On the Role of Entropy in the Protein Folding Process

A Thesis

Submitted to the Faculty

of

Drexel University

by

Travis Hoppe

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

May 2011

Dedications

This thesis is dedicated to Cara Hoppe.

I respect her as my peer,

and cherish her as my wife.

Her love and support made this work possible.

Acknowledgments

The completion of this thesis could not have been done without the tireless support of my advisor, Dr. Jian-Min Yuan. I am deeply indebted for the support he has provided over the years. As a mentor, he helped me develop my skills as a scientist and honed my critical thinking. He is an endless source of new ideas and has taught me how to effectively communicate in the scientific world.

I would also like to thank the physics department at Drexel for helping me prepare for the journey ahead. In particular, both Dr. Michel Vallieres and Dr. Robert Gilmore have been gracious enough to spend many afternoons discussing every interesting theory I've come across. Their insights and enthusiasm for physics and mathematics is inspiring.

I would like to thank my family, my wife, Cara Hoppe, and my children, Hazel and Jackson Hoppe. They cheerfully remind me of the world outside and always bring a smile to my face. Finally, I would like to thank my uncle, Fred Stein, who introduced me to physics and all its wonders at an early age.

List of Figures

Table of Contents

Abstract

On the Role of Entropy in the Protein Folding Process
Travis Hoppe
Dr. Jian-Min Yuan

A protein's ultimate function and activity is determined by the unique three-dimensional structure taken by the folding process. Protein malfunction due to misfolding is the culprit of many clinical disorders, such as abnormal protein aggregations. This leads to neurodegenerative disorders like Huntington's and Alzheimer's disease. We focus on a subset of the folding problem, exploring the role and effects of entropy on the process of protein folding. Four major concepts and models are developed and each pertains to a specific aspect of the folding process: entropic forces, conformational states under crowding, aggregation, and macrostate kinetics from microstate trajectories.

The exclusive focus on entropy is well-suited for crowding studies, as many interactions are non-specific. We show how a stabilizing entropic force can arise purely from the motion of crowders in solution. In addition we are able to make a a quantitative prediction of the crowding effect with an implicit crowding approximation using an aspherical scaled-particle theory.

In order to investigate the effects of aggregation, we derive a new operator expansion method to solve the Ising/Potts model with external fields over an arbitrary graph. Here the external fields are representative of the entropic forces. We show that this method reduces the problem of calculating the partition function to the solution of recursion relations.

Many of the methods employed are coarse-grained approximations. As such, it is useful to have a viable method for extracting macrostate information from time series data. We develop a method to cluster the microstates into physically meaningful macrostates by grouping similar relaxation times from a transition matrix.

Overall, the studied topics allow us to understand deeper the complicated process involving proteins.

Preface

Traditionally, the components of biophysics, biology and physics, have a tangential mode of thought. Biology, the study of living things, is a field governed by qualitative observation and later mathematical models are applied. Prediction comes from the understanding of the various mechanisms that are often intricately coupled. On the other hand, a mature branch of physics is governed by the mathematical form first. Ideas are founded from first principles, from which the physical laws are then derived. If these predictions fail to accurately portray reality the fundamental assumptions are discarded and presumed to be false. The successful theories in physics have a far greater precision than the biological counterparts. How then do the two fields reconcile, the dogma of each being so different?

$$\text{bioPHYSICS} \longleftrightarrow \text{BIOphysics}$$

In this authors opinion, the field of biophysics has truly made possible by a third kind of science. After experiment and theory comes modeling, an approach that has been greatly enhanced by the use of modern day computers. Scientific modeling incorporates rigorous mathematical derivations, yet the results are interpreted as though they were experimental evidence. It is rapidly being recognized as an important and vital component to our scientific endeavors.[4] In some ways modeling takes the best of both forms, allowing each to contribute. This approach has been used in many diverse fields such as meteorology, economics, and our present topic, biophysics.

Biophysics has an implicit scale. Usually the scale spans from collections of cells (on order of microns) to collections of atoms (on order of angstroms). As an example of the types of objects currently studied, Figure 1 shows the usage of the words found in abstracts of presentations and posters at a recent biophysics conference. The three dominant objects that were studied were proteins, cells, and membranes. The larger structures (except in the cases of aggregation) traditionally studied in biology are out of the scope of most biophysical studies.

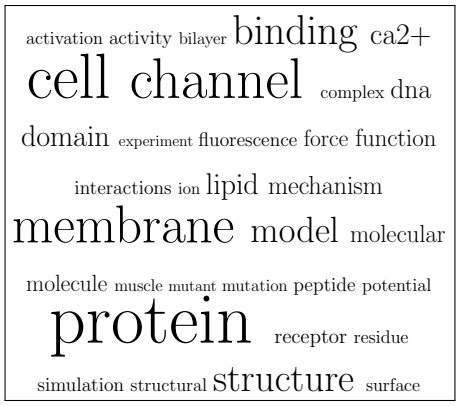


Figure 1: Snapshot of the most commonly used words in the abstracts of the 2010 Biophysical Conference in San Francisco.[1] Larger words correspond to a greater usage. Note that the prominent words, protein, cell, and membrane set the length scale that narrows the focus of the discipline.

As we move to smaller structures we find ourselves in the realm of quantum chemistry and pure physics. In this regime, theories of "first-principle" abound. That is, starting from a set of empirically verified postulates, equations are derived which predict the dynamics and structure of the system. As we move to larger structures we are back in the traditional biological realm, where the role of classification dictates the functionality. For the theoretical and modeling based biophysicist the aim is to derive from first-principle equations. The attempt is to classify and organize the higher-order structures found at this scale. The predictions must ultimately be vetted by experiment studies, thus neither group can work in isolation.

It is here that we begin our discussion, at the crossroads of two diverse fields whose singular aim is no less than the understanding of the phenomena behind life itself.

Chapter 1

Introduction

> *In the drama of life on a molecular scale, proteins are where the action is.*
>
> LESK [5]

In contrast to other important biological molecules, proteins have a remarkable diversity of spatial structures and thus a wide variety of biological functions. Some of the many complex tasks proteins can perform include the ability to [6,7]

- interpret inter-cellular signals such as insulin & EGF in signaling pathways [8,9]
- transport other species [10]
- govern chemical conversion
- control gene expression
- convert chemical energy into mechanical energy
- serve as building blocks to biological structures (i.e. collagen and virus coats) and,
- act as chaperones to help other biological agents function (such as GroEL in bacteria).

As such, the study of proteins, either directly or indirectly, encompasses the core of many microbiological studies. A protein's ultimate function and activity is determined by a process called folding. This folding process guides the protein into a specific three-dimensional form. Unsurprisingly, protein malfunction due to misfolding is the culprit of many clinical disorders, from cancers to abnormal protein aggregations (leading to neurodegenerative disorders like Huntington and Alzheimer [11]).

The study of proteins is paramount to the understanding of life itself. In this thesis we attempt to elucidate some of the properties associated with the folding process. We present a detailed study of the conformational states under various conditions.

1.1 Protein Structure

A protein can be broken down into a hierarchy of structures, each with a different set of motifs describing it. A protein is a polymer linking together a series of amino acids. Like DNA, a protein is composed of a simple alphabet of twenty naturally occurring amino acids (whereas DNA has a four-letter alphabet of base pairs). These amino acids are connected along a backbone of carbon, nitrogen and oxygen atoms. A protein has a primary structure enumerated by the ordered sequence of these letters. These encodings, along with the physiological conditions, uniquely determine the structure of the protein and hence its ultimate function.

Amino acids connected along the backbone cannot assume arbitrary angles with respect to each other. This stereochemistry is vividly displayed on Ramachandran plots.[12] These plots typically illustrate the dihedral (or ϕ-ψ) angles that are allowed due to energetic considerations. This is just one type of the forces which guide the folding process.

At this structural level, certain motifs are seen more often than others. The most frequent structures common to proteins are the α-helices and the β-strands. At this level of hierarchical complexity, the motifs are called the secondary structure. At the next level up, the tertiary structure involves the arrangement of the secondary-structure elements, helix bundles, and β-sheets. These are in turn coalesced into even larger quaternary structures for proteins with multi-domains.

The secondary structures are a good starting point for the study of the protein. In the absence of stabilizing interactions, the protein will be unfolded in a so-called random-coil structure. When hydrogen bonds form between residues along the backbone, the protein will form well-defined (locally periodic) structures of β-sheets and α-helices. Hydrogen-bonds, hydrophobic interactions, van der Walls and electrostatic interactions all contribute to the folded structure of a protein.

Secondary Structures: α-helices and β-sheets

The first of the secondary structures, the α-helix, is stabilized by hydrogen bonds between the CO and HN groups of the backbone four residues down the chain. In theory, there are multiple ways a helix could be formed: both right-handed, left-handed and different 'tightness' of the helix. Almost

all of these other conformations except for the α-helix are energetically unfavorable and are rarely observed in real proteins. In an α-helix all the hydrogen-bonds are made by neighbors that are close along the chain, thus the stabilizing forces involved are often referred to as short-range. These helices are formed very rapidly, on the order of tenths of μs. There are approximately 3.6 residues per helical turn, which from a modeling standpoint, presents a problem if the protein is coarse-grained onto a lattice that can not accommodate the exact structure.

The β-sheet is a lateral packing of β-strands; each strand is a short extended peptide segment. Just like the α-helix, a β-sheet can exist in several different forms, however these other forms are actually found in real proteins. The two main distinctions are parallel and anti-parallel β-sheets. The interactions between the β-stands act over short distances but can be far apart along the length of the chain; they are often referred to as long(er)-range forces when compared to the α-helix. Again, in contrast to α-helices, β-sheets have a wide variability in folding times. Some shorter peptide sequences can form as quickly as $0.5\mu s$, while others can take hours![2]

1.2 Protein folding

Since Anfinsen's discovery in the early 1960's that a denatured protein can spontaneously self-assemble when denaturants are removed,[13] researchers have attempted to deduce which mechanisms and pathways are important to the folding process. One of the main challenges of the field is to predict the final native structure from the primary structure. To put it mildly, this has not been an easy task. Not only does a peptide (small protein) contain hundreds of atoms, it resides in a water based solvent where the effects are mediated out at lengths of several molecules away. A typical molecular dynamics simulation contains thousands of atoms making a full scale quantum computation impossible by todays standards. Classical kinetic approximations are possible, yet consume thousands of hours of computing time for a single trajectory. This is tantamount to a limited observational study; predictions of statistical quantities such as specific heat can only be crudely estimated. These molecular dynamics simulations are, however, the best method for studying dynamics of the system without the need for costly experiments. The mechanism in which protein folding occurs can often be determined out of such *in silico* experiments.

Backing up the veracity of the computational molecular experiments are the physical experiments themselves. The traditional methods of studying an unknown sample such as mass spectrometry, lacked the resolution needed when dealing with complexities of proteins. The more detailed studies of protein structure came from advances in experimental instrumentation. These experimental techniques were sometimes adapted from other fields, but often were invented for the study of biological molecules. X-ray crystallography was created to study the periodic arrangements in crystals but later adapted to study the atomic arrangements of protein structures (for those that could be crystallized). Over the last fifty years, there have been unprecedented experimental advances such as nuclear magnetic resonance (NMR), Förster resonance energy transfer (FRET), circular-dichroism, optical tweezers and atomic force microscopy. All of these techniques contribute to our understanding of the folding process both *in vitro* and in the larger biophysical context.

1.2.1 How can proteins be simplified?

One of the reasons the four levels of structure nomenclatures exist is the inherent simplifications each level implies. For example, at the secondary structure level most α-helices can be expected to have the same energetic and entropic properties. Between the primary structure, the one-dimensional description of the protein as a linear chain of amino acids, and the three-dimensional conformation, there are a host of intermediate coarse-grained models. There is strong motivation for doing such a coarse graining. In a large study of the relevant interaction strengths for each of the twenty amino acids (written in terms of the so-called MJ matrix[14]) the resulting basis set is smaller than twenty. In other words, by diagonalizing the MJ matrix one finds two dominant eigenvectors and three large but slightly smaller eigenvectors. The remaining eigenvalues associated with the other eigenvectors are an order of magnitude smaller than these dominant eigenvectors. This implies that there is a basis for the coarse-graining of the interactions between residues. Indeed, a good approximation to the interactions found in a real protein would be the reduction to only two types, hydrophobic and hydrophilic![15] The remaining interactions can be roughly classified as those belonging to the charged and uncharged types.

1.2.2 Hydrophobicity

Hydrophobicity, literally the fear of water in Greek, plays a central role in protein folding. It was quickly determined that one of the most dominant factors in the folding process was the interaction of certain residues with the solvent. H_2O is an exceptional molecule with an unusually high specific heat compared to other liquids. In addition, its polar nature makes for a non-isotropic fluid with regards to the orientation of the molecules. Residues such as alanine, valine, leucine, isoleucine, phenylalanine, tryptophan and methionine are often found buried in the core of the native state conformation, while the other charged and polar residues are exposed to the solvent.[16] When water encounters a solute it has been known to build hydrogen-bond networks around it.[6] In the case of a protein, it is not altogether clear however if these hydrogen networks are stabilizing or destabilizing. Though early literature often thought it was one of the sole driving forces for folding, depending on the hydrophobicity of a residue it may or may not become part of the hydrogen-bond network. For a comprehensive review of the arguments see Rose.[17]

From an entropic perspective, this presents a facet to the folding problem. A protein will try to minimize the exposure of its hydrophobic amino acids to the solvent. As a first-order effect, hydrophobic resides serve to minimize the accessible surface area.

1.3 Entropy and The Science of Counting

Entropy has a reputation for being an abstract physical concept, mathematically defined as the logarithm of density of states. We will see that a clear intuitive interpretation can be developed using the canonical example of a simple coin. When we flip a coin, the particular trajectory of the flight is irrelevant for the purposes of long-term averages. What matters is the final outcome, heads up or heads down. With one throw, we know very little about the coin, in fact we can't even say with certainty that it will ever land on the other side! However, one of the key assumptions of statistical mechanics (the framework on which many physical ideas of entropy are based) is that of a large sample size. So instead of throwing the coin once, we throw the coin millions[1] of times.

[1] Depending on the type of systems studied, millions can be a gross underestimate. For example, typical calculations that involve macroscopic quantities can easily have 2^{N_A} (where $N_A \approx 10^{23}$) possible states!

Alternatively the conclusions made would be the same if we observed one throw from a million similar coins. Once we've observed this great multitude of throws and tally the results, we can estimate the averages of the system. We call this average the density of states (for the simple coin, it is a count of the frequency of heads and tails). In almost all contexts, the usefulness of the density of states is the same as if it were given in relative proportions. For example instead of our sample experiment with say, 300,000 heads up and 700,000 heads down we could express the density of states as $\{g_{\text{heads}} = 0.3,\ g_{\text{tails}} = 0.7\}$. Entropy then, is the study of the way a system can arrange itself. In short, an entropic theory is a combinatorial one. The formal connection of entropy to statistical mechanics will be derived in detail in Chapter 2.

The underlying subject of this thesis is the study of entropic effects on the protein folding process. In a protein many factors come into play, both energetic and entropic. Entropy may not be the dominant factor in the folding process. In the spirit of teasing out the first principles of protein folding we try to study the entropic effect exclusively. Regardless of the potential, there exist disallowed regions in the conformational space that can be approximated as entropic forces. There are several energetic contributions that effect the free energy of a folded chain. If these effects are simplified to all or nothing (i.e. the black spots on a Ramachandran plot) then they become purely entropic effects. The 'excluded-volume' effect arising from any non-specific hard-core repulsion is another common example of this entropic force.

1.3.1 Levinthal's paradox

It was suggested by Levinthal in 1969 that, by any reasonable measure, the state space for a protein is enormous. There exists however, a unique native state among them which makes folding impossible if the protein were to blindly sample the states.[18] Consider the simplest model, where each amino acid is either in the folded or unfolded states (Levinthal used a more complicated construction, but the result remains the same). In a modest peptide of fifty amino acids this gives 2^{50} different possible microstates. Even if the peptide were to sample 10^{11} of these microstates each second[2] it still would take an impossible amount of time to reach the native conformation. If a protein were to

[2]The time scale for ϕ-ψ rotations is on order of 10^{-11}s

blindly sample the state space this argument might be valid. However, it is very clear (both from experiment, and the mere existence of our biological function) that the folding rate for proteins is much faster than initially suggested by Levinthal's paradox.

It is clear from his original paper where this *gedanken* was proposed that Levinthal did not see this as a paradox at all. In fact, the experimental contradiction to the paradox is what led to a fundamental idea behind protein folding, that of the funneled energy landscape. Levinthal proposed that proteins fold cooperatively, that is, the spontaneous folding happened as a result of guided interactions, much like a rock tumbling down a hill.

This poses an interesting dilemma for the entropic study of protein folding. On one hand, the idea of a funneled energy landscape suggests that a folding pathway captures the essential dynamics of the process. However, it has been shown that landscape itself is not a smooth function, rather it is full of intermediate potential wells and entropic plains that can trap the folding process. As the physiological system's parameters change, such as the temperature or colsolute packing fraction of the system, these intermediates become more numerous and important to the folding process.

1.3.2 Macromolecular Crowding

Observational studies on protein folding have long ignored the true cellular environment, mainly to isolate the effects of a specific interaction or force. However, high concentrations of macromolecules serve to reduce the accessible volume available to the protein folding process. The interactions between the proteins and the crowding agents may be energetic but one often studies the general crowding effect alone due to the non-specific interactions. The reduction in conformational state space then becomes a purely entropic interaction whose contribution has various approximations, such as scaled-particle theory and several integral field theories. By simply confining the protein to hard boundary conditions one can crudely estimate the crowding effect,[19–21] though we have investigated the changes in the conformal space directly using a lattice model and an aspherical scaled-particle theory.[22]

1.4 Chapter Overview

The thesis is broken into four major research chapters, each exploring a different aspect of entropy in the context of protein folding: entropic forces (Ch. 3), conformational states under crowding (Ch. 4), aggregation (Ch. 5), and macrostate kinetics from microstate trajectories (Ch. 6). Along the way we will discuss the implications that our models have on the topics of crowding and protein aggregation. The studies look separately at the solvent, macroscopic crowders and the protein folding pathway itself. A brief summary of each research chapter is given below. In addition, an introduction to the analytical and computational methods is given (Ch. 2).

Entropic forces and Cosolute Flows

In Chapter 3, we consider a simplified treatment of a cosolvent (crowders) acting on a protein. What sets this study apart from others is the prediction of a stabilizing entropic force that arises purely from the *motion* of the cosolvent. It has been long known that an irregular cavity in a fluid will produce an anisotropic distribution of a cosolvent, but we consider under Newtonian assumptions, the averaged velocity profiles as well. The effect of these profiles shows that, in the absence of other effects, irregular conformations are less preferred than their more compact counterparts.

The study of the excluded-volume effects on protein stability and reactions or the stability of colloidal suspensions is an active area of research. Using hard-disc collisional dynamics we investigate whether the presence of a crowding agent can induce a shape change from a non-spherical molecule to a spherical one. We show the averaged density profiles and velocity field of hard-disc crowders with an interior non-circular convex shape as a boundary condition. The density profile is not axially symmetric, consistent with other hard-potential experiments with asymmetry. However, the averaged velocity field was found to have a non-zero curl, implying a region of vorticity without a thermal gradient, advective field or other motivating potential. To explain the occurrence of the vortices, a theoretical model is provided based on the conservation of angular momentum of hard discs at contact. All these results, as well as difference in pressure along the axes, support the fact that as the packing fraction of the crowder rises, increasing force is exerted on an asymmetric

molecule toward a symmetric one.

Conformational States Under Crowding

In Chapter 4 we introduce the idea of the implicit crowding method to study the statistical mechanical behaviors of folding of β-sheet peptides. Using a simple bead-lattice model we are able to consider, separately, the conformational entropy involving the bond angles along the backbone and the orientational entropy associated with the dihedral angles. We use an Ising-like model to partially account for the dihedral angle entropy and implicitly, the hydrogen-bond formations. We also compare our results to recent experiments and find good quantitative agreement on the predicted folded fraction. Based on the predictions from the scaled particle theory we investigate changes in the melting temperature of the protein, suggesting crowding enhanced stability for a variant of trpzip hairpin and a slight instability for the larger β-sheet designed peptides.

Potts Aggregation Models

In Chapter 5 we examine the protein at a different level of complexity. We move from the scale of a single protein to that of an aggregated species through several models. The study is primarily motivated by disorders thought to be caused by aggregated species, such as Alzheimer's disease. We consider a simplified first-order model of cluster (oligomer) growth. The model is sufficient only as coarse approximation, so we develop a methodology to solve a general graph oriented aggregation model. In this chapter we present a new operator expansion method to solve the Ising/Potts model with external fields over an arbitrary graph. This method allows us to present new results on the Potts model problem: a general expansion of a one-dimensional lattice along with suggestions to a generalization of higher dimension. In addition, we solve a wider class of problems with a subgraph recursion relation, with the focus on lattice strips, graphs that are finite in one direction and grow in another.

Clustering and Kinetics

In Chapter 6 we extract macrostate information from the density of states and time series data. We develop a method to cluster the microstates into physically meaningful macrostates. The conforma-

tions are grouped by similar relaxation times from a transition matrix. The method is applied to two very different systems, a frustrated Langevin walk and a lattice model of a β-hairpin. In the random walk we are able to reduce the conformational state space from a continuous two-dimensional potential to a simple linear model with four discrete states. In the β-hairpin model, we are able to extract the folding pathway and show that the defining kinetic pathway is the formation of the turn on the hairpin.

Chapter 2

Computational Methods

In physics, the ultimate goal is to derive a solution from fundamental postulates (first-principles) whose predictions match with empirical observations. However, what it means to 'solve' a system, depends on the techniques applied. As an example, shown in Table 2.1 are some canonical variables sought after in various sub-fields of physics. While the list is necessarily incomplete and inherently

Table 2.1

Classical Mechanics	$(\mathbf{q}, \mathbf{p})(t)$	Positions and momenta as functions of time
Quantum Mechanics	ψ	Wave-functions
Fluid Dynamics	$\mathbf{v}$	Velocity fields of the flow
Chaos Theory	λ, D_α	Lyapunov exponents, fractal dimension
Statistical Mechanics	$\mathcal{Z}$, Ω	Partition functions, density of states

subjective, it illustrates how different each solution can be. Each subfield has its own approximations and limits; it is important that the predictions made are understood within the correct scope. Hence we want to be clear what limits we wish to impose in our present discussion. In the context of this thesis, we concentrate mainly on equilibrium distributions from statistical mechanics and dynamics (kinetics) derived from those results.

The foundations of Statistical Mechanics historically come from thermodynamics, the study of temperature and heat. Entire treatises, books and journals are dedicated to the topic, for the reader interested in a historical background see Lewis.[23] There are three (+1) laws of thermodynamics, the zeroth law so named for the connection to commutativity axiom.

- Zeroth Law : If system $\mathcal{B}$ is in thermal contact and equilibrium with systems $\mathcal{A}$ and $\mathcal{C}$ then $\mathcal{A}$ must be in thermal equilibrium with $\mathcal{C}$.
- First Law : If system $\mathcal{A}$ is isolated, then total energy remains constant. It can however, change forms.
- Second Law : In a closed system, the result of any irreversible process must increase entropy, and remain constant for a reversible process; $S \geq 0$
- Third Law : The entropy of a system approaches zero as the absolute temperature goes to zero.

The Laws of Thermodynamics

The primary consideration of this thesis is the study of entropy involved in the protein folding process. Underpinning all of the discussions in the subsequent chapters is the idea of a canonical ensemble from statistical mechanics. As such, we feel that its derivation from the Hamiltonian under suitable restrictions, is paramount to the ideal of a complete and self-contained thesis. Much of the material in the Section 2.1 is adapted from several of the notable statistical mechanics texts, namely Pathria,[24] Kittel[25] and Dill.[26]

2.1 Statistical Mechanics

Statistical mechanics is a more abstract representation of a system then the physical quantities considered in thermodynamics. Indeed the notion of temperature and states do not have to correspond to directly physically observable quantities (c.f. Shannon's information theory[27]). With the advent of statistical mechanics the three laws of thermodynamics can be formulated from a more fundamental postulate, from which some of the above laws of thermodynamics can be derived.

- Given an isolated system in equilibrium with a finite number of microstates Ω of equal energy, the probability of finding the system in any particular microstate is $1/\Omega$.

THE FUNDAMENTAL POSTULATE OF STATISTICAL MECHANICS

We begin with the Hamiltonian, arguably the most fundamental quantity in all of physics. Let us consider N particles and their instantaneous position $q_i(t)$ and momenta $p_i(t)$ where $i = 1 \dots Nd$ and d is the dimension of the system. We refer to these coordinates $(\mathbf{q}, \mathbf{p}) = (q_1, q_2, \dots, q_{Nd}, p_1, p_2, \dots, p_{Nd})$ as a microstate of the system. The Hamiltonian $\mathcal{H}(\mathbf{q}, \mathbf{p})$ gives a set of d coupled differential equations that govern the dynamical behavior of the system. We refer to this d-dimensional space of coordinates $(\mathbf{q}, \mathbf{p})$ as the phase space of the system. The evolution of the system through phase space is given by the canonical relations

$$\dot{q}_i = \frac{\partial \mathcal{H}(\mathbf{q}, \mathbf{p})}{\partial p_i} \tag{2.1}$$

$$\dot{p}_i = -\frac{\partial \mathcal{H}(\mathbf{q}, \mathbf{p})}{\partial q_i} \tag{2.2}$$

It is worth noting at this point that, if the system is known to have a fixed energy

$$\mathcal{H}(\mathbf{q}, \mathbf{p}) = E \tag{2.3}$$

then the system will trace out a trajectory in phase space that is restricted to a hypersurface of dimension $d - 1$ (if there are no other constants of motion). The phase space splits naturally into two regions, allowed and disallowed, whose particulars depend on the form of the Hamiltonian and the initial conditions. Considering all points in the allowed region, we can form a (possibly time-dependent) density function $\rho(\mathbf{q}, \mathbf{p}; t)$ which determines the number of microstates are found in a given volume of phase space

$$\rho(\mathbf{q}, \mathbf{p}; t)\, d\mathbf{q}\, d\mathbf{p} \tag{2.4}$$

In the context of this thesis, we will consider only density functions that are stationary

$$\frac{\partial \rho(\mathbf{q}, \mathbf{p})}{\partial t} = 0 \tag{2.5}$$

which correspond to systems in equilibrium. Under the assumptions of equilibrium, powerful results can be derived. Since the phase space does not admit "sources" or "sinks", i.e. the total number of points must be conserved, the total inflow and outflows must be the same over time. Letting ω represent an arbitrary volume in phase space and $\mathbf{v}$ be the velocity vector of the points in phase space we get

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{v} = 0 \tag{2.6}$$

which is simply the continuity equation of the system. This is a conservation of the phase space volume, but the generic form is common throughout physics (for example, a similar expression can be found implicitly in Maxwell's equations (c.f. Jackson[28] page 3) for the continuity of charge and current density). Manipulating them further, one can get Louisville's theorem[29]

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + [\rho, H] \tag{2.7}$$

where the Poisson bracket $[\rho, H]$ is defined as

$$[\rho, H] = \sum_{i=1}^{d} \frac{\partial \rho}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial \rho}{\partial p_i} \frac{\partial H}{\partial q_i} \tag{2.8}$$

According to this theorem, if one could travel along with a group of points, one would find that the local volume of points remains constant. Using the equilibrium condition from Equation 2.5 we have the restriction

$$[\rho, H] = 0. \tag{2.9}$$

Ensembles

There is more then one way to satisfy the restriction in Equation 2.9, the choice of which leads to the various statistical ensembles. It is convenient to define a macrostate which is a collection microstates that share common aggregate values. We let the particle number N, volume V and energy E of a system define a macrostate by the tuple (N, V, E), with the number of available microstates in a given macrostate as $\Omega(N, V, E)$.

Placing two systems $\Omega_1(N_1, V_1, E_1)$ and $\Omega_2(N_2, V_2, E_2)$ with a hard but thermally conductive barrier at thermal equilibrium would settle to a composite state Ω_3. In full generality we can set $E_1 = 0$ as a reference point and $E_2 > 0$. Conservation laws tell us that the total volume, particle number and energy are simply $V_3 = V_1 + V_2$, $N_3 = N_1 + N_2$, $E_3 = E_1 + E_2$ but the question remains, how much energy is transferred from the second system to the first? The second law of thermodynamics tells us how this is done. The final number of states $\Omega_3(N_3, V_3, (E_1, E_3))$ is maximized. The rationale behind such an assumption is generally considered valid as, from a purely probabilistic standpoint, the most likely state must be the most common. In addition, the next most probable state is often orders of magnitude less likely then the first, making this specific choice of Ω_3 to correspond to the state at which the system spends "most" of its time in. It is natural then, to identify this state with the equilibrium of the system. We find that the subsystems exchange energy until this point is reached, and hence by the zeroth law of thermodynamics, they must share a common parameter (which we call β for now)

$$\beta \equiv \left(\frac{\partial \ln \Omega(N, V, E)}{\partial E} \right)_{N,V,E=E_{\text{eq}}} \tag{2.10}$$

From the thermodynamic relation

$$\left(\frac{\partial S}{\partial E} \right)_{N,V} = \frac{1}{T} \tag{2.11}$$

and Equation 2.10 we find that

$$\frac{\partial S}{\partial \ln \Omega} = \frac{1}{\beta T} = \text{const.} \tag{2.12}$$

This is identified as the Boltzmann constant, which is related to the gas constant R and Avogadro's number N_A

$$k_B = R/N_A = 1.3806503 \cdot 10^{-23} \frac{\mathrm{m}^2\mathrm{kg}}{\mathrm{s}^2\mathrm{K}} \tag{2.13}$$

This implies the famous relation first discovered by Boltzmann

$$S = k \ln \Omega \tag{2.14}$$

a relation between the entropy and the number of states of the system.

We consider first the microcanonical ensemble, where the particle number N, volume V and energy E are all fixed quantities,[1] of all the ensembles considered this is the most restrictive. In the microcanonical ensemble all states in phase space are equally likely, such that Equation 2.9 is satisfied by

$$\rho(\mathbf{q}, \mathbf{p}) = \text{const.} \tag{2.15}$$

This postulate is commonly referred to as "equal *a priori* probabilities". The equilibrium condition can be satisfied in a more general way, such that the dependence of ρ depends only explicitly on the Hamiltonian. The more general restriction for a stationary ensemble provides a class of density functions. We will see that the class where $\rho(\mathbf{q}, \mathbf{p}) \propto \exp(cH(\mathbf{q}, \mathbf{p}))$ is of particular importance.

The microcanonical ensemble becomes unsatisfactory when considered from a realistic experimental standpoint. For one, it would be hard to keep a real experiment at constant energy. In addition there are few realistic systems in which one can measure the energy *directly*. A more common scenario is to imagine our system of interest to be kept at constant temperature by coupling it with a large reservoir. Let the large reservoir be system A, and the smaller system of interest be system B with macrostate parameters (E_A, T) and (E_B, T) respectively. The energy of the smaller system would be in flux over time and could range from zero to the composite energy of the system $E_{AB} = E_A + E_B = \text{const}$. However, since the reservoir system contains many more states then the

[1] One could also consider the case when the energy lies within a fixed *range* instead $E_0 - \Delta \leq E \leq E_0 + \Delta$, but the same difficulties will still arise in this formulation.

smaller one, and that probability of each microstate at a fixed energy is equal, the probability of any given energy value of the smaller system must be proportional to

$$P(E_B) \propto \Omega(E_A) \equiv \Omega(E_{AB} - E_B) \tag{2.16}$$

We can carry out a Taylor series expansion of the above around the value of $E_A = E_{AB}$ i.e. $E_B = 0$. For reason of convergence we do so around the logarithm instead, giving

$$\ln \Omega(E_A) = \ln \Omega(E_{AB}) + \left(\frac{\partial \ln \Omega}{\partial E'} \right)_{E'=E_{AB}} (E_A - E_{AB}) + \dots \tag{2.17}$$

At equilibrium and from Equation 2.10 we get the fundamental relation for the canonical ensemble

$$P(E_B) \propto \exp(-\beta E_B) \tag{2.18}$$

To normalize Equation 2.18 we simply divide by all possible energy levels. This 'sum over all states function', plays particular importance in studies carried out in this thesis as it essentially (along with the density of states) provides all the thermodynamic information of the system. In language of equilibrium thermodynamics, it is known as the partition function $\mathcal{Z}$ (German: *Zustandssumme*). The partition function then, is the sum of all the events that could occur weighted with their probability over the canonical ensemble. If, as in many physical cases, the energy levels of the system are degenerate, the sum is modified by this multiplicative factor as well. If we let $g(E_i)$ be the degeneracy of energy level E_i one has

$$\mathcal{Z} = \sum_{E_i} g(E_i) \exp(-\beta E_i) \tag{2.19}$$

From this we state the general thermodynamic relations from the canonical ensemble (see Pathria and Beale [24], page 53). With P, U, S, T, $A = U - TS$ as the probability, internal energy, entropy,

temperature and Gibbs free energy respectively we have

$$P(E_i) = g(E_i)\exp(-\beta E_i)/\mathcal{Z} \tag{2.20}$$

$$U = \sum_{E_j} E_j g(E_i)\exp(-\beta E_j)/\mathcal{Z} \tag{2.21}$$

$$A = -kT\ln\mathcal{Z} \tag{2.22}$$

$$S = -k\sum_{E_j} P(E_j)\ln P(E_j) \tag{2.23}$$

The last relation is the most interesting. The implication is that, if we had complete knowledge of the energy level degeneracies $g(E)$, all thermodynamic variables could be computed. The function $g(E)$, up to a normalization constant is known as the density of states. We shall see in Section 2.2.2 that specialized numerical algorithms have been devised that determine this function to high accuracy.

There are physically relevant limits at which the system reduces to particularly simple forms. At high enough temperatures we find that the steady-state distribution is simply the sum over the degeneracy term, $g(E)$

$$\lim_{T\to\infty}\mathcal{Z} = \sum_{E_i} g(E_i) \tag{2.24}$$

and at low temperatures the system is dominated by the ground state

$$\lim_{T\to 0}\mathcal{Z} = g(E_1)\exp(-\beta E_1) \tag{2.25}$$

.

For completeness, we mention a third ensemble, the grand canonical ensemble. In this ensemble, both the energy and the particle numbers are allowed to be exchanged from the reservoir to the smaller system. In this case both the temperature T and the chemical potential μ are fixed giving the equilibrium distribution distribution relation as

$$P(E_B, N_B) \propto \exp(-\alpha N_B - \beta E_B) \tag{2.26}$$

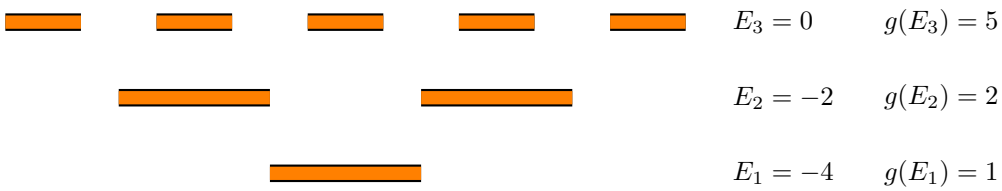


Figure 2.1: Sample three-level system with degeneracy levels shown.

where

$$\alpha = -\frac{\mu}{kT} \tag{2.27}$$

The fugacity of the system is $z = \exp(-\alpha)$. While not explicitly used in this thesis, many of the models presented have obvious extensions to allow a variable particle (or in our case protein) count. A sampling method to determine the density of states would have to count system conformations $g(E_i, N_j)$ but the extension is natural.

