynes et al. [11]) are inconsistent with the actual behavior of the lattice proteins.

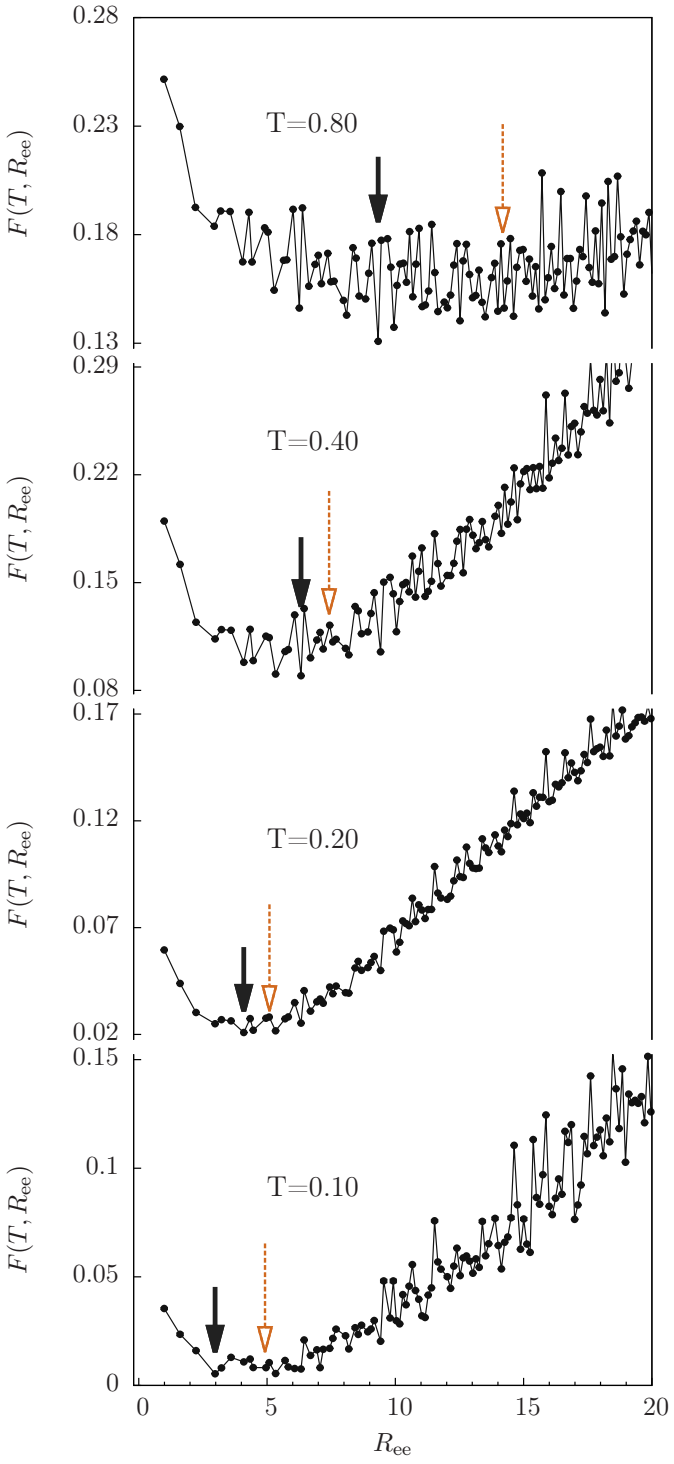


Figure 6.19: Normalized free energy vs end-to-end distance at four different temperatures for the H0P3D124 lattice protein ($\epsilon_{\text{HH}} = 4$, $\epsilon_{\text{H0}} = 2$ and $\epsilon_\theta = -1$). Black, filled arrows indicate the lowest free energy at each temperature, orange arrows point to the mean end-to-end distance at the temperatures. Error bars are smaller than the data points.

Chapter 7

Conclusions

In this work, we applied an advanced Monte Carlo algorithm: (replica-exchange) Wang–Landau sampling with appropriate trial moves to investigate the protein folding problem using simplified lattice protein models.

The classic hydrophobic-polar (HP) model was used originally, and then an "improved" model, the semi-flexible H0P model, was proposed by introducing some simple modifications, rendering the model more realistic without significantly increasing the difficulties of sampling. In the semi-flexible H0P model, we introduced a new type of "neutral" monomer, "0", i.e., neither hydrophobic nor polar and also took into consideration the stiffness of bonds connecting monomers. The semi-flexible H0P model, in comparison with HP model, has significantly reduced ground state degeneracy and rich folding signals as the proteins rearranging into native states from very compact structures at low temperatures.

