Unraveling universal thermodynamic and structural behavior of HP model protein adsorption: A Wang-Landau study

by

Ying Wai Li

(Under the direction of David P. Landau)

Abstract

The hydrophobic-polar (HP) model is a simplified lattice model for simulating protein folding. With a minor modification, it can be used to study general categories of surface adsorption of protein from a coarse-grained perspective. In this work, the thermodynamic behavior and structural properties are studied by means of Wang-Landau sampling complemented by multicanonical sampling. A number of benchmark HP sequences have been considered with different types of surfaces, each of which attracts either: all monomers, only hydrophobic (H) monomers, or only polar (P) monomers, respectively. For some structural "transition" processes, the specific heat only shows obscure or missing signals, and thus a comprehensive analysis is vital in distinguishing structural "transitions" between "phases" for polymeric systems. From the analysis of the combined patterns of various structural observables, e.g., the derivatives of the numbers of surface contacts, together with the specific heat, fundamental, general categories of folding and transition hierarchies have been identified. A connection between the transition categories and the relative surface strengths, i.e., the ratio of the surface attractive strength to the intra-chain attraction among H monomers, has also been inferred. As the classification of transition categories is founded on multiple

benchmark sequences, it is believed that the folding hierarchies and identification scheme are generic for different HP chains interacting with attractive surfaces, regardless of the chain length, sequence, or type of surface attraction.

INDEX WORDS: Monte Carlo simulations, Wang-Landau sampling, protein adsorption, hydrophobic-polar model, HP model, heteropolymers, structural phase transitions

Unraveling universal thermodynamic and structural behavior of HP model protein adsorption: A Wang-Landau study

by

Ying Wai Li

B.Sc., Chinese University of Hong Kong, Hong Kong S.A.R., 2005

M.Phil., Chinese University of Hong Kong, Hong Kong S.A.R., 2007

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

2012

Unraveling universal thermodynamic and structural behavior of HP model protein adsorption: A Wang-Landau study

by

Ying Wai Li

Approved:

Major Professor: David P. Landau

Committee: Steven P. Lewis
Michael Bachmann
Shan-Ho Tsai

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

Center for Simulational Physics
University of Georgia

Unraveling universal thermodynamic and structural behavior of HP model protein adsorption: A Wang-Landau study

Author:

Ying Wai Li

Supervisor:

Prof. David P. Landau

August, 2012

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. David P. Landau. His insightful guidance, constant support and patience have been a tremendous encouragement to me during my PhD studies. I am particularly thankful for his effort in exploring and refining my various abilities, which tremendously enhanced my capabilities and confidence in conducting research. His passion for physics has a great influence on me in pursuing a scientific career.

I am very grateful for other members in my advisory committee: Prof. Steven P. Lewis, who was the teacher of a number of graduate classes, for laying a solid foundation for my theoretical physics background; Prof. Michael Bachmann, for contributing many constructive discussions and suggestions to my research projects; and Dr. Shan-Ho Tsai, for offering assistance on the usage of research computer clusters. Besides, I am thankful to them for reading my dissertation with great care and their helpful comments.

I am also indebted to a number of collaborators. Dr. Thomas Wüst and Dr. Thomas Vogel, two excellent mentors and earnest friends of mine in life, have demonstrated to me that quality research is accomplished by careful thoughts, hard work, and the persistence of striving for excellence. Ms. Busara Pattanasiri, my "little sister" and first research mentee, who worked diligently and suggested many new ideas to our project. I enjoyed our collaborations and treasure the friendship very much.

My gratefulness is also due to many members in our research group and the Center for

Simulational Physics. I have benefited a lot from the scientific (and non-scientific) discussions with Dr. Daniel T. Seaton, Dr. Stefan Schnabel, Dr. Meng Meng, Ms. Siyan Hu and Mr. Dilina Perera. I would like to give special thanks to Mrs. Linda Lee, for her care and advice on different matters. Thanks also to our systems administrators, Mr. Mike Caplinger and Mr. Jeff Deroshia, for their expertise in technical support to our computers and research clusters in the department.

In addition, I wish to express my appreciation to Prof. Lorena Barba from Boston University, for selecting me to be one of the participants of the Pan-American Advanced Studies Institutes Program to learn about parallel computing. I am also obliged to the Anderson family, the Cummings family and the Graduate School at the University of Georgia. The awards and scholarships presented to me were great recognitions for my academic work and research.

Last but not least, I would like to thank my family for their endless support and care throughout all these years. My parents have provided me with excellent education in many aspects; without them I would not have been able to achieve this much. And thanks to my two lovely cats as well, who have witnessed the beginning and the finish of my research project and accompanied me through all the working (plus non-working) nights.

Dr. Landau always says, "life is a Monte Carlo simulation." I am sure I have got a marvelous random number generator to let me meet all these wonderful people and enjoy an unforgettable graduate life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Protein folding and protein adsorption have been important research topics for decades. They gained so much attention both because of the many basic scientific questions they pose that remain unsolved, and also because of their numerous applications[1–4]. In the natural world, numerous biological functions at the cellular level are carried out by proteins. One common example is enzymes, which are a class of proteins that catalyze chemical reactions. Since the function of a protein relies on its structure, a misfolded protein is likely to malfunction or even be toxic. Some illnesses, such as Alzheimer's disease and Mad Cow disease, are known to be caused by aggregated or misfolded proteins. Protein adsorption, in addition, also takes an important role in the biological world. Blood coagulation, for instance, is one of the vital processes in mammals that involves adsorption of blood proteins. Therefore, understanding protein folding and adsorption is the key to revealing the principles of many biological processes and causes of diseases.

Another example of protein adsorption in the medical or biological sciences is the study of protein functions in experiments, which often involves the immobilization of proteins on a solid substrate[5,6]. Protein drug delivery is a potential pharmaceutical application of a protein-substrate complex[7]. In this process, the protein drug is first made adsorbed on a

capsule and then delivered to a desired part of the body for action. After that the protein is released to the target site by desorbing from the carrier and performs the expected medical functions. Thus, understanding how the functions and conformations of a protein are affected by adsorption and desorption is a crucial part of protein drug delivery[8,9].

Other areas where protein adsorption is widely applied include nanotechnology and the fabrication of biomaterials. For instance, the study of adhesion of proteins on solid substrates such as metals, semiconductors, carbon or silica etc., has enabled the synthesis of new materials for biosensors or electronic devices[10–15]. It is also a crucial factor in integrating implanted materials with body tissues[1,16].

Therefore, the study of protein folding and adsorption is a contribution both to the understanding of basic science and to the technological advancement of practical applications. However, our knowledge about protein functions, structure prediction, folding kinetics, dynamics and thermodynamics, folding mechanisms, structural "phases" and transformations, etc., is still at the tip of the iceberg. The exploration of these areas is pioneered by experimental studies nowadays, but the fact that only the "final product" can be obtained and studied in an experiment makes for slow progress in understanding the dynamics and folding processes. From another point of view, the diversity of possible protein sequences and sophisticated interactions among amino acids also complicates the theoretical studies of protein folding - not to mention when the protein interacts with solvent molecules or a substrate where an extra level of complexity enters. It just becomes impossible to obtain an analytical form for the physical quantities for macromolecules like proteins when hundreds of thousands of atoms are involved. Furthermore, since the behavior of proteins can be very different from one another due to the various combinations and sequences of amino acids, even for the same chain length, one cannot apply finite-size scaling[17] to obtain a systematic study of the effect of protein size, unlike some simple models in statistical mechanics.

With the advances in computer capacities, numerical simulations start to be a promising

way to complement on the general problems and to bridge the gap between theoretical and experimental studies. Nevertheless, proteins are sufficiently complicated that attempts to study them numerically rely upon simplifying the problem to one of manageable proportions, yet retaining the fundamental features of the protein. A sensible choice of simulation strategies is also essential to attack a particular problem depending on its nature.

In this work, our intent is to recognize generic thermodynamic and structural behavior in protein adsorption using a minimalistic lattice protein model known as the HP model[18,19]. Furthermore, we would like to see if this behavior is related to some system variables. This question is of fundamental importance because a success in doing so implies that different HP sequences share certain general qualities in structural transformations when brought near to an absorbing substrate. That also means the feasibility of transition behavior prediction given the system settings. Monte Carlo methods, which are capable of exploring a large conformational phase space and giving access to the thermodynamics of a system in equilibrium, are then well suited for this purpose. The combination with simple, coarse-grained protein models is indispensable to enhance computational efficiency and to reduce unnecessary distractions from the atomic details to the recognition of universal behavior.

The arrangement of this dissertation is as follows: Chapter 2 introduces some basics of proteins, methods to study them and questions to be understood. Chapter 3 describes the HP model that we have adopted in our simulations, and the measures that quantify the structural properties in our model. Chapter 4 outlines the Monte Carlo techniques and trial moves employed in our studies. We will start showing our simulation results in Chapter 5 by illustrating how to identify structural changes by a combination of thermodynamic and structural quantities. Chapter 6 is an extension of our identification scheme to some longer HP benchmark sequences, which has allowed for a comprehensive analysis of generic transition behavior. Four transition categories were identified and a correspondence between these categories and some system parameters will also be presented. We will conclude our findings

and offer an outlook for the subject in Chapter 7. Since the quality and the suitability of a pseudo-random number generator is crucial to correctness of the simulation results, we have carried out some simple tests to confirm the validity of the pseudo-random number generator we adopted for use with our algorithm. The results are shown in Appendix A.

Chapter 2

Background

2.1 A brief introduction to proteins

2.1.1 Composition - The (20+2) amino acids

Proteins are biological polymers found in nature. They are essential components which carry out various biological functions in living organisms. These linear polymers are composed of twenty-two types of naturally occurring, genetically coded amino acids (Table 2.1). A protein typically contains 50 to 3000 amino acids. Of these 22 proteinogenic or standard amino acids, twenty are directly encoded by the universal genetic code and they serve as the basic building blocks of proteins[20,21]. The remaining two, selenocysteine and pyrrolysine, are indirectly coded and are incorporated into proteins by special synthetic mechanisms[22,23].

Each amino acid is composed of an amino group ($-NH_2$), a carboxylic acid group ($-COOH$) and an alkyl group ($-R$) which makes up the side chain. All are bonded to the same central carbon atom (C_α). Amino acids are covalently bonded together by a peptide bond between the amino group of one amino acid and the carboxyl group of another. This forms the linear, rigid backbone of the protein.

The only feature that distinguishes the different amino acids is the side chain to which

	Amino acid	**Three-letter code**	**(One-letter) code**
Hydrophobic:	Alanine	*Ala*	(A)
	Glycine	*Gly*	(G)
	Isoleucine	*Ile*	(I)
	Leucine	*Leu*	(L)
	Methionine	*Met*	(M)
	Phenylalanine	*Phe*	(F)
	Proline	*Pro*	(P)
	Valine	*Val*	(V)
Charged polar:	Arginine (+)	*Arg*	(R)
	Aspartic acid (−)	*Asp*	(D)
	Glutamic acid (−)	*Glu*	(E)
	Lysine (+)	*Lys*	(K)
Uncharged polar:	Asparagine	*Asn*	(N)
	Cysteine	*Cys*	(C)
	Glutamine	*Gln*	(Q)
	Histidine	*His*	(H)
	Serine	*Ser*	(S)
	Threonine	*Thr*	(T)
	Tryptophan	*Trp*	(W)
	Tyrosine	*Tyr*	(Y)
The 21^{st} amino acid:	Selenocysteine	*Sec*	(U)
The 22^{nd} amino acid:	Pyrrolysine	*Pyl*	(O)

Table 2.1: The twenty-two genetically encoded amino acids.

the C_α is attached. Side chains can be charged polar, uncharged polar, or non-polar. The first two types are hydrophilic, and the non-polar type is hydrophobic. Hence, the amino acids are classified according to the types of their side chains (Table 2.1).

2.1.2 Levels of protein structures

The enormous combinatorics of the twenty amino acids give rise to multitudinous possible protein sequences to which the structure is closely related, and hence its biological function[20]. There are four levels of complexity in protein structures which are related to each other. The primary structure of a protein simply refers to the protein sequence of the chain, i.e., the order in which the amino acids are arranged. A secondary structure is a local arrangement formed by a peptide* with a regular pattern. The two most common types of secondary structures are the alpha (α) helix and the beta (β) sheet, which is a composite with a few β strands† held together. Tertiary structures are formed by packing secondary structures together to form a compact globule, and a quaternary structure is a macromolecular complex constituted of multiple folded protein chains.

2.1.3 Interactions in a protein

Amino acids interact with one another or with the environment through various bonds or interactions to attain different levels of structures, dictating the major principles of protein folding and adsorption. Apart from the covalent disulfide bridges between cysteine side chains, interactions between amino acids are mostly non-covalent. Ionic bonds are formed as oppositely charged amino acids interact by a transfer of electrons. Secondary structures such as α-helices and β-sheets are substantially formed by the hydrogen bonds between the backbones. Hydrogen bonds can as well associate with other interactions, e.g., reactions

*A peptide is a short chain of two or more amino acids. It usually consists of no more than 50 amino acids.

†A β strand is a peptide of a few amino acids long and its backbone is nearly fully extended.

between charged amino acids and water molecules. In addition, van der Waals forces also exist between molecules to provide complementary attractions.

In an aqueous environment, hydrophobic amino acids group together in order to minimize the disturbance on the hydrogen-bonded networks of water[24]. Hydrophobic residues held together in this manner have been regarded as being "pulled" by their own attraction, the so called "hydrophobic bonds", although it originates from the repulsive force by the water molecules. Such a hydrophobic interaction is believed to be the most significant factor that governs the tertiary structure of proteins[25,26].

2.1.4 Factors that determine protein adsorption

Configurational changes of protein molecules upon surface adsorption depend on both the properties of protein (e.g. sequence, size, thermodynamic stability, etc.) and the surface properties (e.g. materials, polarity, surface roughness, etc.); but how large these changes are and where in the protein molecules they occur remain puzzles to be solved[3,27]. Enormous endeavors have been dedicated to unveil the mysteries in protein folding and adsorption mainly by experimental approaches[28,29].

2.2 Methods of studying protein structures, folding and adsorption

2.2.1 Theoretical approaches

In the theoretical research on protein folding, extensive studies have been made in searching for native states, understanding folding mechanisms and pathways, analyzing relationship between the sequence, structure and function of a protein, etc. Nevertheless, due to the complexity of the problem, the underlying principles for protein structure prediction from

the amino acid sequence are still unclear[18], which makes the above studies difficult. As a result, the theoretical framework for the understanding of protein folding or adsorption largely relies on the building of simple mathematical models, which is a counterpart of that in the study of polymers[30]. For protein adsorption, the majority are the kinetic models like the Langmuir model or the random sequential adsorption (RSA) model[31] for the investigation of adsorption kinetics. A few equilibrium models were proposed for the study of thermodynamics[32,33]. In condensed matter physics, emphasis is also placed on the universal structural properties and "phases", kinetics and dynamics of structural changes.

There are two theories which will be relevant to our study: the free energy landscapes in protein folding and the Flory's theory. The rough free energy landscapes in protein folding is a consequence of all the complicated interactions among the amino acids[34,35]. The free energy as a function of reaction coordinates X_i at a certain temperature T is defined as:

$$F_T(X_1, ...X_i, ...) = -k_B T \ln p(X_1, ...X_i, ...), \tag{2.1}$$

where k_B is the Boltzmann factor, $p(X_1, ...X_i, ...)$ is the probability of finding a macrostate with the reaction coordinates X_i, which are some physical quantities that best describe the system characteristics. There is not a standard way of choosing appropriate reaction coordinates, nevertheless it is a crucial key to unravel the folding behavior, e.g., folding pathways and metastable states, of a protein[36–40]. At high temperatures proteins are distended, but below some characteristic temperature they fold into a "native state" which has the global minimum free energy at a certain finite temperature[41]. The knowledge of the free energy landscape and the thermodynamics of a protein system is essential to completely understand the folding process[35,42–46]. As a result, many theoretical studies have been dedicated to this direction.

Flory's theory[47] deals with the restriction that particles cannot overlap in space for

real polymers. It naturally leads to self-avoiding polymer models, but this also leads to the excluded volume effect: the self-avoidance requirement causes the polymer to occupy a larger volume than the case if it is allowed to overlap. For an unrestricted, freely jointed chain where overlapping particles are permitted, the radius of gyration (a quantity that measures the size of a polymer, to be defined in Section 3.3), R_g, is proportional to $\sqrt{N}$, where N is the number of monomers in the chain. When the excluded volume effect is accounted for, the Flory theory approximated that $R_g \sim N^{3/(D+2)}$, where D is the dimension of space. This theory will come into play for the explanation of the discrepancy in the transition temperatures identified by the specific heat and by the structural quantities in our results.

2.2.2 Experimental approaches

There are three major experimental methods to determine protein structures: x-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy. According to the Protein Data Bank (PDB)[48,49] statistics as of this writing, x-ray diffraction is still the prevailing tools for protein structural determination, with more than 85% of the total structures contributed. About 10% of the structures were determined by NMR, and less than 1% were determined by electron microscopy. A small percentage of structures were determined by hybrid methods using a combination of these techniques.

X-ray crystallography[50] has an atomic resolution of about one Angstrom (Å) to a few Angstroms, making it an ideal tool to determine secondary and tertiary structures where detailed arrangements of atoms are required. However, a serious drawback of x-ray diffraction is that it can only resolve crystal structures with a regular, repeating pattern. As a folded protein is usually a globular lump with irregular surfaces, protein crystals are difficult to prepare. Moreover, crystal growing is time-consuming, and a number of experimental conditions such as temperature, pH, concentrations of protein and solvent, etc., would affect the suitability of the crystal for x-ray diffraction.

NMR spectroscopy[51] is another experimental method to resolve the three-dimensional structure of a protein. Although it does not provide a level of resolution as high as x-ray crystallography, it is able to reveal kinetic or dynamic processes. NMR makes use of the nuclear magnetic moments present in some atoms, e.g., hydrogen (^{1}H), carbon (^{13}C), nitrogen (^{15}N), phosphorus(^{31}P), etc., to obtain some distance constraints between atoms, which in turn can be used to derive the three-dimensional structure. NMR has an advantage of being capable of resolving the protein structure in solution, thus it is able to get a structure closer to a specific physiological environment[52]. However, its application is limited to small proteins as there is an upper limit in the molecular weight that NMR can deal with. As of the year 1999, the upper limit was around 35 kDa[53] ‡, which was equivalent to about 300 residues.

Cryo-electron microscopy (cryo-EM)[54] is less common in the determination of protein structures, due to its relatively low resolution compared to the previous methods. It is mainly used to obtain quaternary structures like viruses, ribosomes or cellular organelles.

For the case of protein adsorption where native structures are not the only interest,many experimental methods are available for different purposes. For instance, ellipsometry[55] and total internal reflection fluorescence (TIRF)[56] can be used to measure the amount and thickness of adsorbed proteins. Optical waveguide lightmode spectroscopy (OWLS)[57] is used to measure adsorption kinetics and conformational changes, which are affected by temperature, pH or electrostatic effects of the environment. Three-dimensional images of the protein adsorbed surface can be obtained by atomic force microscopy (AFM)[12,58]. Some spectroscopic techniques like circular dichroism (CD)[59] and fluorescence measurements[60] are useful in the study of conformations of adsorbed proteins. Readers are directed to some excellent review articles[27,29,61] for a complete survey of experimental methods for probing

‡Masses of proteins are usually measured in Daltons (Da), also known as the unified atomic mass units (u), which has a value of approximately 1.66×10^{-27} kg.

protein adsorption.

2.2.3 Computational approaches

There are two major branches for protein simulations in *silico*: molecular dynamics and Monte Carlo simulations.

Molecular dynamics and Monte Carlo methods

In molecular dynamics, the equations of motion for each particle (it could be an atom or a molecule, depending on the level of simplification) are first determined according to the forces and potentials experienced by the particles. Time is discretized and the new positions and velocities at the following time step are then calculated by integrating the equations of motion. As such, the movements of particles and how the entire system evolves with time can be simulated. For protein simulations, the two most widely-used force fields are CHARMM[62,63] and AMBER[64], which are derived empirically from experiments.

Molecular dynamics is more advantageous than Monte Carlo methods for investigating dynamical processes and simulating with explicit solvents. It is also easier to compare the results with experiments. However, its reliability is highly dependent on the empirical force field that models the interactions at the atomic level. This, in turn, requires proper parametrization of the models and accurate representation of solvation effects[65]. With the computer resources available nowadays, it is still impossible to simulate macromolecules with a reasonably fine time step (say, a femtosecond) to a time scale that is comparable to the realistic folding time, which ranges from microseconds to seconds.

Monte Carlo methods, on the other hand, generate a series of accessible states the system can take on using some trial moves instead of solving the equations of motion. Trial states are accepted according to a certain probability distribution. At the end of the simulation, physical properties of the system can be obtained through a statistical analysis of

the series of configurations generated. As such, Monte Carlo is capable of exploring a larger conformational phase space, allowing one to study a wide range of general problems through statistical mechanics, e.g., thermodynamics, phase diagrams and transitions[17,66]. Nevertheless, an obvious limitation of Monte Carlo methods is that the system dynamics cannot be studied easily.

The use of Monte Carlo is often accompanied with the introduction of simplified models in the simulation of proteins or polymers[17,67]. A coarse-grained model has a much less complicated energy function in contrast to an atomistic model. It thus has an advantage that the calculation of energy is computationally more efficient. Undoubtedly, one severe deficiency of a coarse-grained model is the inability of resembling a realistic protein system from an experimental set-up. But as far as the universal, macroscopic properties are concerned, which is always the major incentive of adopting Monte Carlo methods, coarse-grained models are better suited to serve the purpose due to the absence of distractions from unnecessary details[68]. In the following, we will briefly introduce a few common coarse-grained models for proteins in Monte Carlo simulations, which are closely related to those that are used to simulate polymers.

Overview of coarse-grained models

Most reduced models coarsen the structure of a protein by regarding an amino acid residue as a single, spherical monomer, ignoring the atomistic details within the residue. The chain formed by connecting these monomers together with bonds then represents a polymer or a protein. To simulate a polymeric system, the simplest model is a homopolymer which is a chain composed of only one type of monomer. A heteropolymer (or a copolymer), on the other hand, contains two types of monomers. The homopolymer and the heteropolymer can also be used to simulate a protein, in which case the interaction schemes between monomers are constructed in reference to the realistic interactions between amino acids.

For off-lattice models, a number of degrees of freedom and interactions between a pair of monomers can be incorporated into the model to set restrictions to the orientations and locations of the monomers. Some common ones include:

1. Bond length

   It is the distance between two consecutive monomers. For polymers, it can be viewed as a spring by letting the two monomers interact through a harmonic potential like in a Gaussian chain; or an almost harmonic one like the FENE potential[69]. For proteins, the bond length is usually held fixed due to the rigid backbone formed by covalent bonds.

2. Bond angle

   It is the angle formed by the two bonds connecting to the same monomer, and is a control of the flexibility of a polymer.

3. Torsional angle

   It is the rotation of a bond about the axis formed by the direction of the previous bond.

4. Interaction between non-bonded monomers

   Non-bonded monomers interact with each other mainly through effective potentials. A widely used one is the Lennard-Jones potential.

For lattice models, the degrees of freedom are much reduced. The bond length is typically fixed, the bond angle takes on a value depending on the nature of the lattice. Two-dimensional lattices include the square lattice and the triangular lattice; three-dimensional lattices include the cubic, face-centered cubic, and body-centered cubic lattices. In these examples, a monomer occupies a lattice site and interactions exist between two neighboring, non-bonded monomers.

Though highly simplified compared to a real protein, the off-lattice model is still a challenge to simulate. With various degrees of freedom and interactions, the energy calculation for an off-lattice model is more complicated and computationally expensive when compared to lattice models. The continuous energy levels also induce other simulation difficulties as will be discussed in Section 3.2 of Chapter 3. As a result, we have chosen to concentrate on a minimalistic, lattice model: the hydrophobic-polar (HP) protein model[18,19], which will be introduced in Chapter 3. A related subject in the polymer physics community is the study of block copolymers[70], in which the same types of monomers are arranged in blocks, forming a segment of the polymer.

2.3 Previous work and questions remaining

In spite of its simplicity, the HP model is surprisingly challenging to study. Indeed, the ground state[§] search for an HP sequence turns out to be an NP-complete problem[71,72]. Another complication comes from the uniqueness of every HP sequence. Even for two sequences of the same length (i.e. same system size), their thermodynamics can be different due to the difference in the proportion or order of the H and P monomers. Thus one cannot apply finite-size scaling to the HP model and study the influence of system size to the thermodynamic behavior, unlike many other "traditional" models in statistical mechanics.

The HP model has also created unexpected difficulties in simulations at the algorithmic level. Metropolis method[73] has been proven to be extremely inefficient for obtaining low temperature thermodynamics due to the complexity of the free energy surface[74]. Two approaches have been taken to resolve the problem. One is to invent algorithms only for searching for the ground state configuration, e.g., Monte Carlo with minimization[75], simulated annealing[76], genetic algorithms[77,78], tabu search[79], evolutionary Monte Carlo (EMC)[80],

[§]A ground state is the native state of the system at zero temperature.

fragment regrowth Monte Carlo via energy-guided sequential sampling (FRESS)[81], and many others. Detailed reviews can be found in Refs. [66] and [82]. Apart from Monte Carlo methods, ground state conformation searches for HP sequences can also be performed by tailor-made methods like hydrophobic core construction[83–86].

Another approach allows one to estimate the density of states as a function of energy, $g(E)$, from which the thermodynamic properties of the system can be obtained. Examples include the multi-self-overlap-ensemble simulation (MSOE)[87,88], pruned-enriched Rosenbluth method (PERM)[89] and its variants[90–93], multicanonical chain growth (MCCG)[94–97], equi-energy sampling (EES)[98] and Wang-Landau (WL) sampling[99–103]. As the HP model is a simplified preliminary step towards the real protein folding problem, all these algorithms have used it as a common testing ground for their capabilities.

In terms of studying protein adsorption using the HP model on lattice, similar approaches have been taken to study the energy landscapes, thermodynamics and conformational transitions[104–109]. Previous work by Bachmann and Janke[108], as well as Swetnam and Allen[109], have studied the conformational pseudophases based on the specific heat of individual benchmark HP sequences (a 103mer and a 36mer, respectively). In Ref. [108], some structural quantities and their dependence on temperature for the 103mer were presented. A number of structural "phases" have been recognized, which are similar to those identified using a homopolymer (see Refs. [110], [111] and references therein). A few quantities (the contact numbers, which will be discussed in Chapter 3, Section 3.3 after the introduction of the model), have also been identified as good "order parameters" for signaling structural changes. These have enabled the investigation of the free energy landscapes and folding channels for the adsorption of the 103mer[112,113]. Nevertheless, many areas remain relatively unexplored in several aspects:

1. lack of studies for other and longer sequences;

2. the existence of a "universal behavior" for different HP sequences is unknown;

3. lack of thorough investigation of the free energy landscapes for sequences other than the 103mer;

4. folding channels for other sequences have not been investigated; and thus it is unclear whether "universal" folding channels which are common to different sequences exist.

These questions are indeed interrelated. In this work, we are going to address items 1, 2 and 4 by studying the combination of the thermodynamic and structural behavior of the HP sequences using Monte Carlo simulations. Multiple HP benchmark sequences, with lengths between 36 and 103 monomers, are investigated and compared to "fill the holes" left by the above studies. To the best of our knowledge, this work is the first complete interpretation of protein adsorption that integrates analyses from multiple HP sequences. Although one is tempted to think there is no "universal" behavior for the HP model because of the uniqueness of the sequence and the failure of using finite-size scaling, it is indeed observed from our careful analyses and interpretation of structural transitions for multiple HP sequences.

Chapter 3

The Hydrophobic-Polar (HP) Protein Model

3.1 Specifics of the model

Out of the 22 proteinogenic amino acids, the 20 standard ones can roughly be classified as either hydrophobic or polar depending on the nature of their side chains. The tendency of the non-polar residues to stay away from water molecules has been identified as the key driving "force" in forming tertiary structures. The hydrophobic-polar (HP) model[18,19] is a coarse-grained lattice model for proteins that captures this hydrophobic effect. In this model, an amino acid residue is treated as a single monomer and is classified into either: the hydrophobic (H) type, or the polar (P) type. A protein is thus represented by a heteropolymer of N connected monomers performing a self-avoiding walk on a rigid lattice.

An attractive interaction exists only between a pair of non-bonded nearest neighboring H monomers. This attraction is denoted by ε_{HH} in our discussion, and the magnitude indicates the ability of the H monomers to pull themselves together as determined by the insolubility of the protein in an aqueous environment. In other words, the solvent quality is intrinsically

considered by the model.

In reality, there are other interactions that govern protein folding, e.g., attractions between hydrophobic and polar residues, and that between polar residues. Nevertheless, their magnitudes are relatively weak compared to the hydrophobic attraction and are thus neglected in the model. Factors like hydrogen bonds, charges and acidity of amino acids which are important in forming secondary structures are also not handled in this scope.

In the case of protein adsorption, the binding of the protein with an attractive substrate contributes to the energy externally in addition to the internal interactions within the polymer. A slight modification to the model is then necessary. We have considered three types of surface fields in view of the setting of the HP model: (i) a surface attracts only H monomers with strength ε_{SH}, (ii) a surface attracts only P monomers with strength ε_{SP}, and (iii) a surface interacts with both H and P monomers with equal strength, i.e., $\varepsilon_S = \varepsilon_{SH} = \varepsilon_{SP} \neq 0$. The energy function of the modified model thus takes the general form of[108]:

$$E = -\varepsilon_{HH} n_{HH} - \varepsilon_{SH} n_{SH} - \varepsilon_{SP} n_{SP}, \tag{3.1}$$

where n_{HH} denotes the number of interacting pairs of H monomers, n_{SH} the number of surface contacts with the H monomers and n_{SP} the number of surface contacts with the P monomers. The negative signs in front of each term ensure that it is energetically favorable when the monomers interact or come in contact with the surface.