Example three-level system

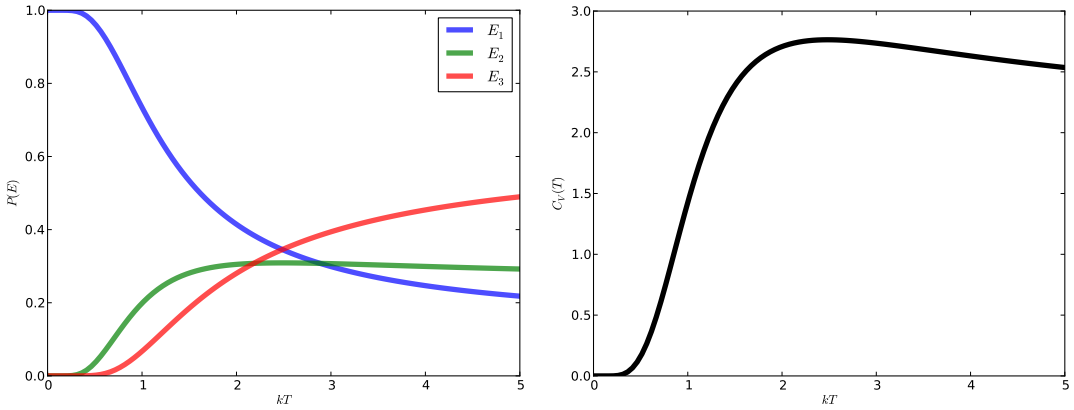
To demonstrate the concepts, we will use a sample system with the energy levels shown in Figure 2.1. In this system there are three discrete energy levels ($E_1 = -4$, $E_2 = -2$, $E_3 = 0$) with a state degeneracy of $g(E_1) = 1$, $g(E_2) = 2$, $g(E_3) = 5$. The partition function is

$$\mathcal{Z} = e^{4\beta} + 2e^{2\beta} + 5 \tag{2.28}$$

Giving the probability of each state

$$\begin{aligned} P(E_1) &= \frac{e^{4\beta}}{\mathcal{Z}} \\ P(E_2) &= \frac{2e^{2\beta}}{\mathcal{Z}} \\ P(E_3) &= \frac{5e^{0\cdot\beta}}{\mathcal{Z}} = \frac{5}{\mathcal{Z}} \end{aligned} \tag{2.29}$$

A plot of the probabilities and the corresponding specific heat in units of $1/\beta = kT$ are shown in Figure 2.2. The peak in the specific heat is a nice example of the so-called Schottky anomaly,[30] as



(a) $P(E)$ **(b)** C_V

Figure 2.2: Probabilities for each state of the sample three-level system (see Figure 2.1) and the specific heat as a function of $1/\beta = kT$.

normally the specific heat is monotonically increasing with temperature. In this case however, the peak corresponds to a loose 'melting' of the system, when the system undergoes a transition from a full occupancy of the ground-state to the general disordered state found at high temperatures. In this case, since the number of energy levels are very small, the effect is small and correspondingly the peak is quite broad. We will see more physically relevant examples with sharper peaks in the later chapters when we consider larger systems.

2.2 Monte Carlo Methods

In statistical mechanics we are not integrating a set of differential equations (as one would in the Classical or Quantum), but determining the distribution that satisfies the steady-state properties of a given Hamiltonian at a specific temperature. If the system is small and enumerable, such as the one we encountered in Section 2.1, we can evaluate the exact form of $\mathcal{Z}$. If the system displays self-similarity such as the Ising/Potts models 1D, 2D or over Cayley trees sometimes an analytical solution can be found.[31] In general however, the size state-space is too large and complex (and in the continuous case, too rough to discretize at a reasonable level) to completely solve for the states of the system. In general, one must use sampling methods to solve for the distribution. The etymology of the phrase 'Monte Carlo' connects the random sampling of a physical system with the games of

chance (whose profits also rely on random distributions) in the celebrated city of the same name in Monaco.

2.2.1 Metropolis-Hastings

While many different adaptations of a Monte Carlo exist, the method most often used in physics simulations is the Metropolis-Hastings.[32] The Metropolis-Hastings algorithm is essentially a Markov transition matrix (a concept fully explained in Section 2.4). For each pair of valid conformations (ξ_A, ξ_B) of the system (which need not be finite), there exists a corresponding matrix element that represents the transition probability. Not every state has to be connected to every other state, the only requirement is that connections between the conformations are ergodic, in that eventually every state will be reached from every other state. There exists a function called the move set which relates, from a given state, the accessible states to it.

The full Metropolis Hastings algorithm is a more general method of sampling then will be presented here, which will be confined to the simulation of physically relevant numerical experiments. From the canonical ensemble, the probability of observing a single conformation is $P(\xi_A) = \exp(-\beta E_A)/\mathcal{Z}$. When considering the ratio of two states, the partition function cancels leaving

$$P(\xi_A \to \xi_B) = \frac{\exp(-\beta E_A)}{\exp(-\beta E_B)} = \exp(-\beta \Delta E) \tag{2.30}$$

with $\Delta E = E_A - E_B$. Hence, by using a move set function and starting from an initial conformation ξ_0 one proceeds to sample the state space by accepting or rejecting new conformations based on the probability in Equation 2.30.

This method, and others like it, generate a canonical distribution at a fixed temperature. The primary disadvantage is the lack of data over a significant range of temperatures, making multiple simulations necessary for the calculation of thermodynamic quantities. At low temperatures the algorithm can get stuck in local minima, severely impacting convergence rates. Various methods have been proposed to counteract this shortcoming, simulated annealing[33] being the most popular. In addition, near phase transitions the system can exhibit a critical slowing down, making the

convergence extremely slow.[34]

2.2.2 The Wang-Landau Density of States Method

Wang-Landau (WL) sampling[35] is a generic algorithm to calculate the relative density of states (DOS) for a given system. The algorithm starts off with the initial *a priori ansatz* that all conformational states are equally likely $\Omega(\xi) = 1$, where ξ is a conformation of the system. Traditionally the calculation of the density of states was computed as a function of the energy of the system. Like others, we extend the use the Wang-Landau method to determine the the density of states for a set of conformations of the system rather than the energy explicitly.[36,37] This allows the system to be sampled for all values of the free variables simultaneously as the conformations can be turned into energies later.

In the WL method, the density of states is iteratively refined, crudely at first to ensure a large sampling, and then with greater precision as Ω converges. Similar to a typical Monte-Carlo simulation the algorithm has an acceptance rate. However, unlike traditional Metropolis-Hastings simulations, the process can not be modeled as a Markov chain since the transition matrix itself is iteratively refined. The WL acceptance rate is

$$P(\xi_A \to \xi_B) = \min\left(1, \frac{\Omega(\xi_A)}{\Omega(\xi_B)} \frac{n_{B\to A}/n_B}{n_{A\to B}/n_A}\right) \tag{2.31}$$

Where n_A is the number of outgoing moves from states A and $n_{A\to B}$ is the number of moves from A to B (similarly defined for n_B and $n_{B\to A}$). If the moves are reversible then $\frac{n_{B\to A}}{n_{A\to B}} = 1$. These factors are necessary for detailed balance if the move set chosen has a variable number of moves from each state. Once any particular state had been selected the density of states is modified by $\Omega(\xi) \to f\Omega(\xi)$ where f is a constant that is slowly reduced to unity during the simulation.

It is important to note that the process is iterative, which has its advantages and disadvantages. Since the process uses the previous value of the density of states to determine the next refinement, excellent speedups can be obtained if one has any prior information about the distribution of the conformational degeneracies of the system. This may come from a previous simulation or general

knowledge (for example, the 2D Ising model is roughly quadratic in log space). However, unlike the traditional Metropolis-Hastings models, the iterative refinement of the density states makes the process unable to be mapped to a Markov chain. This severely limits the analytic treatment that can be applied when discussing the convergence rates and saturation of errors. Several adaptions have been proposed to improve the algorithm. One can minimize the saturation of errors using an N-fold way process[38] with a steadily decreasing constant. One can also modify the move set in an optimal way to improve convergence.[39]

If the state space is not discrete, additional considerations have to be made. For the algorithm to generate a histogram, the continuous state space has to be discretized. Without prior knowledge of the system this can be difficult. If the level spacing is to small then the algorithm may fail to converge in a reasonable amount of time. If the level spacing is too large it may obscure the results, especially where the derivative of the density of states is large. To partially mediate the problem, we provide an approximation to the specific heat (a typically sought after parameter) in Appendix A.

2.3 Networks and Graphs

This section introduces the idea of a graph. A graph G is a collection of vertices $\{v_1, v_2, v_3, ...\} \in \mathcal{V}$ and edges $\{e_1, e_2, e_3, ...\} \in \mathcal{E}$. An edge is defined by the two vertices connecting to it, $e_i = (v_a, v_b)$. If the graph is directed then the ordering matters, i.e. the edge (n_a, n_b) points only from n_a to n_b. If the graph is undirected then the an edge is considered to point both ways. The edge set $\mathcal{E}$ for a graph contains only unique entries, as opposed to multi-graphs where this restriction is relaxed. Edges can carry a weight which we associate with each edge $\{e_1 : w_1, ..., e_i : w_2\}$. A vertex can also have a weight, and is associated the same way, $\{v_1 : n_1, ..., v_i : n_2\}$. If a graph has an edge connection to itself $e_i = (n_j, n_j)$ this edge is called a loop. Undirected, loop-free graphs are called simple graphs.

A useful device for keeping track of graphs is the adjacency matrix. In this matrix, the non-zero

entries are those where an edge connects vertices v_i to v_j

$$\mathbf{A}_{ij} = \begin{cases} 1 & : \quad e_k = (v_i, v_j) \in \mathcal{E} \\ 0 & : \quad \text{otherwise} \end{cases}$$

As an example, see Figure 2.3, for a simple graph whose adjacency matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \tag{2.32}$$

The adjacency representation is computationally useful, with it various graph properties can be

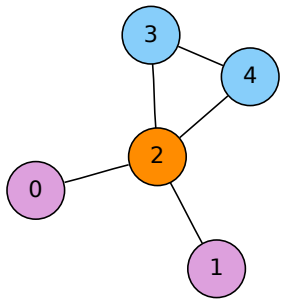


Figure 2.3: Pictorial representation of the graph given by the adjacency matrix, Equation 2.32. The vertices are colored according to the walking polynomial defined in Equation 2.34.

calculated with ease. For example, one can show that $(\mathbf{A}^2)_{ij}$ is the number of paths from v_i to v_j in exactly two steps. This holds in general, $(\mathbf{A}^n)_{ij}$ is the number of paths from v_i to v_j in exactly n steps. This connection between repeated iterations of multiplication makes simple graphs a natural study in the context of Markov matrices.

2.3.1 Graph Operations

We will find it useful to perform two specific operations on the structure of the graph itself. The first, *edge removal*, is simple. We define the operation $G_{-e_{ij}}$ to be the graph constructed from the edge set of G with the edge (v_i, v_j) removed. It is worthwhile to note that edge removal may split the graph into two disconnected subgraphs. If this happens we denote the subgraphs G_1 and G_2 as $G_{-e_{ij}} = G_1 \sqcup G_2$. The second operation, *edge contraction*, is denoted by $G_{/e_{ij}}$. In edge contraction vertices v_i, v_j are merged into a new vertex v_k. All vertices that previously joined to v_i or v_j are now joined to the new vertex v_k. This operation can turn a simple graph into a non-simple one due to the presence of multiple edges. If this is the case, then the graph lacks a nice representation of an adjacency matrix (but it still can be referred to through its edge set $\mathcal{E}$). Representations of the two operations are shown in Figure 2.4.

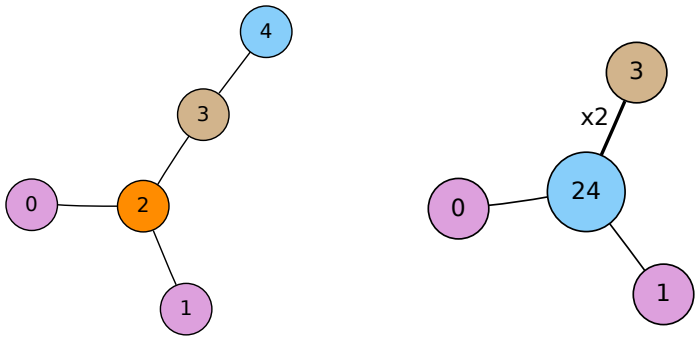


Figure 2.4: Examples of edge removal (left) and edge contraction (right) of the edge (v_2, v_4) on the graph shown in Figure 2.3. Each operation removes exactly one edge. In this case edge contraction has made the graph non-simple due to the presence of a multi-edge (indicated by the x2.

2.3.2 Graph Isomorphism

The ability to quickly determine if two graphs are isomorphic has attracted the attention of physicists, chemists, computer scientists and mathematicians for years. It is a long-standing open problem in both pure mathematics and its application. Graph structures are a common theme throughout this thesis, making the study of graph isomorphism relevant to many of the computations. For an

approximate answer the problem is considered solved for most cases with the software *nauty*.[40] On a purely theoretical level, the exact solution is unique in terms of computational complexity. It is proposed that the problem is outside of the $\mathcal{P}$ vs $\mathcal{NP}$ class.[41] Despite the growing number of invariants proposed, the graph isomorphism problem remains unsolved. Its ability to occupy the scientific world has been enough to label it as a plague on the community.[42]

Nevertheless, we propose another invariant, the walking polynomials, and do so with three goals in mind. The first is to expand upon the arsenal graph theorists have to determine basic properties of a particular graph. The second is the identification that the walking polynomials lend themselves naturally to a coloring scheme that allows the viewer to quickly identify the symmetries of the graph. Finally it is conjectured that the walking polynomials serve as a polynomial-time test to determine if two graphs are isomorphic.

The motivation for the algorithm comes naturally from the study of diffusive processes. The finite limit of a diffusive process is a collection of random walks along a graph. We can count (and solve) for a closed form expression for the number of closed loops along a graph starting at any particular vertex. A generating function can be constructed, one that encodes the path length of *all* walks. It is with these generating functions we construct the invariant known as the walking polynomial.

Walking Polynomials

Given a graph with its associated adjacency matrix $\mathbf{A}$ define $\mathbf{S} = (\mathbf{I} - \mathbf{A}z)$ and the walking polynomial matrix

$$\mathbf{W}(z) = \mathbf{S}^{-1} = (\mathbf{I} - \mathbf{A}z)^{-1} \tag{2.33}$$

Which can be computed without taking the inverse via Cramer's rule

$$\mathbf{W}_{ij} = \frac{\text{minor}_{ij}\mathbf{S}}{\det(\mathbf{S})} \tag{2.34}$$

The set of diagonal elements $\{\mathbf{W}_{11}, \mathbf{W}_{22}, \ldots, \mathbf{W}_{nn}\}$ represent the generating functions[2] for the number of return paths to that vertex. Let an ordered set of these diagonal elements be called $\mathcal{W}$. The choice of ordering is irrelevant, only that it unambiguously sorts rational polynomials. We refer to $\mathcal{W}$ as the walking polynomial for short. We claim that $\mathcal{W}$ is a polynomial time computable invariant.

Take two graphs G, G' with adjacency matrices $\mathbf{A}$, $\mathbf{A}'$, the matrices $\mathbf{S} = \mathbf{I} - \mathbf{A}z$, $\mathbf{S}' = \mathbf{I} - \mathbf{A}'z$ and their walking polynomial invariants $\mathcal{W}$, $\mathcal{W}'$. We say that two walking polynomials are equivalent $\mathcal{W} \sim \mathcal{W}'$ if and only if for every $w_i \in \mathcal{W}, v_i \in \mathcal{W}$ we have $w_i = v_i$. By definition, graphs G and G' are isomorphic if and only if there exists a permutation matrix $\mathbf{P}$ such that

$$\mathbf{P}\mathbf{A}\mathbf{P}^{-1} = \mathbf{A}' \tag{2.36}$$

Proof:

If G and G' are isomorphic then $\mathcal{W} \sim \mathcal{W}'$. Since G and G' are isomorphic let $\mathbf{P}$ be the permutation matrix $\mathbf{P}\mathbf{A}\mathbf{P}^{-1} = \mathbf{A}'$ and $\sigma(\mathbf{P}) \in \{-1, 1\}$ be the the parity of the permutation. The denominator of each of the terms in $\mathcal{W}, \mathcal{W}'$ are related by

$$\det(\mathbf{S}) = (\mathbf{P}) \det(\mathbf{S}') \tag{2.37}$$

Given that the permutation matrix maps the numbers $(1, 2, \ldots, N) \to (k_1, k_2, \ldots, k_N)$ we have

$$\text{minor}_{i,i}(\mathbf{S}) = \sigma(\mathbf{P}) \text{minor}_{k_i,k_i}(\mathbf{S}') \tag{2.38}$$

Thus for each element $w \in \mathcal{W}$ there must be an element $w' \in \mathcal{W}'$ such that $w = w'$ and hence $\mathcal{W}$ is

[2] It is often the generating functions themselves one is interested in, not any particular term. However, one can extract the n^{th} term from the infinite series by differentiating n times to the desired term and setting all higher powers to zero.

$$f_{ij}[n](z) = \frac{1}{n!}\left[\frac{\partial^n}{\partial z^n} f_{ij}(z)\right]_{z=0} = \frac{1}{n!}\left[\frac{\partial^n}{\partial z^n} \frac{p_{ij}(z)}{q(z)}\right]_{z=0} \tag{2.35}$$

If the generating function is a rational function and a closed form is sought, one can perform a partial fraction decomposition. Usually, known relations can reduce the solution to a closed form from this point. In addition, this exposes the poles for asymptotic analysis.

a graph invariant. □

2.4 Markov Matrix

A Markov matrix $\mathbf{M}$ is a matrix where each element $\mathbf{M}_{ij}$ describes the probability of a transition from $i \to j$ in a single time step. As such each row represents a probability distribution

$$\sum_j \mathbf{M}_{ij} = 1 \tag{2.39}$$

$$0 \leq \mathbf{M}_{ij} \leq 1 \tag{2.40}$$

We say that a matrix is positive if all matrix elements are strictly greater then zero ($\forall i, j \quad \mathbf{A}_{ij} > 0$). A matrix is called primitive if there is a $k > 0$ such that $\mathbf{A}^k$ is a positive matrix. Physically, this implies that the matrix becomes well-mixed, and as we will see shortly, allows for the definition of a unique steady-state. If we view the Markov matrix as a graph whose edges are are weighted according to the matrix entries, we see that a disconnected graph implies a non-positive matrix. This typically, is not a problem as the matrix can be broken into separate subgraphs whose solutions are independent. There are however, insidious pathological matrices such as

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{2.41}$$

Here we see that $\mathbf{P}^k = \mathbf{P}$ for all odd k, implying that an initial condition in either state will elastically jump and never mix. Connected matrices such as $\mathbf{P}$ are rare in practice however, since a tiny random perturbation to each of the matrix elements $\mathbf{P}_{ij} \to \mathbf{P}_{ij} + \epsilon_{ij}$ (that preserves the row sums value of unity and keeps all elements non-negative) will force the matrix to become primitive.

Powers of the matrix describe longer jumps, i.e. $(\mathbf{M}^5)_{ij}$ describes the probability of ending up at state j starting at state i after five time steps. The process can be continued indefinitely, but for any primitive matrix, all initial conditions will converge to the same steady state vector. The largest eigenvalue of a primitive Markov matrix is unique and equal to unity, a consequence of the

Perron-Frobenius theorem.[43] Due to the restriction on the row sums, all other eigenvalues are real. The associated eigenvector is then the unique steady-state vector. Intuitively, with only knowledge of the spectra of the eigenvalues, one can deduce this must be so. A large matrix power can be Schur decomposed into a matrix of eigenvectors $\mathbf{V}$ and a diagonal matrix $\mathbf{\Lambda}$ of the eigenvalues by $\mathbf{M}^k = \mathbf{V}\mathbf{\Lambda}^k\mathbf{V}^{-1}$. Since, for diagonal matrices we have $(\mathbf{\Lambda}^k)_{ij} = (\mathbf{\Lambda}_{ij})^k$ the appearance of a unique steady-state vector is consequence of the fact that the smaller eigenmodes will decay with time, leaving only the dominant eigenmode.

If the underlying graph of the matrix is disconnected, there will be a multiplicity k in this unit eigenvalue, where k is the number of disconnected pieces in the graph. Since the graph can be permuted into block diagonal form (with k blocks), the degeneracy in the unit eigenvalue corresponds to the unique solution of each separate subgraph.

As an example, let's put weights on our graph (Figure 2.3) from the previous section. We will use this matrix as an example to explain the concepts discussed below. To make matters simple, let's assume the probability of leaving (or staying) at any particular vertex is equiproportional, giving us the Markov matrix

$$\mathbf{M} = \begin{bmatrix} 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix} \tag{2.42}$$

As stated, all possible states of the system can be expressed by powers of this matrix. It is useful however to specify a specific state of the system, given as a distribution of over the vertices. For example, consider the initial condition such that all states are localized on vertex v_1 (given by the second row of the matrix). We would represent this condition by the vector

$$\mathbf{v}_0 = [0, 1, 0, 0, 0]^T \tag{2.43}$$

One iteration of the Markov matrix with this initial condition would give

$$\mathbf{M}\mathbf{v}_0 = \mathbf{v}_1 = [0, 1/2, 1/2, 0, 0]^T \tag{2.44}$$

And subsequent iterations of the matrix give

$$\mathbf{M}^2\mathbf{v}_0 = \mathbf{v}_2 = [1/10, 7/20, 7/20, 1/10, 1/10]^T \tag{2.45}$$

$$\mathbf{M}^3\mathbf{v}_0 = \mathbf{v}_3 = [3/25, 49/200, 217/600, 41/300, 41/300]^T \tag{2.46}$$

$$\mathbf{M}^{50}\mathbf{v}_0 = \mathbf{v}_5 \approx [.133, .133, .333, .200, .200] \tag{2.47}$$

There are a couple of things of interest going on here. The third power of the matrix has strictly positive entries, implying that the matrix is primitive and has a unique steady state distribution. The Schur decomposition is

$$\mathbf{M} = \mathbf{V}\mathbf{\Gamma}\mathbf{V}^{-1} \tag{2.48}$$

$$\mathbf{\Gamma}_{ii} = \{1, 0.592, 1/2, 0, -0.225\} \tag{2.49}$$

This appearance of an eigenvalue of unity is not surprising since we know that matrix is primitive. The magnitudes of the other entries set the time-scales for the decay of the other eigenmodes. Even in such a simple system the convergence to steady-state is not monotonic. As we've seen, all modes with $\lambda_i < 1$ are transient and do not effect the steady state distribution. However, the modes close to 1 may persist for long times (relative to a characteristic time). The state vector π, where the net flow away and to each vertex is zero can be stated as the vector satisfying the condition such that

$$\pi\mathbf{M} = \pi \tag{2.50}$$

i.e. the left-eigenvector associated with the unit eigenvalue.

2.5 Master Equation (ME)

The master equation is a set of first-order differential equations that describe the time-evolution of a discrete system. It was first introduced by Wolfgang Pauli[44] and later used by Glauber to study the kinetics of the Ising model.[45] The master equation relates the change in the probabilities of the occupation of the states $\Xi = \{\xi_1, \xi_2, \ldots, \xi_k\}$ through a vector $\mathbf{p}(t) = [p_1(t;\xi_1), p_2(t;\xi_2), \ldots]$. From the ME, one can derive many relations used in stochastic dynamics such as the Fokker-Planck equation and Langevin equations. Let the conditional probability of changing from states $\xi_i \to \xi_j$ be w_{ij}. The master equation is

$$\frac{d}{dt}\mathbf{p}(t) = \sum_{j \neq i} \left(w_{ij}\mathbf{p}_j - w_{ji}\mathbf{p}_i \right) \tag{2.51}$$

The rate of change of the probabilities depends only on the current state of the system, i.e. the master equation describes a memory-less or Markovian process. Total probability is conserved

$$\frac{d}{dt} \sum_{p_i(t) \in \mathbf{p}(t)} p_i(t) = 0 \tag{2.52}$$

The master equation can be cast into matrix form

$$\frac{d}{dt}\mathbf{P}(t) = -\mathbf{W}\mathbf{P}(t) \tag{2.53}$$

where $\mathbf{W}$ is the stochastic matrix

$$\mathbf{W}_{ij} \equiv \begin{cases} w_{ij} & : i \neq j \\ -\sum_{k \neq j} w_{kj} & : i = j \end{cases} \tag{2.54}$$

By definition the columns sums of $\sum_i \mathbf{W}_{ij} = 0$ for all j, which is simply a restatement of the conservation of probability. Since the equations are only coupled first-order differentials, their solution

is simply

$$\mathbf{P} = e^{\lambda t}\mathbf{v} \tag{2.55}$$

$$\mathbf{W}\mathbf{v} = \lambda\mathbf{v} \tag{2.56}$$

In general, $\mathbf{W}$ is not symmetric and therefore may have complex eigenvalues. However, any matrix with only real elements is guaranteed to satisfy

$$\mathbf{W}\mathbf{v} = \lambda\mathbf{v} \rightarrow \mathbf{W}\mathbf{v}^* = \lambda^*\mathbf{v}^* \tag{2.57}$$

Since the elements of $\mathbf{W}$ are always real, the eigenvalues and their associated eigenvectors may be real, or complex conjugate pairs.

For an ergodic system with rate matrix $\mathbf{W}$, the Perron-Frobenius theorem tells us that there is a unique eigenvalue $\lambda_1 = 0$ and all of the other eigenvalues must have a strictly negative real part. Furthermore the associated eigenvector $\mathbf{v}_1$ must have only non-negative components. This eigenvector has a connection to the steady state probability and ensures that the probabilities are non-negative. All initial conditions of an ergodic system (excluding pathological cases) will approach a steady state vector π. A stronger condition is that of detailed balance which holds, if for each pair i, j

$$\mathbf{W}_{ij}\pi_j = \mathbf{W}_{jk}\pi_k \tag{2.58}$$

implying that around any closed cycle of states there is no net flow of probability. The master equation has a formal solution

$$p_i(t) = \sum_j \sum_k \mathbf{V}_{ik} \exp\left(\lambda_k t\right) \mathbf{V}^{-1}_{kj} p_j(0) \tag{2.59}$$

For a stationary process, i.e. a Markovian one, the transition probability depends only on the time

interval between two events. Therefore we can write the solution as

$$p_i(t) = [\mathbf{V} \exp(\Gamma t) \mathbf{V}^{-1}] p(0) \tag{2.60}$$

$$= \mathbf{M}(t) p(0) \tag{2.61}$$

Where $\mathbf{M}(t)$ is the matrix exponential of the rate transition matrix. The exponential of a matrix always exists and can be found with a Taylor expansion.

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \frac{1}{3!}\mathbf{A}^3 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{A}^k \tag{2.62}$$

If the matrix is diagonalizable it can be found by

$$e^{\mathbf{A}} = \mathbf{V} e^{\mathbf{\Gamma}} \mathbf{V}^{-1} \tag{2.63}$$

Which is trivial since the exponential of a diagonal matrix is

$$\left(e^{\mathbf{\Gamma}}\right)_{ij} = e^{\mathbf{\Gamma}_{ij}} \delta_{ij} \tag{2.64}$$

This gives the important relation between a Markov matrix and a rate matrix acting as its generator

$$\mathbf{M} = \exp(t\mathbf{W}). \tag{2.65}$$

Chapter 3

Entropic forces and Cosolute Flows

3.1 Experimental Motivation

When considering the structural dynamics and functionality of macromolecules, the effective volume of the surrounding environment is a critical factor. Conditions in typical cellular environments are often filled with other molecules whose size is on the same order as macromolecules. As the environment gets crowded, the mutual impenetrability of the particles gives rise to excluded volume effects and has significant consequences for protein stability and folding rate,[46–54] chaperonin action[55] and amyloid fibril formation.[56] Crowded environments may also enhance the rate of protein aggregation[57]. A comprehensive review of macromolecular crowding can be found in[58].

Along this line, two recent experiments are the motivation for the present work. The first is the single-molecule measurement of the mechanical force required to unfold a protein molecule (ubiquitin) using atomic force microscopy (AFM). It was found that at fixed pulling speed the unfolding force increased progressively as more crowders are added.[59] Our model here assumes the globular protein is mainly spherical in shape and when stretched under the AFM, elliptic. Furthermore, while the protein molecules are under stretch they are not free to move. Therefore, for simplicity we consider the limiting case that the stretched protein is an infinitely heavy hard ellipse which is stationary under collisions with the crowder molecules. In this paper, we have restricted ourselves to a 2D system.

The effect caused by the addition of cosolutes observed in these experiments is the result of the interplay of the entropic effect of the macromolecule, cosolute and water. Since the cosolutes added are typically hydrophilic, their addition causes an increase in the total packing fraction of the aqueous solution, leading to a stronger crowding effect.[60] Nevertheless, a single-component system is still a useful tool to elucidate the physical nature of the entropic excluded volume effect. Our

work considers the dynamics of a large heavy solute immersed in a bath of hard-spheres to model these observations of crowding-enhanced protein stability.

Another related experiment showed that a crowded environment can induce shape change in protein molecules (Borrelia burgdorferi VIsE) from a non-spherical native structure to a more compact non-native spherical structure.[61] Associated with this shape change, the function of protein may change as a result. An interesting question to ask is then: Can a non-spherical molecule experience an intra-molecular attraction, somewhat similar to the role of surface tension, change its shape into a spherical one due to the presence of crowders? An objective of the present article is to try to answer this question by studying the collision dynamics between hard-disc crowders and a hard ellipse.

The first study of excluded-volume effects, also known as the depletion force, was originally derived by Asakura and Oosawa[62] for hard spheres. AO theory assumes that the density of the crowders is uniform for all permissible volumes and the excluded volume is simply the volume of the offset shape. The offset for a single convex particle is defined as the surface extended some distance normally from the surface. The excluded volume modifies the partition function by restricting the space available to the particles. The closer two particles are, the more volume is available to the remaining crowders and hence a stronger depletion force. By simple geometric arguments, AO theory gives an attractive interaction for hard discs that scales monotonically with the inter-particle separation. Essentially this is the zeroth-order term in integral equation theories based on statistical mechanics. The true spatial distribution is dependent on pair-wise correlations, which themselves are dependent on three-point correlations, etc. To move beyond AO theory requires detailed knowledge of these higher-order correlation functions. The Ornstein-Zernike equation, an open equation that gives the exact distribution function, can be solved analytically under proper closure relations such as the Percus-Yevick (PY) relation for hard spheres,[63] and hard ellipses.[64] The PY closure gives good results for low size asymmetry and packing fractions, outside this domain it can lead to pathological results for the density profiles. An alternative closure relation, hypernetted-chain (HNC)[65] has been shown to give results more consistent with numerical simulations.[66]

This tendency of hard-disc or hard-sphere objects to feel an inter-particle entropic force has

previously been observed in physical experiments,[67,68] theoretical predictions,[69,70] and computer simulations.[71] The impenetrability of the molecules, a nonspecific steric repulsion, forms the basis of this entropic force. The loss of entropy between the two particles is overcome by the gain in the entropy of the remaining particles.

An interesting effect resulting from the theory of depletion force is the tendency for particles to move along surfaces of decreasing curvature for convex surfaces (and equivalently, increasing curvature for concave surfaces). The depletion zone a crowder makes at contact is greater for a flatter surface, hence the particle should, on average, feel a force in this direction. Studies of various geometries using integral theories[72,73] and density functional theory[74,75] have confirmed this fact. This has been experimentally observed for colloidal particles that feel repulsion near a sharp edge.[76] Manipulation of colloidal structures has an obvious appeal, but typically these studies focus on the density profiles and not the resulting velocity patterns that arise from the entropic interaction. In this chapter we provide evidence of the entropic flows arising from hard potential surfaces.

If a crowder tends to move along regions of decreasing depletion areas, one should observe a flow of entropic origin surrounding non-uniformly curved surfaces. If we vary the geometry of a fixed macromolecule, say from a circle to an ellipse, it is possible to change these density profiles and resulting flows. We investigate this effect by varying the shape of the ellipse as well as the different packing fractions and the relative size of the crowders.

In a biological system, water cannot be regarded as an inert background.[77] The presence of a solute generates an excluded volume not only for the other solutes but also for water molecules. To exclusively investigate the entropic excluded-volume effect, water molecules can also be modeled as hard spheres. In a strict sense, "crowders" should be treated as a multi-component system.[60] However, considering water molecules explicitly in a multi-component system would greatly increase the complexity of the computation. We shall therefore focus on the depletion effects of one-component crowders on a non-spherical body in the present study.

This chapter covers the computational method first, detailing the design parameters and the implementation of the discrete molecular dynamics. This is followed by the results of the simulations,

along with a quantitative analysis on the boundary condition itself. The final section outlines the importance of the velocity fields along with their connections to the original motivating experiments.

3.2 Computer simulations

3.2.1 Simulation design

In the present work, we examine the flow of hard-discs around a hard ellipse fixed at the origin. The entire simulation is done using discrete molecular dynamics (DMD) as each potential collision can be predicted analytically. Initially the time of first collision for all disc pairs is found, along with the interaction against the interior boundary condition. This list is chronologically ordered and the simulation is integrated to the first collision. The collision is handled, conserving energy and impact angle and the collisions for the interacting discs are recalculated with the new velocity vectors.

Calculating a collision between hard-discs is trivial. Given two velocity vectors, initial positions and radii of r_a, r_b, the discs first intersect when the positions are exactly a distance of $r_a + r_b$ apart. The exact time of collision can be reduced to a quadratic equation, whose discriminant is zero when the discs do not collide.

The collision of a disc with any other closed surface is, in general, a difficult problem to solve analytically. If the shape is convex, then the disc can intersect with the surface at most four times. The first point of collision can be found by setting the discriminant to zero, corresponding to a multiplicity of the roots, or identically the first point of collision. The collision can then be found by plugging this solution into the general solution of the fourth order intersection polynomial. This collision can also be visualized as the first point of intersection between a line and the offset of the convex shape. If the offset has a simple form, the problem simplifies greatly. For a disc, the offset *is* another disc, whose collision is trivial. For an ellipse the offset shape is complicated, often requiring an iterative solution. The problem of two translating, rotating ellipses can be solved however, by reducing the problem to the roots of a simpler eighth-order polynomial.[78]

3.2.2 Boundary conditions

Using a dimensionless unit of length L the system of interest is enclosed in a two-dimensional square box centered at the origin. The side length was $2L$ using periodic boundary conditions. Each hard-disc was given an initial velocity vector in a random direction whose magnitude was drawn from Gaussian distribution. The interior boundary condition was an ellipse, whose offset was defined as the surface extended normally a distance r. The ellipse defined with axes E_a, E_b was parametrized as

$$\vec{Q}(\phi) = \begin{bmatrix} E_a \cos\phi \\ E_b \sin\phi \end{bmatrix} \tag{3.1}$$

The first point of contact made by a hard-disc of radius r with the ellipse is the intersection of its velocity ray with the ellipse offset

$$\vec{Q}_{offset} = \vec{Q} + r\hat{N} \tag{3.2}$$

Where $\hat{N}$ is the outward unit vector normal to the surface. The parametric form for the offset shape is thus

$$\vec{Q}(\phi) = \begin{bmatrix} E_a \cos\phi + grE_b \cos\phi \\ E_b \sin\phi + grE_a \sin\phi \end{bmatrix} \tag{3.3}$$

where $g = ((E_b \cos\phi)^2 + (E_a \sin\phi)^2)^{-1/2}$.

3.3 Results

The crowders consisted of a fixed number $N = 50$ of homogeneous hard-discs whose radius varied according to packing fraction values of $\Pi = 0.10$ to $\Pi = 0.30$ with fixed aspect ratio of $k \equiv E_a/E_b$. For comparison, cellular interiors show approximately 20-30% volume occupation by macromolecules.[79]

Statistical averages of the velocity and density fields were taken by dividing the system into 150^2 square cells. Snapshots of the fields were taken after 1/5 of the total simulation time to allow for an initial thermalizing of the system. The resulting density patterns shown in FIG. 3.1 are complex, but similar to other studies of hard-solutes near boundary conditions. As an example, consider in

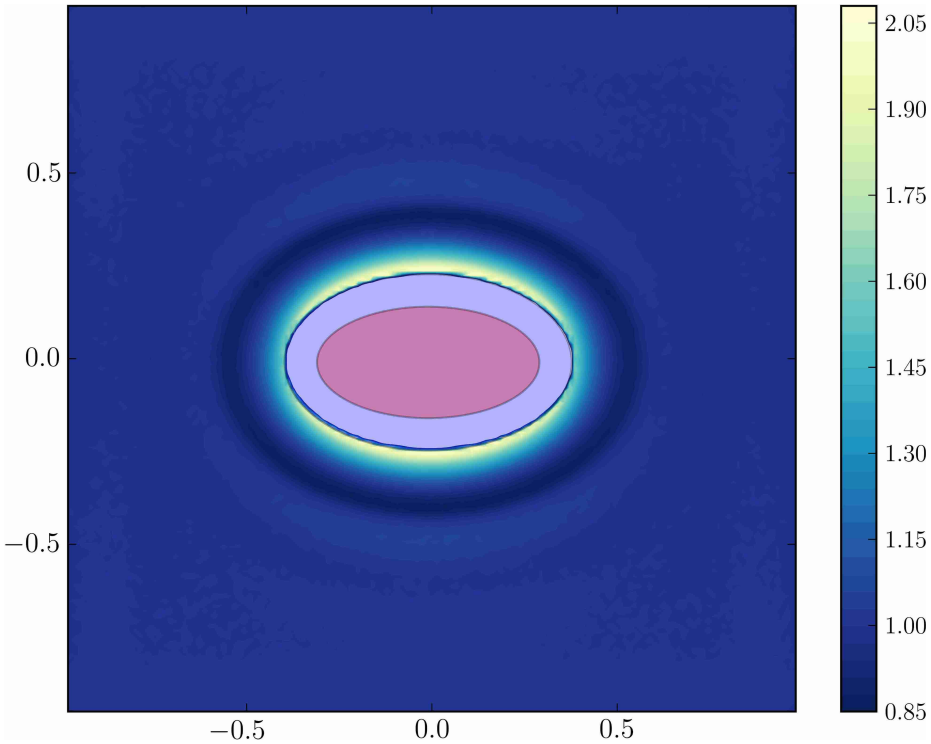


Figure 3.1: Contours of averaged density profiles for an aspect ratio $k = 2.0$, packing fraction $\Pi = 0.30$, and relative crowder radius $r_c/L = 0.04370$. The density has been normalized such that the bulk density is unity. Both the hard-ellipse and an approximation of the depletion zone are shown schematically.