Ground state structures, as one of the most important properties of proteins, are essential in understanding many biological problems of interest. For estimating the ground state degeneracy of lattice protein models, we developed a heuristic method which combines flat-histogram sampling with an efficient structure database and enables us to gain detailed insight in systems of sizes far beyond those accessible by exact enumeration approaches,

while also working as effectively as such methods for short sequences. The reliability of this method has been demonstrated by comparing its results with those from other enumeration methods, including the determination of the exact density of states for some short sequences.

We carried out a thorough investigation of the effect of single-site mutation (SSM) on two long, designed HP proteins: HP3D42 and HP3D67, which result in 42 and 67 mutated sequences respectively after SSM. By systematically studying 109 mutated sequences plus two original sequences, we discovered that many mutations do not affect the protein significantly in any regard, including the ground state degeneracy and energy. On the other hand, very sensitive positions in the primary structure exist, where mutations can drastically change the folding process and low-energy structures. Remarkably, both observations coincide with experimental discoveries for real proteins [95, 96], confirming the adequacy of simple, generic models for certain problems. In addition, we find that the thermal stability of mutated sequences is likely to be lower than original sequences, from the observation of ground state population and specific heat. The reason is that even though the ground state degeneracy may increase dramatically after mutations, the degeneracy of the first and second excited states grow with the same rate or even faster.

Lastly, we uncovered folding funnels for two lattice protein models that are mapped from the protein Ribonuclease A, consisting of 124 amino acids. In order to obtain the results of high resolution, it is only possible to employ parallel computing with advanced Monte Carlo methods such as replica-exchange Wang–Landau sampling. We find the HP model has a relatively shallow free energy minimum, reflecting the high ground state degeneracy, while the H0P model develops a clear, rough free energy funnel with a relatively low degeneracy ground state. Unlike the schematic figures in the literature, we find an asymmetric folding funnel that changes shape substantially as the temperature decreases, and even the location of the free energy minimum shifts. While the HP and H0P models are simplified descriptions of a real protein, neither the mapping nor the interactions were tuned to produce a special

free energy structure. We, thus, believe that the general characteristics of the folding funnels found in our study (particularly for the H0P model) will persist in a more realistic description of protein folding.

Bibliography

[1] K.A. Dill and J.L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–6, 2012.

[2] A. Goate, M. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, L. James, R. Mant, P. Newton, K. Rooke, P. Roques, C. Talbot, M. Pericak-Vance, A. Roses, R. Williamson, M. Rossor, M. Owen, and J. Hardy. Segregation of a missense mutation in the amyloid precursor protein gene with familial alzheimer's disease. *Nature*, 349:704–706, 1991.

[3] E. Leroy, R. Boyer, G. Auburger, B. Leube, G. Ulm, E. Mezey, G. Harta, M. J. Brownstein, S. Jonnalagada, T. Chernova, A. Dehejia, C. Lavedan, T. Gasser, P. J. Steinbach, K. D. Wilkinson, and M. H. Polymeropoulos. The ubiquitin pathway in Parkinson's disease. *Nature*, 395:451–452, 1998.

[4] A.R. Davidson. A folding space odyssey. *Proc. Natl. Acad. Sci.*, 105(8):2759–60, 2008.

[5] C.G. Roessler, B.M. Hall, W.J. Anderson, W.M. Ingram, S.A. Roberts, W.R. Montfort, and M.H.J. Cordes. Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proc. Natl. Acad. Sci.*, 105(7):2343–8, 2008.

[6] P.A. Alexander, Y. He, Y. Chen, J. Orban, and P.N. Bryan. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc. Natl. Acad. Sci.*, 104(29):11963–8, 2007.

[7] K.A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–9, 1985.

[8] K.F. Lau and K.A. Dill. Sequence Spaces of Proteins. *Macromolecules*, 22(i):3986–3997, 1989.

[9] J. Kubelka, J. Hofrichter, and W.A. Eaton. The protein folding speed limit. *Current Opinion in Structural Biology*, 14(1):76 – 88, 2004.

[10] K.A. Dill and H.S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4:10–19, 1997.

[11] P.G. Wolynes, J.N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995.

[12] C. Branden and J. Tooze. *Introduction to protein structure.* Garland Science, 2nd edition, 1999.

[13] K.A. Dill. The meaning of hydrophobicity. *Science*, 250:297, 1990.

