Multi-scale modeling in molecular self-assembly

by

Tien Minh Phan

M.S., Ho Chi Minh University of Education, Vietnam, 2012

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

The study of molecular self-assembly has attracted considerable interest over the decades due to its wide-ranging applications in chemistry, materials science, and biology. The challenge in studying molecular self-assembly is that it involves complex spatiotemporal scales ranging from short-lived microscopic events to lifelong macroscopic architectures. This work aims to investigate the molecular self-assembly in different systems to better understand the macroscopic properties of microscopic activities. To overcome the complexity of time scale and length scale, we employed a combination of computer simulations and simple analytic theories to develop multi-scale models to study the systems of interest. We find that: (1) in the slow growth regime of crystalline solids, impurity particles can speed up crystal growth with minimal impact on the final product, (2) in the self-assembly of peptides into amyloid fibrils the conformational entropy plays an important role in the transition from nucleation to elongation, (3) in biomolecule condensates, the surface tension arises from the competition between binding energy and configurational entropy. These results highlight the power of multi-scale models to interpret macroscopic physical observables in terms of fundamental microscopic mechanisms in molecular self-assembly processes.

Multi-scale modeling in molecular self-assembly

by

Tien Minh Phan

M.S., Ho Chi Minh University of Education, Vietnam, 2012

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Jeremy D. Schmit

Copyright

Tien Minh Phan

2021

Abstract

The study of molecular self-assembly has attracted considerable interest over the decades due to its wide-ranging applications in chemistry, materials science, and biology. The challenge in studying molecular self-assembly is that it involves complex spatiotemporal scales ranging from short-lived microscopic events to lifelong macroscopic architectures. This work aims to investigate the molecular self-assembly in different systems to better understand the macroscopic properties of microscopic activities. To overcome the complexity of time scale and length scale, we employed a combination of computer simulations and simple analytic theories to develop multi-scale models to study the systems of interest. We find that: (1) in the slow growth regime of crystalline solids, impurity particles can speed up crystal growth with minimal impact on the final product, (2) in the self-assembly of peptides into amyloid fibrils the conformational entropy plays an important role in the transition from nucleation to elongation, (3) in biomolecule condensates, the surface tension arises from the competition between binding energy and configurational entropy. These results highlight the power of multi-scale models to interpret macroscopic physical observables in terms of fundamental microscopic mechanisms in molecular self-assembly processes.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation is the end product of my long journey at K-State. On this journey, I have received enormous help and advice from a lot of people. I would like to thank everyone even though I cannot mention all of them here.

First and foremost, I am very much indebted to my major advisor, Professor Jeremy D. Schmit. Words can not express how lucky I am to have Professor Schmit as a mentor and a friend during my time in graduate school. Many of the ideas and methods represented in this work took shape from countless, insightful, and timely discussions with Professor Schmit over the past five years. He also helped me every step on the way to grow as a scientist in training, such as writing a concise paper, organizing random ideas into a story, delivering an insightful presentation without losing the audience. He is always eager and patient to answer every single question, even a very silly one. His sense of humor and his attitude towards science and life make me think more positively in my "downtime". I wish I would have more time to learn from him.

I would like to thank Professors Christopher M. Sorensen, Paul E. Smith, and Matthew J. Berg for serving on my dissertation committee and giving me valuable advice and assessment during my research. Also, I am grateful to Professor Mingjun Wei for serving as the outside chairperson for my dissertation defense meeting. I would also like to thank Dr. Stephen Whitelam, research scientist at the Molecular Foundry at Lawrence Berkeley National Lab, for collaborating with us on the catalyst project. His expertise and defensive scientific style of writing have been a valuable example to me.

I am very grateful for the support from the Physics Department at Kansas State University during my time working as a graduate teaching assistant. I was very fortunate to work with Peter Nelson during this time. I learned a lot of good things from him, especially how to fix any broken equipment in the stockroom. I owe special thanks to Ms. Brandi Lohman,

a great boss and dedicated teacher. I learned from her how to explain difficult concepts in an understandable way to students.

Thank you to all members of the Schmit group: Tam, Kamal, Nelson, and Yuba. Your company and the valuable discussions have been an inspiration to go to work, and will always remain one of the high points of my PhD experience. Many thanks to Kamal "Sir", who was with me during my difficult time after we spent an "unforgettable" first summer and moved to the Schmit group.

Last but not least, I would like to thank my parents, my brothers, and my beloved wife for all their sacrifice, support, and encouragement. This would not have been possible without their unconditional love.

Dedication

To my parents, my brothers, and my loving wife.

Chapter 1

Introduction

1.1 Motivation

The term molecular self-assembly describes dynamical processes in which a collection of particles organize themselves, without external driving forces, into ordered patterns or structures[1–5]. Molecular self-assembly is the basic phenomena ranging from the construction of nano-materials[1] to the formation of large-scale functional structures to generate molecular machinery of life[6]. There are several reasons for interest in studying molecular self-assembly. First, self-assembly is an intriguing physics problem of obtaining order from disorder. Second, understanding self-assembly of living cells is fundamental to understanding life on Earth and the possibility of life elsewhere in the universe[7–9]. Third, self-assembly offers great potentials for making new and useful materials. For example, on the molecular scale, DNA-mediated interactions have been exploited to assemble a variety of ordered crystalline structures[10–12]. On the colloidal scale, patchy particles with anisotropic interactions[13–15] and particles with controllable geometrical shapes[16,17] also enable the assembly of novel structures. Finally, an important class of molecular self-assembly in biology is protein aggregation of which the typical end-product is amyloid fibril, structurally dominated by cross-β sheet structures. The aggregation process has been linked to over 50 human diseases[18], most notably Alzheimer's and Parkinson's diseases. Understanding mechanisms of protein aggregation will shed light

on the development of therapeutic strategies for aggregation diseases.

The processes of molecular self-assembly span a wide range of spatial and temporal scales and the choice of approach to study them depends on the question asked. In many cases the best way is an experimental technique. However, a central problem in this approach is how to interpret the macroscopic observables measured from the experiments in terms of the microscopic mechanisms to reveal the dynamic behaviors of the system. To bridge the gap between microscopic and macroscopic scales, theoretical methods together with modeling and simulation have made tremendous advances in the last few decades to complement experimental molecular biology investigations. Molecular dynamics (MD) simulation, for example, is an invaluable tool to study biomolecules *in silico*. MD simulations now can reach the timescales of microseconds for systems having tens of thousands of atoms or even millisecond timescales for some systems with scalable codes[19–21] and specialized computers[22,23]. Although these simulations can provide microscopic details of the system and may seem both large and long at the atomic scale, at the real scale of molecular self-assembly they are only a small piece of the overall picture.

This dissertation is an attempt to use coarse-graining simulations together with analytic theories to develop multi-scale models to access longer time scale and provide new physical insights into different self-assembly systems. This combination helps us understand how layer-by-layer growth of crystalline solids induced by impurities, explore the role of conformational entropy in the transition from nucleation to elongation of amyloid aggregation and investigate the entropic and energetic contributions to the surface tension of biomolecule condensates.

1.2 Dissertation outline

The dissertation is structured as follows:

Chapter 2 introduces biological concepts relevant to this work. We start the discussion with proteins and their classification as well as the interactions at the amino acid level. Next we describe the hierarchical levels of protein structures and the Levinthal's paradox and a

solution to the protein folding problem. Then, we briefly summarize the protein aggregation processes and pathological relevance. Finally, we end the chapter with a short discussion on the nucleation mechanisms, which are widely applied to many fields of science.

Chapter 3 summarizes the computational methods used in this work. We introduce the classical molecular dynamics simulation, a powerful tool employed to study biomolecular structure, dynamic and function. Next, we review the Monte Carlo method, another powerful computer simulation technique used to study equilibrium properties of molecular systems. We also describe the kinetic Monte Carlo technique, which is a solution to the "time-step" problem in molecular dynamics simulations for some processes in nature.

Chapter 4 provides an in-depth description on the role of impurity particles in the growth of crystalline solids. Impurities are known to hinder growth by poisoning the crystal surface. Here we find a surprising result to the contrary: in the slow growth regime, impurities can accelerate crystal growth, with minimal impact upon the final crystal quality. In this respect they act almost as a catalyst. We demonstrate this effect using simulations, and present scaling arguments that indicate the mechanism to be broadly applicable. For instance, the mechanism is likely to play an important role in the crystallization of anisotropic particles such as biomolecules.

Chapter 5 explores how conformational entropy limits the transition from nucleation to elongation in amyloid aggregation. The formation of amyloid fibrils in Alzheimer's disease and other neurodegenerative disorders is limited by a slow nucleation step due to the entropic cost to initiate the ordered cross-β structure. We find that the optimal degree of order in a nucleus depends on protein concentration. Low concentration systems require more ordered nuclei to capture infrequent monomer attachments. The nucleation phase transitions to the elongation phase when the β-sheet core becomes large enough to overcome the initiation cost, at which point further ordering becomes favorable and the nascent fibril efficiently captures new molecules.

Chapter 6 investigates the entropic and energetic contributions to biomolecule condensate surface tension. We study models ranging from a single component system that associates by purely energetic nearest-neighbor interactions, to a two-component system that mimics

the entropy dominated mechanism. Between these limits is an intermediate case of a two-component system with an energy-dominated attraction mechanism. We use lattice simulations and analytic theory to understand how network connectivity affects the mechanism of attraction and surface tension.

Chapter 7 provides a summary of what we have learned and of our potential future directions.

References

[1] G. M. Whitesides, J. P. Mathias, and C. T. Seto, Science **254**, 1312 (1991).

[2] M. F. Hagan and D. Chandler, Biophysical journal **91**, 42 (2006).

[3] A. W. Wilber, J. P. Doye, A. A. Louis, E. G. Noya, M. A. Miller, and P. Wong, The Journal of chemical physics **127**, 08B618 (2007).

[4] B. Bozorgui, D. Meng, S. K. Kumar, C. Chakravarty, and A. Cacciuto, Nano letters **13**, 2732 (2013).

[5] J. Madge and M. A. Miller, The Journal of chemical physics **143**, 044905 (2015).

[6] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).

[7] J. Cohen, Science **267**, 1265 (1995).

[8] J. Trevors, Antonie van Leeuwenhoek **72**, 251 (1997).

[9] C. De Duve and R. De Neufville, *Blueprint for a cell: the nature and origin of life* (Carolina Biological Supply Company, 1991).

[10] K. T. Nam, S. A. Shelby, P. H. Choi, A. B. Marciel, R. Chen, L. Tan, T. K. Chu, R. A. Mesch, B.-C. Lee, M. D. Connolly, et al., Nature materials **9**, 454 (2010).

[11] T. Gibaud, E. Barry, M. J. Zakhary, M. Henglin, A. Ward, Y. Yang, C. Berciu, R. Oldenbourg, M. F. Hagan, D. Nicastro, et al., Nature **481**, 348 (2012).

[12] A. Stannard, J. C. Russell, M. O. Blunt, C. Salesiotis, M. del Carmen Giménez-López, N. Taleb, M. Schröder, N. R. Champness, J. P. Garrahan, and P. H. Beton, Nature chemistry **4**, 112 (2012).

[13] A. B. Pawar and I. Kretzschmar, Macromolecular rapid communications **31**, 150 (2010).

[14] D. J. Kraft, J. Groenewold, and W. K. Kegel, Soft Matter **5**, 3823 (2009).

[15] S. Jiang, Q. Chen, M. Tripathy, E. Luijten, K. S. Schweizer, and S. Granick, Advanced materials **22**, 1060 (2010).

[16] L. Rossi, S. Sacanna, W. T. Irvine, P. M. Chaikin, D. J. Pine, and A. P. Philipse, Soft Matter **7**, 4139 (2011).

[17] S. Sacanna, M. Korpics, K. Rodriguez, L. Colón-Meléndez, S.-H. Kim, D. J. Pine, and G.-R. Yi, Nature communications **4**, 1 (2013).

[18] C. M. Dobson, T. P. Knowles, and M. Vendruscolo, Cold Spring Harbor perspectives in biology **12**, a033878 (2020).

[19] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, Journal of computational chemistry **26**, 1781 (2005).

[20] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1**, 19 (2015).

[21] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, et al., PLoS computational biology **13**, e1005659 (2017).

[22] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, et al., Communications of the ACM **51**, 91 (2008).

[23] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, et al., in *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (IEEE, 2014), pp. 41–53.

Chapter 2

Background

2.1 Proteins

The term "protein" (Greek: *proteios*, primary) was coined by the Swedish chemist Jacob Berzelius in 1838. Proteins are the most fundamental ingredient of a living organism. They have remarkably structural diversity and catalytic activity[1], which play a central role in nearly all biological processes. Proteins are linear polymer molecules of amino acids, which are linked into a peptide chain by forming peptide bonds, and their chemical properties are determined by the interactions of the side chains and the peptide backbone. Amino acids are the building blocks of proteins. All of the 20 natural common amino acids have a basic structure consisting of a central carbon atom (C_α) linked to a hydrogen atom (H), a carboxyl group (COOH), an amino group (NH_2) and a side chain that distinguishes one amino acid from the others (Gly – G lacks a side chain). Amino acids are classified into groups based on the unique chemical nature of their side chains, which are hydrophobic, charged and polar (Fig. 2.1). The first group comprises strictly hydrophobic side chains, which can be subdevided into two categories: aliphatic with linear chains (Ala – A, Val – V, Leu – L, Ile – I, and Met – M) and aromatic with ring structures (Phe – F, Tyr – Y, and Trp – W). Proline (Pro – P) is often considered hydrophobic amino acid and shares many properties with the aliphatic group. When proline is in a peptide bond, it does not have a hydrogen

on the α amino group, so it cannot donate a hydrogen bond to stabilize an α-helix or a β-sheet. The proline ring limits degrees of freedom around the dihedral angle, which may lead to a reduction of conformational entropy of the polypeptide chains. Proline can also form favourable stacking interactions with aromatic systems[2,3]. The four charged residues Asp – D, Glu – E, Lys – K and Arg – R, form the second group. The third group consists of polar side chains: Ser – S, Thr – T, Cys – C, His – H, Asn – N, and Gln – Q.

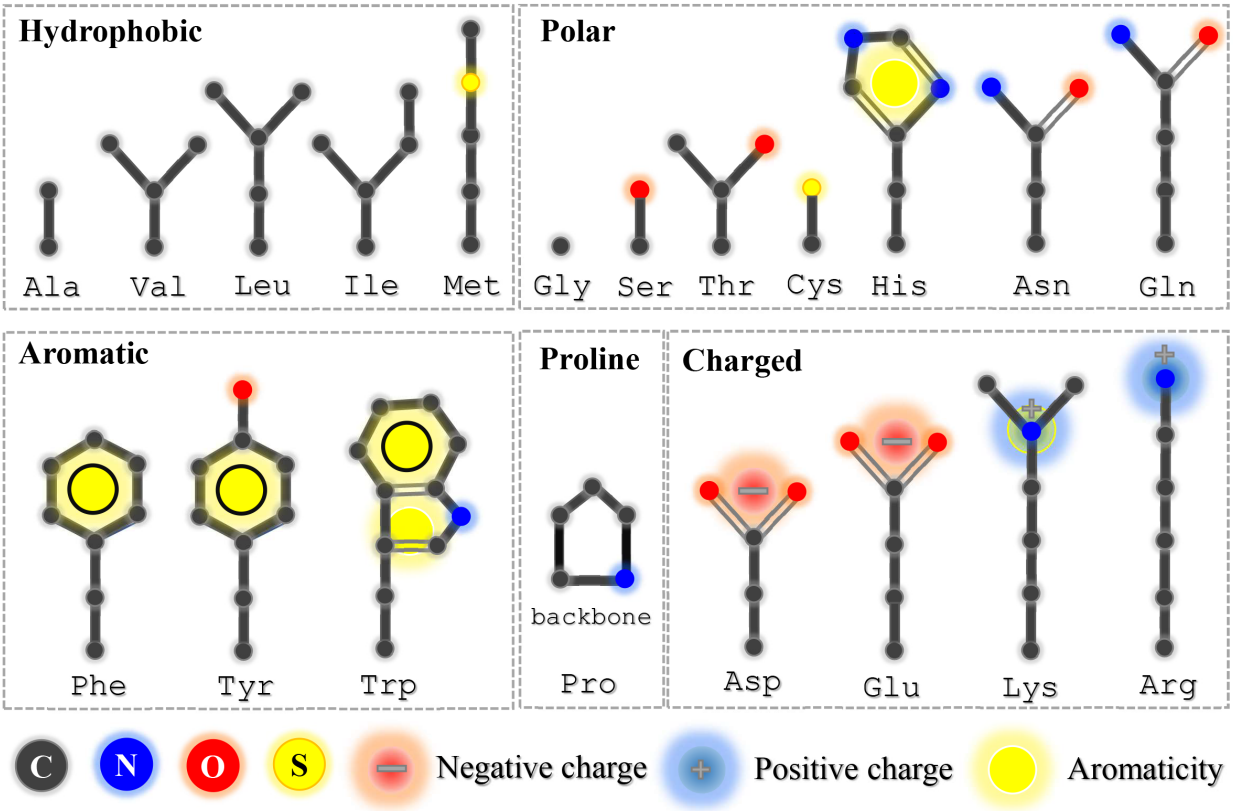

Figure 2.1: Schematic representations of twenty natural amino acids (hydrogen atoms are not shown) with charge and aromaticity color-coded in blue/red and yellow, respectively. Adapted from Ref. Martin and Holehouse[4].

While chemical properties of amino acid side chains are key to understanding the large-scale characteristic of proteins, the polypeptide backbones are vital to protein physical chemistry. The polypeptide backbones provide hydrogen bonding with donors and acceptors in carbonyl oxygen and amide proton, respectively. Moreover, three backbone dihedral angles defines the intrinsic flexibility of proteins. The peptide backbone can thus be regarded as a flexible, polar homopolymer.

The interactions between amino acids mainly stem from electrostatics:[5,6]

Covalent bond ($\sim 100\ k_{\mathrm{B}}T$): chemical bond due to the mutual sharing pairs of electron between atoms.

Coulomb (~ 1 – $10\ k_{\mathrm{B}}T$): interaction between charged atoms. In biological environment, the ions are not in vacuum but in the presence of surrounding solvent involving large gradients in the dielectric constant (≈ 80 in water compared to ≈ 5 in the interior protein). *Hydrogen bond*, a Coulombic interaction which are ubiquitous at the heart of several biological phenomena, such as the formation of α-helix and β-sheet secondary structures in protein (Section 2.2). It is attractive, highly-directional, non-bonded interaction between a polar hydrogen atom ("donor") and an electronegative atom (e.g., nitrogen, oxygen) with a nonbonding orbital ("acceptor"); the hydrogen atom must be covalently bonded to another electronegative atom to leave it with a partial positive charge[5,7].

van der Waals ($\lesssim 1\ k_{\mathrm{B}}T$): interactions, which act between all pairs of atoms and do not depend on charges, describe a strong repulsion when two atoms are too close together and an attraction when they are a bit further apart.

Hydrophobic effect ($\lesssim 1\ k_{\mathrm{B}}T$:) entropic driving force for self-association of non-polar groups in water; the hydrophobic effect describes the tendency of water to minimize its contacts with nonpolar substances due to a hydrophobic group which enforces constraints on the hydrogen-bond network in its vicinity[5,8].

Among these interactions, covalent bonds are the strongest, others are relatively weak that are comparable with thermal fluctuation at room temperature. Therefore, thermal fluctuations will have a central role in the protein structures and entropic effects are a major driving force in protein folding and dynamics.

2.2 Levinthal's paradox and protein folding funnels

Proteins are synthesized on ribosomes and often released as unstructured chains. Each chain then folds into a specific three dimensional (3D) structure, which is encoded in the amino acid sequence, either by itself or with the help from chaperone molecules. This 3D

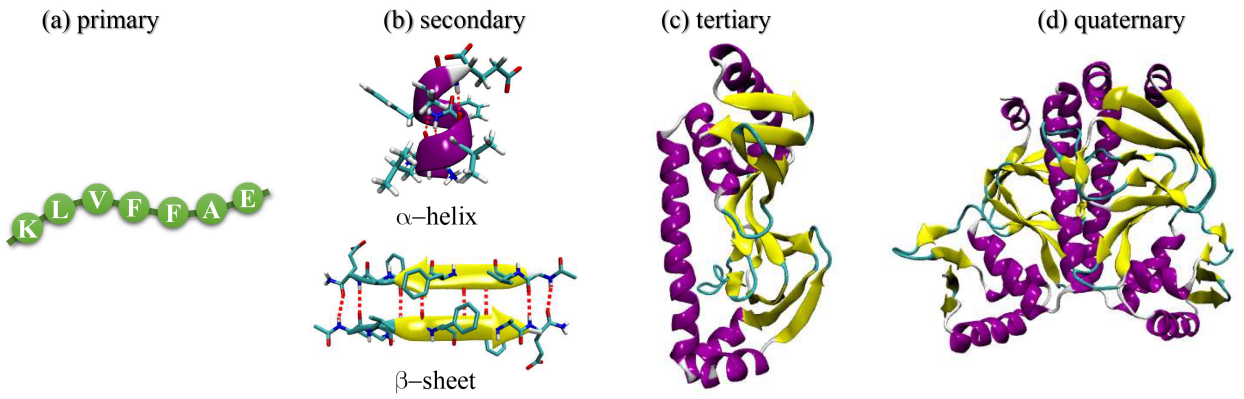

Figure 2.2: (a) An example of primary structures or sequences ($A\beta_{16-22}$). (b) The secondary structures are the helices and sheets (β-strands are shown as arrows and the red dashed lines indicate hydrogen bonds). (c) Tertiary structure of protein (2CGP). (d) Quaternary structure with two chains (1CGP).

folded structure (also known as native state) is essential to the protein's proper biological function. Protein structure can be described in terms of four hierarchical levels, primary, secondary, tertiary, and quaternary (Fig. 2.2). The primary structure is a linear amino acid sequence of its polypeptide chain. The secondary structure refers to α-helices and β-sheets, two main types of ordered substructures in proteins, which have particularly stable hydrogen-bonded arrangements of amino acids. The tertiary structure includes all aspects of the 3D arrangement of its secondary structures including its connecting turns, loops, or coiled segments. The quaternary structure describes how various polypeptide chains come together to form a multi-subunit complex.