FIG. 3.2 the density distributions plotted along the major and minor axes. Each curve is zero inside the depletion zone, then it exhibits characteristic oscillations on the length scale of the crowder diameter. The reason for this is well known, the crowder statistically prefers to form shells around immobile barriers. It is also clear from the earlier discussion that the density on-contact should be larger for the minor (flatter) axis of the ellipse, as the depletion force there is greater.

Interestingly, when we plot the time-averaged velocity field near the hard ellipse the field exhibits four vortices. Since the system has four-fold symmetry (up to a sign change in the curl of the velocity field), we show only a portion of the upper quadrant of the velocity in FIG. 3.3. The velocity field clearly exhibits a single vortex, one that remains stable in both size and location for the duration of the simulation. The net angular momentum of the system is still zero as each vortex has a counter-rotating partner, but the distribution of the momentum has been partitioned along the quadrant lines. This is an initially surprising result, as such flows are usually caused by a thermal gradient,

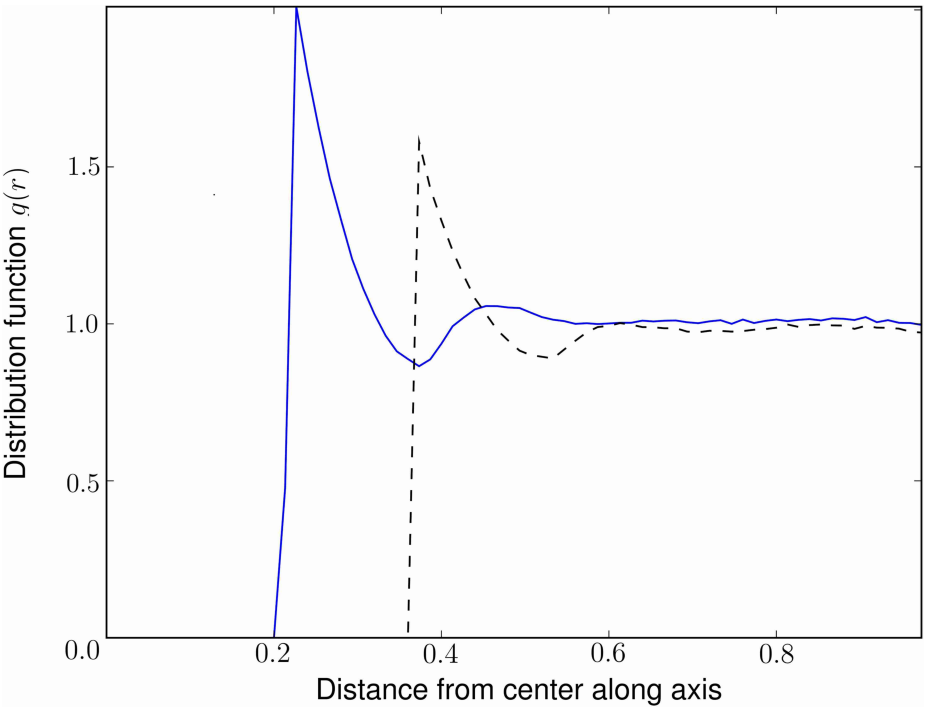


Figure 3.2: Generalized distribution function along each elliptical axis for the parameters aspect ratio $k = 2.0$, packing fraction $\Pi = 0.30$, and relative crowder radius $r_c/L = 0.04370$. The distribution function $g(r)$, is a measure of the average density at a point, normalized to one at the bulk density. The position of the function is to be taken from the origin along the specified axis, with the distance in units of L. The blue (solid) and black (dashed) curves denote the values along the minor and major axis respectively. The density on-contact is significantly greater at the minor axis of the ellipse. The curves exhibits characteristic oscillations at precisely the crowder diameter.

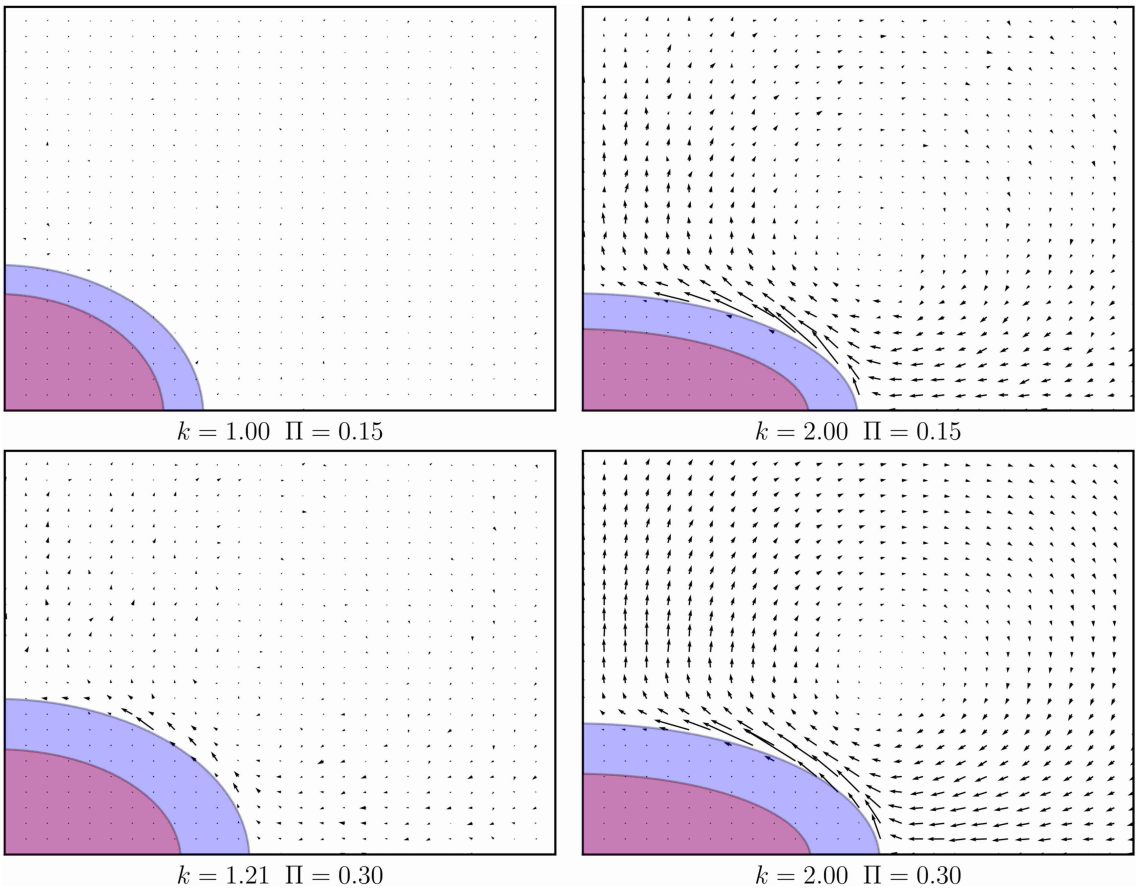


Figure 3.3: Time-averaged velocity vectors in the upper right quadrant of the simulation for different parameters (shown on chart). For comparison of vector magnitudes the largest arrow in the lower-right graph corresponds to a velocity that is 6.4% and 11.8% of the average and median respectively of the initial velocity distribution. The general magnitude of the vectors is highly dependent on the aspect ratio, and disappears completely for a circle ($k = 1$). Both the hard-ellipse and an approximation of the depletion zone are shown schematically.

advective field or other potential. The entropic flows observed here are completely the result of treating the ellipse as a hard boundary (rather than a free particle), which serves to redistribute the angular momentum of the system. This is discussed further in next section. The exterior periodic boundary conditions serve to enclose the flow. It is unknown if the flows obtained are still valid in an infinite bath ($L \to \infty$). The deviation from uniformity of both the density and velocity fields occurs whenever $k \neq 1$. As the the aspect ratio approaches unity ($k \to 1$) the pattern generation takes longer to develop implying that there are no observed phase transitions.

Motivated by the time-averaged velocity contours which seem to imply a compressing force on the

ellipse along the x-axis and an expanding force along the y-axis, an averaged pressure was calculated by recording all momentum changes of the crowder against the ellipse. The average pressure ratio, a heuristic measure designed to measure the tendency of the ellipse to deform, is

$$R \equiv \sum_{r,\Delta p \in C} \frac{1}{k} \frac{\Delta p_x \, \mathbf{sgn}\,(r_x)}{\Delta p_y \, \mathbf{sgn}\,(r_y)} \tag{3.4}$$

where the sum extends over the set C of all recorded collisions of the crowders against the ellipse, $r, \Delta p$ represent the position and the change of momentum vectors over the duration of the collision, respectively, with the subscripts indicating the component along the indicated direction, and **sgn** is the sign function.

R is a positive function over both increasing aspect ratio and crowder density as show in FIG. 3.4. Furthermore it shows the scaling of R is roughly quadratic over an increase of the aspect ratio. Values of $R > 1$, more pressure against the major axis, suggest a return to a spherical shape. The results obtained show that the presence of crowders near a fixed aspherical molecule is not the favored state, one that becomes increasingly unfavorable in both crowded conditions and aspect ratio.

3.4 Theoretical Analysis

As a first approximation, we consider an ideal crowder whose initial distribution is constant at ρ and velocities are drawn from a Boltzmann distribution with temperature T. Under a hard-potential the angle of impact at the point of collision along the tangent surface is conserved. When a crowder particle impacts the boundary of the fixed ellipse the resulting trajectory conserves energy but not momentum. We define $\hat{N}(\phi)$ and $\hat{T}(\phi)$ to be the normal and tangent unit vectors from the

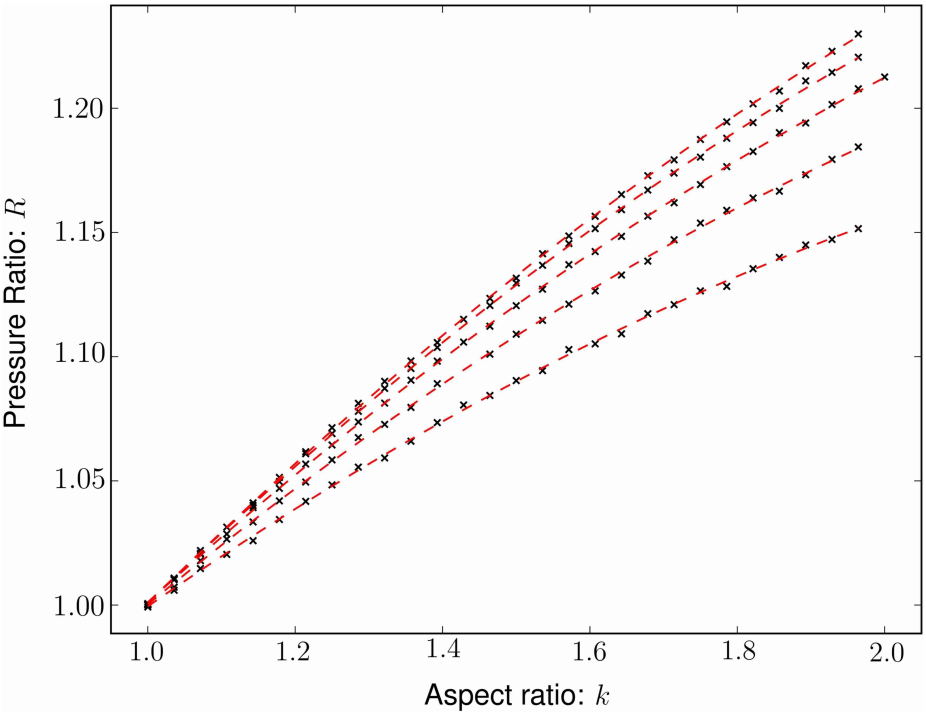


Figure 3.4: Plot of the pressure ratio R as a function of the aspect ratio k for various values of fixed packing fraction Π. Each point on the graph represents a complete simulation with different parameters. The values of the fixed packing fraction include $\Pi = [0.10, 0.15, 0.20, 0.25, 0.30]$. A quadratic best fit curve shown for each value of Π with the bottom curve corresponding to $\Pi = 0.10$ and the other curves following sequentially. Note that each curve starts at $R = 1.00$ regardless of Π since the aspect ratio starts at $k = 1.0$.

parametrization of the ellipse

$$\hat{N}(\phi) = g \begin{bmatrix} E_b \cos\phi \\ E_a \sin\phi \end{bmatrix} \tag{3.5}$$

$$\hat{T}(\phi) = g \begin{bmatrix} E_a \sin\phi \\ -E_b \cos\phi \end{bmatrix} \tag{3.6}$$

The incoming and outgoing velocities $\vec{v}$, $\vec{v}'$ for are simply

$$\vec{v}' = \vec{v} - 2\left(\vec{v}\cdot\hat{N}(\phi)\right)\hat{N}(\phi) \tag{3.7}$$

We can calculate the average angular momentum change for a given ϕ by rotating the incident velocity from 0° to 180° with respect to the tangent. In two dimensions, the cross product of any two vectors $\vec{A}, \vec{B}$ becomes a scalar quantity $\ell = \left[\vec{A}\times\vec{B}\right]_{\hat{z}} = A_x B_y - B_x A_y$. We can define the scalar change of angular momentum with respect to the origin at the point $\vec{Q}(\phi)$ along the ellipse with incident velocity $\vec{v}$

$$\begin{aligned} &\Delta\ell(\vec{v},\phi) = \ell_f - \ell_i = \\ &\left[\vec{Q}(\phi)\times\left(\vec{v} - 2\left(\vec{v}\cdot\hat{N}(\phi)\right)\hat{N}(\phi)\right)\right]_{\hat{z}} - \left[\vec{Q}(\phi)\times\vec{v}\right]_{\hat{z}} \end{aligned} \tag{3.8}$$

Assuming the crowder is uniformly incident from all angles, the average change in angular momentum at a given point along the ellipse can be found by direct integration of an incoming unit velocity vector

$$\begin{aligned} \langle\Delta L(\phi)\rangle =& \frac{1}{\pi}\int_0^\pi \Delta\ell\left(R(\theta)\hat{T},\phi\right)d\theta \\ =& \frac{1}{\pi}\frac{-2(E_a^2 - E_b^2)\cos\phi\sin\phi}{\sqrt{E_b^2\cos^2\phi - E_a^2\cos^2\phi + E_a^2}} \end{aligned} \tag{3.9}$$

Where θ is the incident angle with respect to the tangent point of contact at ϕ, and $R(\theta)$ is the

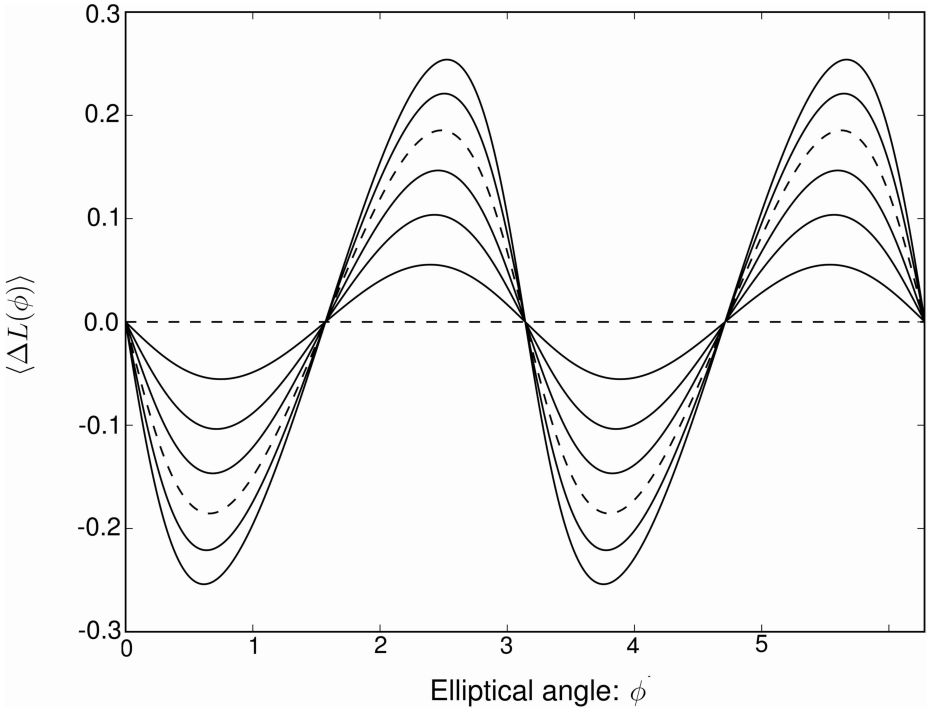


Figure 3.5: Plot of $\langle \Delta L(\phi) \rangle$ for various aspect ratios as a function of the elliptical parameter ϕ. With area fixed at $E_a E_b \pi = 1$, aspect ratios are plotted from $k = E_a/E_b = 1$ to $k = 2$ in increments of 1/6. The dashed lines indicate the two curves $k = 1, 5/3$, with the flat line corresponding to the circle.

rotation matrix

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{3.10}$$

Lacking further information on the ϕ dependence of the velocity distribution we assume that this distribution is the same at all points. As such, an integration over a range of velocities drawn from a Boltzmann distribution at temperature T would scale the expression by a constant factor. As expected, the expression reduces to $\langle \Delta L(\phi) \rangle = 0$ for circles. For aspect ratios where $k \neq 1$, we see in FIG. 3.5 that there are four regions of non-zero $\langle \Delta L(\phi) \rangle$ separated by the symmetries of the ellipse. If the crowder particles are allowed to interact, regions of alternating $\langle \Delta L(\phi) \rangle$ indicate the potential to develop counter-rotating flows in each quadrant. This approximation suffers from several drawbacks; an implicit assumption that the density is constant, that all impact angles θ have an equal probability, and the point-size radius of the crowders. However, this simplified calculation

does suggest the fact that vortex flow generation and the non-conservation of angular momentum of the system are related.

3.5 Discussion

Molecular dynamics simulations on the microscopic level have long been shown to exhibit behaviors on the macroscopic scale.[80] In continuous systems where the flow is obstructed, the non-linear advective term $(\vec{v} \cdot \nabla)\vec{v}$ in the Navier-Stokes equation will convert linear momentum into angular momentum.[81] With a high enough Reynolds number, laminar flow around the obstruction becomes unstable and creates pairs of counter-rotating vortices. Theoretical calculations for flow past a sharp-edge plate predict some vortex formation at any non-zero Reynolds number.[82] The fact that vortices exist for any aspect ratio other than unity in our simulation agrees with these results, the difference here is that our system lacks any initial thermal, velocity or density gradients.

The findings have important implications for all hard-potential models with curvature in the boundary conditions. Related to the original AFM experiment,[59] the assumption that the protein has an elliptical shape leads naturally to the questions addressed in our experiment. Namely, does the addition of crowders induce shape changes for a fixed non-spherical molecule? The pressure ratio observed suggested that a return to a spherical shape, one that increased with both ellipticity and packing fraction. This effect is consistent with experimental observations[59,61] which show the tendency of macromolecules to return to a spherical shape at progressively higher packing fractions. In the context of the AFM experiment, our simulations agree that the presence of crowders led to a greater unfolding force required.

For the simulations of protein folding and unfolding processes, models of Brownian dynamics are often used. The equivalence of Brownian and hard-disc dynamics has been previously investigated in the literature.[83,84] In a hard-potential simulation all particles maintain an infinite memory of their previous trajectories. Contrast this with the Langevin dynamics which include a viscous damping term along with a random component. In these dynamics the particles maintain only a partial memory of their previous trajectories. In the current simulation, reducing the memory of the particles (by including a viscous drag and random force component) reduces the effect the elliptical

boundary condition has on the surrounding environment. In the extreme case of Brownian motion, the density profiles extend only as far as the Brownian steps themselves.

The point of the above observation is to highlight a crucial difference between hard-sphere simulations with and without memory. The distinction may be irrelevant for a rarefied hard-sphere gas such as argon, but may play a role in the crowded molecular environment. While the role of memory (or lack thereof) in biological simulations is important, we feel that the conclusions drawn from these simulations are still applicable. To the extent that DMD and Brownian dynamics are equivalent, our result can be used to explain both the AFM and shape change experiments.

Chapter 4

Conformational States Under Crowding

4.1 Experimental Motivation

The idealization of dilute conditions in conventional *in vitro* biophysical experiments has long been recognized to ignore the important aspect of crowding. In typical cellular environments, the experimental, computational, and theoretical results have shown crowding agents to affect the stability of proteins and the rates of protein folding. The measured and predicted effects of crowding, however, are varied and seem to be dependent on both the protein and the crowders themselves. For a summary of the recent developments in the field since 2004, see the excellent review by Zhou, Rivas and Minton.[85] Crowding effects are still actively being explored, with most efforts focusing on the entropic effects to elucidate the response common to all crowding agents. While energetic interactions may exist between the protein and the crowding agents, a simplified yet effective treatment of a crowder is that of a steric, inert particle, affecting the entropy of the protein's conformational state according to its compactness and shape.[61,86]

In one limit, where the crowding particles are much larger in volume and mass than the protein, good results have been obtained by approximating the crowders through localization. Typically this is modeled as confinement between parallel plates, or in spherical or cylindrical cavities.[19–21] In these approximations, the conformations that are extended beyond the confinement wall are excluded. In the other limit, where the crowders approach the size of a typical residue, the effects of excluded volume dominate. Here, not only are the extended conformal states of the protein highly perturbed, but intermediate states are also proportionally less favored to those states that are compact.[87]

In this chapter, we investigate the effects of crowders on the folding properties of β-sheet proteins using a lattice model. While lattice-based approaches are numerous, those that connect their results to physical experiments are less so. We are motivated by the experiments done by Gai and cowork-

ers[2,3,88] and examine our model in light of their observations. We recognize however, that simple lattice-bead models capture only a portion of the conformational entropy; typically only the gross features of the backbone are correctly modeled. To accommodate for the orientational entropy of the dihedral or ϕ-ψ angles, we introduce an Ising-like model. Each bead along the chain is associated with a binary internal state, with an interaction potential requiring the energetic contributions of two beads to have the same state. Since the correct configuration of the dihedral angles is essential for proper hydrogen-bond formations, this extra degree of freedom gives a physically motivated Hamiltonian that implicitly includes the hydrogen-bond contacts. We model the positional entropy of the backbone by projecting it onto a face-centered cubic (fcc) lattice. We use a Gō-like Hamiltonian to model the native connections and to ensure the existence of a unique ground state. With the conformational entropy defined we propose a model that attempts to capture the salient aspects of macro-molecular crowding using a detailed density of states (DOS) calculation.

Once the DOS has been determined, we use the results of the scaled particle theory (SPT) to approximate the effects of crowding on several β-sheet proteins. The use of the SPT to study the conformational states of the protein folding process is, of course, not new and has been studied previously, see[47,53,89,90] and the references cited therein. We model the protein as right circular cylinder since the native state of β-sheets are naturally disk-like. The crowders, Ficoll 70, are modeled as sphereocylinders accounting for their observed elongation.[91,92] Our treatment is unique among the previous studies in that we use the Wang-Landau algorithm to determine density of states for the positional and orientational entropies, separately. This gives us an accurate measurement of the crowding effects across the conformations of the density of states and the ability to compute thermodynamic quantities to high accuracy.

The organization of this chapter is as follows. In Section 4.2 we introduce the Hamiltonian and the effective free energies associated with crowding and the dihedral angles. In this section, we explicitly define the cost, both enthalpically and entropically for each conformation. We demonstrate how the density of states can be factored into two terms, greatly speeding up the calculation using the Wang-Landau algorithm. We then describe the experimental results in Section 4.3 and fit our

model to the experimental observations. We use the SPT to determine the effect of crowders on the thermodynamic quantities and discuss the implications. Finally, we use the results to make predictions for future experiments.

4.2 Methods

Our protein is coarse grained to a chain of beads and projected as a self-avoiding walk onto a fcc lattice, with a 'bead' representing an amino acid residue. The fcc lattice was chosen over the traditional cubic lattice to provide more degrees of freedom. Previous works have found the fcc lattice to be a more natural fit to the secondary structures of α-helices and β-sheets.[93,94] The choice of lattice is not arbitrary, as higher coordination numbers and different symmetries may better represent the underlying structure. For a summary on the effect of lattice choice see Pierri *et al.*[95]

Two beads are considered nearest-neighbors if they are on adjacent sites on the lattice. With the underlying lattice being defined by a set of primitive vectors $\mathbf{e}$, we define two lattice points $\mathbf{x}_i, \mathbf{x}_j$ to be nearest neighbors iff there exists a vector $\mathbf{v} \in \mathbf{e}$ such that $\mathbf{x}_i = \mathbf{x}_j + \mathbf{v}$. For convenience, we define the twelve lattice steps in Cartesian coordinate space (x, y, z) that form the base set of a face-centered cubic lattice

$$\mathbf{e} = l \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 0 & 0 \end{bmatrix}$$

These twelve vectors define the nearest-neighbors for a given lattice point. Here $l = 3.8$Å is the length scale of the lattice, which is the average spacing between two C_α atoms. Let the set of all backbone conformations be denoted by $\mathcal{C}$ with the vector $\mathbf{c} \in \mathcal{C}$ representing an individual conformation on the lattice.

Our model Hamiltonian is a modification of the Gō-model,[96] where the only energetic contributions are either from the attraction of the residues that are predefined native contacts or the repulsion of the nonnative ones. The Gō model, primarily a model of minimal frustration, typically

ignores the potential from non-native contacts. Models with the repulsive terms added,[97,98] create a frustrated energy landscape since more structural information is encoded in the Hamiltonian. In addition, the high coordination number of the fcc lattice does not always admit a unique native state for some structures without the nonnative term. We let $\mathbf{G}$ be a symmetric matrix of native contacts, $\mathbf{G}_{ij} = 1$ iff positions i and j are native contacts of the protein, otherwise $\mathbf{G}_{ij} = 0$.

In addition to the Gō-like native contacts we further require that the beads have the correct orientations. To achieve this, each bead has a binary internal state, representing the correct (or incorrect) range of values of its dihedral angles. This is similar to the ideas presented in the Muñoz-Eaton (ME) model[99] where each amino acid is allowed two internal states, folded or unfolded. While ME model has been solved exactly in a restricted form[100] and incorporated into more extensive models,[101] we exploit the fact that the ME model generates a density of internal states that is easily decoupled with the positional microstates, thus leading to more precise estimate of the thermodynamic variables. The permutations of these internal states generate an ensemble of microstates; let the set of all such state sequences be denoted $\mathcal{S}$, with the vector $\mathbf{s} \in \mathcal{S}$ representing a particular sequence of the internal states. It will be useful to refer to the total number of folded beads for a state $\sigma = \sum_{i=1}^{L} \mathbf{s}_i$, with the unfolded/folded states defined, indexed by $s_i = 0/s_i = 1$, and amino acid residue count L.

Our Hamiltonian depends on the number of native and non-native contacts of all the beads on the lattice and on their internal state

$$\mathcal{H}(\mathbf{c}, \mathbf{s}) = -\sum_{i=1}^{L} \sum_{j=i+2}^{L} \omega_{ij}[J_{+}\mathbf{s}_i\mathbf{s}_j\mathbf{G}_{ij} - J_{-}(1 - \mathbf{G}_{ij})], \tag{4.1}$$

where $\omega_{ij} = 1$ iff residues i and j are nearest-neighbors on the lattice and J_{+}, J_{-} represent the strength of the Gō model's native and non-native contacts, respectively. A more intuitive form can

be written by counting the number of contributions from the native k_+ and non-native k_- contacts:

$$\begin{aligned} \mathcal{H}(\mathbf{c},\mathbf{s}) &= -J_+ k_+ + J_- k_- \\ k_+ &= \sum_{i=1}^{L}\sum_{j=i+2}^{L} \omega_{ij}\mathbf{s}_i\mathbf{s}_j\mathbf{G}_{ij} \\ k_- &= \sum_{i=1}^{L}\sum_{j=i+2}^{L} \omega_{ij}(1-\mathbf{G}_{ij}) \end{aligned} \tag{4.2}$$

There are two entropic effects incorporated into the free energy, the crowding and the dihedral angle restrictions. The free energy of a state $(\mathbf{c},\mathbf{s})$ is

$$\mathcal{F}(\mathbf{c},\mathbf{s}) = \mathcal{H}(\mathbf{c},\mathbf{s}) - \beta\Delta\psi(\sigma) - \beta\Delta\mu(\mathbf{c}). \tag{4.3}$$

Here $\beta = 1/k_BT$, $\beta\Delta\psi(\sigma)$ is the free energy term associated with the entropy of the dihedral angle orientation and $\beta\Delta\mu(\mathbf{c})$ is an entropic cost of inserting the protein into a solution of crowders (both terms to be defined in later sections). By modeling the crowders implicitly as hard-particles there is only entropic cost for insertion, the term is truly a free energy contribution. If however, the crowder specifically interacts with the protein this contribution must be included in the Hamiltonian. The model admits three fitting parameters (J_+, J_-, h), with h setting the energy scale of the dihedral angle term $\beta\Delta\psi(\sigma)$. Additionally, the crowding term $\beta\Delta\mu(\mathbf{c})$, depends on the concentration and the geometry of the crowders.

The positional conformation $\mathbf{c}$ determines the number of non-native contacts k_-, thus we take $\Omega(\mathbf{c},\sigma,k_+)$ as the density of states. The partition function can be factored by summing up to the maximum number of native contacts k_+^*:

$$\begin{aligned} \mathcal{Z} &= \sum_{\mathbf{c}\in\mathcal{C}}\sum_{\mathbf{s}\in\mathcal{S}} e^{-\beta\mathcal{F}(\mathbf{c},\mathbf{s})} \\ &= \sum_{\mathbf{c}\in\mathcal{C}}\sum_{\sigma=0}^{L}\sum_{k_+=0}^{k_+^*}\Big[\Omega(\mathbf{c},\sigma,k_+)e^{\beta\Delta\psi(\sigma)+\beta\Delta\mu(\mathbf{c})}\sum_{\mathbf{s}\in\mathcal{S},\Sigma\mathbf{s}_i=\sigma} e^{-\beta\mathcal{H}(\mathbf{c},\mathbf{s})}\Big] \end{aligned} \tag{4.4}$$

4.2.1 Conformational Entropy of Dihedral Angles

Associating an entropic cost with the correct dihedral angles is an idea that goes back to the original Zimm-Bragg (ZB)[102] and Lifson-Roig (LR) models.[103] These models were first used for helix-coil transitions, and later extended to include sheets.[104–106] Letting each bead have an internal state, native or non-native, allows us to capture some of the detail present in more complex models, yet still retain the simplicity inherent in lattice models. It is the lack of spatial degrees of freedom that separate the ZR, LR type models from the one presented here. In our model each conformation defines a new Ising-like sub-problem, where we consider the entropy associated with the ensemble of 'spins' of only the nearest neighbor contacts. Our model is actually the generalized variant of the spin systems, commonly referred to as the Potts model. There are two major distinctions between the Ising and Potts models; the spin directions are not necessarily restricted to two states and the strength of spins in contact are determined by an interaction matrix. We still retain the two state model, folded/unfolded, but our interaction matrix has only a single non-zero term, contributing only when both spins are in the folded state. This is evident by the $\mathbf{s}_i\mathbf{s}_j$ term in the Hamiltonian since $\mathbf{s}_i, \mathbf{s}_j \in \{0, 1\}$. However, each folded residue comes at a price, the entropic cost of restricted dihedral angles for a single bead is

$$\beta\Delta\psi(\sigma) = h\sum_{i=1}^{L}\mathbf{s}_i = h\sigma. \tag{4.5}$$

4.2.2 Reduction of State Space - Conformational Decoupling

Since the Hamiltonian, and thus the free-energy, is a function of both positional conformations $\mathbf{c}$, and the orientational conformations σ, k_+, it would seem that a calculation of the full density of states $\Omega(\mathbf{c}, \sigma, k_+)$, is necessary. We will show however, that given a positional conformation, one can decouple the internal states by grouping similar conformations into isomorphic macrostates. Consider the matrix

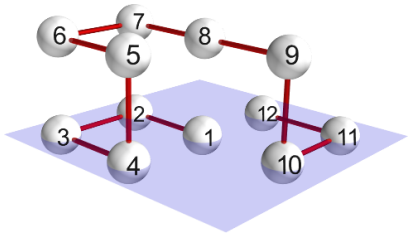
$$\mathbf{X}_{ij}(\mathbf{c}) = \mathbf{G}_{ij}\omega_{ij}(\mathbf{c}) \tag{4.6}$$

of the positive energetic contributions to the Hamiltonian when one ignores the internal states. We can map this symmetric matrix to a simple, but possibly disjoint graph, $\mathbf{X}_{ij} \rightarrow g(\mathbf{c})$, by observing that $\mathbf{X}$ is an adjacency matrix. Let $\mathrm{Aut}(g(\mathbf{c}))$ define the automorphism group to which $\mathbf{c}$ belongs. Each $\mathrm{Aut}(g(\mathbf{c}))$ is a permutation group, whose members are related by mapping the vertices of the graph onto itself through permutation such that the resulting graph is isomorphic to the original. While all graphs in the same automorphism group are isomorphic to each other, we should note that not every graph is physically realizable in our lattice model due to the restriction that no two beads can occupy the same lattice site. The key to decomposing the density of states, however, is the grouping of the graphs, and hence the conformations. Each automorphism group defines a finite-graph for a spin-state system. The density of states for this finite-graph system is calculated by considering, over all possible 'spins' of the system, the restricted counts of σ and k_+. Since the problem of finding the internal states is identical for all members of a particular automorphism group we can decouple the density of states as

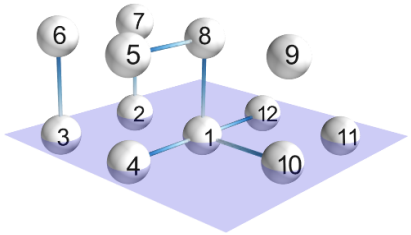
$$\Omega(\mathbf{c}, \sigma, k_+) = \Omega_1(\mathbf{c})\, \Omega_2(\sigma, k_+; \mathbf{X}(\mathbf{c})) \tag{4.7}$$

As an illustrative example of the decoupling method consider a 12 bead homopolymer defined over a cubic lattice where every connection is favorable ($\mathbf{G}_{ij} = 1$ if $|i - j| > 1$) in the particular conformation shown in Figure 4.1a. All of the native connections, which in this case are simply all nearest neighbors, are shown in Figure 4.1b. Abstracting the representation to a graph in Figure 4.1c shows the finite system that solves the density of states over the Potts model. Usually, when solving a Potts/Ising-type system the underlying graph has a high degree of symmetry (cubic lattices or Cayley trees are common examples). For a typical graph produced by our model this symmetry is broken, forcing us to numerically compute $\Omega_2(\sigma, k_+; \mathbf{X}(\mathbf{c}))$. However, the number of edges in the graph determine the maximum number of favorable connections. Since this number is small, the convergence of the Wang-Landau algorithm on this portion of the DOS is rapid.

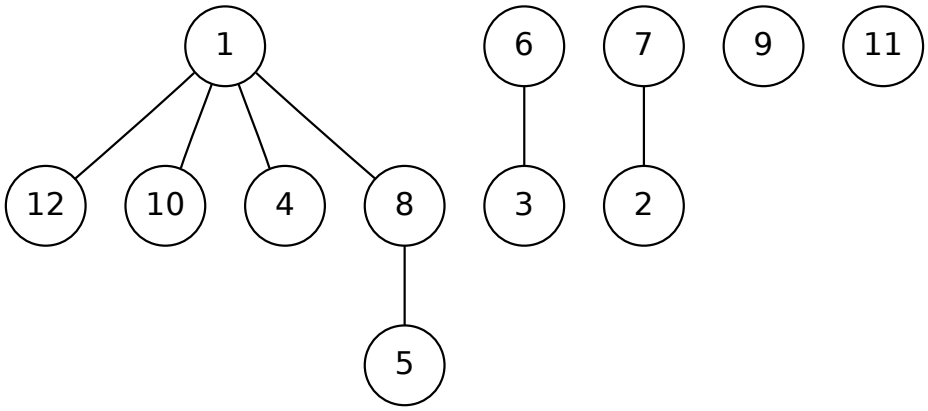
In general, grouping a particular $\mathbf{c}$ to its automorphism group requires solving the graph isomorphism problem multiple times. While specialized algorithms exist,[107] the computational solution



(a) Backbone **c**



(b) Native Contacts **X**(**c**)



(c) Connection Graph $\mathbf{X}(\mathbf{c}) \rightarrow g(\mathbf{c})$

Figure 4.1: Sample homopolymer (with all connections favorable) on a cubic lattice with (a) the backbone, and (b) the energetically favorable connections. The problem of finding the density of states for the internal conformations of each bead for this conformation is then reduced to solving the density of states of the Potts model over the graph shown in (c). Note that the labels in (c) are shown only to guide the eye, all valid permutations of the indices belong to $\mathrm{Aut}(g(\mathbf{c}))$ and hence define the same Potts sub-problem. The **G** model in the Hamiltonian would make some of the connections in $\mathbf{X}(\mathbf{c})$ unfavorable, further simplifying the problem. Additionally, our model is defined over a fcc lattice, while the example shown above is a cubic lattice for illustrative purposes.

is unique in its complexity class and lacks a simple invariant that definitively determines isomorphism.[108] The problem is greatly simplified if the energetic matrix **G** permits only a few, highly degenerate sets of graphs. The Hamiltonian defined for β-sheets happens to be one of these favorable partitions. Since the β-sheet structure is essentially planar, moving perpendicular to the strand direction identifies a column of connections. When abstracted to a graph, the only structure possible is that of a linear chain (unary tree), whose length is limited by the number of strands. The set of conformations $\mathcal{C}$ for β-sheets create graph structures that have a high degeneracy and consequently low cardinality in the set of unique automorphism groups. Additionally, checking for graph isomorphism of linear chains is trivial since these graphs can be uniquely determined by a

single number, the chain length. These two facts combined significantly speed the computation of the decoupled density of states $\Omega_1(\mathbf{c})\Omega_2(\sigma, k_+; \mathbf{X}(\mathbf{c}))$.