[14] W. Kauzmann. Some factors in the interpretation of protein denaturation. In C.B. Anfinsen, M.L. Anson, K. Bailey, and J.T. Edsall, editors, *Advances in Protein Chemistry*, volume 14, page 1. Academic Press, 1959.

[15] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[16] E. Haber and C.B. Anfinsen. Side-chain interactions governing the pairing of half-cystine residues in ribonuclease. *J. Biol. Cbem.*, 237:1839–1844, 1962.

[17] C. Levinthal. Are there pathways for protein folding. *J. Chim. Phys*, 65:44–45, 1968.

[18] J. Drenth. *Principles of protein x-ray crystallography*. Springer, 3rd edition, 2006.

[19] G.S. Rule and T.K. Hitchens. *Fundamentals of protein NMR spectroscopy*. Springer, 1st edition, 2005.

[20] D.C. Rapaport. *The Art of Molecular Dynamics Simulation*. Cambridge University Press, 2ed edition, 2004.

[21] D.P. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, 4th edition, 2014.

[22] M. Bachmann. *Thermodynamics and Statistical Mechanics of Macromolecular Systems*. Cambridge University Press, 1st edition, 2014.

[23] V. Castells, S. Yang, and V.T. Paul. Surface-induced conformational changes in lattice model proteins by Monte Carlo simulation. *Phys. Rev. E*, 65(3):031912, 2002.

[24] M. Bachmann and W. Janke. Substrate specificity of peptide adsorption: A model study. *Phys. Rev. E*, 73:020901, 2006.

[25] A. Swetnam and M.P. Allen. Selective adsorption of lattice peptides on patterned surfaces. *Phys. Rev. E*, 85:062901, 2012.

[26] M. Radhakrishna, S. Sharma, and S.K. Kumar. Enhanced Wang–Landau sampling of adsorbed protein conformations. *J. Chem. Phys.*, 136(11):114114, 2012.

[27] Y.W. Li, T. Wüst, and D.P. Landau. Generic folding and transition hierarchies for surface adsorption of hydrophobic-polar lattice model proteins. *Phys. Rev. E*, 87(1):012706, 2013.

[28] R. Bonaccini and F. Seno. Simple model to study insertion of a protein into a membrane. *Phys. Rev. E*, 60(6 Pt B):7290–8, 1999.

[29] G. Ping, J.M. Yuan, M. Vallieres, H. Dong, Z. Sun, Y. Wei, F.Y. Li, and S.H. Lin. Effects of confinement on protein folding and protein stability. *J. Chem. Phys.*, 118(17):8042, 2003.

[30] B. Pattanasiri, Y.W. Li, D.P. Landau, T. Wüst, and W. Triampo. Conformational transitions of a confined lattice protein: A Wang–Landau study. *J. Phys.: Conf. Ser.*, 402:012048, 2012.

[31] A. Irbäck and C. Troein. Enumerating Designing Sequences in the HP Model. *J. Biol. Phys.*, 28(1):1–15, 2002.

[32] M. Bachmann and W. Janke. Density of states for HP lattice proteins. *Acta Phys. Pol. B*, 34(10):4689–4697, 2003.

[33] R. Schiemann, M. Bachmann, and W. Janke. Exact sequence analysis for three-dimensional hydrophobic-polar lattice proteins. *J. Chem. Phys.*, 122(11):114705, 2005.

[34] S.L. Narasimhan, A.K. Rajarajan, and L. Vardharaj. HP-sequence design for lattice proteins–An exact enumeration study on diamond as well as square lattice. *J. Chem. Phys.*, 137(11):115102, 2012.

[35] R. Backofen and S. Will. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints*, 11:5, 2006.

[36] E.M. OToole and A.Z. Panagiotopoulos. Monte Carlo simulation of folding transitions of simple model proteins using a chain growth algorithm. *J. Chem. Phys.*, 97(11):8644, 1992.

[37] T.C. Beutler and K.A. Dill. A fast conformational search strategy for finding low energy structures of model proteins. *Protein Sci.*, 5:2037–2043, 1996.

[38] P. Grassberger. Pruned-enriched Rosenbluth method: Simulations of θ polymers of chain length up to 1 000 000. *Phys. Rev. E*, 56(3):3682–3693, 1997.

[39] H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler. New Monte Carlo Algorithm for Protein Folding. *Phys. Rev. Lett.*, 80(14):3149–3152, 1998.

[40] M. Bachmann and W. Janke. Multicanonical Chain-Growth Algorithm. *Phys. Rev. Lett.*, 91(20):208105, 2003.