Protein folding is a complex process in which a newly synthesized protein finds its way to reach a unique stable conformation rather than one of countless alternatives. From a statistical point of view, if a protein has 100 amino acid residues and each residue can adopt two possible orientations for a trivial model, we obtain 2^{100} possible conformations. Let us assume that the conversion from one configuration to another takes 100 picoseconds, then it would take 5×10^8 years to extensively search among all conformations for the native state. Here is the central question, since the number of possible conformations for just a small polypeptide chain is astronomically large, how can a given protein find its unique native structure in an accessible time? Furthermore, it is quite surprising that real proteins fold

rapidly, often less than 1 second[9]. This puzzle is known as the Lavinthal's paradox[10].

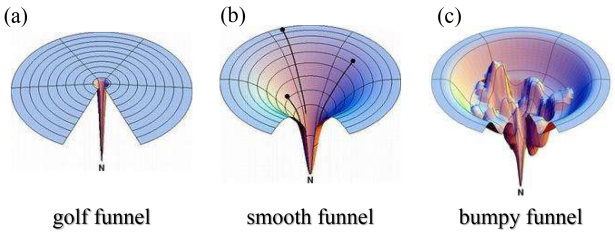

Figure 2.3: (a) A "golf-course" energy landscape, (b) an ideal funnel landscape, and (c) a bumpy energy landscape. The energy landscape figures are licensed under CC BY 4.0.

The solution to this paradox involves the proposal of the folding energy landscape on which the folding pathway occurs[11–14]. An energy landscape is a mathematical function $G(x_1, x_2, ..., x_L)$ of independent variables $x_1, x_2, ..., x_L$, which are a chain conformational degrees of freedom such as the geometric features described by bond lengths, bond angles, intermolecular distances, etc. In the protein folding problem, G is a high-dimensional free energy surface. Levinthal's paradox describes an extensive search occurs on a "golf-course" free energy landscape (Fig. 2.3a), which is high-dimensional space, flat everywhere except for the minimum of the native state. On this flat surface, any random pathway to the localized well is equally probable which may lead to an almost infinite time. Nevertheless, statistical mechanical theories derived in the 1980s and 1990s showed that folding landscapes do not look like golf courses, instead they prefer funnel-like shape[15–17] (Fig. 2.3b). Folding funnels provide a "new view" and a simple way of understanding protein folding. In a funnel-like energy landscape, there is a competition between enthalpy and entropy resulting in small free energy of just 1–10 $k_{\mathrm{B}}T$[11,18]. The energetic requirements for folding an unstructured protein into its native conformation, among other things, are the formation of hydrophobic contacts and hydrogen bonds, which leads to a strong loss in entropy and a big gain in energy. An ideal funnel landscape (Fig. 2.3b) is large and open on the top and becomes small and confined at the bottom. When a protein starts folding it increases the number of intramolecular contacts and lowers its internal free energy, which leads to a reduction of conformational search. Fig. 2.3c shows a bumpy landscape with kinetic traps and some narrow pathways to the native state, illustrating multi-state folding. In this funnel, there is also a bottleneck region corresponding to an ensemble of conformations of transi-

tion state[19–21]. To continue going down hill, the protein needs to overcome the transition state by breaking different contacts for example. The folding process on a bumpy funnel is fundamentally slower than on a smooth one due to kinetic traps and energy barriers.

2.3 Protein aggregation

In a cell, protein folding takes place in a complex and highly crowded environment, which can dramatically affect folding and association between proteins when the conditions deviate from the physiological optimum. Undoubtedly, chaperone molecules and folding catalysts are able to mitigate the complex misbehavior of the folding to provide some protection for incompletely folded protein[22,23] and accelerate the slow steps in the folding process including peptidyl-prolyl isomerases[24,25]. However, under the changes of temperature, pH in the cellular environment or mutation, post translational modifications in the proteins, misfolding events may occur during the search for the native conformation, and misfolded proteins may further self-assemble into potentially toxic aggregate structures that may be harmful to the cell[26]. When proteins aggregate, they may form insoluble dense supramolecular assemblies known as amyloid fibrils, and once formed they are essentially indestructible under physiological conditions. Intracellular and extracellular fibrillar aggregates in cells are a characteristic feature of many common types of neurodegenerative diseases such as Alzheimer's and Parkinson's diseases[27–29]. Yet, it is noteworthy that the protein aggregation is not always problematic, amyloids have also been connected with biological functions[26,30] and functional amyloids are found in several organs[26].

"Amyloid" (Greek, *amylon*, starch) was coined initially in a botanical context by Schleiden[32], and later transferred into medicine to describe human-pathogenic deposits by Virchow and others[33]. Amyloid fibrils can be described as a cross-β structure of polypeptide chains. Cross-β structures represent intertwined layers of β-sheets extending in a direction parallel to the fibril axis, with perpendicular hydrogen bonds[34–36] (Fig. 2.4a). The discovery of cross-β structures in all amyloid fibrils was initially shown in 1968 by X-ray diffraction measurements[37]. More recently, studies of peptide microcrystals have revealed the interactions

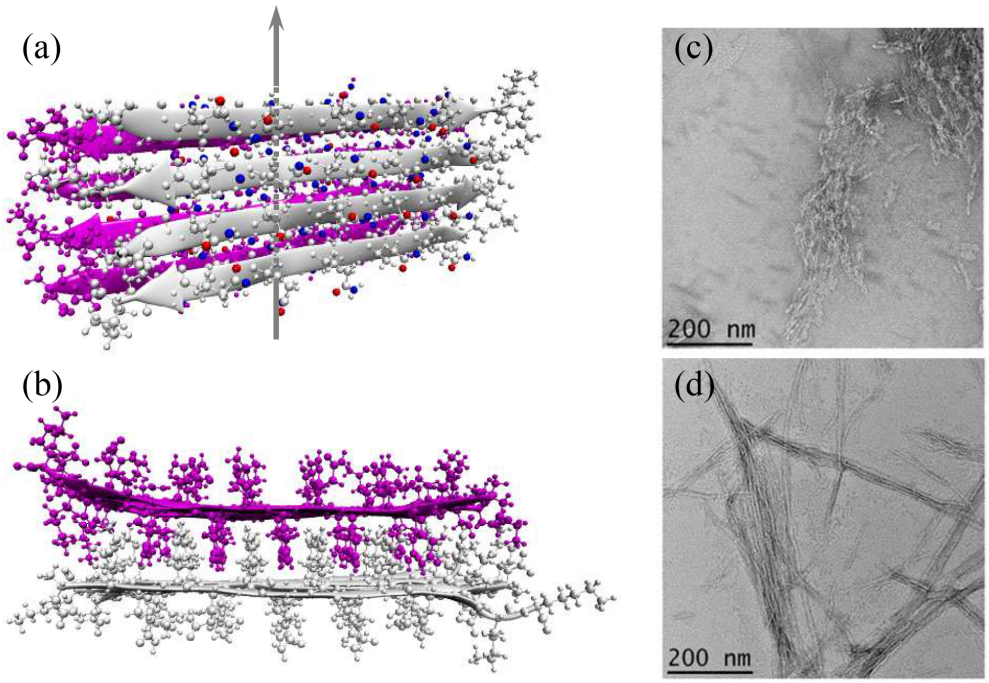

Figure 2.4: (a) Structure of cross-β spine and (b) side view of representative steric zipper structure of anti-parallel β-sheets of polyglutamine $D_2Q_{10}K_2$ (each chain consists of 10 glutamine amino acids and the two amino acids of aspartic acid (D) or lysine (K) amino acids are capped with each end). Oxygen atoms and nitrogen atoms are in red and blue respectively, and the long fibril axis is shown. (c, d) show Transmission Electron Microscopy (TEM) images of polyglutamine (Q10) aggregates, which cluster into dense meshworks composed of small, thin fibrils (c) and long, ribbon-like fibrils (d). Figures (c, d) reprinted with permission from Ref. Punihaole et al.[31]. Copyright 2016 American Chemical Society.

in cross-β sheets, so-called steric zippers[35,38], which represent pairs of self-complementary β-sheets that are interdigitated (Fig. 2.4b). Zippers have been suggested to constitute, in the context of full-length polypeptide chains, the structural spine of amyloid fibril[38] (Fig. 2.4c,d).

Experimental studies of protein aggregation typically show a sigmoidal curve of growth kinetics[40–44], where two relatively flat regions are connected by a steep transition zone (Fig. 2.5). The region before the transition zone is known as the lag phase or nucleation period, in which partially folded monomers associate to form primary nuclei. The steep transition zone is often called the growth phase or elongation phase as the nuclei reach the critical size at which point small fibrils emerge and elongate. The final flat region is referred to as the plateau phase and represents a steady state with an equilibrium monomer concentration.

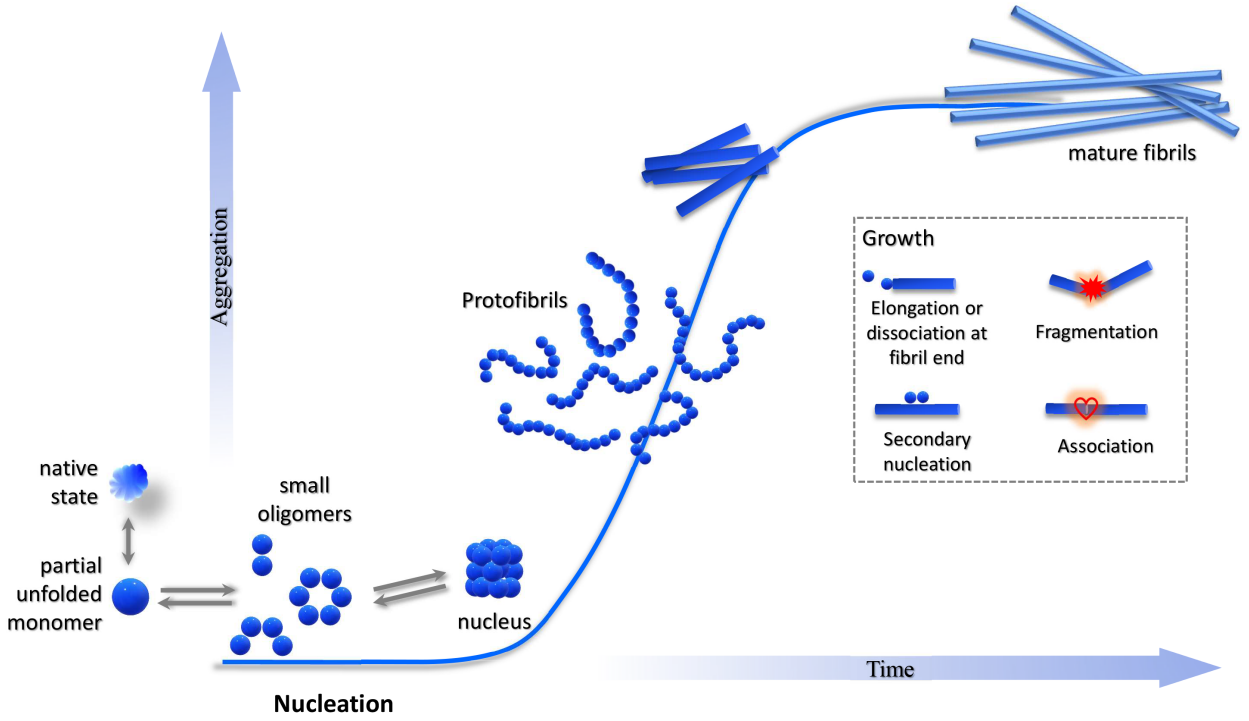

Figure 2.5: The sigmoidal growth profile of fibril nucleation mechanism, including secondary nucleation and fragmentation processes. Adapted form Ref. Morriss-Andrews and Shea[39].

This sigmoidal appearance is a characteristic of nucleated self-assembly reactions[45]. In the growth phase, there are primary growth and secondary growth processes. The former is attributed to fibril-end elongation by a dock-lock mechanism[46–49], in which the monomer first forms an initial contact with the fibril template (dock step), then it rearranges and changes conformation until it fully aligns with the structure (lock step). The latter is associated with a range of processes such as lateral growth, fragmentation, and association[50–53].

Protein aggregation processes are not yet well understood but a few guiding principles have been revealed over the years[54].

- Charged proteins repel each other, thus proteins are more prone to aggregate when their net charge is zero at the isoelectric point pI[55].
- Adding salt can assist in shielding the charges and weakening the repulsion between two proteins that have the net charge of the same sign, therefore it can promote the aggregation process[56].

- Hydrophobic interactions are the major driving forces in sticking proteins to each other. Aggregation is produced when hydrophobic regions of partially folded proteins, which are expected to be buried in the native state, are exposed to the cellular environment[57–59].

In neurodegenerative aggregation diseases, neuronal loss, neuroinflammation, synaptic alterations are some typical features that occur in widely varying parts of the brain[60–62]. Although affected regions of the brains vary among diseases[63], protein misfolding and aggregation are key events in each disorder. Therefore, a potential, comprehensive therapy should focus on the casual protein misfolding events in the disease initiation or target disease-modifying strategies that prevent the formation of protein aggregates. Fortunately, research studies in aggregation diseases over the years have yielded the sweet fruits. In June 2021, the first new drug for Alzheimer's disease for nearly two decades was approved by the US Food and Drug Administration. The approved therapy, which has the molecular name aducanumab, targets the underlying cause of Alzheimer's rather than its symptoms[64–66]. Although many scientists are skeptical about the sufficient evidence of the effectiveness of aducanumab for the disease, it gives hopes to patients and sheds light for other treatments in the future.

2.4 Nucleation

The phenomenon of nucleation is ubiquitously observed in many different systems, from everyday life examples such as Diet Coke and Mentos eruption, CO_2 bubbles in a glass of soda, formation of cloud and snow to science such as nanomaterials[67], polymerization processes[68], protein and mineral crystallization[69–71], and initiation of neurodegenerative diseases[72–77]. Nucleation is the initial step in the formation of a new thermodynamic phase from a high free energy parent phase to an ordered, well-organized structure or pattern with lower free energy. Nucleation can be homogeneous[78,79], in the same type of particles, or heterogeneous[80,81], in the presence of foreign species.

Many theories have been developed to describe the homogeneous nucleation process. They are based on phenomenological, kinetic, and microscopic approaches. In phenomenological theory, the free energy formation of clusters is calculated by using the macroscopic quantities. In contrast, the kinetic theory avoids using the macroscopic surface tension and exploits the molecular interactions instead[82–84]. The microscopic approaches including Monte Carlo simulations and molecular dynamics simulations try to obtain nucleation rates starting from the potential energy of interactions among particles[85–87]. Classical nucleation theory (CNT), based on phenomenological approach, has been the standard theory used to describe homogeneous nucleation for many decades since it successfully captures the qualitative features of the nucleation phenomena and gives reasonable predictions of the nucleation rates.

In CNT, the driving force required for nucleation is referred to as *supersaturation* and is defined as the chemical potential difference between the two different phases (e.g. vapor-liquid or liquid-solid). For crystallization from solution, we have

$$\Delta\mu = \mu_s - \mu_c, \tag{2.1}$$

where μ_s is the chemical potential of a molecule in solution and μ_c is the chemical potential of the molecule in the bulk crystal, k_{B} is the Boltzmann constant, and T is the absolute temperature. When $\Delta\mu > 0$, the solution is said to be supersaturated, which means the nucleation is possible, while $\Delta\mu < 0$ the solution is undersaturated and dissolution may occur.

According to CNT, the work necessary to form a cluster of n monomers is the free energy difference between the initial and final states (the bulk free energy) plus the energy required to form an interface between the nucleus and the solution (the surface energy). Assuming a spherical shape, this can be written as

$$\begin{aligned} \Delta G &= -n\Delta\mu + 4\pi r^2\sigma \\ &= -\frac{4}{3}\pi\frac{r^3}{\nu}\Delta\mu + 4\pi r^2\sigma, \end{aligned} \tag{2.2}$$

where r is the radius of the nucleus, ν is the volume that each molecule occupies in the crystal, and σ is the surface tension.

Fig. 2.6a shows a plot of ΔG as a function of nucleus size r; this represents the energetic barrier ($\Delta G^{\ddagger}$) that needs to be surpassed to achieve the nucleation. The value of r at the free energy maxima ($r^{\ddagger}$) is the critical nucleus size[88,89]. The formation of a small nucleus is energetically unfavorable due to the high entropic cost but when the nucleus reaches the critical size the further growth is favorable due to the enthalpic gain. The critical nucleus size describes the balance between entropic and energetic contributions to the nucleation, as nuclei with smaller size than the critical size will easily dissolve back to the solution whereas larger nuclei can surpass the fluctuations to form stable structures and undergo further growth.

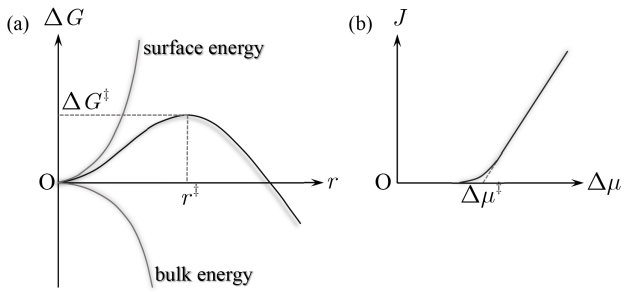

Figure 2.6: (a) Total free energy as a function of nucleus size. (b) Nucleation versus supersaturation (critical supersaturation is shown).

In CNT, the central quantity is the nucleation rate, which describes the number of nuclei formed per unit of time per unit of volume. It can be expressed by an Arrhenius-like equation,

$$J = A \exp\left(\frac{-\Delta G^{\ddagger}}{k_{\mathrm{B}} T}\right), \tag{2.3}$$

where A is the pre-exponential factor and also depends on supersaturation. Fig. 2.6b shows a typical plot of nucleation rate as a function of supersaturation. The nucleation rate is almost zero until a critical supersaturation is obtained, after that the rate increases exponentially. This critical supersaturation defines the metastable zone where the new phase can proceed without nucleation. The CNT was originally developed by the work of Volmer and Weber[90], Becker and Doring[91], and Frankel[92] to describe the condensation of vapor into a liquid, and later extended to other liquid-solid systems by employing the "analogy" such as crystallization from melts and solutions as well as amyloid peptide condensation.

An important assumption in CNT described above is that the nuclei are formed directly

in solution without precursors or intermediates. This single-step nucleation, however, may be prohibited by strong energy barriers required to surmount the interfacial surface tension of the nucleus[93,94]. Moreover, experiments of the crystallization process for several materials, such as proteins[95–97], minerals[98] and colloids[99], as well as computer simulations[93,100–104] have shown that the nucleation process proceeds through a liquid-liquid phase transition to provide oligomeric intermediates before reaching a thermodynamically stable state. In these cases, non-classical nucleation or two-step nucleation suggests that the first step towards a critical nucleus is the formation of a sufficient-sized cluster of solute molecules, followed by reorganization or conformational change of that cluster into ordered structures or patterns. Thorough reviews of the non-classical nucleation theory can be found in Erdemir et al.[105], Karthika et al.[106], Cubillas and Anderson[107].

References

[1] C. Lad, N. H. Williams, and R. Wolfenden, Proceedings of the National Academy of Sciences **100**, 5607 (2003).

[2] N. J. Zondlo, Accounts of chemical research **46**, 1039 (2013).

[3] L. Biedermannova, K. E. Riley, K. Berka, P. Hobza, and J. Vondrasek, Physical Chemistry Chemical Physics **10**, 6350 (2008).

[4] E. W. Martin and A. S. Holehouse, Emerging Topics in Life Sciences (2020).

[5] R. Milo and R. Phillips, *Cell biology by the numbers* (Garland Science, 2015).

[6] T. Bereau, *Unconstrained structure formation in coarse-grained protein simulations* (2011), dissertation.

[7] Y. Marechal, *The hydrogen bond and the water molecule: The physics and chemistry of water, aqueous and bio-media* (Elsevier, 2006).

[8] C. Tanford, *The hydrophobic effect: formation of micelles and biological membranes 2d ed* (J. Wiley., 1980).

[9] M. Karplus, Folding and design **2**, S69 (1997).

[10] C. Levinthal, Journal de chimie physique **65**, 44 (1968).

[11] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Structure, Function, and Bioinformatics **21**, 167 (1995).

[12] K. A. Dill and H. S. Chan, Nature structural biology **4**, 10 (1997).

[13] C. M. Dobson and P. J. Hore, nature structural biology **5**, 504 (1998).

[14] A. R. Dinner, A. Šali, L. J. Smith, C. M. Dobson, and M. Karplus, Trends in biochemical sciences **25**, 331 (2000).

[15] K. A. Dill, Biochemistry **24**, 1501 (1985).

[16] J. D. Bryngelson and P. G. Wolynes, Proceedings of the National Academy of sciences **84**, 7524 (1987).

[17] P. E. Leopold, M. Montal, and J. N. Onuchic, Proceedings of the National Academy of Sciences **89**, 8721 (1992).

[18] J. N. Onuchic and P. G. Wolynes, Current opinion in structural biology **14**, 70 (2004).

[19] C. J. Camacho and D. Thirumalai, Proceedings of the National Academy of Sciences **90**, 6369 (1993).

[20] H. S. Chan and K. A. Dill, The Journal of chemical physics **100**, 9238 (1994).

[21] H. S. Chan and K. A. Dill, The Journal of chemical physics **99**, 2116 (1993).

[22] A. L. Fink, Physiological reviews **79**, 425 (1999).

[23] F. U. Hartl, A. Bracher, and M. Hayer-Hartl, Nature **475**, 324 (2011).

[24] S. F. Göthel and M. Marahiel, Cellular and molecular life sciences CMLS **55**, 423 (1999).

[25] R. B. Freedman, T. R. Hirst, and M. F. Tuite, Trends in biochemical sciences **19**, 331 (1994).

[26] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).

[27] R. W. Carrell and D. A. Lomas, The Lancet **350**, 134 (1997).

[28] C. M. Dobson, Trends in biochemical sciences **24**, 329 (1999).

[29] C. Soto, FEBS letters **498**, 204 (2001).

[30] T. P. Knowles and M. J. Buehler, Nature nanotechnology **6**, 469 (2011).

[31] D. Punihaole, R. J. Workman, Z. Hong, J. D. Madura, and S. A. Asher, The Journal of Physical Chemistry B **120**, 3012 (2016).

[32] M. J. Schleiden, Annalen der Physik **119**, 391 (1838).

[33] R. Virchow, Archiv für pathologische Anatomie und Physiologie und für klinische Medicin **6**, 416 (1854).

[34] M. Sunde, L. C. Serpell, M. Bartlam, P. E. Fraser, M. B. Pepys, and C. C. Blake, Journal of molecular biology **273**, 729 (1997).