4.2.3 Implicit Crowding Effects

We study the effects of crowders to our protein system using the scaled particle theory. In one of the original formulations of SPT [109] one calculates the work done to expand a spherical cavity of radius R in a hard sphere fluid of radius r_c. The centers of the fluid particles are excluded from the cavity region, which implies that meaningful values of R must be greater than $-r_c$. This work $W(R)$ is the configurational part of the chemical potential for a single solute particle $\Delta\mu(\mathbf{c}(R))$. The relationship between the probability of any particular configuration and the work required to create it is $P(R) = \exp(-\beta\Delta W(R))$, where $P(R)$ is the probability that there are no centers of fluid particles of radius r_c in the spherical region $(R + r_c)$. If we approximate both our protein and crowders as hard-spheres, we can calculate the work required to create the proper cavity. One first calculates the probability of finding a cavity whose size can accommodate only a single spherical solute particle. This yields

$$W(-r_c \leq R \leq 0) = -kT \ln(1 - (4/3)\pi\eta(R + r_c)^3), \tag{4.8}$$

where η is the number density of the solvent. One expands $W(R)$ around $R = 0$ to the second order (noting that the leading term for large R must be the pressure-volume term $4/3\pi pR^3$ with p as the pressure of the fluid) and obtains

$$\begin{aligned} \beta\Delta\mu(\mathbf{c}(R)) = \quad & \beta(\Delta\mu(\mathbf{c}(0)) + \Delta\mu'(\mathbf{c}(0))R \\ & + \tfrac{1}{2}\Delta\mu''(\mathbf{c}(0))R^2 + \tfrac{4}{3}\pi pR^3) \end{aligned} \tag{4.9}$$

These terms can be computed by the continuity of $W(R)$ and its first two derivatives at $R = 0$. The pressure can be found by substituting in the exact solution of the Percus-Yevick equation, yielding a density approximation as a function of the packing fraction ϕ. This density route is not unique

among thermodynamic pathways. Expressions have been worked out for both compressibility and viral routes.[110] Each of these pathways amount to a smoothing in the structural information of the fluid as one 'turns-on' the density field. The compressibility and viral routes tend to yield better approximations to the solvation free energy, giving

$$\begin{aligned}(\beta\Delta\mu)_{\text{spherical}} &= \frac{\phi(-2+7\phi-11\phi^2)}{2(1-\phi)^3} - \ln(1-\phi) \\ &+\frac{18\phi^3}{(1-\phi)^3}\left(\frac{R}{2r_c}\right) - \frac{18\phi^2(1+\phi)}{(1-\phi)^3}\left(\frac{R}{2r_c}\right)^2 \\ &+\frac{8\phi(1+\phi+\phi^2)}{(1-\phi)^3}\left(\frac{R}{2r_c}\right)^3\end{aligned} \tag{4.10}$$

The above treatment by SPT assumes, however, that the cavity created is spherical, a condition that is not rigorously satisfied for the crowders nor the proteins examined in this study. The native states of the proteins studied in this work can be well approximated by a right circular cylinder, as the β-sheet structures are disk-like. Additionally our crowder, Ficoll 70, is known to have an elongated shape[91] and recent predictions[92] model them as sphereocylinders with diameter of 28 Å and an end-to-end length of 184 Å. The extension of SPT to work with aspherical mixtures can be expressed through the activity coefficient.[111,112] The activity coefficient γ_i is the measure of the deviation of the ith species at the actual composition of the solution from the chemical potential of an ideal solution as given by the equation,

$$RT\ln\gamma_i \equiv \mu_i - \mu_i^0 \tag{4.11}$$

where μ_i^0 is the chemical potential of a reference state. For hard-convex particles the non-ideality of a particular species of interest can be found by computing an expression as a function of the volume V_i, surface area S_i and the Kihara support function H_i of that species,[92] given by

$$\begin{aligned}\ln\gamma_i &= -\ln(1-\langle V\rangle) + \tfrac{H_i\langle S\rangle + S_i\langle H\rangle + V_i\langle 1\rangle}{1-\langle V\rangle} \\ &+\tfrac{H_i^2\langle S\rangle^2 + 2V_i\langle H\rangle\langle S\rangle}{2(1-\langle V\rangle)^2} + \tfrac{V_i\langle H^2\rangle\langle S\rangle^2}{3(1-\langle V\rangle)^3},\end{aligned} \tag{4.12}$$

where $\langle X \rangle \equiv \Sigma \rho_i X_i$ and $\langle 1 \rangle \equiv \Sigma \rho_i$ are the averages over the different species, with ρ_i as the number density of that species. For a right circular cylinder and sphereocylinder respectively the Kihara support functions are $H_{\text{sphereocylin}} = r\pi/4 + L/2$ and $H_{\text{cylinder}} = r + L/4$ with r as the radius and L as the length of the cylindrical section.

In the process of calculating the density of states, we can sample the conformations to determine the parameters for the activity coefficient. For each conformation of the peptide we calculate a best fit circular cylinder by pointing the axis of the cylinder along the largest principle axis of the bead positions then scale the radius and length so all beads fit inside. Using Eq. 4.11 and 4.12 we can determine the free energy due to crowders in our system.

4.2.4 Implementation of the Wang-Landau Method

For a review of the Wang-Landau method and a definition of the terms see Chapter 2.2.2. All of the computations carried out in this chapter took $f_0 = e^1 \approx 2.71828$, $f_{\text{final}} = e^{-9}$, and $f_{i+1} = \sqrt{f_i}$. Each time a conformation was selected, the DOS is updated along with a histogram of visits for that conformation $H(\xi)$. The factor f was reduced when $H(\xi)$ was no less then 90% of the average number of visits for all conformations $\langle H \rangle$. Once the factor f had been reduced we reset the histogram of visits for all conformations, $H(\xi) = 0$, and began the process again. Each conformation is still directly related to a numerical energy. By calculating the DOS for conformations we can delay the calculation of the energy. This has the advantage that multiple simulations are not required for each set of the system parameters (J_+, J_-, h, ϕ, r_c).

Our move set consists of pull moves, which were first defined over a cubic lattice [113] and later for triangular lattice models. [114] Pull moves are an ergodic, reversible move set that modify the positional conformations by moving the chain along a path defined by a pair of beads adjacent in chain sequence $(i, i+1)$. Since the number of moves are finite and easily computable, we can quickly determine the factors necessary for detailed balance (i.e. n_A, $n_{A \to B}$, etc).

When converged, the WL method gives a flat histogram. That is, the averaged fraction of time spent at each macrostate approaches the same constant. Here we define a macrostate as the set of all conformations with the same energy level. Not every microstate is visited during the

simulation, nor would it be possible due to the exponential growth in the DOS as a function of chain length. We assume that the visits to each state are ergodic, subsequent visits to a macrostate will visit each conformation an equal number of times over long averages. This idea is reasonable when we consider detailed balance is obeyed for the Monte-Carlo simulation, and is employed by Wust and Landau.[37] We exploit this observation to determine a probability distribution for a second observable as a function of the first. For instance, we can step through the conformations to compute the probability distribution of the activity coefficient for the protein in a particular conformation $\mathbf{c}$. Doing so prevents the need for a multiplicative increase in the density of states (thus speeding up the convergence of the WL algorithm), yet it still provides us with a reasonable estimate of an extended DOS $\Omega(\mathbf{c}, r_g)$.

4.3 Results

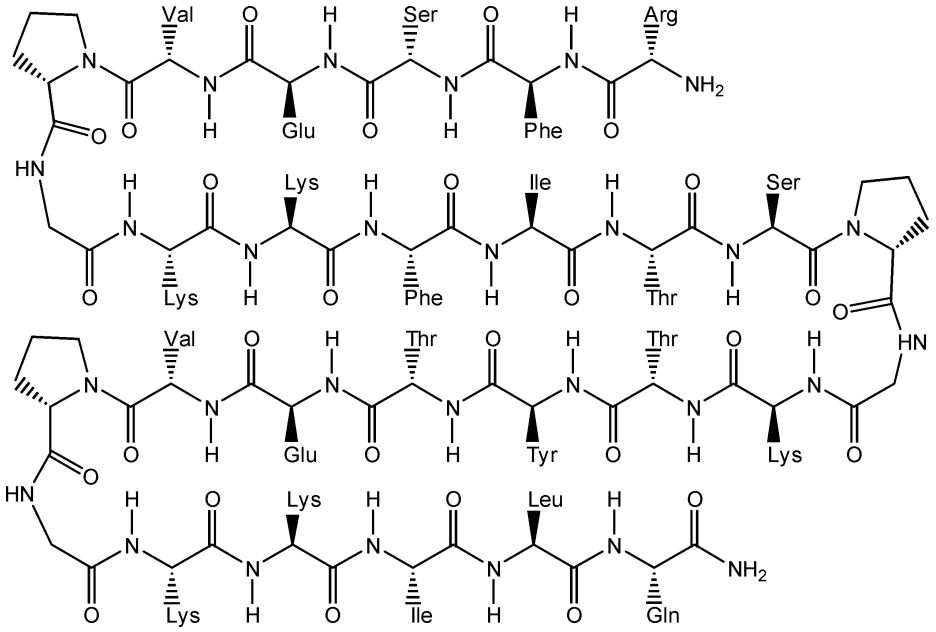
4.3.1 Model Calibration

Unlike α-helices, the β-sheet motif has been difficult to study experimentally due to its propensity to aggregate. Recently there has been a spate of designed peptides that exhibit the β-sheet motif, albeit with extremely broad thermal transitions.[115] We consider and describe below, three experimentally designed β-sheets peptides used in this study. These designed proteins were specifically chosen to test the model against their readily available experimental measurements.

The first peptide (sequence: `RFSEV`^{D}`[PG]KKFITS`^{D}`[PG]KTYTEV`^{D}`[PG]KKILQ`, nicknamed ^{D}P^{D}P^{D}P) is a 28-residue chain with a natural four strand β-sheet structure. This designed peptide was studied experimentally by Xu *et. al.* [2] as an extension of the peptide ^{D}P^{D}P-II first proposed by Gellman and co-workers.[116] A schematic model of the native state for the peptide is shown in Figure 4.2a. The second peptide (sequence: `RFIEV`^{D}`[PG]KKFITS`^{D}`[PG]KTYTE`, nicknamed ^{D}P^{D}P) is a 20-residue chain with a natural three strand β-sheet secondary structure.[117] The final peptide (seq: `GEWTWAD[AT]KTWTWTE`, nicknamed trpzip4-m1), is a 16-residue variant of the tryptophan zippers studied by Cochran *et. al.*[115] and later by Du *et. al.*[3] Compared to the designed peptides, the stability of the tryptophan zippers is significantly higher (with a difference of approximately $\Delta G_{\text{unf}} \approx 1.0$ kcal mol^{-1} at 298 K). A schematic model of the native state for this peptide is shown

in Figure 4.3a.

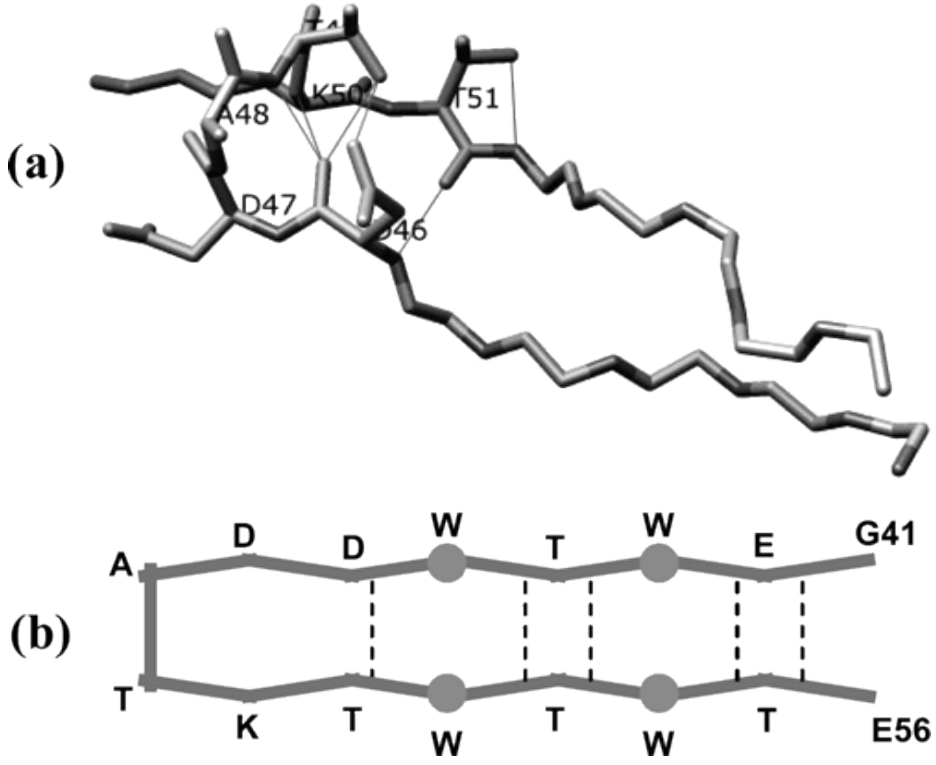
An artifact of the fcc lattice forces the β-hairpin to be made over an odd number of lattice sites, thus we replace the Pro-Gly residues in the first two designed proteins and the Ala-Thr residues in the third peptide trpzip4-m1, with a combined residue (denoted with a square bracket, e.g.. [PG]). The experimental measurements on these peptides have been carried out using temperature jump experiments, for details see[2,3].



(a) Chemical Representation

Figure 4.2: Native state of the peptide ${}^{D}\mathrm{P}{}^{D}\mathrm{P}{}^{D}\mathrm{P}$ represented schematically. The Pro-Gly amino acid residues at each end are combined to a single bead to allow for the proper turn structure on the fcc lattice. Image used with permission from Gai.[2]

We use the WL method to calculate the conformations of the positional $\Omega_1(\mathbf{c})$ and orientational $\Omega_2(\sigma, k_+; \mathbf{X}(\mathbf{c}))$ density of states. The model is calibrated by fitting J_+, J_- and h to the data of three experiments.[2,3,88] We calculate the fraction of β-sheet contacts, observable from the experiments, by taking the expectation of $\langle \alpha\Theta(\alpha)\rangle$. Here $\langle\cdot\rangle$ is the standard Boltzmann average, Θ the Heaviside step function and $\alpha(\mathbf{c})$ measures how close the protein is to its native state. Since the experimental data measures the fraction of β-sheet contacts (inferred from a circular dichroism measurement), we



(a) Chemical Representation

Figure 4.3: Native state of the trpzip-m1 peptide represented schematically. This image was generated from the NMR structure of trpzip4 (PDB code 1LE3, structure 1). The dashed lines represent the hydrogen bonds. Image used with permission from Gai.[3]

calibrated our model with physically similar observable,

$$\alpha(\mathbf{c}) = (k_+ - k_-)/k_+^*. \tag{4.13}$$

This fraction of β-sheet contacts was used to calibrate the three free parameters. We note that without the non-native term J_- in the Hamiltonian in Eq. (4.2) the fraction of sheet contacts would be, as usual $\langle k_+/k_+^* \rangle$. In Figure 4.4 we show the fits of the two proteins trpzip4-m1 and ${}^{D}\mathrm{P}^{D}\mathrm{P}$ with the fitting parameters given in Table 4.1. The fits are quite good, encouraging us to make predictions about the system behavior as a function of crowding packing fraction. It is worth noting that the three and four stranded designed β-sheets had the best fits with $J_- = 0$, implying that additional stabilization provided by the term was needed only to model trpzip4-m1. This may not be surprising when considering the larger melting points and the broad thermal transitions of the

Table 4.1: The fit parameters of the free energy (Eq. 4.3) to the model for each peptide. Here J_+, J_- are the strengths of the native and nonnative bonds and h sets the energy scale of the dihedral angle term.

	h	J_+	J_-
Trpzip4-m1	9.396	6526.8	4691.6
Three-stranded	5.103	2228.8	0.0
Four-stranded	4.319	1751.3	0.0

designed proteins versus that of the smaller β-hairpin peptide (listed in Table 4.2).

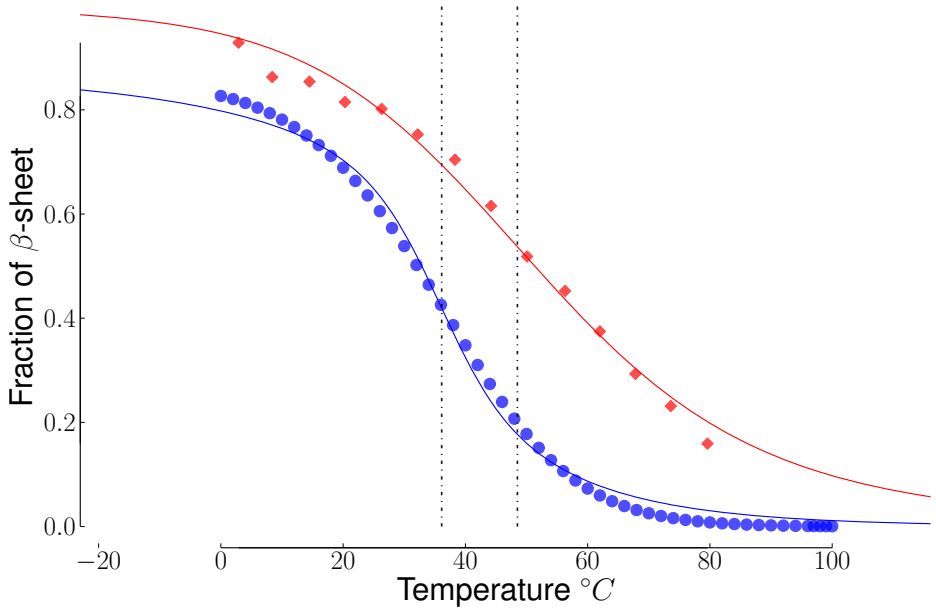


Figure 4.4: Experimental data of fraction folded versus temperature for trpzip4-m1 (blue circles) and the three stranded β-sheet $^{D}\mathrm{P}^{D}\mathrm{P}$ (red diamonds). Model fits are shown with dashed lines of the same color. The thin (black) vertical lines are shown to mark the critical temperatures at 36.1 and 48.5 °C for three-stranded and trpzip4-m1 peptides respectively. The fit for the four stranded β-sheet $^{D}\mathrm{P}^{D}\mathrm{P}^{D}\mathrm{P}$ is similar to the three strand and is not shown for clarity.

4.3.2 Effects of Crowders

In this section we use the model defined above to study the effects of crowders on peptide/protein structures and stability. For one of the peptides we have the experimental results on crowding effects to compare with. In the paper by Mukherjee *et. al.*[88] a significant change (approximately 12°C) in the melting point of trpzip4-m1 was observed under crowded conditions. The crowder chosen for this experiment was Ficoll 70 (F70) at a concentration of 200mg/ml. F70 is a compact, highly cross-linked branched co-polymer of sucrose and epichlorohydrin[118] with an average molecular

weight of 70,000. At 200mg/ml, 300mg/ml the packing fractions are approximately $\phi = 0.13$ and $\phi = 0.20$ respectively.[119,120] We study the effects of crowders by considering the specific heat, $C_V(T) = \beta^2 \left(\langle E^2 \rangle - \langle E \rangle^2 \right)$ and note that in all cases, we observe only a single maxima. We identify this maxima as the melting temperature T_C (alternatively, $\frac{\partial C_V}{\partial T}|_{T_C} = 0$).

Heat capacity as a function of temperature for trpzip4-m1 is shown in Figure 4.5 while the melting points for all peptides listed in Table 4.2. As expected, trpzip4-m1 displays crowding enhanced stability with the change of critical temperature $\Delta T_c = [1.03, 1.65]^\circ C$ at packing fractions $\phi = [0.13, 0.20]$ respectively. However, the three and four stranded β-sheets exhibit a slight decrease in their critical temperatures with ϕ, indicating an entropically based *instability* caused by the crowders.

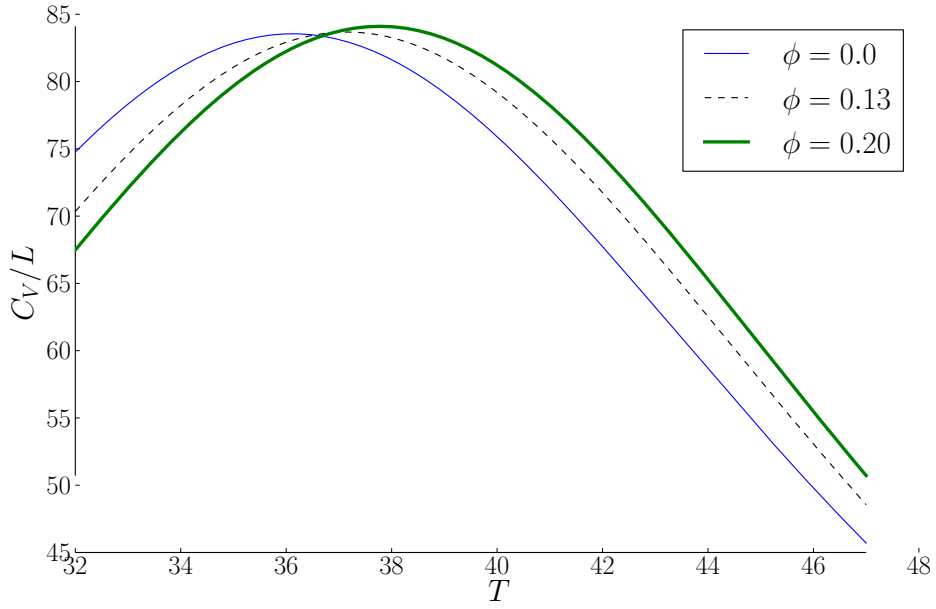


Figure 4.5: Specific heat per residue count for the protein trpzip4-m1 in the presence of crowders.

The native state for the three and four stranded β-sheets are highly aspherical. When considering the entropic effects of crowders, the system prefers compact conformations that minimize the excluded volume effect. As a consequence of this, the native state ceases to be the minimum free energy conformation at large enough ϕ. Crowding-induced conformational change of the native state has been observed experimentally in a recent work by Dhar *et. al.*[120] who studied phosphoglycerate kinase (PGK) with the same crowders as our simulations (Ficoll 70). In this study, the conformational

Table 4.2: List of the experimental and model melting points (given in $^{\circ}C$) for each peptide. The experimental melting points are taken from [2,3,88] in a dilute solution without crowders. The calculated melting points from the model are given at the listed values of the packing fraction. Not shown is the experimental value of trpzip4-m1 in the Ficoll 70 solution of 200 mg/ml with $T_C = 44.0 \pm 0.2$ °C.

	Experimental	$\phi = 0$	0.13	0.20
Trpzip4-m1	32.1 ± 0.9	36.12	37.15	37.76
Three-stranded	52.6 ± 0.4	48.49	48.23	48.10
Four-stranded	50.5 ± 0.8	49.18	49.01	48.94

states were changed dramatically with crowders; an optimal non-zero packing fraction of crowders was found to increase the protein's activity. In our simulation, there was no shift to a new distinct native state at higher crowding concentrations. Rather, we observed a gradual shift towards more compact conformations at the expense of breaking energetically favored bonds, a general collapse of the β-sheet. As an example of the native conformation, which is enthalpically favored, versus an entropically favored one see Figure 4.6. Previous studies that showed crowding enhanced stability often dealt with globular wild-type proteins, whose natural environment required them to operate in crowded conditions. In contrast to the PGK study, the two larger peptides in our study were not wild-type, rather they were designed and studied because of the fact that they folded into beta-like conformations at realistic temperatures without aggregation. This suggests further experimentation on the designed peptides to determine if the destabilization of the native state against those of the unfolded and intermediate states under crowded conditions can be observed experimentally.

In order to assess the effect on the conformational states we examine the Boltzmann averaged excess chemical potential from the native state as a function of temperature and ϕ

$$\beta \langle \Delta\mu_{\text{ex}}(T) \rangle = \langle \ln \gamma_i - \ln \gamma_N \rangle . \tag{4.14}$$

Figure 4.7 of this free energy term for trpzip4-m1 illuminates several interesting structural features from an ensemble perspective. At large temperatures we see that this excess chemical potential approaches a constant, proportional to the change in the unfolded states due to the crowders. Conversely, at very low temperatures crowders have no effect on the only viable conformation, the

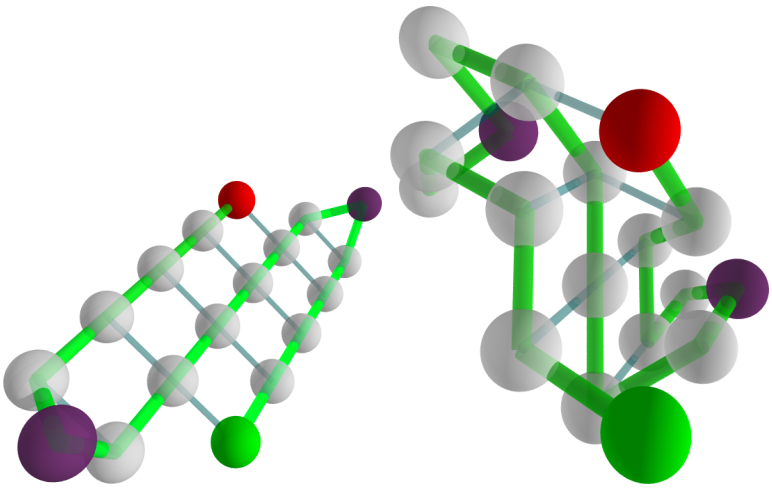


Figure 4.6: Example of native state (top) and intermediate state (bottom) of the three-stranded peptide. The top state has ten bonds (shown as thin blue lines) while the bottom has eight bonds, making the native state favored energetically. However the ratio of activity coefficients $\ln \gamma_1 / \ln \gamma_2$ is 1.13 at 200 mg/ml ($\phi = 0.13$) favors the eight bond structure due to the entropic crowding effects. The C and N terminus marked with red and green beads respectively and the combined Pro-Gly amino acid residues are purple beads.

native state. Near the folding transition temperature, the effect is large and non-linear.

4.4 Discussion

In this chapter we developed a coarse-graining model for proteins by combining the Ising-like state information of the dihedral angles of the modeled β-sheet structures with a fcc lattice model. The density of states has a favorable decoupling which enabled us to use the Wang-Landau method to determine the partition function accurately, giving excellent quantitative agreement with previous *in vitro* experiments. Using our model and the predictions of SPT we found crowding-induced stability, in qualitative agreement with experiment for the smaller peptide, trpzip4-m1. The effect predicted by this model by showed a modest change of about $\approx 1°$C, in contrast the the large change observed in the experiments of Mukherjee.[88] We note however, that these coarse-grained models are approximations and selectively ignore various interactions. The study presented here is an entropic one. If the crowders have enthalpic interactions with the peptide, then these predicted effects will be incomplete. We attribute this underestimation to effects that cannot be explained by excluded volume effects alone.

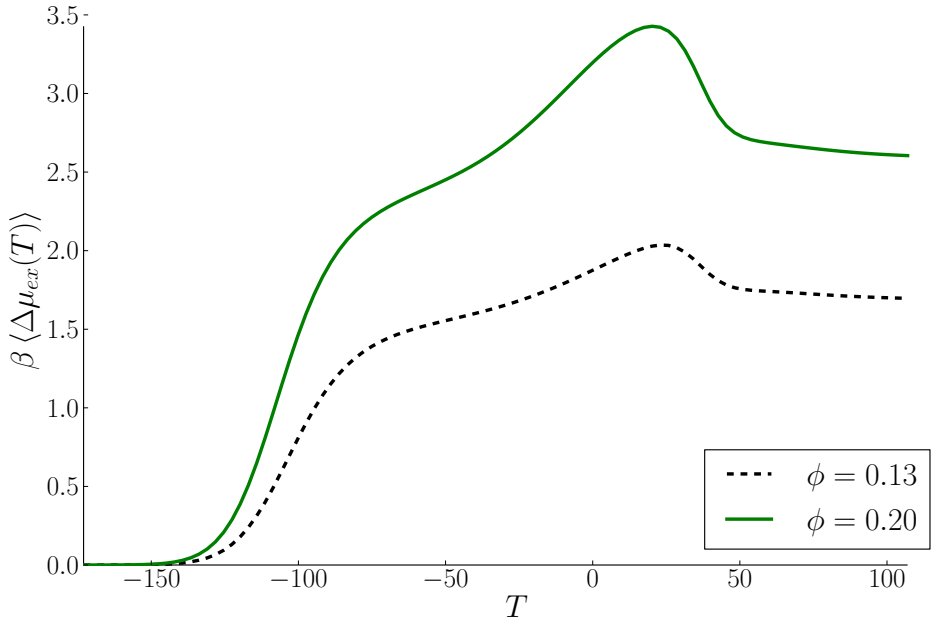


Figure 4.7: Excess chemical potential (defined in Eq. 4.14) for the protein trpzip4-m1 in the presence of crowders.

We found that that the model predicted instability for the designed three- and four-stranded β-sheet peptides. This is consistent with the observation that their native state does not minimize excluded volume effects (it is disk-like rather then globular). This observation alone however, is not sufficient to predict of crowding based instability. Even if the native state is non-ideal, one has to consider the entire ensemble of states as a whole. This was possible using the Wang-Landau method which allowed us to accurately determine the density of states under the constraints of our model.

The extension of the Gō-like contact map to a finite graph presented here is not limited to the β-sheet motif. This entropic model of conformational states can be extended to α-helices or a mixture of secondary structures, as long as the contact graph structure can be decomposed into simple degenerate forms.

Chapter 5

Aggregation Models

5.1 Introduction and Motivation

As we have seen in the previous chapters, the folding properties of a single protein are not only affected by the surrounding solution, but also the conformational state space available through confinement and excluded volume interactions. Up until this point we have considered the protein folding process as an independent event, including only an entropic solute interaction. In reality, there exist multiple biological events that feature enthalpic protein-protein interactions. Focusing on a subset, we apply our techniques to the study of aggregated species to model proteins that form composite structures which are larger and more complex than the sum of the individual elements.

Our study of aggregated species is primarily motivated by the devastating neurodegenerative disorder known as Alzheimer's disease. Alzheimer's is a progressive disorder that causes a deterioration of memory and cognitive function. The disease is the leading cause of dementia,[121] effecting over thirty-five million people worldwide. The principal risk factor for contracting the disease is age, with the incidence doubling every five years after age 65. At 80 years old, one in six people have the disease.

Plaques of amyloid-β (Aβ) fibrils are recognized as an important pathological feature in Alzheimer's disease. At normal physiological parameters, Aβ is known to spontaneously self-aggregate into oligomers of two to six peptides. These in turn coalesce into larger forms. Aβ can then grow into fibrils that form insoluble β-pleated sheets of amyloid plaques. However, it is the soluble oligomers and smaller amyloids that have been shown to be the most neurotoxic. This is evidenced by correlating the levels of the oligomerization with toxicity to synapses in brain-slice preparations.[122] This is interesting, since the total amount of Aβ plays less of a role than the intermediate growth phase.[123] Structural information for the Aβ peptide and its oligomers and fibril structure has been

slow,[124] but recent results are beginning to emerge that detail these key features of the aggregation process.[125] What is needed, is a general model that can incorporate these findings for Aβ and other aggregation problems. Aggregation studies of these proteins, and indeed all proteins in general, are essential to our understanding of the disease.

It is the aim of this chapter to develop and investigate several aggregation models with the tools introduced in this thesis. We begin with Section 5.2, where we develop a simplified first-order model. In it, we make several simplifying assumptions on the conformations and energies of the aggregated species. While the model is useful to develop intuition, we find that it is too restrictive and unable to provide a complete description about the aggregation process. To that end we generalize the first-order model in Section 5.3 by reformulating the aggregation problem as the Potts model in an external field over an arbitrary graph. The main result of this chapter is the development and application of this solution as an operator expansion. With this method we are able to present new results on the Potts model problem. In Sections 5.5 and 5.6 we derive results for a one-dimensional 'ladder' with thickness one and two respectively. Finally we present results in Section 5.7 for a wider class of problems with a subgraph recursion relation, albeit in the absence of external fields.

5.2 First-order Aggregation

Just as we coarse-grained the amino acid residues in Chapter 4 to study the folding properties of a single protein, as we study the aggregation process we will make a suitable approximation for the larger length scale of aggregated species. In this section we will approximate an entire protein as one unit on a graph or lattice. Aggregation then, is a study of nearest neighbor contacts over this graph.

Consider a simple interaction model over a square lattice with side length L and $N = L^2$ sites with the Ising Hamiltonian

$$\mathcal{H} = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j \tag{5.1}$$

The lattice spacing is such that only a single particle can be occupied at any given site, hence $\sigma_i \in \{0, 1\}$ corresponds to a lattice site being occupied or not. The interaction strength J, is related

to the propensity for two proteins to be found next to each other. The summation extends over all nearest neighbors with a global restraint on the number of particles in the system ($m \leq N$)

$$\sum_i \sigma_i = m \tag{5.2}$$

Ultimately we want to set $N \to \infty$ corresponding to the thermodynamic limit. Thus we scale the interaction strength with the number if lattice sites as

$$J \propto \ln N \tag{5.3}$$

The object of our study is a cluster or an oligomer, a collection of m proteins that are connected through a path of nearest neighbors. For a given m we partition the cluster configurations into combinations of clusters of different sizes and shapes. For example if $m = 3$ we consider the partitions $[1,1,1],[1,2],[3]$ which correspond to 3 monomers, a monomer and a dimer, and a trimer, respectively. To the leading order in powers of N, the number of ways these configurations can be placed on the lattice are N^3, N^2 and N, respectively. We assume an individual k-mer is compact, it always its lowest energy state. On a 2D lattice this corresponds to an energy as a function of the cluster size k

$$\mathcal{H}(k) = J \begin{cases} k + \lfloor k/2 - 2 \rfloor & : \quad k > 1 \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.4}$$

where $\lfloor x \rfloor$ is the floor function. The partition function can then be written as

$$\mathcal{Z} = \sum_{p \in \text{partitions(m)}} N^{|p|} \exp\left(-\beta \sum_{k \in p} \mathcal{H}(k) \right) \tag{5.5}$$

This leads to several interesting properties. When m is even it appears that there is a single first-order phase transition at $\beta_1 = \frac{2m-2}{3m-4}$. When m is odd and $m > 3$ there are phase transitions at β_1 and $\beta_2 = 1$. These phase-transitions happen as the spontaneous collapse of the roots of a m^{th} order polynomial equation. The type of phase-transition is logarithmic in N.

5.2.1 First-order pentamer model

We illustrate the method with a worked example of the first-order theory with $m = 5$. The set of partitions of 5 particles (monomers to pentamers) is given by

$$\text{partitions}(5) = \{[1,1,1,1,1],[1,1,1,2],[1,2,2],[1,1,3],[2,3],[1,4],[5]\} \tag{5.6}$$

The energy of a single polymer in its configuration of maximally compact shape is given by Equation 5.4: the monomers have energy proportional to 0, dimers 1, trimers 2, tetramers 4, 5-mers 5. Scaling the energy as in Equation 5.3, this gives the partition function as

$$\mathcal{Z} = N^5 + N^4 \mathrm{e}^{\beta \ln N} + 2\,N^3 \mathrm{e}^{2\,\beta \ln N} + N^2 \mathrm{e}^{3\,\beta \ln N} + N^2 \mathrm{e}^{4\,\beta \ln N} + N\,\mathrm{e}^{5\,\beta \ln N} \tag{5.7}$$

Using units where $k_B = 1$, the specific heat is $C_V = \beta^2 \frac{\partial^2}{\partial \beta^2} \ln \mathcal{Z}$, we seek the discontinuities of C_V as $N \to \infty$. Dropping the factor of β^2 this amounts to finding the discontinuities $C_V(\beta_C) = \infty$

$$\beta_C = \lim_{N \to \infty} \frac{1}{\ln N} \ln(\text{Roots}(f_5(x))) \tag{5.8}$$

with the polynomial $f_5(x)$ defined as

$$f_5(x) = n^5 + n^4 x + 2n^3 x^2 + n^2 x^3 + n^2 x^4 + n x^5 \tag{5.9}$$

The coefficients to this polynomial come from the terms in the energy function. For this particular problem the solutions to equation 5.9 are

$$\beta_C = \left\{1, \frac{3}{4}\right\} \tag{5.10}$$

With a four-fold degeneracy on the 3/4 root (see Figure 5.1).