Our simulations were performed on a three-dimensional simple cubic lattice, with periodic boundary conditions imposed in the x and y directions. The attractive surface is represented by an xy-plane placed at $z = 0$. A non-interacting steric wall is placed at $z = h_w = N + 1$ to confine the polymer in a way that it can contact both walls with its ends only when it is a vertical straight chain. A schematic setting of the model is shown by Figure 3.1. The purpose of putting in a non-attractive wall is to reduce the vertical translational degrees

of freedom of the polymer, so as to restrict the number of desorbed configurations in the simulation that are translationally equivalent when the size of the simulation box is large. Nevertheless, this steric wall also introduces an entropic effect which affects the adsorption process. Detailed discussions will be given in Chapter 5, Section 5.5.1.

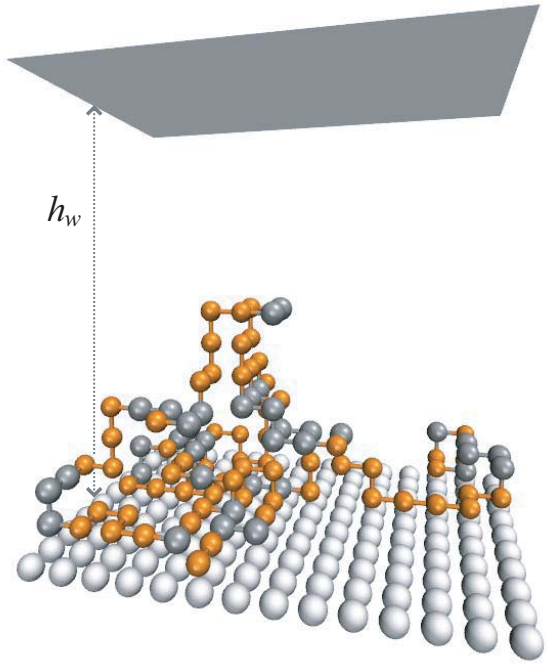


Figure 3.1: A schematic diagram showing the model used in this work. The gray spheres represent hydrophobic monomers, orange spheres represent polar monomers, faint spheres are the attractive molecules of the substrate and the solid top surface is non-attractive.

3.2 Advantages and drawbacks of lattice models

The calculation of the energy of a lattice model is fast and straight-forward as the monomers are restricted on a grid. When counting the number of interacting nearest neighbors, only $2 \times D$ steps are needed for every monomer, where D is the dimension of the simulation box. As such, there are six directions to check on a cubic lattice and four directions on a square lattice, instead of typically $N - 1$ steps per monomer for an off-lattice model when every monomer has to be walked through and checked to see if it falls into the interactive range.

A self-avoiding check also becomes trivial on a lattice, as one only needs to verify that

there are no two monomers occupying the same lattice site. This is a more sophisticated and time-consuming task for an off-lattice case.

Furthermore, integer arithmetic can be used for the energy calculation and integer data type can be used for the storage of coordinates since the energy and coordinates of monomers in our system are discrete. This eliminates the chance of introducing round-off or truncation errors for floating-point numbers as in off-lattice models, where the decision on the sizes of an energy bin and the energy range unavoidably introduces artifacts into the simulations. This also implies an absence of the problem of energy-binning in obtaining the density of states $g(E)$ and the histogram $H(E)$ in Wang-Landau sampling (which will be introduced in Chapter 4). All these advantages allow for the simulations of longer chains compared to off-lattice models.

One unavoidable drawback of using a lattice is the introduction of an unnaturally fixed bond length and bond angle. As there is only nearest neighbor interaction in the HP model, another artificial defect is that every monomer is not able to interact with half of the monomers within the chain due to the cubic lattice arrangement. For instance, a monomer labeled i can only have nearest neighbors which have labels $i + 2j + 1$. This artifact exists merely in a cubic lattice, and can be overcome by employing, for example, a triangular lattice, a face-centered cubic lattice, or a bond-fluctuation model. The AB model[114], which is literally the off-lattice version of the HP model, can also be adopted to eliminate the lattice effects completely.

3.3 Structural quantities

Besides contributing to the system's energy, the three contact numbers entered in Eq. (3.1), n_{HH}, n_{SH} and n_{SP}, are also useful "order" parameters that give quantitative measures of the structure of a conformation. They are identified to be good "order parameters" for

this system[112,113]. While n_{HH} gives the number of hydrophobic pairs and measures the energy contribution from the internal interactions, n_{SH} and n_{SP} give the numbers of surface contacts and thus the energy contributions arising from the surface-monomer interactions. It is equivalent to investigating different energy components instead of the total energy of the system, and it becomes apparent how different transition processes affect the energy fluctuations.

It is, therefore, natural to investigate the thermal derivatives of the ensemble averages of these quantities so as to decompose their contributions to the energy fluctuations. These include the derivative of the average number of H-H interactions, $d\langle n_{HH}\rangle/dT$, and those of the numbers of surface contacts, $d\langle n_{SH}\rangle/dT$ and $d\langle n_{SP}\rangle/dT$. A peak in $d\langle n_{HH}\rangle/dT$ signals the construction of H-H interactions to form a hydrophobic core (H-core formation). Peaks in $d\langle n_{SH}\rangle/dT$ and $d\langle n_{SP}\rangle/dT$ provide information about the formation of surface contacts, which is associated with the adsorption process as well as the "flattening" of the structure due to surface attraction. All these transition processes will be discussed in more detail in Chapter 5 and Chapter 6.

Other structural quantities which are essential in understanding non-energetic properties of the system include the radius of gyration,

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_{cm})^2}, \tag{3.2}$$

and the end-to-end distance,

$$R_{ee} = |\vec{r}_N - \vec{r}_1|, \tag{3.3}$$

where $\vec{r}_i$ is the position of monomer i, $\vec{r}_{cm}$ in Eq. (3.2) is the average location of all the monomers: $\vec{r}_{cm} = \frac{1}{N}\sum_{i=1}^{N}\vec{r}_i$.

The radius of gyration R_g measures how far away the monomers are distributed relative to

their average; while the end-to-end distance R_{ee} gives the distance between both ends of the polymer. Both quantities measure the degree of extension of the polymer and are essential properties in quantifying structural transformations. In many cases, a rapid decrease in R_g and R_{ee} with the decrease in temperature is an indication of the collapse transition, where an extended chain collapses to a globule structure.

3.4 HP benchmark sequences

A number of HP sequences have been proposed in the literature for simulations of proteins or algorithm testing purposes. Each of them is specially designed to attain either a two-dimensional or a three-dimensional ground state configuration. Since the ground state search of these sequences is itself a challenging task as well as a "competition" among research groups, these sequences have thus been deemed to be a benchmark for testing new algorithms. Some common benchmark sequences are listed in Table 3.1. The lowest energy for each sequence, which is essentially the number of H-H interaction pairs found in the ground state, was obtained by HPstruct in the CPSP-tools package[115,116] (developed based on a constraint-based algorithm in Ref. [86]) and was also presented in the table as a reference.

Seq. code	Sequence	Lowest energy
2D:		
2D36[77]	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	−14
2D50[77]	$H_2PHPHPHPH_4PHP_3HP_3HP_4HP_3HP_3HPH_4PHPHPHPH_2$	−21
2D60[77]	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	−36
2D64[77]	$H_{12}PHPHP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HP_2H_2P_2H_2P_2HPHPH_{12}$	−42
2D85[117]	$H_4P_4H_{12}P_6H_{12}P_3H_{12}P_3H_{12}P_3HP_2H_2P_2H_2P_2HPH$	−53
2D100a[118]	$P_6HPH_2P_5H_3PH_5PH_2P_4H_2P_2H_2PH_5PH_{10}PH_2PH_7P_{11}H_7P_2HP-H_3P_6HPH_2$	−48
2D100b[118]	$P_3H_2P_2H_4P_2H_3PH_2PH_2PH_4P_8H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2P-HP_2HPH_3P_6H_3$	−50
3D:		
3D42[83]	$PH_2PHPH_2PHPHP_2H_3PHPH_2PHPH_3P_2HPHPH_2PHPH_2P$	−34
3D48[84]	$PHPHP_4HPHPHP_2HPH_6P_2H_3PHP_2HPH_2P_2HPH_3P_4H$	−34
3D67[83]	$PHPH_2PH_2PHP_2H_3P_3HPH_2PH_2PHP_2H_3P_3HPH_2PH_2PHP_2H_3-P_3HPH_2PH_2PHP_2H_3P$	−56
3D88[119]	$PHPH_2PH_2PHP_2H_2P_2HP_2HP_2HP_2HP_2HP_2H_2P_2H_3P_2H_3P_2H_3-P_2H_3P_2HPH_2PH_2PHP_2HP_2HP_2H_2P$	−72
3D103[120]	$P_2H_2P_5H_2P_2H_2PHP_2HP_7HP_3H_2PH_2P_6HP_2HPHP_2HP_5H_3P_4-H_2PH_2P_5H_2P_4H_4PHP_8H_5P_2HP_2$	−58
3D124[120]	$P_3H_3PHP_4HP_5H_2P_4H_2P_2H_2P_4HP_4HP_2HP_2H_2P_3H_2PHPH_3P_4-H_3P_6H_2P_2HP_2HPHP_2HP_7HP_2H_3P_4HP_3H_5P_4H_2PHPHPHPH$	−79
3D136[120]	$HP_5HP_4HPH_2PH_2P_4HPH_3P_4HPHPH_4P_{11}HP_2HP_3HPH_2P_3H_2-P_2HP_2HPHPHP_8HP_3H_6P_3H_2P_2H_3P_3H_2PH_5P_9HP_4HPHP_4$	−83

Table 3.1: Common benchmark HP sequences designed for 2D and 3D simulations. The energy is measured in units of ε_{HH}.

Chapter 4

Monte Carlo Techniques and Trial Moves

Many Monte Carlo methods lay their foundation on statistical physics. As some basic principles are often required for the generation and analysis of the simulation results, we will start this chapter by reminding the readers of a few fundamental concepts that will enter our calculations later in the next two chapters. We will then describe our simulation techniques for the rest of the chapter. The discussions on statistical mechanics and Monte Carlo techniques are kept concise as full descriptions can be found in a number of standard textbooks and "classic" literature. Techniques which are specific to our implementation and model will be discussed more extensively.

4.1 A brief review on statistical physics

4.1.1 Calculation of thermodynamic quantities

The partition function Z at a particular temperature T is defined as:

$$Z = \sum_{\mathbf{x}} e^{-E[\mathbf{x}]/k_B T}, \tag{4.1}$$

where $E[\mathbf{x}]$ is the energy of the system which in turn depends on the state $\mathbf{x}$, k_B is the Boltzmann constant, and the sum runs over all the states that the system can take. One can also count the number of states of the same energy E to give the energy degeneracy, also known as the energy density of states, $g(E)$. The partition function can also be expressed in terms of it:

$$Z = \sum_{E} g(E) e^{-E/k_B T}, \tag{4.2}$$

where the sum runs over all the energy levels. Since $g(E)$ does not depend on temperature T, one may calculate Z at any temperature with a single computation of $g(E)$, which then gives access to thermodynamic quantities at any temperature. For example, the average energy $\langle E \rangle$ and the heat capacity C_V are calculated as:

$$\langle E \rangle = \frac{1}{Z} \sum_{E} E g(E) e^{-E/k_B T}, \tag{4.3}$$

$$C_V = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2}. \tag{4.4}$$

The specific heat is defined as C_V/N accordingly.

For a structural parameter, Q, which could be one of those in Section 3.3, the partition function Z and its expectation value can be obtained likewise from the two-dimensional

density of states, $g(E, Q)$:

$$Z = \sum_{E,Q} g(E, Q) e^{-E/k_B T}, \tag{4.5}$$

$$\langle Q \rangle = \frac{1}{Z} \sum_{E,Q} Q g(E, Q) e^{-E/k_B T}. \tag{4.6}$$

It is also informative to calculate the thermal derivative of the average structural quantity from Eq. (4.6) by finite difference:

$$\frac{d \langle Q \rangle}{dT} \approx \frac{\Delta Q}{\Delta T}. \tag{4.7}$$

It measures the fluctuation of $\langle Q \rangle$ as temperature varies, which often accompanies a structural transition.

4.1.2 Remarks on "phase transitions"

Theoretically, phase transitions exist only in the thermodynamic limit, i.e., when the system size is infinitely large. It is because physical quantities for a finite system are smooth functions of temperature, they cannot have true singularities which signify phase transitions. Traditionally, finite size scaling is applied to study phase transitions by extrapolating finite size results to the thermodynamic limit. However, finite size scaling does not exist for our HP model simulations because the thermodynamics depends not only on the chain length but also on other features of the HP chain like the proportion or sequence of H and P monomers. Therefore, we emphasize that the "phase transitions" in this work are indeed pseudo-phase transitions, or more precisely, structural transformations for finite systems.

The thermodynamics of the structural quantities in addition to the specific heat, C_V, are essential in identifying "transitions" between different structural "phases". In cases where the specific heat shows ambiguous signals, structural quantities help clarifying the types of transition taking place at different temperatures. In some cases distinct signals might

be missing in the specific heat, whereas structural quantities are more reliable to identify structural transitions.

4.2 Metropolis Sampling

Metropolis sampling[73] is a groundwork in Monte Carlo methods and is widely used in statistical physics. It is a special case of the importance sampling, where the outputs are distributed non-uniformly.

At the beginning of the simulation, an initial state of the model system to be simulated is first set up, and the initial energy E_0 is calculated. The temperature T of the simulation is also specified. Next, a new trial state is generated and its energy, E_{trial}, is calculated. The change in energy is then $\Delta E = E_{trial} - E_{old}$, where E_{old} is the energy of the current state. The probability for the system to transform from the current state to the trial state is determined by:

$$W(E_{old} \to E_{trial}) = \min\left(1, e^{-\Delta E/k_B T}\right), \tag{4.8}$$

where k_B is the Boltzmann's constant. If $\Delta E < 0$, the proposed move will automatically be accepted. If $\Delta E > 0$, a random number $r \in [0, 1]$ needs to be generated. If $r < W(E_{old} \to E_{trial})$, the trial state will be accepted, otherwise the old state will be kept. The above procedure is repeated until a desired number of Monte Carlo steps have been performed. At the end of the simulation, the averages of all physical quantities of interest are calculated.

Suppose P_{old} and P_{trial} are the probability of finding a microstate with energy E_{old} and E_{trial} respectively. If a transition probability fulfills the detailed balance condition such that:

$$P_{old} W(E_{old} \to E_{trial}) = P_{trial} W(E_{trial} \to E_{old}), \tag{4.9}$$

then the generated series of accepted states will distribute according to P. Since the above

Metropolis procedure satisfies the detailed balance condition when

$$P_{old} = e^{-E_{old}/k_BT} \qquad \text{and} \qquad P_{trial} = e^{-E_{trial}/k_BT},$$

the states generated are then distributed according to the Boltzmann distribution at a certain temperature T.

One unavoidable drawback of Metropolis sampling is that it is easily trapped in metastable states of the simulated system. This makes the simulations at low temperature practically intractable, and one must be careful in determining the temperature region in which Metropolis results are unreliable. Figure 4.12 at the end of this chapter shows one such example. More discussion will be given in Section 4.5.4.

4.3 Wang-Landau (WL) Algorithm

4.3.1 Description of the algorithm

Wang-Landau (WL) sampling[99,100,121] is an iterative Monte Carlo algorithm to estimate the density of states, $g(E)$, by generating a series of configurations. The simulation begins with an initial guess for $g(E)$, which can be any sensible estimation with $g(E) > 0$ or merely a simple guess as $g(E) = 1, \forall E$. We also accumulate a histogram $H(E)$ for the same energy range, which will help determine the flow of the simulation at a later time. It is set to be zero at the beginning, i.e., $H(E) = 0, \forall E$.

The initial configuration is a horizontal straight chain a few lattice spacings above the bottom surface. There are no interactions within the chain nor with the surface, so that the initial energy is $E = 0$ according to Eq. (3.1). We have performed some control experiments for a short sequence that began with different initial configurations, and the results were consistent with those obtained with the horizontal structure to within the error bars. We

thus believe that it is valid to start our simulations with a straight chain configuration.

Suppose the original configuration has an energy of E_{old}. Then a trial configuration of energy E_{trial} is generated by some Monte Carlo trial moves (which will be introduced in Section 4.5). The acceptance probability, $P(E_{old} \to E_{trial})$, is inversely proportional to $g(E_{trial})$:

$$P(E_{old} \to E_{trial}) = \min\left(1, \frac{g(E_{old})}{g(E_{trial})}\right). \tag{4.10}$$

Therefore, if $g(E_{old}) > g(E_{trial})$, the proposed move will be accepted automatically. Otherwise a random number $r \in [0, 1]$ is drawn, the trial configuration will be accepted if $r < P(E_{old} \to E_{trial})$ (then $E = E_{trial}$); or else the old configuration will be restored ($E = E_{old}$) and the trial one will be rejected. $g(E)$ is then modified by multiplying the existing value by a modification factor f, which has an initial value of $f_0 = e^1 = 2.71828...$ at the beginning of the simulation. $H(E)$ is also accumulated such that:

$$g(E) \to g(E) \times f, \tag{4.11}$$

$$H(E) \to H(E) + 1. \tag{4.12}$$

$H(E)$ is indeed a statistic to keep track of the number of visits to an energy E within an iteration interval. Note that if a trial move is rejected and the old state is restored, E_{old} has to be counted again. The updates of configurations, $g(E)$ and $H(E)$ repeat until $H(E)$ is sufficiently "flat" over the entire energy range. The simulation is then brought to the next iteration: $H(E)$ is reset to zero and f is reduced, $f \to \sqrt{f}$, but $g(E)$ is retained. A "flat" histogram refers to a histogram $H(E)$ for which all entries are not less than $p \times H_{ave}$, where H_{ave} is the average of all entries in $H(E)$ and p is called the flatness criterion, which is a tunable parameter for achieving a desired accuracy, and $0 < p < 1$. The larger the value, the more accurate the results are. The iteration goes on and $g(E)$ is continuously modified during the simulation until the modification factor is smaller than some predefined value,

f_{final}. At this point $g(E)$ should have converged to its true value after normalization.

For our system, $g(E)$ is normalized in such a way that

$$\sum_{E} g(E) = 1. \tag{4.13}$$

It is because, unlike the Ising model for which the ground state degeneracy is exactly known to be 2, the energy degeneracy is an unknown for our system and cannot be used as a reference for normalization. Nevertheless, it does not affect the correctness of the thermodynamics, as the partition function will cancel out any scaling factor brought by $g(E)$ in the calculations of the ensemble averages.

For all the simulations presented in this work, rather stringent parameters were used in order to obtain accurate estimates for $g(E)$. The flatness criterion was set to be $p = 0.8$ for the simulations of 2D36 and 3D48, and $p = 0.6$ for 3D67 and 3D103. $g(E)$ was estimated iteratively until the natural log of the modification factor, $\ln(f)$, was brought below 10^{-8} in all cases.

4.3.2 Considerations specific to our simulations

(i) **WL sampling as a random walk in energy space**

WL sampling is an efficient and robust simulation method as it focuses on the estimation of $g(E)$ which is independent of temperature. Thus, all thermodynamics at any temperature can be calculated from one single simulation. It performs a random walk in energy space by accepting a trial configuration with a probability proportional to the reciprocal of $g(E)$ instead of the Boltzmann factor e^{-E/k_BT} in traditional Metropolis sampling, so that there is not a problem for the random walker being trapped in local minima of the free energy. WL sampling is, therefore, able to simulate systems possessing rough free energy landscapes, i.e., the free energy as a function of a specific reaction

coordinate (which could be a structural quantity, for instance) has many maxima and minima.

(ii) **Detailed balance issue**

As $g(E)$ is constantly modified, the algorithm does not satisfy detailed balance during the early iterations. However, by the end of the simulation, $g(E)$ does not change as rapidly as in the beginning when the modification factor approaches unity. By Eq. (4.10), it is straight-forward to arrive at:

$$\frac{1}{g(E_{old})}P(E_{old} \to E_{trial}) = \frac{1}{g(E_{trial})}P(E_{trial} \to E_{old}), \tag{4.14}$$

where $1/g(E_{old})$ and $1/g(E_{trial})$ are the probabilities of finding a microstate with energy E_{old} and E_{trial} respectively. The WL algorithm thus converges toward detailed balance at a later stage of the simulation.

The convergence of $g(E)$ in WL sampling is proven to be achieved by an optimization procedure rather than by means of Markov chain Monte Carlo arguments[122]. Recently, the convergence has also been derived from the detailed balance condition and improvements have been suggested[10].

(iii) **"Self-adaptive" energy levels**

The knowledge of the full energy range is essential in the WL algorithm for the examination of the flatness of the histogram, but the ground state energy is *a priori* unknown for the HP model. To overcome this difficulty, the following procedure was used: every time a new energy E_{new} was found, it was marked as "visited" and $g(E_{new})$ was set to g_{min}, i.e., the minimum entry among all previously visited energy levels. The flatness of the histogram is checked only for those energy levels which have been marked visited. With this self-adaptive procedure, new regions of conformational space can be explored

simultaneously as the random walker would not spend a vast amount of time on this newly found energy level trying to catch up with other entries.

(iv) **Performance improvements of WL sampling**

Since the birth of WL sampling, variations are proposed from time to time to boost the performance of WL sampling in general[122]. Recently, some improvements have been suggested to speed up and ensure the convergence of WL sampling in the simulation of lattice polymers or proteins[109,123,124].

We also noticed that WL sampling is rather computationally intensive when applied to complex systems, even for a lattice protein model like the HP model in this study. The sampling algorithm of the HP model can be accelerated through the implementation of specialized coding techniques[125]. For more general applications, a parallel realization of WL sampling has been designed and is under development at the time of writing[126].

4.4 Multicanonical (MUCA) Sampling

4.4.1 Description of the algorithm

In the second stage of our simulation, we estimated the joint density of states, $g(E, Q)$, by multicanonical (MUCA) sampling[127,128]. This two-dimensional density of states can be used to obtain the thermodynamics of a structural observable Q. Although it is feasible to sample $g(Q)$ or $g(E, Q)$ all over again using a one-dimensional or a two-dimensional random walk in WL sampling if only one structural quantity is concerned, it becomes extremely time-consuming if several of them are of interest. A more efficient way is to make use of the prior knowledge of $g(E)$ and perform a multicanonical sampling. In this process, trial states are accepted or rejected according to a weight, $W(E) \propto 1/g(E)$, where this $g(E)$ is the density of states in energy obtained previously by the one-dimensional WL sampling.

Unlike the previous WL process, $W(E)$ is held fixed throughout the whole multicanonical production procedure.

Again, suppose the original configuration has an energy of E_A. A new trial state of energy E_B, and a random number, $r \in [0, 1]$, are generated. The decision of whether it will be accepted depends on the same acceptance probability as Eq. (4.10), i.e., if $r < P(E_{old} \to E_{trial})$. If the new configuration is taken, any desired structural quantity Q will be evaluated, the corresponding two-dimensional histogram, $H(E, Q)$, will be accumulated as in Eq. (4.12). If the old configuration is retained, E and Q will be counted again. The simulation is brought to an end when a predetermined number of Monte Carlo steps have been performed. The joint density of states, $g(E, Q)$, is then obtained by reweighting $H(E, Q)$ with $g(E)$:

$$g(E, Q) = g(E)H(E, Q), \tag{4.15}$$

followed by normalization:

$$\sum_{E,Q} g(E, Q) = 1. \tag{4.16}$$

As such, we can obtain $g(E, Q)$ for as many Q as desired in a single MUCA production run.

4.4.2 Considerations specific to our simulations

One important factor governing the accuracy of $g(E, Q)$ obtained by MUCA is the number of Monte Carlo steps, which is not at all obvious to determine. In our simulations, a few measures were taken to ensure sufficient sampling in this MUCA production stage so that we are confident that the phase space is reasonably sampled:

(i) the number of visits to the ground state exceeds a certain preset value, say, 10^4 times

(ii) the entire energy range has to be covered

(iii) the average energy, specific heat or other thermodynamic quantities calculated from

the resulting $g(E, Q)$ agree with the ones calculated from the $g(E)$ obtained by WL sampling

(iv) the random walker has not been "stuck" at a particular energy for a long time. This can be confirmed by investigating the time series (energy as a function of Monte Carlo time) of the simulation

4.5 Monte Carlo trial moves for lattice polymers

4.5.1 "Traditional" moves and their limitations

Traditional Monte Carlo trial moves for lattice polymers either change a conformation locally or globally. Local moves include the end-flip (Figure 4.1), kink-flip (Figure 4.2) and crankshaft (Figure 4.3). They generate new configurations fairly similar to the old ones as most parts of the polymer remains unchanged, inducing long correlation times in the simulation. The pivot move (Figure 4.4) is the most common non-local move. It does not share the same problem with the local moves, but it is ineffective in generating new states from dense conformations, making for its high rejection rate.

(a) (b)

Figure 4.1: End flips.

Figure 4.2: A kink flip.

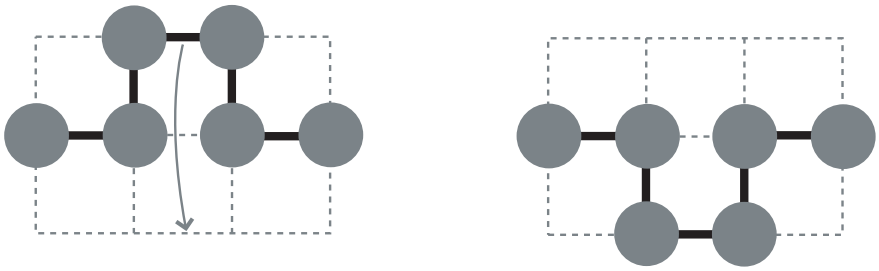


Figure 4.3: A crank shaft move.

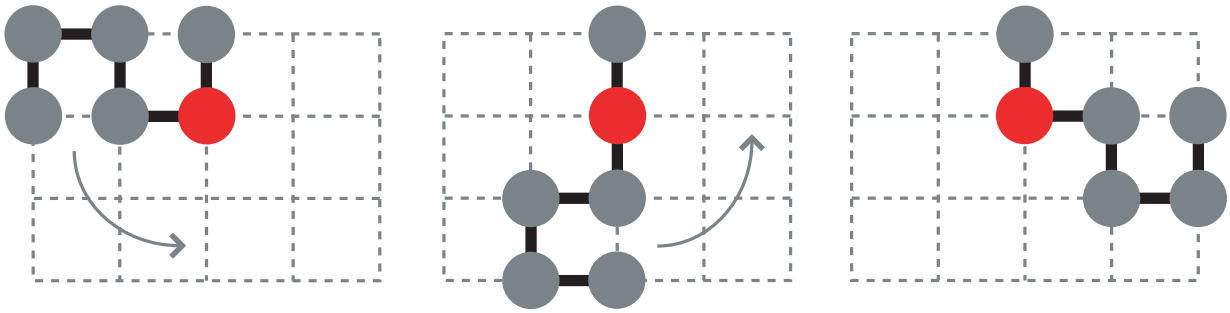


Figure 4.4: Pivot moves.

It has been found that two inventive trial moves, pull moves[79] and bond-rebridging moves[129], are able to eradicate the problems. They work particularly well together with WL sampling in search of the global energy minimum conformations and the determination of the density of states for lattice polymers[102,103]. The ability of reaching low energy states allows for a more thorough survey of the conformational space; thus a higher resolution of $g(E)$ and more precise thermodynamic quantities, especially in the low temperature regime, can be obtained. This is of particular importance for systems with longer chain lengths and more complex energy landscapes[130]. We therefore adopted the same strategy for our

simulations in this work.

4.5.2 Pull moves

Pull moves were originally proposed to combine with the tabu algorithm in search of new minimum energy configurations for the 2D HP model[79]. Later it was found to be equally effective when combined with WL sampling[103].

The robustness of pull moves comes from two important properties which have been proven mathematically: reversibility and ergodicity. Reversibility means that for any move in a move set M applied to a configuration to form a new one, there is some move in M that can restore the configuration to the original one. Ergodicity refers to the fact that any configuration is reachable from any other valid configuration through a sequence of pull moves in M. Theoretically, all microstates in phase space are equally probable over a long period of time with the use of pull moves.

Description of the move

The way that pull moves are designed gives them a good balance between local and non-local moves since depending on the starting conformation and the part of the N monomer chain where a pull move is performed, it can displace only one monomer or up to $N-1$ monomers of the entire chain. We now illustrate the implementation of pull move in a square lattice, but it can be easily generalized to a cubic lattice:

1. For monomer i located at (x_i, y_i) and monomer $i+1$ located at (x_{i+1}, y_{i+1}), denote an unoccupied site L so that it is adjacent to monomer $i+1$ and diagonally adjacent to monomer i. Monomers i and $i+1$ together with site L form the three corners of a square. If there are no empty adjacent sites to denote as site L, a pull move cannot be performed, and the procedure starts over again with another monomer.

2. Denote the fourth corner of the square as site C. Check that it has to be empty, or else it must be (x_{i-1}, y_{i-1}), i.e., it is occupied by monomer $i-1$. Figure 4.5 shows the notations defined up to this point.

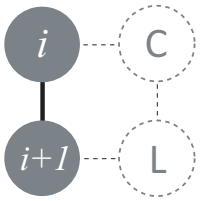


Figure 4.5: Annotations used for a pull move.

3. If $C = (x_{i-1}, y_{i-1})$, move monomer i to site L and the move is done (see Figure 4.6). It is equivalent to a local kink-flip move.

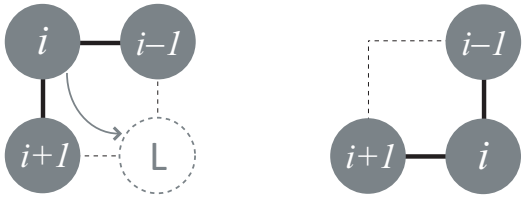


Figure 4.6: A pull move when site C is preoccupied by monomer $i-1$.

4. If C is empty, move monomer i to site L and monomer $i-1$ to site C. Check if monomer $i-1$ can be connected to monomer $i-2$. If so, the move completes (as shown in Figure 4.7); otherwise, proceed to the next step.