[35] R. Nelson, M. R. Sawaya, M. Balbirnie, A. Ø. Madsen, C. Riekel, R. Grothe, and D. Eisenberg, Nature **435**, 773 (2005).

[36] T. Lührs, C. Ritter, M. Adrian, D. Riek-Loher, B. Bohrmann, H. Döbeli, D. Schubert, and R. Riek, Proceedings of the National Academy of Sciences **102**, 17342 (2005).

[37] E. Eanes and G. Glenner, Journal of Histochemistry & Cytochemistry **16**, 673 (1968).

[38] M. R. Sawaya, S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. Wiltzius, H. T. McFarlane, et al., Nature **447**, 453 (2007).

[39] A. Morriss-Andrews and J.-E. Shea, Annual review of physical chemistry **66**, 643 (2015).

[40] F. Oosawa, S. Asakura, K. Hotta, N. Imai, and T. Ooi, Journal of Polymer Science **37**, 323 (1959).

[41] F. Oosawa and M. Kasai, Journal of molecular biology **4**, 10 (1962).

[42] F. A. Ferrone, J. Hofrichter, and W. A. Eaton, Journal of molecular biology **183**, 591 (1985).

[43] F. A. Ferrone, J. Hofrichter, and W. A. Eaton, Journal of molecular biology **183**, 611 (1985).

[44] W.-F. Xue, S. W. Homans, and S. E. Radford, Proceedings of the National Academy of Sciences **105**, 8926 (2008).

[45] P. Arosio, T. P. Knowles, and S. Linse, Physical Chemistry Chemical Physics **17**, 7606 (2015).

[46] W. P. Esler, E. R. Stimson, J. M. Jennings, H. V. Vinters, J. R. Ghilardi, J. P. Lee, P. W. Mantyh, and J. E. Maggio, Biochemistry **39**, 6288 (2000).

[47] P. H. Nguyen, M. S. Li, G. Stock, J. E. Straub, and D. Thirumalai, Proceedings of the National Academy of Sciences **104**, 111 (2007).

[48] M. Schor, J. Vreede, and P. G. Bolhuis, Biophysical journal **103**, 1296 (2012).

[49] M. Schor, A. S. Mey, F. Noe, and C. E. MacPhee, The journal of physical chemistry letters **6**, 1076 (2015).

[50] D. Thirumalai, D. Klimov, and R. Dima, Current opinion in structural biology **13**, 146 (2003).

[51] J. E. Straub and D. Thirumalai, Annual review of physical chemistry **62**, 437 (2011).

[52] T. P. Knowles, C. A. Waudby, G. L. Devlin, S. I. Cohen, A. Aguzzi, M. Vendruscolo, E. M. Terentjev, M. E. Welland, and C. M. Dobson, Science **326**, 1533 (2009).

[53] T. C. Michaels and T. P. Knowles, The Journal of chemical physics **140**, 214904 (2014).

[54] I. Bahar, R. L. Jernigan, and K. A. Dill, *Protein actions: Principles and modeling* (Garland Science, 2017).

[55] K. L. Shaw, G. R. Grimsley, G. I. Yakovlev, A. A. Makarov, and C. N. Pace, Protein Science **10**, 1206 (2001).

[56] E. Ruckenstein and I. L. Shulgin, Advances in colloid and interface science **123**, 97 (2006).

[57] D. March, V. Bianco, and G. Franzese, Polymers **13**, 156 (2021).

[58] K. V. Biza, K. C. Nastou, P. L. Tsiolaki, C. V. Mastrokalou, S. J. Hamodrakas, and V. A. Iconomidou, PloS one **12**, e0173163 (2017).

[59] S. K. Chaturvedi, M. K. Siddiqi, P. Alam, and R. H. Khan, Process Biochemistry **51**, 1183 (2016).

[60] J. B. Martin, New England Journal of Medicine **340**, 1970 (1999).

[61] M. F. Tuite and R. Melki, *Protein misfolding and aggregation in ageing and disease: molecular processes and therapeutic perspectives* (2007).

[62] C. Soto and L. D. Estrada, Archives of neurology **65**, 184 (2008).

[63] T. N. Shamsi, T. Athar, R. Parveen, and S. Fatima, International journal of biological macromolecules **105**, 993 (2017).

[64] H. Fillit and A. Green, Nature Reviews Neurology pp. 1–2 (2021).

[65] A. Mullard, Nature (2021).

[66] J. Sevigny, P. Chiao, T. Bussière, P. H. Weinreb, L. Williams, M. Maier, R. Dunstan, S. Salloway, T. Chen, Y. Ling, et al., Nature **537**, 50 (2016).

[67] E. Mendez-Villuendas and R. K. Bowles, Physical review letters **98**, 185503 (2007).

[68] R. J. Young and P. A. Lovell, *Introduction to polymers* (CRC press, 2011).

[69] O. Galkin and P. G. Vekilov, Journal of Crystal Growth **232**, 63 (2001).

[70] J. M. Garcıa-Ruiz, Journal of Structural Biology **142**, 22 (2003).

[71] A. E. Van Driessche, M. Kellermeier, L. G. Benning, and D. Gebauer, *New perspectives on mineral nucleation and growth: from solution precursors to solid materials* (Springer, 2016).

[72] D. Kashchiev and S. Auer, The Journal of chemical physics **132**, 06B602 (2010).

[73] S. Auer, P. Ricchiuto, and D. Kashchiev, Journal of molecular biology **422**, 723 (2012).

[74] R. D. Hills Jr and C. L. Brooks III, Journal of molecular biology **368**, 894 (2007).

[75] A. R. Hurshman, J. T. White, E. T. Powers, and J. W. Kelly, Biochemistry **43**, 7365 (2004).

[76] F. Massi and J. E. Straub, Proteins: Structure, Function, and Bioinformatics **42**, 217 (2001).

[77] E. T. Powers and D. L. Powers, Biophysical journal **91**, 122 (2006).

[78] P. Wagner and R. Strey, The Journal of chemical physics **80**, 5266 (1984).

[79] C.-H. Hung, M. J. Krasnopoler, and J. L. Katz, The Journal of chemical physics **90**, 1856 (1989).

[80] D. Turnbull, The Journal of Chemical Physics **18**, 198 (1950).

[81] X. Liu, The Journal of Chemical Physics **112**, 9949 (2000).

[82] J. L. Katz and M. D. Donohue, Advances in Chemical Physics **40**, 137 (1979).

[83] S. L. Girshick and C.-P. Chiu, The journal of chemical physics **93**, 1273 (1990).

[84] E. Ruckenstein and Y. Djikaev, Advances in colloid and interface science **118**, 51 (2005).

[85] G. K. Schenter, S. M. Kathmann, and B. C. Garrett, Physical review letters **82**, 3484 (1999).

[86] H.-M. Shim, J.-K. Kim, H.-S. Kim, and K.-K. Koo, Crystal growth & design **14**, 5897 (2014).

[87] M. Sekine, K. Yasuoka, T. Kinjo, and M. Matsumoto, Fluid dynamics research **40**, 597 (2008).

[88] D. Kashchiev, *Nucleation* (Elsevier, 2000).

[89] J. W. Mullin, *Crystallization* (Elsevier, 2001).

[90] M. Volmer and A. Weber, Z. phys. chem **119**, 277 (1926).

[91] R. Becker and W. Döring, Annalen der Physik **416**, 719 (1935).

[92] J. Frenkel, The Journal of Chemical Physics **7**, 538 (1939).

[93] P. R. ten Wolde and D. Frenkel, Science **277**, 1975 (1997).

[94] P. G. Vekilov, Nanoscale **2**, 2346 (2010).

[95] A. Sauter, F. Roosen-Runge, F. Zhang, G. Lotze, R. M. Jacobs, and F. Schreiber, Journal of the American Chemical Society **137**, 1485 (2015).

[96] A. Sauter, F. Roosen-Runge, F. Zhang, G. Lotze, A. Feoktystov, R. M. Jacobs, and F. Schreiber, Faraday discussions **179**, 41 (2015).

[97] Y. Liu, X. Wang, and C. B. Ching, Crystal growth & design **10**, 548 (2010).

[98] E. M. Pouget, P. H. Bomans, J. A. Goos, P. M. Frederik, N. A. Sommerdijk, et al., Science **323**, 1455 (2009).

[99] J. Savage and A. Dinsmore, Physical review letters **102**, 198302 (2009).

[100] G. Nicolis and C. Nicolis, Physica A: Statistical Mechanics and its Applications **323**, 139 (2003).

[101] A. Gavezzotti, Chemistry–A European Journal **5**, 567 (1999).

[102] K. G. Soga, J. R. Melrose, and R. C. Ball, The Journal of chemical physics **110**, 2280 (1999).

[103] J. D. Shore, D. Perchak, and Y. Shnidman, The Journal of Chemical Physics **113**, 6276 (2000).

[104] A. Šarić, Y. C. Chebaro, T. P. Knowles, and D. Frenkel, Proceedings of the National Academy of Sciences **111**, 17869 (2014).

[105] D. Erdemir, A. Y. Lee, and A. S. Myerson, Accounts of chemical research **42**, 621 (2009).

[106] S. Karthika, T. Radhakrishnan, and P. Kalaichelvi, Crystal Growth & Design **16**, 6663 (2016).

[107] P. Cubillas and M. W. Anderson, Zeolites and catalysis: synthesis, reactions and applications pp. 1–55 (2010).

Chapter 3

Computational Methods

Computer simulations provide a unified picture of the microscopic length and time scales and the macroscopic properties of biological systems measured in the laboratory. For example, given an interatomic interaction model from an experiment-based guess, simulations can make "exact" predictions of bulk and detailed properties subject to limitations. We can test and fine-tune the model by comparing various properties obtained from the simulations with experimental results. Simulations can also act as a useful toolbox to test against analytic theories. Theoretical predictions usually rely on several assumptions that require the input parameters which may be inaccessible through experiments. In these situations, we can perform simulations on the computer to test the validity of the theory in conditions that are difficult or impossible in the laboratory (for example, extremes of temperature or pressure). However, powerful simulations will not magically provide valid results, one should use this tool wisely and be aware of some potential pitfalls. The two most common simulation techniques are Classical Molecular Dynamics (MD) and Monte Carlo (MC) Simulations. In MD, atomic motion is simulated by solving Newton's equations of motion simultaneously for all atoms in the system. MD simulations can be used for both equilibrium and transport properties of the system. MC simulations, on the other hand, do not rely on the equations of motion. They mostly focus on the equilibrium states and thus do not provide direct information about the dynamics of the system. In addition to equilibrium properties, MC

idea can also be exploited to deal with dynamical properties, this method is called kinetic Monte Carlo (kMC). KMC takes the known transition rates among states as input parameters to simulate the time evolution of natural processes. This chapter is devoted to brief introductions of MD, MC simulations and kinetic MC simulations used in the dissertation.

3.1 Classical Molecular Dynamics Simulations

The basic idea behind an MD simulation is simple. Given the structure of a biomolecular system (the relative coordinates of the constituent atoms), by calculating the net force exerted on each atom by all of other atoms one can use the Newtonian laws to predict the spatial position as a function of time.

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i, \qquad F_i = -\frac{\partial U}{\partial r_i}, \tag{3.1}$$

where m_i and r_i are the mass and position of the i^{th} atom of the system, respectively. For this purpose, we need to calculate the net force F_i acting on the atoms. In classical MD simulations where electrons are not treated explicitly, the forces rely on an empirically derived force field $U(r^N)$, where r^N represents the coordinates of N atoms. This potential energy is a collection of mathematical functions and parameters that describe the interactions among atoms in the system. All common force fields can be divided into bonded interactions and nonbonded interactions (Fig. 3.1). The mathematical form[1,2] can be written as

$$\begin{aligned} U(r^N) &= \sum_{\text{bonds}} \frac{k_\ell}{2}(\ell_i - \ell_{i,0})^2 + \sum_{\text{angles}} \frac{k_\theta}{2}(\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) \\ &+ \sum_{i=1}^{N}\sum_{j\neq i}^{N} \left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right). \end{aligned} \tag{3.2}$$

In Eq. 3.2, the first three terms deal with the specific internal degrees of freedom within the molecules, which are bond stretching, angle bending and bond rotating. The first term is a harmonic potential between bonded atoms that gives the contribution to the energy

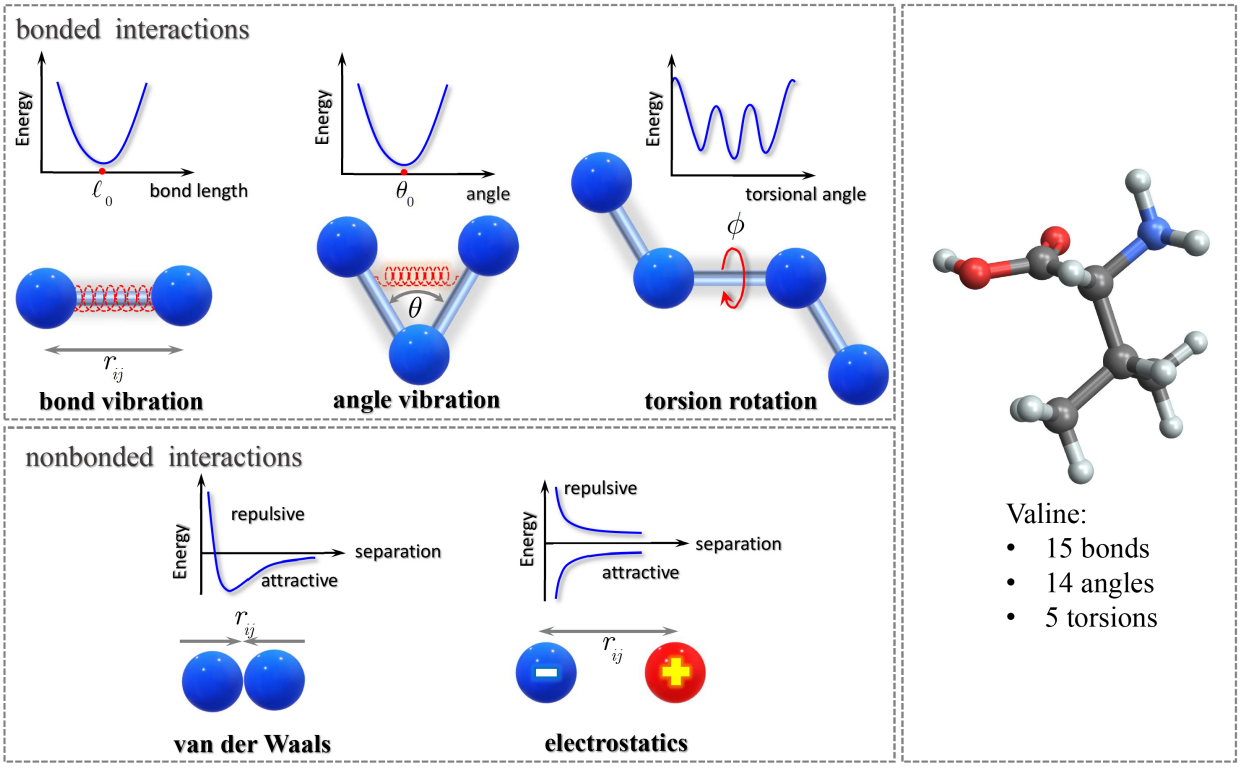

Figure 3.1: Schematic representation for bonded interactions including covalent bond-stretching, angle-bending and torsion-rotating, and nonbonded interactions which are based on neighbor-lists and consist of long-range van der Waals and Coulomb interactions. Example of interactions in valine, a small amino acid residue.

when the bond length ℓ_i deviates from the equilibrium value $\ell_{i,0}$. Similarly, the second term is a harmonic potential in the valence angles of the molecules. The third term is a torsional potential that models how the energy changes as a bond rotates. The last two terms represent nonbonded interactions. The fourth contribution is a Lennard-Jones potential representing the van der Waals interactions, and the last term is the Coulomb electrostatic potential. Note that the first four terms deal with mainly short-ranged interactions, the last one refers to long-ranged interaction. Parameters for the potential are usually obtained by fitting to either *ab initio* calculations or experimental data.

Significant progress has recently been made in development of force fields that allow the reliable modeling of biomolecules using MD simulations. Choosing a proper force field is the vital step when carrying out MD simulations due to possible biases different force fields have toward certain types of secondary structure[3–6]. For example, one force field may be fully

validated with specific experimental data, but that is typically not possible for validation against different structures and other physical properties from a large number of independent and fully validated experiments because experiments have their own error sources. While it is difficult to obtain a completely transferable force field, refinement and modification of the existing ones have led to significant improvements in agreement with experimental data[7–10].

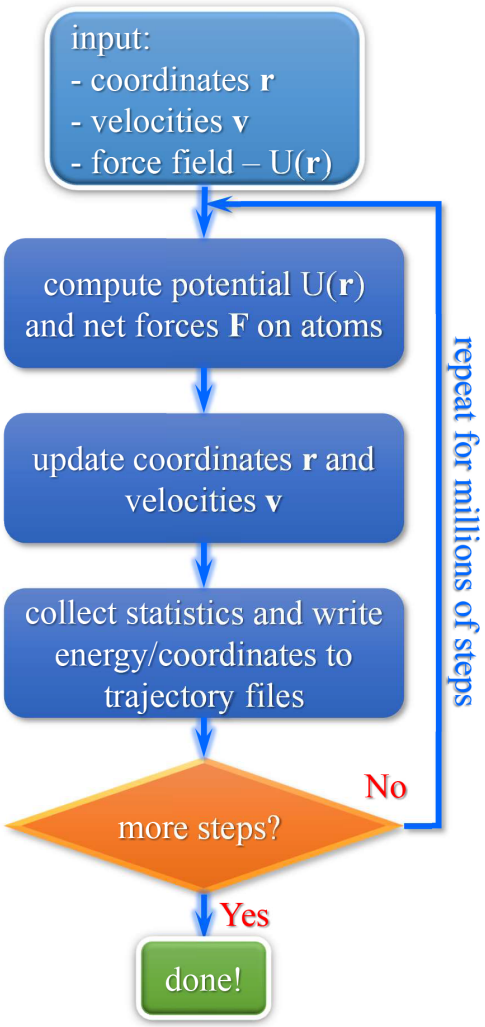

Figure 3.2: A typical procedure of an MD simulation (adapted from Ref. Lindahl[1]).

In MD simulations, solving Newton's equations of motions analytically is impossible task due to the complex form of the potential energy function and the large number of atoms in the system. Therefore, various numerical integration algorithms have been developed to solve the equations of motion[11,12]. After each integration, the updated coordinates are then used to evaluate the potential energy again (as shown in the flow chart Fig. 3.2). Repeatedly updating the coordinates and velocities of atoms step-by-step through the time results in a three dimensional trajectory describing the atomistic configuration of the system during the simulated time interval. To ensure numerical stability, the time steps in MD simulations must be shorter than the system's fastest motions, for example bond vibrations in proteins' motion take 1-2 femtoseconds (10^{-15} s). Otherwise, it violates the "small-step" assumption in Taylor expansions of positions[13], which causes nonphysical energies and accelerations[14].

In most situations, we need to simulate bulk systems such as a solid crystal or protein solution. The presence of "walls" in a simulation box would cause a profound effect on the properties of the system since there are unwanted interactions with the boundary surfaces. Periodic boundary conditions

(PBCs) are then used to minimize edge effects in a finite simulation box and allow us to simulate bulk systems. In PBCs, particles have copies of themselves inside every periodic repetition of the simulation box. In this way, the simulation box and its images will occupy the whole space and thus mimic the bulk phase. Note that PBCs are simply an approximation to the bulk behavior and they are not effective to simulate an infinitely sized system. In protein simulations, the box size needs to be sufficiently large to avoid interactions between protein and its images. The cubic box is mostly widely used, however other simulation cell geometries are possible: rectangular cuboid, hexagonal prism, truncated octahedron, etc. All of these geometries will regularly tile space and thus can serve to replicate the infinite number of periodic images.

At the end of an MD simulation, we obtain a trajectory, which is a time series of the system's coordinates. There may be profound fluctuations in the trajectory during the simulated time interval. Proper calculations of averages and variances of conformational properties are needed to identify important states of motion. To ensure sufficient and accurate statistics, one should either generate multiple trajectories or long individual ones. A useful measure of fluctuations for the entire group of atoms of interest, for example the whole protein molecule, is the root-mean-square deviation (RMSD) in atomic coordinates as a function of time with respect to the initial state,

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\vec{r}_i(t) - \vec{r}_i(0)]^2} \tag{3.3}$$

The convergence of RMSD during a simulation is determined by seeking a plateau in RMSD(t). Another useful quantity is the RMSD between two conformations A and B,

$$\mathrm{RMSD}(A, B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\vec{r}_i(A) - \vec{r}_i(B)]^2} \tag{3.4}$$

where $\vec{r}_i(A)$ and $\vec{r}_i(B)$ are the positions of atom i in the two respective conformations, after their optimal superposition.

Nowadays, improvements in computing hardware and in algorithms have allowed MD simulations access to biologically meaningful timescales. Current software for MD simulations, such as NAMD[15], GROMACS[16] and OPENMM[17], utilizing efficient algorithms on GPUs are highly efficient. Indeed, performing MD simulations is now relatively straightforward because simulation packages are usually open source and the computational resources are increasingly accessible. Researchers now just focus on determining which questions can be addressed by simulations, designing the systems, running simulations and interpreting the results.

3.2 Monte Carlo Methods

Unlike MD, Monte Carlo methods are stochastic in nature - the time progression of the atomic positions proceeds randomly and is not predictable given a set of initial conditions. MC is an equilibrium method that aims for low free energy states and rigorously generates correct thermodynamic properties within the constructive design. MC and MD approaches are complementary. MD is a natural choice if kinetic properties are of interest. Otherwise, MC methods have some attractive features: (1) MC methods are not limited to "small-time" step approximations of equations of motion, (2) MC methods require only energy calculations, which can handle continuous and discrete intermolecular potentials efficiently, (3) MC methods can explore broadly conformational space by offering flexibility in choosing random moves, (4) MC methods easily deal with different thermodynamics ensembles.