In the original work by Yang and Lee,[126] they considered the grand partition function in terms of the fugacity. By extending the fugacity to the complex plane they used it to show that an Ising-like

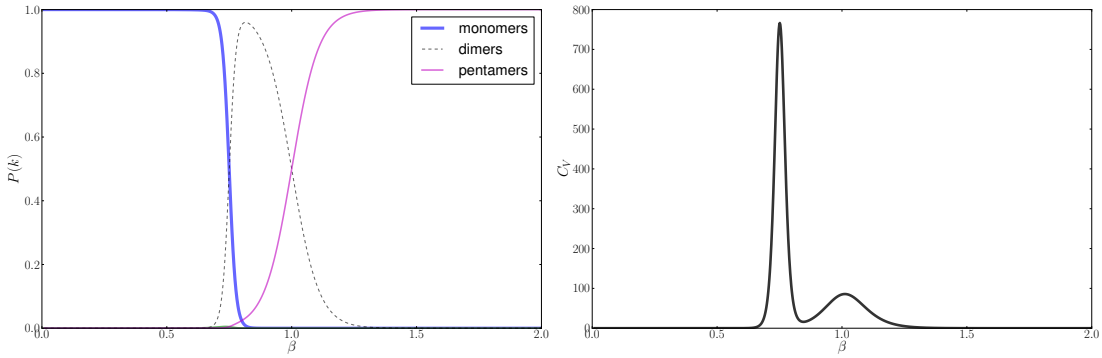


Figure 5.1: For the first-order aggregation model, the first graph shows the probability of each k-mer as a function of β for the different conformations for the specific case of $m = 5$. The second graph shows the specific heat as a function of β for large N ($N = 10^{16}$). The large peak represents the first aggregation of monomers to dimers while the smaller peak signifies the collapse of the system to the lowest energy state of the pentamer. The trimers and tetramers have a low probability for all temperatures and are only visible near the transition temperature.

model will have complex roots on a unit circle. The expansion above is a continuation of this idea, except the parameter extended to the complex plane is β and the zeros are known as Fisher zeros.[127] In this case (as in many others), the zeros do not lie on a unit circle and produce intricate structures on the complex plane.

In Figure 5.2 we plot the phase angle of $\mathcal{Z}$ when temperature is continued onto the complex plane. The zeros of this function signify phase changes of the system. These zeros are visible on the complex plots by noting that around the poles the phase will change by a multiple of 2π.

We list the values of the function f_n for several small n

$$f_1(n) = n \tag{5.11}$$

$$f_2(n) = n^2 + nx$$

$$f_3(n) = n^3 + n^2x + nx^2$$

$$f_4(n) = n^4 + xn^3 + 2x^2n^2 + x^4n$$

$$f_5(n) = n^5 + xn^4 + 2x^2n^3 + x^3n^2 + x^4n^2 + x^5n$$

$$f_6(n) = n^6 + xn^5 + 2x^2n^4 + 2x^3n^3 + x^4n^3 + x^4n^2 + 2x^5n^2 + x^7n$$

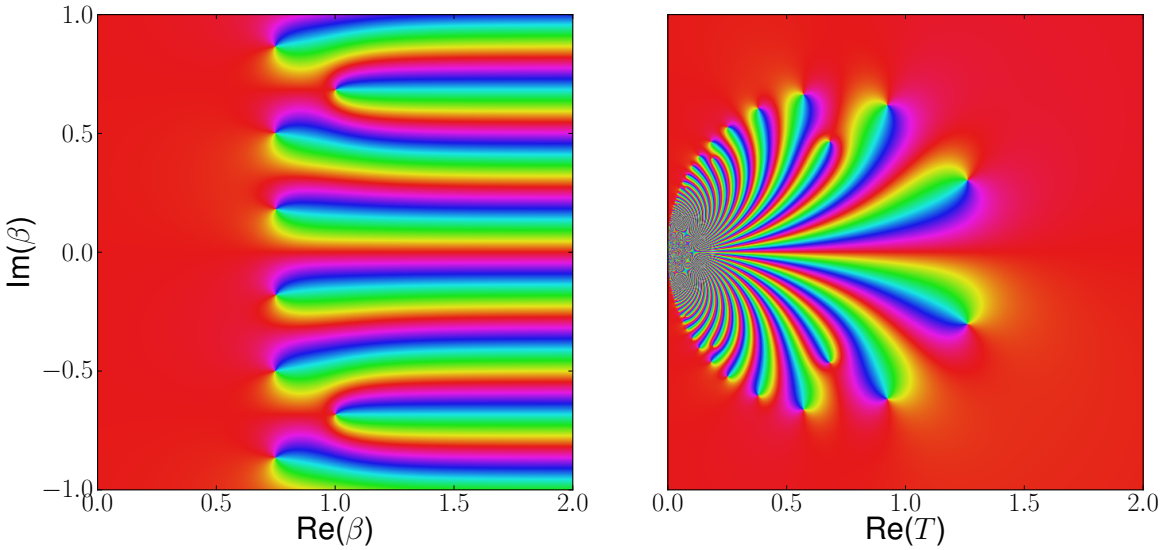


Figure 5.2: Partition function of Equation 5.7 with β analytically continued onto the complex plane. The colors of the graph mark the phase angle of the function, thus simple poles are marked by a complete change of color (due to Cauchy's theorem). The zeros at $\beta = 1$ and $3/4$ can clearly be seen, but they have not yet collapsed onto the real plane due to finite N (here $N = 10^4$). Increasing N alters the density of the zeros until they touch the real axis as N goes to infinity.

5.2.2 Drawbacks of the First-Order Model

The model is interesting, but lacks the fidelity needed to model real aggregated proteins due to the approximations made. For one, the treatment of a protein as a single unit with uniform bond-strengths prevents the expression of any non-isotropic growth patterns. This is a severe limitation, as many fibril growths and aggregated species have a preferred growth direction after a minimal nucleation size. Along those lines, the assumption that the protein will aggregate only in the most compact and hence lowest energy conformation (via Equation 5.4), may be valid for protein folding but is certainly invalid for many types of aggregated species.

The scaling relation, Equation 5.3, was introduced in an ad-hoc manner to help retain a finite convergence of the critical temperatures. The motivating idea behind the approximations was to reduce the problem to the expression of a single polynomial whose properties could be studied analytically. Equation 5.11 lists several values of the polynomial series f_m, but we did not find any

interesting properties of the system. It is unknown if there is a deeper relation between this general aggregation model and this series.

Despite the shortcomings of this simple aggregation model, it provides an instructive tool to develop a more general method. To account for these approximations, we introduce more degrees of freedom into the system. We let the conformations take arbitrary shapes and potentials through a Potts model and define the boundary conditions by the graph the system is embedded in. We attempt to develop a solution to the Potts model over an arbitrary graph and discuss the possibilities this opens for modeling in the next section.

5.3 Potts-Model over Arbitrary Graphs

Computing the partition function $\mathcal{Z}$ of a model with short-range interactions is a standard problem in statistical mechanics. Indeed, the Ising/Potts model has attracted attention from fields as diverse as condensed matter to biophysics to economics.[31] The prototypical example is the Potts model, defined over a graph G with N vertices, edge set $\mathcal{E}$, and vertex set $\vec{\sigma}$.

$$\vec{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_N] \tag{5.12}$$

Each vertex $\sigma_i \in \vec{\sigma}$ has a 'spin' property.[1] In contrast to the Ising model, each spin can point in q different directions

$$\sigma_i \in \{1, 2, 3, \ldots, q\}$$

The Hamiltonian for the Potts model on a simple graph can be defined as

$$\mathcal{H} = -J \sum_{\langle ij \rangle} \delta(\sigma_i, \sigma_j) - h \sum_{\langle i \rangle} \delta(\sigma_i, \alpha) \tag{5.13}$$

where the first summation goes over all edges, while the second summation goes over each vertex. Here δ is the Dirac delta function, J is the strength of the bond interaction and h is the strength of the external field (which acts only on spins pointing in direction α). The system is fully specified by

[1] We use the term spin to keep with the original terminology of the Ising model. In our case the spin is not related to the electronic states of an atom, but rather a marker of the internal conformation states of the protein.

the arrangement of the spin at each vertex. Let $\sigma' \in \vec{\sigma}$ denote a particular set of spin conformations from the set of all arrangements. The partition function for the system is a sum over all such possible spin arrangements

$$\mathcal{Z} = \sum_{\sigma' \in \vec{\sigma}} e^{-\beta \mathcal{H}(\sigma')} \tag{5.14}$$

We can expand the expression by noting that

$$\begin{aligned} \mathcal{Z} &= \sum_{\sigma' \in \vec{\sigma}} e^{\beta J \sum_{\langle ij \rangle} \delta(\sigma'_i, \sigma'_j)} e^{\beta h \sum_{\langle i \rangle} \delta(\sigma'_i, \alpha)} \\ &= \sum_{\sigma' \in \vec{\sigma}} \prod_{\langle ij \rangle} \left[1 + v\delta(\sigma'_i, \sigma'_j)\right] \prod_{\langle i \rangle} \left[1 + u\delta(\sigma'_i, \alpha)\right] \end{aligned}$$

With the understanding that $v = e^{\beta J} - 1$, and $u = e^{\beta h} - 1$. This gives a polynomial in u and v that reduces down to several other well known polynomials such as the Tutte and chromatic polynomial.[128]

By definition, the partition function is the weighted sum over all possible states. Quite often physical problems can be approximated by a discrete (or countably infinite) set of interactions. The graph representing the interactions needs not be regular, and finding the exact solution has many direct benefits. The subgraph expansion method presented here is somewhat abstract, thus it will be worthwhile to have a toy system to illustrate each step of the process. We will apply the methods to a the simple three vertex graph illustrated in Figure 5.3.

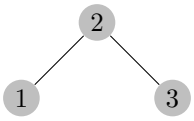


Figure 5.3: Simple three vertex graph used in Section 5.3 to illustrate the subgraph decomposition.

We first consider the case where there is no external field, $u = 1$. Expanding out the partition function (with $\sigma_{ij} \equiv \delta(\sigma_i, \sigma_j)$ and $\bar{\sigma}_i \equiv \delta(\sigma_i, \alpha)$) for our toy model gives

$$\mathcal{Z}_{\text{toy}} = \sum_{\sigma' \in \vec{\sigma}} (1 + v\sigma_{12})(1 + v\sigma_{23}) \tag{5.15}$$

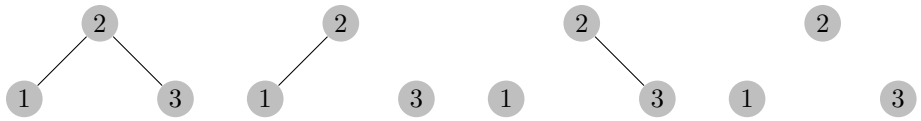


Figure 5.4: The four possible subgraphs of a graph with edge set $\mathcal{E} = \{(1,2),(2,3)\}$. The cluster counting functions (Eq. 5.30) associated with each subgraph from left to right are χ_{32}, $\chi_{21}\chi_{10}$, $\chi_{21}\chi_{10}$, χ_{10}^3.

$$= \sum_{\sigma' \in \vec{\sigma}} 1 + v(\sigma_{12} + \sigma_{23}) + v^2(\sigma_{12}\sigma_{23}) \tag{5.16}$$

Observe in Figure 5.4 the four subgraphs of $A \subseteq G$. These four graphs correspond exactly to the four terms of the partition function of Eq. 5.16. Each subgraph divides the graph into a collection of clusters, disjoint pieces that are not connected. Each term in the expansion survives only if *all* of the delta functions inclusive to that cluster survive. For the empty graph with no edges, there are clearly q^3 ways of choosing the spins so the term survives. For the vertices with one edge, or identically two clusters, there are only q^2 ways to choose the spins as the requirement of a cluster with one edge forces two spins to be the same. For the subgraph with all the edges there are only q ways to arrange the spins as each vertex must have the same spin. Thus each subgraph has a multiplicity over the spin states of $q^{k(A)}$ where $k(A)$ is the number of clusters in the subgraph.

Note in the expansion of $(1 + v\sigma_{ij})$, each time a σ term is picked up, a v term is as well and each σ term corresponds to an edge in the subgraph. With $A \subseteq G$ representing a particular subgraph and $e(A)$ denoting the number of edges in a subgraph, the partition function can be written as a sum over all the subgraphs rather than the vertices and edges

$$\mathcal{Z} = \sum_{A \subseteq G} q^{k(A)} v^{e(A)} \tag{5.17}$$

This is the expansion first discovered by Fortuin-Kasteleyn.[129] When this equation is applied to our toy system we get

$$\mathcal{Z}_{\text{toy}} = \sum_{\sigma' \in \vec{\sigma}} 1 + v(\sigma_{12} + \sigma_{23}) + v^2(\sigma_{12}\sigma_{23}) \tag{5.18}$$

$$= \sum_{\sigma' \in \vec{\sigma}} \left(1 \prod_{j=1}^{q} \prod_{j'=1}^{q} \prod_{j''=1}^{q} \delta(\sigma_1, j)\delta(\sigma_2, j')\delta(\sigma_3, j'') \right.$$
$$+ 2v \prod_{j=1}^{q} \prod_{j'=1}^{q} \delta(\sigma_1, j)\delta(\sigma_2, j)\delta(\sigma_3, j')$$
$$\left. + \; v^2 \prod_{j=1}^{q} \delta(\sigma_1, j)\delta(\sigma_2, j)\delta(\sigma_3, j) \right)$$

when the delta terms are taken into account this gives

$$\mathcal{Z}_{\text{toy}} = q^3 + 2vq^2 + v^2 q \tag{5.19}$$

We can compute the partition function in the presence of an external field by a similar trick. Let us first consider the second subgraph in Figure 5.4, the graph with only a single edge joining vertices 2 and 3. There are two clusters $k(A) = 2$, the cluster with one edge and vertices v_1, v_2 and the singleton cluster with vertex v_3. We consider the expansion

$$\prod_{i=1}^{2} [1 + u\delta(\sigma_i', \alpha)] = [(1 + u\bar{\sigma}_2)(1 + u\bar{\sigma}_3)]\,[(1 + u\bar{\sigma}_1)]$$
$$= \left[1 + u(\bar{\sigma}_2 + \bar{\sigma}_3) + u^2(\bar{\sigma}_2\bar{\sigma}_3)\right]\,[(1 + u\bar{\sigma}_1)]$$

For the first bracketed term to survive in a subgraph expansion, all spins must be pointing in the same direction (namely that of α) which gives a degeneracy of q for the subgraph cluster. If, say $\bar{\sigma}_2 \neq \alpha$ then the term reduces to 1. Let the set of clusters in subgraph A be $C_A = \{c_1, c_2, \ldots, c_{k(A)}\}$ and the number of vertices in a cluster as $n(c)$, we can rewrite the expansion as a product of the clusters:

$$\prod_{c \in C_A}^{k(A)} \left[w^{n(c)} - 1 + q \right] \tag{5.20}$$

Using the simpler form with $w = e^h$. This gives us the expansion by Wu[130] for the Potts model

with an external field acting over a single spin by summing over the subgraphs:

$$\mathcal{Z} = \sum_{A \subseteq G} v^{e(A)} \prod_{c \in C_A}^{k(A)} \Big[w^{n(c)} - 1 + q \Big] \tag{5.21}$$

5.3.1 Generalization to arbitrary field

It is quite simple to let a field act in multiple directions. Instead of having a single α with strength h, we let $\vec{a} = \{h_1, h_2, \cdots, h_q\}$ be the vector that denotes the strength of the field acting on each spin direction. Since the fields are independent of each other (i.e. the contribution from h_1 is independent of h_2, etc.) the derivation is identical giving

$$\sum_{A \subseteq G} v^{e(A)} \prod_{c \in C_A}^{k(A)} \left[q + \sum_{j=1}^{q} \Big[(u_j + 1)^{n(c)} - 1 \Big] \right] \tag{5.22}$$

Letting $w_i = e^{h_i} = u_i + 1$:

$$\begin{aligned} \mathcal{Z} &= \sum_{A \subseteq G} v^{e(A)} \prod_{c \in C_A}^{k(A)} \left[q + \sum_{j=1}^{q} w_j^{n(c)} - \sum_{j=1}^{q} 1 \right] \\ &= \sum_{A \subseteq G} v^{e(A)} \prod_{c \in C_A}^{k(A)} \left[\sum_{j=1}^{q} w_j^{n(c)} \right] \end{aligned} \tag{5.23}$$

As an explicit example, if $q = 2$, $w_1 = e^h$, $w_2 = e^{-h}$, $w_2^{-1} = w_1 = w$, we have the standard Ising model in a field. The energy level spacing between the parallel and anti-parallel spins is 1 rather then 2, but this can be fixed by scaling J.

$$\mathcal{Z}_{\text{Ising}} = \sum_{A \subseteq G} v^{e(A)} \prod_{c \in C_A}^{k(A)} \Big[w^{n(c)} + w^{-n(c)} \Big] \tag{5.24}$$

5.3.2 Generalization to specific bonded interaction

Here we expand the first term when we consider that every interaction does not have the same strength. In this generalization of the Potts model we let the energy of a bonded pair be

$$E_{ij} = J_i \delta(\sigma_i, \sigma_j) = J_i \sigma_{ij} \tag{5.25}$$

Where $\vec{J} = \{J_1, J_2, \cdots, J_q\}$ is a vector of the interaction strengths between spins of the same alignment. Ignoring the field term for a moment (as the term is independent in this partition function), our new Hamiltonian is

$$\mathcal{H} = -\sum_{\langle ij \rangle} \sum_{k}^{q} J_k \sigma_{ij} \delta(i,k) \tag{5.26}$$

The partition function, in a form analogous to those described in previous sections, is

$$\mathcal{Z} = \sum_{\sigma' \in \vec{\sigma}} \left[\prod_{k}^{q} \prod_{\langle ij \rangle} \left(1 + v_k \sigma'_{ij} \sigma'_{ik}\right) \right] \tag{5.27}$$

Where $v_i = e^{\beta J_i} - 1$. For a given cluster in each subgraph $A \subseteq G$, the spins must all be aligned for the term to be non-zero. Each cluster, therefore, contributes for each edge and for all J_i. Thus:

$$\mathcal{Z} = \sum_{A \subseteq G} \prod_{c \in C_A}^{k(A)} \sum_{j}^{q} v_j^{e(c)} \tag{5.28}$$

5.3.3 Full Subgraph Generalization

A q-state Potts model, defined over a graph G with specific bonded interactions defined by $\vec{J}$ and an external field term acting on each spin direction with strength $\vec{a}$ has a partition function expressed as the sum over its subgraphs. When considering the product over the subgraphs we note that each associated subgraph has to have all the spins aligned, thus terms like $v_1 w_2 = 0$ in a subgraph

multiplication. This gives

$$\mathcal{Z} = \sum_{A \subseteq G} \prod_{c \in C_A}^{k(A)} \sum_{j}^{q} \left(w_j^{n(c)} v_j^{e(c)} \right) \tag{5.29}$$

Since we are counting properties of the clusters, we define the cluster counting function to count a cluster with n vertices and e edges as

$$\chi_{ne} = \sum_{j}^{q} \left(w_j^n v_j^e \right) \tag{5.30}$$

A particular subgraph is the product of its cluster counting functions, while the subgraph expansion is the sum of all such products. In Figure 5.4 we list these terms for each subgraph of our toy model as an example. The full expansion of the Potts model with external field can then be expressed as a multivariate polynomial of the cluster counting functions

$$\mathcal{Z} = \sum_{A \subseteq G} \prod_{c \in C_A}^{k(A)} \chi_{n(c),e(c)} \tag{5.31}$$

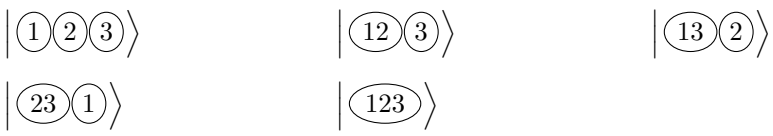
5.4 Operator approach

Given a graph with N vertices labeled $v_1, v_2, \ldots, v_N$, it will be useful to refer to a subset of these vertices. Let the i^{th} subset be denoted $p_i = \{v_{i_1}, v_{i_2}, \ldots\}$. We associate with this subset two additional pieces of information and call it a subgraph partition, or simply a partition for short. These pieces of information are running tally of the edges and vertices as the operators process the vertices. These counts, e_i and n_i, are not representative of current state of the partition, rather they represent placeholders for a counting algorithm.

We denote the i^{th} partition by $P_i \equiv \textcircled{p_i}_{n_i}^{e_i}$. For brevity in notation we omit the indices $\textcircled{b_i}_0^0 = \textcircled{b_i}$ if the block has a zero count $e_i = n_i = 0$. A basis state is specified by a collection of these partitions. Given a state $|\psi\rangle$ with k partitions

$$|\psi\rangle = \left| \sum_{i=1}^{k} \textcircled{p_i}_{n_i}^{b_i} \right\rangle = \left| \sum_{i=1}^{k} P_i \right\rangle \tag{5.32}$$

For example, the 5 basis states for three vertices $1, 2, 3$ are



Each of these partitions represent a possible connection in the graph. A linear superposition of the basis states is simply called a state.

The idea is to write down a set of linear operators, that when applied to a specific graph, they reproduce all the subgraphs along with the prefactors needed for the partition function calculation. Our study was motivated by a recent paper by Bendini and Jacobsen[131] who gave an operator expansion to handle the Potts model without external fields. Their operators were

- The join operator

$$\mathbf{J}_{ij} \left| \textcircled{i}\textcircled{j} \right\rangle = \left| \textcircled{ij} \right\rangle \tag{5.33}$$

$$\mathbf{J}_{ij}^2 = \mathbf{J}_{ij} \tag{5.34}$$

- The deletion operator

$$\mathbf{D}_i \left| \textcircled{i}\textcircled{j}\ldots \right\rangle = q \left| \textcircled{j}\ldots \right\rangle \tag{5.35}$$

$$\mathbf{D}_i \left| \textcircled{ij}\ldots \right\rangle = \left| \textcircled{j}\ldots \right\rangle \tag{5.36}$$

Note that the deletion operator applies the factor q when the cluster is destroyed, thus it counts the number of clusters in the subgraph. This gives, as expected, a factor of $q^{k(A)}$. When a vertex is processed, for each edge (i, j) connected to it they applied the operator $(\mathbf{1} + v\mathbf{J}_{ij})$ and finally the operator $\mathbf{D}_i$ at the end of all the joins.

5.4.1 Operators with External Fields

To compute $\mathcal{Z}$ with a field term we need to keep track of the number of vertices and edges in each cluster. Unfortunately the previous formulation by Bendini and Jacobsen destroys this information. We modify the state vectors so that each partition has two additional piece of information, $n(c)$ and $e(c)$. The new operators are defined to keep track of the pieces of information for each cluster. In all, we have four operators: vertex addition, removal, partition joins and global vertex renaming. The vertex renaming renaming operator is optional, but it facilitates the creation of composite operators that are useful for graphs with symmetry. The operators are

- Adding a new vertex: Adding a vertex to the state simply creates a new partition with a single vertex. Since this marks the beginning of a new cluster, the edge and vertex count is set to zero.

$$\mathbf{A}_i \,|\, \psi\rangle = \left| \overset{0}{\underset{0}{\left(\{v_i\}\right)}} + \psi \right\rangle \tag{5.37}$$

- Joining two partitions: If $v_i \in p_1$ and $v_j \in p_2$ are the vertices to be joined, the partitions are combined by a set union and the vertices and edge counts are totaled together. An extra edge is added to reflect the new edge added to the partition.

$$\mathbf{J}_{ij} \,|\, \psi\rangle = \left| \overset{e_1+e_2+1}{\underset{n_1+n_2}{\left(p_1 \cup p_2\right)}} + \psi - P_1 - P_2 \right\rangle \tag{5.38}$$

- Vertex removal: Removing a vertex can have two effects. Let $v_i \in p_1$ be the vertex to be removed. If the partition has more vertices than the one being removed $|p_1| > 1$, we remove the vertex and increment the vertex count by one.

$$\mathbf{D}_i \,|\, \psi\rangle = \left| \overset{e_1}{\underset{n_1+1}{\left(p_1 / \{v_i\},\right)}} + \psi - P_1 \right\rangle \tag{5.39}$$

  If the partition is a singleton (that is it contains only v_i), once this last vertex is removed the partition is complete. We can now count its contribution as a full cluster. The singleton partition $\overset{e_1}{\underset{n_1}{\left(p_1\right)}}$ picks up a factor of $\chi_{n_1+1 e_1}$ with the n_1+1 term coming from the final vertex

removal.

$$\mathbf{D}_i \,|\, \psi\rangle = \chi_{n_1+1,e_1} \,|\, \psi - P_1\rangle \tag{5.40}$$

- Vertex relabeling: If the vertices are labeled by $\mathcal{V} = \{v_j, v_{j+1}, v_{j+2}, \ldots\}$, the operator $\mathbf{T}_i \,|\, \psi\rangle$ shifts the labeling of all the vertices by a constant factor, $\mathcal{V} \to \mathcal{V}' = \{v_{i+j}, v_{i+j+1}, v_{i+j+2}, \ldots\}$.

$$\mathbf{T}_i \,|\, \psi(v_1, v_2, v_3, \ldots)\rangle = |\, \psi(v_{1+i}, v_{2+i}, v_{3+i}, \ldots)\rangle \tag{5.41}$$

An individual operator is commutative with others of its type, i.e. $\mathbf{J}_{ij}\mathbf{J}_{mn} = \mathbf{J}_{mn}\mathbf{J}_{ij}$ but two different types of operators are not. This makes physical sense, one can not join a vertex until it is created.

5.4.2 Vertex Processing

With the operators defined, we now need a way to process the vertices of a graph. With the set of edges defined as $\mathcal{E}$ the most naïve approach would be

$$\sum_{v_i \in V} \mathbf{D}_i \sum_{(v_i,v_j)\in\mathcal{E}} (\mathbf{I} + \mathbf{J}_{i,j}) \sum_{v_i \in V} \mathbf{A}_i \,|\, \emptyset\rangle \tag{5.42}$$

Where $\emptyset$ is the empty graph. This is however, horribly inefficient. If there are any common subgraphs, this process ignores any inherent symmetries. A more efficient method would remove vertices as soon as all their contributions were considered. This approach can be encapsulated with a composite operator $\mathbf{L}$. Taking the vertices in lexicographic order we follow the following prescription at the ith step: add the target vertex and all vertices sharing an edge with the target vertex where $i < j$, delete the target vertex and repeat for all vertices

$$\mathbf{L}_i = \mathbf{D}_i \sum_{(v_i,v_j)\in E} (\mathbf{I} + \mathbf{J}_{i,j}) \left[\sum_{(v_i,v_j)\in E, i<j} \mathbf{A}_j \right] \mathbf{A}_i \tag{5.43}$$

An optimal method of vertex processing would use a tree-decomposition of a graph. In a tree-decomposition, the vertices are mapped onto a separate tree graph (a graph with no cycles).[132] That is, the vertices of the tree graph $\{u_1, u_2, \ldots\}$ contain subsets of the vertices of the original graph

$u_1 = \{v_{j_1}, v_{j_2}, \ldots\}$. The vertices are mapped such that each vertex pair representing an edge on the original graph is contained within some vertex in the tree graph. Additionally, if vertex v_1 is contained in both u_1 and u_2 than each vertex along the unique path connecting u_1 and u_2 must also contain v_1. The tree vertex with the largest cardinality $\max(|u_i|)$ is known as the tree-width. While it is $\mathcal{NP}$-hard to determine the optimal tree-decomposition of a graph,[133] once known many graph algorithms that are $\mathcal{NP}$-hard can be computed in polynomial time, bounded by the tree-width.[134]

Given a tree decomposition of the graph, the vertex processing outlined in Equation 5.43 could be greatly improved. To do so however, would require a new operator, one that fuses the partition when joined together on the tree-decomposition. In this chapter we only work with the lexicographic order of processing. For our purposes it straightforward to apply the simpler lexicographic order to the larger ladder graphs considered in the later sections.

Worked example of Figure 5.3

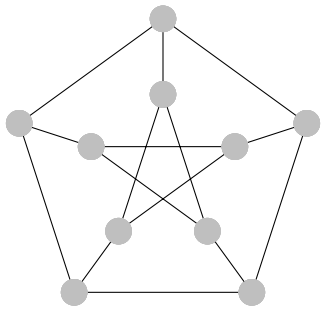
We now work through our toy graph as an example of the operators with a standard path decomposition. The full process is

$$|\,\psi_3\rangle = \mathbf{D}_3\mathbf{D}_2(\mathbf{1}+\mathbf{J}_{23})\mathbf{A}_3\mathbf{D}_1(\mathbf{1}+\mathbf{J}_{12})\mathbf{A}_2\mathbf{A}_1\,|\,\emptyset\rangle$$

The operators applied at each step:

$$\begin{aligned}
\mathbf{A}_2\mathbf{A}_1\,|\,\emptyset\rangle &= \Big|\,\textcircled{1}\textcircled{2}\,\Big\rangle = |\,\psi_0\rangle \\
\mathbf{D}_1(\mathbf{1}+\mathbf{J}_{12})\,|\,\psi_0\rangle &= \mathbf{D}_1\left(\Big|\,\textcircled{1}\textcircled{2}\,\Big\rangle + \Big|\,\textcircled{12}^{\,1}\,\Big\rangle\right) = \chi_{10}\Big|\,\textcircled{2}\,\Big\rangle + \Big|\,\textcircled{2}^{\,1}_{\,1}\,\Big\rangle = |\,\psi_1\rangle \\
\mathbf{D}_2(\mathbf{1}+\mathbf{J}_{23})\mathbf{A}_3\,|\,\psi_1\rangle &= \mathbf{D}_2\left(\chi_{10}\Big|\,\textcircled{2}\textcircled{3}\,\Big\rangle + \Big|\,\textcircled{2}^{\,1}_{\,1}\textcircled{3}\,\Big\rangle + \chi_{10}\Big|\,\textcircled{23}\,\Big\rangle + \Big|\,\textcircled{23}^{\,2}_{\,1}\,\Big\rangle\right) \\
&= (\chi_{10}^2+\chi_{21})\Big|\,\textcircled{3}\,\Big\rangle + \chi_{10}\Big|\,\textcircled{3}^{\,1}_{\,1}\,\Big\rangle + \Big|\,\textcircled{3}^{\,2}_{\,2}\,\Big\rangle = |\,\psi_2\rangle \\
\mathbf{D}_3\,|\,\psi_2\rangle &= \chi_{10}^3 + 2\chi_{21}\chi_{10} + \chi_{32}
\end{aligned}$$

Subgraph decomposition of the Petersen Graph



The Petersen graph was constructed to be the smallest bridgeless cubic graph with no three-edge-coloring.[135] It is a well known graph that often appears as an example or counterexample of graph properties.[136] As such, we feel the it makes for a good example of the application of the methods. Listed below is the full subgraph decomposition for the Petersen graph, a full solution for the Potts model on it with arbitrary bond strengths and external fields

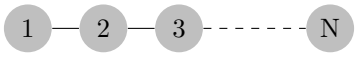
$$
\begin{aligned}
\mathcal{Z}_{\text{petersen}} = {} & (\chi_{1,0}^{10} + \chi_{10,15}) + 6(\chi_{5,5}^{2} + \chi_{2,1}^{5}) + 10\chi_{1,0}\chi_{9,12} + 12\chi_{1,0}^{5}\chi_{5,5} \\
& + 15(\chi_{1,0}^{8}\chi_{2,1} + \chi_{1,0}^{2}\chi_{8,10} + \chi_{10,14} + \chi_{8,10}\chi_{2,1}) + 30(\chi_{1,0}^{7}\chi_{3,2} + \chi_{3,2}\chi_{7,8} + \chi_{1,0}^{3}\chi_{7,8}) \\
& + 60(\chi_{3,2}\chi_{5,5}\chi_{2,1} + \chi_{1,0}\chi_{5,5}\chi_{4,3} + \chi_{1,0}\chi_{7,8}\chi_{2,1} + \chi_{1,0}^{3}\chi_{5,5}\chi_{2,1} \\
& + \chi_{1,0}\chi_{5,5}\chi_{2,1}^{2} + \chi_{1,0}^{2}\chi_{3,2}\chi_{5,5} + \chi_{5,4}\chi_{5,5} + \chi_{2,1}^{2}\chi_{6,6}) \\
& + 70(\chi_{1,0}^{4}\chi_{6,6} + \chi_{4,3}\chi_{6,6} + \chi_{1,0}^{6}\chi_{4,3}) + 75\chi_{1,0}^{6}\chi_{2,1}^{2} + 80\chi_{4,3}\chi_{2,1}^{3} \\
& + 90\chi_{1,0}^{2}\chi_{2,1}^{4} + 105\chi_{10,13} + 110\chi_{1,0}\chi_{3,2}^{3} + 120\chi_{1,0}\chi_{9,11} + 135\chi_{3,2}^{2}\chi_{2,1}^{2} + 145\chi_{1,0}^{4}\chi_{2,1}^{3} \\
& + 150(\chi_{4,3}^{2}\chi_{2,1} + \chi_{3,2}^{2}\chi_{4,3} + \chi_{8,9}\chi_{2,1} + \chi_{1,0}\chi_{3,2}\chi_{6,6}) + 165\chi_{1,0}^{4}\chi_{3,2}^{2} \\
& + 180(\chi_{1,0}^{5}\chi_{5,4} + \chi_{8,9}\chi_{1,0}^{2}) + 210(\chi_{1,0}^{2}\chi_{2,1}\chi_{6,6} + \chi_{5,4}^{2}) \\
& + 240(\chi_{1,0}\chi_{3,2}\chi_{2,1}^{3} + \chi_{1,0}^{5}\chi_{3,2}\chi_{2,1} + \chi_{3,2}\chi_{7,7}) + 270\chi_{1,0}^{2}\chi_{4,3}^{2} + 300\chi_{1,0}^{3}\chi_{7,7} \\
& + 315\chi_{6,5}\chi_{2,1}^{2} + 360\chi_{6,5}\chi_{4,3} + 420\chi_{1,0}^{4}\chi_{4,3}\chi_{2,1} + 435\chi_{1,0}^{4}\chi_{6,5} + 445\chi_{10,12} \\
& + 480(\chi_{3,2}\chi_{5,4}\chi_{2,1} + \chi_{1,0}^{3}\chi_{3,2}\chi_{4,3} + \chi_{1,0}^{2}\chi_{3,2}^{2}\chi_{2,1}) + 510\chi_{1,0}^{3}\chi_{3,2}\chi_{2,1}^{2} \\
& + 540\chi_{1,0}\chi_{7,7}\chi_{2,1} + 570\chi_{1,0}^{2}\chi_{4,3}\chi_{2,1}^{2} + 600(\chi_{1,0}\chi_{5,4}\chi_{2,1}^{2} + \chi_{1,0}\chi_{5,4}\chi_{4,3}) \\
& + 615\chi_{8,8}\chi_{2,1} + 630(\chi_{9,10}\chi_{1,0} + \chi_{3,2}\chi_{7,6}) + 720\chi_{5,4}\chi_{1,0}^{2}\chi_{3,2}
\end{aligned}
$$

$$+ 780(\chi_{1,0}^3\chi_{5,4}\chi_{2,1} + \chi_{4,3}\chi_{1,0}\chi_{3,2}\chi_{2,1}) + 840\chi_{1,0}\chi_{3,2}\chi_{6,5} + 855\chi_{1,0}^2\chi_{8,8}$$

$$+ 950\chi_{1,0}^3\chi_{7,6} + 1080\chi_{8,7}\chi_{2,1} + 1230(\chi_{1,0}^2\chi_{6,5}\chi_{2,1} + \chi_{10,11}) + 1560\chi_{1,0}\chi_{7,6}\chi_{2,1}$$

$$+ 1710\chi_{1,0}^2\chi_{8,7} + 1780\chi_{1,0}\chi_{9,9} + 2000\chi_{10,9} + 2172\chi_{10,10} + 2400\chi_{1,0}\chi_{9,8}$$

5.5 One-dimensional $1N$ Ladder



We apply the method to the one-dimensional strip that has a thickness of one lattice unit. The system displays a similarity as we move along the strip. Moving horizontally, this defines the operator

$$\mathbf{L} = \mathbf{T}_{-1}\mathbf{D}_1(\mathbf{I} + \mathbf{J}_{12})\mathbf{A}_2 \tag{5.44}$$

here $\mathbf{I}$ is the identity operator and $\mathbf{T}$ is the shift operator, Equation 5.41. To determine the behavior of the system, we apply $\mathbf{L}$ to an arbitrary state. At each iteration, there is only one item in the partition, representing the cluster that has not been fully counted yet. We define a notational shorthand for this term as $\left|\textcircled{1}_a^b\right\rangle = g_{ab}$. Since our solution will be constructed with generating functions we use the following notation, if Z_n is the partition function for a chain of length n then $Z(y)$ is its generating function

$$Z_n = \left.\frac{\partial^n Z(y)}{\partial y^n}\right|_{y=0} \tag{5.45}$$

The partition function for a $1N$ ladder expressed in the operator notation is

$$Z_n = \mathbf{D}_1\mathbf{L}^{N-1}g_{00} \tag{5.46}$$

While a single application of the operator is

$$\mathbf{L}g_{ab} = g_{a+1,b+1} + \chi_{a+1,b}g_{00} \tag{5.47}$$

Let $\phi_n = \mathbf{L}^n g_{00}$ be the intermediate step of our calculation, i.e. the n^{th} application of the operator. Defining the coefficients of ϕ_n as

$$\phi_n = \sum_{ab \geq 0} c_{nab} g_{ab} \tag{5.48}$$

These coefficients c_{nab} are zero for $a < 0$, $b < 0$ or $n < 0$ since the operators can only increment a and b but never reduce them below the starting value of zero. We have in terms of ϕ

$$\begin{aligned} \phi_n &= \sum_{ab \geq 0} c_{nab} g_{ab} \\ &= \mathbf{L}\phi_{n-1} = \sum_{ab} c_{n-1,ab} \mathbf{L} g_{ab} \\ &= \sum_{ab \geq 0} c_{n-1,ab} (g_{a+1,b+1} + \chi_{a+1,b} g_{00}) \end{aligned} \tag{5.49}$$

By inverting the equation, we can solve for the coefficients

$$c_{nab} = \begin{cases} \sum_{ij} c_{n-1,i,j} \chi_{i+1,j} & a = b = 0 \\ c_{n-1,a-1,b-1} & \text{otherwise} \end{cases} \tag{5.50}$$

Since $c_{nab} = c_{n-1,a-1,b-1}$ for nonzero a, b all terms of the type $a \neq b$ are zero. To see why, consider as an example c_{n32}

$$c_{n32} = c_{n21} = c_{n10} = c_{n,0,-1} = 0 \tag{5.51}$$

Thus we write $a = b = k$. We also note that the recurrence relation, when applied multiple times can be rewritten as $c_{nk} = c_{n-k,0}$. This gives

$$\begin{aligned} c_{n,0} &= \sum_{ij} c_{n-1,i,j} \chi_{i+1,j} \\ &= \sum_{k} c_{n-1,k} \chi_{k+1,k} \\ &= \sum_{k} c_{n-k-1,0} \chi_{k+1,k} \end{aligned} \tag{5.52}$$

We can solve for these terms by the use of generating functions. Let

$$C = \sum_{n\geq 0} y^n c_{n,0} \tag{5.53}$$

$$X = \sum_{n\geq 0} y^n \chi_{n+1,n} \tag{5.54}$$

Then by multiplying Equation 5.52 by y^n and summing on both sides (starting at $n = 1$) we get

$$\sum_{n\geq 1} y^n c_{n0} = \sum_{n\geq 1} y^n \sum_k c_{n-k-1,0}\chi_{k+1,k} \tag{5.55}$$

$$yC - c_{00} = yCX$$

$$C = \frac{1}{y(1-yX)}$$

by use of the Cauchy Product rule.[2] To get the partition function, we simply need to delete the last vertex:

$$Z_n = \mathbf{D}_1\phi_n = \sum_{ab} c_{nab}\mathbf{D}_1 g_{ab} \tag{5.56}$$

$$= \sum_{ab} c_{nab}\chi_{a+1,b}$$

$$= \sum_k c_{nk}\chi_{k+1,k}$$

$$= c_{n+1,0}$$

Finally we have the complete generating function for the $1N$ ladder

$$Z(y) = yC = \frac{1}{1-yX} \tag{5.57}$$

GF for 1N Ladder

[2]The Cauchy Product rule gives the product of two generating functions $A = \sum_{n=0}^{\infty} a_n$, $B = \sum_{n=0}^{\infty} b_n$ as $AB = C = \sum_{n=0}^{\infty} c_n$ where $c_n = \sum_{k=0}^{n} a_k b_{n-k}$.