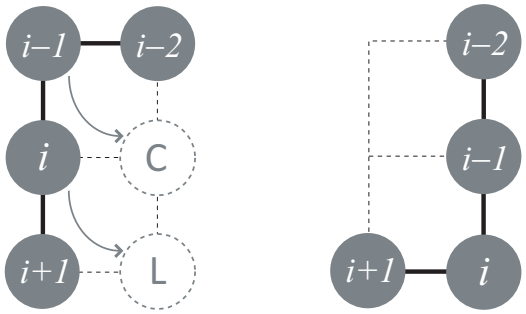


Figure 4.7: A pull move when site C is adjacent to monomer $i-2$.

5. Starting from monomer $j = i - 2$ down to monomer 1, pull the monomers two spaces up the chain, i.e., move monomer j to (x_{j+2}, y_{j+2}), repeat it for monomer $j - 1$ and so on, until a valid configuration is formed. See Figure 4.8 for such an example.

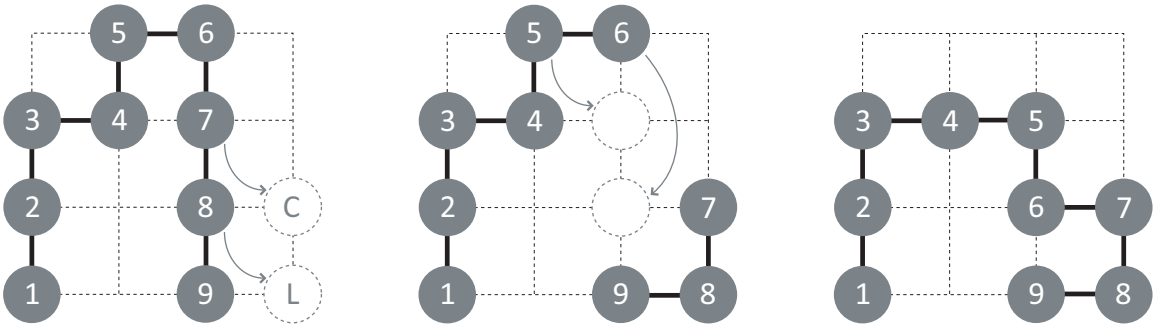


Figure 4.8: A pull move that causes part of the chain to relocate.

The detailed balance consideration

The violation of detailed balance condition can be introduced by a non-symmetric probability matrix that is used to generate the trial moves, inducing a possible bias in the final estimate of $g(E)$[131]. It is then important to correct Eq. (4.14) for unequal move ratios for pull moves in order to eliminate the problem. An extra factor is inserted to the acceptance probability of moving from state A (with energy E_A) to state B (with energy E_B):

$$P(A \to B) = \min\left(1, \frac{g(E_A)}{g(E_B)} \frac{n_{B\to A}/n_B}{n_{A\to B}/n_A}\right), \tag{4.17}$$

where $n_{A\to B}$ is the number of pull moves that transform A to B; n_A is the number of possible pull moves which can be performed on state A; $n_{B\to A}$ and n_B are defined likewise. The proven reversibility of pull moves has ensured that $n_{A\to B} = n_{B\to A}$, which causes these two terms to cancel out. Eq. (4.17) then becomes:

$$P(A \to B) = \min\left(1, \frac{g(E_A)}{g(E_B)} \frac{n_A}{n_B}\right). \tag{4.18}$$

When a pull move is chosen to generate a new configuration, a list of all possible moves that can be performed on state A is first constructed to obtain n_A. A move is selected at random from the list, generating state B. n_B is then obtained by counting the number of all possible moves that can be performed on state B. $P(A \to B)$ can thus be calculated. This list construction procedure is computationally intensive, and it slows down the simulation by approximately an order of magnitude. For the sake of the correctness of our results, we have kept this procedure in all of our simulations to guarantee detailed balance.

4.5.3 Bond rebridging moves

When the conformation of a lattice polymer is dense, it is extremely difficult to perform a global move as it is unlikely to find vacant lattice sites around to accommodate a group of monomers at a time without overlapping with each other. The bond rebridging move was proposed to resolve the difficulty[129]. It is termed a "long range move" in the original paper for its capability of generating a very different compact conformation from another compact one, which makes it powerful in exploring different parts of the conformational phase space.

Description of the move

We now describe the move using a square lattice. Again, it can be generalized to a cubic lattice in a similar manner.

1. Pick two consecutive monomers, i and $i+1$, randomly. The displacement vector between them gives the local direction of the chain.

2. Choose a unit vector at random which has a direction normal to the one formed by the selected monomers (there are two such directions on a square lattice and four on a cubic lattice).

3. If the neighbors of monomers i and $i+1$ in the chosen direction are occupied by two other adjacent, bonded monomers, two parallel strands are found. Proceed to the next step in this case. Otherwise the process starts over again.

4. Denote the sites neighboring to monomers i and $i+1$ as j and k respectively as shown in Figure 4.9(a). If $k - j = -1$, the two strands are anti-parallel and move type 1 will be performed. Otherwise, the two strands are parallel and move type 2 will be performed.

5. **Type 1 move:**

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

   (b) Make a link between i and j, and a link between $i+1$ and k. The chain is now broken into a segment and a loop.

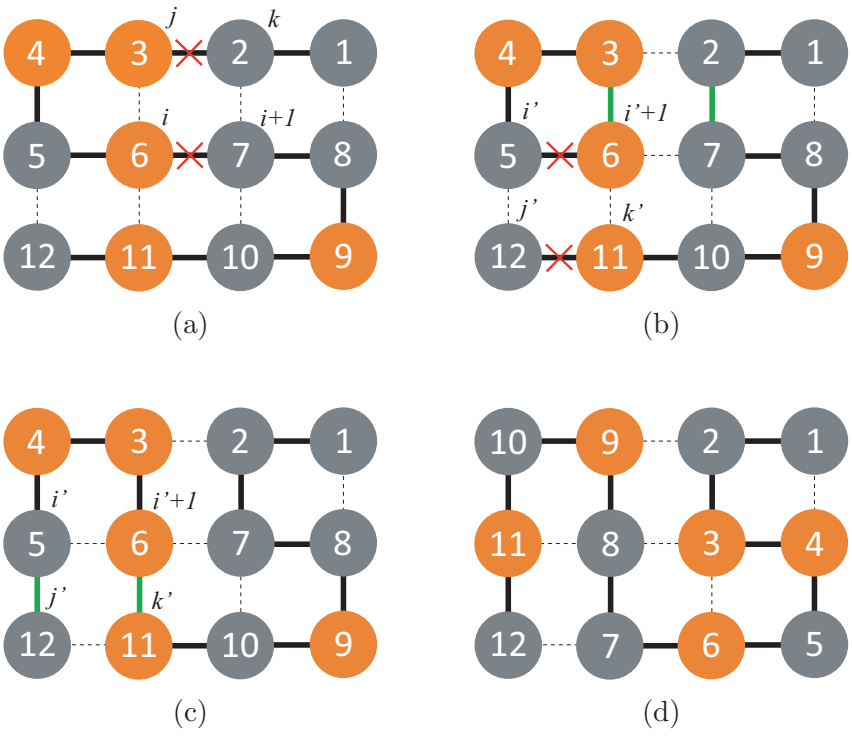


Figure 4.9: Type 1 bond rebridging move.

(c) Choose two consecutive monomers i' and $i'+1$ in the loop at random to form a vector, and see if a segment parallel (or anti-parallel) to it can be found. If so, do the "cut-and-join" again like steps (a) - (b); if not, a move cannot be performed and the whole process starts over again.

(d) Renumber the monomers to restore the sequence of the chain; reassign the H/P type accordingly.

Type 2 move:

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

(b) Make a link between i and j, and a link between $i+1$ and k.

(c) Renumber the monomers.

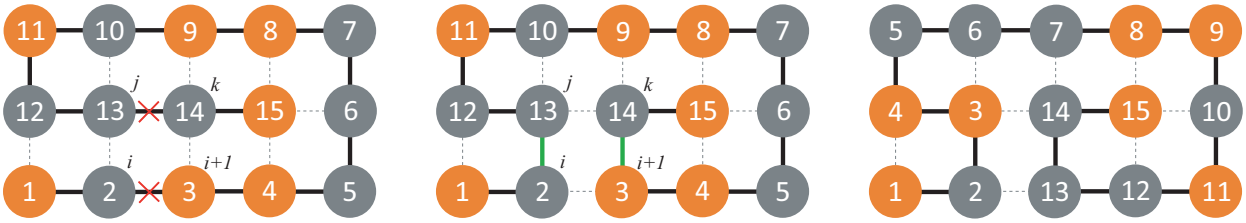


Figure 4.10: Type 2 bond rebridging move.

The bond rebridging move can also be performed to an end monomer as an end move as illustrated by the following (Figure 4.11):

1. An end monomer is chosen at random and denoted as i.

2. Out of the three neighboring sites which are not directly connected to the end monomer, choose one at random. If it is empty, try another site; otherwise, denote it as j.

3. Join monomers i and j.

4. A bond between j and one of its neighbors is cut so that a valid configuration is formed.

5. Renumber the monomers to restore the sequence of the chain; reassign the H/P type accordingly.

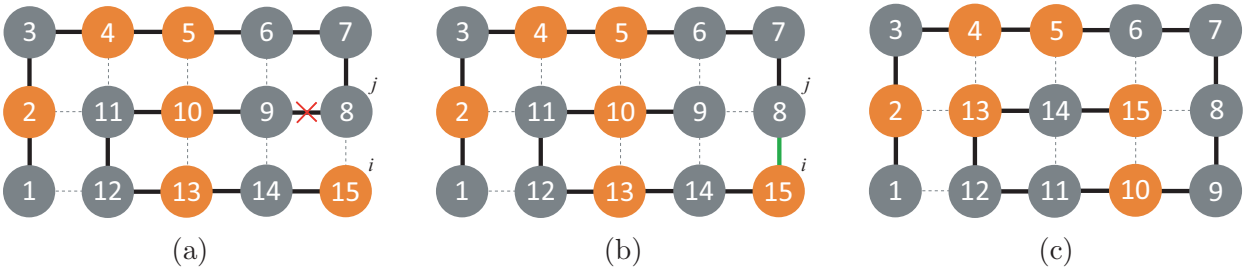


Figure 4.11: Bond rebridging move involving an end monomer.

A bond rebridging move is effective for high density polymers as it rearranges the bonds rather than displacing the monomers. And as the move itself obeys detailed balance, it is unnecessary to correct the acceptance probability in Eq. (4.10) as for pull moves. One obvious drawback of the bond rebridging move, however, is the non-ergodicity, as movements of monomers are not involved.

4.5.4 Combination of different moves and algorithms

The two trial moves are called with different probabilities. Bond rebridging move transforms a polymer from one compact state to another without uncoiling, making it more efficient than pull moves in dealing with densely packed polymers. It also makes a huge energy difference between consecutive moves; yet its acceptance rate is rather low because there are only a small number of possible moves available for a given configuration. This drawback is compensated for with a higher calling ratio. In our simulations, every time a new configuration is to be generated, there is an 80% chance that bond rebridging moves would be used and only a 20% chance pull moves would be used.

However, it should be noted that these non-traditional trial moves alone are not able to give correct low temperature thermodynamics if they are not combined with a suitable Monte Carlo algorithm. As an illustration, we have compared WL with Metropolis sampling in obtaining the specific heat of the 2D36 sequence interacting with a very weak attractive surface in Figure 4.12. The two transition peaks at low temperature are clearly missing in the Metropolis case, even though a very large number of trial moves (10^8) was used! Although Metropolis sampling gave a seemingly correct answer with small statistical errors, the low temperature results are obviously wrong, for they fail to predict the adsorption and "flattening" processes (which will be introduced in Chapter 5) at low temperature, even when combined with pull moves and bond rebridging moves.

We thus stress that an appropriate combination of both the sampling method and trial updates is crucial in obtaining correct results from a Monte Carlo simulation.

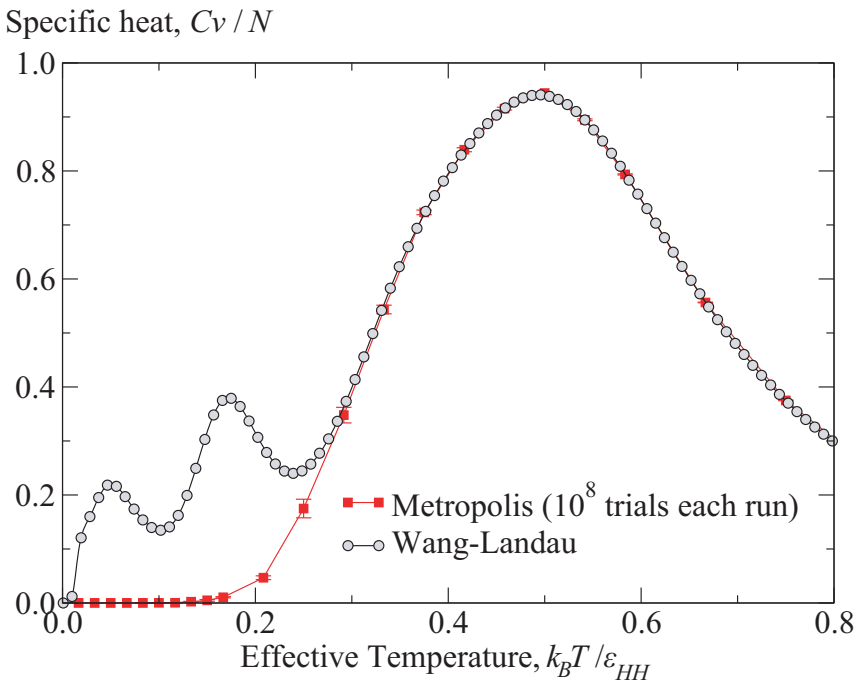


Figure 4.12: Comparison between Wang-Landau and Metropolis sampling in obtaining the specific heat of 2D36 interacting with a very weak attractive surface, for which $\varepsilon_S/\varepsilon_{HH} = 1/12$. 15 independent runs were performed for both algorithms to obtain statistical errors (which are not shown as they are smaller than the data points).

Chapter 5

Identifying Structural "Transitions" by Thermodynamic and Structural Quantities in a Short HP Sequence

In this chapter, we will present our earliest simulation results for two short HP benchmark sequences, 2D36 and 3D48. We will discuss and compare some fundamental physical properties of 2D36, both in the absence and presence of a substrate. We will then demonstrate how these structural changes, or "phase transitions", can be identified utilizing thermodynamic quantities like the specific heat, together with the thermal dependence of some structural parameters introduced in Chapter 3. This chapter serves both as a presentation of our results, and as an illustration of our method of analysis. Some of the results have been published in Refs. [132], [133] and [134].

For the examples shown in this chapter, 15 independent runs were performed with the simulation parameters given in Chapter 4 to obtain statistical errors. CPU time grows with the size of the energy range of the systems: it takes about 15 minutes to finish a simulation on an AMD Opteron dual-core 2.2 GHz processor for the surface-free case using 19 energy

bins; 5-10 hours for the strong attractive surface case with 51 energy bins; but around 10 days are generally needed for the weak attractive surface where there are 242 energy bins in the full energy range.

5.1 The density of states

The immediate output from a Wang-Landau sampling is the density of states in energy, $g(E)$. Figure 5.1 shows some typical densities of states for sequence 3D48 in the absence of a surface (both in two and three dimensions), and in the presence of a strongly attractive surface ($\varepsilon_S/\varepsilon_{HH} = 1$ or 2). The difference in the magnitude of $g(E)$ grows with the chain length of the sequence. For 3D48 as shown in Figure 5.1, the densities of states span a wide range of about 30 orders of magnitude for the adsorption cases. For 3D103 (which is not shown here), the density of states spans an even wider range of about 60 orders of magnitude. This huge difference in the degeneracy of high and low energies is one of the reasons why even a small lattice polymeric system is difficult to simulate. This also explains why the ground state search is a challenge.

Figure 5.2 shows an interesting $g(E)$ obtained for 3D36, interacting with a very weakly attractive surface ($\varepsilon_{HH} = 12, \varepsilon_S = 1$). The $g(E)$ looks like saw-teeth with spikes every multiple of ε_{HH}. This is a typical shape for a $g(E)$ of a system with a weakly attractive surface when ε_{SH} or $\varepsilon_{SP} < \varepsilon_{HH}$. To understand why a spike exists, we now take the system in Figure 5.2 as an example. The second spike from the right occurs at $E = -12$, and there are two ways of achieving it: (i) by forming one H-H interaction and no surface interaction, i.e., $n_{HH} = 1, n_{SH} = n_{SP} = 0$; (ii) by forming no H-H interaction but 12 surface interactions, i.e., $n_{HH} = 0, n_{SH} + n_{SP} = 12$. Therefore, there are more combinations of getting an energy which is a multiple of ε_{HH}, and this extra degree of freedom comes from the formation of H-H interactions. Another way of seeing this is to count the number of spikes. In Figure

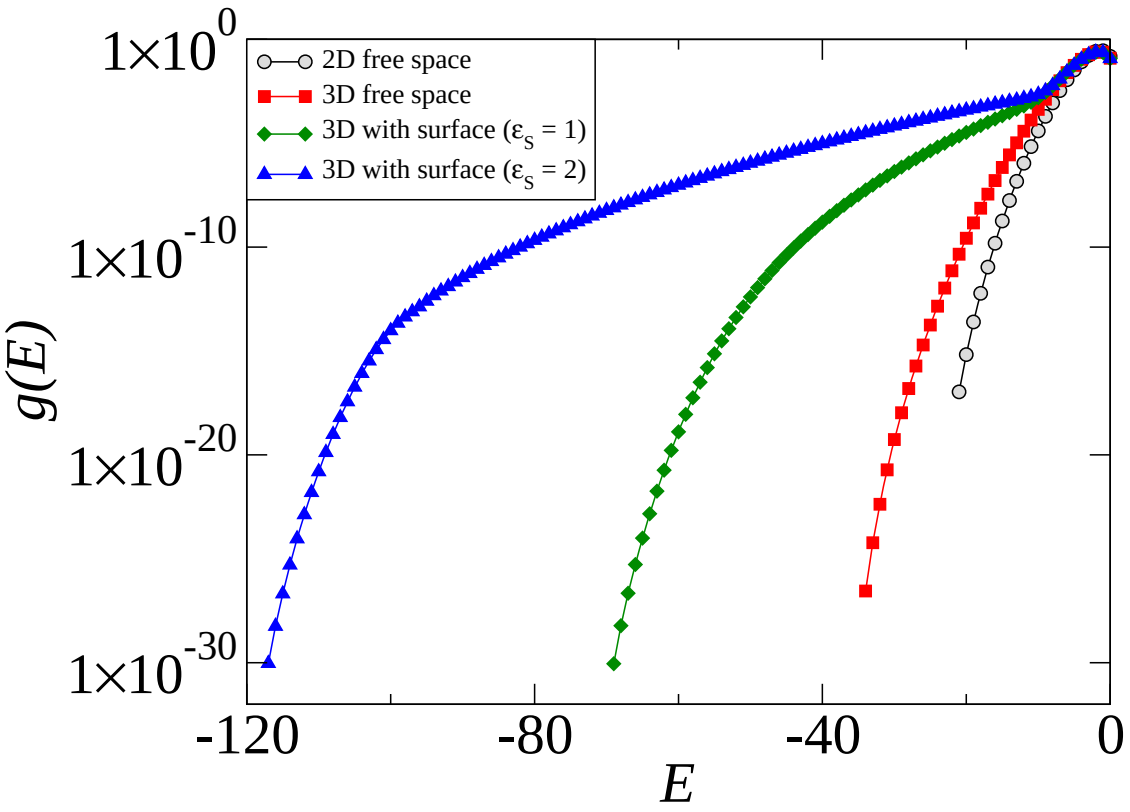


Figure 5.1: The densities of states in energy for 3D48 in free space and with two strongly attractive surfaces. Errors are smaller than the data points.

5.2, there are 18 spikes excluding the one at $E = 0$. In a 3D free space case without the attractive surface, the ground state energy is found when $n_{HH} = 18$ (as will be discussed in the following section). This agreement with the number of spikes shows another evidence that the spikes correspond to the numbers of n_{HH} available in the system.

Next we would like to stress the importance of obtaining $g(E)$ with lowest energy states in the calculation of thermodynamics. Here, we look at an example demonstrated by a simulation for 3D48 interacting with a relatively weakly attractive surface, in which $\varepsilon_{HH} = 2, \varepsilon_{SH} = 1, \varepsilon_{SP} = 0$. Ten individual runs were started at the beginning, six of which found the ground state $E = -79$, but four of which only discovered the first excited state $E = -78$.

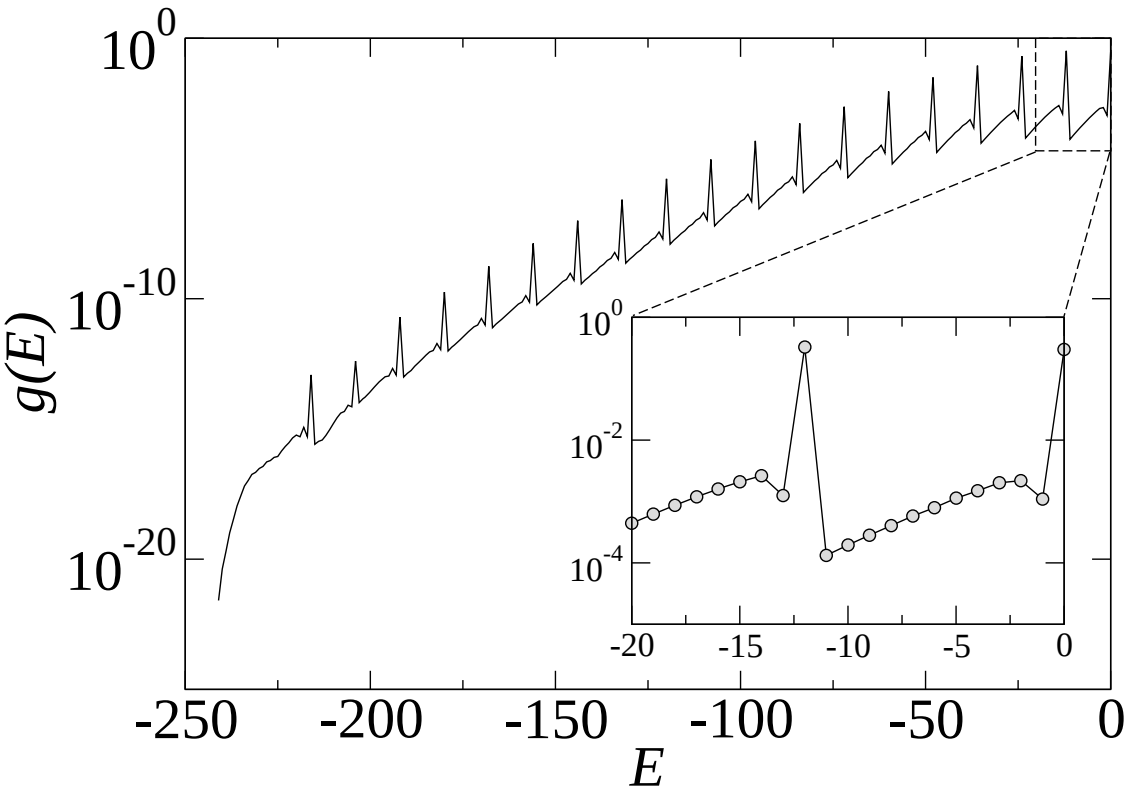


Figure 5.2: The densities of states in energy for 2D36 interacting with a very weakly attractive surface ($\varepsilon_{HH} = 12, \varepsilon_S = 1$). Errors are smaller than the data points. The inset shows a close-up of one of the saw-teeth in the high energy region.

These runs were then divided into two groups and the specific heat curves C_V/N were calculated separately and were compared in Figure 5.3. From the top inset of the figure, one can see that the transition peak height and temperature are not much affected, as both results agree with each other to within the error bars. However, from the lower inset of the figure, which shows C_V/N at a very low temperature, a shoulder exists for the one calculated from the $g(E)$ with the ground state, and it is clearly missing for the one calculated from the $g(E)$ without the ground state. For only a difference of whether the ground state has been included in the calculation, a significant discrepancy in the low temperature thermodynamics is resulted. This is also a major cause of the difficulty in studying the low temperature

bel

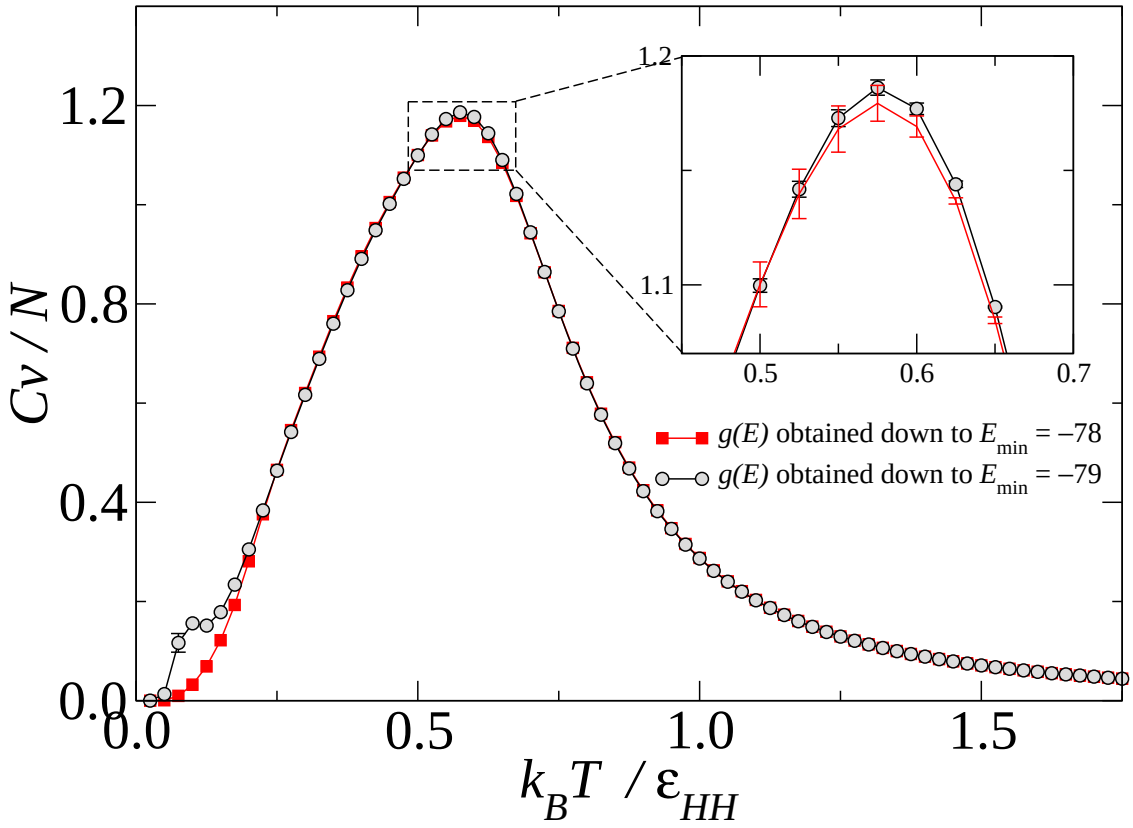


Figure 5.3: Effect of the range of density of states on the low temperature regime of the specific heat C_V/N for 3D48 interacting with a weak attractive surface ($\varepsilon_{HH} = 2, \varepsilon_{SH} = 1, \varepsilon_{SP} = 0$). Four runs obtained $g(E)$ down to $E = -78$, while six runs obtained $g(E)$ down to $E = -79$. The inset magnifies the peak region of the specific heat.

5.2 Structural changes in the absence of a substrate

The thermodynamics and structural transitions of an HP chain folding freely on a 3D cubic lattice without an adsorbing substrate have been studied extensively in some existing work. Generally speaking, the acquisition of the ground state structure is a two-stage process: the coil-globule transition at a higher temperature, and a globule-ground-state transition at a lower temperature. These can be verified by investigating the equilibrium states as temperature decreases.

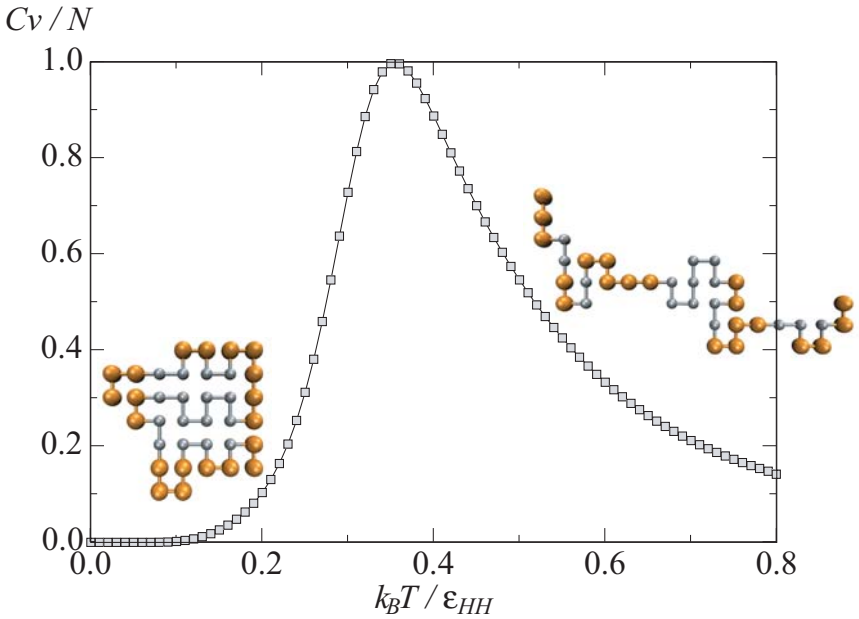
At high temperature, the HP polymer has a random, extended coil structure. As the temperature is lowered, it undergoes a coil-globule transition to form a fairly compact structure, in which a hydrophobic core is primarily constructed in the center and polar residues residing on the outside. Nonetheless, this globule is not perfectly packed and some loose monomers can still be found. This hydrophobic core forming process involves rapid formation of H-H pairs that causes a vast decrease in energy and thus a pronounced peak in the specific heat.

Further decrease in temperature induces the globule-ground-state transition, where the globule is first partially unwound then collapses again to form a compact ground state, where the hydrophobic core attains an optimal shape that minimizes the energy. This process causes some energy fluctuations but not as much as the coil-globule transition, leading to a smaller peak or a shoulder in the specific heat.