As we know, Monte Carlo simulations use random moves to explore the configuration space to find out some information about the space. In 1953, Nicholas Metropolis and coworkers[18] proposed a new sampling procedure which incorporates a temperature of the system to calculate the Boltzmann average of a property of the system. This MC method is called Metropolis Monte Carlo (MMC) simulation. For more detailed descriptions of MMC simulation the textbook by Frenkel and Smit[19] is an excellent resource. Here, to keep the math and the notations simple, let us consider a one dimensional free energy landscape

$\mathcal{U}(x)$[14]. The probability distribution is written as

$$p(x) = \mathcal{Z}^{-1} e^{-\beta \mathcal{U}(x)}, \tag{3.5}$$

where

$$\mathcal{Z} = \int e^{-\beta \mathcal{U}(x)} dx \tag{3.6}$$

is the partition function and $\beta = 1/k_{\mathrm{B}}T$. The average value of an observable $A(x)$ is given by

$$\langle A \rangle = \int A(x) p(x) dx = \frac{\int A(x) e^{-\beta \mathcal{U}(x)} dx}{\int e^{-\beta \mathcal{U}(x)} dx} \tag{3.7}$$

If we need to estimate this average based on a relatively small number of samples compared to the large configuration space of x, we can define

$$\langle A \rangle_{\text{estimate}} = \sum_{i=1}^{N} A(x_i) p(x_i) \approx \langle A \rangle \tag{3.8}$$

To compute $\langle A \rangle$ over a probability distribution $p(x)$, uniform sampling will give an estimate of average according to Eq. 3.8. However, it is not usually useful and efficient, because we easily miss the most populated region of the distribution (for example region between x_2 and x_3 in Fig. 3.3).

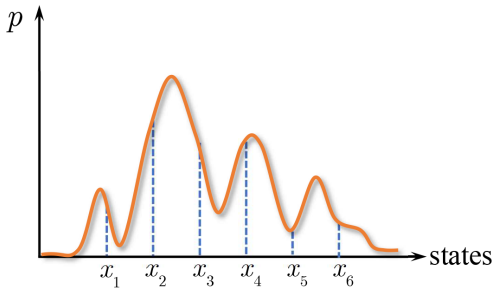

Figure 3.3: Uniformed sampling can waste time in the regions of low population instead of focusing on the important ones.

MMC approach is to focus on sampling near states that seemingly have the highest probabilities, which avoids wasting much time in poorly populated regions. The generation of configurations according to a distribution is called *importance sampling*. The challenge question is how to generate configurations randomly according to distribution $p(x)$. It turns out that we can importance-sample configurations using a statistical construct called a *Markov chain* of states. A Markov chain describes

a stochastic process in which the state of a system changes randomly with time and has no memory of previous states. At each step in time, the system can move randomly to another state.

In the context of Monte Carlo simulation, let us consider the basic procedure to generate a Markov chain. Start at step i, where the system is in configuration x_i. This configuration is randomly perturbed to generate a new atomic configuration. For example, the particle is randomly picked and displaced its coordinates by small random amounts. In general, these perturbations are termed MC *moves*. The new configuration is considered a proposed new state of the system x_j with a trial selection probability or rate α_{ij}. The configuration at the next step is then either the proposed configuration if accepted with probability $\mathcal{P}_{ij}^{acc}$ or the original configuration if rejected with the probability $1-\mathcal{P}_{ij}^{acc}$. The acceptance or rejection of the proposed moves is performed in such a way that configurations are generated according to $p(x)$ in the long-time limit. The overall transition rate from state i to state j is thus given by the transition matrix $w_{ij} = \alpha_{ij}\mathcal{P}_{ij}^{acc}$. The process is repeated over and over again to generate a trajectory of configurations. In this way, Monte Carlo moves can be propagated in time according to pre-specified configurational probabilities.

The key point in this procedure is to decide how to accept or reject proposed configurations in our simulations. We need to choose the acceptance criterion in such a way that our long-time trajectory correctly generates configurations with the desired probability distribution $p(x)$. We can do this if the overall transition probabilities w_{ij} satisfies the *detailed balance* condition

$$w_{ij}\mathcal{P}_i = w_{ji}\mathcal{P}_j, \tag{3.9}$$

which implies that the desired distribution $p(x)$ is a stationary state. This condition applies a constraint to the transition and state probabilities for every pair of states. If we impose the detailed balance condition, the acceptance probability should obey

$$\frac{\mathcal{P}_{ij}^{acc}}{\mathcal{P}_{ji}^{acc}} = \frac{\alpha_{ij}\mathcal{P}_i}{\alpha_{ji}\mathcal{P}_j} \tag{3.10}$$

This equation now gives us a starting point for correctly performing our Monte Carlo simulation. Several possible forms for the acceptance probabilities $\mathcal{P}_{ij}^{acc}$ satisfy the detailed balance condition in Eq. 3.9. The simplest and most commonly used corresponds to the Metropolis criterion[20]

$$\mathcal{P}_{ij}^{acc} = \min\left(1, \frac{\mathcal{P}_j}{\mathcal{P}_i}\right), \tag{3.11}$$

where the symmetric trial configuration selection rates are used, $\alpha_{ij} = \alpha_{ji}$. So-called symmetric Monte Carlo moves have move proposal probabilities that are equal in the forward and reverse directions. Using Eq. 3.5, the acceptance probability $\mathcal{P}_{ij}^{acc}$ for a transition from state i to state j can be written as

$$\mathcal{P}_{ij}^{acc} = \min\left(1, e^{-\beta(U_j - U_i)}\right). \tag{3.12}$$

The min function is incorporated into this criterion. If $U_j < U_i$, the acceptance probability is always one. Otherwise, it is less than one. Thus, this move specifies that we should always move downhill in energy if we can, an aspect which helps reach equilibration faster in Monte Carlo simulations than alternative criteria. The rule above must be applied equally to the reverse move.

To summarize, a general approach to any MC simulation involves the following steps:

- Choose the system potential energy function, $\mathcal{U}(\mathbf{r}^N)$.
- Choose the statistical-mechanical ensemble of interest. This uniquely specifies the probabilities $\mathcal{P}_i$ with which each microstate i should be sampled. In the canonical ensemble, $\mathcal{P}_i \propto e^{-\beta\mathcal{U}_i}$.
- Choose the MC move set, which is a collection of rules for how the system is allowed to transition from state i to state j. These moves uniquely specify the move proposal probabilities α_{ij}. For symmetric moves, $\alpha_{ij} = \alpha_{ji}$.
- Determine the appropriate acceptance criterion. Typically we use the Metropolis cri-

terion. The acceptance criterion then follows directly from the relation,

$$\mathcal{P}_{ij}^{acc} = \min\left(1, \frac{\alpha_{ji}\mathcal{P}_{ji}}{\alpha_{ij}\mathcal{P}_{ij}}\right). \tag{3.13}$$

- Carry out the simulation using the determined acceptance criterion. Equilibration must first be achieved by propagating the system for several relaxation times.

- Ensemble property averages are computed from trajectory averages. The average value of any configurational property in the ensemble of interest then follows from a simple average over the "time"-progression of the production phase of the simulation

$$\langle A \rangle = \frac{1}{N_{\text{tot}}} \sum_{i=1}^{N_{\text{tot}}} A(\mathbf{r}_i^N). \tag{3.14}$$

3.3 Kinetic Monte Carlo

Although MC simulations are efficiently used to study equilibrium properties of physical systems, non-equilibrium or "dynamic" MC simulations are also of interest. Researchers began to study MC algorithms for evolving systems from state to state in the 1960s, but the terminology kinetic Monte Carlo (kMC) was settled in the 1990s. Note that the kMC method is fundamentally the same as the dynamic MC or Gillespie algorithm. As discussed in Section 3.1, MD simulation is the premier tool to simulate dynamical evolution of atomic systems by propagating Newton's equations of motion forward in time. However, the "time-step" problem limits the length of the MD trajectory and the total simulation time to the timescales of microseconds to milliseconds due to the fact that accurate integration requires time steps short enough (10^{-15} seconds) to resolve the atomic vibrations. KMC attempts to conquer this limitation by exploiting the fact that long-time dynamics of the system generally consists of diffusive jumps from one state to another. If these transitions are treated directly, kMC can be used to efficiently model a wide range of dynamical processes at vastly longer timescales.

For many atomic systems, the dynamics can be characterized as a sequence of infrequent transitions from one state (or potential basin) to another. Once a system is caught in a certain basin, it stays there for a long time compared to the time of one vibrational period and it loses the memory of the previous basin where it came from. A successful transition from state i to state j is characterized by a rate constant k_{ij}, which depends only on the shape of the potential basin i and is independent of history. This transition rate satisfies the property of a Markov chain. Since the transition out of state i only depends on the rate constants, we can propagate the system efficiently and correctly from state to state by designing a simple stochastic procedure. If we can determine these transition rates for every potential basin in the system, this state-to-state trajectory will be indistinguishable from a vibrational trajectory generated from an MD simulation.

Given all the rate constants for escape from one state, the probability of the system stays in this state (or the survival probability) can be written as

$$p_{survival}(t) = \exp(-k_{tot}t), \tag{3.15}$$

where k_{tot} is the total of the rate constants. This survival probability can be used to obtain the probability distribution $p(t)$ for the time of first escape from the basin

$$\int_0^{t'} p(t)dt = 1 - p_{survival}(t') \tag{3.16}$$

Taking the time derivative of the right hand side of Eq. 3.16 gives the probability distribution for of the first escape (also known as first-passage-time distribution in kMC procedure),

$$p(t) = k_{tot} \exp(-k_{tot}t). \tag{3.17}$$

The average time for escape τ is

$$\tau = \int_0^{\infty} tp(t)dt = \frac{1}{k_{tot}}. \tag{3.18}$$

The total rate constant in Eq. 3.18 is the sum over the rates of all possible pathways to escape out of the basin,

$$k_{tot} = \sum_j k_{ij}. \tag{3.19}$$

For each pathway, we also have the exponential first-escape-time distribution

$$p_{ij}(t) = k_{ij} \exp(-k_{ij}t), \tag{3.20}$$

although only one of the pathways can be the first to drive the system out of the basin potential. Given the above equations, we are ready to present the kMC algorithm.

The algorithm can briefly describes in two steps[21,22]: The first step is to choose the transition among all possible ones, the second step is to determine the time it takes for this transition. At each iteration, two random numbers in the interval [0, 1], R_1 and R_2, are generated and used to determine which transition will occur and the amount of time required. Given the rate $k_1, k_2, ..., k_n$ for all possible transitions from the current state and the sum of these rates, k_{tot}, transition $i+1$ is selected when

$$\frac{\sum_{j=1}^{i-1} k_j}{k_{\text{tot}}} < R_1 < \frac{\sum_{j=1}^{i} k_j}{k_{\text{tot}}} \tag{3.21}$$

The second random number is set equal to the cumulative distribution function

$$R_2 = \int_0^t P(t')dt' \tag{3.22}$$

to determine the time that elapses before transition $i+1$ occurs. Given that the probability function follows the single exponential distribution, the resulting explicit formula for this time is

$$t = -\frac{1}{k_{\text{tot}}} \ln(1 - R_2) \tag{3.23}$$

The chosen state and elapsed time are appended to the trajectory, at which point the new set of accessible states is determined.

References

[1] E. R. Lindahl, in *Molecular modeling of proteins* (Springer, 2008), pp. 3–23.

[2] A. D. MacKerell Jr et al., Computational biochemistry and biophysics pp. 7–38 (2001).

[3] R. B. Best and J. Mittal, Proteins: Structure, Function, and Bioinformatics **79**, 1318 (2011).

[4] R. B. Best, N.-V. Buchete, and G. Hummer, Biophysical journal **95**, L07 (2008).

[5] P. L. Freddolino, S. Park, B. Roux, and K. Schulten, Biophysical journal **96**, 3772 (2009).

[6] J. Mittal and R. B. Best, Biophysical journal **99**, L26 (2010).

[7] E. J. Sorin and V. S. Pande, Biophysical journal **88**, 2472 (2005).

[8] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, et al., Journal of computational chemistry **24**, 1999 (2003).

[9] A. D. MacKerell Jr, Journal of computational chemistry **25**, 1584 (2004).

[10] R. B. Best and G. Hummer, The journal of physical chemistry B **113**, 9004 (2009).

[11] L. Verlet, Physical review **159**, 98 (1967).

[12] R. W. Hockney, S. Goel, and J. Eastwood, Journal of Computational Physics **14**, 148 (1974).

[13] R. Zhou, *Molecular modeling at the atomic scale: methods and applications in quantitative biology* (CRC Press, 2014).

[14] I. Bahar, R. L. Jernigan, and K. A. Dill, *Protein actions: Principles and modeling* (Garland Science, 2017).

[15] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, Journal of computational chemistry **26**, 1781 (2005).

[16] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1**, 19 (2015).

[17] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, et al., PLoS computational biology **13**, e1005659 (2017).

[18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, The journal of chemical physics **21**, 1087 (1953).

[19] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, vol. 1 (Elsevier, 2001).

[20] D. P. Landau and K. Binder, *A guide to Monte Carlo simulations in statistical physics* (Cambridge university press, 2014).

[21] Z. Jia, A. Beugelsdijk, J. Chen, and J. D. Schmit, The Journal of Physical Chemistry B **121**, 1576 (2017).

[22] A. F. Voter, in *Radiation effects in solids* (Springer, 2007), pp. 1–23.

Chapter 4

Catalyst-like role of impurities in speeding layer-by-layer growth

Molecular self-assembly is usually done at low supersaturation, leading to low rates of growth, in order to allow time for binding mistakes to anneal. However, such conditions can lead to prohibitively long assembly times where growth proceeds by the slow nucleation of successive layers. Here we use a lattice model of molecular self-assembly to show that growth in this regime can be sped up by impurities, which lower the free-energy cost of layer nucleation. Under certain conditions impurities behave almost as a catalyst in that they are present at high concentration at the surface of the assembling structure, but at low concentration in the bulk of the assembled structure. Extrapolation of our numerics using simple analytic arguments suggests that this mechanism can reduce growth times by orders of magnitude in parameter regimes applicable to molecular systems.

4.1 Introduction

The difficulty of achieving reliable self-assembly is one of controlling timescales[1–5]. While it is relatively easy to design a system in which the desired product is the thermodynamic ground state, it is more difficult to ensure that relaxation to equilibrium happens on observ-

able timescales. If a structure grows more rapidly than its component pieces can sample their positional and conformational degrees of freedom then these components become trapped in non-optimal states. This is the case for simple components, such as colloids, and complex components, such as biomolecules[6–9]. It is useful to arrange for the free-energy difference between the desired structure and the starting solution to be small, so that structures grow slowly enough that their constituent particles have time to relax to their preferred configurations[10–17]. A small free-energy difference can be achieved under conditions of small supercooling or low supersaturation. However, while such conditions help to avoid trapped states composed of improperly bound molecules, they exacerbate another kinetic trap, the long induction time associated with nucleation[18–21]. This kinetic trap can also impair growth when growth occurs in a layer-by-layer fashion, because nucleation is the rate-determining step for each stage of growth.

We use computer simulations of growing three-dimensional (3D) lattice-based structures to show that impurity particles can dramatically speed up layer-by-layer growth at low supersaturation, with little effect on the purity of the grown structure. Impurities are generally regarded as problematic, because they have the potential to arrest growth by "poisoning" the growth front[22,23]. However, we find that impurities can *speed* nucleation in the layer-by-layer growth regime, by lowering the free-energy cost of 2D layer nuclei and providing extra nucleation sites[24,25]. Impurities appear in the final 3D structure in low concentration, and in this respect behave almost as a catalyst.

Simple scaling results explain this catalyst-like mechanism, and suggest that it should be relevant to a wide range of molecular and nanoscale systems. Let $\Delta\epsilon$ be the energy difference between a particle-particle bond and a particle-impurity bond, and let $z_{\rm b}$ and $z_{\rm s} \approx z_{\rm b}/2$ be the bulk- and surface coordination numbers of the structure. If the time intervals between successive nucleation events are long, then a fraction $f_{\rm s} \approx \exp(-\beta z_{\rm s}\Delta\epsilon)$ of surface particles will be impurities [here $\beta \equiv 1/(k_{\rm B}T)$]. Impurities can be numerous enough to lower the barrier to 2D nucleation, and therefore substantially increase the layer-by-layer growth rate, which scales as the exponential of this barrier. Impurities near the growth front can exchange with solution before the front moves away, leading to a bulk impurity

fraction $f_b \approx \exp(-\beta z_b \Delta\epsilon) < f_s$. For large $\beta\Delta\epsilon$ this effect is akin to that of a catalyst, in that impurities can be abundant at the growth front, substantially increase the growth rate, and yet reside in the final structure in much smaller number. This speed-up of growth is reminiscent of the nucleation enhancement of colloidal clusters by liquid-vapor critical fluctuations[26], in the sense that impurities serve as a source of fluctuations that promote a desired ordering process.

4.2 Model

We demonstrate this effect using a lattice model of two-component growth introduced previously[27,28]. Lattice sites can be vacant (white), or occupied by blue or red particles; these represent crystal and impurity particles, respectively. We refer to a blue structure as a crystal. Contacts between nearest-neighbor blue particles contribute a favorable binding energy $-\epsilon_b < 0$, while blue-red and red-red contacts contribute a less favorable energy $-\epsilon_r < 0$ ($\epsilon_b > \epsilon_r$). White sites carry an energy penalty of μ. The quantity $\Delta\mu \equiv 3\epsilon_b - \mu$, which we call the supersaturation, is the bulk free-energy difference between an all-white state and an all-blue state; when $\Delta\mu > 0$ there exists a thermodynamic driving force to grow a crystal from solution. We carried out Monte Carlo simulations of this model on a 3D cubic lattice of 12×12 sites in the xy plane. Periodic boundary conditions were applied in this plane, and the crystal was seeded with 3 blue layers. The other direction, z, is the growth direction.

We evolved the model using the discrete-time Monte Carlo dynamics considered previously[27,28] (reproduced for completeness in Appendix). To allow access to long timescales we carried out an additional set of simulations in which we imposed a solid-on-solid (SOS) restriction[29,30]: for sites with given values of (x, y) we proposed Monte Carlo moves only at two sites, the occupied site with the largest value of z and its neighboring unoccupied site. This restriction reduces the number of moves required to observe growth by a factor of order the length of the system [Fig. 4.6(a)]. It also artificially prevents vacancies within the solid, leading to a restricted equilibrium in which bulk vacancies do not exist [a]. However, in the

[a] This restriction also prevents non-physical vacancy-catalyzed annealing in the interior of the solid[31].

regime studied here the equilibrium vacancy concentration is very small[32] and, as a result, the differences between our results in the presence and absence of the SOS constraint are negligible [Fig. 4.6(b)]. Here we present results obtained with the constraint.

4.3 Impurities speed growth

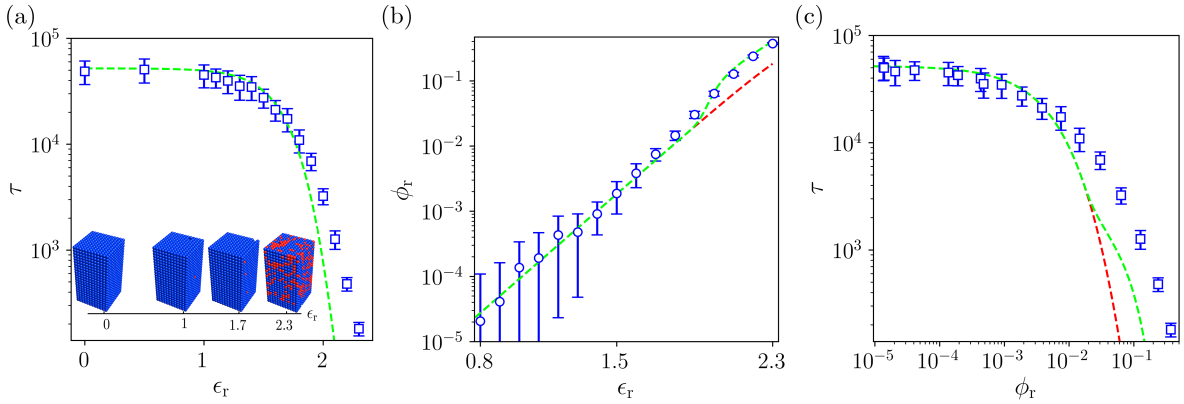

Figure 4.1: (a) Layer addition time and representative snapshots and (b) impurity fraction for $\epsilon_b = 2.55, \Delta\mu = 0.25$. The dashed line in (a) is the prediction of Eq. 4.1, and the dashed lines in (b) and (c) are the predictions of Eq. 4.10 (red) and Eq. 4.11 (green). For impurity binding energies $\epsilon_r < 1.9$ ($\phi_r < 10^{-2}$) impurity relaxation is sufficiently fast that the solid composition can be approximated by the equilibrium result (red), whereas for large binding energies additional impurities become trapped by the advancing growth front (green). (c) Parametric plot of the data in panels (a) and (b) showing the layer addition time as a function of impurity fraction.

In Fig. 4.1 we show the mean time to grow one layer of the crystal, and the impurity fraction in the bulk, for various values of the impurity interaction ϵ_r (the impurity-free case corresponds to the limit $\epsilon_r \to -\infty$). Simulations were stopped when 20 layers were deposited (we define a layer as an (x, y) plane in which at least half the sites are occupied by colored particles). The growth time is defined as the average number of Monte Carlo moves required to complete a layer. The impurity fraction is defined as the number of red particles divided by the number of colored particles. We see that the growth time (eventually) decreases as the impurity binding energy increases, and the grown structure contains an increasing number of impurities. As we shall show, by varying conditions it is possible to have the growth time decrease more rapidly than the impurity fraction increases.

To estimate the growth time of the crystal we focus our discussion on the layer-by-layer growth regime at low temperature, where growth is limited by the nucleation of new layers on the crystal surface. When the time for 2D nucleation is much longer than the time for the resulting postcritical cluster to grow to completion, the layer growth time τ scales as

$$\tau \sim \exp(G_{\max}), \tag{4.1}$$

where $G_{\max}$ is the free energy of the critical 2D cluster (here and subsequently we work in units such that $k_{\mathrm{B}}T = 1$). Eq. 4.1 is valid when the layer completion time is short compared to the nucleation time, the regime on which we focus (more generally, see Ref. Saito[29]). To estimate $G_{\max}$ we consider a $k \times k$ cluster on a flat blue surface [b]. Each particle incurs a chemical potential cost μ, so the chemical potential cost of the cluster is $k^2\mu = 3\epsilon_{\mathrm{b}}k^2 - k^2\Delta\mu$. Each of the k^2 particles in the cluster makes one bond with the layer below it, and there are $2k(k-1)$ in-plane bonds. Thus the total bonding energy is $(-\epsilon_{\mathrm{b}}) \times (3k^2 - 2k)$. Adding to this the chemical potential cost gives the energy cost for making a $k \times k$ square:

$$G(k) = 2k\epsilon_{\mathrm{b}} - k^2\Delta\mu. \tag{4.2}$$

For nonzero supersaturation this function has a maximum at $k_\star = \epsilon_{\mathrm{b}}/\Delta\mu$. The critical cluster therefore contains $k_\star^2 = (\epsilon_{\mathrm{b}}/\Delta\mu)^2$ particles, and the corresponding energy barrier is $G(k_\star) = \epsilon_{\mathrm{b}}^2/\Delta\mu$.