This is a powerful result as it enumerates all of the subgraphs by counting the degeneracy of those that share the same number of edges and vertices. Essentially this graph is solved for any combination of external fields and bonds strengths. The first terms from the series are

$$Z_1 = \chi_{10} \tag{5.58}$$

$$Z_2 = \chi_{10}^2 + \chi_{21} \tag{5.59}$$

$$Z_3 = \chi_{10}^3 + 2\chi_{21}\chi_{10} + \chi_{32} \tag{5.60}$$

$$Z_4 = \chi_{10}^4 + 3\chi_{21}\chi_{10}^2 + 2\chi_{10}\chi_{32}\chi_{21}^2 + \chi_{43} \tag{5.61}$$

5.5.1 Special Case $1N$ Ladder: Potts

The standard Potts model in the absence of an external field with the parameters $w_j = 1, v_j = v$ gives by Equation 5.23 an explicit form of the cluster counting function

$$\chi_{ab} = \sum_{j=1}^{q} v^b = qv^b \tag{5.62}$$

which is similar to the Fortuin-Kasteleyn expansion. This gives the partition function in terms of a generating function

$$Z(y) = \frac{yv - 1}{yv - 1 + yq} \tag{5.63}$$

and closed form

$$Z_n = q(v + q)^{n-1} \tag{5.64}$$

This clearly has no phase transitions since $v_c + q = e^{\beta_c J} - 1 + q = 0$ has no real solutions for $q > 1$, but at $n \to \infty$ when $T \to 0$ there is a phase transition. The specific heat in units of $k_{\mathrm{B}} = 1$, $C_V = \beta^2 \frac{\partial^2}{\partial \beta^2} \ln Z_n$ is

$$C_V = \frac{(n-1)\, J^2 e^{J\beta}(q-1)\beta^2}{\left(e^{J\beta} - 1 + q\right)^2} \tag{5.65}$$

While there are no true phase transitions in this system, there are semi-critical β_c (temperatures which the specific heat peaks $\frac{\partial C_V}{\partial \beta} = 0$, the Schottky anomalies). These temperatures are

$$\beta_c = \frac{2(e^r - 1 - q)}{J(e^r + 1 - q)} \tag{5.66}$$

where r is the root to the following transcendental equation

$$qr - re^r + 2q - r + 2e^r - 2 = 0 \tag{5.67}$$

A plot of the specific heat for $n = 10^6$, $q = 2$, and $J = -1$ is shown in Figure 5.5. The semi-critical transition was numerically calculated to be $\beta_C = 2.3994$.

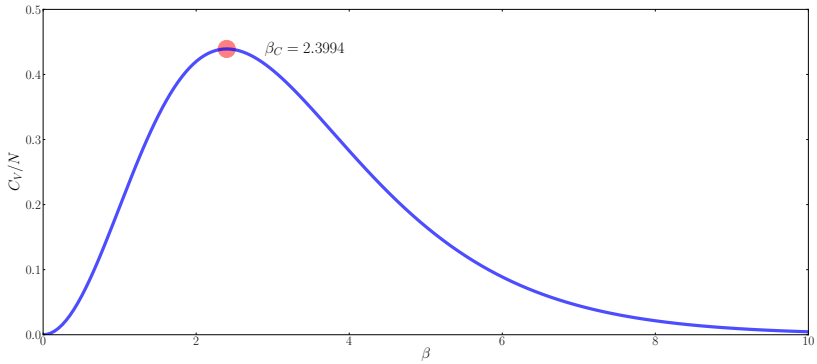


Figure 5.5: Specific heat per chain length for the one-dimensional Potts model with parameters $n = 10^6$, $q = 2$, and $J = -1$. The semi-critical β was calculated from Equation 5.66.

To visualize the behavior of the system on a larger scope, we can analytically continue the temperature onto the complex plane. In Figure 5.6 we can visually inspect the behavior of the zeros on the complex plane. In the picture the semi-critical behavior (since we are plotting at finite n) is visible at $T = 0$.

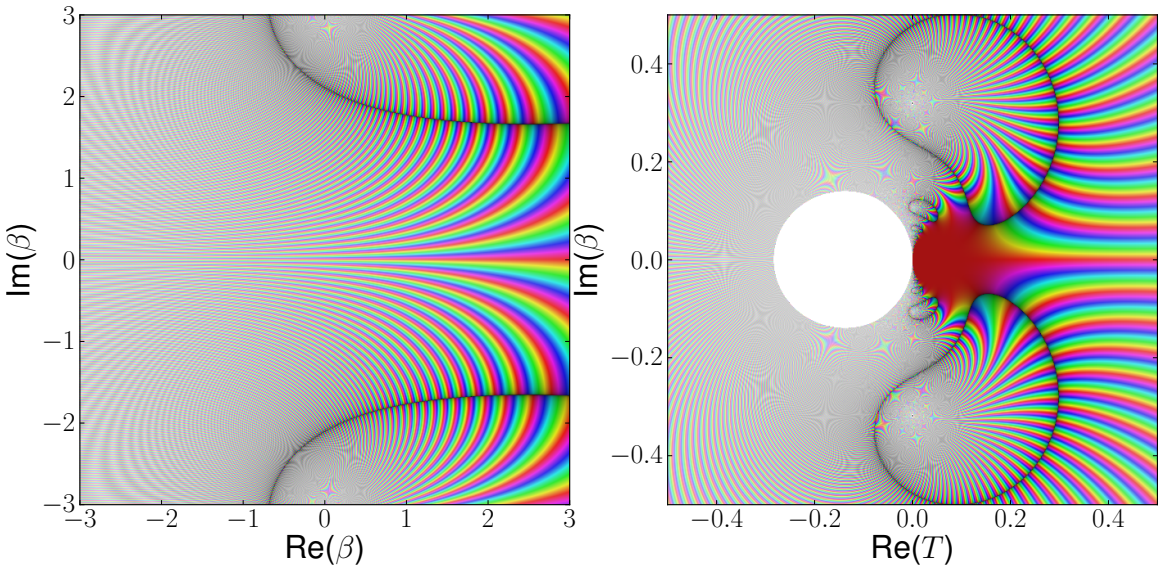


Figure 5.6: Phase angle of the partition function for the $1N$ ladder with $J = -1$, $q = 2$ where the temperature is continued onto the complex plane. The left plot has $\beta = 1/kT$ to show the behavior of the high temperatures and the right plot shows the low temperature behavior. The points where the phase changes rapidly signify zeros of the partition function and hence first-order phase changes. The lighter areas on the low temperature plot are large values of the partition function (not zeros).

5.5.2 Special Case $1N$ Ladder: Potts with Magnetic Field

The standard Potts model with a field term only acting on the first state has the parameters of $w_j = w\delta(j, 1)$ and $v_j = v$. This gives

$$\chi_{a+1,a} = w^{a+1}v^a + \sum_{j=2}^{q} v^a = w^{a+1}v^a + (q-1)v^a \tag{5.68}$$

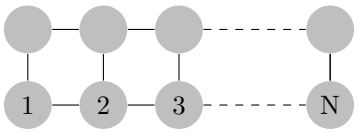
$$X = \sum_{n=0} y^n \chi_{n+1,n} = \frac{-w}{wvy - 1} - \frac{q}{yv - 1} + \frac{1}{yv - 1} \tag{5.69}$$

Giving the generating function as

$$Z(y) = \frac{(wvy - 1)(yv - 1)}{y^2v^2w - yv - wvy + 1 - wy + qwvy^2 - yq + y} \tag{5.70}$$

While resulting expression for $Z_n = \partial_y^n Z(y)\big|_{y=0}$ is too complicated to evaluate explicitly, any particular term can be extracted with ease.

5.6 One-dimensional $2N$ Ladder



We give the general recurrence relation for the $2N$ ladder. The generating function solution remains an open problem. With a one dimensional strip two lattice units thick the operator is

$$\mathbf{L} = \mathbf{T}_{-2}\mathbf{D}_2\mathbf{D}_1(\mathbf{I}+\mathbf{J}_{34})(\mathbf{I}+\mathbf{J}_{13})(\mathbf{I}+\mathbf{J}_{23})\mathbf{A}_3\mathbf{A}_4 \tag{5.71}$$

There are now two general states which we have to consider. The first is when the vertices do not belong to the same partition $\left|\textcircled{1}_{b_1}^{a_1}\textcircled{2}_{b_2}^{a_2}\right\rangle$ and the second when they do belong to the same partition $\left|\textcircled{12}_{b_3}^{a_3}\right\rangle$. We let f_{a_1,b_1,a_2,b_2} and g_{a_3,b_3} represent a shorthand for the two states respectively. The action of the operator is on f is

$$\begin{aligned}\mathbf{L}f_{a_1,b_1,a_2,b_2} &= f_{a_1+1,b_1+1,a_2+1,b_2+1} + \chi_{a_1+1,b_1}f_{0,0,a_2+1,b_2+1} \\ &+ \chi_{a_2+1,b_2}f_{a_1+1,b_1+1,0,0} + \chi_{a_1+1,b_1}g_{a_2+1,b_2+2} \\ &+ \chi_{a_2+1,b_2}g_{a_1+1,b_1+2} + g_{a_1+a_2+2,b_1+b_2+3} \\ &+ \chi_{a_1+1,b_1}\chi_{a_2+1b_1}g_{0,1} + \chi_{a_1+1,b_1}f_{0,0,0,0}\end{aligned} \tag{5.72}$$

And the action on g is given by

$$\begin{aligned}\mathbf{L}g_{a_3,b_3} &= f_{a_3+2,b_3+1,0,0} + f_{0,0,a_3+2,b_3+1} \\ &+ 3g_{a_3+2,b_3+2} + g_{a_3+2,b_3+3} \\ &+ \chi_{a_3+2b_3}g_{0,1} + \chi_{a_3+2,b_3}f_{0,0,0,0}\end{aligned} \tag{5.73}$$

As before we let ϕ_n be the n^{th} application of $\mathbf{L}$ against the initial state

$$\phi_n = \sum_{i,j\geq 0} c_{n,i,j}g_{i,j} + \sum_{i_1,i_2,j_1,j_2\geq 0} d_{n,i_1,j_1,i_2,j_2}f_{i_1,j_1,i_2,j_2} \tag{5.74}$$

with the initial state

$$\phi_0 = g_{01} + f_{0000} \tag{5.75}$$

We note that by symmetry we have

$$d_{n,i_1,j_1,i_2,j_2} = d_{n,i_2,j_2,i_1,j_1} \tag{5.76}$$

Using that information we invert the equations

$$\begin{aligned} c_{n01} &= \sum_{ij} \chi_{i+2,j} c_{n-1,i,j} \\ &+ \sum_{i_1,i_2,j_1,j_2} \chi_{i_1+1,j_1} \chi_{i_2+1,j_2} d_{n-1,i_1,i_2,j_1,j_2} \end{aligned} \tag{5.77}$$

$$\begin{aligned} c_{nab} &= 3c_{n-1,a-2,b-2} + c_{n-1,a-2,b-3} \\ &+ 2\sum_{i,j} \chi_{i+1,j} d_{n-1,i,j,a-2,b-2} \\ &+ 2\sum_{i,j} d_{n-1,i,j,a-i-2,b-j-3} \end{aligned} \tag{5.78}$$

$$d_{naa00} = c_{n-1,a-2,a-1} + S_1(a,a) + S_2(a) \tag{5.79}$$

$$d_{nab00} = c_{n-1,a-2,b-1} + S_1(a,b)$$

$$d_{n,i_1,j_1,i_2,j_2} = d_{n-1,i_1-1,j_1-1,i_2-1,j_2-1}$$

With

$$S_1(a,b) = \sum_{k=0}^{n-1} \chi_{k+1,k} d_{n-1-k,a-1-k,b-1-k,0,0} \tag{5.80}$$

$$S_2(a) = \sum_{i=a} \sum_{j=a} \chi_{i+1,j} d_{n-a,i-a-1,j-a-1,0,0} \tag{5.81}$$

5.7 Edge contraction methods

In the previous section we developed a method to determine the full partition function for an arbitrary graph with a field term using an operator expansion. While this is the most general method, it does limit the depth of the results one can obtain due to the complicated nature of the recursion relations. If we simplify the Hamiltonian by dropping the field term ($h = 0$ in Eq. 5.13), the solution to a wider class of graphs becomes tractable.

An equivalent evaluation to $\mathcal{Z}$ is the recursion relation mediated by graph reductions of edge contraction and removal (see Section 2.3 for an explanation of these methods). There does not appear to be a general formula for the Potts model with an external field, but by relaxing this constraint there is the well-known relation [128]

$$\mathcal{Z}(G) = w_{ij}\mathcal{Z}(G_{/e_{ij}}) + \mathcal{Z}(G_{-e_{ij}}) \tag{5.82}$$

$$\mathcal{Z}(G_1 \sqcup G_2) = \mathcal{Z}(G_1)\mathcal{Z}(G_2) \tag{5.83}$$

$$\mathcal{Z}(\emptyset) = 1 \tag{5.84}$$

$$\mathcal{Z}(S) = q \tag{5.85}$$

Here $/e_{ij}$ is edge contraction, $-e_{ij}$ is edge removal, and w_{ij} is the weight of the edge joined by vertices i, j. $G_1 \sqcup G_2$ represents two graphs that do not share any edges. S is the singleton graph, with only one vertex and no edges and $\emptyset$ is the empty graph.

We can immediately derive the results for the $1N$ chain in the absence of an external field. As a pedagogical bonus, the resulting derivations can be done graphically. With a chain of length n, an edge weight of v for each sequential vertex along the chain, the partition function for the n^{th} chain

is

$$Z(n) = \quad = v\Big(\quad\Big) + \Big(\quad\quad\Big) \tag{5.86}$$

$$= vZ(n-1) + qZ(n-1)$$

$$= Z(n-1)(v+q)$$

Since the recursion relation is linear, it can be solved for in closed form giving exactly Equation 5.64.

$$Z(n) = q(v+q)^{n-1} \tag{5.87}$$

$2N$ Ladder - Directional bonds

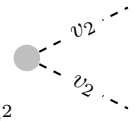
We now attempt to solve the $2N$ chain with bond strengths that vary in along the short or long edge of the ladder. The strengths for the vertical and horizontal directions respectively are J_1 and J_2 (where $v_1 = e^{J_1\beta} - 1$ and $v_2 = e^{J_2\beta} - 1$)

$Z(n) =$ (5.88)

$= v_1$ $+$ (5.89)

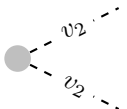
$= v_1 v_2$ $+ Z(n-1)((v_2+q)^2 + v_1(v_2+q))$ (5.90)

$= v_1 v_2^2$ $+ Z(n-1)((v_2+q)^2 + v_1(v_2+q) + v_1 v_2)$ (5.91)

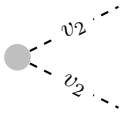


$$= v_1 v_2^2 \qquad (1 + v_1) + Z(n-1)((v_2+q)^2 + v_1(v_2+q) + v_1 v_2) \tag{5.92}$$

We notice however that the final graph is one that has already been seen before in the expansion, just shifted by one term in the recurrence



$$Z(n-1) = v_1 \qquad + Z(n-2)(v_2+q)^2 \tag{5.93}$$



$$= \frac{1}{v_1}(Z(n-1) - Z(n-2)(v_2+q)^2) \tag{5.94}$$

This gives the recursion relation of

$$Z(n) = Z(n-1)((q+v_2)^2 + v_1(q+2v_2)) + (Z(n-1) - Z(n-2)v_2^2(q+v_2)^2(1+v_1) \tag{5.95}$$

The initial conditions are

$$Z(1) = q(v_1 + q) \tag{5.96}$$

$$Z(2) = ((v_1+q)(v_2+q))^2 - (v_1 v_2)^2(q-1) \tag{5.97}$$

This again is a linear recurrence relation, albeit with two terms, that can be solved for rather easily. The closed form solution however is rather verbose and not particularly illuminating we provide the (slightly) more compact generating function

$$Z(y) = \frac{a}{b} \tag{5.98}$$

$$\begin{aligned} a =& -yq(-yv_1v_2^2 - yv_1v_2q - yv_1^2v_2 - yqv_1^2 - v_1 - yq^2v_1 - q + v_1^2v_2^2y + 2yv_1v_2^2q \\ &+ yv_1^2v_2q + yv_1v_2q^2 + v_1^3v_2^2y + v_1^2v_2^2yq) \\ b =& v_1v_2^4y^2 + v_1^2v_2^4y^2 - 2yv_2q - yv_1v_2 - 2yv_1v_2^2 + 1 + 2v_1v_2^3y^2q + 2v_1^2v_2^3y^2q + \\ &y^2v_1v_2^2q^2 + v_1^2v_2^2y^2q^2 - yv_1v_2q - yv_2^2 - yq^2 - v_1^2v_2^2y \end{aligned}$$

GF for 2N Ladder with no field term

Using the closed form solution we can look for the discontinuities attempting to find phase transitions. The solutions for β where $Z(n) = 0$ only have negative values for those on the real axis. It seems that the other roots also do not have real, positive solutions for $q \geq 2$, implying that the system has no mathematical phase transitions in the physical parameter space. This is best visualized in Figure 5.7.

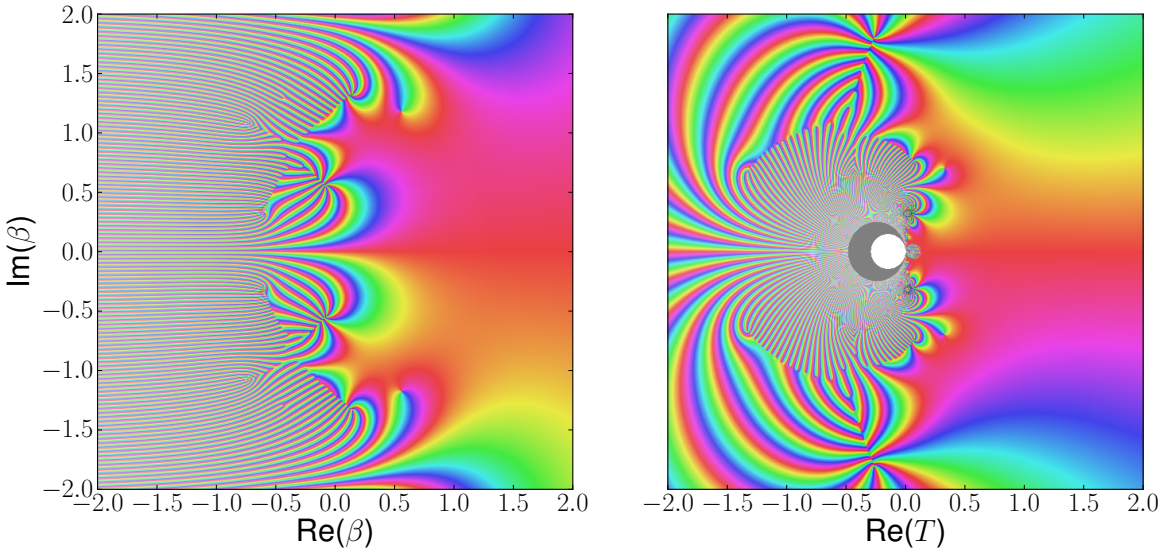
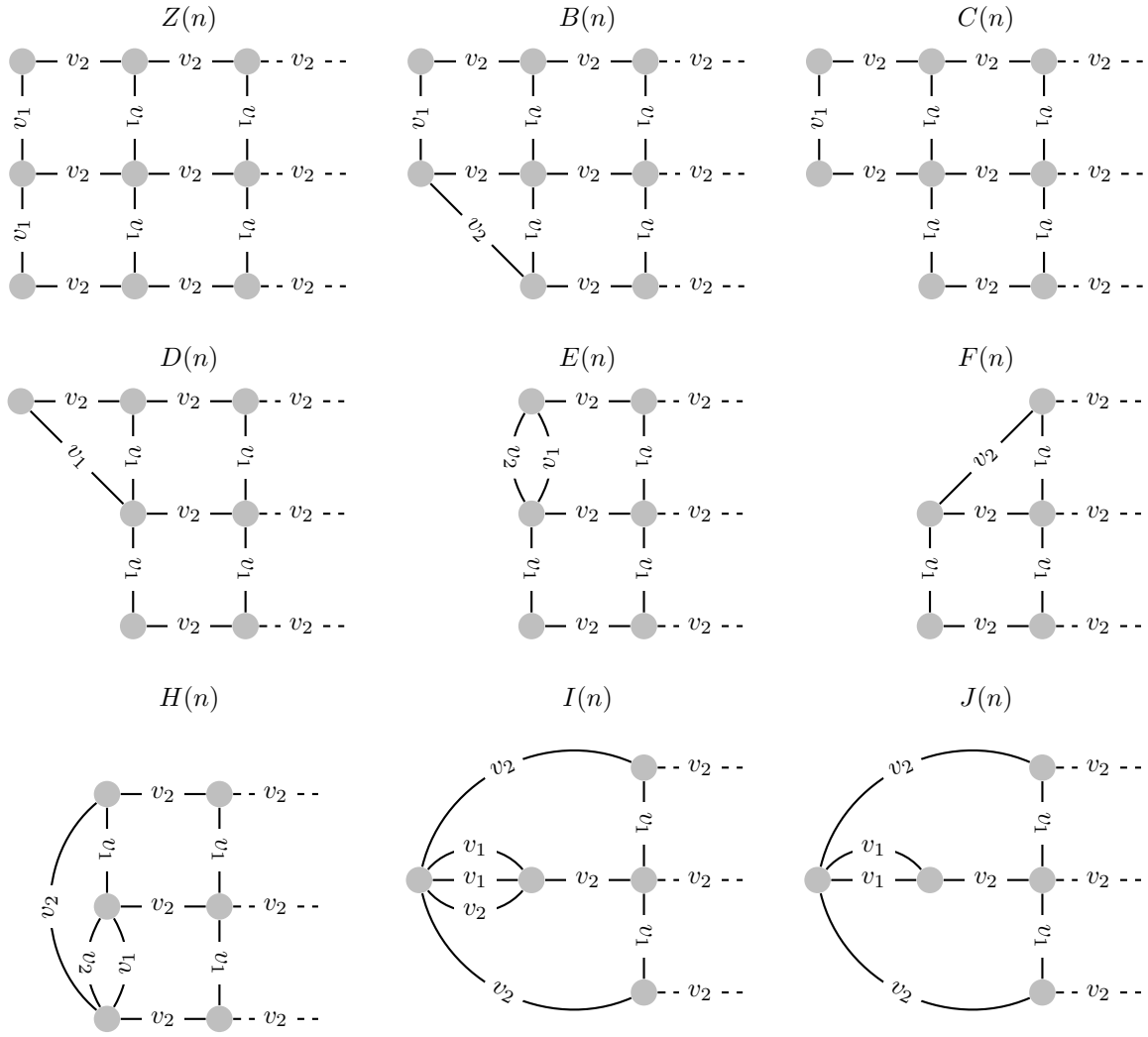


Figure 5.7: Phase angle of the partition function for the $2N$ ladder with $J_1 = -2, J_2 = -1$ where the temperature is continued onto the complex plane. The left plot has $\beta = 1/kT$ to show the behavior of the high temperatures and the right plot shows the low temperature behavior. The points where the phase changes rapidly signify zeros of the partition function and hence phase changes. The gray areas and on the low temperature plot are large values of the partition function (not zeros) where the evaluation failed due to an overflow error. The lighter areas are large values of the partition function (not zeros).

$3N$ Ladder - Directional bonds

We give the general recurrence relations for a $3N$ ladder with bonds strengths that vary in direction. The strength for the vertical and horizontal directions respectively are again J_1 and J_2 where $v_1 = e^{J_1\beta} - 1$ and $v_2 = e^{J_2\beta} - 1$. It is to cumbersome to write the terms graphically in each equation, so we assign a symbol for each graph in the expansion indexed by the length of the ladder.



This gives a set of coupled recurrence relations

$$Z(n) = v_1 B(n) + (q + v_2) C(n) \tag{5.99}$$

$$C(n) = v_1 D(n) + (q + v_2)^2 Z(n-1) \tag{5.100}$$

$$D(n) = v_1 E(n) + (q + v_2) Z(n-1) \tag{5.101}$$

$$E(n) = v_2(1+v_1)F(n) + Z(n-1) \tag{5.102}$$

$$B(n) = v_1 G(n) + (q+v_2)D(n) \tag{5.103}$$

$$G(n) = v_2 E(n) + (q+v_2)Z(n-1) + v_2 H(n) \tag{5.104}$$

$$H(n) = v_2 I(n) + E(n) \tag{5.105}$$

$$I(n) = v_2(1+v_1)^2 G(n-1) + J(n) \tag{5.106}$$

$$J(n) = v_1(1+v_1)G(n-1) + v_1 G(n-1) \tag{5.107}$$
$$+ (v_2+q)(Z(n-2)(1+(v_2+q))J(n-1))$$

$$F(n) = B(n-1) \tag{5.108}$$

These equations could, in theory, be solved to get a recurrence relation on in terms of $\{Z(n), Z(n-1), Z(n-2), \ldots\}$. We conjecture that the process could continue indefinitely for larger ladders, though we have not explored the $4N$ ladders or higher.

5.8 Remarks

In this chapter we developed new methods for solving the Ising/Potts model over arbitrary graphs and sought closed form solutions to ladder graphs. The ladder graphs have a natural connection to amyloid fiber aggregation with a preferred growth direction since one can take into account different bond strengths along each direction. These models permit only a few free parameters J_1, J_2, q and for the graphs solved with external fields h. The results presented here open avenues of future work to connect the analytical models to concrete experimental data of real amyloid fiber growth by fitting these parameters.

Chapter 6

Clustering and Kinetics

Macrostates, collections of microstates that share similar characteristics, are a useful tool in the study of any system, for instance, in setting up coarse-graining models. For proteins in particular, with a small enough set of macrostates one can usually deduce the folding pathway with even a rudimentary kinetic model. This poses a problem for a system with many degrees of freedom; how does one go about finding the relevant macrostates in the first place? In this chapter we develop a plan for partitioning the microstates into broad, meaningful macrostates from time-series data or detailed microstate information. We begin with an outline of the methodology and elaborate on the details through two models: a kinetic trajectory taken from a frustrated Langevin system and a small, idealized β-hairpin embedded on a two-dimensional lattice. We will see that the macrostate clustering method developed here provides a robust tool in the analysis of dynamical systems.

6.1 Macrostate Clustering

In theory, the equations of motion along with initial conditions impart complete knowledge of any system. In reality, the situation is more complex. Few systems can be solved analytically, and these solutions are often idealized models of the underlying mechanics. Topological characterizations can usually be made (such as Lyapunov exponents), but these convey a solution only in the broadest sense. What is needed is geometrical measure, a way of partitioning the phase space to a more manageable subset. If our system is near or at thermal equilibrium this is tantamount to clustering the microstates into physically meaningful macrostates. The idea is not new, many of the early pioneers of biophysical models have been advocating different methodologies for macrostate classification.[137,138] It would be useful to have a general purpose algorithm for constructing the macrostates from information obtainable from the equations of motion, or a statistical description of the system.

In this chapter we propose a viable method for macrostate clustering (MSC for short). The MSC algorithm takes as input a kinetic trajectory and a single parameter. We convert the kinetic trajectory into a Markov matrix and use its eigenvectors to make this classification. Since an embeddable Markov matrix is a mapping from a system of linear first-order differential equations, the MSC algorithm can be considered a first-order approximation to the motion.

With a system specified by its Hamiltonian $\mathcal{H}$ and intensive properties, the first step is a suitable identification of the microstates $\{\xi_1, \xi_2, \ldots, \xi_N\}$. For continuous systems, the accessible state space must be partitioned.[1] A trajectory is obtained by integrating the equations of motion, making note of the conformation at equal time intervals. An incidence matrix is formed by counting the number of times microstate ξ_i moved to ξ_j. This matrix can be converted to a proper Markov matrix by normalizing the row sums. Alternatively, if one had complete knowledge of the microstates and the transition rates between them, one could convert this matrix into a Markov matrix without the loss of any information, that is, the Markov matrix is a snapshot of a continuous time Markov process. With Markov matrix $\mathbf{M}$, compute the Schur eigendecomposition

$$\mathbf{M} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}^{-1} \tag{6.1}$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$. Since $\mathbf{M}$ is a Markov matrix, $\lambda_1 = 1$ corresponds to the slowest mode of relaxation of the system, i.e. the steady-state. The other eigenvalues, being smaller, represent faster modes of relaxation (see Section 2.4 for more details on Markov matrices). The influence of the modes on the relaxation of the system is exponential $\propto e^{\lambda}$, thus the slower modes not only represent the dominant flows of the system, they are less influenced by the initial conditions. This makes them prime candidates for macrostate discovery. The fastest mode included in this consideration is the single free parameter in the MSC algorithm, ϕ. Let this parameter $0 \leq \phi \leq 1$ be the eigenvalue threshold cutoff. Take a cut of the eigenvector matrix $\mathbf{V}$, retaining only those vectors where $\lambda_i > \phi$. At $\phi = 1$ only steady-state is retained, while at $\phi = 0$ all

[1] In practice this is easily adjustable even after a trajectory has been obtained and does not change the results of the MSC algorithm greatly. With a finer partitioning method, more resolution is obtained at the cost of under-sampling. This consequently increases the running time.

modes have equal influence. If the number of eigenvalues remaining is k then

$$\mathbf{U} = \text{cut}(\mathbf{V}) \tag{6.2}$$

is a N by k matrix. While the columns are the eigenvectors, each row $u_i \in \mathbf{U}$ represents microstate ξ_i contribution to the remaining modes. We call the vectors u the eigenflows of the system. We can imagine each of the N vectors as points in the space of $\mathbb{R}^k$. This gives the central idea behind the MSC algorithm. Microstates should be grouped into macrostates by clustering their eigenflows. Eigenvectors are unique up to a sign, so the clustering must take this into account. Formally, we are looking to define a metric $\mathbf{g}(u_i, u_j)$, such that two eigenflows belong to the same cluster if, for some small ϵ

$$\mathbf{g}(u_i, u_j) = \|u_i - u_j\| < \epsilon \tag{6.3}$$

The idea has physical appeal, points near a local minima in phase space should all have similar eigenflows. Transition points in the system should be different, they should split the flow across the boundaries giving distinct components in their eigenflows compared to the local minima. In practice we found the main stumbling block was twofold: the choice of a proper metric (Equation 6.3) and a suitable algorithm to group the elements into clusters implied by $\mathbf{g}_{ij}$. Initially we choose the Euclidean metric $\|\cdot\|_2$ and used various clustering algorithms to group the states. The method was only partially successful. The greatest shortcoming was the introduction of several new ad-hoc parameters into the system from the clustering algorithms. We eventually settled on a simpler classification that relied on no new parametrization. Let

$$\mathbf{sgn}\,(u) = [\mathbf{sgn}\,(u_1), \mathbf{sgn}\,(u_2), \ldots, \mathbf{sgn}\,(u_k)] \tag{6.4}$$

be the sign signature of a vector, where we define

$$\mathbf{sgn}\,(x) = \begin{cases} -1 & : \quad x < 0 \\ \\ 1 & : \quad x \geq 0 \end{cases} \tag{6.5}$$

Two microstates belong to the same cluster (and hence macrostate) if and only if their sign signatures are equal. We assign each of the microstates u_i to clusters in $\mathcal{C}' = \{c_1, c_2, \ldots, c_{2^k}\}$. Not every cluster will be populated, since it is possible for $2^k > N$. Let $\mathcal{C}$ be the set of non-empty clusters and $|\mathcal{C}|$, $|c_i|$ count the number of clusters and the number of microstates within cluster c_i respectively.

6.1.1 Quasi-steady-state

Once the clusters have been identified, we can present a macroscopic version of the original Markov matrix. Since macrostate behavior is often dominated by rare transitions between macrostates, we will assume that each macrostate relaxes to a local steady-state. We call this the quasi-steady-state. For each cluster we first compute the quasi-steady-state transition probability.