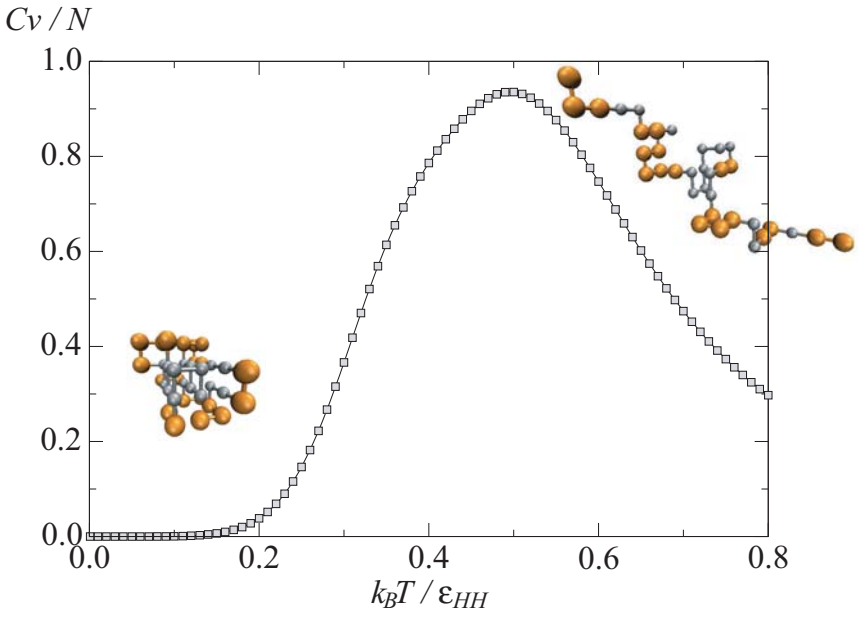
However, depending on the HP sequence, the globular phase might not exist in some cases. The acquisition of the ground state is then a single process of coil-ground-state transition. To obtain a verification of the above descriptions, we simulated the sequences 2D36 and 3D48 in a two-dimensional and a three-dimensional free space. For 2D36, we will compare its fundamental structural transitions in the vicinity of an absorbing bottom surface in the following sections. Results for other larger sequences will be presented in Chapter 6.

The specific heat and typical states before and after the transition occurs are shown in Figure 5.4. 2D36 undergoes a collapse transition as temperature decreases in both the 2D and 3D cases, resulting in a compact hydrophobic core with polar residues residing on the exterior to screen the core from the polar solvents. However, the folding to the ground state from a random coil in the 3D space is a single process instead of the two-step process as expected. It is probably because the sequence is too short, and so the construction of a hydrophobic core seems to be highly cooperative and straight-forward.

We then turn to the specific heat for 3D48 as shown in Figure 5.5. In the 3D free space case, the coil-globule transition is signaled by the peak at $k_BT/\varepsilon_{HH} \approx 0.5$, while the acquisition of ground state from a globule is signaled by the shoulder at $k_BT/\varepsilon_{HH} \approx 0.25$. This is, of course, in complete agreement with the aforementioned studies.



(a)



(b)

Figure 5.4: The specific heat and typical states before and after the coil-globule transition for 2D36. The HP chain is displayed with orange (larger) polar and grey (smaller) hydrophobic residues.
(a) The two-dimensional free space case. The ground state energy is found to be $E = -14$; equivalently, 14 H-H bond pairs are formed ($n_{HH} = 14$).
(b) The three-dimensional free space case. The ground state energy is found to be $E = -18$ (i.e., $n_{HH} = 18$).

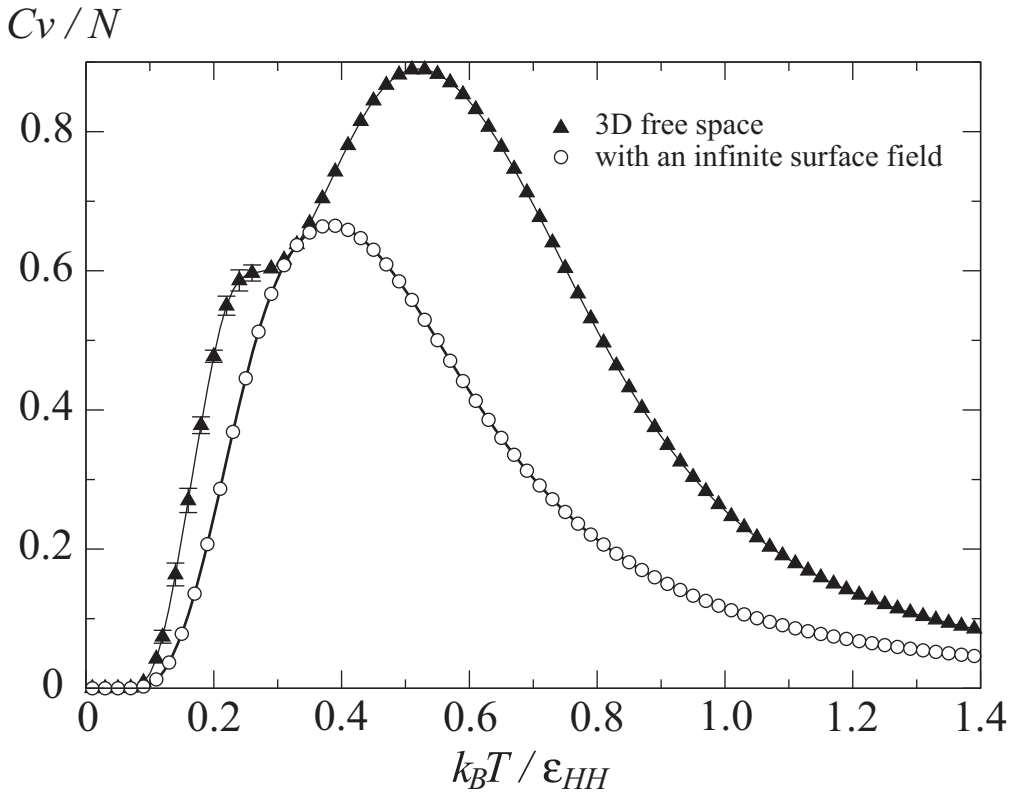


Figure 5.5: Comparison of the specific heat for 3D48 in a two-dimensional and a three-dimensional free space.

5.3 Limiting behavior in the presence of a substrate

The two cases in free space together serve as the "asymptotic" folding behavior in two limiting cases. Since 2D36 was originally designed for folding on a square lattice to attain a square hydrophobic core in the ground state configuration[77], it is therefore an ideal test case for comparing the ground state configurations when the chain interacts with a strongly attractive surface in a 3D space.

The 2D free space case resembles the limit of a surface with infinite attractive strength, since it is equivalent to restricting all monomers of the HP chain on the surface to yield a film-like, two-dimensional ground state. The 3D case corresponds to another limit where the surface is extremely weakly attractive. It also sets a reference to the high temperature thermodynamics for the adsorption cases when the HP protein is not interacting with the surface.

These limiting cases are also useful in visualizing upper and lower bounds for thermodynamic observables, and they serve as an aid to understanding the details of folding behavior. Besides the specific heat, another demonstrative quantity is the averaged radius of gyration per monomer, $\langle R_g \rangle / N$. In Figure 5.6, we show the $\langle R_g \rangle / N$ for 3D48 interacting with a surface attracting all monomers. The radii of gyration for the two limiting cases are plotted in the same figure.

Drawing a simple connection to the self-avoiding random walk on square and cubic lattices, it is obvious that $\langle R_g \rangle$ is largest when all the monomers are forced to sit on the surface to form planar structures, while $\langle R_g \rangle$ is smallest when the HP chain is allowed to fold freely in a three-dimensional space to form 3D structures. For this reason, the 2D and 3D free space cases give the upper and lower bound of $\langle R_g \rangle$, respectively, for the cases where the HP chain interacts with a surface, as seen in the lower panel of Figure 5.6. Generally speaking, when the HP chain is placed near a surface of finite attractive strength, it remains as an

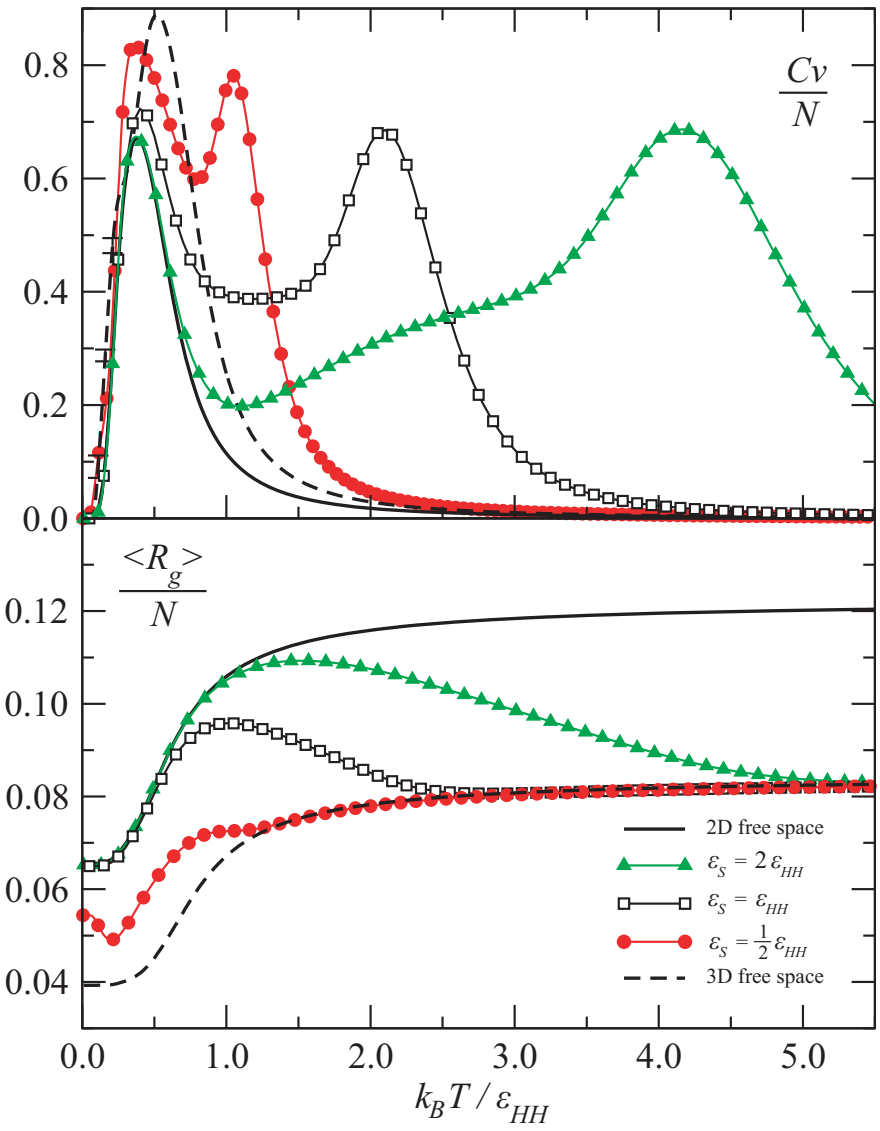


Figure 5.6: **Upper panel:** Specific heat C_V/N as a function of the effective temperature k_BT/ε_{HH} for 3D48 interacting with a surface which attracts all monomers with different strengths. **Lower panel:** Averaged radii of gyration per monomer, $\langle R_g \rangle /N$, as a function of k_BT/ε_{HH}, for 3D48 interacting with a surface which attracts all monomers with different strengths. Note that k_BT is scaled with the internal attraction strength, ε_{HH}, so as to compare different systems in the same energy scale. In this manner, any difference in quantities comes solely from the surface strength ε_S. Errors smaller than the data points are not shown.

extended coil at high temperature as if the surface is absent. The radii of gyration for all cases thus coincide with the 3D, surface-free one.

As the temperature decreases, the HP chain interacting with a stronger attractive surface ($\varepsilon_S/\varepsilon_{HH} = 2$) starts the adsorption process the earliest at $k_BT/\varepsilon_{HH} \approx 5.0$ as its $\langle R_g \rangle$ "departs" from the lower bound and begins to approach the upper bound. Such an adsorption transition is clearly signaled by the peak in C_V centered at $k_BT/\varepsilon_{HH} \approx 4.25$. At $k_BT/\varepsilon_{HH} \approx 1.0$, $\langle R_g \rangle$ merges with the upper bound signifying a complete adsorption of all monomers. The formation of a hydrophobic core in which the number of intra-chain H-H interactions, n_{HH}, is maximized, then takes place entirely on the surface until the ground state is reached at zero temperature. This process in the low temperature regime is identical to the one in two-dimensional free space, as indicated by the complete agreement in the radii of gyration and the coincidence of the peaks at $k_BT/\varepsilon_{HH} \approx 0.5$ observed in C_V.

The thermodynamics for the surface with $\varepsilon_S/\varepsilon_{HH} = 1$ is qualitatively similar to that of the former case except that it requires a lower adsorption temperature. Since the radii of gyration for both surface types end up with the same value as the upper bound at $k_BT/\varepsilon_{HH} = 0$, one may expect that the ground state conformations for both systems are two-dimensional. This has been confirmed by measuring the number of surface contacts ($n_{SH} = n_{SP} = 24$, meaning the entire chain is in contact with the surface) and the number of H-H interactions ($n_{HH} = 21$, which is the same as the ground state of the 2D limiting case).

While the two peaks in the specific heat for the surface with $\varepsilon_S/\varepsilon_{HH} = \frac{1}{2}$ tend to give the impression that this situation has the same qualitative folding behavior as the previous cases, the shape of the radius of gyration clearly distinguishes it from the others, apart from showing that the ground state is now three-dimensional. This is the first sign that the specific heat alone is not sufficient to reveal all structural transition behavior. Indeed, the transition hierarchy, i.e., the order of occurrence of different transition processes, is different from the other two in this case. This can only be verified by examining other structural

parameters, as we shall see in the following sections.

5.4 Structural transitions in the vicinity of a weakly adsorbing surface

With a surface that attracts both H and P monomers, the HP chain exhibits a much richer range of structural "phase transitions" due to the competition between surface adsorption and attraction within the polymer. The different structural "phase transitions" are best illustrated by considering the 2D36 sequence interacting with a very weak attractive surface ($\varepsilon_S = \frac{1}{12}\varepsilon_S$). The top portion of Figure 5.7 shows the specific heat as compared to that of the three-dimensional free space case, while the temperature dependence of the structural properties is shown in the bottom portion.

The height of the non-attractive wall is set to be $h_w = N + 1 = 37$, i.e., there are 36 layers between the two horizontal surfaces, the 36mer can touch both surfaces with its ends only when it is a fully stretched, vertical chain.

While there is only a single peak corresponding to the coil-globule transition at $k_BT/\varepsilon_{HH} \approx 0.5$ for a free chain, three distinct peaks are observed in the case with a weakly attractive surface. These maxima correspond to three basic phase transitions respectively (from high to low temperature): (i) hydrophobic core (H-core) formation, (ii) adsorption, and (iii) "flattening" of an adsorbed structure. From the comparison with the free space case, it is obvious that the two transitions in the low temperature regime are due to the influence of the attractive substrate. The individual transitions are explained in the following subsections.

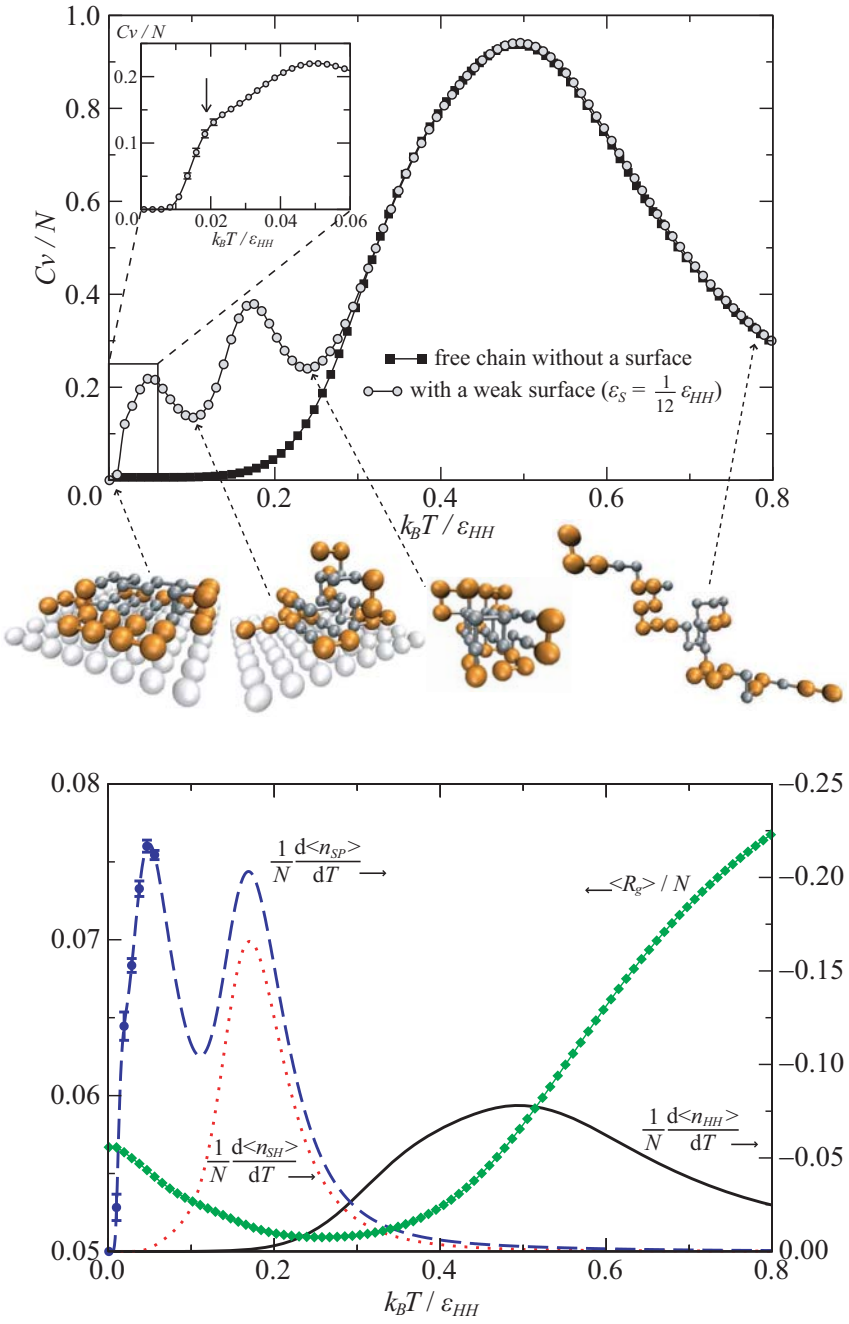


Figure 5.7: **Upper panel:** The specific heat of sequence 2D36 interacting with a weakly attractive surface ($\varepsilon_S = \frac{1}{12}\varepsilon_{HH}$, $h_w = 37$) and without the presence of the surface. Typical configurations are shown for several different temperatures. **Lower panel:** The radius of gyration per monomer, and the thermal derivatives of the numbers of H-H contacts as well as surface contacts. Horizontal arrows besides the labels indicate the scales that the quantities are using. For both graphs, error bars smaller than the data points are not shown.

5.4.1 Hydrophobic core formation / collapse transition

Referring to Figure 5.7, the specific heat shows the largest peak at the effective temperature $k_BT/\varepsilon_{HH} \approx 0.5$. This is the same coil-globule transition as the one found in free space, which is clearly demonstrated by the overlap of the specific heat from the two cases. During this stage, the HP polymer transforms from an extended chain-like structure to a compact, but desorbed, globule. Typical structures are much the same as the collapsed states of the free chain in the absence of the surface, as shown in the same figure. The radius of gyration is rapidly decreasing and the number of H-H contacts is rapidly increasing (which results in a peak in its thermal derivative) upon cooling. Both pieces of information support the idea that they are the very same coil-globule transformation.

5.4.2 Adsorption

The middle peak at $k_BT/\varepsilon_{HH} \approx 0.18$ signals protein adsorption, during which the compact HP globule "docks" at the surface with the hydrophobic core remaining intact. As seen in a typical configuration of this kind in Figure 5.7, the "freshly" adsorbed globule spans several layers vertically, and the total energy of the system is lowered slightly due to the contact with the surface. The thermal derivatives of the numbers of surface contacts for both H and P monomers thus show a peak at the same temperature due to the rapid increase in the surface contacts.

5.4.3 Flattening

Further decrease in temperature brings the system to the third transition at $k_BT/\varepsilon_{HH} \approx 0.05$ where the system maximizes the number of surface interactions without sacrificing an intact, energetically minimized hydrophobic core. Forming H-H contacts is immensely more energetically favorable than forming surface contacts with a large value of ε_{HH}.

The thermal derivative of the number of surface-P contacts has maxima at the same temperatures as both of the low temperature specific heat peaks. The radius of gyration also increases at low temperature because the protein is flattening out on the surface and becoming rather two-dimensional in shape.

One interesting feature is a subtle shoulder at $k_BT/\varepsilon_{HH} \approx 0.02$ in the specific heat. This shoulder is a "crossover" signal owing to a transition from the ground state to the first few excited states at low temperature. This behavior is similar to that in the case reported in the investigation of freezing and collapse of homopolymers[102,135], for which a signal is caused by the big difference in the numbers of possible configurations for the ground state and the first excited states. The difference for our case is that this excitation is an effect purely due to the existence of the surface, as the same weak shoulder is also found in $\frac{d\langle n_{SP}\rangle}{dT}$. In addition, the excitation from the ground state to the first few excited states can only be due to the decrease in one surface-P contact.

Typical states at $k_BT/\varepsilon_{HH} = 0$ and $k_BT/\varepsilon_{HH} \approx 0.02$ suggest that the difference in the numbers of available ground states and first few excited states is most probably due to the additional shape of a hydrophobic core available at $k_BT/\varepsilon_{HH} \approx 0.02$, resulting in a much larger number of first excited states available than the ground states. At $k_BT/\varepsilon_{HH} = 0$ where the polymer occupies just ground states, only a rectangular core is able to maximize the number of surface-P contacts that minimizes the energy. This gives the ground state energy $E = -241$ ($n_{HH} = 18, n_{SH} = 8, n_{SP} = 17$). An example of such a structure is shown in the leftmost configuration in Figure 5.7. At $k_BT/\varepsilon_{HH} \approx 0.02$ where first and second excited states dominate, two hydrophobic core shapes are observed: a rectangular and an "L-shape" as shown in Figure 5.8. Since these two cores have the same energetic contribution ($n_{HH} = 18$, with 8 H monomers interacting with the surface for both cases), the excitation from the ground state to the first few excited states can involve a change in the hydrophobic core shape in addition to losing a surface-P contact.

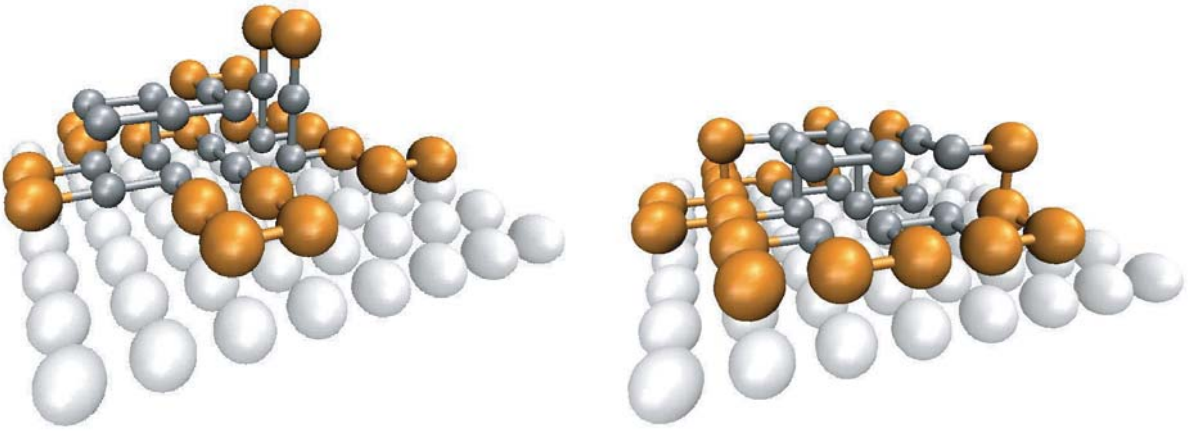


Figure 5.8: First excited states of 2D36 interacting with a very weakly attractive surface, with energy $E = -240$ ($n_{HH} = 18, n_{SH} = 8, n_{SP} = 16$).

To summarize at this point, the three transitions, namely, hydrophobic core formation, adsorption and flattening, can be identified by comparing the peak positions of the specific heat and those of the structural quantities. The thermal derivative of $\langle n_{HH} \rangle$ peaks for H-core formation, while the thermal derivatives of $\langle n_{SH} \rangle$ and $\langle n_{SP} \rangle$ peak for adsorption at a higher temperature, and flattening at a lower temperature due to the fact that the flattening process has to take place after the protein is adsorbed on cooling. In some cases, it is also possible that flattening is signaled by a shoulder, or only a peak in either $\frac{d\langle n_{SH} \rangle}{dT}$ or $\frac{d\langle n_{SP} \rangle}{dT}$ but not both.

5.5 Structural transitions in the vicinity of a strongly adsorbing surface

If the surface is strongly attractive, the transition behavior is quite different from the previous case as shown in Figure 5.9. For the surface strength we have chosen to illustrate here ($\varepsilon_S = 2\varepsilon_{HH}$, $h_w = 37$), only two peaks with a weak bump in between are seen in the specific

heat. The peak at $k_BT/\varepsilon_{HH} \approx 3.9$ indicates an adsorption transition, and the peak at $k_BT/\varepsilon_{HH} \approx 0.35$ corresponds to the H-core formation that takes place completely on the surface. The bump at $k_BT/\varepsilon_{HH} \approx 2.0$ then signals the flattening stage.

Understanding this in conjunction with the snapshots of the typical states, we see that at high temperature where the specific heat shows a peak, the extended HP chain first adsorbs on the surface until a significant number of monomers touch it. This is in agreement with the peaks in the derivatives of the numbers of surface contacts at the same temperature.

In the temperature range $k_BT/\varepsilon_{HH} \approx 3$ down to $k_BT/\varepsilon_{HH} \approx 1$, the partly adsorbed chain flattens itself out until all monomers come into contact with the surface. Again, the specific heat and the derivatives of the number of surface contacts echo each other by revealing convex bumps at the same temperature. The radius of gyration increases rapidly in this temperature range as the structure becomes more and more planar. However, the chain remains extended without forming many H-H contacts during this stage, and no signal is found from the derivative of the number of H-H contacts. As a third step the fully adsorbed, yet expanded, two-dimensional chain undergoes a collapse transition at $k_BT/\varepsilon_{HH} \approx 0.35$ to maximize the number of H-H contacts before adopting a film-like, compact structure. This is clearly signaled by the sharp peaks in the specific heat and the derivative of the number of H-H contacts at that temperature.

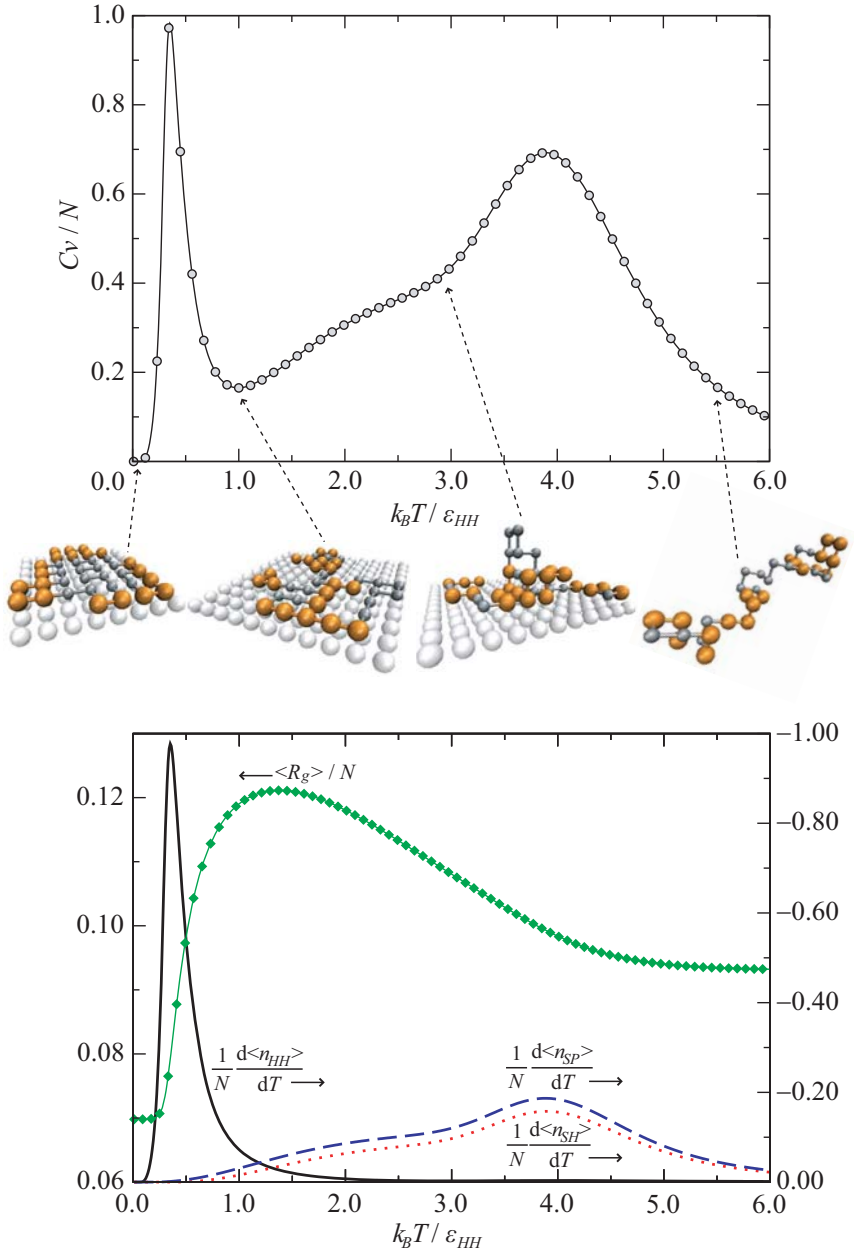


Figure 5.9: **Upper panel:** The specific heat of sequence 2D36 interacting with a strongly attractive surface ($\varepsilon_S = 2\varepsilon_{HH}$, $h_w = 37$). Typical configurations are shown for several different temperatures. **Lower panel:** Radius of gyration and thermal derivatives of the numbers of H-H contacts as well as surface contacts. Horizontal arrows besides the labels indicate the scales that the quantities are using. For both graphs, error bars are not shown as all are smaller than the data points.