To understand how this result changes in the presence of impurities (red particles), consider the following simple argument. Let a lattice site be surrounded by z blue particles and $6-z$ white particles, and let p be the probability that an isolated particle is a crystalline one as opposed to being an impurity (in simulations we model an equimolar mixture of crystal- and impurity particles, and so set $p = 1/2$). At that lattice site, in a mean-field approximation, the thermal weight of a blue particle is $p\mathrm{e}^{z\epsilon_{\mathrm{r}}}$; the thermal weight of a red particle

[b] An accurate expression for 2D nuclei of irregular shape can be found in Ref. Ryu and Cai[33]; approximating nuclei as squares incurs numerical errors, but captures important trends of barrier height with model parameters.

is $(1-p)\mathrm{e}^{z\epsilon_\mathrm{b}}$; and the thermal weight of a vacancy is e^{μ}. Thus the equilibrium fraction of colored particles is

$$f_1 = \frac{(1-p)\mathrm{e}^{z\epsilon_\mathrm{r}} + p\mathrm{e}^{z\epsilon_\mathrm{b}}}{(1-p)\mathrm{e}^{z\epsilon_\mathrm{r}} + p\mathrm{e}^{z\epsilon_\mathrm{b}} + \mathrm{e}^{\mu}} = \frac{\mathcal{G}p\mathrm{e}^{z\epsilon_\mathrm{b}}}{\mathcal{G}p\mathrm{e}^{z\epsilon_\mathrm{b}} + \mathrm{e}^{\mu}}, \tag{4.3}$$

where $\mathcal{G} \equiv 1 + (p^{-1} - 1)\mathrm{e}^{-z\Delta\epsilon}$ and $\Delta\epsilon \equiv \epsilon_\mathrm{b} - \epsilon_\mathrm{r}$. The corresponding expression in the absence of impurities is

$$f_2 = \frac{p\mathrm{e}^{z\epsilon_\mathrm{b}}}{p\mathrm{e}^{z\epsilon_\mathrm{b}} + \mathrm{e}^{\mu}}. \tag{4.4}$$

Comparison of f_1 and f_2 indicates that $\mathcal{G}$ functions as an effective degeneracy for blue particles. Alternatively, we can consider that the effective blue-particle interaction energy in the presence of impurities is larger than in their absence, i.e. $\mathrm{e}^{z\epsilon_\mathrm{eff}} = \mathcal{G}\mathrm{e}^{z\epsilon_\mathrm{b}}$, giving

$$\epsilon_\mathrm{eff} = \epsilon_\mathrm{b} + \frac{1}{z}\ln\left[1 + (p^{-1} - 1)\mathrm{e}^{-z\Delta\epsilon}\right]. \tag{4.5}$$

The argument leading to Eq. 4.2 can now be modified, by replacement of ϵ_b with ϵ_eff in the bond-energy reward term, to estimate the energy cost $G_\mathrm{eff}(k) = G(k) + \Delta G(k)$ required to make a $k \times k$ cluster in a solution of particles and impurities:

$$\Delta G(k) = \frac{k(2-3k)}{z}\ln\left[1 + (p^{-1} - 1)\mathrm{e}^{-z\Delta\epsilon}\right]. \tag{4.6}$$

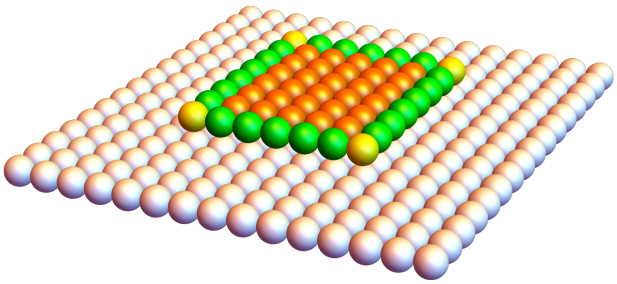

Figure 4.2: A square nucleus on a flat surface. Particles with 4 in-plain bonds, 3 in-plain bonds, and 2 in-plain bonds at the 4 corners are in orange, green, and yellow, respectively.

To estimate the mean coordination number z as a function of k, note that in a $k \times k$ cluster we have $(k-2)^2$ particles with 4 in-plane bonds, $4(k-2)$ particles with 3 in-plane bonds, and 4 corner particles with 2 in-plane bonds (Fig. 4.2). Each particle makes one extra bond with the substrate. Thus the average coordination number is $z(k) = 5 - 4/k$.

Inserting $z(k)$ into Eq. 4.6 gives

$$\Delta G(k) = -\frac{k^2(3k-2)}{5k-4} \ln\left[1 + (p^{-1}-1)\mathrm{e}^{-\Delta\epsilon(5-4/k)}\right]. \tag{4.7}$$

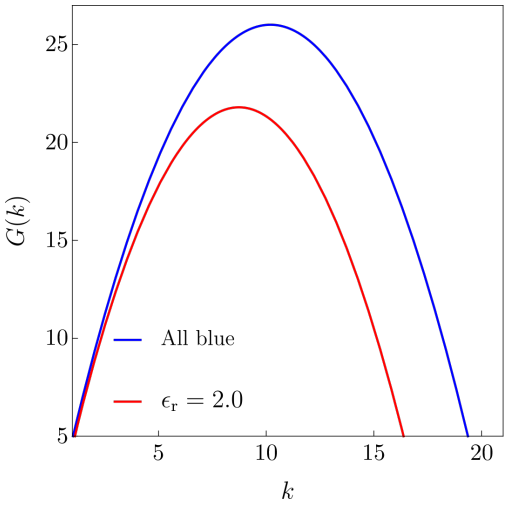

Figure 4.3: Free energy barrier with (red curve) and without (blue curve) the presence of impurities ($\epsilon_b = 2.55$, $\epsilon_r = 0.2$, and $\Delta\mu = 0.25$).

The right-hand side of Eq. 4.7 describes the impurity-induced reduction in the energy cost of a $k \times k$ cluster (we recover the no-impurity case in the limit $\Delta\epsilon \to \infty$). For small $\Delta\mu$ the function $G_{\mathrm{eff}}(k)$ will take its maximum at a value of $k \gg 1$. In this regime we can expand Eq. 4.7 to get $G_{\mathrm{eff}}(k) \approx 2k\epsilon_b - k^2\mu_{\mathrm{eff}}$, which has the same form as the impurity-free expression Eq. 4.2 but with effective supersaturation

$$\Delta\mu_{\mathrm{eff}} = \Delta\mu + \frac{3}{5}\ln\left[1 + (p^{-1}-1)\mathrm{e}^{-5\Delta\epsilon}\right]. \tag{4.8}$$

The free-energy barrier to layer nucleation in the presence of impurities can then be estimated as

$$G_{\max} \approx \frac{\epsilon_b^2}{\Delta\mu_{\mathrm{eff}}}. \tag{4.9}$$

Note that the reduction to the nucleus free energy (Fig. 4.3) enters through the bulk term, not the surface term as is typical in models of heterogeneous nucleation at a surface.

We next consider the fraction of impurities involved during growth (in the parameter regime in which this fraction is small). For a lattice site surrounded by z blue particles, the equilibrium fraction of red particles is

$$\phi(z) = \frac{(1-p)\mathrm{e}^{z\epsilon_\mathrm{r}}}{(1-p)\mathrm{e}^{z\epsilon_\mathrm{r}} + p\mathrm{e}^{z\epsilon_\mathrm{b}}} = \frac{1-p}{1-p+p\,\mathrm{e}^{z\Delta\epsilon}}. \tag{4.10}$$

This fraction is smaller in the interior of the crystal, where the impurity makes $z_\mathrm{b} = 6$ blue contacts, than at the surface.

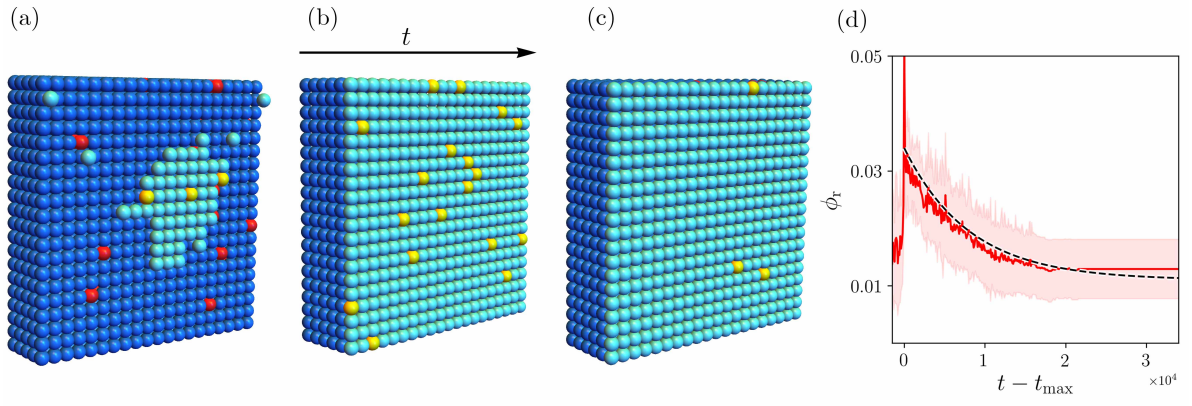

Figure 4.4: Impurities are incorporated in each layer and gradually anneal to a more ordered structure. Snapshots of the annealing of a representative layer (in that layer only, blue particles are colored light blue, and red particles are colored yellow) (a) shortly after nucleation, (b) upon completion of the layer, and (c) after the growth front has moved away. Subsequent layers have been omitted for clarity. (d) Time progression of the impurity content in a layer (averaged over 10 simulations). The decay of the impurity fraction after reaching a peak value (at $t = t_\mathrm{max}$) approaches the estimate Eq. 4.11 (dashed line).

As the completed layer becomes covered by new particles, it will evolve toward the bulk defect concentration. The timescale for this relaxation, τ_r, is the timescale for a fivefold-coordinated particle at the surface to unbind, and so we estimate $\tau_\mathrm{r} \propto \mathrm{e}^{5\epsilon_\mathrm{r}}$. Provided the layer addition time τ is longer than this, we estimate the impurity fraction in a newly completed layer as

$$\phi_\mathrm{r} \approx \phi(5)\mathrm{e}^{-\tau/\tau_\mathrm{r}} + \phi(6). \tag{4.11}$$

This annealing process is illustrated in Fig. 4.4. The snapshots (a–c) and time-trace (d) show that impurity particles are present at higher concentration at the growth front than in the bulk of the structure. The relaxation of the impurity fraction from the surface- to the bulk equilibrium concentration occurs in a manner consistent with Eq. 4.11; see panel (d).

4.4 Identifying the parameter regime in which impurities are of most benefit

The preceding analysis confirms that impurities speed layer nucleation, via Eq. 4.1, Eq. 4.8, and Eq. 4.9, and make the equilibrium solid less pure, via Eq. 4.10. Impurities are most beneficial when the former effect is as large as possible, and the latter effect as small as possible. To make the bulk equilibrium impurity concentration Eq. 4.10 small we want $\Delta\epsilon$ large; we then want $\Delta\mu$ small, so that the second term in Eq. 4.8 remains significant.

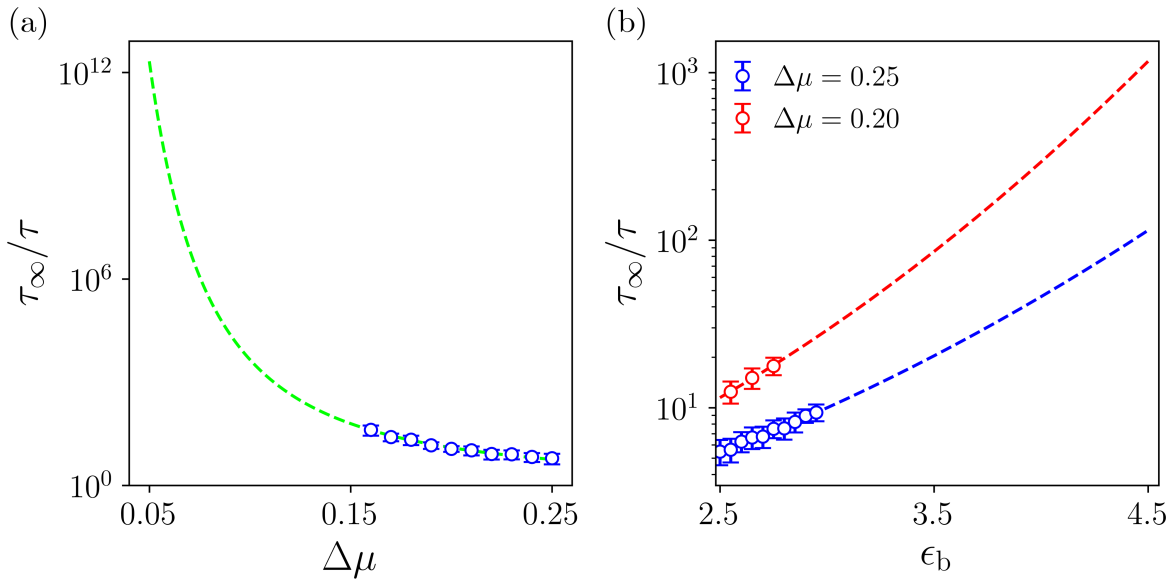

Figure 4.5: Ratio of the growth time τ in the presence of impurities to that in the impurity-free case, τ_∞, as a function of (a) supersaturation and (b) binding energy. The beneficial effect of impurities is most pronounced in the presence of small supersaturation and large binding energies. In both panels, parameters are chosen so that the bulk equilibrium impurity fraction is always 1%. The dashed lines are the predictions of Eq. 4.1.

In Fig. 4.5 we show that these predictions are consistent with our simulations: a crystal of a certain impurity fraction grows more rapidly than its impurity-free counterpart, and this effect is much enhanced as supersaturation is reduced. Our predictions also suggest that impurities can be orders of magnitude more effective in parameter regimes that are inaccessible to our simulations but which describe molecular systems.

4.5 Conclusions

Impurities are often considered to be problematic when attempting to grow crystals, but we have shown that layer-by-layer growth can be dramatically sped up by impurities with little impact on the quality of the final structure. Our computer simulations and simple scaling arguments suggest that this effect will be most pronounced under conditions of low supersaturation and low temperature. Such conditions are often required for the crystallization of highly anisotropic molecules, for which the probability of crystalline (or productive) binding is small. For example, proteins must sample an ensemble of $\simeq 10^4 - 10^5$ states in order to find the crystallographic state[34–36]. Given many ways of misbinding, growth must be slow (and so supersaturation must be low) in order to allow time for error correction. Furthermore, a large binding energy is needed to offset the entropic advantage of the disordered ensemble[31]. This combination of large binding energies and low supersaturation leads to high surface tension and long nucleation times, precisely the region in which impurities are expected to be beneficial [Fig. 4.5(b)]. Indeed, this mechanism may provide an explanation for the utility of non-specific binding enhancers in protein crystallization[37–39], such as depletants, in the layer-by-layer growth regime.

Appendix

Lattice model Monte Carlo simulations

The unrestricted Monte Carlo protocol proceeds as follows. At each step of the simulation a site was chosen at random. If the chosen site was white then we proposed with probability p (resp. $1-p$) to make it blue (resp. red). If the chosen site was red or blue then we proposed to make it white. No red-blue interchange was allowed. These proposals were accepted with probabilities

$$
\begin{aligned}
\mathrm{R} \to \mathrm{W} &: \min(1, (1-p)\exp(-\Delta E)); \\
\mathrm{W} \to \mathrm{R} &: \min(1, (1-p)^{-1}\exp(-\Delta E)); \\
\mathrm{B} \to \mathrm{W} &: \min(1, p\exp(-\Delta E)); \\
\mathrm{W} \to \mathrm{B} &: \min(1, p^{-1}\exp(-\Delta E)),
\end{aligned}
\tag{4.12}
$$

where ΔE is the energy change resulting from the proposed move. This change was calculated from the lattice energy function

$$
E = \sum_{\langle i,j \rangle} \epsilon_{C(i)C(j)} + \sum_i \mu_{C(i)}. \tag{4.13}
$$

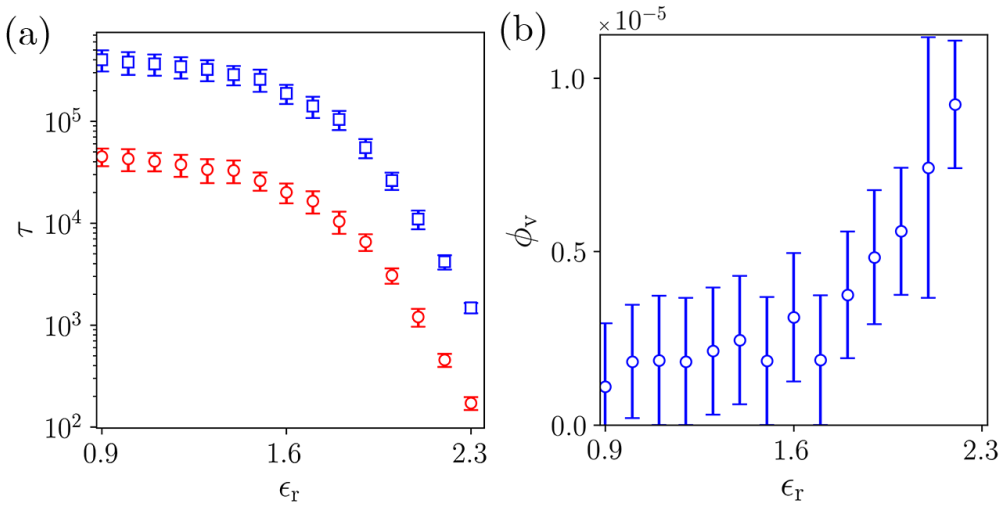

Figure 4.6: The growth time (a) and average fraction of vacancies (b) in the bulk as a function of impurity binding energy. In (a), the SOS restriction (red) reduces the the number of moves required to observe growth (by a factor of order the length of the system) compared with the unrestricted Metropolis Monte Carlo simulation (blue). (b) shows the fraction of vacancies in the bulk, averaged over 100 simulations, in the absence of the SOS restriction. These small vacancy fractions show that the effect of imposing the SOS restriction (which eliminates vacancies) is slight.

The first sum runs over all distinct nearest-neighbor interactions. The second sum runs over all sites. The index $C(i)$ describes the color of site i and is W (white), B (blue), or R (red); $\epsilon_{C(i)C(j)}$ is the interaction energy between colors $C(i)$ and $C(j)$ (this is zero if either site is white); and the chemical potential $\mu_{C(i)}$ is μ, $\ln p$ and $\ln(1-p)$ for W, B, and R,

respectively. In the main text we set $p = 1/2$ in order to model an equimolar mixture of crystal- and impurity particles.

In the main text we describe a solid-on-solid (SOS) restricted protocol in which Monte Carlo moves are performed only at the growth front. This protocol, which does not allow vacancies to become incorporated into the 3D structure, results in a different equilibrium than the unrestricted protocol. However, in the parameter regime we probe the difference is slight, because few vacancies appear in the unrestricted protocol (Fig. 4.6), and the presence or absence of the restriction does not qualitatively affect our conclusions.

References

[1] S. Zhang, Nature biotechnology **21**, 1171 (2003).

[2] J. C. Huie, Smart Materials and Structures **12**, 264 (2003).

[3] B. A. Lindquist, R. B. Jadrich, and T. M. Truskett, *Communication: Inverse design for self-assembly via on-the-fly optimization* (2016).

[4] S. Whitelam and R. L. Jack, Annual Review of Physical Chemistry **66**, 143 (2015).

[5] A. W. Long and A. L. Ferguson, Molecular Systems Design & Engineering **3**, 49 (2018).

[6] A. W. Wilber, J. P. Doye, A. A. Louis, E. G. Noya, M. A. Miller, and P. Wong, The Journal of chemical physics **127**, 08B618 (2007).

[7] D. Rapaport, Physical Review Letters **101**, 186101 (2008).

[8] M. F. Hagan and D. Chandler, Biophysical journal **91**, 42 (2006).

[9] T. K. Haxton and S. Whitelam, Soft Matter **8**, 3558 (2012).

[10] K. Kremer, Journal of Aerosol Science **9**, 243 (1978).

[11] D. Stauffer, Journal of Aerosol Science **7**, 319 (1976).

[12] H. Trinkaus, Phys. Rev. B **27**, 7372 (1983).

[13] J. Schmelzer, A. Abyzov, and J. Möller, The Journal of Chemical Physics **121**, 6900 (2004).

[14] J. Schmelzer, J. Schmelzer Jr, and I. Gutzow, The Journal of Chemical Physics **112**, 3820 (2000).

[15] R. Scarlett, J. Crocker, and T. Sinno, The Journal of Chemical Physics **132**, 234705 (2010).

[16] A. Kim, R. Scarlett, P. Biancaniello, T. Sinno, and J. Crocker, Nature Materials **8**, 52 (2008).

[17] R. Scarlett, M. Ung, J. Crocker, and T. Sinno, Soft Matter **7**, 1912 (2011).

[18] N. M. Dixit and C. F. Zukoski, Phys. Rev. E **66**, 051602 (2002).

[19] S. A. Kulkarni, S. S. Kadam, H. Meekes, A. I. Stankiewicz, and J. H. ter Horst, Crystal Growth & Design **13**, 2435 (2013).

[20] T. E. Paxton, A. Sambanis, and R. W. Rousseau, Langmuir **17**, 3076 (2001).

[21] J. K. Dhont, C. Smits, and H. N. Lekkerkerker, Journal of Colloid and Interface Science **152**, 386 (1992), ISSN 0021-9797.

[22] G. Ungar, E. G. R. Putra, D. S. M. de Silva, M. A. Shcherbina, and A. J. Waddon, Advances in polymer science **180**, 45 (2005).

[23] T. Schilling and D. Frenkel, Journal of Physics: Condensed Matter **16**, S2029 (2004).

[24] M. van der Leeden, D. Kashchiev, and G. van Rosmalen, Journal of Crystal Growth **130**, 221 (1993), ISSN 0022-0248.