Let $\mathbf{M}_{c_i}$ be the square matrix with $|c_i|^2$ elements that are pulled from cluster c_i. This cluster is approximated as an isolated system, as we've only retained the flows within this cluster. We row normalize the matrix, discarding any rows (and their associated column) where the row sum is zero. Such states are possible within a transient cluster since it is conceivable that the trajectory from a fixed microstate does not flow to any other microstate in the cluster. The quasi-steady-state vector $\pi(c_i)$ is given by the left-eigenvector with eigenvalue one of the reduced renormalized Markov matrix

$$\pi(c_i)\mathbf{M}_{c_i} = \pi(c_i) \tag{6.6}$$

Once we have computed the quasi-steady-states we use them as initial conditions to approximate macrostate behavior. Over a single time step these initial conditions will flow into other clusters giving us macrostate kinetics. Let b_i be a vector of initial conditions from the full system. This vector has N elements that correspond to those in $\pi(c_i)$, i.e. it is the initial condition of quasi-steady-state for cluster c_i and zero for all the elements of the other clusters. We start the system in b_i and advance it one time step $b_i\mathbf{M}$ to get the Markov macrostate matrix $\mathbf{S}$.

$$\mathbf{S}_{ab} = \sum_{i \in c_b}^{|c_b|} (b_a \mathbf{M})_i \tag{6.7}$$

The matrix $\mathbf{S}$ gives us new macrostate information. In addition to simply identifying the

macrostates, the matrix identifies kinetic pathways by giving us the transition rate to each cluster. By starting at some set of initial conditions, we can trace a pathway along the clusters by continuously applying **M** to the set. From here we can discard irrelevant clusters to a particular problem, and identify intermediate macrostates along the way. The identification of intermediates would be difficult with the original matrix; no single intermediate microstate dominates the statistics. When the states are clustered together, their similar eigenflows tend to move in tandem. These newly identified macrostates could present a new reaction coordinate or kinetic pathway. In the two subsections below, we apply the formalism to two illustrative cases.

6.2 Frustrated Langevin Walk

In this example we work directly with time-series data from a kinetic trajectory. The trajectory is obtained by integrating the equations of motion of a random walk in an external field. We will see that only the time-series data are needed, knowledge of the Hamiltonian or a suitable reaction coordinate is not required to determine the macrostate classification. This example is particularly useful in view of the popularity of molecular dynamics (MD) simulations in the biophysical community.[139] In MD simulations, one often has many degrees of freedom, hence developing a general model is often a non-trivial task. The ability to partition the conformational space into physically relevant macrostates of lower dimension would greatly facilitate a deeper understanding of the system.

We model a particle in a viscous fluid subject to stochastic noise in a Langevin-like equation.[140] There are kinetic traps - minima in the potential that localize the motion until the particle can escape by thermal diffusion. Because of this, we call the system a frustrated Langevin walk (FLW).

The traps are modeled as Gaussian wells. To simplify the expression of the equations of motion we define an isotropic Gaussian well that is centered at (x_R, y_R) by

$$L(x_R, y_R) = -\exp\left(-\sqrt{(x - x_R)^2 + (y - y_R)^2}\right) \tag{6.8}$$

The potential, shown in Figure 6.1, is the contribution of three wells

$$V(\mathbf{x}) = L(0,0) + L(2,2) + L(0,-2) \tag{6.9}$$

with $\mathbf{x} = (x, y)$. Since the potential at a given point is the sum of the Gaussian wells, the depth at each of the three minima have the relation $V(0,0) < V(0,-2) < V(2,2)$.

In addition to the potential, the particle is subject to stochastic noise and viscous damping

$$m\ddot{\mathbf{x}} = -\nabla V(\mathbf{x}) - \lambda\dot{\mathbf{x}} + \eta(t) \tag{6.10}$$

The dot notation indicates a derivative with respect to time. The damping coefficient λ, is a measure of the frictional forces acting on the system proportional to the velocity. The stochastic term η, is an approximation to the microscopic collisions the particle would experience in a real fluid. The force has a correlation time of

$$\langle \eta_i(t)\eta_j(t')\rangle = \frac{2\lambda}{\beta}\delta_{i,j}\delta(t-t') \tag{6.11}$$

We choose the parameters $\beta = m = 1$, $\lambda = 0.07$ as they were sufficiently weak to bound the motion, but strong enough to occasionally facilitate a transition over the potential barriers. We integrated the system for a total time of $t_{\max} = 10^6$ using a time step of $\Delta = 0.2$ with a standard Runge-Kutta 4th order integrator.[2] The trajectory $\mathcal{T}$, is an ordered list

$$\mathcal{T} = \{\mathbf{x}_{t_0}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_{\max}}\} \tag{6.13}$$

[2] Integrating a stochastic differential equation properly is a subtle, difficult task. Since most integration schemes rely on a Taylor expansion, it becomes problematic when there are force terms that are nowhere differentiable. In general, one must expand with the time-ordered exponential T

$$T\left[\exp i\int_0^{\delta t}\mathcal{L}(s)ds\right] \tag{6.12}$$

where $\mathcal{L}$ is the Liouvillian.[141] Since our focus is on the method of macrostate approximation, we estimate the stochastic force by adding a random perturbation to the force each step of the integration. A deeper study of the system in question would require a proper handling of this issue by considering a suitable time-ordered integration scheme.[142–144]

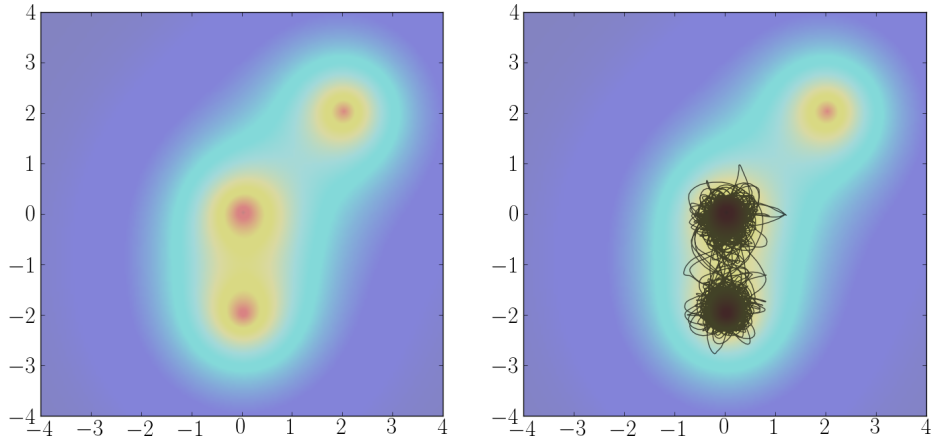


Figure 6.1: (Left) Contour map of the three well Gaussian potential, Equation 6.9. (Right) A sample ($0.2t_{\max}$) of the FLW trajectory shown by integrating Equation 6.10. The motion is localized in the wells, escaping only when random fluctuations push it over a potential barrier.

A sample trajectory is plotted in Figure 6.1. The accessible coordinate space $\{\xi_1, \xi_2, \ldots, \xi_N\}$, is given by discretizing the interval $\mathbf{x} \in [-4, 4]$ into $N = 45^2$ squares . For a given position, let $\Xi(\mathbf{x})$ return the discretized position in coordinate space. For example, if $\mathbf{x}_i \in \xi_j$ then $\Xi(\mathbf{x_i}) = \xi_j$. We choose a lag time of 5Δ to simulate the collection of time-series data collected at larger intervals than the integrating time-step. Next we computed an incidence matrix $\mathbf{C}_{ij}$. This matrix counts the number of times over the lag time the trajectory went from ξ_i to ξ_j

$$\mathbf{C}_{ij} = \sum_{i=5}^{t_{\max}/\Delta} \delta(\Xi(\mathbf{x}_{t_{k-5}}), \xi_i)\ \delta(\Xi(\mathbf{x}_{t_k}), \xi_j) \tag{6.14}$$

A Markov matrix $\mathbf{M}$ is constructed from the incidence matrix by normalizing the row sums

$$\mathbf{M}_{ij} = \frac{\mathbf{C}_{ij}}{\sum_k \mathbf{C}_{kj}} \tag{6.15}$$

It is worth nothing that the resulting Markovian matrix is only a linear approximation to the actual dynamics. Especially in the regions between pairs of the wells the may not hold. However, it is a useful first-guess a the transition rates that can later be refined. The Markov matrix was diagonalized and clusters were created with an eigenvalue cutoff of $\phi = 0.85$. This partitioned the conformational space into four clusters, shown in Figure 6.2.

With the clusters identified, we permuted the Markov matrix to reflect our knowledge of the clusters. In Figure 6.3 we see that the original matrix has no particular structure, while the permuted matrix shows a nearly block diagonal structure. The non-zero elements off the block diagonal indicate transitions between the clusters, while those within the block indicate movement within the cluster. The macrostate approximation identified all three local minima and the largest crossing barrier. Our choice of ϕ was not arbitrary, it was selected to be large enough to grab the semi-stable states, yet small enough to have broad macrostates. When ϕ is lowered, the state space partitioned into finer clusters. This may or may not be helpful in determining a kinetic pathway, the macrostates need to be connected to a physically relevant set of macro coordinates. Therefore, while a lower ϕ gives more information, not all of it is useful. For example, when a slightly lower ϕ is used there are two new macrostates: the first is a barrier crossing between the bottom wells (labeled c_1 and c_2 in Figure 6.2), and the second is a ring shaped boundary around the upper well c_0. The first is well-understood, it clearly represents the boundary between two more dominant macrostates; this barrier is useful for characterizing the system. The second is also easy to explain; there is simply a transient state that moves away from the top cluster only to quickly return. However, knowledge of this transient state does little to further our understanding of the system. Thankfully the cluster calculation is quick compared to the generation of time-series data for even a modest system.

To reduce the degrees of freedom of the system, we suggest a new kinetic pathway. With the above choice of ϕ, all clusters have physical significance. Clusters c_0, c_1, and c_2 are the local minima while cluster c_3 is a barrier. Our model should reflect the kinetics of these four macrostates. We first computed the macrostate kinetic matrix $\mathbf{S}$ from Equation 6.7 by solving for the quasi-steady-state

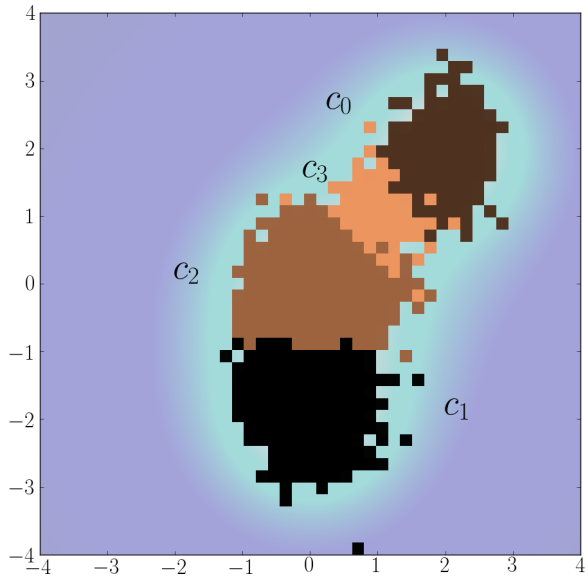


Figure 6.2: Clusters found by partitioning the eigenflows of the FLW trajectory (Equation 6.10). Here $c_i(n)$ denotes the i^{th} cluster with n conformational states in it. The original potential is overlaid. The algorithm found all three potential minima and identified a section as a crossing barrier.

flows. This matrix was converted into a rate transition matrix by the relation[3] $t\mathbf{W} = \log \mathbf{S}$ giving

$$t\mathbf{W}_{\text{reduced}} = 10^{-4} \begin{bmatrix} -29 & & & 29 \\ & -61 & 61 & \\ & 37 & -41 & 4 \\ 3181 & & 1198 & -4377 \end{bmatrix} \tag{6.16}$$

For clarity, blank entries are zero. Taking the entries of this matrix, we can now propose a simple discretized linear model for the movement in the conformational space.

[3] While one can compute the logarithm of a matrix, there are issues such as embeddability and uniqueness that do not occur in the matrix exponential. Matrix $\mathbf{A}$ is said to be embeddable if there exists a matrix $\mathbf{B}$ such that $\mathbf{A} = \exp(\mathbf{B})$ and $\exp(t\mathbf{B})$ is Markov for all $t \geq 0$. This is equivalent to taking a snapshot of an autonomous finite state Markov process that evolves continuously in time.[145] In addition, the matrix logarithm is not necessary unique. If the matrix is diagonalizable and the spectrum of eigenvalues are all strictly greater than zero, than the logarithm is unique and is given by $\log \mathbf{A} = \mathbf{V} \log(\mathbf{\Gamma}) \mathbf{V}^{-1}$. This is the Schur decomposition where the logarithm is applied to the diagonal matrix of eigenvalues. For the FLW walk, the matrix $\mathbf{S}$ satisfied both constraints, hence the rate matrix obtained was unique.

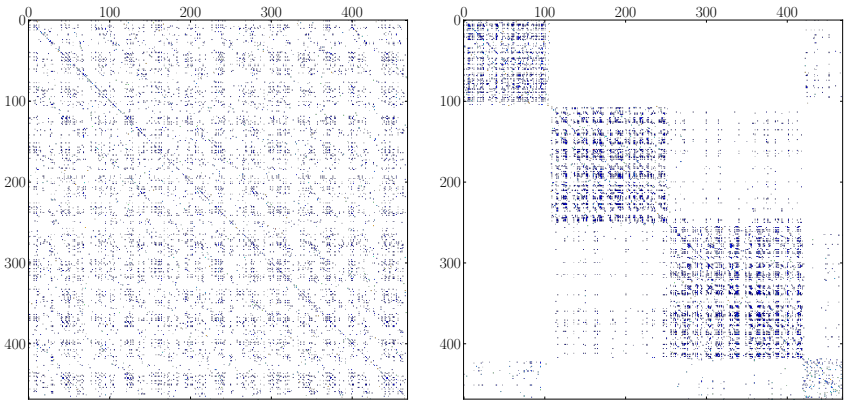
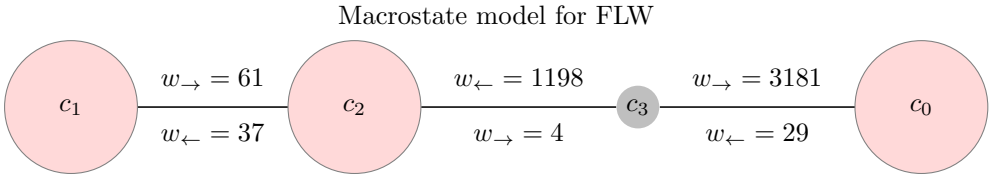


Figure 6.3: Markov transition matrix for the FLW system. The left matrix shows the states in the original, unsorted form. On the right the states have been permuted such that the states that belong to the same cluster are in adjacent rows. The block diagonal structure is clearly evident.



The macrostate characterization is complete. A new reaction coordinate for the system is suggested by this model, one that moves along the average coordinates for each macrostate. Since the macrostates are only approximations deduced from trajectories, their primary use is a categorical tool. One can also estimate first-crossing times and other kinetic predictions using various classical reaction-rate theories.[146]

6.3 Folding Pathway of a Beta-Hairpin

For our second example illustrating an application of the MSC algorithm, we choose a system where the evaluation of the macrostates is not as straightforward as the FLW. Here the microstates are the various conformations of a simple lattice peptide. For a peptide, coarse-graining of the conformational space is usually achieved by various reaction coordinates such as fraction of native contacts or radius of gyration. These are selected with an *a posteriori* knowledge of the physical and relevant coordinates. Here we profess ignorance of the system to see if the MSC algorithm can help determine a proper reaction coordinate.

The general properties and behavior of lattice peptides have been detailed in the previous chapters (c.f. Section 4.2). The peptide in our current study is constructed such that it displays both an interesting folding pathway and is a reasonable approximation to the known folding properties of a ten residue β-hairpin. Each of the ten residues has a property $r_i \in \{\mathrm{H}, \mathrm{P}, \mathrm{C}_+, \mathrm{C}_-, \mathrm{N}\}$ in a HPC bead-model Hamiltonian (hydrophobic, hydrophilic, charged residues) where N is a neutral non-interacting residue. Two residues in contact on the lattice contribute to the Hamiltonian

$$\mathcal{H} = \sum_i \sum_{j>i+3} \delta(v_i, v_j) \left(E_{\mathrm{HP}}(i,j) + E_{\mathrm{HH}}(i,j) + E_{\mathrm{PP}}(i,j) + E_{\mathrm{C_+C_-}}(i,j) \right) \tag{6.17}$$

Here $\delta(v_i, v_j)$ is 1 if and only if residues v_i and v_j are nearest neighbors on the lattice. The energies E_{AB} are defined by an interaction strength ϵ_{AB} and are non-zero if and only if residues v_i and v_j have the property AB (the ordering is unimportant). The magnitudes of the various interactions are

$$\begin{aligned} \epsilon_{\mathrm{HP}} &= 0.1 \\ \epsilon_{\mathrm{PP}} &= -0.5 \\ \epsilon_{\mathrm{HH}} &= -0.6 \\ \epsilon_{\mathrm{C_+C_-}} &= -1.5 \end{aligned} \tag{6.18}$$

A picture of the unique native state is shown in Figure 6.4, with energy -3.2. This lattice Hamiltonian is one of the simplest β-hairpin models possible that captures the biological essence, namely:

- Favorable turn formation: One of the defining observed characteristics of the hairpin is the favorable internal energy the residues make at the turn. This is primarily due to the formation of a salt bridge.
- Realistic relative energetic magnitudes: The energies are chosen such that $-\epsilon_{\mathrm{HP}} < \epsilon_{\mathrm{PP}} < \epsilon_{\mathrm{HH}} < \epsilon_{\mathrm{C_+C_-}}$.
- Hydrophobic core: By including some unfavorable HP interactions we define a clear hydropho-

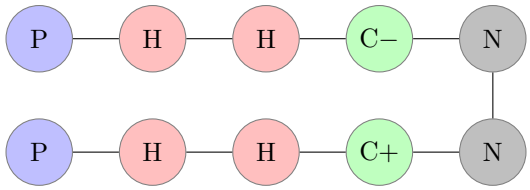


Figure 6.4: Native state of the ten residue two-dimensional lattice β-hairpin. Residues marked with H, P, C+, C−, N are hydrophobic, hydrophilic, charged, uncharged and neutral respectively.

bic core that stabilizes at the center of the protein. In addition, this creates an enthalpic barrier to reach the unique native state if the protein misfolds.

Previously we've seen that sampling the state space can be done efficiently with Monte-Carlo techniques. In this case however, the state space is small enough to enumerate completely. Accounting for rotational and mirror-image symmetries there are only $N = 714$ possible conformations of the system. The conformations are linked to each other by a move set of rigid rotations; every point on the chain was a permissible pivot to a new conformation as long as the resulting chain did not self-intersect. Since the chain length is less than 12, it is known that rigid rotations for a two-dimensional lattice peptide provide sufficient ergodicity.[147] Numbering the conformations $\Xi = \{\xi_1, \xi_2, ..., \xi_N\}$ we define an adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij} = 1$ if and only if conformation ξ_j can be reached from conformation ξ_i by exactly one rigid rotation.

We compute the transition rate from two states using both Metropolis and Glauber dynamics.[45] In both methods one needs to compute the energy difference $\Delta E_{ij} \equiv \mathcal{H}(\xi_j) - \mathcal{H}(\xi_i)$. For $i \neq j$ the transition rate matrix is specified by

- Metropolis dynamics

$$\mathbf{W}_{ij} = \min(1, e^{-\beta \Delta E_{ij}}) \tag{6.19}$$

- Glauber dynamics

$$\mathbf{W}_{ij} = \frac{e^{-\beta \Delta E_{ij}}}{1 + e^{-\beta \Delta E_{ij}}} \tag{6.20}$$

For $i = j$, $\mathbf{W}_{ii} = -\sum_{k \neq j} \mathbf{W}_{kj}$ (see Section 2.5 for additional details on the master equation

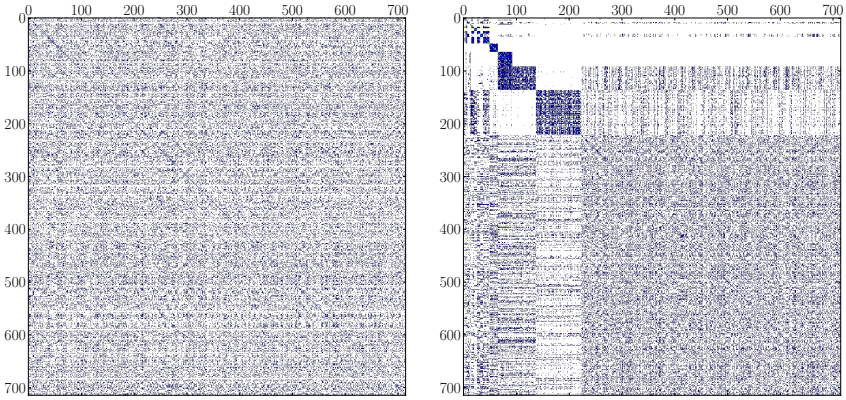


Figure 6.5: Markov transition matrix for the β-hairpin system using Glauber dynamics. The left matrix shows the states in the original, unsorted form where the states were enumerated by rigid rotations. On the right, the matrix has been permuted such that the states that belong to the same cluster are in adjacent rows. The block diagonal structure is clearly evident, with a large fraction of the states grouped into a single cluster.

formulation). We take the exponential of this matrix to get a Markovian one, $\mathbf{M} = \exp(t\mathbf{W})$ with $t = 0.5$ and $\beta = 4$. While different rates were obtained using Glauber and Metropolis dynamics, there was no discernible difference in the results of the MSC algorithm. As in the previous problem with the FLW system, we show the original and permuted Markov matrices in Figure 6.5. The block diagonal structure can be seen for small isolated clusters, but a large majority of conformations reside in the large block. Examining the states in this cluster we see that these are the random coil states.

6.3.1 Folding Nucleation

In this system there were 22 identified clusters, though not all of them were physically relevant. To see why, we start the system in an initial state of the fully extended random coil b_c. We advance the system in time via $b_c\mathbf{S}^t$ and plot the occupational probabilities of the clusters. In Figure 6.6 we see the long-time behavior of the system. In the figure we order the clusters by the number of microstates they contain to illustrate the magnitude of the largest cluster. The ordering of the clusters is unimportant, any permutation would give the same macrostate information. This cluster c_{21}, contained 493 microstates and had the initial state as one of its members. However the system quickly and cooperatively folded into two different clusters with a much smaller number of

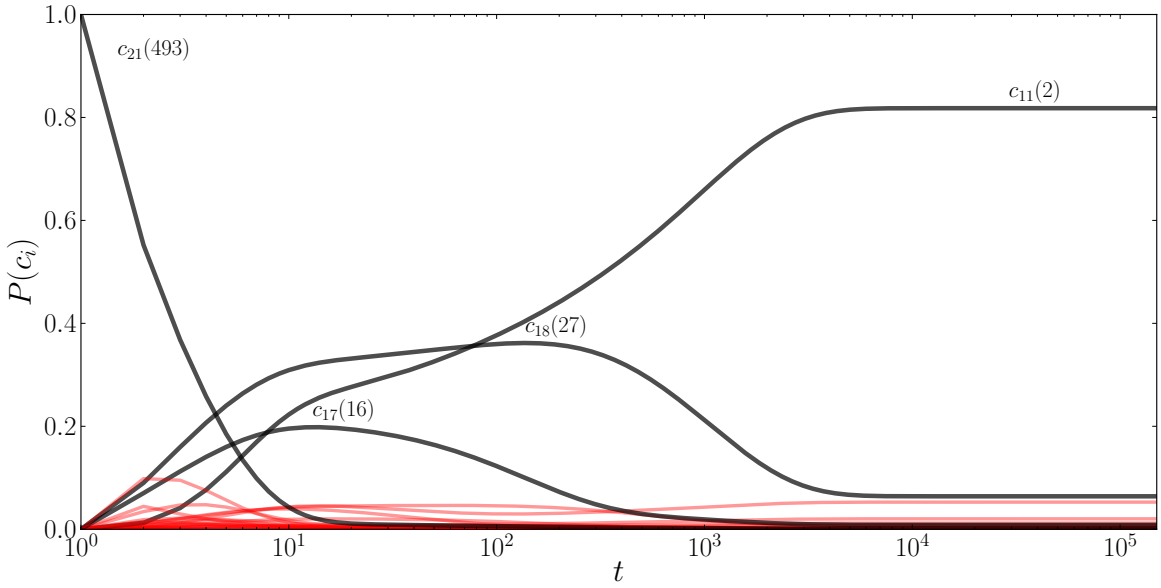
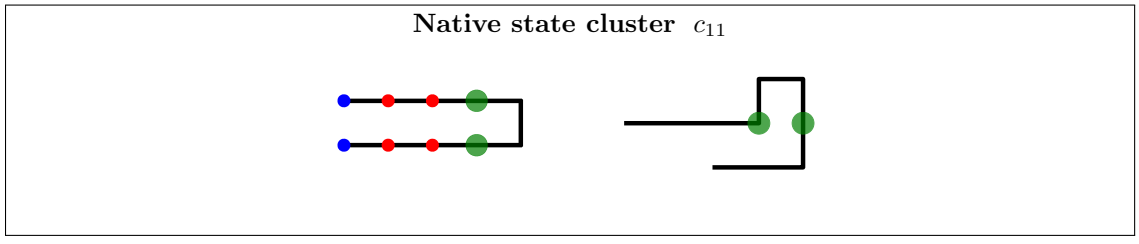


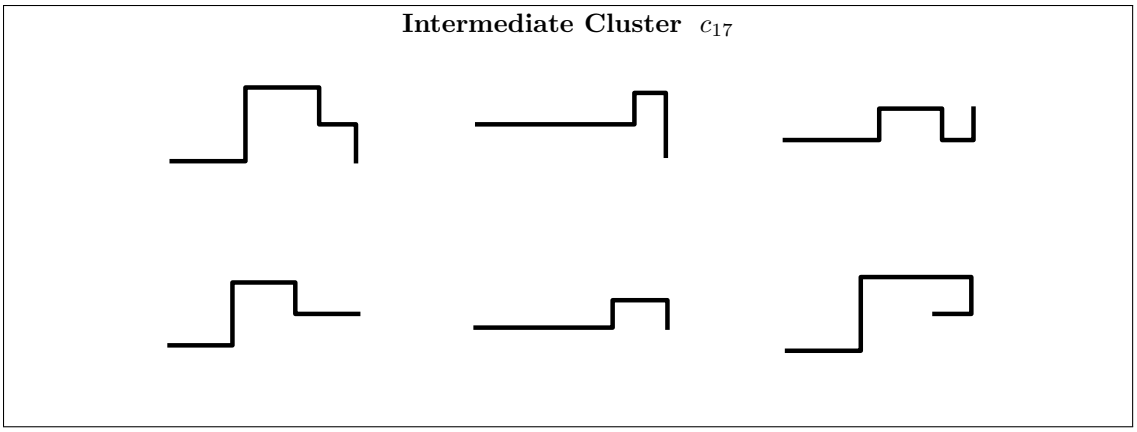
Figure 6.6: Cluster kinetics for the β-hairpin over a logarithmic time axis. Starting at the completely unfolded state, the motion spans many decades before it folds in the native state. For clarity, the clusters with a significant population are labeled and shown in black. The other clusters are shown in red.

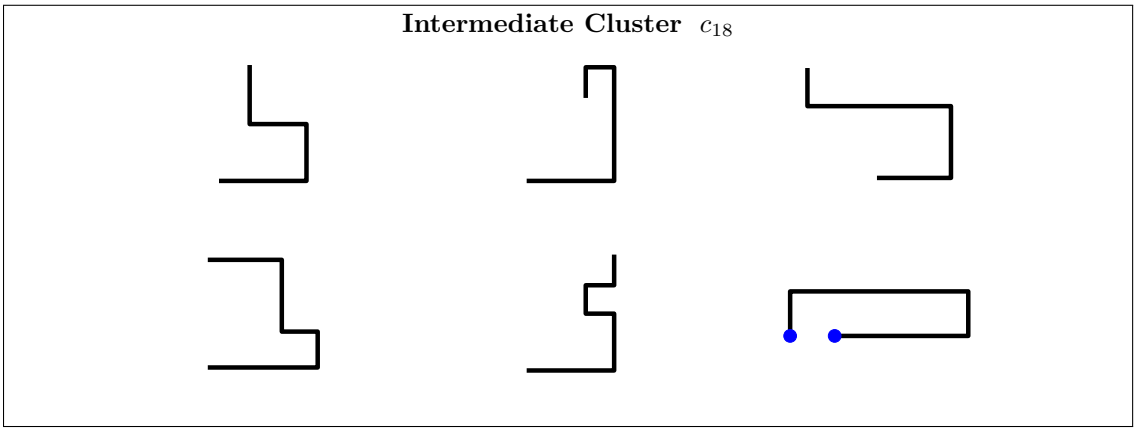
microstates. These clusters c_{17} and c_{18} had 16 and 27 microstates respectively. We identified them as intermediate states of the folding processes as they eventually decayed before reaching a final cluster c_{11} with only two microstates.

With the interesting clusters identified, we can examine them for any shared property. The microstates are shown at different scales for visual clarity (long, narrow microstates may appear smaller than globular ones). Shown below is the final cluster c_{11}, clearly containing native state.



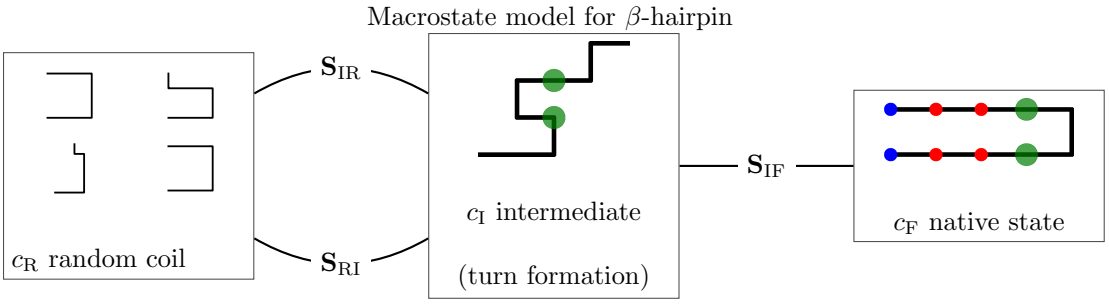
The presence of an additional microstate is not unexpected. As we saw before in the FLW walk, local minima are often associated with the same cluster. All of the microstates in the other two intermediate clusters have one particular thing in common, the formation of the salt bridge. Listed below are six representative conformations in each of the two intermediate clusters





Clearly the folding pathway is dominated by the formation of the turn. This implies that the formation of the turn in the β-hairpin is a nucleation event, a necessary first step for the folding process.

The folding kinetics can be computed using the quasi-steady-steady approximation. Grouping the intermediate states as $c_{\mathrm{I}} = \{c_{17}, c_{18}\}$, the final state $c_{\mathrm{F}} = \{c_{11}\}$ and all other clusters as random coil states c_{R} we have



The forward and backwards transitions were calculated to be

$$\mathbf{S} = \begin{bmatrix} 0.54042 & 0.39912 & 0.06046 \\ 0.00713 & 0.99266 & 0.00021 \\ 0.00005 & 0.00001 & 0.99995 \end{bmatrix} \tag{6.21}$$

with the rows ordered as $[c_{\mathrm{R}}, c_{\mathrm{I}}, c_{\mathrm{F}}]$.

We stress that these predicted transitions are a coarse-graining of the real kinetics since we have assumed the states equilibrate before transitioning into a different macrostate. Nevertheless, the approximation is validated by the observation that at the observed temperature, the slowest eigenmodes dominate the kinetics.

6.4 Remarks

When the number of microstates is large, the eigendecomposition could take an inordinate amount of computation. Thankfully, once ϕ is chosen, only the largest eigenpairs need to be found. In general, the full eigendecomposition is unnecessary. There exist techniques like the Lanczos routine that extract only the larger eigenpairs.[148] These techniques work well with the matrices used in the MSC algorithm; the Lanczos algorithm works with non-Hermitian matrices and excels when the matrices are sparse and the eigenvalues are not degenerate.

It is worth noting that the macrostates obtained did not use the equations of motion, only the time-series data. This gives the technique wide applicability, extending beyond the domain of biophysical simulations. While the systems studied here were small, the MSC method should scale well to larger more complicated sets of conformational microstates. The limit of the application appears to be the ability of the researcher to extract meaningful data out of the clusters once they have been found. For example, in a larger off-lattice model of protein folding, the full state space of each of the $3N$ residues' coordinates may be unnecessary. In this case, a relative set of microstates may be more useful, say the $2N - 2$ dihedral angles.

It would be interesting to pair the results obtained with studies such as Max Caliber, a variational principle for the dynamics of nonequilibrium statistical mechanics.[149,150] Along similar lines, insight

might be gained by comparing the clusters against the so-called 'thermodynamic length', a measure of the distance between equilibrium thermodynamics states. This is a Riemannian metric that defines the distance between two states as the number of natural fluctuations along that path.[151] Here the fluctuations are defined through the metric $\mathbf{g}_{ij}$

$$\mathbf{g}_{ij} = \frac{\partial \langle X_i \rangle}{\partial \lambda^j} = \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle \tag{6.22}$$

where the λ's are the time dependent and configuration independent conjugate variables and the X's are the time independent functions of the configurations (e.g. in an isothermal-isobaric ensemble X would contain the internal energy and volume and λ would contain the temperature and pressure). The length of a curve parametrized by t, from 0 to τ is

$$\mathcal{L} = \int_0^\tau \sqrt{\frac{d\lambda^i}{dt} \mathbf{g}_{ij} \frac{d\lambda^j}{dt}} dt \tag{6.23}$$

Since the integration is path-dependent, minimized paths provide a variational approach to important relaxation properties of the system. It is possible that the clusters defined here have a natural connection to this length, perhaps as a first order approximation.

Chapter 7

Final Remarks

In this thesis, we have developed several new methods and models, each an attempt to provide insight into the role of entropy in the protein folding process. These studies and others like it, are paramount to a full understanding of protein models. The breadth of the research chapters, depletion forces, crowding, aggregation, and macrostate clustering, emphasizes the fact that protein folding process is still very much an open problem.

Each of the four research chapters highlight particular suggestions for model improvements, but there are a few overarching directions that seem to be common to all of them. First and foremost, we can't underestimate the need for more comparisons with experimental data. While Chapters 3 and 4 were designed around previously measured data, the model theories have only been extrapolated to different systems, but never tested. The models in Chapters 5 and 6 have yet to be applied and tested by experimental data outside of a few small systems. This presents an opportunity for both new experimental measurements and extensions of the models to existing data.

In addition to the experimental connection, many of the models present a simplified treatment of the enthalpic terms. This was motivated by the idea to study the effects of entropy exclusively. Many of the calculations however, would have been much more difficult if a more realistic potential had been used. While we feel that a simplified Hamiltonian can capture the essence of the entropic forces, it does present an incomplete picture of physical reality. While the conclusions may remain the same, the supporting evidence could only be strengthened by a more realistic potential. The smaller more focused studies presented here serve an important role as building blocks for a more complete description of the biophysical process.

It is quite possible to combine molecular dynamics simulations with some of the ideas presented here. A first step would be the replacement of the Gō-like or HPC model with the empirically based

residue to residue contact matrix defined by Miyazawa and Jernigan.[152] In addition the lattice restriction could be removed, or at very least, extended to more complicated geometries. Independently, the macrostate clustering algorithm can serve as a starting point for molecular dynamic simulations, guiding the research to the salient points in conformational space. The time-series method in Chapter 6 can be applied to the multitude of molecular dynamics trajectories from other numerical experiments.

Each day, new observations are being made that seem to highlight the importance of an *in vivo* model; crowding and aggregation processes are often incomplete without them. Our study here was an attempt to simplify the process at various levels of complexity, so as to give a deeper understanding of the protein folding problem.

Bibliography

[1] Biophysical society annual meeting. San Francisco, CA, 2010.

[2] Yao Xu, Michelle R Bunagan, Jia Tang, and Feng Gai. Probing the kinetic cooperativity of beta-sheet folding perpendicular to the strand direction. *Biochem.*, 47(7):2064–2070, 2008.

[3] Deguo Du, Matthew J Tucker, and Feng Gai. Understanding the mechanism of beta-hairpin folding via phi-value analysis. *Biochem.*, 45(8):2668–2678, 2006.

[4] Daniel A Reed, Ruzena Bajcsy, Manuel A Fernandez, Jose-Marie Griffiths, Randall D Mott, Jack Dongarra, Chris R Johnson, Alan S Inouye, William Miner, Martha K Matzke, and Terry L Ponick. Computational science: Ensuring america's competitiveness. Technical report, 2005.

[5] Arthur M. Lesk. *Introduction to Protein Architecture: The Structural Biology of Proteins.* Oxford University Press, USA, 1 edition, 2001.

[6] Matsudaira Lodish Berk. *Molecular Cell Biology Fifth Edition.* 2004.

[7] Pablo Echenique. Introduction to protein folding for physicists. *Contemp. Phys.*, 48(2):81, 2007.

[8] D.L Miller and A. Ghosh. *Computational Systems Biology.* PhD thesis, Drexel, 2008.

[9] D.J. Miller and A. Ghosh. A systems approach for the prediction of wild type MAPK pathway response to targeted drugs. In *Eighth International Conference on Systems Biology*, 2007.

[10] M.N. Zakharov. *A microrheological study of sickle hemoglobin polymerization.* PhD thesis, Drexel, 2009.

[11] Vladimir N Uversky. Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front. Biosci.*, 14:5188–5238, 2009.

[12] GN Ramachandran, C Ramakrishnan, and V Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Bio.*, 7:95–99, 1963.

[13] C B Anfinsen, E Haber, M Sela, and F H White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci.*, 47:1309–1314, 1961.