[41] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger. Growth algorithms for lattice heteropolymers at low temperatures. *J. Chem. Phys.*, 118(1):444, 2003.

[42] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger. Growth-based optimization algorithm for lattice heteropolymers. *Phys. Rev. E*, 68(2):021113, 2003.

[43] T. Prellberg and J. Krawczyk. Flat Histogram Version of the Pruned and Enriched Rosenbluth Method. *Phys. Rev. Lett.*, 92(12):120602, 2004.

[44] J.L. Zhang and J.S. Liu. A new sequential importance sampling method and its application to the two-dimensional hydrophobic-hydrophilic model. *J. Chem. Phys.*, 117(7):3492, 2002.

[45] J. Zhang, S.C. Kou, and J.S. Liu. Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *J. Chem. Phys.*, 126(22):225101, 2007.

[46] W. Tang and Q. Zhou. Finding multiple minimum-energy conformations of the hydrophobic-polar protein model via multidomain sampling. *Phys. Rev. E*, 86(3):031909, 2012.

[47] R. Unger and J. Moult. Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, 231:75, 1993.

[48] R. König and T. Dandekar. Improving genetic algorithms for protein folding simulations by systematic crossover. *BioSystems*, 50(1):17–25, 1999.

[49] F. Liang and W.H. Wong. Evolutionary Monte Carlo for protein folding simulations. *J. Chem. Phys.*, 115(7):3374, 2001.

[50] A. Shmygelska and H.H. Hoos. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinf.*, 6:30, 2005.

[51] F. Wang and D.P. Landau. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.*, 86(10):2050–2053, 2001.

[52] F. Wang and D.P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E*, 64:1–16, 2001.

[53] D.P. Landau, S.-H. Tsai, and M. Exler. A new approach to Monte Carlo simulations in statistical physics: Wang–Landau sampling. *Am. J. Phys.*, 72:1294, 2004.

[54] T. Vogel, Y.W. Li, T. Wüst, and D.P. Landau. Generic, hierarchical framework for massively parallel Wang–Landau sampling. *Phys. Rev. Lett.*, 110:210603, 2013.

[55] T. Vogel, Y.W. Li, T. Wüst, and D.P. Landau. Scalable replica-exchange framework for Wang–Landau sampling. *Phys. Rev. E*, 90:023302, 2014.

[56] E. Bornberg-Bauer. Chain growth algorithms for hp-type lattice proteins. In *Proceedings of the First Annual International Conference on Computational Molecular Biology*, RECOMB '97, pages 47–55, New York, NY, USA, 1997. ACM.

[57] T. Hoque, M. Chetty, and A. Sattar. Extended HP model for protein structure prediction. *J. Comput. Biol*, 16(1):85–103, 2009.

[58] J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105 – 132, 1982.

[59] E.E. Lattman, K.M. Fiebig, and K.A. Dill. Modeling Compact Denatured States. *Biochemistry*, 33:6158–6166, 1994.

[60] G. Shi, T. Wüst, Y.W. Li, and D.P. Landau. Protein folding of the H0P model: a parallel Wang–Landau study. *J. Phys.: Conf. Ser.*, 640:012017, 2015.

[61] P.D. Thomas and K.A. Dill. Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation. *Protein Science*, 2:2050–2065, 1993.

[62] U. Bastolla and P. Grassberger. Phase Transitions of Single Semistiff Polymer Chains. *J. of Stat. Phys*, 89:1061, 1997.

[63] J. Krawczyk, A.L. Owczarek, and T. Prellberg. Semi-flexible hydrogen-bonded and non-hydrogen bonded lattice polymers. *Physica A*, 388:104–112, 2009.

[64] T. Wüst and D.P. Landau. Optimized Wang–Landau sampling of lattice polymers: ground state search and folding thermodynamics of HP model proteins. *J. Chem. Phys.*, 137(6):064903, 2012.

[65] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys*, 21:1087, 1953.

[66] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8:3, 1998.

[67] M. Lüscher. A portable high-quality random number generator for lattice field theory simulations. *Comput. Phys. Commun*, 79:100, 1994.

[68] Y.W. Li. *Unraveling universal thermodynamic and structural behavior of HP model protein adsorption*. PhD thesis, The University of Georgia, 2012.

[69] N. Madras and A.D. Sokal. The pivot algorithm: a highly efficient Monte Carlo method for the self-avoiding walk. *Journal of Statistical Physics*, 50(1-2):109–186, 1988.