[25] R. M. Ginde and A. S. Myerson, Journal of Crystal Growth **126**, 216 (1993), ISSN 0022-0248.

[26] P. R. ten Wolde and D. Frenkel, Science **277**, 1975 (1997).

[27] S. Whitelam, R. Schulman, and L. Hedges, Physical Review Letters **109**, 265506 (2012).

[28] S. Whitelam, L. O. Hedges, and J. D. Schmit, Physical Review Letters **112**, 155504 (2014).

[29] Y. Saito, *Statistical Physics of Crystal Growth* (World Scientific, 1996).

[30] J. M. Kim and J. M. Kosterlitz, Phys. Rev. Lett. **62**, 2289 (1989).

[31] S. Whitelam, Y. R. Dahal, and J. D. Schmit, The Journal of chemical physics **144**, 064903 (2016).

[32] S. Whitelam, The Journal of Chemical Physics **149**, 104902 (2018).

[33] S. Ryu and W. Cai, Physical Review E **81**, 030601 (2010).

[34] D. Asthagiri, A. Lenhoff, and D. Gallagher, Journal of Crystal Growth **212**, 543 (2000).

[35] A. M. Kierzek, W. M. Wolf, and P. Zielenkiewicz, Biophysical journal **73**, 571 (1997).

[36] J. D. Schmit and K. Dill, Journal of the American Chemical Society **134**, 3934 (2012).

[37] P. G. Vekilov and J. I. D. Alexander, Chemical reviews **100**, 2061 (2000).

[38] R. Giegé, The FEBS journal **280**, 6456 (2013).

[39] S. D. Durbin and G. Feher, Annual Review of Physical Chemistry **47**, 171 (1996).

Chapter 5

Conformational entropy limits the transition from nucleation to elongation in amyloid aggregation

The formation of amyloid fibrils in Alzheimer's disease and other neurodegenerative disorders is limited by a slow nucleation step due to the entropic cost to initiate the ordered cross-β structure. While the nucleation barrier can be lowered if the molecules maintain conformational disorder, clusters with little secondary structure provide a poor binding surface for new molecules. To understand these opposing factors, we used all-atom simulations to parameterize a lattice model that treats each amino acid as a binary variable with β-sheet and non-β states. We find that the optimal amount of secondary structure in a critical nucleus depends on protein concentration. Low concentration systems require more ordered nuclei to capture infrequent monomer attachments. Our model explains the transition from the nucleation phase to elongation as the point where the β-sheet core becomes large enough to overcome the initiation cost, at which point β-strand elongation becomes favorable and the nascent fibril efficiently captures new molecules.

5.1 Introduction

The assembly of proteins into amyloid fibrils causes numerous neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease, and prion disorders[1]. *In vitro* experiments show that the conversion to the fibril state is limited by nucleation events, indicative of a free

energy barrier to initiate the fibril state. In the case of isotropic particle condensation, the nucleation barrier arises because particles at the periphery have sacrificed the translational entropy of the dilute phase, but only form a fraction of the favorable interactions available to interior particles[2]. This interaction deficit, usually described as a surface tension, becomes a smaller fraction of the free energy as the cluster grows larger.

Surface tension does not limit 1D assemblies because the surface energy does not depend on the cluster size[3]. Early attempts to explain amyloid nucleation identified β-sheet layering as a second assembly dimension[4–9]. In Ref. Šarić et al.[10], the authors show that the conformational conversion also plays an important role in converting monomers to β-prone state. We used all-atom simulations to observe the conformational change in two cases: when the two disordered molecules initiate a β-structure and when a molecule joins an established template Fig. 5.1(a, b). The simulation rates give rise to two kinds of bond free energies, repulsion and attraction, corresponding to the two configurations, which is consistent with the entropy interpretations[11–13]. The initiation of β-structure from the two disordered monomers requires both molecules to loose conformational entropy, which leads to net repulsive. However, for a molecule joining an established fibril the template is rigid and the entropy loss is limited solely to the incoming molecule. This event is more favorable and results in net attractive.

We used the free energy model and lattice simulations to explore the transition between the disordered initiation and the retention of incoming molecules with a large β-sheet template. To access nucleation timescales, we adopt a multi-scale approach in which all-atom simulations are used to parameterize a lattice model. Nuclei consist of a β-sheet core surrounded by disordered tails. The shape of the β-sheet core depends on both the strength of attraction and the protein concentration, consistent with classical nucleation theory (CNT) applied at the amino acid level. Concentrated solutions favor clusters with shorter β-strands, while lower concentrations favor longer β-strands. The transition from transient binding during nucleation to nearly irreversible binding during elongation is explained by the enlargement of the β-core after it reaches a critical size, which provides a stronger binding surface for incoming molecules.

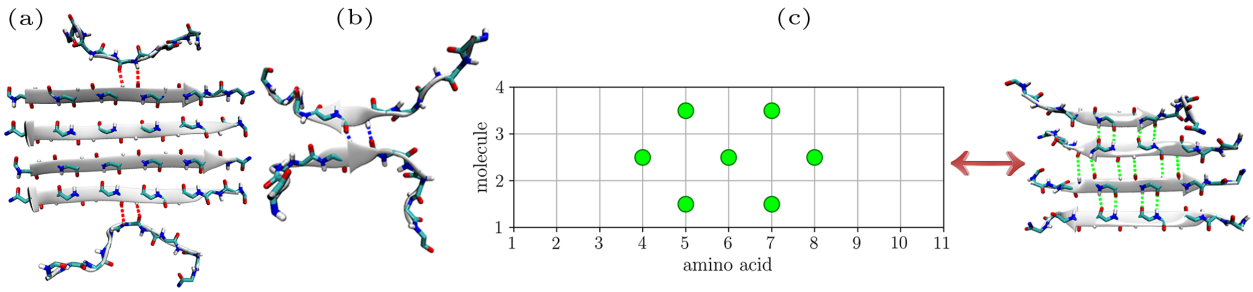

Figure 5.1: MD snapshots from the sampling of strong and weak H-bonds (side chains are not shown for clarity). Kinetic parameters for strong bonds are sampled from the terminal molecules on an established cluster (a), while weak bonds are sampled using a dimer that is harmonically restrained at the central amino acids (b). (c) Schematic of the mapping between the lattice representation and the atomistic representation.The lattice has a width given by the number of amino acids per molecule and a height given by the current number of molecules in the cluster. Lattice sites (represented by vertical lines) can be occupied, representing an H-bond between connected amino acids, or empty, indicating that at least one of the adjacent amino acids is in the random coil state. The alternating direction of H-bonds (seen in the atomistic view) means that only every-other site can have a bond.

5.2 Model

We use two sets of bonding rate constants, depending on the position in the cluster. *"Strong"* bonds form between a disordered amino acid and an established β-strand. *"Weak"* bonds form between two disordered amino acids. These bonds result in the loss of conformational entropy from both backbones, a greater penalty than bonds with a pre-existing strand[11,12]. The free energy of a square cluster can be written in terms of the total bonds N, an energy penalty coming from the conformational change, and the number of molecules M in the cluster (see Appendix for derivation),

$$F = N\epsilon_s + \gamma\ell + \mu M, \tag{5.1}$$

where $\ell = N/(M-1)$ is the average length of β-strands, μ and $\gamma = \epsilon_w - \epsilon_s$ serve as the line tensions in the vertical and horizontal directions, respectively. $\mu = -k_\mathrm{B}T \ln c/c_0$ (c_0 is a reference concentration) is the chemical potential and ϵ_w, ϵ_s are the free energy of the weak bonds and strong bonds, respectively. Strong and weak bond kinetic parameters were assessed using all-atom simulations of hexamer and dimer assemblies (Fig. 5.1a, b)

with harmonic restraints applied to the non-sampled bonds (see SI). Here we present results for the AMBER14SB force field, which has intermediate affinity of the three tested (see SI). AMBER14SB weak bonds have free energy $\epsilon_w = 0.71\ k_\mathrm{B}T$, and the strong bonds, $\epsilon_s = -0.93\ k_\mathrm{B}T$ provide enough attraction to stabilize a single-layered β-sheet. The difference $\epsilon_w - \epsilon_s = 1.64\ k_\mathrm{B}T$ is in good agreement with estimates for the entropic cost of secondary structure[13].

For fixed N, the optimal cluster dimensions are $\ell = (\mu N/\gamma)^{1/2}$ and $M = (\gamma N/\mu)^{1/2}$, which can be used to find the free energy maximum

$$F^\ddagger = \frac{\gamma\mu}{|\epsilon_s|}, \tag{5.2}$$

$$N^\ddagger = \frac{\gamma\mu}{\epsilon_s^2}, \tag{5.3}$$

at which point the cluster has dimensions

$$M^\ddagger = \frac{\gamma}{|\epsilon_s|} + 1, \tag{5.4}$$

$$\ell^\ddagger = \frac{\mu}{|\epsilon_s|}. \tag{5.5}$$

Using the bond free energy from AMBER14SB gives $M^\ddagger = 2.76$. From Fig. 5.4a, we obtain the reference concentration $c_0 = 44.34\ \mu$M by fitting the data to $N^\ddagger$ in Eq. 5.3. This value can be used in the expression for $\ell^\ddagger = k_\mathrm{B}T\ln(c/c_0)/(\epsilon_s)$, which captures the change in β-ordering as a function of concentration (Fig. 5.4b).

5.3 Results and Discussion

Lattice model captures molecule addition and β-sheet formation. Our lattice model is based on a Markov State Model (MSM) developed to study fibril elongation[14,15]. In those works the conformational search was discretized using two reaction coordinates: 1) the alignment of an incoming molecule with the template and 2) the number of β-sheet H-bonds. Here we remove the alignment complication by considering polyglutamine, motivated by the

aggregation-prone region of huntingtin protein[16–23].

The lattice model, illustrated in Fig. 5.1(c), evolves by the Gillespie algorithm[24]. Lattice model monomers have 11 amino acids and form anti-parallel β-sheets, which is more stable than parallel β-sheets for polyglutamine[25]. Each peptide unit is modeled as a binary variable with states representing β-sheet and non-β conformations. Each amino acid can form a pair of H-bonds, which are mapped to a single bond in the lattice model [Figs. 5.1(c)]. Peptide units at the periphery of the β-core fluctuate between β-sheet and non-β states at rates measured from the all-atom model. New molecules can add to either end of the β-sheet at a concentration-dependent rate approximated by the Smoluchowski formula for an absorbing sphere

$$k_{\text{add}} = 4\pi\sigma D_{\text{m}} c \tag{5.6}$$

where c is the protein concentration, $\sigma = 1.75$ nm is the radius of the sphere, approximated by half the length of an extended monomer, and D_{m} is the diffusion coefficient of the monomer, $1.79 \times 10^{-10}\text{m}^2/\text{s}$[26].

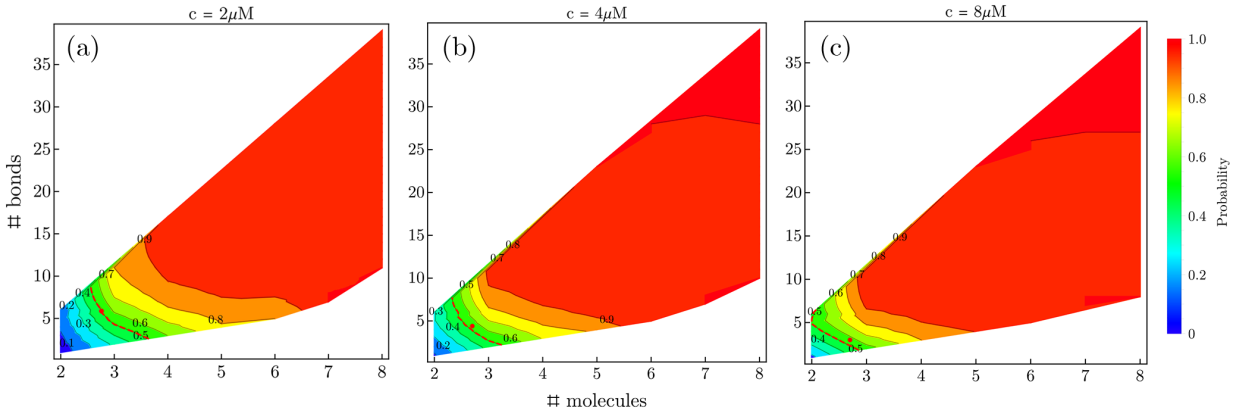

Figure 5.2: Probability of successful nucleation trajectories with parameters from the AMBER14SB force field as a function of the number of molecules and the total number of H-bonds in the cluster. Increasing concentration shifts the transition surface (50%, red dotted line) to smaller clusters and reduces the need for β-structure due to increased monomer deposition rate.

The committor is used to identify the transition state ensemble. The committor is used to identify the transition state ensemble. The committor is defined as the probability that a nucleus with a specified number of bonds reaches the cutoff size of 8 molecules before

completely dissolving. If a molecule breaks all bonds with the cluster it is considered to return to free solution. Our model does not consider molecules that associate with the cluster without backbone H-bonds. These non-β contacts are a negligible contribution to the association time during elongation[15]. The lower β-content during nucleation will increase the importance of non-β states, however, we expect the primary effect would be to enhance the local concentration, thereby increasing the rate new molecules reach the ends of the β-sheet. More pronounced effects will occur at high concentrations where disordered oligomers are stable[10,27,28].

Established β-structure helps capture new molecules. The assembly will grow when the attachment rate is greater than the detachment rate and shrink when the detachment rate is greater. While the attachment rate is a function of concentration (Eq. 5.6), detachment is limited by the rupture of favorable interactions. To gain intuition, consider an Arrhenius model in which the molecular detachment rate scales as $e^{-n|\epsilon_s|/k_{\mathrm{B}}T}$, where n is the number of strong bonds between the cluster and the departing molecule. In the dimer $n = 0$, so the detachment rate is large. In the elongation phase n is given by the molecule length so the detachment rate is small.

Fibril nucleation is slow because it is unfavorable to form a template large enough to capture new molecules. However, the attachment rate increases with concentration, reducing the required template size. This is seen in Fig. 5.2 which shows the committor as a function of the number of molecules M and the number of bonds N. At high concentration (Fig. 5.2c, 5.4a) nucleation is more likely than dissolution for a cluster containing $M \simeq 3$ and $N \simeq 3$. But, at lower concentration (Fig. 5.2a, 5.4a) a 50% committor is not reached until $M \simeq 3$ and $N \simeq 6$. The trend toward larger N at lower concentrations is consistent with CNT. We refer the 50% committor surface as the transition state, whereas state at the free energy saddle point are referred to as "critical" (Fig. 5.3b). The free energy contours are plotted along with the lowest free energy path with the red dots showing the saddle points. Increasing concentration has the effect of tipping the pathway with the saddle point consisting the same critical molecules but fewer H-bonds.

Nucleation at low concentration requires highly ordered clusters. Fig. 5.4b

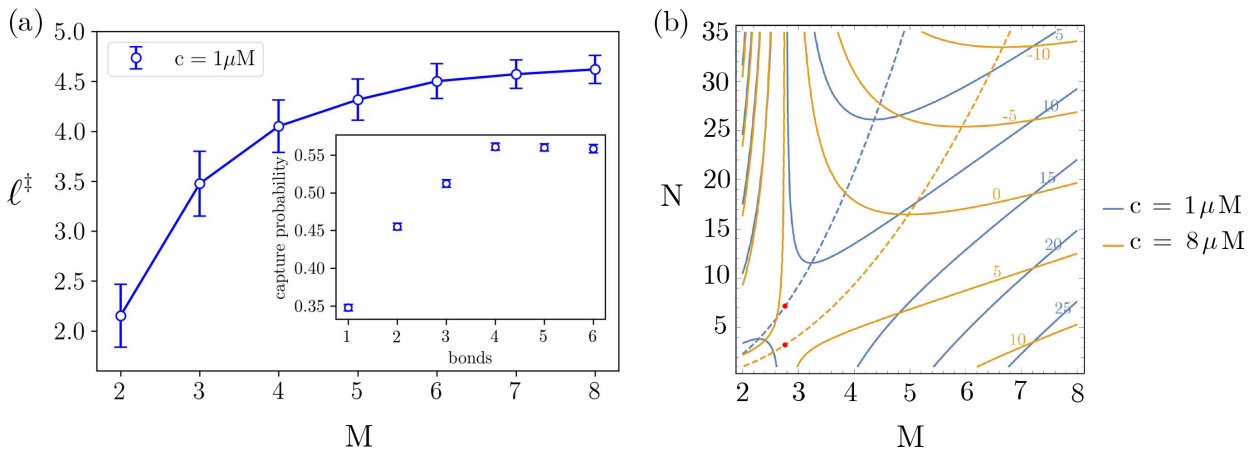

Figure 5.3: (a) Average length of β-strands, ℓ, on successful nucleation trajectories. β-strands asymptotically approach the maximum length above the critical size $M > 3$, but remain short at $M = 2$ where additional secondary structure is unfavorable. (inset) The capture probability, defined as the probability a new molecule arrives before the previous one detaches, depends on the size of the available template. Large templates allow the new molecule to form more bonds, which increases their retention time. (b) Contour plot of cluster free energy, Eq. 5.1 as a function of the number of molecules, M, and the number of intermolecular H-bonds, N. Increasing the concentration lower the overall free energy, decreases the slope of the lowest free energy path (dotted lines), and shifts the free energy saddle point (red dots) to lower values of N.

shows the number of H-bonds per molecule in clusters on the transition surface. We see the trend that low c requires high secondary structure content (~ 4 bonds per molecule) in transition clusters, while poorly ordered clusters (~ 2 bonds/molecule) are enough at high c. Also, the concentration behavior depends on the intermolecular attraction strength. This can be seen from the more aggregation-prone AMBER99SB-ILDN force field, which shows high concentration behavior at 4 μM (Fig. 5.4b). In contrast, AMBER14SB has a larger detachment rate and requires $c \simeq 8$ μM to nucleate from low order ($N/M < 3$) clusters. Note that the concentration changes the line tensions, which controls the cluster shape. The anisotropic line tensions are similar to many crystals.

The elongation phase begins when there are enough molecules that β-strand extension is favorable. Eq. 5.1 provides an explanation for the transition from transient binding during nucleation to efficient capture during elongation. For $M < M^{\ddagger}$ lateral growth of the β core is unfavorable, $(\partial F/\partial N)_M > 0$. This keeps β-strands short, which limits their ability to bind incoming molecules. However, for $M > M^{\ddagger}$, increasing N is favorable.

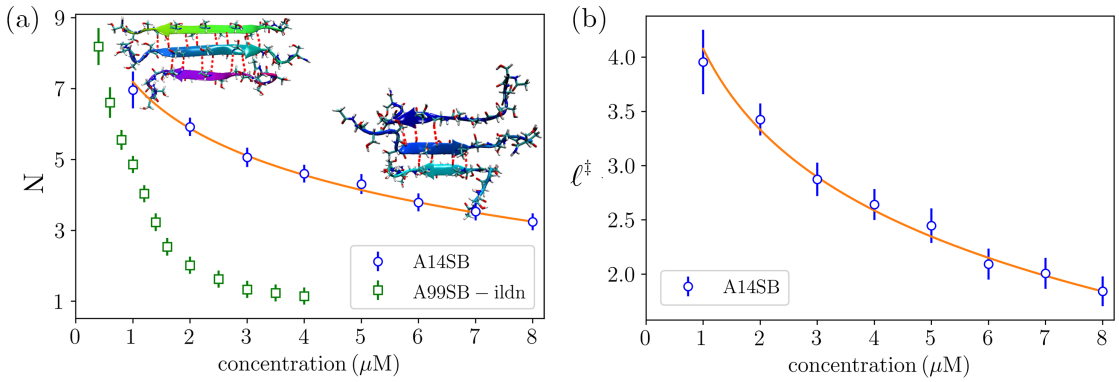

Figure 5.4: (a) Solution concentration determines the shape of the β-sheet core in critical nuclei as seen by the number of H-bonds in transition clusters (defined by the 50% committor). Line shows fit to Eq. 5.3. (inset left) Low concentration nuclei have extensive β-structure to provide a strong binding surface for newly docked molecules. (inset right) Higher concentration nuclei have shorter β-strands because the higher deposition rate places lower demands on retaining new additions. (b) The number of H-bonds per molecule in the transition cluster. Line shows Eq. 5.5.

Therefore, adding molecules leads to an elongation of β-strands. This leads to an increase in capture efficiency (Fig. 5.3a), which signifies the beginning of the elongation phase.

To understand the relationship between secondary structure and the retention of newly added molecules, we computed the capture probability as a function of the template size. For each template we conducted 10000 lattice simulations using parameters from the AMBER14SB force field. Each trial was initiated with a dimer containing 1-6 H-bonds. The cluster was allowed to recruit a third molecule and the simulation was terminated when either a forth molecule attached to the cluster or the third molecule broke all bonds. The capture probability is given by the ratio of these two outcomes. To acquire sufficient statistics, we performed 10 sets of simulations. The capture probability (Fig. 5.3, inset) increases from 0.35 to 0.55 as the template size increases from 2 to 4, but increases in the template size have a minimal effect of capture efficiency. The maximum value of 0.55 is due to the fact that newly bonded molecules have nearly equal probabilities to either break the single H-bond, or form additional bonds.

The increase in the capture probability is illuminating when compared to the extent of β-sheet structure in nucleating clusters. Fig. 5.3a (main panel) shows the average β-strand length, $\langle N/M \rangle$, as a function of M for successful trajectories. The β-strands are very short

for $M = 2$, consistent with the unfavorable free energy to increase N (Eq. 5.1). It becomes favorable for β-strands to elongate when $M > 3$. Accordingly, above this value there is only an asymptotic increase in $\langle N/M \rangle$. Comparing the steep rise in $\langle N/M \rangle$ between $M = 2$ and $M = 4$ to the inset of Fig. 5.3a, we see that this region of expanding secondary structure will result in a corresponding increase in the capture probability.

Disorder in the nucleation phase allows promiscuity in cross-seeding. Aβ_{16-22} reduces the lag time of Aβ_{1-40} despite the fact that molecules are not mixed in the resulting fibrils[29]. This is explained by our results showing that the β-core contains only a portion of the aggregating molecules, so the nucleus will not be sensitive to molecule length (Fig. 5.5). Therefore, the lag time reduction by protein mixtures will depend on c because at lower c the ordered portion of the nucleus will be larger and more sensitive to mismatches. Impurity molecules incorporated during nucleation will be sites susceptible to fragmentation, enabling removal of the defect during elongation.