[14] Sanzo Miyazawa and Robert L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[15] K A Dill. Theory for the folding and stability of globular proteins. *Biochem.*, 24(6):1501–1509, 1985.

[16] CN Pace, BA Shirley, M McNutt, and K Gajiwala. Forces contributing to the conformational stability of proteins. *Fed. Am. Soc. Exp. Bio.*, 10(1):75 –83, 1996.

[17] G D Rose and R Wolfenden. Hydrogen bonding, hydrophobicity, packing, and protein folding. *Anal. Rev. Biophys. and Biomol. Struct.*, 22:381–415, 1993.

[18] Levinthal. Are there pathways for protein folding? *J. Chem. Phys.*, 65(1):44–45, 1968.

[19] Jeetain Mittal and Robert B. Best. Thermodynamics and kinetics of protein folding under confinement. *Proc. Natl. Acad. Sci.*, 105(51):20233–20238, 2008.

[20] Wei Wang, Wei-Xin Xu, Yaakov Levy, E. Trizac, and P. G. Wolynes. Confinement effects on the kinetics and thermodynamics of protein dimerization. *Proc. Natl. Acad. Sci.*, 106(14): 5517–5522, 2009.

[21] Huan-Xiang Zhou and Ken A. Dill. Stabilization of proteins in confined spaces. *Biochem.*, 40 (38):11289–11293, 2001.

[22] Travis Hoppe and Jian-Min Yuan. Protein folding with implicit crowders: A study of conformational states using the Wang-Landau method. *J. Phys. Chem. B*, 115(9):2006–2013, 2011.

[23] Christopher J.T Lewis. *Heat and Thermodynamics: A Historical Perspective*. Greenwood, 2007.

[24] R K Pathria and Paul D. Beale. *Statistical Mechanics, Second Edition*. Butterworth-Heinemann, 2 edition, 1996.

[25] Charles Kittel and Herbert Kroemer. *Thermal Physics*. W. H. Freeman, second edition edition, 1980.

[26] Ken A. Dill and Sarina Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry & Biology*. Garland Science, 1 edition, 2002.

[27] C. E Shannon. A mathematical theory of communication. *ACM Mobile. Comp. Rev.*, 5:3–55, 2001.

[28] John David Jackson. *Classical Electrodynamics Third Edition*. Wiley, 3 edition, 1998.

[29] J. Liouville. *J. Math. Pure Appl.*, 3:342, 1838.

[30] W. Schottky. Zur statistischen fundamentierung der chemischen thermodynamik. *Annalen der Physik*, 373(14):481–544, 1922.

[31] F. Y. Wu. The potts model. *Rev. Mod. Phys.*, 54(1):235, 1982.

[32] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087, 1953.

[33] S Kirkpatrick, C D Gelatt, and M P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[34] Hideo Yahata and Masuo Suzuki. Critical slowing down in the kinetic ising model. *J. Phys. Soc. Jpn.*, 27:1421–1438, 1969.

[35] Fugao 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(5):056101, 2001.

[36] Bryan A. Patel, Pablo G. Debenedetti, Frank H. Stillinger, and Peter J. Rossky. The effect of sequence on the conformational stability of a model heteropolymer in explicit water. *J. Chem. Phys.*, 128(17):175102, 2008.

[37] Thomas Wust and David P. Landau. Versatile approach to access the low temperature thermodynamics of lattice polymers and proteins. *Phys. Rev. Lett.*, 102(17):178101–4, 2009.

[38] R E Belardinelli and V D Pereyra. Wang-Landau algorithm: a theoretical analysis of the saturation of the error. *J. Chem. Phys.*, 127(18):184105, 2007.

[39] Yong Wu, Mathias Körner, Louis Colonna-Romano, Simon Trebst, Harvey Gould, Jonathan Machta, and Matthias Troyer. Overcoming the slowing down of flat-histogram monte carlo simulations: Cluster updates and optimized broad-histogram ensembles. *Phys. Rev. E*, 72(4): 046704, 2005.

[40] B. D. Mckay. Nauty users guide. `http//cs.anu.edu.au/~bdm/nauty/`.

[41] Uwe Schoning. Graph isomorphism is in the low hierarchy. *J. Comp. Sys. Sci.*, 37:312–323, 1988.

[42] Ronald C. Read and Derek G. Corneil. The graph isomorphism disease. *J. Graph Theor.*, 1 (4):339–363, 1977.

[43] Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[44] W. Pauli, P.J.W. Debye, and A. Sommerfeld. *Probleme der modernen physik*. S. Herzel, 1928.

[45] Roy J. Glauber. Time-Dependent statistics of the ising model. *J. Math. Phys.*, 4(2):294, 1963.

[46] Allen P. Minton. How can biochemical reactions within cells differ from those in test tubes? *J. Cell Sci.*, 119(14):2863–2869, 2006.

[47] Guanghui Ping, Guoliang Yang, and Jian-Min Yuan. Depletion force from macromolecular crowding enhances mechanical stability of protein molecules. *Polymer*, 47(7):2564–2570, 2006.

[48] Jorg Rosgen, Montgomery Pettitt, and David Wayne Bolen. Protein folding, stability and solvation structure in osmolyte solutions. *Biophys. J.*, 89:2988–2997, 2005.

[49] David L. Pincus and D Thirumalai. Crowding effects on the mechanical stability and unfolding pathways of ubiquitin. *J. Phys. Chem. B*, 113(1):359–68, 2009.

[50] Margaret S. Cheung and D. Thirumalai. Effects of crowding and confinement on the structures of the transition state ensemble in proteins. *J. Chem. Phys. B*, 111(28):8250–8257, 2007.

[51] Margaret S. Cheung, Dmitri Klimov, and D. Thirumalai. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proc. Natl. Acad. Sci.*, 102(13):4753–4758, 2005.

[52] Loren Stagg, Shao-Qing Zhang, Margaret S. Cheung, and Pernilla Wittung-Stafshede. Molecular crowding enhances native structure and stability of alpha / beta protein flavodoxin. *Proc. Natl. Acad. Sci.*, 104(48):18976–18981, 2007.

[53] Allen P Minton. Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited. *Biophys. J.*, 88(2):971–985, 2005.

[54] Kenji Sasahara, Peter McPhie, and Allen P Minton. Effect of dextran on protein stability and conformation attributed to macromolecular crowding. *J. Mol. Biol.*, 326(4):1227–37, 2003.

[55] Huan-Xiang Zhou. Protein folding and binding in confined spaces and in crowded solutions. *J. Mol. Recognit.*, 17(5):368–375, 2004.

[56] Masahiro Kinoshita. Ordered aggregation of big bodies with high asphericity in small spheres: a possible mechanism of the amyloid fibril formation. *Chem. Phys. Lett.*, 387(1-3):54–60, 2004.

[57] R John Ellis and Allen P Minton. Protein aggregation in crowded environments. *Biol. Chem.*, 387(5):485–97, 2006.

[58] Huan-Xiang Zhou, Germn Rivas, and Allen P. Minton. Macromolecular crowding and confinement: Biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.*, 37, 2008.

[59] Jian-Min Yuan, Chia-Lin Chyan, Huan-Xiang Zhou, Tse-Yu Chung, Haibo Peng, Guanghui Ping, and Guoliang Yang. The effects of macromolecular crowding on the mechanical stability of protein molecules. *Protein Sci.*, 17(12):2156–2166, 2008.

[60] Ryo Akiyama, Yasuhito Karino, Yasuhiro Hagiwara, and Masahiro Kinoshita. Remarkable solvent effects on depletion interaction in crowding media: Analyses using the integral equation theories. *J. Phys. Soc. Jpn.*, 75:064804, 2006.

[61] Dirar Homouz, Michael Perham, Antonios Samiotakis, Margaret S. Cheung, and Pernilla Wittung-Stafshede. Crowded, cell-like environment induces shape changes in aspherical protein. *Proc. Natl. Acad. Sci.*, 105(33):11754–11759, 2008.

[62] Sho Asakura and Fumio Oosawa. Interaction between particles suspended in solutions of macromolecules. *J. Polym. Sci.*, 33(126):183–192, 1958.

[63] M. S. Wertheim. Exact solution of the Percus-Yevick integral equation for hard spheres. *Phys. Rev. Lett.*, 10(8):321–323, 1963.

[64] D. A. Ward and F. Lado. Structure, thermodynamics, and orientational correlations of the nematogenic hard ellipse fluid from the Percus-Yevick equation. *Mol. Phys.*, 63:623–638, 1988.

[65] J. M. J. van Leeuwen, J. Groeneveld, and J. de Boer. New method for the calculation of the pair correlation function. *Physica*, 25(7-12):792–808, 1959.

[66] Masahiro Kinoshita, Shin ya Iba, Ken Kuwamoto, and Makoto Harada. Interaction between macroparticles in Lennard-Jones fluids or in hard-sphere mixtures. *J. Chem. Phys.*, 105(16): 7177–7183, 1996.

[67] J. C. Crocker, J. A. Matteo, A. D. Dinsmore, and A. G. Yodh. Entropic attraction and repulsion in binary colloids probed with a line optical tweezer. *Phys. Rev. Lett.*, 82(21):4352, 1999.

[68] Y. N. Ohshima, H. Sakagami, K. Okumoto, A. Tokoyoda, T. Igarashi, K. B. Shintaku, S. Toride, H. Sekino, K. Kabuto, and I. Nishio. Direct measurement of infinitesimal depletion force in a colloid-polymer mixture by laser radiation pressure. *Phys. Rev. Lett.*, 78(20):3963, 1997.

[69] Riina Tehver, Amos Maritan, Joel Koplik, and Jayanth R. Banavar. Depletion forces in hard-sphere colloids. *Phys. Rev. E*, 59(2):R1339, 1999.

[70] Y. Mao, M. E. Cates, and H. N. W. Lekkerkerker. Depletion force in colloidal systems. *Physica A*, 222(1-4):10–24, 1995.

[71] Thierry Biben, Peter Bladon, and Daan Frenkel. Depletion effects in binary hard-sphere fluids. *J. Phys. Condens. Mat.*, 8(50):10799–10821, 1996.

[72] Masahiro Kinoshita. Interaction between big bodies with high asphericity immersed in small spheres. *Chem. Phys. Lett.*, 387(1-3):47–53, 2004.

[73] Masahiro Kinoshita. Spatial distribution of a depletion potential between a big solute of arbitrary geometry and a big sphere immersed in small spheres. *J. Chem. Phys.*, 116(8): 3493–3501, 2002.

[74] R. Roth, B. Gotzelmann, and S. Dietrich. Depletion forces near curved surfaces. *Phys. Rev. Lett.*, 83(2):448, 1999.

[75] R. Roth, R. Evans, and S. Dietrich. Depletion potential in hard-sphere mixtures: Theory and applications. *Phys. Rev. E*, 62(4):5360, 2000.

[76] A. D. Dinsmore, A. G. Yodh, and D. J. Pine. Entropic control of particle motion using passive surface microstructures. *Nature*, 383(6597):239–242, 1996.

[77] Masahiro Kinoshita. Roles of entropic excluded-volume effects in colloidal and biological systems: Analyses using the three-dimensional integral equation theory. *Chem. Eng. Sci.*, 61(7): 2150–2160, 2006.

[78] Yi king Choi, Wenping Wang, Yang Liu, and Myung-Soo Kim. Continuous collision detection for two moving elliptic disks. *IEEE Trans. Rob. Autom.*, 22(2):213–224, 2006.

[79] R. John Ellis. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.*, 26(10):597–604, 2001.

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

[81] L. Hannon, G. C. Lie, and E. Clementi. Molecular dynamics simulation of flow past a plate. *J. Sci. Comput.*, 1(2):145–150, 1986.

[82] T. Miyagi and T. Kamei. The standing vortex behind a disk normal to uniform flow at small reynolds number. *J. Fluid Mech.*, 134:221–230, 1983.

[83] Felix Hofling, Tobias Munk, Erwin Frey, and Thomas Franosch. Critical dynamics of ballistic and brownian particles in a heterogeneous environment. *J. Chem. Phys.*, 128(16):164517–13, 2008.

[84] Tobias Gleim, Walter Kob, and Kurt Binder. How does the relaxation of a supercooled liquid depend on its microscopic dynamics? *Phys. Rev. Lett.*, 81(20):4404, 1998.

[85] Huan-Xiang Zhou, Germn Rivas, and Allen P. Minton. Macromolecular crowding and confinement: Biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.*, 37(1):375–397, 2008.

[86] Travis Hoppe and Jian-Min Yuan. Entropic flows, crowding effects, and stability of asymmetric proteins. *Phys. Rev. E*, 80(1):011404, 2009.

[87] A P Minton. Effect of a concentrated "inert" macromolecular cosolute on the stability of a globular protein with respect to denaturation by heat and by chaotropes: a statistical-thermodynamic model. *Biophys. J.*, 78(1):101–109, 2000.

[88] Smita Mukherjee, Matthias M Waegele, Pramit Chowdhury, Lin Guo, and Feng Gai. Effect of macromolecular crowding on protein folding dynamics at the secondary structure level. *J. Mol. Biol.*, 393(1):227–236, 2009.

[89] Jeetain Mittal and Robert B. Best. Dependence of protein folding stability and dynamics on the density and composition of macromolecular crowders. *Biophys. J.*, 98(2):315–320, 2010.

[90] Jyotica Batra, Ke Xu, Sanbo Qin, and Huan-Xiang Zhou. Effect of macromolecular crowding on protein binding stability: Modest stabilization and significant biological consequences. *Biophys. J.*, 97(3):906–911, 2009.

[91] D. Asgeirsson, D. Venturoli, E. Fries, B. Rippe, and C. Rippe. Glomerular sieving of three neutral polysaccharides, polyethylene oxide and bikunin in rat. effects of molecular size and conformation. *Acta Physiol.*, 191(3):237–246, 2007.

[92] Adedayo A Fodeke and Allen P Minton. Quantitative characterization of polymer-polymer, protein-protein, and polymer-protein interaction via tracer sedimentation equilibrium. *J. Phys. Chem. B*, 114(33):10876–10880, 2010.

[93] Myron Peto, Taner Z Sen, Robert L Jernigan, and Andrzej Kloczkowski. Generation and enumeration of compact conformations on the two-dimensional triangular and three-dimensional fcc lattices. *J. Chem. Phys.*, 127(4):044101, 2007.

[94] Piotr Pokarowski, Andrzej Kolinski, and Jeffrey Skolnick. A minimal physically realistic protein-like lattice model: designing an energy landscape that ensures all-or-none folding to a unique native state. *Biophys. J.*, 84(3):1518–1526, 2003.

[95] Ciro Leonardo Pierri, Anna De Grassi, and Antonio Turi. Lattices for ab initio protein structure prediction. *Proteins*, 73(2):351–361, 2008.

[96] H Abe and N Go. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. *Biopolymers*, 20(5): 1013–1031, 1981.

[97] Thomas A. Knotts IV, Nitin Rathore, and Juan J. de Pablo. An entropic perspective of protein stability on surfaces. *Biophys. J.*, 94(11):4473–4483, 2008.

[98] Trinh Xuan Hoang and Marek Cieplak. Molecular dynamics of folding of secondary structures in go-type models of proteins. *J. Chem. Phys.*, 112(15):6851, 2000.

[99] Victor Munoz and William A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci.*, 96(20):11311–11316, 1999.

[100] Pierpaolo Bruscolini and Alessandro Pelizzola. Exact solution of the Muoz-Eaton model for protein folding. *Phys. Rev. Lett.*, 88(25 Pt 1):258101, 2002.

[101] Hoi Sung Chung and Andrei Tokmakoff. Temperature-dependent downhill unfolding of ubiquitin. II. modeling the free energy surface. *Proteins*, 72(1):488–497, 2008.

[102] B. H. Zimm and J. K. Bragg. Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.*, 31(2):526, 1959.

[103] Shneior Lifson and A. Roig. On the theory of HelixCoil transition in polypeptides. *J. Chem. Phys.*, 34(6):1963, 1961.

[104] Wayne L. Mattice and Harold A. Scheraga. Matrix formulation of the transition from a statistical coil to an intramolecular antiparallel beta sheet. *Biopolymers*, 23(9):1701–1724, 1984.

[105] Liu Hong. A statistical mechanical model for antiparallel beta-sheet/coil equilibrium. *J. Chem. Phys.*, 129(22):225101, 2008.

[106] John S. Schreck and Jian-Min Yuan. Exactly solvable model for helix-coil-sheet transitions in protein systems. *Phys. Rev. E*, 81(6):061919, 2010.

[107] B. D McKay. Practical graph isomorphism. *Congr. Numer.*, 30(30):4787, 1981.

[108] Johannes Kbler, Uwe Schning, and Jacobo Torn. *The graph isomorphism problem: its structural complexity*. Birkhauser Verlag, 1993.

[109] J. L. Lebowitz, E. Helfand, and E. Praestgaard. Scaled particle theory of fluid mixtures. *J. Chem. Phys.*, 43(3):774, 1965.

[110] Yng-Gwei Chen and John D. Weeks. Different thermodynamic pathways to the solvation free energy of a spherical cavity in a hard sphere fluid. *J. Chem. Phys.*, 118(17):7944, 2003.

[111] Tom Boublk. Statistical thermodynamics of convex molecule fluids. *Mol. Phys.*, 27(5):1415, 1974.

[112] A P Minton. Molecular crowding: analysis of effects of high concentrations of inert cosolutes on biochemical equilibria and rates in terms of volume exclusion. *Method Enzymol.*, 295:127–149, 1998.

[113] Neal Lesh, Michael Mitzenmacher, and Sue Whitesides. A complete and effective move set for simplified protein folding. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 188–195, Berlin, Germany, 2003. ACM.

[114] Hans-Joachim Bockenhauer, Abu Dayem Ullah, Leonidas Kapsokalivas, and Kathleen Steinhofel. A local move set for protein folding in triangular lattice models. In *Algorithms in Bioinformatics*, volume 5251, pages 369–381. Springer Berlin, 2008.

[115] Andrea G. Cochran, Nicholas J. Skelton, and Melissa A. Starovasnik. Tryptophan zippers: Stable, monomeric -hairpins. *Proc. Natl. Acad. Sci.*, 98(10):5578–5583, 2001.

[116] Faisal A. Syud, Heather E. Stanger, Heather Schenck Mortell, Juan F. Espinosa, John D. Fisk, Charles G. Fry, and Samuel H. Gellman. Influence of strand number on antiparallel beta-sheet stability in designed three- and four-stranded beta-sheets. *J. Mol. Biol.*, 326(2):553–568, 2003.

[117] F Michael Hudson and Niels H Andersen. Measuring cooperativity in the formation of a three-stranded beta sheet (double hairpin). *Biopolymers*, 83(4):424–433, 2006.

[118] Daniele Venturoli and Bengt Rippe. Ficoll and dextran vs. globular proteins as probes for testing glomerular permselectivity: effects of molecular size, shape, charge, and deformability. *Am. J. Physiol. Renal. Physiol.*, 288(4):F605–613, 2005.

[119] Peter N. Lavrenko, Olga I. Mikriukova, and Olga V. Okatova. On the separation ability of various ficoll gradient solutions in zonal centrifugation. *Anal. Biochem.*, 166(2):287–297, 1987.

[120] Apratim Dhar, Antonios Samiotakis, Simon Ebbinghaus, Lea Nienhaus, Dirar Homouz, Martin Gruebele, and Margaret S. Cheung. Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding. *Proc. Natl. Acad. Sci.*, 107(41):17586 – 17591, 2010.

[121] Henry W. Querfurth and Frank M. LaFerla. Alzheimer's disease. *New Engl. J. Med.*, 362(4): 329–344, 2010.

[122] Dominic M. Walsh, Matthew Townsend, Marcia B. Podlisny, Ganesh M. Shankar, Julia V. Fadeeva, Omar El Agnaf, Dean M. Hartley, and Dennis J. Selkoe. Certain inhibitors of synthetic amyloid beta-Peptide (Abeta) fibrillogenesis block oligomerization of natural abeta and thereby rescue Long-Term potentiation. *J. Neurosci.*, 25(10):2455–2462, 2005.

[123] Lih-Fen Lue, Yu-Min Kuo, Alex E. Roher, Libuse Brachova, Yong Shen, Lucia Sue, Thomas Beach, Janice H. Kurth, Russel E. Rydel, and Joseph Rogers. Soluble amyloid [beta] peptide concentration as a predictor of synaptic change in alzheimer's disease. *Am. J. Pathol.*, 155(3): 853–862, 1999.

[124] Andrey V Kajava, John M Squire, and David A D Parry. Beta-structures in fibrous proteins. *Adv. Pro. Chem.*, 73:1–15, 2006.

[125] Victor A Streltsov, Joseph N Varghese, Colin L Masters, and Stewart D Nuttall. Crystal structure of the amyloid-beta p3 fragment provides a model for oligomer formation in alzheimer's disease. *J. of Neurosci.*, 31(4):1419–1426, 2011.

[126] T. D. Lee and C. N. Yang. Statistical theory of equations of state and phase transitions. II. lattice gas and ising model. *Phys. Rev.*, 87(3):410, 1952.

[127] ME Fisher. The Nature of Critical Points, Lectures in Theoretical Physics Vol. *University of Colorado Press*, 1965.

[128] Benny Godlin, Emilia Katz, and Johann A Makowsky. Graph polynomials: From recursive definitions to subset expansion formulas. *arXiv*, 0812.1364, 2008.

[129] P. W. Kasteleyn and C. M. Fortuin. *J. Phys. Soc. Jpn. Suppl.*, 26:11–14, 1969.

[130] F. Y. Wu. Duality transformation in a many-component spin model. *J. Math. Phys.*, 17(3): 439, 1976.

[131] Andrea Bedini and Jesper Lykke Jacobsen. A tree-decomposed transfer matrix for computing exact potts model partition functions for arbitrary graphs, with applications to planar graph colourings. *J. Phys. A*, 43(38):385001, 2010.

[132] Neil Robertson and P. D. Seymour. Graph minors. III. planar tree-width. *J. Comb. Theor. B*, 36(1):49–64, 1984.

[133] Stefan Arnborg, Derek G Corneil, and Andrzej Proskurowski. Complexity of finding embeddings in a k-tree. *SIAM J. Alg. and Disc. Meth.*, 8:277–284, 1987.

[134] Stefan Arnborg and Andrzej Proskurowski. Linear time algorithms for NP-hard problems restricted to partial k-trees. *Disc. Appl. Math.*, 23(1):11–24, 1989.

[135] J Petersen. Sur le thorme de Tait. *L'Intermdiaire des Mathmaticiens*, 5:225–227, 1898.

[136] J.A. Bondy and U.S.R. Murty. *Graph theory with applications*, volume 290. MacMillan, 1976.

[137] S Banu Ozkan, Ken A Dill, and Ivet Bahar. Computing the transition state populations in simple protein models. *Biopolymers*, 68(1):35–46, 2003.

[138] Huan-Xiang Zhou. A minimum-reaction-flux solution to master-equation models of protein folding. *J. Chem. Phys.*, 128(19):195104, 2008.

[139] Martin Karplus and J. Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.*, 9(9):646–652, 2002.

[140] T William, Y. Kalmykov, P. Waldron, and J T Coffey. *The Langevin Equation*. World Scientific Pub Co Inc, 2004.

[141] Masuo Suzuki. General theory of higher-order decomposition of exponential operators and symplectic integrators. *Phys. Lett. A*, 165:387–395, 1992.

[142] Daniel Goldman and Tasso J. Kaper. Nth-Order operator splitting schemes and nonreversible systems. *SIAM J. Num. Anal.*, 33(1):349, 1996.

[143] Andrea Ricci and Giovanni Ciccotti. Algorithms for brownian dynamics. *Mol. Phys.*, 101: 1927–1931, 2003.

[144] Giovanni Ciccotti and Galina Kalibaeva. Deterministic and stochastic algorithms for mechanical systems under constraints. *Phil. Trans. Math.*, 362(1821):1583–1594, 2004.

[145] E B Davies. Embeddable markov matrices. *arXiv*, 2010.

[146] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Rev. Mod. Phys.*, 62(2):251, 1990.

[147] Hue Sun Chan and Ken A. Dill. The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.*, 92(5):3118, 1990.

[148] Louis Komzsik. *The Lanczos Method: Evolution and Application*. Society for Industrial Mathematics, 1987.

[149] Gerhard Stock, Kingshuk Ghosh, and Ken A Dill. Maximum caliber: a variational approach applied to two-state dynamics. *J. Chem. Phys.*, 128(19):194102, 2008.

[150] E. T. Jaynes. The minimum entropy production principle. *Anal. Rev. Phys. Chem.*, 31: 579–601, 1980.

[151] Gavin E. Crooks. Measuring thermodynamic length. *Phys. Rev. Lett.*, 99(10):100602, 2007.

[152] Sanzo Miyazawa and Robert L Jernigan. RL: an empirical energy potential with a reference state for protein folding and sequence recognition. proteins. pages 36—357, 1999.

Appendix A

Macro state approximation to the specific heat

Often times algorithms such as Wang-Landau (see Section 2.2.2) have convergence properties proportional to the cardinality of the density of states. For many systems characterized by a low degeneracy in the lowest energy levels, the particular form of the density of states at high energies is not important when calculating derivatives of the partition function. Therefore, we find it useful to approximate these higher energy states as a single macrostate. We derive a result for the error introduced by approximating a portion of the density of states $g(x)$ as a single macro state. While the results presented here are for a continuous one-dimensional density of states, the generalization to higher dimensions or discretized states is straightforward.

We start with the expression of the partition function, assuming the energy range is continuous and spans the interval A to B

$$\mathcal{Z} = \int_A^B g(x)e^{-x\beta}dx \tag{A.1}$$

We approximate a small portion of the larger energies as a single macrostate in the interval $B(1-\epsilon)$ to B. Over this interval we assume the energy is the constant value x_E. The most straightforward choice for x_E is a weighted average over the approximated states

$$x_E = \frac{\int_{B(1-\epsilon)}^B xg(x)e^{-x\beta}dx}{\int_{B(1-\epsilon)}^B g(x)e^{-x\beta}dx} \tag{A.2}$$

The approximated form of the partition function becomes

$$\mathcal{Z}_\epsilon = \int_A^{B(1-\epsilon)} g(x)e^{-x\beta}dx + \int_{B(1-\epsilon)}^B g(x)e^{-x_E\beta}dx \tag{A.3}$$

$$= \int_A^{B(1-\epsilon)} g(x)e^{-x\beta}dx + e^{-x_E\beta}\int_{B(1-\epsilon)}^B g(x)dx \tag{A.4}$$

From here, we can ask how accurate this approximation is. By considering small ϵ from the term $B(1-\epsilon)$, we take a Taylor expansion around $\epsilon = 0$,

$$\mathcal{Z}_\epsilon = \sum \frac{\epsilon^n}{n!} \left[\frac{\partial \mathcal{Z}_\epsilon}{\partial \epsilon^n} \right]_{\epsilon=0} \tag{A.5}$$

When evaluated, this has the surprising result that the first two non-constant terms are zero

$$\mathcal{Z}_\epsilon = \mathcal{Z}\epsilon^0 + 0\epsilon^1 + 0\epsilon^2 + \Delta\epsilon^3 + O(\epsilon^4) \tag{A.6}$$

implying that our approximation is very good indeed.[1] The third order term is

$$\Delta = -\frac{B^3 g(B) e^{-B\beta} \beta^2}{24} \tag{A.10}$$

At very high temperatures we can approximate Δ by expanding around $\beta = 0$

$$\lim_{\beta \to 0} \Delta = -\frac{B^3 g(B) \beta^2}{24} + O(\beta^3) \tag{A.11}$$

If we have the explicit form for the density of states such as $g(x) = e^{\gamma x}$ (typically for Ising like

[1] In order to take the derivatives we recall that taking the derivative over the limits of integration can be done by

$$\frac{\partial}{\partial a} \int_{g(a)}^{b} f(x)dx = \frac{\partial}{\partial a} \left[F(b) - F(g(a)) \right] \tag{A.7}$$

$$= -\left(\frac{dF(t)}{dt}\bigg|_{t=g(a)} \right) \left(\frac{dg(a)}{da} \right) \tag{A.8}$$

$$= -f(g(a)) \left(\frac{dg(a)}{da} \right) \tag{A.9}$$

where F is the antiderivative of f.

systems) we can compute the expansion around ϵ to many more terms

$$\begin{aligned}\mathcal{Z}_\epsilon(g(x)=e^{\gamma x}) = \mathcal{Z} & \\ & -(1/24)B^3e^{B(\gamma-\beta)}\beta^2\epsilon^3 \\ & +(1/48)B^4\beta^2e^{B(\gamma-\beta)}(\gamma-\beta)\epsilon^4 \\ & -(1/5760)B^5\beta^2e^{B(\gamma-\beta)}(-72\beta\gamma+33\beta^2+28\gamma^2)\epsilon^5+ \\ & +(1/11520)B^6\beta^2e^{B(\gamma-\beta)}(45\beta^2\gamma-40\beta\gamma^2+8\gamma^3-13\beta^3)\epsilon^6+O(\epsilon^7)\end{aligned} \tag{A.12}$$

Worked Example : Gaussian DOS

To illustrate the power of the method we show an example at various levels of approximation. We choose a density of states that is shaped like a Gaussian, i.e. a high degeneracy of intermediate states and a low probability of the extreme energies. This is characteristic of the standard Ising model; roughly, an inverted parabolic shape is found when the log of the density of states is plotted against the energies of the system. The chosen Hamiltonian is linear, again in connection with the Ising model. For convenience and clarity in plotting, we shift all energies by a constant factor $\mu = 6$, so that the largest accessible state has zero energy

$$\mathcal{H}(x) = \begin{cases} x-\mu & 0 \le x \le \mu \\ \infty & \text{otherwise} \end{cases} \tag{A.13}$$

$$g(x) = \exp\left(\frac{-(x-\mu)^2}{2}\right) \tag{A.14}$$

In Figures A.1, A.2, A.3, A.4, we have plotted the approximate specific heat CV_ϵ against the exact value as a function of β. The leftmost graph shows the area of the density states that was considered a single macrostate. At low values of ϵ the approximation is nearly indistinguishable from the exact curve. In Figure A.3, $\epsilon = \frac{1}{2}$, that we have approximated half of the state space *as a single macrostate* and still achieved a reasonable approximation to the specific heat. However, if ϵ is pushed to high the resulting approximations eventually break down, ultimately creating new spurious phase changes

(Figure A.4).

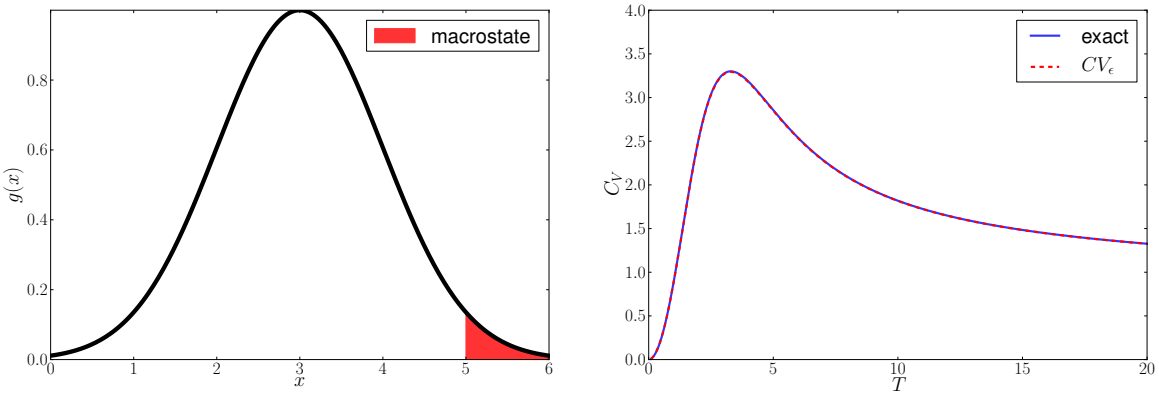


Figure A.1: CV_ϵ with $\epsilon = 1/6$. Approximately 2.1% of $g(x)$ is considered as a single macrostate (see text for details).

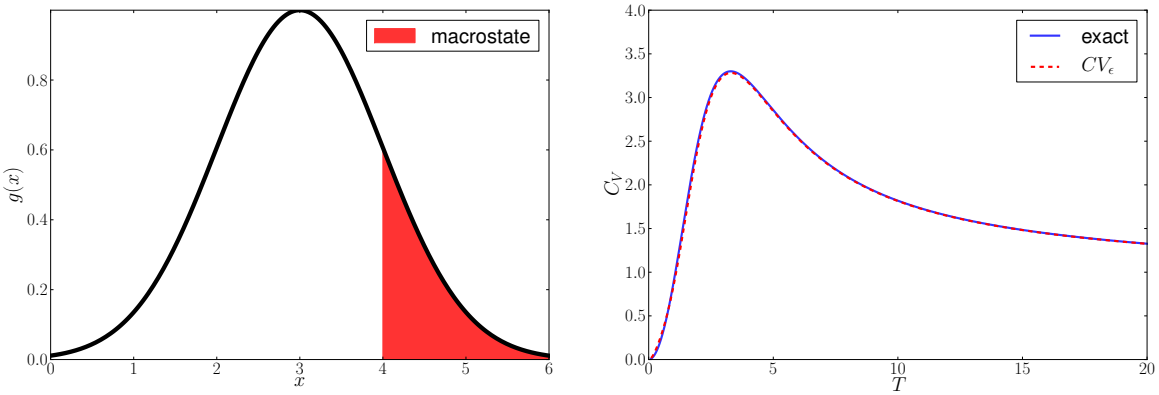


Figure A.2: CV_ϵ with $\epsilon = 1/3$. Approximately 15.7% of $g(x)$ is considered as a single macrostate (see text for details).

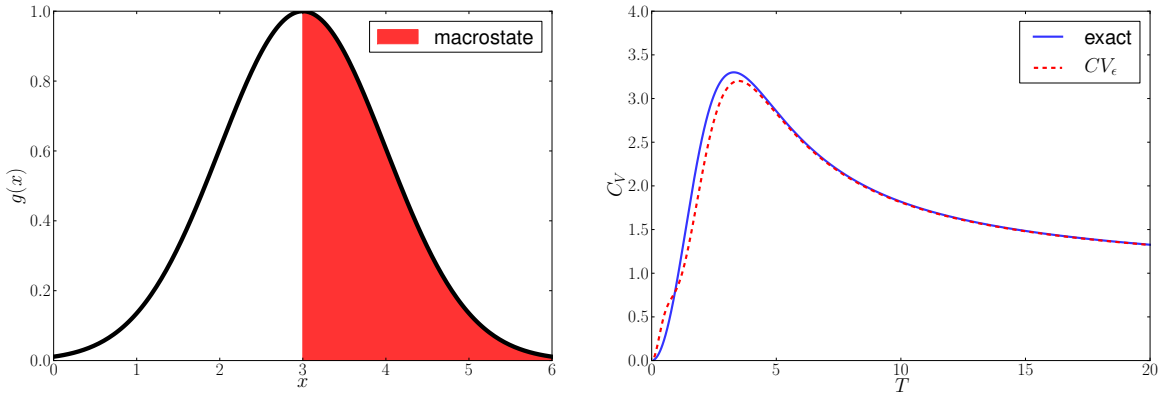


Figure A.3: CV_ϵ with $\epsilon = 1/2$. 50.0% of $g(x)$ is considered as a single macrostate (see text for details).

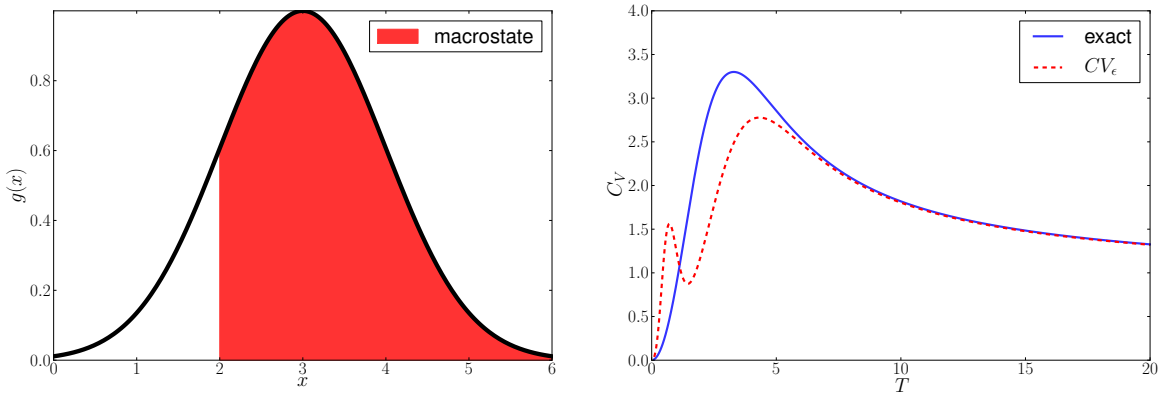


Figure A.4: CV_ϵ with $\epsilon = 2/3$. Approximately 84.2% of $g(x)$ is considered as a single macrostate (see text for details).