5.5.1 Entropic effects of surface separation in a simulation

There are also cases where the flattening process combines with the adsorption process. In this case no signal is observed between adsorption and H-core formation. An example is shown in Figure 5.10, where a slightly weaker surface attractive field is used (it is still a strongly attractive surface but not as strong as the previous one, with $\varepsilon_S = \varepsilon_{HH}$, $h_w = 37$).

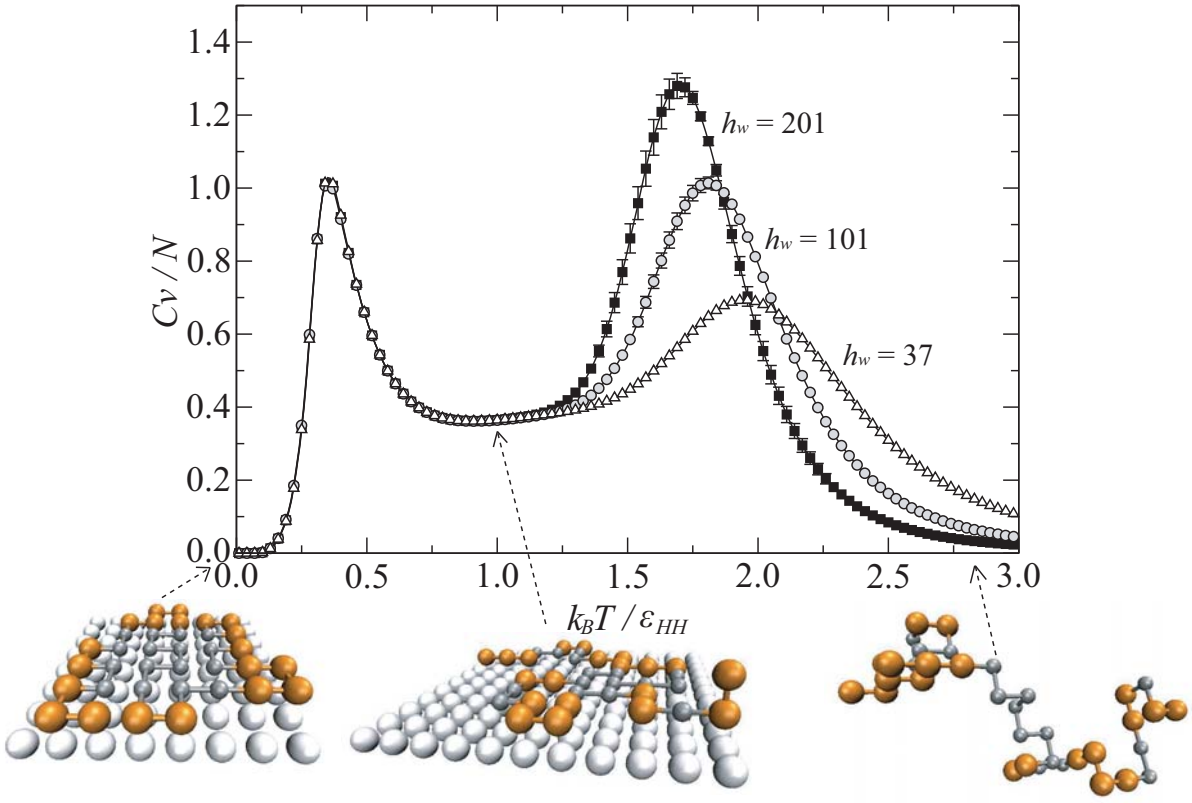


Figure 5.10: The specific heat of 2D36 interacting with a strongly attractive surface ($\varepsilon_S = \varepsilon_{HH}$, $h_w = 37$). Error bars smaller than the data points are not shown.

Figure 5.10 also shows the influence of the entropic effects on the transition behavior due to the height h_w of the non-interacting wall bounding the chain from above. When the two surfaces are made further apart, i.e., h_w becomes larger, the system can take on more possible configurations due to the additional vertical translational degree of freedom. The entropy for a macrostate, $S = -k_B \log \Omega$ (where Ω is the thermodynamic probability, which can be interpreted as the number of accessible configurations for that macrostate),

thus increases with h_w. This change in entropy as a result of a change in the number of possible configurations is termed the entropic effect.

The desorption-adsorption transition thus shows a systematic dependence on h_w. A smaller h_w restricts the vertical movements of the HP chain to a larger extent, resulting in a smaller entropy gain (and thus a less pronounced peak in the specific heat) as less translational variations of the same configuration are allowed. The chain is also more likely to be on the attractive surface, resulting in a higher adsorption temperature. This dependence of the adsorption transition peak on h_w was also reported recently[136], where an off-lattice homopolymer model is used to study polymer adsorption.

The low temperature collapse, however, is obviously not affected at all, as it takes place on the attractive surface regardless of where the steric upper surface is placed.

5.6 Effect of surface attraction on the structural transitions

As shown in the few examples here, structural transitions may occur in different orders for different surface attractive strengths when temperature changes. Some transitions may not give distinct signals in specific heat but merely in particular structural parameters. It is thus essential to analyze specific heat together with structural quantities to identify various "phases".

With longer HP chains, our interpretation of selected structural parameters suggests that the hierarchies of structural phase transitions for this model can be generalized into a few categories, for which the occurrence can be closely related to the surface attractive field strength. In the following chapter, we are going to explore these in great detail.

Chapter 6

Generic Transition Hierarchies

In the last chapter, we have illustrated that it is crucial to identify structural "phase transitions" by analyzing the thermodynamics of some structural quantities in addition to the specific heat using a short sequence 2D36 as an example. Three major structural changes, H-core formation, adsorption, and "flattening", are observed. In this chapter, we shall first demonstrate, by sequence 2D36 again, that these structural transitions occur at different temperatures when the surface attractive field strength varies. The transitions arrange themselves in different orders of occurrence, which lay the foundation of the hierarchies of the structural phase transitions and has been published in Ref. [137]. To build a connection to some existing work, we will also distinguish a few basic structural "phases" which were identified previously[108,112,113]. They are, namely, adsorbed compact (AC), adsorbed globular (AG), adsorbed expanded (AE), desorbed compact (DC), and desorbed expanded (DE) phases. Note that as these phases were defined primarily based on the transitions identified by the specific heat, they are not perfect classifications for the phases found in our study. It is because we have defined transitions using structural quantities as well. As a result, two structures before and after a transition in our study might belong to the same phase according to the existing definitions of phases.

Next, we shall show in deeper detail, with evidence provided by some longer benchmark sequences, that the same hierarchical changes could also be observed by varying the surface attraction. That leads to the idea of grouping similar transition hierarchies into a single category. By doing so, a few general categories can be identified, and their relations to the surface attractions have emerged. These results have been published in Ref. [134].

6.1 Identification of transition hierarchies

Recalling the order of occurrence of the structural transitions in Section 5.4 where the 2D36 interacts with a very weakly attractive surface, H-core formation takes place at the highest temperature, adsorption follows, and flattening occurs at the lowest temperature (see Figure 6.1). This is referred to as Category IV in our following discussions.

When the surface attraction becomes stronger, the three basic transitions occur in a very narrow temperature window so that only one peak in the specific heat is observed as seen in Figure 6.2. In the example we are showing here, these three transitions can be identified clearly by the structural parameters and show a sequence of adsorption, H-core formation and flattening with the decrease in temperature. However, we stress that this order could be different depending on the surface attraction. If the surface attraction is weaker, the order of transitions would be closer to Category IV. Similarly, with the surface attraction a little stronger, the order of transitions would be closer to Category II as will be discussed in the following. As long as these transitions occur at almost the same temperature and give only one single peak in the specific heat, we group these types of orders of transitions together as Category III.

By making the surface attraction even stronger, the transitions begin to be separately distinguishable again. Figure 6.3 shows the thermodynamics for 2D36 interacting with a surface of moderate attractive strength ($\varepsilon_S = \frac{1}{2}\varepsilon_{HH}$). Two peaks are present in the specific

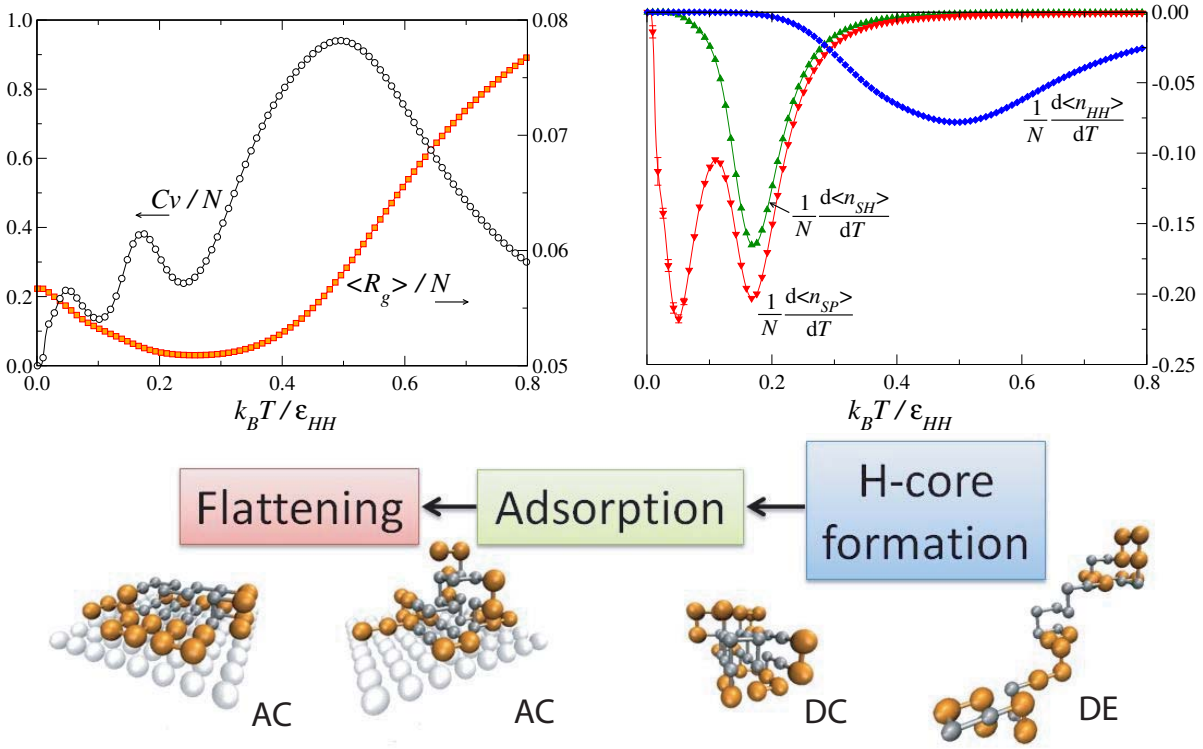


Figure 6.1: Specific heat and structural quantities of 2D36 interacting with a very weak attractive surface ($\varepsilon_S = \frac{1}{12}\varepsilon_{HH}$). Typical conformations with their classified phases are also shown. This is a typical example for the Category IV transition hierarchy.

heat, with adsorption taking place at $k_BT/\varepsilon_{HH} \approx 1.0$, a hydrophobic core forms at a slightly higher temperature ($k_BT/\varepsilon_{HH} \approx 0.35$) than that of flattening ($k_BT/\varepsilon_{HH} \approx 0.27$). The last two processes are still indistinguishable by the C_V peak at the lower temperature. In terms of the transition sequence, H-core formation and adsorption have swapped places compared to Category IV. On cooling, a three-dimensional, adsorbed but extended structure is first formed after adsorption which takes place at the highest temperature. The lowest energy state with a two-dimensional hydrophobic core is achieved after the combined action of H-core formation and flattening. As these two processes closely overlap, no intermediate states could be singled out between them.

Further increase in surface attractive strength shifts the H-core formation to an even lower

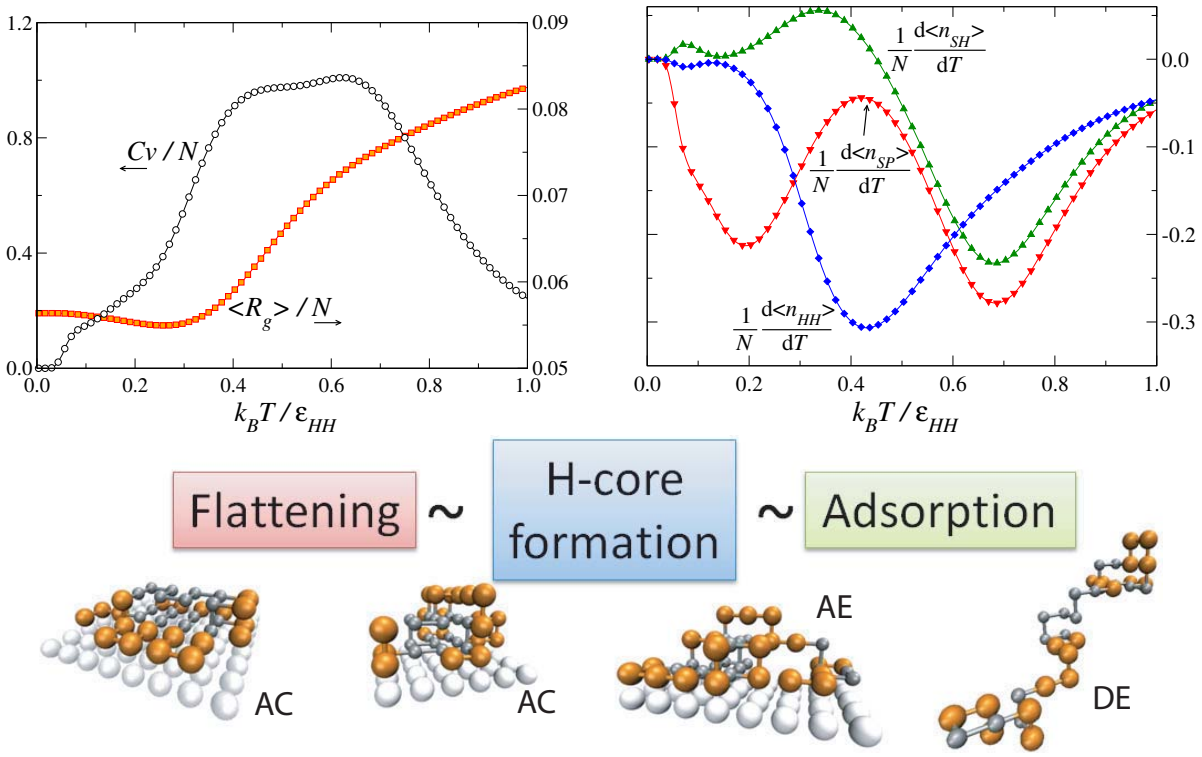


Figure 6.2: Specific heat and structural quantities of 2D36 interacting with a weak attractive surface ($\varepsilon_S = \frac{1}{3}\varepsilon_{HH}$). Typical conformations with their classified phases are also shown. This is a typical example for the Category III transition hierarchy.

temperature as shown in Figure 6.4, where a strong attractive surface is used ($\varepsilon_S = 2\varepsilon_{HH}$). In this case, there are also two peaks in the specific heat with a weak bump in between. A comparison with the structural properties clearly distinguishes the three basic transitions, which now occur at well separated temperatures. Adsorption takes place at $k_BT/\varepsilon_{HH} \approx 4.0$; H-core formation at $k_BT/\varepsilon_{HH} \approx 0.4$; the bump occurring between $k_BT/\varepsilon_{HH} \approx 1.0$ and $k_BT/\varepsilon_{HH} \approx 3.0$ is a signal of flattening.

With this transition ordering, the desorbed, extended protein first adsorbs on the surface to form a three-dimensional, adsorbed yet extended structure. After the flattening process, most of the monomers contact with the surface but the chain is still not compact. The H-core formation finally takes place on the surface, forming a two-dimensional ground state

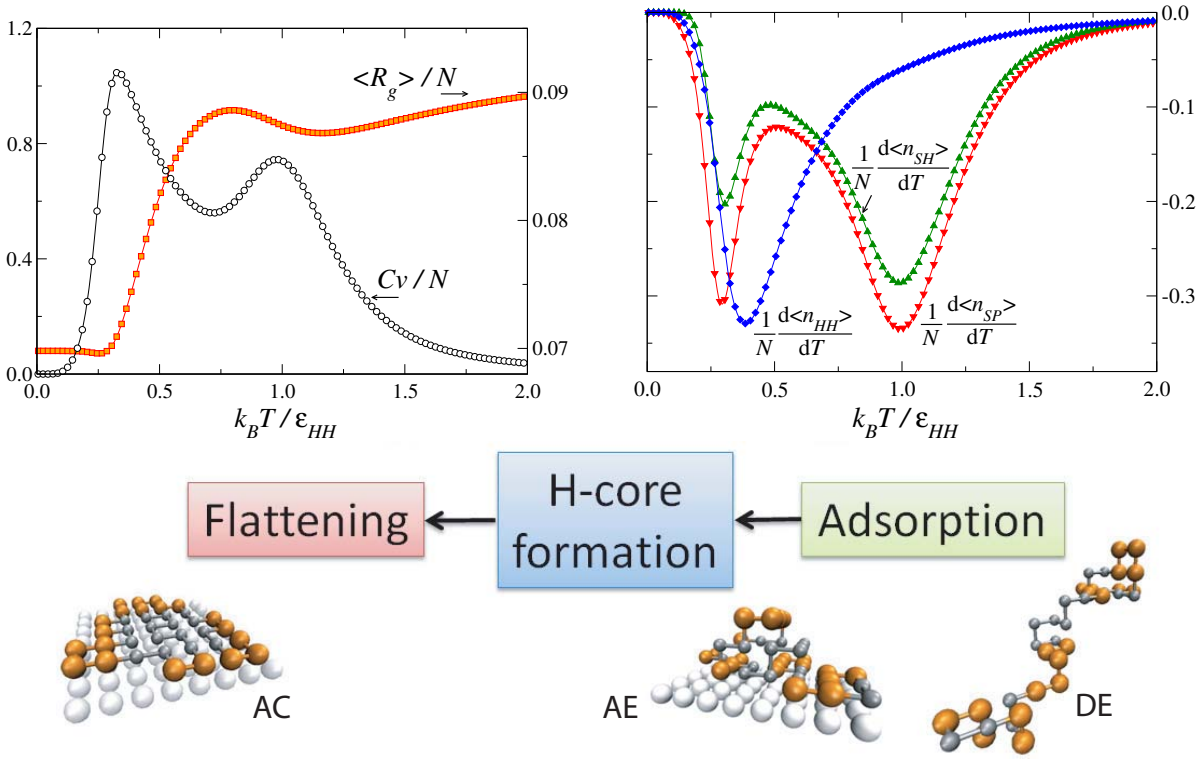


Figure 6.3: Specific heat and structural quantities of 2D36 interacting with a moderately attractive surface ($\varepsilon_S = \frac{1}{2}\varepsilon_{HH}$). Typical conformations with their classified phases are also shown. This is a typical example for the Category II transition hierarchy.

with a hydrophobic core. Comparing with Category II, H-core formation and flattening have swapped places in this case, which we regard as the Category I transition.

As we have seen in these examples from 2D36, the three basic transitions, H-core formation, adsorption, and "flattening", occur at different temperatures when the surface attractive strength varies, giving rise to a different order in structural changes. As a consequence, an extended, desorbed protein goes through a different path in conformational space towards the acquisition of compact, adsorbed ground states. Structures of the intermediate and ground states thus vary from case to case and are completely dependent on this order (or hierarchy) of transitions.

The four major transition categories in the previous examples were identified from a

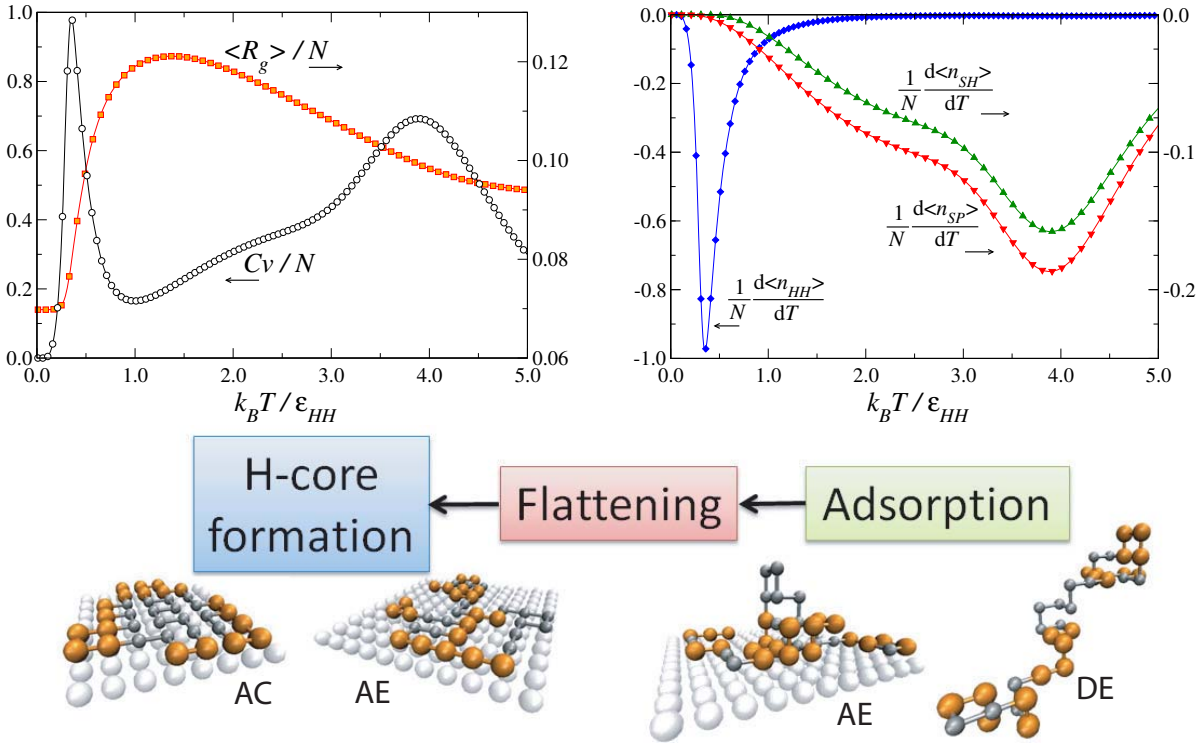


Figure 6.4: Specific heat and structural quantities of 2D36 interacting with a strong attractive surface, in which $\varepsilon_S = 2\varepsilon_{HH}$. Typical conformations with their classified phases are also shown. This is a typical example for the Category I transition hierarchy.

combination of the specific heat, average radius of gyration $\langle R_g \rangle$, and the thermal derivative of the average number of H-H interactions, $d\langle n_{HH} \rangle / dT$, and those of the numbers of surface contacts, $d\langle n_{SH} \rangle / dT$ and $d\langle n_{SP} \rangle / dT$. Nevertheless, since all these changes in structural parameters are interrelated to each other, the classification of transition categories can be simplified as considering only the combined patterns of C_V and $\langle R_g \rangle$, where the order of occurrence of different folding process is "encoded":

Category I C_V shows two peaks, a bump between the peaks might be possible, $\langle R_g \rangle$ shows a maximum between these two peaks

Category II C_V shows two peaks, $\langle R_g \rangle$ decreases upon cooling. In the very low tempera-

ture regime, it might rise back up a little to form a minimum when the temperature approaches zero

Category III C_V shows only one peak with possible shoulders, $\langle R_g \rangle$ decreases on cooling

Category IV C_V shows three distinct peaks, $\langle R_g \rangle$ decreases upon cooling. In the very low temperature regime, it might rise back up a little to form a minimum when the temperature approaches zero

Nevertheless, this simplified classification scheme should only be used as a coarse outline when details are not of importance. Later in Section 6.7, we shall show that this classification outline is not able to pick up a subtle crossover between two categories.

We also observe that $\langle R_{ee} \rangle$ behaves quite similarly as $\langle R_g \rangle$ but is less reliable at low temperature where compact structures are mainly found. For this reason our analysis relies on $\langle R_g \rangle$ rather than on $\langle R_{ee} \rangle$.

6.2 Comprehensive analysis of longer HP sequences

So far, the transition hierarchies are identified by a comprehensive analysis of different structural parameters of the systems, as illustrated by examples of a rather short sequence, 2D36. We extended the analysis to some longer benchmark HP sequences interacting with different surface attractions. Specifically, three other HP sequences, 3D48, 3D67 and 3D103, were first simulated and their transition categories were determined (see Table 6.1). The thermodynamic properties will be reported in Sections 6.3 − 6.5. Based on the category information we obtained from these three sequences, we have set up a classification scheme for the transition categories related to the surface attractive strength and the internal H-H interaction strength, namely, $(\varepsilon_{SH} + \varepsilon_{SP})/\varepsilon_{HH}$.

The second data analysis stage involves the "validation" of the inferred rule set forth by

the three aforementioned sequences. We have simulated two other benchmark sequences, 3D48.1 and 2D64 (folding in a three-dimensional space), categorized their transition behaviors and saw if the results match with the prediction. These results will be presented in Section 6.8.

Note that only Categories I to III are realized from the systems in Table 6.1. It is because the numbers of energy bins become larger and larger when the surface becomes weaker and weaker, and the computation time grows with the number of energy bins. It has become very computationally inefficient to simulate a long sequence with a weakly attractive surface.

surface	ε_{HH}	ε_{SH}	ε_{SP}	3D48		3D67		3D103	
Free space without surface:									
2D	1	/	/	-21		-29		-32	
3D	1	/	/	-34		-56		-58	
Surfaces attract all monomers:									
A1	1	1	1	-69	(I)	-96	(I)	-135	(I)
A2	1	2	2	-117	(I)	-163	(I)	/	
A$1/2$	2	1	1	-93	(II)	-132	(II)	-167	(I/II)
Surfaces attract only H monomers:									
H1	1	1	0	-49	(II)	-72	(II)	-80	(II)
H2	1	2	0	-73	(I)	-108	(I)	/	
H$1/2$	2	1	0	-79	(III)	-118	(II)	-128	(III)
Surfaces attract only P monomers:									
P1	1	0	1	-48	(II)	-69	(II)	-100	(I/II)
P2	1	0	2	-71	(I)	-91	(I)	/	
P$1/2$	2	0	1	-79	(III)	-123	(II)	-150	(II)

Table 6.1: Systems simulated using the sequences 3D48, 3D67 and 3D103. Different attractive surface types and strengths are abbreviated in the surface labels (A, H or P stand for the surface types, the numbers stand for the ratio between ε_{SH} or ε_{SP} and ε_{HH}). The lowest energy found during the estimation of $g(E)$ for each system is reported, with the Roman number in the parentheses denoting the classification of transition categories.

6.3 Category I: folding behavior with a strongly attractive surface

Figure 6.5 shows a typical transition of Category I demonstrated by 3D67 interacting with surface A2, for which $\varepsilon_S = 2\varepsilon_{HH}$. It is characterized by two pronounced peaks in C_V, with $\langle R_g \rangle$ attaining its maximum between them as seen in the upper panel of the figure. The nature of transitions to which the two peaks in C_V correspond are identified by comparing them with $d\langle n_{HH} \rangle / dT$ and $d\langle n_{SH} \rangle / dT$ in the lower panel. Since the surface attracts both types of monomers equally, $d\langle n_{SP} \rangle / dT$ shows similar behavior as $d\langle n_{SH} \rangle / dT$ and thus is not shown in the figure. The peak at $k_BT/\varepsilon_{HH} \approx 2.2$ in C_V represents the desorption-adsorption transition where $d\langle n_{SH} \rangle / dT$ peaks at the same temperature. The peak at $k_BT/\varepsilon_{HH} \approx 0.2$ represents the H-core formation as $d\langle n_{HH} \rangle / dT$ also shows a peak at that position. $\langle R_g \rangle$ decreases most rapidly during this process when temperature is lowered.

A closer look at C_V in Figure 6.5 shows a weak bump between $k_BT/\varepsilon_{HH} \approx 0.5$ and 1.75. The same phenomenon is also observed in $d\langle n_{SH} \rangle / dT$ (and $d\langle n_{SP} \rangle / dT$), suggesting that this is a region where the HP chain keeps forming contacts with the surface until it adsorbs completely on the surface, which is the "flattening" of the structure. It is also evidenced by the continuous increase in the average radius of gyration $\langle R_g \rangle$ with the decrease of temperature. When the surface attraction is sufficiently strong, "flattening" occurs right after the chain is adsorbed to the surface but before the H-core formation. A similar case is shown by the 3D48 interacting with the A2 surface ($\varepsilon_S = 2\varepsilon_{HH}$) as shown in Figure 6.6, in which a bump for "flattening" is also observed in C_V between the adsorption and the H-core formation peak.