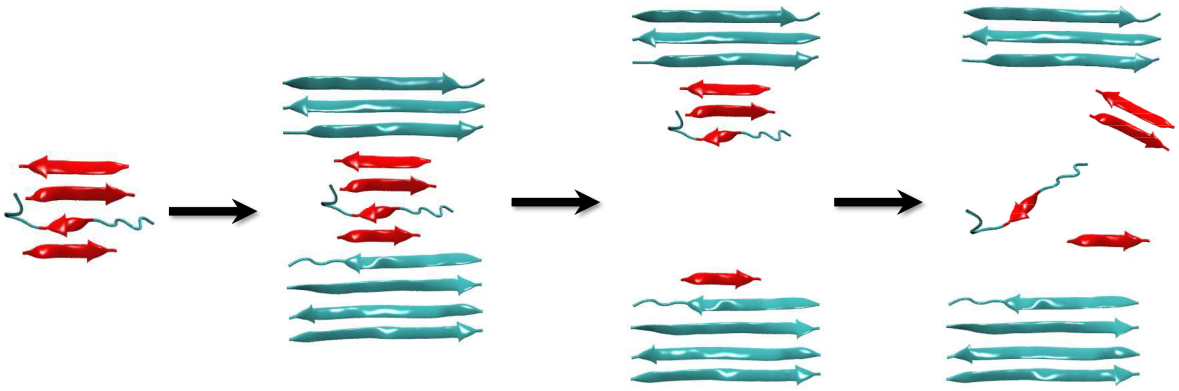

Figure 5.5: Small ordered nucleus can provide templates for other molecules to nucleate and growth. This β-core is also a weak spot and is more prone to fragmentation and let the fibrils continue to grow.

The nucleation mechanism in our simulations blurs the line between one-step nucleation, where condensation and ordering coincide, and two-step nucleation, where condensation precedes ordering[2,30–32]. We find β-sheet ordering occurs concurrently with initial cluster formation and, in fact, is necessary for molecule retention. However, the molecules remain mostly disordered. Previous simulations of fibril nucleation have shown a two-step mechanism for most cases[10]. However, the model in that work represented molecules using an

all-or-nothing conversion between ordered and disordered states, hence high concentration or non-specific attractions were necessary to surmount the conversion barrier. Our work shows that a partial conversion to the β-state is an important mechanism to lower the nucleation barrier.

Our coarse-grained representation only accounts for intermolecular backbone H-bonds, allowing for analysis of secondary structure formation. However, sidechain and non-β backbone contacts are potentially significant in at least two ways. First, omitting non-β interactions prevents completely disordered clusters. This is less of an issue at low concentrations where disordered oligomers are unstable[33–35]. But, disordered binding will, in general, lower the free energy of pre-nucleation clusters[36]. We suggest that in cases where β-sheets nucleate from disordered clusters[10,27] the alignment (rotational) entropy will play a similar role as μ in Eq. 5.1. Second, our model does not include the stacking of β-sheets via steric zipper interactions[37]. The absence of single-layer fibrils in experiments suggests that steric zippers are necessary for fibril stability. This implies that the stability of single-layer sheets in our simulations is a result of force fields that overly stabilize protein-protein interactions[38–41]. Should this be a force field artifact, it is fortuitous for our study because it allows us to study β-sheet initiation without the complication of multiple layers.

5.4 Conclusion

There are many different mechanisms for fibril nucleation, including homogenous nucleation, secondary nucleation catalyzed by fibrils, and heterogeneous nucleation at impurities or interfaces[42–46]. All of these pathways must surmount a free energy barrier that arises from the fact that immature clusters lack the stabilizing interactions of established fibrils so the entropic costs of condensing and ordering are incompletely compensated. Our work shows that the conformational entropy contribution to the barrier is reduced by limiting the extent of secondary structure in the cluster and that the optimal amount of structure depends on the concentration of free protein. This finding should apply to heterogeneous and secondary nucleation pathways as well.

Appendix

Molecular dynamics simulation details

To simulate strong bonds in the all-atom model we used a β-sheet consisting of six monomers. The backbone atoms of the four internal molecules were harmonically restrained, along with the central amino acid of the terminal molecules, using a force constant of 10 kcal/mol/Å^2. The system was heated to 900 K for 100 ps generating an ensemble of 10 initial states. We then monitored H-bond transitions around the anchored amino acid (Fig. 5.1a).

Weak bonds were assessed using an anti-parallel dimer with harmonic restraints on the central amino acids (Fig. 5.1b). We ran 10 replicas for each starting state lasting for 100 ns each. H-bond transition rates were obtained as described in[14].

H-bond transition rates depend on the sidechains and the length of the disordered chains adjacent to the H-bond (the free chain length, FCL)[14,15]. We neglected the FCL effect here due to the lack of alignment effects in the lattice model and the short length of the peptides (Fig. 5.6).

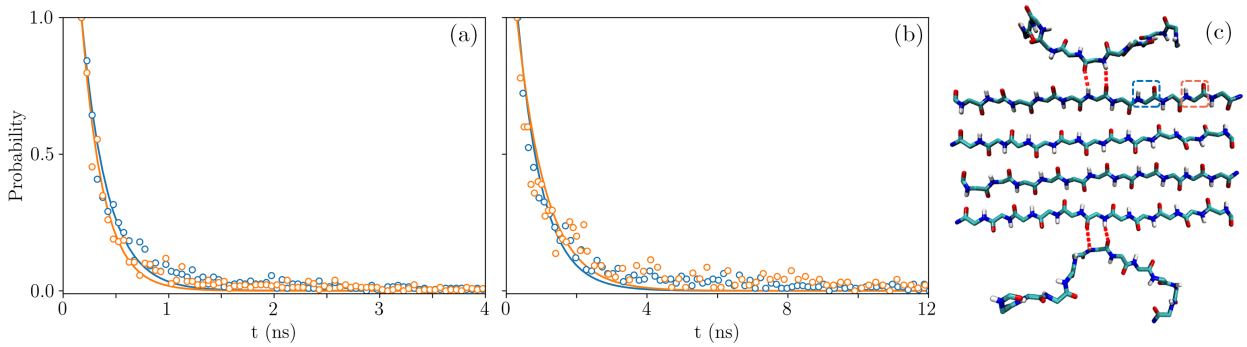

Figure 5.6: The time distributions of the unbound (a) and bound state (b) of the two H-bond pairs adjacent to the retrained bond with different free chain lengths highlighted on (c). They generally follow single exponentials (solid lines).

All simulations were performed with OpenMM 7.3.1[47] using three common force fields, CHARMM36m[41], AMBER99SB-ILDN[48] and AMBER14SB[49], with the TIP3P[50] water model. These force fields are suggested candidates for amyloid peptide assembly based on the study of Aβ_{16-22} dimer in[51]. For all simulations, long-range electrostatic interactions were treated

with particle mesh Ewald (PME), with both direct-space PME and Lennard-Jones potentials making use of a 10 Å cutoff; the Lennard-Jones potential was switched to zero at the cutoff over a switch width of 1.5 Å to ensure continuity of potential and forces. PME used a relative error tolerance of 10^4 at the cutoff to automatically select the α smoothing parameter, and the default algorithm in OpenMM was used to select Fourier grid spacing (which selected a grid spacing of 0.8 Å in each dimension). All bonds to hydrogen were constrained to a within a fractional error of 10^{-8} of the bond distances using CCMA[52,53], and waters were rigidly constrained with SETTLE[54]. OpenMM's long-range analytical dispersion correction was used to avoid pressure artifacts from truncation of the Lennard-Jones potential. Simulations were run at 300 K with a Monte Carlo barostat with 1 atm external pressure and Monte Carlo update interval of 25 steps. Langevin dynamics was used with a 2 fs time steps and collision of 1 ps^{-1}.

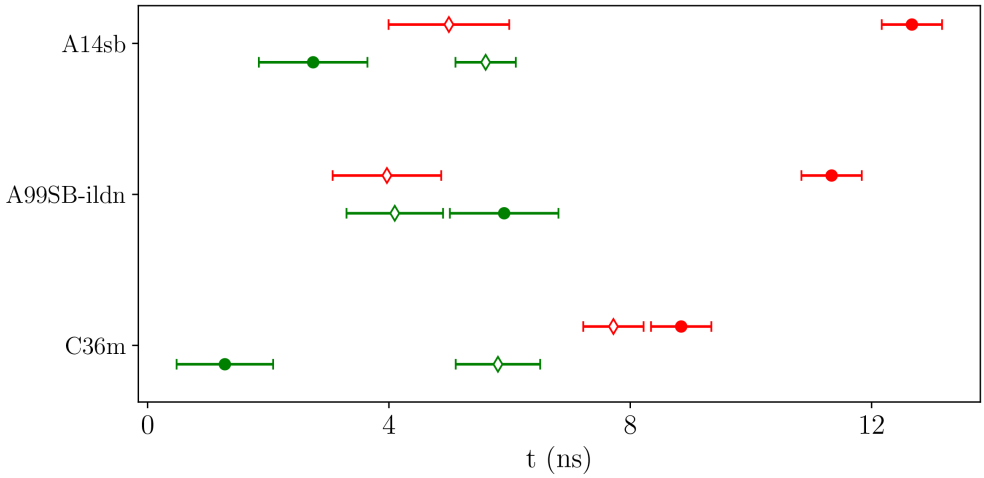

Figure 5.7: Average H-bond pair transtion times associated with three different force fields. The open and closed markers represent the times of unbound and bound state of the H-bond pair respectively. Red indicates strong bonds and green indicates weak bonds.

Coordinates were saved every 5 ps and the rate constants between two states, β-sheet and non-β conformations, were measured using the method in[14]. A transition involves either the formation or breakage of a pair of backbone H-bonds between two residues. The distances between amide hydrogen and carbonyl oxygen pair d_{H-O} were computed using pytraj[55] and monitored for transitions. A backbone H-bond is considered broken when d_{H-O} exceeds 3.5

Å and formed when d_{H-O} is shorter than 2.5 Å. In addition to the 1 Å gap in formation and breakage cutoff distances, a five-frame (25 ps) running average of d_{H-O} and elimination of fast transition less than 300 ps were used to suppress spurious high frequency fluctuations in the detection of backbone H-bond transitions.

We used three force fields to measure the kinetic rate constants. To compare force fields, it is convenient to compute the free energy of the H-bonds using the detailed balance relation

$$\frac{k_{\text{on}}}{k_{\text{off}}} = \mathrm{e}^{-|\epsilon|/k_{\text{B}}T}. \tag{5.7}$$

	C36m	A99SB-ILDN	A14SB
strong bond	−0.14	−1.05	−0.93
weak bond	1.51	−0.37	0.71

Table 5.1: Strong and weak bond free energies calculated from kinetic parameters usin Eq. 5.7. Strong bonds have attractive free energies ranging from 0.14 to 1.05 $k_{\text{B}}T$, while weak bonds range from an attraction of 0.37 $k_{\text{B}}T$ to a repulsion of 1.51 $k_{\text{B}}T$.

Despite the fact that each of these force fields has been favorably evaluated for simulations of polyglutamine[56–60], they yielded widely different results (Table S1). Rather than assess the relative accuracy of these results, we use these different values as representatives of sequences in different regimes of fibril stability. AMBER99SB-ILDN is the most attractive, and in fact, even the weak bonds are net attractive by 0.37 $k_{\text{B}}T$. In this case conformational entropy cannot contribute to the barrier. The least attractive force field is CHARMM36m, which was developed in response to overly compact ensembles in simulations of disordered proteins[41]. CHARMM36m has a strong bond affinity 0.14 $k_{\text{B}}T$, which is too low to observe nucleation in our model and suggests that steric zipper interactions would be necessary to form stable fibrils. AMBER14SB has an intermediate affinity with weak bonds $\epsilon_w = 0.71 k_{\text{B}}T$ that are sufficiently repulsive to create a free energy barrier, yet the strong bonds $\epsilon_s = -0.93 k_{\text{B}}T$ provide enough attraction to stabilize a single-layered β-sheet. Both the AMBER14SB and CHARMM36m force fields are remarkably close to previous studies that estimated 1.86 $k_{\text{B}}T$

for the difference between strong and weak bonds[11–13].

Lattice simulation details

We compute the committor as a function of the number of molecules and the total number of H-bonds using 10000 lattice simulations at each monomer concentration. Each simulation was initiated with a dimer with one H-bond and terminated when there were 15 molecules in the cluster or all H-bonds were broken. The formation and breakage of H-bonds in the dimer are described by *weak* bond rate constants since neither strand is previously in β-conformation. The third molecule brings the possibility for the formation of strong bonds, provided one of the peptides forming the new bond is previously in the β-conformation[11]. A new molecule can attach to either end of the β-sheet when it has at least two backbone H-bonds.

Cluster free energy derivation

The free energy of a nucleus contains three terms:

$$F = E_{\text{bonds}} - \Delta\text{S} + \mu M, \tag{5.8}$$

the first term is the energy form all H-bonds formed in the nucleus, the second term is the energy penalty from the loss of conformational entropy when forming new H-bonds (S is energy coming from the conformational entropy), the last term is the energy required to recruit new molecules (μ is the chemical potential and M is the number of molecules as shown in Eq. 5.1).

Fig. 5.8 shows the number of H-bonds and conformational entropy units in a cluster. Red lines are the H-bonds; one blue open circle represents the restrained region of the backbone around the H-bond when it has another neighbor and two blue open circle is equivalent to one unit of a conformational entropy. We starts form 2 molecules with one contact (one H-bond), then increase the H-bonds with the fixed the number of molecules on the horizontal

direction and add more molecules on the vertical one. By counting the number of H-bonds, the number of chemical potentials and the conformational loss in each configuration of a square nucleus, we can express Eq. 5.8 as

$$F = N\epsilon_{\text{bond}} - (2\ell + M - 2)\text{S} + M(\mu + 2\text{S}). \tag{5.9}$$

Rearrange and absorbed the constant to the chemical potential term we can arrive at Eq. 5.1, where $\gamma = -2\text{S}$, which is the energy difference between a strong bond and a weak bond.

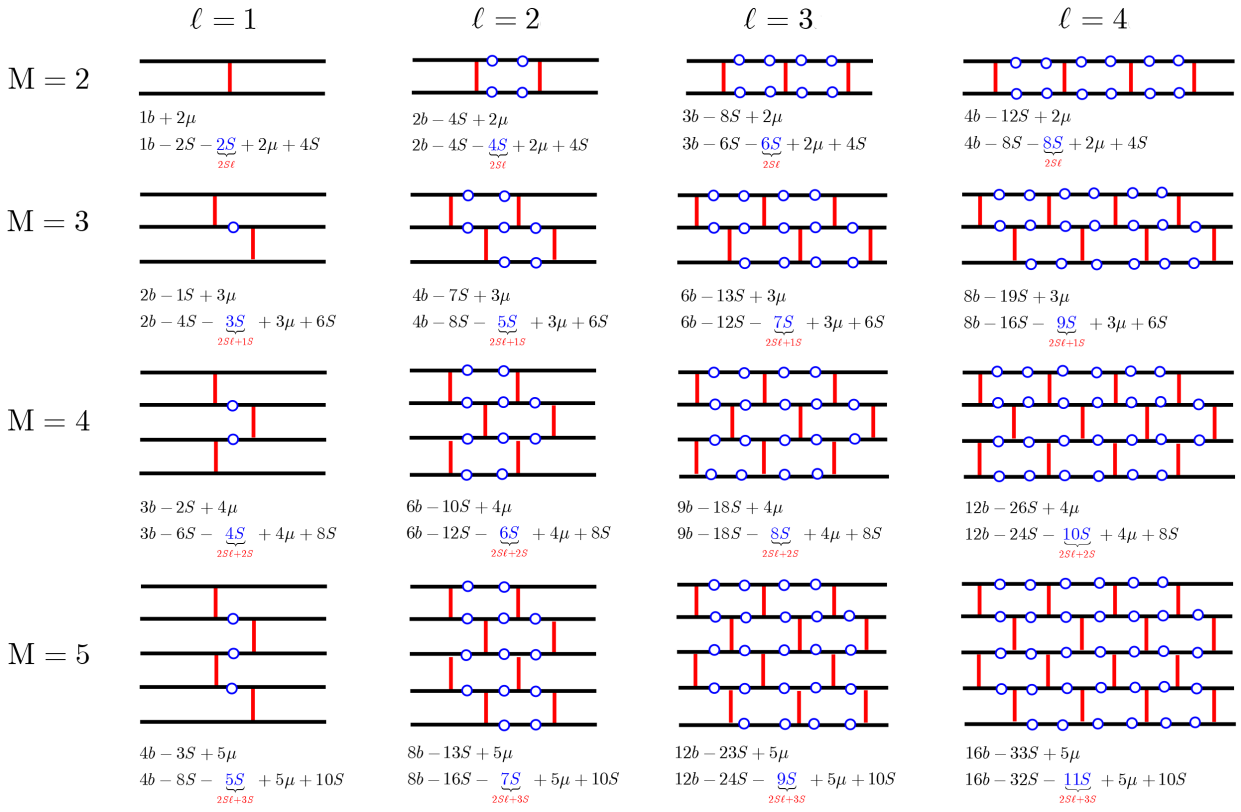

Figure 5.8: Free energy of square clusters with dimensions of M molecules and ℓ H-bonds per molecule. Red lines are the H-bonds with energy b and blue open circles present a conformational entropy unit in the regions stabilized by 2 H-bonds on 2 different rows.

References

[1] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).

[2] D. Erdemir, A. Y. Lee, and A. S. Myerson, Accounts of chemical research **42**, 621 (2009).

[3] J. Chen, E. Zhu, J. Liu, S. Zhang, Z. Lin, X. Duan, H. Heinz, Y. Huang, and J. J. De Yoreo, Science **362**, 1135 (2018).

[4] S. Auer, C. M. Dobson, M. Vendruscolo, and A. Maritan, Physical review letters **101**, 258101 (2008).

[5] J. Zhang and M. Muthukumar, The Journal of chemical physics **130**, 01B610 (2009).

[6] J. A. Luiken and P. G. Bolhuis, Physical Chemistry Chemical Physics **17**, 10556 (2015).

[7] J. A. Luiken and P. G. Bolhuis, The Journal of Physical Chemistry B **119**, 12568 (2015).

[8] F. X. Smit, J. A. Luiken, and P. G. Bolhuis, The Journal of Physical Chemistry B **121**, 3250 (2017).

[9] J. Haaga, J. Gunton, C. N. Buckles, and J. Rickman, The Journal of chemical physics **148**, 045106 (2018).

[10] A. Šarić, Y. C. Chebaro, T. P. Knowles, and D. Frenkel, Proceedings of the National Academy of Sciences **111**, 17869 (2014).

[11] L. Zhang and J. D. Schmit, Physical Review E **93**, 060401 (2016).

[12] L. Zhang and J. D. Schmit, Israel journal of chemistry **57**, 738 (2017).

[13] K. Ghosh and K. Dill, Journal of the American Chemical Society **131**, 2306 (2009).

[14] Z. Jia, A. Beugelsdijk, J. Chen, and J. D. Schmit, The Journal of Physical Chemistry B **121**, 1576 (2017).

[15] Z. Jia, J. D. Schmit, and J. Chen, Proceedings of the National Academy of Sciences **117**, 10322 (2020).

[16] G. Bates, P. S. Harper, and L. Jones, *Huntington's disease*, 45 (Oxford University Press, USA, 2002).

[17] M. Arrasate and S. Finkbeiner, Experimental neurology **238**, 1 (2012).

[18] H. Y. Zoghbi and H. T. Orr, Annual review of neuroscience **23**, 217 (2000).

[19] F. O. Walker, The Lancet **369**, 218 (2007).

[20] K. Kar, M. Jayaraman, B. Sahoo, R. Kodali, and R. Wetzel, Nature structural & molecular biology **18**, 328 (2011).

[21] S. Chen, F. A. Ferrone, and R. Wetzel, Proceedings of the National Academy of sciences **99**, 11884 (2002).

[22] M. Chen, M. Tsai, W. Zheng, and P. G. Wolynes, Journal of the American Chemical Society **138**, 15197 (2016).

[23] M. Chen and P. G. Wolynes, Proceedings of the National Academy of Sciences **114**, 4406 (2017).

[24] D. T. Gillespie, The journal of physical chemistry **81**, 2340 (1977).

[25] D. Punihaole, R. J. Workman, Z. Hong, J. D. Madura, and S. A. Asher, The Journal of Physical Chemistry B **120**, 3012 (2016).

[26] D. Punihaole, R. S. Jakubek, R. J. Workman, L. E. Marbella, P. Campbell, J. D. Madura, and S. A. Asher, The Journal of Physical Chemistry B **121**, 5953 (2017).

[27] A. Šarić, A. K. Buell, G. Meisl, T. C. Michaels, C. M. Dobson, S. Linse, T. P. Knowles, and D. Frenkel, Nature physics **12**, 874 (2016).

[28] J. Krausser, T. P. Knowles, and A. Saric, bioRxiv (2020).

[29] S. J. Bunce, Y. Wang, K. L. Stewart, A. E. Ashcroft, S. E. Radford, C. K. Hall, and A. J. Wilson, Science Advances **5**, eaav8216 (2019).

[30] P. R. ten Wolde and D. Frenkel, Science **277**, 1975 (1997).

[31] P. G. Vekilov, Crystal Growth & Design **4**, 671 (2004).

[32] L. F. Filobelo, O. Galkin, and P. G. Vekilov, The Journal of chemical physics **123**, 014904 (2005).

[33] T. T. Phan and J. D. Schmit, Biophysical Journal **118**, 2989 (2020), ISSN 0006-3495.

[34] J. D. Schmit, K. Ghosh, and K. Dill, Biophysical journal **100**, 450 (2011).

[35] C. F. Lee, J. Loken, L. Jean, D. J. Vaux, et al., Physical Review E **80**, 041906 (2009).

[36] T. M. Phan, S. Whitelam, and J. D. Schmit, Physical Review E **100**, 042114 (2019).

[37] M. R. Sawaya, S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. Wiltzius, H. T. McFarlane, et al., Nature **447**, 453 (2007).

[38] M. R. Jensen and M. Blackledge, Proceedings of the National Academy of Sciences **111**, E1557 (2014).

[39] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. MacKerell Jr, Journal of chemical theory and computation **8**, 3257 (2012).

[40] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, The journal of physical chemistry B **119**, 5113 (2015).

[41] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, Nature methods **14**, 71 (2017).

[42] M. Törnquist, T. C. Michaels, K. Sanagavarapu, X. Yang, G. Meisl, S. I. Cohen, T. P. Knowles, and S. Linse, Chemical Communications **54**, 8667 (2018).