[70] N. Lesh, M. Mitzenmacher, and S. Whitesides. A complete and effective move set for simplified protein folding. *in RECOMB (Berlin, Germany, 2003)*, page p. 188, 2003.

[71] J.M. Deutsch. Long range moves for high density polymer simulations. *J. Chem. Phys.*, 106(21):8849, 1997.

[72] T. Wüst and D.P. Landau. Versatile Approach to Access the Low Temperature Thermodynamics of Lattice Polymers and Proteins. *Phys. Rev. Lett.*, 102(17):178101, 2009.

[73] G. Shi, A.C.K. Farris, T. Wüst, and D.P. Landau. Folding in a semi-flexible lattice model for Crambin. *J. Phys.: Conf. Ser.*, 686:012001, 2016.

[74] Y.W. Li, T. Vogel, T. Wüst, and D.P. Landau. A new paradigm for petascale Monte Carlo simulation: Replica exchange Wang–Landau sampling. *J. Phys.: Conf. Ser.*, 510:012012, 2014.

[75] T. Vogel, Y.W. Li, T. Wüst, and D.P. Landau. Exploring new frontiers in statistical physics with a new, parallel Wang-Landau framework. *J. Phys.: Conf. Ser.*, 487:012001, 2014.

[76] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications, 1970.

[77] B.A. Berg and T. Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Rev. B*, 267:249, 1991.

[78] B.A. Berg and T. Neuhaus. Multicanonical Ensemble: A New Approach to Simulate First-Order Phase Transitions. *Phys. Rev. Lett.*, 68:9, 1992.

[79] K. Yue and K.A. Dill. Sequence-structure relationships in proteins and copolymers. *Phys. Rev. E*, 48(3), 1993.

[80] R. Schiemann, M. Bachmann, and W. Janke. Exact enumeration of three-dimensional lattice proteins. *Comp. Phys. Comm.*, 166(1):8–16, 2005.

[81] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms, 3rd ed.* MIT Press, Cambridge, MA, 2009.

[82] K. Yue and K.A. Dill. Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci.*, 92(1):146–50, 1995.

[83] M. Bachmann and W. Janke. Thermodynamics of lattice heteropolymers. *J. Chem. Phys.*, 120(14):6779–91, 2004.

[84] K.F. Lau and K.A. Dill. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci.*, 87(2):638–42, 1990.

[85] D. Shortle, H.S. Chan, and K.A. Dill. Modeling the effects of mutations on the denatured states of proteins. *Protein Sci.*, 1(2):201–15, 1992.

[86] C. Holzgräfe, A. Irbäck, and C. Troein. Mutation-induced fold switching among lattice proteins. *J. Chem. Phys.*, 135(19):195101, 2011.

[87] F.J. Blanco, I. Angrand, and L. Serrano. Exploring the conformational properties of the sequence space between two proteins with different folds: an experimental study. *J. Mol. Biol.*, 285(2):741–53, 1999.

[88] S. Dalal and L. Regan. Understanding the sequence determinants of conformational switching using protein design. *Protein Sci.*, 9(9):1651–9, 2000.

[89] T.A. Anderson, M.H.J. Cordes, and R.T. Sauer. Sequence determinants of a conformational switch in a protein structure. *Proc. Natl. Acad. Sci.*, 102(51):18344–9, 2005.

[90] X.I. Ambroggio and B. Kuhlman. Design of protein conformational switches. *Curr. Opin. Struc. Biol.*, 16(4):525–30, 2006.

[91] Y. He, Y. Chen, P. Alexander, P.N. Bryan, and J. Orban. NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc. Natl. Acad. Sci.*, 105(38):14412–7, 2008.

[92] P.A. Alexander, Y. He, Y. Chen, J. Orban, and P.N. Bryan. A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci.*, 106(50):21149–54, 2009.

[93] A. Schmidt, M. Teeter, E. Weckert, and V.S. Lamzin. Crystal structure of small protein crambin at 0.48Å resolution. *Acta Crystallographica Section F*, 67(4):424–428, Apr 2011.

[94] J. Shimada, E.L. Kussell, and E.I. Shakhnovich. The folding thermodynamics and kinetics of Crambin using an all-atom monte carlo simulation. *J. of Mol. Biol.*, 308(1):79 – 95, 2001.

[95] S. Rackovsky. Spectral Analysis of a Protein Conformational Switch. *Phys. Rev. Lett.*, 106(24):248101, 2011.

[96] J.U. Bowie, J.F. Reidhaar-Olson, W.A. Lim, and R.T. Sauer. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, 247:1306, 1990.