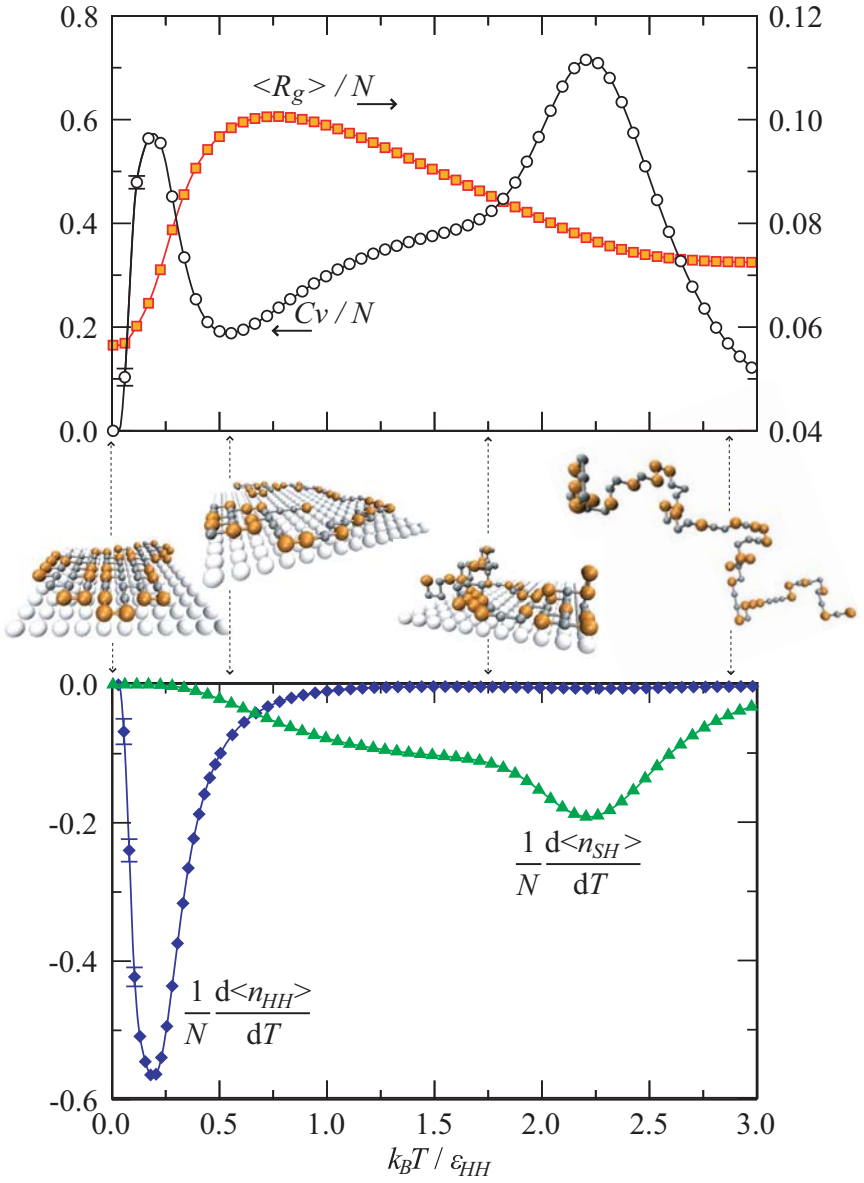


Figure 6.5: Thermodynamics of 3D67 interacting with surface A2 ($\varepsilon_S = 2\varepsilon_{HH}$), which shows a typical Category I transition.
Upper panel: Specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH}. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Middle panel: Typical configurations at different temperatures.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH} \rangle /dT$, and that of the average number of surface contacts of H monomers per monomer, $(1/N)d\langle n_{SH} \rangle /dT$, as a function of k_BT/ε_{HH}.

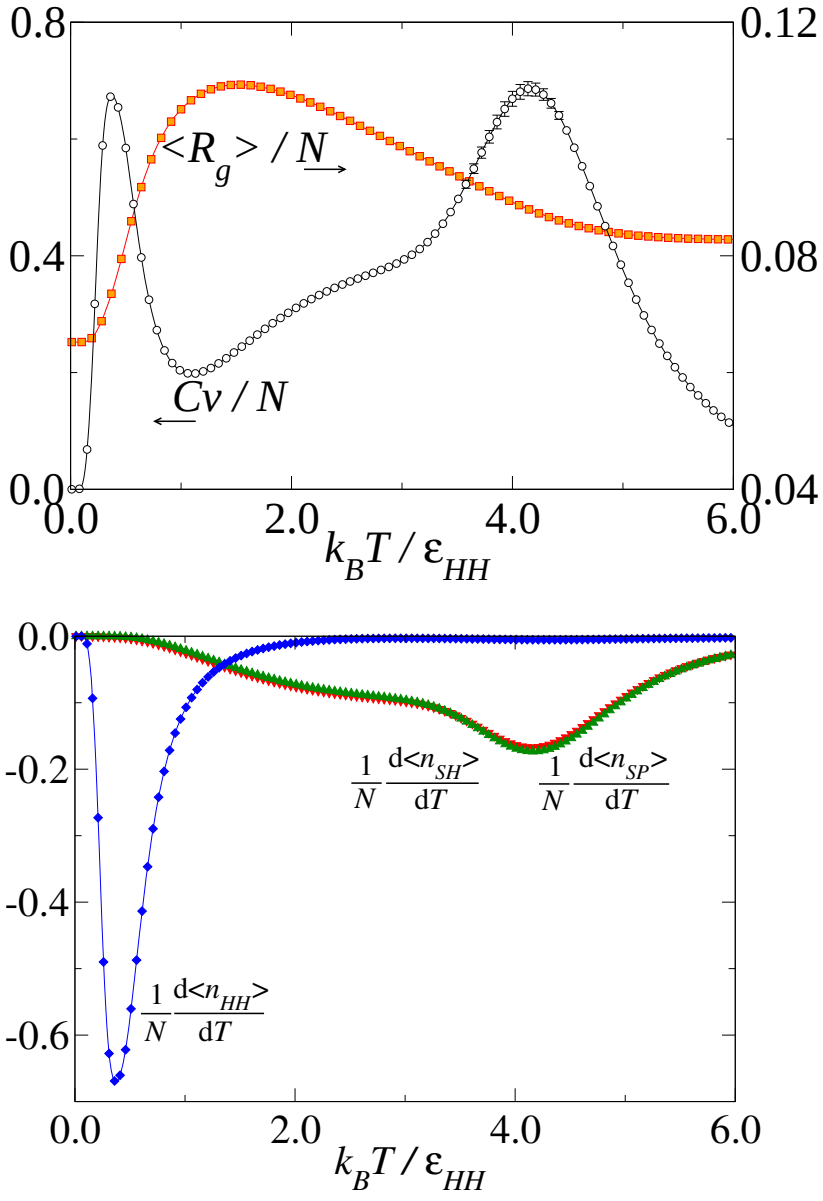


Figure 6.6: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D48 interacting with surface A2 ($\varepsilon_S = 2\varepsilon_{HH}$), another example of a Category I transition in which a flattening bump is present. The horizontal arrows beside the labels indicate the axes to which the quantities refer.

Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH}\rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH}\rangle /dT$ and $(1/N)d\langle n_{SP}\rangle /dT$, as a function of k_BT/ε_{HH}.

However, there are cases where this "flattening" bump is not observed in C_V, as seen from two other examples for 3D67 (Figure 6.7) and 3D103 (Figure 6.8) with different surface attractions. The flattening process might have been "integrated" within adsorption, or it simply does not cause energy fluctuations to give an obvious signal in C_V. In the latter case, signals can be found in other structural quantities like $d\langle n_{SH}\rangle/dT$ or $d\langle n_{SP}\rangle/dT$.

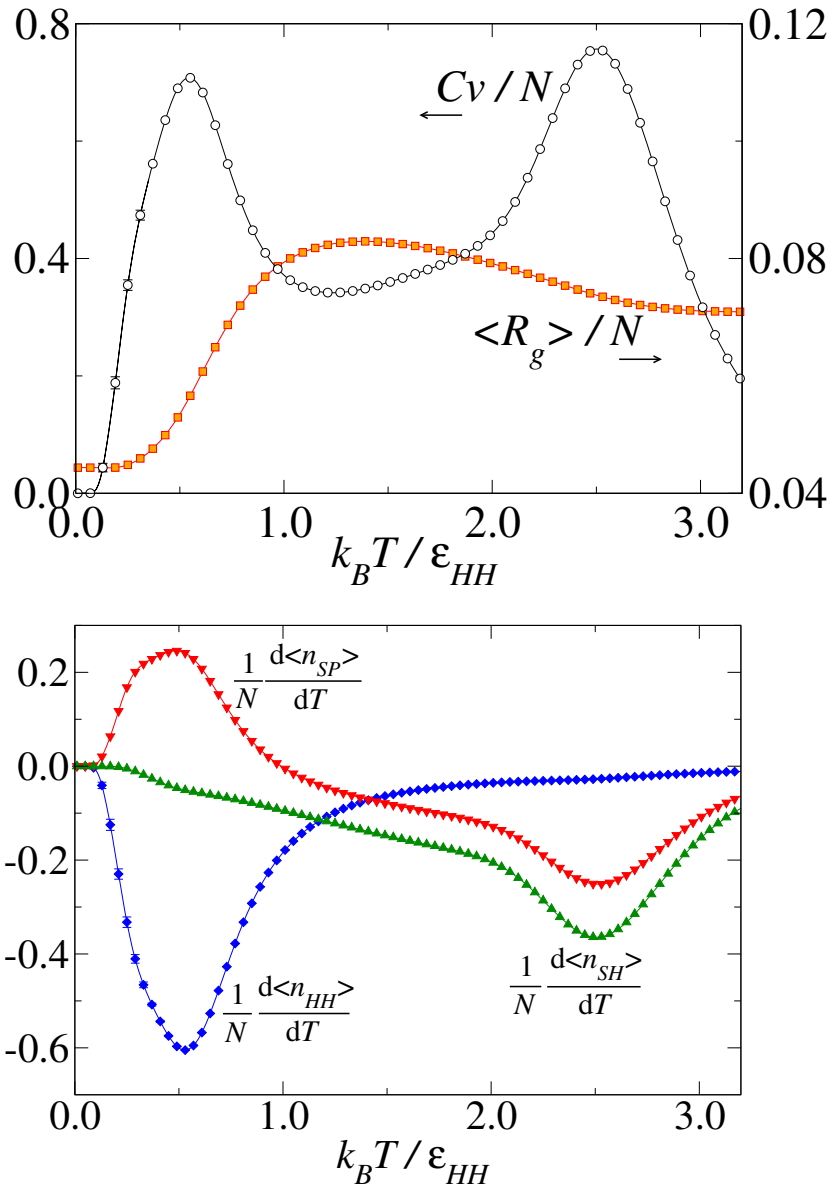


Figure 6.7: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D67 interacting with surface H2 ($\varepsilon_{SH} = 2\varepsilon_{HH}, \varepsilon_{SP} = 0$), an example of a Category I transition in which a "flattening" bump is not observed in C_V. The horizontal arrows beside the labels indicate the axes to which the quantities refer.

Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH} \rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH} \rangle /dT$ and $(1/N)d\langle n_{SP} \rangle /dT$, as a function of k_BT/ε_{HH}.

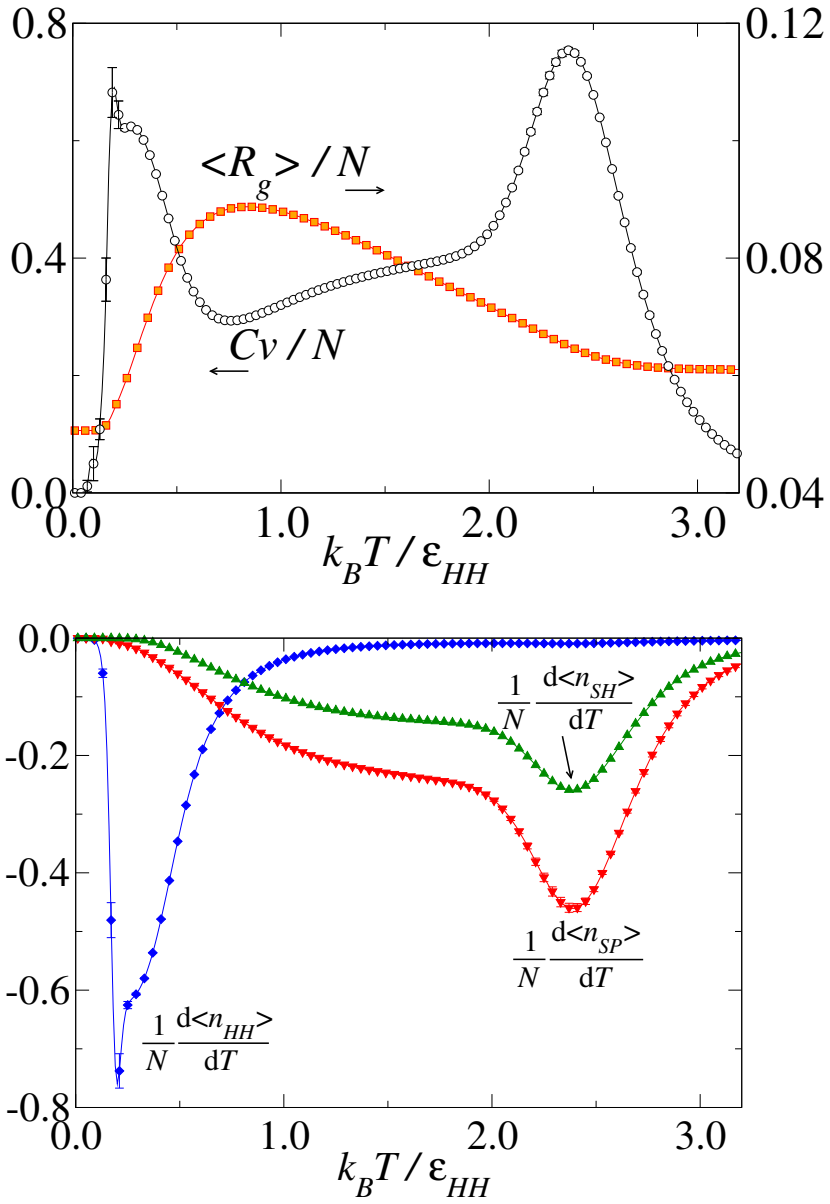


Figure 6.8: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D103 interacting with surface A1 ($\varepsilon_S = \varepsilon_{HH}$), another example of a Category I transition in which a "flattening" bump is not observed in C_V. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH}\rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH}\rangle /dT$ and $(1/N)d\langle n_{SP}\rangle /dT$, as a function of k_BT/ε_{HH}.

6.4 Category II: folding behavior with a moderately attractive surface

Figure 6.9 shows the thermodynamics for the 3D103 with surface P1/2, a typical case in Category II. Similar to Category I, systems in Category II also show two pronounced peaks in C_V and identification of structural transitions depends on the thermal derivatives of $\langle n_{HH} \rangle$, $\langle n_{SH} \rangle$ and $\langle n_{SP} \rangle$. The peak at $k_B T/\varepsilon_{HH} \approx 0.85$ represents the desorption-adsorption transition as identified by the peaks in $d\langle n_{SH} \rangle / dT$ and $d\langle n_{SP} \rangle / dT$. Another peak at $k_B T/\varepsilon_{HH} \approx 0.42$ indicates the H-core formation as signaled by a peak in $d\langle n_{HH} \rangle / dT$.

Interesting observations at low temperature are revealed by the thermodynamics of $d\langle n_{HH} \rangle / dT$, $d\langle n_{SH} \rangle / dT$ and $d\langle n_{SP} \rangle / dT$ as shown in the lower panel of Figure 6.9. During the H-core formation at $k_B T/\varepsilon_{HH} \approx 0.42$ where $d\langle n_{HH} \rangle / dT$ peaks, troughs are observed in $d\langle n_{SH} \rangle / dT$ and $d\langle n_{SP} \rangle / dT$. This is a process of "thickening", during which some of the surface attachments have to be broken to facilitate the construction of H-H interactions.

When the temperature is further lowered to $k_B T/\varepsilon_{HH} \approx 0.25$, a subtle shoulder could barely be seen in C_V and $\langle R_g \rangle$ stays still on cooling; $d\langle n_{SP} \rangle / dT$, however, shows a clear peak. This suggests that surface contacts for the P monomers are established, demonstrating the flattening effect. Eventually the structures with minimal possible energy are attained but they no longer span as many layers vertically as at higher temperature. These structures are not completely planar as in Category I, as forming surface contacts is not always more energetically favorable than forming hydrophobic interactions.

Another major feature that differentiates Category II from Category I is the absence of a maximum for $\langle R_g \rangle$ between the two peaks in C_V. It decreases upon cooling until the very low temperature regime. The difference arises from the fact that the flattening of structures occurs at a lower temperature than the H-core formation in the vicinity of a less attractive surface, giving rise to another transition hierarchy than Category I.

Two possibilities for $\langle R_g \rangle$ are then observed when the temperature is further lowered: (a) it keeps descending as in Figure 6.9; (b) it rises back up until $T = 0$, forming a minimum below the H-core formation temperature as the 3D48 does in Figure 6.10.

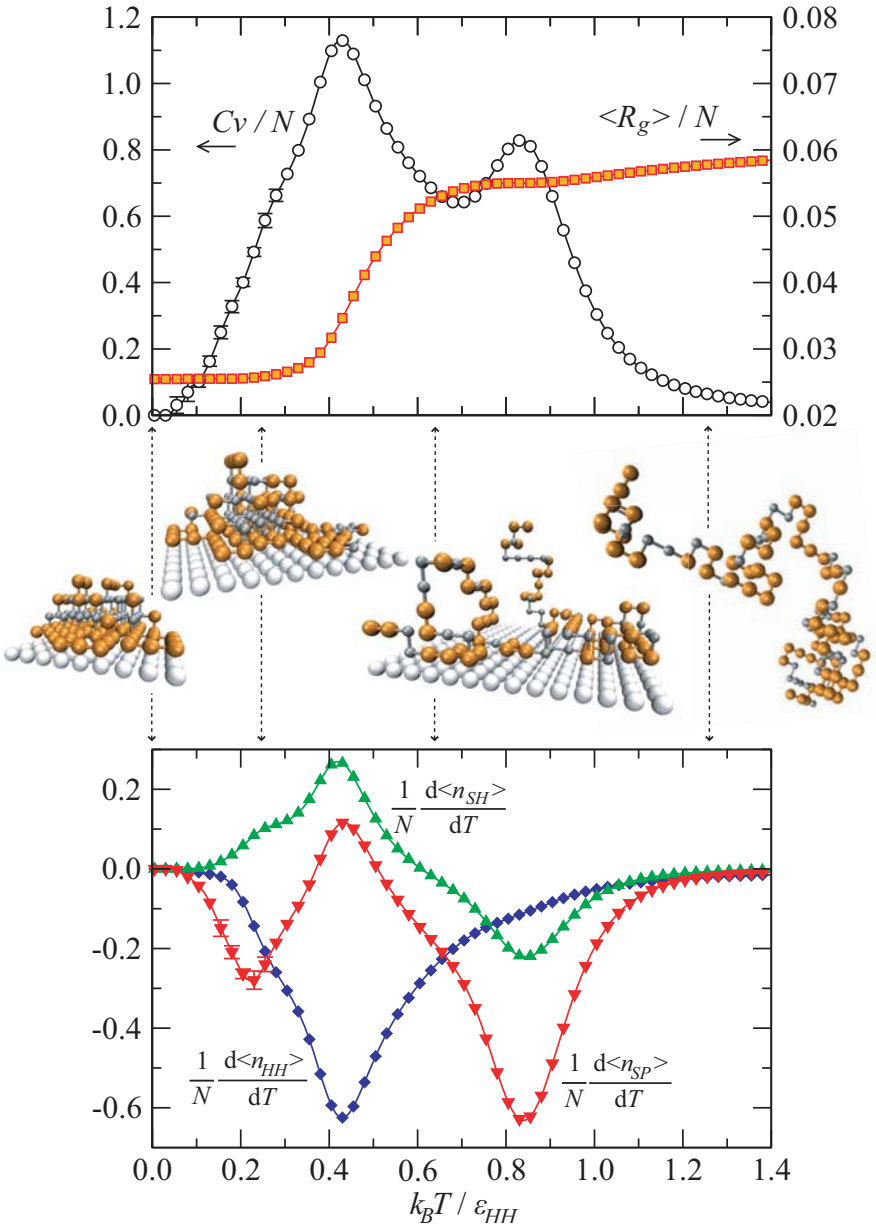


Figure 6.9: Thermodynamics of the 3D103 interacting with surface P1/2 ($\varepsilon_{SH} = 0, \varepsilon_{SP} = \frac{1}{2}\varepsilon_{HH}$), which shows a typical Category II transition.
Upper panel: The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH}. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Middle panel: Typical configurations at different temperatures.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH} \rangle /dT$, and those of the numbers of surface contacts, $(1/N)d\langle n_{SH} \rangle /dT$ and $(1/N)d\langle n_{SP} \rangle /dT$, as a function of k_BT/ε_{HH}, respectively.

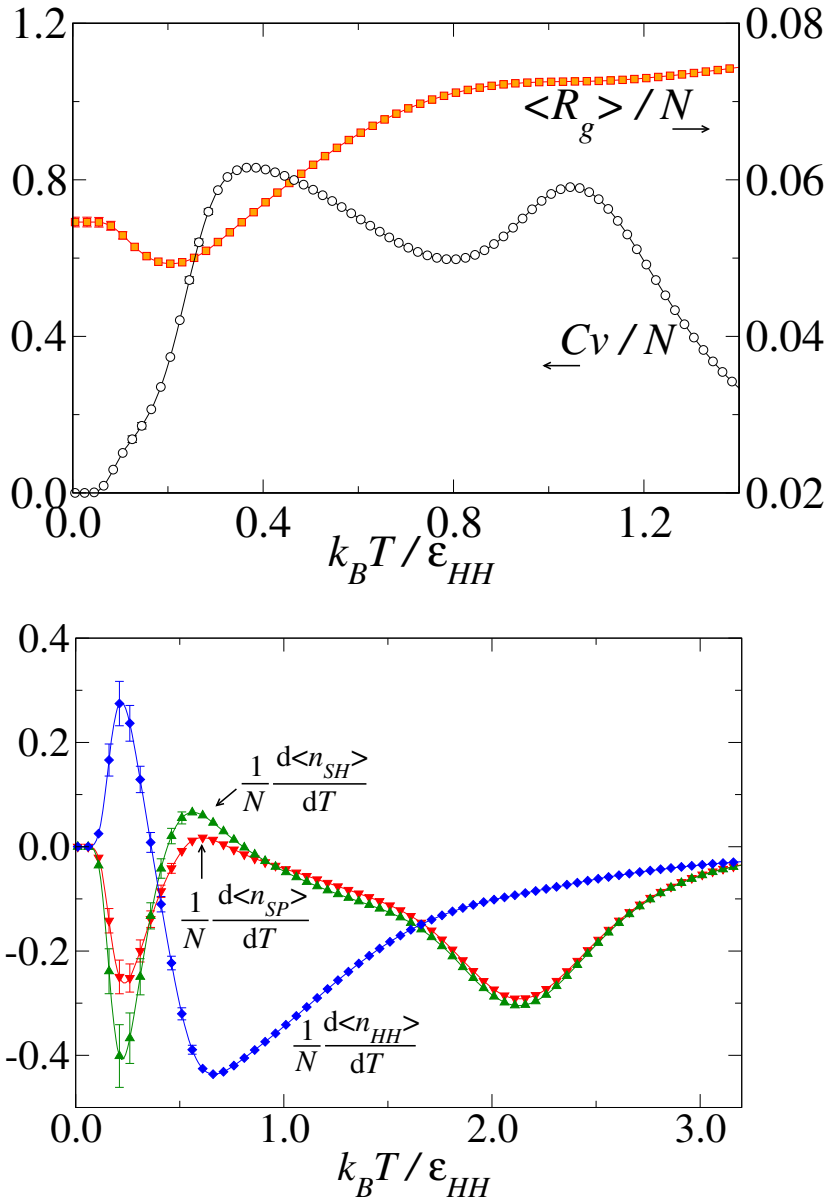


Figure 6.10: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D48 interacting with surface A1/2 ($\varepsilon_S = \frac{1}{2}\varepsilon_{HH}$), another example of a Category II transition in which a "kink" is formed in the very low temperature regime of $\langle R_g \rangle /N$ upon cooling. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH} \rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH} \rangle /dT$ and $(1/N)d\langle n_{SP} \rangle /dT$, as a function of k_BT/ε_{HH}.

In some other examples of this transition category (Figure 6.11 and Figure 6.12), C_V only has two major transition peaks, sometimes with a subtle shoulder or a spike merged into the peaks as a result of a combination of various events. Individual investigation of structural measures is thus essential to segregate different structural changes.

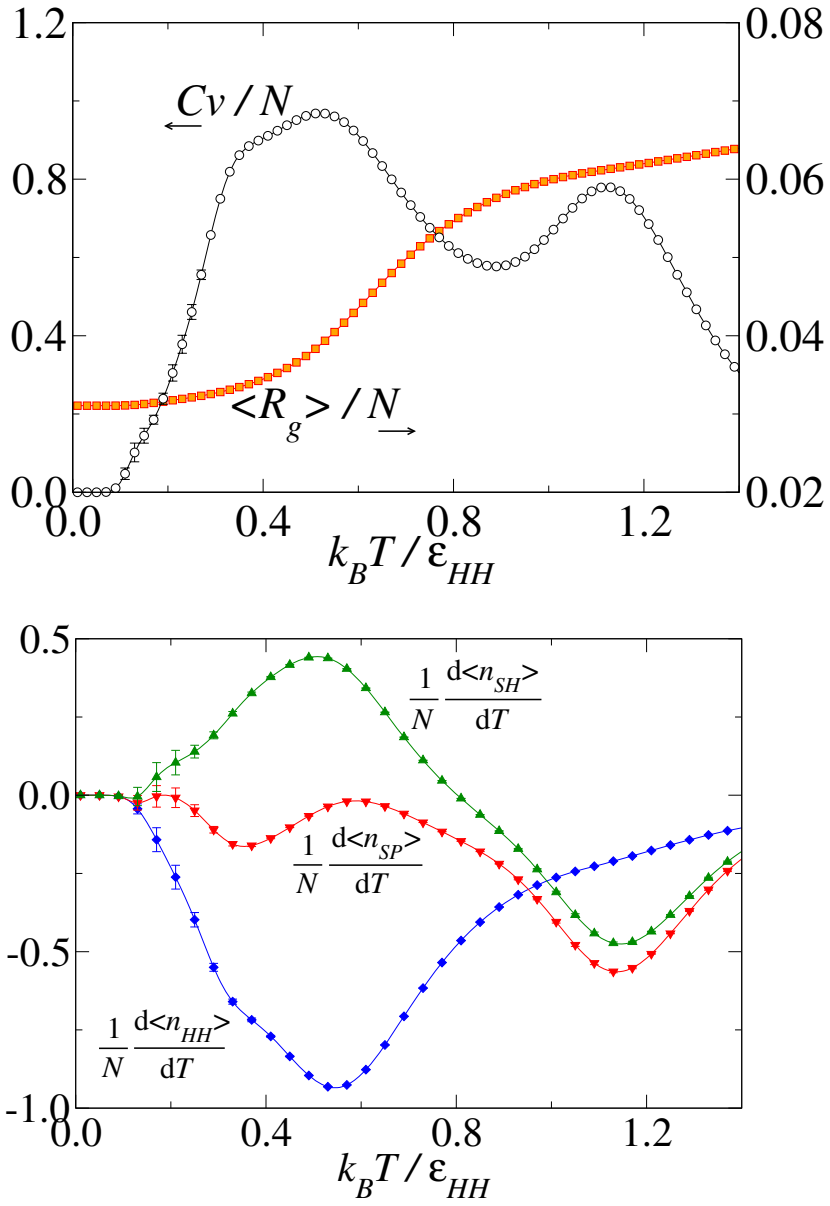


Figure 6.11: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D67 interacting with surface P1 ($\varepsilon_{SH} = 0, \varepsilon_{SP} = \varepsilon_{HH}$). This is an example of a Category II transition without a "kink" formed in the very low temperature regime of $\langle R_g \rangle /N$ upon cooling. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH} \rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH} \rangle /dT$ and $(1/N)d\langle n_{SP} \rangle /dT$, as a function of k_BT/ε_{HH}.

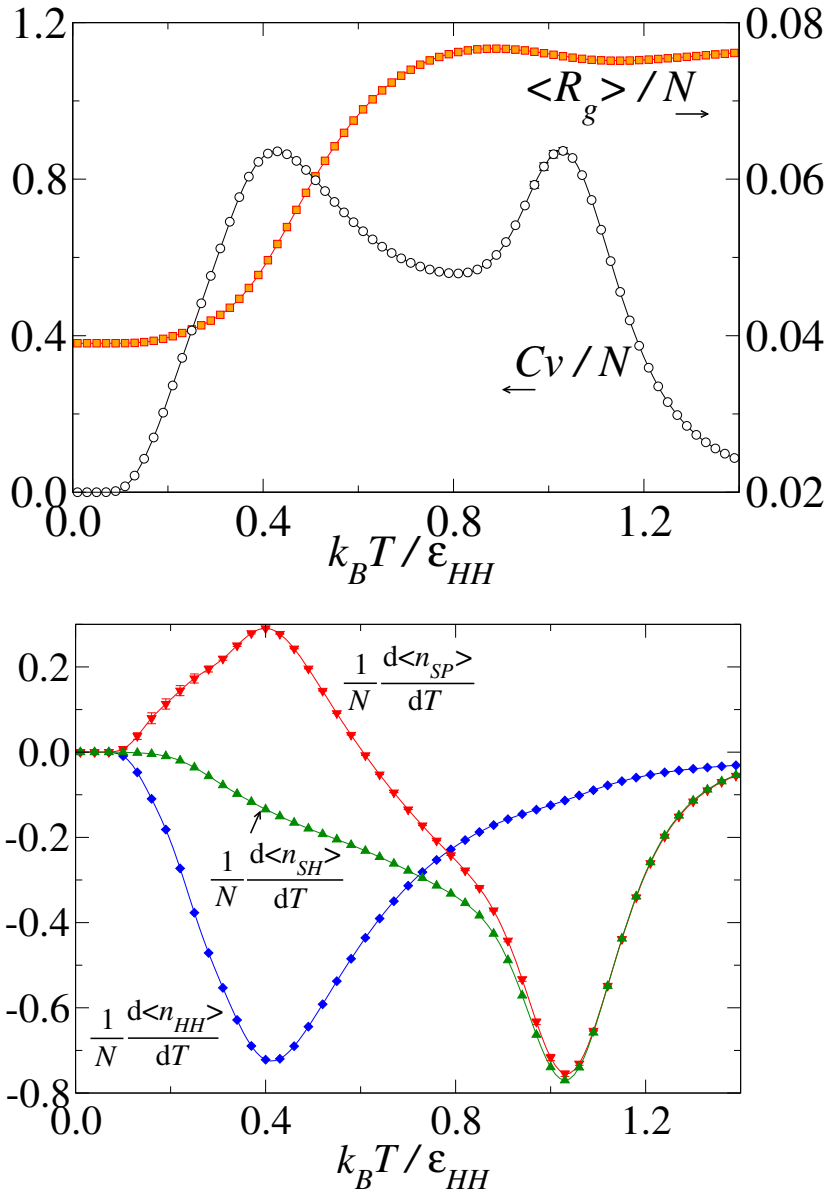


Figure 6.12: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D103 interacting with surface H1 ($\varepsilon_{SH} = \varepsilon_{HH}, \varepsilon_{SP} = 0$). This is another example of a Category II transition without a "kink" formed in the very low temperature regime of $\langle R_g \rangle /N$ upon cooling. The horizontal arrows beside the labels indicate the axes to which the quantities refer.

Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH} \rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH} \rangle /dT$ and $(1/N)d\langle n_{SP} \rangle /dT$, as a function of k_BT/ε_{HH}.

6.5 Category III: folding behavior with a weakly attractive surface

When the surface attractive strength further reduces, the adsorption and flattening temperatures decrease accordingly. Category III is identified when the adsorption transition coincides with H-core formation, giving a single peak in C_V associated with a shoulder in some cases like the example shown in Figure 6.13. The thermodynamics of Category III transitions looks similar to that in 3D free space: $\langle R_g \rangle$ decreases upon cooling and C_V peaks at nearly the same temperature in both cases, except that a higher peak results for Category III. Since adsorption and H-core formation now occur almost together at nearby temperatures, more conformational degrees of freedom are introduced by the surface interactions, this higher entropy gain results in a larger C_V.

Details of transitions are again provided by $d\langle n_{HH} \rangle / dT$, $d\langle n_{SH} \rangle / dT$ and $d\langle n_{SP} \rangle / dT$. From the positions of peaks in $d\langle n_{SH} \rangle / dT$ and $d\langle n_{SP} \rangle / dT$, one may identify adsorption at $k_B T/\varepsilon_{HH} \approx 1.25$ and flattening at $k_B T/\varepsilon_{HH} \approx 0.5$ respectively. $d\langle n_{HH} \rangle / dT$ demonstrates a wide peak across the adsorption and flattening temperatures, which suggests that the hydrophobic core is formed roughly in the temperature range $k_B T/\varepsilon_{HH} \approx 0.5 - 1.5$. Instead of producing individual peaks in C_V, the signals of adsorption and flattening are "bridged" and smoothed out by the H-core formation, giving only a peak with a shoulder in C_V. In some cases, e.g., Figure 6.14 and Figure 6.15, the shoulder may appear at different locations. It could become a spike or it could be absent. In these cases, one needs to rely on the structural quantities again to separate signals for various transitions as illustrated in the previous discussions.

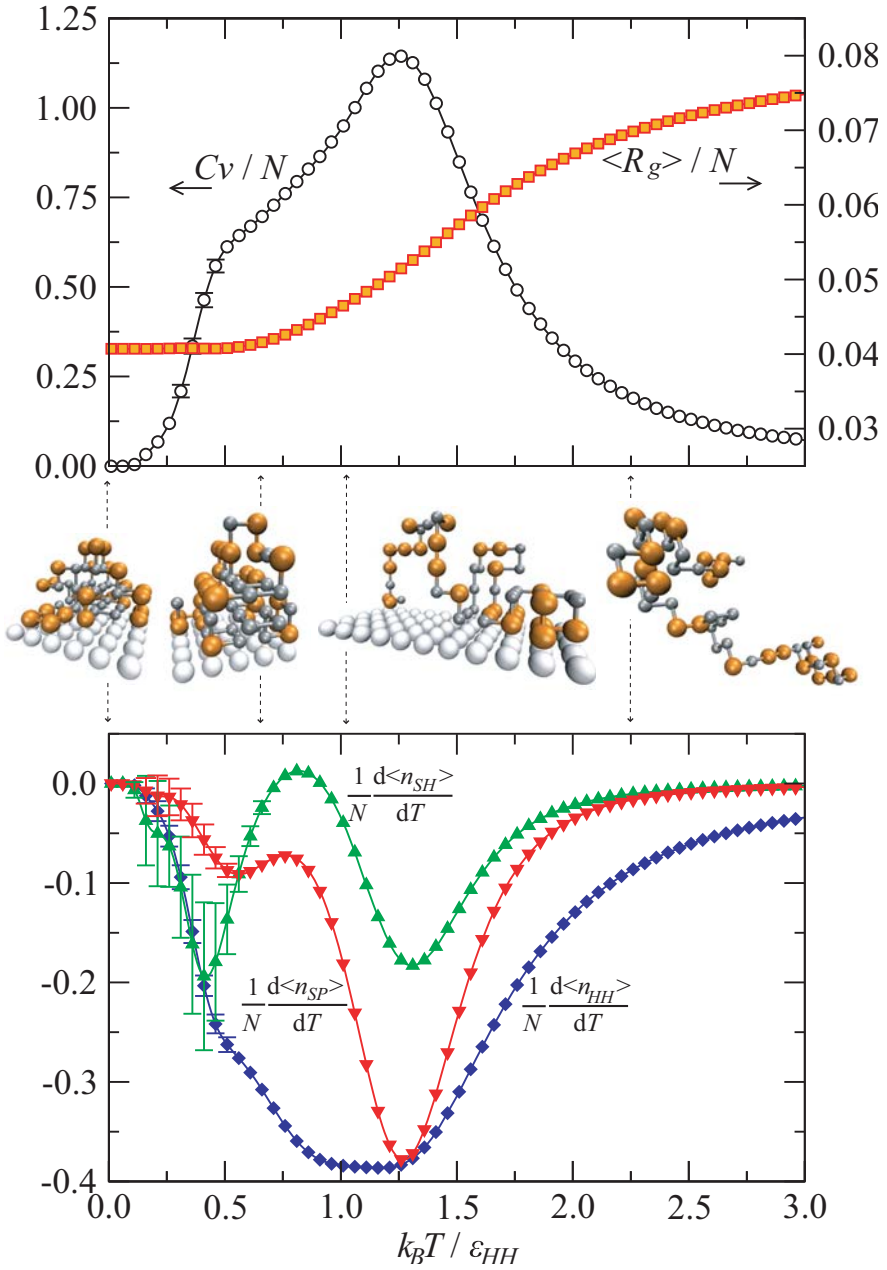


Figure 6.13: Thermodynamics of the 3D48 interacting with surface P1/2 ($\varepsilon_{SH} = 0, \varepsilon_{SP} = \frac{1}{2}\varepsilon_{HH}$), which shows a typical Category III transition.
Upper panel: The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$ as a function of the effective temperature k_BT/ε_{HH}. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Middle panel: Typical configurations at different temperatures.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH}\rangle /dT$, and those of the numbers of surface contacts, $(1/N)d\langle n_{SH}\rangle / dT$ and $(1/N)d\langle n_{SP}\rangle /dT$, as a function of k_BT/ε_{HH}, respectively.

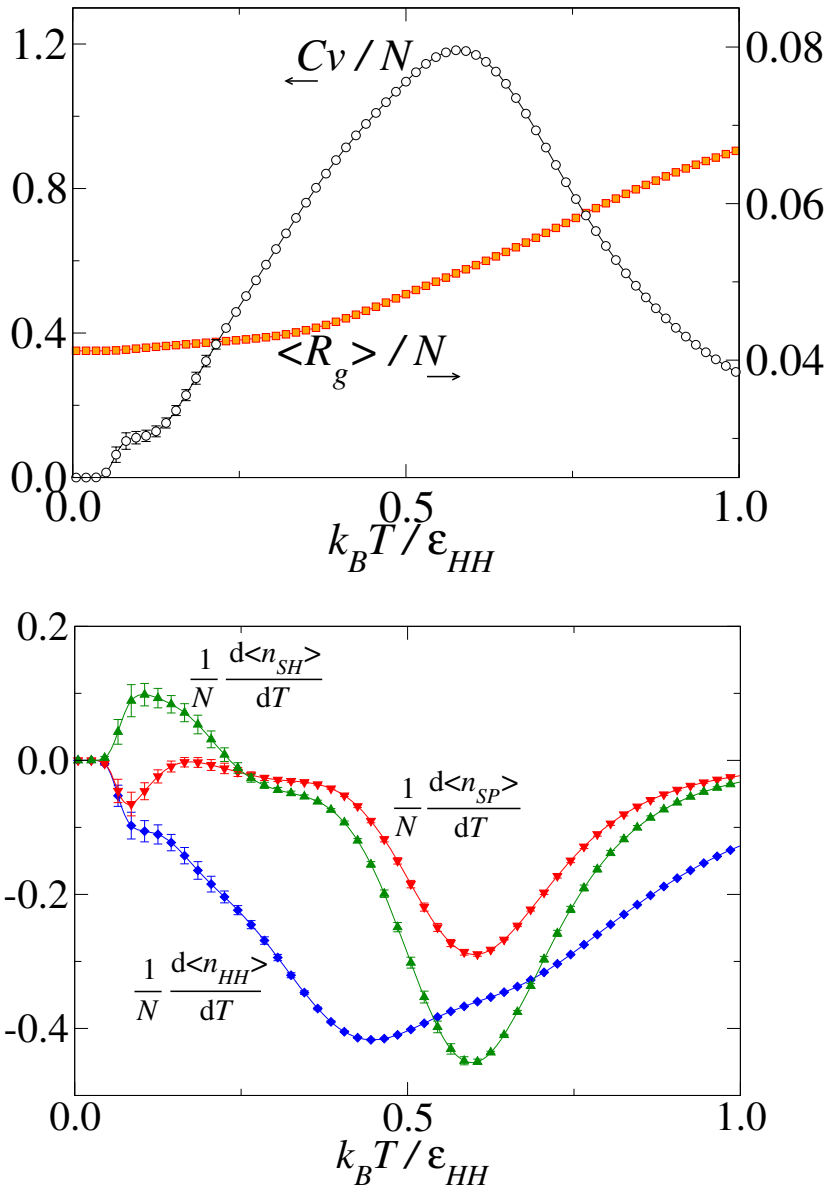


Figure 6.14: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D48 interacting with surface H1/2 ($\varepsilon_{SH} = \frac{1}{2}\varepsilon_{HH}, \varepsilon_{SP} = 0$). This is an example of a Category III transition with a shoulder in C_V/N at a very low temperature. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH}\rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH}\rangle /dT$ and $(1/N)d\langle n_{SP}\rangle /dT$, as a function of k_BT/ε_{HH}.

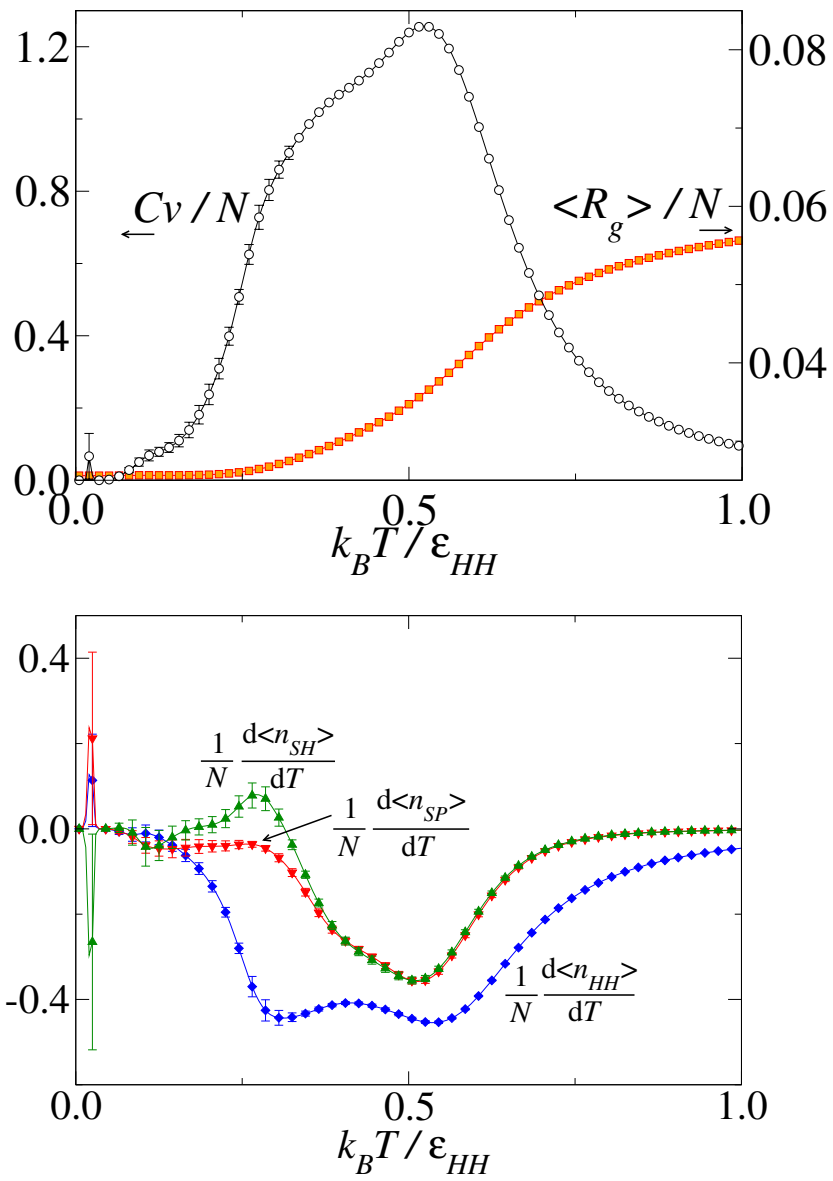


Figure 6.15: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D103 interacting with surface H1/2 ($\varepsilon_{SH} = \frac{1}{2}\varepsilon_{HH}, \varepsilon_{SP} = 0$). This is another example of a Category III transition with a shoulder in C_V/N. The horizontal arrows beside the labels indicate the axes to which the quantities refer.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH}\rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH}\rangle /dT$ and $(1/N)d\langle n_{SP}\rangle /dT$, as a function of k_BT/ε_{HH}.

6.6 Category IV: folding behavior with a very weakly attractive surface

For weaker attractive surfaces, adsorption and flattening occur at even lower temperatures. They become distinguishable from the H-core formation which takes place at a higher temperature, forming two or even three distinct peaks in C_V. We generally classify systems with this transition hierarchy as Category IV. However, since it is computationally too expensive to simulate longer sequences with a very weakly attractive surface due to the larger number of energy bins required, data are only available from the 2D36 sequence.

Although one cannot prove rigorously that there are no other transition categories between Categories III and IV, it is probably not likely by considering the following "hand-waving" argument. Assume that adsorption, flattening and H-core formation are the only three dominant transition processes for HP protein adsorption. There are $3! = 6$ ways of permuting these three processes. But since there is a constraint that flattening has to take place after (i.e., at a lower temperature than) adsorption, that rules out three permutations and only three of them are physical upon cooling:

- adsorption $\rightarrow$ flattening $\rightarrow$ H-core formation;
- adsorption $\rightarrow$ H-core formation $\rightarrow$ flattening;
- H-core formation $\rightarrow$ adsorption $\rightarrow$ flattening.

The first two possibilities have already been identified as Categories I and II respectively. Furthermore, allowing the three transition processes to occur nearly together gives an extra possibility as Category III. The last possibility is then Category IV, and no other hierarchies are possible. Without loss of generality, we thus believe that Category IV is the only transition hierarchy possible for systems with a very weakly interacting surface.

6.7 Crossover between two categories

There are also a few individual cases where the transition behavior demonstrates dual properties of two categories, as seen in the two examples obtained from the 3D103 sequence interacting with surface A1/2 (Figure 6.16) and surface P1 (Figure 6.17).

At first glance, the combined pattern of C_V and $\langle R_g \rangle$ in Figure 6.16 suggests that it belongs to Category I, where flattening occurs in the temperature range $k_BT/\varepsilon_{HH} \approx 1.0$ to 2.5. However, further investigation of the thermal derivatives of the numbers of surface contacts reveals the presence of another flattening process at $k_BT/\varepsilon_{HH} \approx 0.4$, which accounts for the little spike in C_V and a kink in $\langle R_g \rangle$ at the same temperature. This is obviously a Category II feature where flattening occurs at a lower temperature than the construction of an hydrophobic core. Indeed, what happens in this system is that the flattening process splits into two pieces and takes place both before and after the H-core formation, yielding a crossover order of transitions between Categories I and II, i.e., adsorption $\rightarrow$ flattening $\rightarrow$ H-core formation $\rightarrow$ flattening. This phenomenon thus accounts for the mixed signals in the thermodynamic and structural quantities.

Figure 6.17 shows another example of a mixed Categories I and II transition, in which the low temperature flattening at $k_BT/\varepsilon_{HH} \approx 0.13$ causes a small shoulder in C_V and a kink in $\langle R_g \rangle$ respectively.

A remark in identifying this kind of rare events is that the simplified classification method of analyzing only the combined pattern of C_V and $\langle R_g \rangle$ as proposed in Section 6.1 have failed to recognize the low temperature flattening process. The two examples shown above would have been regarded as Category I if other structural parameters were not looked into. The simplified method should thus be used merely as a rough guide with caution, and it is more secure to analyze as many structural properties as possible when categorizing the structural transition behavior.

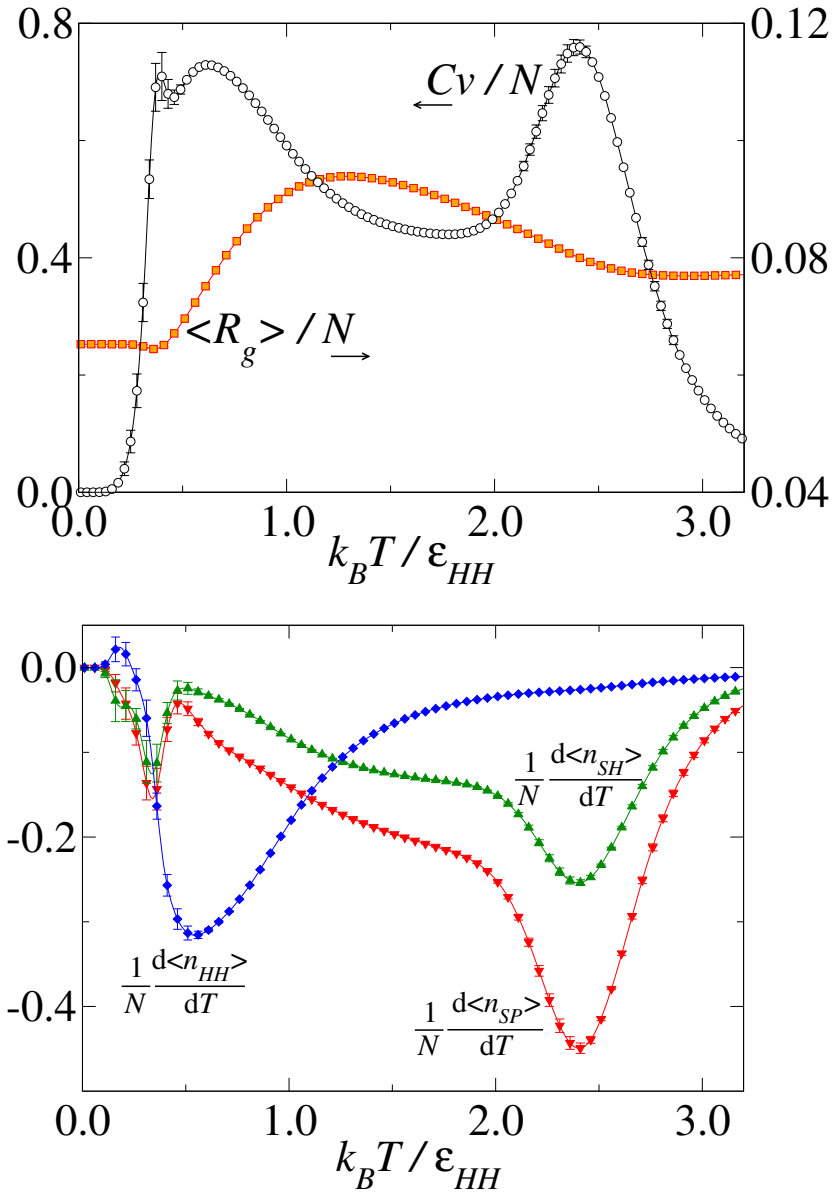


Figure 6.16: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D103 interacting with surface A1/2 ($\varepsilon_S = \frac{1}{2}\varepsilon_{HH}$). This is an example of a dual behavior of Category I and II. In both figures, the horizontal arrows beside the labels indicate the axes to which the quantities refer.

Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH}\rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH}\rangle /dT$ and $(1/N)d\langle n_{SP}\rangle /dT$, as a function of k_BT/ε_{HH}.

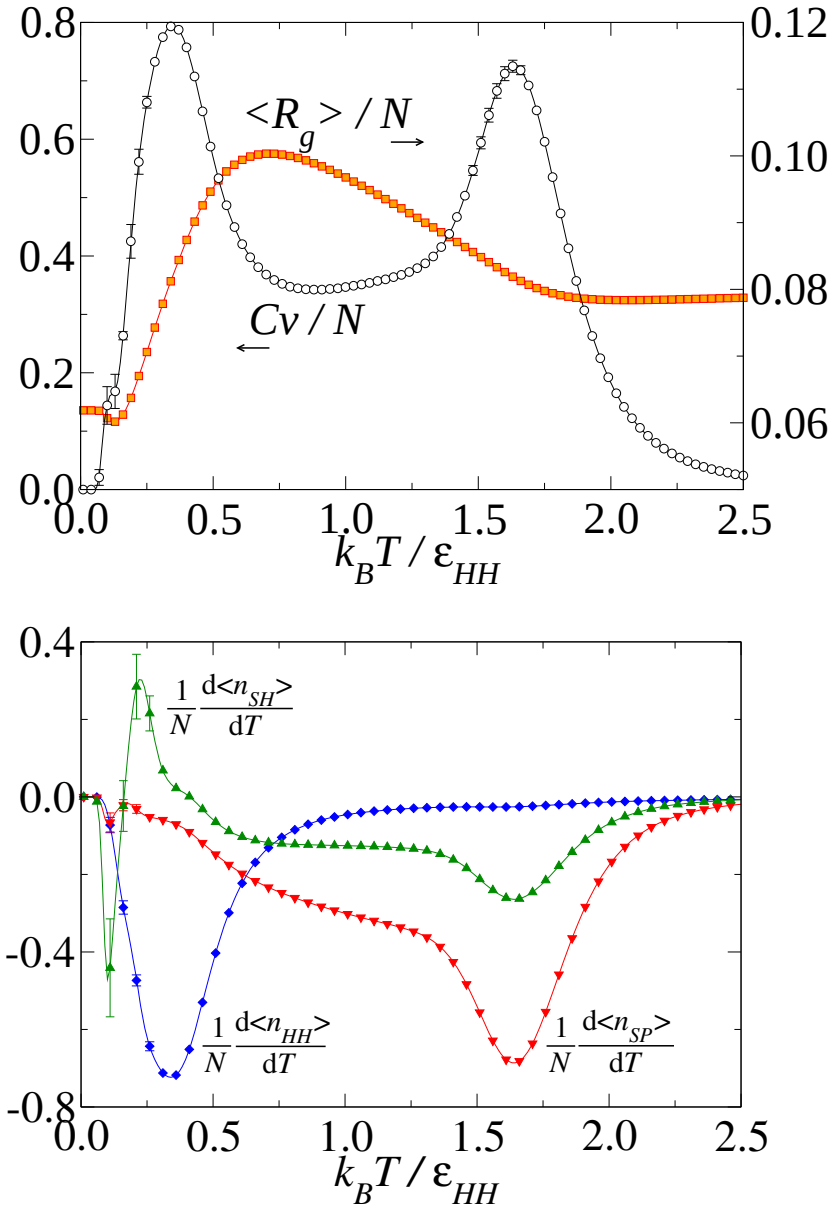


Figure 6.17: **Upper panel:** The specific heat, C_V/N, and the average radius of gyration per monomer, $\langle R_g \rangle /N$, as a function of the effective temperature k_BT/ε_{HH} for 3D103 interacting with surface P1 ($\varepsilon_{SH} = 0, \varepsilon_{SP} = \varepsilon_{HH}$). This is another example of a dual behavior of Category I and II. In both figures, the horizontal arrows beside the labels indicate the axes to which the quantities refer.
Lower panel: Derivatives of the average numbers of H-H contacts per monomer, $(1/N)d\langle n_{HH} \rangle /dT$, and that of the average numbers of surface contacts per monomer, $(1/N)d\langle n_{SH} \rangle /dT$ and $(1/N)d\langle n_{SP} \rangle /dT$, as a function of k_BT/ε_{HH}.

6.8 Classification of categories using relative surface attractive strengths

The described classification scheme effectively generalized the folding behavior into a few transition hierarchies. One step further, we investigate if these hierarchies are related to any intrinsic system parameter(s), or a combination of them, for the general HP lattice protein adsorption problem. If it is the case, this implies that the folding behavior is almost certainly predictable once the system is set up.

Recall our model setting; there are seven tunable system parameters, namely:

- h_w: the separation between the attractive bottom surface and the neutral upper wall;
- ε_{HH}: attraction strength of the internal H-H interactions;
- ε_{SH}, ε_{SP}: attraction strengths between the bottom surface and the H or P monomers, respectively;
- N: chain length;
- N_H, N_P: numbers of H or P monomers in a chain.

Making use of these quantities as a starting point, we observe that the dominating factor determining the transition category is the relative surface attraction. Specifically, it is the ratio between $\varepsilon_{SH} + \varepsilon_{SP}$ and ε_{HH}. We have also investigated other possible quantities which might be used for categorizing the transition hierarchies, e.g., proportion of the H and P monomers with respect to the chain length, N_H/N and N_P/N, or the relative surface attraction weighted by these factors, $(N_H/N)\varepsilon_{SH} + (N_P/N)\varepsilon_{SP}$ compared to ε_{HH}. For these cases, no explicit association with the hierarchies can be be concluded as precisely and elegantly as $(\varepsilon_{SH} + \varepsilon_{SP})/\varepsilon_{HH}$.

$(\varepsilon_{SH}+\varepsilon_{SP})/\varepsilon_{HH}$	**Category I**	**Category II**	**Category III**
>1	A1: 36, 48.1, 48.9, 64, 67, 103 A2: 36, 48.9, 64, 67 H2: 48.1, 48.9, 64, 67 P2: 48.1, 48.9, 64, 67		
$=1$		A1/2: 103 P1: 103 A1/2: 36, 48.1, 48.9, 64, 67 H1: 48.1, 48.9, 64, 67, 103 P1: 48.1, 48.9, 64, 67	
<1		H1/2: 67 P1/2: 67, 103	H1/2: 48.1, 48.9, 64, 103 P1/2: 48.1, 48.9 A1/3: 36, 64

Table 6.2: Distribution of transition categories with respect to the relative surface attractions. The abbreviations refer to the surface types introduced in Table 6.1. The numbers are the short forms of the benchmark sequences (e.g. 36 stands for 2D36, 48.1 and 48.9 correspond to Seq. 1 and 9 among the ten "Harvard sequences" respectively. 48.9 is also the 3D48 we have introduced in the previous text.)

Table 6.2 shows the distribution of transition categories against the relative surface attractions for systems with various chain lengths and surface types. Ideally, a perfect correspondence between the transition categories and the relative surface attractions is implied if only the diagonal compartments are filled in the table. In reality, as thermodynamic subtleties vary from sequence to sequence, some off-diagonal compartments are also occupied

(e.g., some systems with $(\varepsilon_{SH} + \varepsilon_{SP})/\varepsilon_{HH} < 1$ show Category II behavior).

A few systems also reveal "category duality" in which case the thermodynamics have dual properties from two consecutive categories (e.g. 3D103 interacting with surfaces A1/2 or P1 as discussed in Section 6.7). Nonetheless, the generality of our classification scheme is still apparent and allows for the inference of the following basic rules:

1. Category I occurs for surfaces which are strongly attractive, i.e., $(\varepsilon_{SH}+\varepsilon_{SP})/\varepsilon_{HH} > \approx 1$

2. Category II occurs when the hydrophobic internal attraction is approximately comparable to the surface attractions, i.e., $(\varepsilon_{SH} + \varepsilon_{SP})/\varepsilon_{HH} \sim 1$

3. Category III can only occur when surface strengths are relatively weak compared to ε_{HH}, i.e., $(\varepsilon_{SH} + \varepsilon_{SP})/\varepsilon_{HH} < 1$

Although the conditions for Category IV to take place are not concluded from the results presented here for the longer HP sequences, from the results of 3D36 in Section 6.1 and the argument in Section 6.6, we believe that it should follow Category III when the surface attractions becomes even weaker, without loss of generosity.

We stress that this classification is an inference based on multiple HP sequences of various lengths and attributes. In Table 6.2, we have also included results from 2D36 and two other sequences, which were used as a "validation data set" for the adequacy of the classification scheme. They are 3D48.1 (another 48mer: $HPH_2P_2H_4PH_3P_2H_2P_2HPH_3PHPH_2P_2H_2P_3H$-$P_8H_2$), and 2D64. 2D64 was originally proposed to test a 2D genetic algorithm just like 2D36[77], whereas the 3D48.1 is Seq. 1 of the ten "Harvard testing sequences" [84]. All these results fall into the diagonal compartments of the table, reinforcing that our classification scheme is applicable to other sequences interacting with an absorbing substrate without loss of generality. This is also a breakthrough in the understanding of adsorption properties of lattice proteins: instead of behaving individually, the thermodynamics of HP proteins

do follow common patterns in structural transitions when they interact with an absorbing substrate.

6.9 Remarks on the structural measures and categories

Our results demonstrate that a comprehensive analysis on C_V with a set of appropriate structural quantities is essential to shed light on recognizing structural transformations, especially those subtle ones for which C_V alone provides insufficient information. We also note that in identifying phase transitions, the peaks observed in structural quantities and those in C_V might be slightly off. One possible explanation could be the finite size effects: the cause for $d\langle R_g\rangle/dT$'s peak being slightly off compared to that of C_V's[138]. Nevertheless, this does not affect our identification scheme much as the difference in transition temperatures is sufficiently small compared to the difference in temperature scales for different phase transitions.

Chapter 7

Conclusions and Outlook

In this work, the structural phase transitions of protein adsorption was studied with a coarse-grained lattice model, the hydrophobic-polar (HP) model, interacting with a surface which either attracts all monomers, only hydrophobic monomers or only polar monomers of the chain. Wang-Landau sampling was employed to obtain the one-dimensional energy density of states, while multicanonical sampling was used to estimate the two-dimensional densities of states in energy and different structural parameters. Two inventive Monte Carlo trial moves, pull moves and bond-rebridging moves, have been implemented in combination with these two sampling algorithms to enhance the ability of obtaining a thorough survey of the conformational space. This resulted in densities of states which are more accurately estimated, and this allowed for the calculation of the thermodynamic and structural quantities to a higher resolution, especially in the low temperature regime. Ground state energies for the system simulated were also reported as "side products".

The first stage of the study of structural "phase transitions" involved the identification of transition processes for a 36mer. Three structural transitions, namely, H-core formation, adsorption, and flattening, have been identified. We have illustrated that instead of using specific heat alone, a comprehensive analysis of other structural parameters alongside with

it, is essential in revealing the folding and adsorption behavior. We found that the radius of gyration and the derivatives of the numbers of surface contacts are particularly informative in our case.

With the capability of distinguishing different structural changes, the second stage of our study concerned about the generalization of the orders of occurrence of these transformations, i.e., the "phase transition" hierarchies. We have generalized four main types of transition hierarchies, which were first identified from the 36mer case, and were then reinforced by the extensive results of a 48mer, a 67mer and a 103mer.

In the next stage we found that the occurrence of these transition hierarchies depends, in principle, on the attractive strengths of the surface relative to the internal hydrophobic attraction, regardless of the surface type, chain length or composition of H and P monomers of an HP sequence. Two other benchmark sequences, another 48mer and a 64mer, have been used to confirm the validity of our classification scheme.

Although there were a few rare crossover cases in which dual properties of two categories were observed, the classification scheme proposed in this work provides a general and representative picture of the thermodynamics of HP proteins interacting with an adsorbing substrate. Classifying transition hierarchies by a comprehensive analysis of combined patterns of specific heat and appropriate structural parameters also sets a paradigm of approaching similar systems of large conformational and sequence spaces, for instance, HP proteins interacting with two confining, attractive surfaces[139–141].

However, further investigation is necessary to determine if there is a more rigorous relation between the transition categories and the relative surface attraction. More statistics from longer chains, or chains of the same length but different H and P composition would help clarifying the problem. The next question is whether the same conclusion can be drawn, or what discrepancies will be found, for other lattice models with other energy functions, i.e., different interactions between monomers and with the surface. Another important question

is whether the classification scheme can be applied to off-lattice models. In this case, thermodynamics of other structural parameters, e.g., the gyration tensor, density profile or any suitable ones, should also be examined in verifying and improving the classification scheme. All these together are essential in determining the effectiveness of using different simplified protein models in computer simulations to study protein adsorption from a macroscopic perspective.

Appendix A

Tests of Random Number Generators

The quality of the pseudo-random number generator is indispensable in obtaining reliable results in Monte Carlo simulations. Nevertheless, even "high quality" generators which are able to pass a number of statistical tests can yield systematically incorrect results when used for specific algorithms. One example is the combination of R250 with the Wolff algorithm[142]. On the other hand, since these random number generators are based on deterministic recursion rules, it is not unreasonable to find them problematic under certain conditions[143]. It is, therefore, necessary to check if a pseudo-random number generator works fine for the system being simulated.

In all our simulations, we have used RANLUX[144] as the random number generator. It uses a lagged-fibonacci-with-skipping algorithm, has a long period of about 10^{171}, and is recommended by the GNU Scientific Library (GSL) for its "reliable source of uncorrelated numbers" and "strongest proof of randomness"[145]. In addition, we adopted the double precision version provided by the GSL, `gsl_rng_ranlxd2`, which produces double precision outputs.

Since RANLUX has been tested extensively and is widely accepted to be good, we focused on the question of whether it is suitable for the simulation of the HP model using Wang-

Pseudo-random number generator	Simulation time (days)
RANLUX	7.06 ± 1.58
Mersenne Twister	6.05 ± 1.39

Table A.1: Average time for simulating 2D36 interacting with a very weakly attractive surface ($\varepsilon_S = \frac{1}{12}\varepsilon_{HH}, h_w = 37$) on an IBM Power4 1.3GHz processor, using different random number generators. The average value is based on 15 individual runs.

Landau sampling. One simple way is to compare it with another pseudo-random number generator. If they yield agreeable results, we are more confident that both random number generators are fine as the probability of two generators getting the same wrong results is quite low. In Figure A.1, the $g(E)$ of 2D36 interacting with a very weakly attractive surface as obtained by RANLUX is compared to that obtained by another high quality pseudo-random number generator, Mersenne Twister[146]. 15 runs were performed for each algorithm to obtain the statistical errors. Despite the roughness of the $g(E)$, the results from both methods agree extremely well with each other throughout the entire energy range. Only tiny discrepancies are observed for the lowest two energy states where the error bars overlap, as shown in the inset of Figure A.1. We then believe that RANLUX would not incur a systematic error (at least not noticeable if there was one) to our simulations.

A drawback of RANLUX is its sluggishness. During the generation of random numbers, RANLUX discards at least half of the numbers it generated. The higher the "luxury" level, the more numbers it discards from the sequence. Table A.1 shows a comparison in computation time of simulating the aforementioned system. For a small system like this, RANLUX takes approximately 17% more time than Mersenne Twister, although this difference falls within the variance range. To ensure the quality of our work we have used RANLUX for all of our simulations. Nevertheless, as the time difference will grow larger with the system size when more random numbers are drawn, one should take this consideration into account if simulation time is a concern.

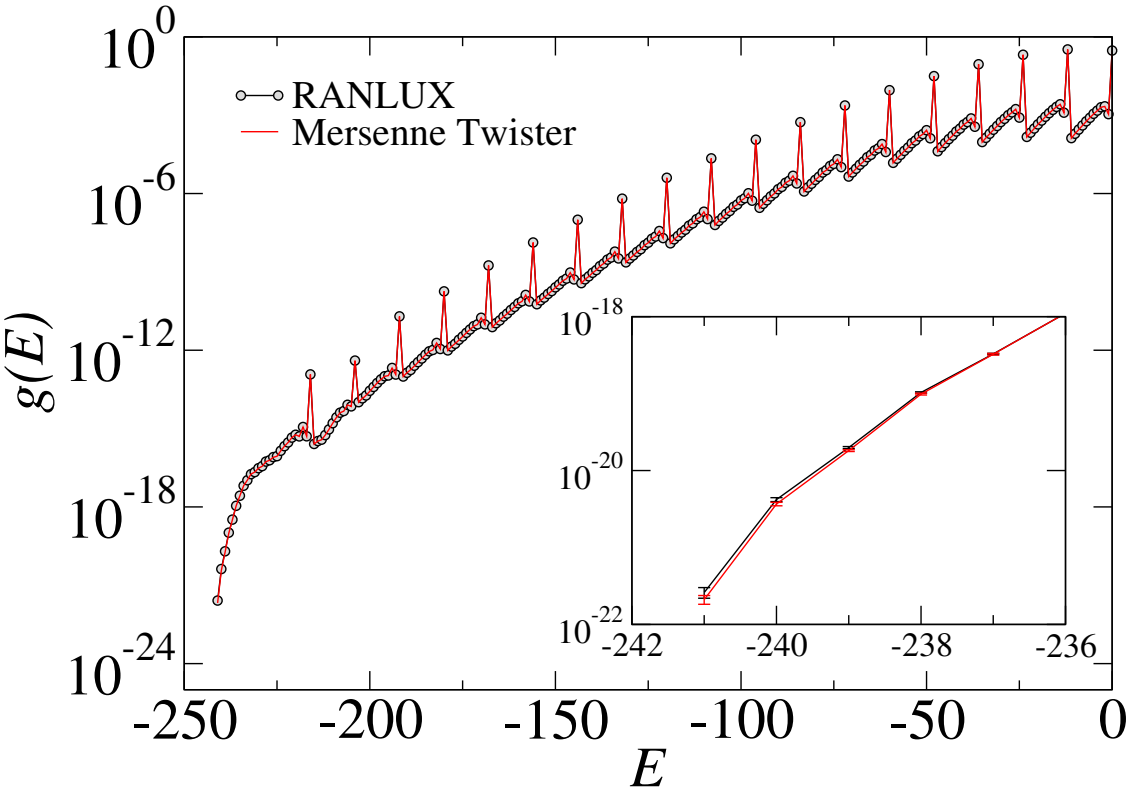


Figure A.1: The densities of states in energy for 2D36 interacting with a very weakly attractive surface ($\varepsilon_S = \frac{1}{12}\varepsilon_{HH}$), obtained by two random number generators: Mersenne twister and RANLUX. Statistical errors are obtained from 15 runs for each method. They are smaller than the data points and are not shown in the major panel.

Bibliography

[1] T. A. Horbett and J. L. Brash, *Protein at interfaces II: Fundamentals and applications* (Washington DC: American Chemical Society, 1995)

[2] M. Sarikaya, C. Tamerler, A. K. Y. Jen, K. Schulten, and F. Baneyx, *Molecular biomimetics: nanotechnology through biology*. Nat. Mater. **2**, 577 (2003)

[3] V. Hlady and J. Buijs, *Protein adsorption on solid surfaces*. Curr. Opin. Biotech. **7**, 72 (1996)

[4] J. J. Gray, *The interaction of proteins with solid surfaces*. Curr. Opin. Colloid Interface Sci. **5**, 315 (2000)

[5] G. MacBeath and S. L. Schreiber, *Printing proteins as microarrays for high-throughput function determination*. Science **289**, 1760 (2000)

[6] E. Phizicky, P. I. Bastiaens, H. Zhu, M. Snyder, and S. Fields, *Protein analysis on a proteomic scale*. Nature **422**, 208 (2003)

[7] S. S. Davis and L. Illum, *Polymeric microspheres as drug carriers*. Biomaterials **9**, 111 (1988)

[8] W. R. Gombotz and D. K. Pettit, *Biodegradable polymers for protein and peptide drug-delivery*. Bioconjugate Chem. **6**, 332 (1995)

[9] C. Pinholt, R. A. Hartvig, N. J. Medlicott, and L. Jorgensen, *The importance of interfaces in protein drug delivery - why is protein adsorption of interest in pharmaceutical formulations?* Expert Opin. Drug Deliv. **8**, 949 (2011)

[10] G. Brown, K. Odbadrakh, D. M. Nicholson, and M. Eisenbach, *Convergence for the Wang-Landau density of states*. Phys. Rev. E **84**, 065702(R) (2011)

[11] S. R. Whaley, D. S. English, E. L. Hu, P. F. Barbara, and A. M. Belcher, *Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly*. Nature **405**, 665 (2000)

[12] K. Goede, P. Busch, and M. Grundmann, *Binding specificity of a peptide on semiconductor surfaces*. Nano Lett. **4**, 2115 (2004)

[13] M. Bachmann, K. Goede, A. G. Beck-Sickinger, M. Grundmann, A. Irbäck, and W. Janke, *Microscopic mechanism of specific peptide adhesion to semiconductor substrates*. Angew. Chem. Int. Ed. **49**, 9530 (2010)

[14] R. Fernandez-Lafuente, P. Armisén, P. Sabuquillo, G. Fernández-Lorente, and J. M. Guisán, *Immobilization of lipases by selective adsorption on hydrophobic supports*. Chem. Phys. Lipids **93**, 185 (1998)

[15] R. S. Kane and A. D. Stroock, *Nanobiotechnology: Protein-nanomaterial interactions*. Biotechnol. Prog. **23**, 316 (2007)

[16] J. D. Andrade and V. Hlady, *Protein adsorption and materials biocompatibility: A tutorial review and suggested hypotheses*. Adv. Polym. Sci. **79**, 1 (1986)

[17] D. P. Landau and K. Binder, *A Guide to Monte Carlo simulations in statistical physics* (Cambridge University Press, 2009), 3rd edn.

[18] K. A. Dill, *Theory for the folding and stability of globular proteins*. Biochemistry **24**, 1501 (1985)

[19] K. F. Lau and K. A. Dill, *A lattice statistical mechanics model of the conformational and sequence spaces of proteins*. Macromolecules **22**, 3986 (1989)

[20] C. Branden and J. Tooze, *Introduction to protein structure* (Garland Science, 1999), 2nd edn.

[21] T. E. Creighton, *Proteins: structures and molecular properties* (W. H. Freeman, 1992), 2nd edn.

[22] A. Böck, K. Forchhammer, J. Heider, and C. Baron, *Selenoprotein synthesis: an expansion of the genetic code*. Trends Biochem. Sci. **16**, 463 (1991)

[23] J. F. Atkins and R. Gesteland, *The 22nd amino acid*. Science **296**, 1409 (2002)

[24] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular biology of the cell* (Garland Science, 1994), 3rd edn.

[25] W. Kauzmann, *Some factors in the interpretation of protein denaturation*. Adv. Protein Chem. **14**, 1 (1959)

[26] K. A. Dill, *The meaning of hydrophobicity*. Science **250**, 297 (1990)

[27] M. Rabe, D. Verdes, and S. Seeger, *Understanding protein adsorption phenomena at solid surfaces*. Adv. Colloid Interface Sci. **162**, 87 (2011)

[28] J. J. Ramsden, *Experimental methods for investigating protein adsorption-kinetics at surfaces*. Q. Rev. Biophys. **27**, 41 (1994)

[29] V. Hlady, J. Buijs, and H. P. Jennissen, *Methods for studying protein adsorption*. Methods Enzymol. **309**, 402 (1999)

[30] A. Kolinski and J. Skolnick, *Reduced models of proteins and their applications*. Polymer **45**, 511 (2004)

[31] P. Weroński, *Application of the extended RSA models in studies of particle deposition at partially covered surfaces*. Adv. Colloid Interface Sci. **118**, 1 (2005)

[32] R. C. Chatelier and A. P. Minton, *Adsorption of globular proteins on locally planar surfaces: models for the effect of excluded surface area and aggregation of adsorbed protein on adsorption equilibria*. Biophys. J. **71**, 2367 (1996)

[33] A. P. Minton, *Effects of excluded surface area and adsorbate clustering on surface adsorption of proteins I. Equilibrium models*. Biophys. Chem. **86**, 239 (2000)

[34] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Funnels, pathways and the energy landscape of protein folding: a synthesis*. Prot. Struct. Funct. Genet. **21**, 167 (1995)

[35] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, *Theory of protein folding: the energy landscape perspective*. Ann. Rev. Phys. Chem. **48**, 545 (1997)

[36] V. S. Pande, A. Y. Grosberg, T. Tanaka, and D. S. Rokhsar, *Pathways for protein folding: is a new view needed?* Curr. Opin. Struct. Biol. **8**, 68 (1998)

[37] V. S. Pande and D. S. Rokhsar, *Folding pathway of a lattice model for proteins*. Proc. Natl. Acad. Sci. USA **96**, 1273 (1999)

[38] S. Schnabel, M. Bachmann, and W. Janke, *Two-state folding, folding through intermediates, and metastability in a minimalistic hydrophobic-polar model for proteins*. Phys. Rev. Lett. **98**, 048103 (2007)

[39] H. Arkin, *Determination of the structure of the energy landscape for coarse-grained off-lattice models of folding heteropolymers*. Phys. Rev. E **78**, 041914 (2008)

[40] M. P. Taylor, W. Paul, and K. Binder, *Two-state protein-like folding of a homopolymer chain*. Physics Procedia **4**, 151 (2010)

[41] C. B. Anfinsen, *Principles that govern folding chains*. Science **181**, 223 (1973)

[42] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, *Principles of protein folding–a perspective from simple exact models*. Protein Sci. **4**, 561 (1995)

[43] K. A. Dill, *Polymer principles and protein folding*. Protein Sci. **8**, 1166 (1999)

[44] A. Sali, E. Shakhnovich, and M. Karplus, *How does a protein fold?* Nature **369**, 248 (1994)

[45] A. Sali, E. Shakhnovich, and M. Karplus, *Kinetics of protein folding: A lattice model study of the requirements for folding to the native state*. J. Mol. Biol. **235**, 1614 (1994)

[46] T. Lazaridis and M. Karplus, *Thermodynamics of protein folding: a microscopic view*. Biophys. Chem. **100**, 367 (2003)

[47] P. Flory, *Statistical mechanics of chain molecules* (Wiley, New York, 1969), 1st edn.

[48] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The protein data bank*. Nucl. Acids Res. **28**, 235 (2000)

[49] Protein Data Bank: www.pdb.org

[50] J. Drenth, *Principles of protein x-ray crystallography* (Springer, 2006), 3rd edn.

[51] G. S. Rule and T. K. Hitchens, *Fundamentals of protein NMR spectroscopy* (Springer, 2005), 1st edn.

[52] K. Wüthrich, *NMR studies of structure and function of biological macromolecules (Nobel Lecture)*. J. Biomol. NMR **27**, 13 (2003)

[53] H. Yu, *Extending the size limit of protein nuclear magnetic resonance*. Proc. Natl. Acad. Sci. USA **96**, 332 (1999)

[54] J. Frank, *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state* (Oxford University Press, USA, 2006), 2nd edn.

[55] B. Liedberg, B. Ivarsson, P.-O. Hegg, and I. Lundstöm, *On the adsorption of β-lactoglobulin on hydrophilic gold surfaces: studies by infrared reflection-adsorption spectroscopy and ellipsometry*. J. Colloid Interface Sci. **114**, 386 (1986)

[56] B. K. Lok, Y. L. Cheng, and C. R. Robertson, *Total internal reflection fluorescence: a technique for examining interactions of macromolecules with solid surfaces*. J. Colloid Interface Sci. **91**, 87 (1983)

[57] C. Calonder, Y. Tie, and P. R. V. Tassel, *History dependence of protein adsorption kinetics*. Proc. Natl. Acad. Sci. USA **98**, 10664 (2001)

[58] D. C. Cullen and C. R. Lowe, *AFM studies of protein adsorption. I. Time-resolved protein adsorption to highly oriented pyrolytic graphite*. J. Colloid Interface Sci. **166**, 102 (1994)

[59] S. M. Chen, F. A. Ferrone, and R. Wetzel, *Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation*. Proc. Natl. Acad. Sci. USA **99**, 11884 (2002)

[60] M. C. L. Maste, W. Norde, and A. J. W. G. Visser, *Adsorption-induced*

conformational changes in the serine proteinase savinase: a tryptophan fluorescence and circular dichroism study. J. Colloid Interface Sci. **196**, 224 (1997)

[61] K. Nakanishi, T. Sakiyama, and K. Imamura, *On the adsorption of proteins on solid surfaces, a common but very complicated phenomenon*. J. Biosci. Bioeng **91**, 233 (2001)

[62] B. R. Brooks, R. E. Bruccoleri, D. J. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations*. J. Comput. Chem. **4**, 187 (1983)

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

[64] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. J. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*. J. Am. Chem. Soc. **117**, 5179 (1995)

[65] R. A. Latour, *Molecular simulation of protein-surface interactions: Benefits, problems, solutions, and future directions*. Biointerphases **3**, FC2 (2008)

[66] U. H. E. Hansmann and Y. Okamoto, *New Monte Carlo algorithms for protein folding*. Curr. Opin. Struct. Biol. **9**, 177 (1999)

[67] F. Ganazzoli and G. Raffaini, *Computer simulation of polypeptide adsorption on model biomaterials*. Phys. Chem. Chem. Phys. **7**, 3651 (2005)

[68] J. R. Banavar and A. Maritan, *Physics of proteins*. Annu. Rev. Biophys. Biomol. Struct. **36**, 261 (2007)

[69] H. R. Warner, *Kinetic theory and rheology of dilute suspensions of finitely extendible dumbbells*. Ind. Eng. Chem. Fundam. **11**, 379 (1972)

[70] K. Binder and M. Müller, *Monte Carlo simulation of block copolymers*. Curr. Opin. Colloid Interface Sci. **16**, 463 (1991)

[71] B. Berger and T. Leighton, *Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete*. J. Comput. Biol. **5**, 27 (1998)

[72] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, *On the complexity of protein folding*. J. Comput. Biol. **5**, 423 (1998)

[73] 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)

[74] G. H. Paine and H. A. Scheraga, *Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. I. Backbone structure of enkephalin*. Biopolymers **24**, 1391 (1985)

[75] Z. Q. Li and H. A. Scheraga, *Monte Carlo-minimization approach to the multiple-minima problem in protein folding*. Proc. Natl. Acad. Sci. USA **84**, 6611 (1987)

[76] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*. Science **220**, 671 (1983)

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

[78] J. T. Pedersen and J. Moult, *Genetic algorithms for protein structure prediction*. Curr. Opin. Struct. Biol. **6**, 227 (1996)

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

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

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

[82] X. Zhao, *Advances on protein folding simulations based on the lattice HP models with natural computing*. Applied Soft Computing **8**, 1029 (2008)

[83] K. Yue and K. A. Dill, *Forces of tertiary structural organization in globular-proteins*. Proc. Natl. Acad. Sci. USA **92**, 146 (1995)

[84] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, *A test of lattice protein-folding algorithms*. Proc. Natl. Acad. Sci. USA **92**, 325 (1995)

[85] K. Yue and K. A. Dill, *Sequence-structure relationships in protein and copolymers*. Phys. Rev. E **48**, 2267 (1993)

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

[87] Y. Iba, G. Chikenji, and M. Kikuchi, *Simulation of lattice polymers with multi-self-overlap ensemble*. J. Phys. Soc. Jpn **67**, 3327 (1998)

[88] Y. Iba, G. Chikenji, and M. Kikuchi, *Multi-self-overlap ensemble for protein folding: Ground state search and thermodynamics*. Phys. Rev. Lett. **83**, 1886 (1999)

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

[90] H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler, *New Monte Carlo algorithm for protein folding*. Phys. Rev. Lett. **80**, 3149 (1998)

[91] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler, *Testing a new Monte Carlo algorithm for protein folding*. Proteins **32**, 52 (1998)

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

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

[94] M. Bachmann and W. Janke, *Multicanonical chain-growth algorithm*. Phys. Rev. Lett. **91**, 208105 (2003)

[95] M. Bachmann and W. Janke, *Thermodynamics of lattice heteropolymers*. J. Chem. Phys. **120**, 6779 (2004)

[96] T. Prellberg and J. Krawczyk, *Flat histogram version of the pruned and enriched Rosenbluth method*. Phys. Rev. Lett. **92**, 120602 (2004)

[97] T. Prellberg, J. Krawczyk, and A. Rechnitzer, *Polymer simulations with a at histogram stochastic growth algorithm*. In Computer simulation studies in condensed-matter physics XVII. Proceedings of the 17th workshop on recent developments in computer simulation studies in condensed matter physics. p. 122 (2006)

[98] S. C. Kou, J. Oh, and W. H. Wong, *A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling*. J. Chem. Phys. **124**, 244903 (2006)

[99] F. Wang and D. P. Landau, *Efficient, multiple-range random walk algorithm to calculate the density of states*. Phys. Rev. Lett. **86**, 2050 (2001)

[100] 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**, 056101 (2001)

[101] F. Wang and D. P. Landau, *Determining the density of states for classical statistical models by a flat-histogram random walk*. Comput. Phys. Commun. **147**, 674 (2002)

[102] T. Wüst and D. P. Landau, *The HP model of protein folding: A challenging testing ground for Wang-Landau sampling*. Comput. Phys. Commun. **179**, 124 (2008)

[103] T. Wüst and D. P. Landau, *Versatile approach to access the low temperature thermodynamics of lattice polymers and proteins*. Phys. Rev. Lett. **102**, 178101 (2009)

[104] S. M. Liu and C. A. Haynes, *Energy landscapes for adsorption of a protein-like HP chain as a function of native-state stability*. J. Colloid Interface Sci. **284**, 7 (2005)

[105] S. M. Liu and C. A. Haynes, *Mesoscopic dynamic Monte Carlo simulations of*

the adsorption of proteinlike HP chains within laterally constricted spaces. J. Colloid Interface Sci. **282**, 283 (2005)

[106] V. Castells, S. Yang, and P. R. V. Tassel, *Surface-induced conformational changes in lattice model proteins by Monte Carlo simulation*. Phys. Rev. E **65**, 031912 (2002)

[107] V. Castells and P. R. V. Tassel, *Conformational transition free energy profiles of an adsorbed, lattice model protein by multicanonical Monte Carlo simulation*. J. Chem. Phys. **122**, 084707 (2005)

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

[109] A. D. Swetnam and M. P. Allen, *Improved simulations of lattice peptide adsorption*. Phys. Chem. Chem. Phys. **11**, 2046 (2009)

[110] M. Bachmann and W. Janke, *Conformational transitions of nongrafted polymers near an adsorbing substrate*. Phys. Rev. Lett. **95**, 058102 (2005)

[111] M. Bachmann and W. Janke, *Substrate adhesion of a nongrafted flexible polymer in a cavity*. Phys. Rev. E **73**, 041802 (2006)

[112] M. Bachmann and W. Janke, *Thermodynamics of protein folding from coarse-grained models' perspective*. Lect. Notes. Phys. **736**, 203 (2008)

[113] M. Bachmann and W. Janke, *Minimalistic hybrid models for the adsorption of polymers and peptides to solid substrates*. Physics of Particles and Nuclei Letter **5**, 243 (2008)

[114] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, *Toy model for protein folding*. Phys. Rev. E **48**, 1469 (1993)

[115] M. Mann, S. Will, and R. Backofen, *CPSP-tools - Exact and complete algorithms for high-throughput 3D lattice protein studies*. BMC Bioinformatics **9**, 230 (2008)

[116] M. Mann, C. Smith, M. Rabbath, M. Edwards, S. Will, and R. Backofen, *CPSP-web-tools: a server for 3D lattice protein studies*. Bioinformatics **25**, 676 (2009)

[117] R. König and T. Dandekar, *Improving genetic algorithms for protein folding simulations by systematic crossover*. Biosystems **50**, 17 (1999)

[118] R. Ramakrishnan, B. Ramachandran, and J. F. Pekny, *A dynamic Monte Carlo algorithm for exploration of dense conformational spaces in heteropolymers*. J. Chem. Phys. **106**, 2418 (1997)

[119] C. Beutler and K. A. Dill, *A fast conformational search strategy for finding low energy structures of model proteins*. Protein Sci. **5**, 2037 (1996)

[120] E. E. Lattman, K. M. Fiebig, and K. A. Dill, *Modeling compact denatured states of proteins*. Biochemistry **33**, 6158 (1994)

[121] 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)

[122] C. Zhou and R. N. Bhatt, *Understanding and improving the Wang-Landau algorithm*. Phys. Rev. E **72**, 025701(R) (2005)

[123] A. D. Swetnam and M. P. Allen, *Improving the Wang-Landau algorithm for polymers and proteins*. J. Comput. Chem. **32**, 816 (2011)

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

[125] 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., in press (2012)

[126] T. Vogel, Y. W. Li, T. Wüst, and D. P. Landau, (in preparation)

[127] B. A. Berg and T. Neuhaus, *Multicanonical algorithms for 1st order phase-transitions*. Phys. Lett. B **267**, 249 (1991)

[128] B. A. Berg and T. Neuhaus, *Multicanonical ensemble - A new approach to simulate 1st-order phase-transitions*. Phys. Rev. Lett. **68**, 9 (1992)

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

[130] T. Wüst and D. P. Landau, (private communications)

[131] O. Collet, *Conformational rigidity in a lattice model of proteins*. Phys. Rev. E **67**, 061912 (2003)

[132] Y. W. Li, T. Wüst, and D. P. Landau, *Monte Carlo simulations of the HP model (the "Ising model" of protein folding)*. Comput. Phys. Commun. **182**, 1896 (2011)

[133] T. Wüst, Y. W. Li, and D. P. Landau, *Unraveling the beautiful complexity of simple lattice model polymers and proteins using Wang-Landau sampling*. J. Stat. Phys. **144**, 638 (2011)

[134] Y. W. Li, T. Wüst, and D. P. Landau, *Generic folding and transition hierarchies for surface adsorption of HP lattice model proteins*. Submitted to Phys. Rev. E (2012)

[135] T. Vogel, M. Bachmann, and W. Janke, *Freezing and collapse of flexible polymers on regular lattices in three dimensions*. Phys. Rev. E **76**, 061803 (2007)

[136] M. Möddel, W. Janke, and M. Bachmann, *Systematic microcanonical analyses of polymer adsorption transitions*. Phys. Chem. Chem. Phys. **12**, 11548 (2010)

[137] Y. W. Li, T. Wüst, and D. P. Landau, *Surface adsorption of lattice HP proteins: Thermodynamics and structural transitions using Wang-Landau sampling*. JPCS, in press (2012)

[138] S. Sharma and S. K. Kumar, *Finite size effects on locating conformational transitions for macromolecules*. J. Chem. Phys. **129**, 134901 (2008)

[139] B. Pattanasiri, Y. W. Li, D. P. Landau, T. Wüst, and W. Triampo, *Conformational transitions of a confined lattice protein: A Wang-Landau study*. JPCS, in press (2012)

[140] B. Pattanasiri, Y. W. Li, D. P. Landau, and T. Wüst, *Wang-Landau simulations of adsorbed and confined lattice proteins*. Int. J. Mod. Phys. C, in press (2012)

[141] B. Pattanasiri, Y. W. Li, D. P. Landau, T. Wüst, and W. Triampo, in preparation

[142] A. M. Ferrenberg, D. P. Landau, and Y. J. Wong, *Monte Carlo simulations: hidden errors from "good" random number generators*. Phys. Rev. Lett. **69**, 3382 (1992)

[143] W. Janke, *Pseudo random numbers: generation and quality checks.* In Quantum simulations of complex many-body systems: From theory to algorithms, Lecture notes **10**, 447 (2002)

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

[145] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU scientific library reference manual* (Network Theory Ltd., 2009), 3rd edn.

[146] 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)