[43] W.-F. Xue, S. W. Homans, and S. E. Radford, Proceedings of the National Academy of Sciences **105**, 8926 (2008).

[44] S. I. Cohen, S. Linse, L. M. Luheshi, E. Hellstrand, D. A. White, L. Rajah, D. E. Otzen, M. Vendruscolo, C. M. Dobson, and T. P. Knowles, Proceedings of the National Academy of Sciences **110**, 9758 (2013).

[45] G. Meisl, X. Yang, E. Hellstrand, B. Frohm, J. B. Kirkegaard, S. I. Cohen, C. M. Dobson, S. Linse, and T. P. Knowles, Proceedings of the National Academy of Sciences **111**, 9384 (2014).

[46] S. I. Cohen, R. Cukalevski, T. C. Michaels, A. Šarić, M. Törnquist, M. Vendruscolo, C. M. Dobson, A. K. Buell, T. P. Knowles, and S. Linse, Nature chemistry **10**, 523 (2018).

[47] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, et al., PLoS computational biology **13**, e1005659 (2017).

[48] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins: Structure, Function, and Bioinformatics **78**, 1950 (2010).

[49] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, Journal of chemical theory and computation **11**, 3696 (2015).

[50] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of chemical physics **79**, 926 (1983).

[51] V. H. Man, X. He, P. Derreumaux, B. Ji, X.-Q. Xie, P. H. Nguyen, and J. Wang, Journal of chemical theory and computation **15**, 1440 (2019).

[52] E. Barth, K. Kuczera, B. Leimkuhler, and R. D. Skeel, Journal of computational chemistry **16**, 1192 (1995).

[53] P. Eastman and V. S. Pande, Journal of chemical theory and computation **6**, 434 (2010).

[54] S. Miyamoto and P. A. Kollman, Journal of computational chemistry **13**, 952 (1992).

[55] D. R. Roe and T. E. Cheatham III, Journal of chemical theory and computation **9**, 3084 (2013).

[56] A. M. Fluitt and J. J. de Pablo, Biophysical journal **109**, 1009 (2015).

[57] J. Wen, D. R. Scoles, and J. C. Facelli, PloS one **12** (2017).

[58] À. Gómez-Sicilia, M. Sikora, M. Cieplak, and M. Carrión-Vázquez, PLoS computational biology **11** (2015).

[59] S. Cote, G. Wei, and N. Mousseau, The journal of physical chemistry B **116**, 12168 (2012).

[60] E. Starikov, H. Lehrach, and E. Wanker, Journal of Biomolecular Structure and Dynamics **17**, 409 (1999).

Chapter 6

Entropic and energetic contributions to biomolecule condensate surface tension

6.1 Introduction

Numerous cellular compartments form by the spontaneous condensation of biomolecules in a process that resembles liquid-liquid phase separation (LLPS). Although these structures may contain hundreds of molecular components, often only a few are needed to recapitulate condensation *in vitro*. These "scaffolds" are usually multivalent, polymer-like molecules that drive condensation by forming many weak intermolecular interactions[1]. While early work focused on the utility of these condensed states as a compartmentalization mechanism, more recently it has become apparent that these intermolecular interactions are responsible for creating essential structure on the atomic[2,3], network[4–6], and organelle scales[7,8]. In particular, many liquid organelles have a multi-layered organization with separate compartments maintained by a hierarchy of surface tension[7].

The utility of surface tension in cellular organization highlights the need to understand the mechanism of surface tension in multi-component phase-separated structures. A useful

point of reference is the simplest case of attractive spheres. In this adhesion-driven, single-component system the surface tension arises from the fact that particles on the surface of a cluster have fewer attractive interactions to compensate for the translational entropy penalty that comes from condensation. Thus, the surface tension can be considered the result of unsatisfied bonds on the cluster surface.

A contrasting case is that of condensates formed from binary mixtures of SPOP and DAXX[5,9]. Under conditions where SPOP is present at stoichiometric excess, DAXX drives condensation by forming cross-links between rod-like SPOP assemblies. However, there is minimal change in the interaction energy upon condensation because DAXX is flexible enough to satisfy its binding sites on a single SPOP assembly when cross-linking partners are not available[5]. Instead, the driving force for condensation arises from the configurational entropy of binding. This is because the are many more ways for DAXX to satisfy its binding sites in a dense SPOP condensate compared to a dilute solution. From this example, we can extract two important lessons. First, when modeling the surface tension of biomolecule condensates it is important to account for molecular entropy, not just binding energy. Second, the balance between entropic and energetic driving forces, which is controlled by the network connectivity, provides a complementary mechanism to hydrophobic/hydrophilic content to establish discrete liquid compartments.

Here we employ a combination of lattice simulations and analytic theory to understand how network connectivity affects the mechanism of attraction and the surface tension. We study models ranging from a single component system that associates by purely energetic nearest-neighbor interactions, to a two component system that mimics the entropy dominated mechanism of SPOP/DAXX. Between these limits is an intermediate case of a two-component system with an energy-dominated attraction mechanism. We find three contributions to the surface tension. The first is an entropy dominated mechanism appearing in all three systems that results from the reduced translational entropy of molecules near the surface of the condensate. The second is an energy dominated mechanism, similar to the missing bonds between attractive spheres, that appears only in the purely energy-driven system. This mechanism is absent in the two component systems where it is replaced by

a purely entropic mechanism resulting from the reduced number of binding partners at the periphery of the condensate.

6.2 Model

We study the association of rigid rods on a lattice inspired by recent experimental results on the SPOP/DAXX and actin/fibrilarin systems[9,10]. SPOP and actin both polymerize to form rod-like assemblies that are rigid on the polymerization scale explored experimentally. For our purposes the rod systems are convenient because they break the system symmetry, which allows us to identify multiple contributions to the surface tension. We explore three attraction mechanisms to drive condensation of the rods.

1) In the simplest version of the model the rods have attractive interactions of affinity ϵ_i between adjacent sites on the lattice. This attraction mechanism does not include end-to-end interactions. This model provides an easily calculated baseline and allows us to identify effects arising solely from the rod geometry (Fig. 6.1a).

2) In the second model there is no direct interaction between the rods apart from hard-core repulsion. Instead, interaction is mediated by cross-linking molecules. These molecules occupy two adjacent sites on the lattice in the x, y, or z direction. Each lattice site can be in one of three states: a dual occupancy state with both a rod and cross-linker subunit, which is assigned an energy ϵ_s, and zero energy states where the site contains only a rod subunit, only a cross-linker subunit, or is empty. To account for excluded volume, dual occupancy by the same type of subunit is not allowed. This model is inspired by the SPOP/DAXX system in which the DAXX cross-linkers have the flexibility to find binding partners in all directions, including multiple sites on the same rod (Fig. 6.1b).

3) The third model is identical to the second except that cross-linker molecules are only allowed to extend in the x and y directions, perpendicular to the rods which are oriented in the z direction. This mimics a system with rigid cross-linkers that extend away from the rods. We label the binding energy in this model ϵ_a, reflecting the actin system that inspired it (Fig. 6.1c).

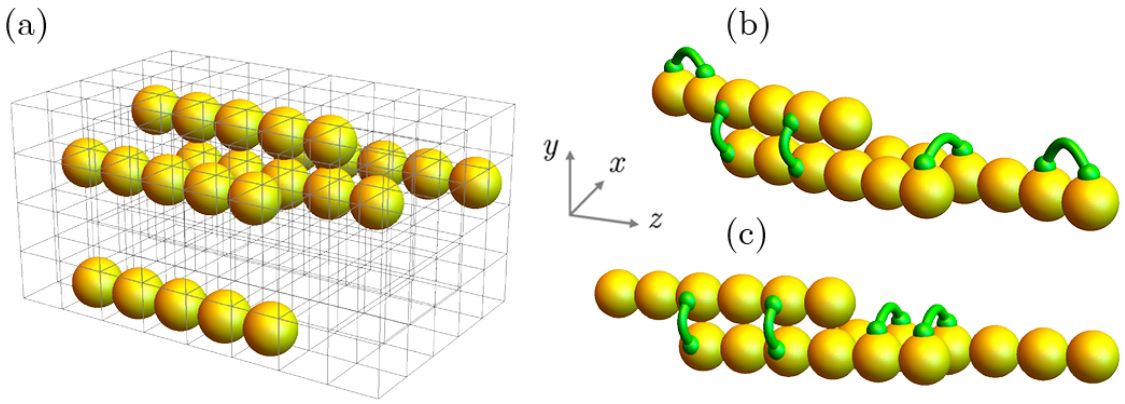

Figure 6.1: Three attraction models to drive condensation of rods: (a) implicit cross-linking model in which each rod orienting on the z direction contains ℓ yellow spheres, (b) SPOP/DAXX model with the presence of cross-linking molecules in green, occupying two adjacent sites on the lattice in the x, y, or z direction, and (c) Actin/Filamin model in which cross-linking molecules are only allowed to extend in the x and y direction, perpendicular to the rods.

To model the assemblies of the simplest case, we compute the configuration entropy using a lattice model with energy,

$$\frac{F}{k_{\mathrm{B}}T} = \epsilon N L z \rho - \ln \Omega, \tag{6.1}$$

where ϵ is the attractive energy of each unit in a rod, z is a coordination number, N is the number of rods, L is the length of of rods, ρ is the density of the assembly, and Ω is number of ways to arrange N rods and m voids between the rods on a row (Fig. 6.2). The first term is a mean field binding energy and the second term accounts for the configuration entropy. $\ln \Omega$ can be expressed as

$$\begin{aligned} \ln \Omega &= \ln \frac{(N+m)!}{N!m!} \\ &= \ln(M+m)! - \ln M! - \ln m! \\ &= (N+m)\ln(N+m) - N \ln N - m \ln m \\ &= N \ln\left(1 + \frac{m}{N}\right) + m \ln\left(1 + \frac{N}{m}\right). \end{aligned} \tag{6.2}$$

The surface tension of this solid assembly can be approximated by using the pinning effect, which prevents the rods from protruding from the surface, and accounting for the

missing bonds on the surface of the assembly. Eq. 6.1 can be rewritten as

$$\frac{F}{k_{\mathrm{B}}T} = \epsilon NLz\rho - (N-2)\ln\left(1+\frac{m}{N}\right) - m\ln\left(1+\frac{N}{m}\right) - 2\ln\left(1+\frac{m}{N}\right), \tag{6.3}$$

The last term $2\ln(1+m/N)$ in Eq. 6.3 accounts for the pinning effect by assuming the two outermost rods are fixed at the tips of the assembly. This leads to the gain of translational entropy of the pinned rod (red rod at each tip of the assembly in Fig. 6.2), $\ln(2\delta+1)$ where δ is overhang length (Fig. 6.3). Therefore, the contribution due to the pinning effect to the surface tension is $\ln(1+m/N)+\ln(2\delta+1)$. Another contribution to the surface tension is

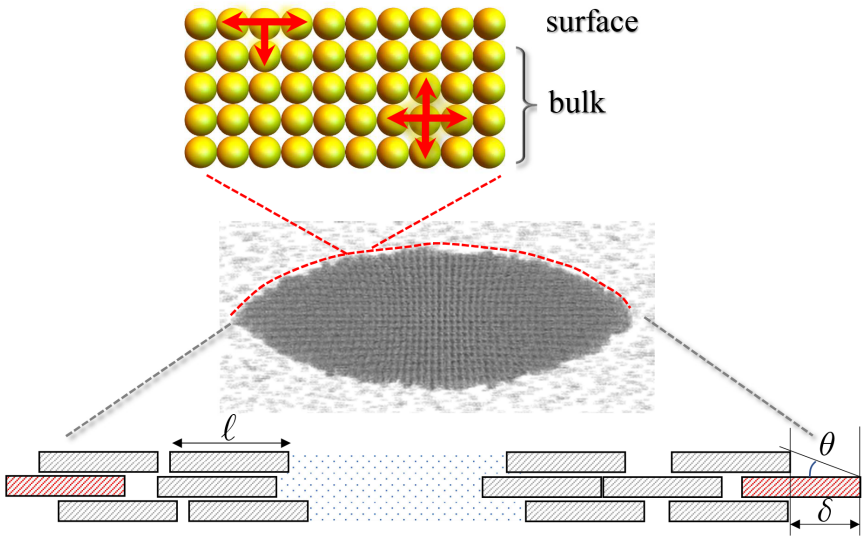

Figure 6.2: Schematic representative of missing bond mechanism at the surface boundary and the pinning effect. In the middle is a snapshot of the spindle-like cluster from the simulation. On the top illustrates the missing bond mechanism, in which the surface particles have one nearest neighbor less than the bulk particles because they have one side exposed at the surface. At the bottom shows the pinning effect, where the two outermost rods are stationary.

approximated by accounting for the missing bonds at the surface. A surface is defined as the boundary between the condensed phase and a solution of vapor. *Surface tension* is the energy cost of increasing the surface area of the system[11]. When the droplet changes shape, its surface gets larger relative to its volume. Water tends to form spherical droplets because deviations away from spherical shapes are opposed by the surface tension. Surface particles have one nearest neighbor less than the bulk particles due to being exposed to a different

environment at the surface. To account for this, we use the energetic term $\epsilon_i \rho \delta$, where ϵ_i (here and subsequently we work in the units such that $k_B T = 1$) is a favorable binding energy and $\rho = N\ell/(N\ell + m)$ is the density of the assembly. With these contributions, the surface energy per overhang per surface area, $a\sqrt{a^2 + \delta^2}$ where a is dimension of a unit cell in the lattice or the diameter of the yellow spheres in Fig. 6.1 ($a = 1$ for simplicity), can be expressed as

$$\sigma(\delta) = \frac{\delta \epsilon_i \rho + \ln(1 + m/N) + \ln(2\delta + 1)}{\sqrt{1 + \delta^2}}. \tag{6.4}$$

The first and the second terms in the numerator account for the energetic and the entropic contributions to the surface energy, respectively. The third term is the correction due to the pinning effect. Eq. 6.4 can be rewritten in terms of assembly density ρ and the angle θ using the geometry shown in Fig. 6.2, $\tan\theta = 1/\delta$,

$$\sigma(\theta) = \left(\epsilon \rho \tan^{-1}\theta + \ln[1 + \ell(\rho^{-1} - 1)] + \ln(2\tan^{-1}\theta + 1)\right) \sin\theta. \tag{6.5}$$

This expression will be used to compare with the simulations.

6.3 Results and Discussion

6.3.1 Coarse-grained simulations

The analytic theory will be complimented with two sets of coarse-grained simulations. Lattice simulations are conducted using the Monte Carlo algorithm. We consider each molecule as a rod-like particle of length ℓ with each representing an occupied site in the lattice. The first set of simulations treats the cross-linking interaction implicitly. Rods are treated in the canonical ensemble, contacts between nearest neighbors contribute a favorable binding energy $-\epsilon_i < 0$. Fig. 6.3a shows an implicit cross-linking simulation box, where the rods align along the z-direction and can move along all six directions. In the second set of simulations, the cross-linking molecules are treated in the grand canonical ensemble, but the number of rods are fixed. The cross-linkers can add to or leave the lattice satisfying

the *detail balance* (discussed in section 3.2). The interactions between nearest neighbors are only allowed when the two ends of the same cross-linker connect these sites (Figs. 6.1b,c). These contacts contribute a favorable energy $-\epsilon_j < 0$. The cross-linkers can orient in 6 directions and 4 directions in the SPOP/DAXX model (Fig. 6.1b) and Actin/Filamin model (Fig. 6.1c), respectively. In the latter model, the cross-linkers are not allowed to orient in the same direction of the alignment of the rod-like particles. Both simulations will employ Monte Carlo (MC) sampling with rod-like particles treated in the canonical ensemble and cross-linking particles treated in the grand canonical respectively.

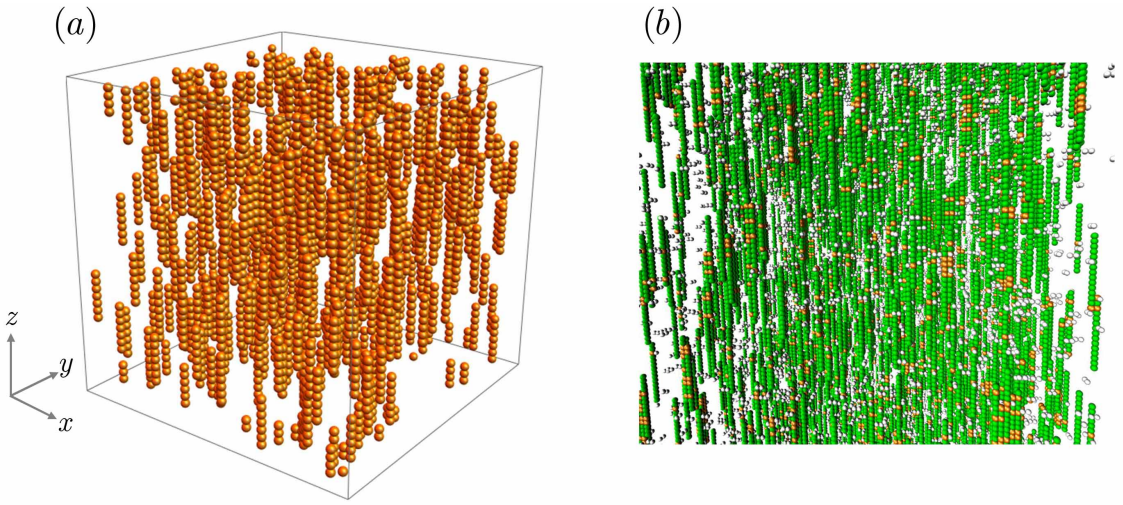

Figure 6.3: (a) Implicit cross-linking simulation box. The rod-like particles orient in the z-direction; they can translate in all six directions and interact with the nearest neighbors in the simulation box (end-to-end interactions are not allowed). (b) Snapshot of the box in the explicit cross-linking simulation. A cross linker has two white ends, which can orient in 6 directions (SPOP/DAXX) or 4 directions (Actin/Filamin). Rod-like particles are in green and they do not interact with each other unless the cross-linkers connect them (the orange units show the sites of the rods occupied by one end of the cross-linkers)

6.3.2 Implicit cross-linking simulations provide microscopic properties of equilibrium assembly

We carried out the implicit cross-linking simulations in a three dimensional (3D) cubic lattice of $300 \times 300 \times 650$ sites and periodic boundary conditions applied in all directions. Fig. 6.4a shows a snapshot of an equilibrium elongated assembly. The similarity to the shape in

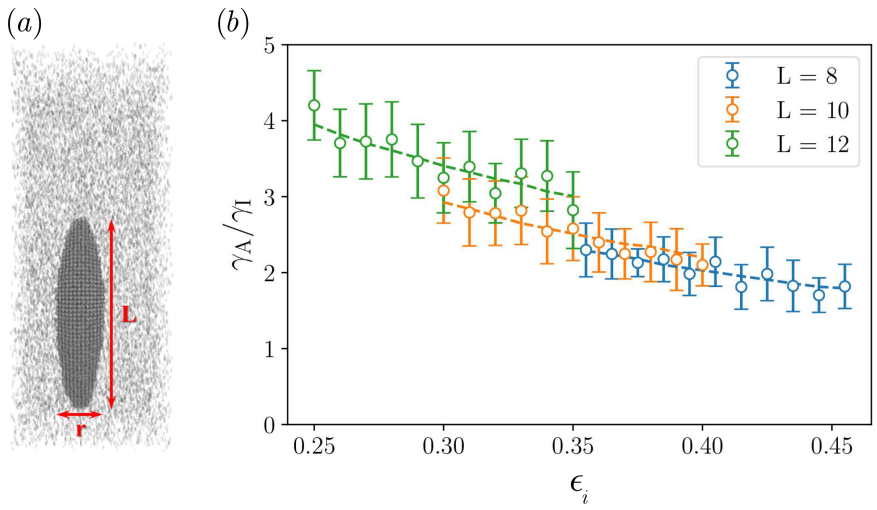

Figure 6.4: (a) Major and minor axis lengths of an aseembly in a snapshot taken from the implicit cross-linking simulation. (b) The ratio of anisotropic to isotropic surface tension at various the binding energies and rod lengths. Open circles with errorbars show the simulation data calculated from Eq. 6.4 and dashed lines are the predictions from analytic theory.

Weirich et al.[10] suggests that the model has the physics necessary to replicate the experiments. The equilibrium aspect ratios of the assemblies, L/r (Fig. 6.4a) where L and r are the major and minor axes lengths respectively, were measured in the simulations. We can use these ratios to extract the isotropic and anisotropic contributions to the surface tension using the continuum theory in Refs. Weirich et al.[10], Prinsen and van der Schoot[12],

$$\frac{L}{r} = \begin{cases} 2\omega^{1/2}, & \text{if } \omega \geq 1 \\ 1+\omega, & \text{if } 0 \leq \omega < 1 \end{cases} \tag{6.6}$$

where ω is the ratio of the anisotropic to isotropic surface tension, γ_A/γ_I. The anisotropic contribution arises from the elongated shape of the rods while the isotropic contribution comes from the interactions. At a specific rod length, increasing binding energy helps the system to nucleate and reach equilibrium faster. This leads to the decreasing of the aspect ratio and the surface tension ratio while at lower binding affinity requires more binding sites at the surface which leads to more elongated shape. Since implicit cross-linking is a pure energy driven system, shorter rods require more binding energy per unit than longer

rods in order to form the assemblies. These trends are shown in Fig. 6.4b, plotting the surface tension ratio as a function of binding energy with differing rod lengths. For each binding energy, at the end of the simulation we used the Hoshen-Kopelman algorithm[13] (see Appendix) to find the largest cluster and measure the the major and minor axes lengths. To compare with the simulations, the ratio of anisotropic surface tension to the isotropic surface tension in our analytic can be approximately written as

$$\frac{\gamma_A}{\gamma_I} \approx \frac{\sigma(\theta_m) + \sigma(\theta_0)}{\sigma(\theta_0)} \tag{6.7}$$

where $\theta_m \approx \pi/2$ and $\theta_0 \approx \tan^{-1}(1/\ell)$ are the angle near the tip ($\delta \ll 1$) and the small angle at the side ($\delta \simeq \ell$) of the assembly, respectively. This approximation is in good agreement with the simulation data for elongated assemblies with sharp tips. For assemblies with more rounded tips, we need a more rigorous calculation, involving functional minimization of Eq. 6.1.

6.3.3 Contribution of cross-linking entropy to the surface tension

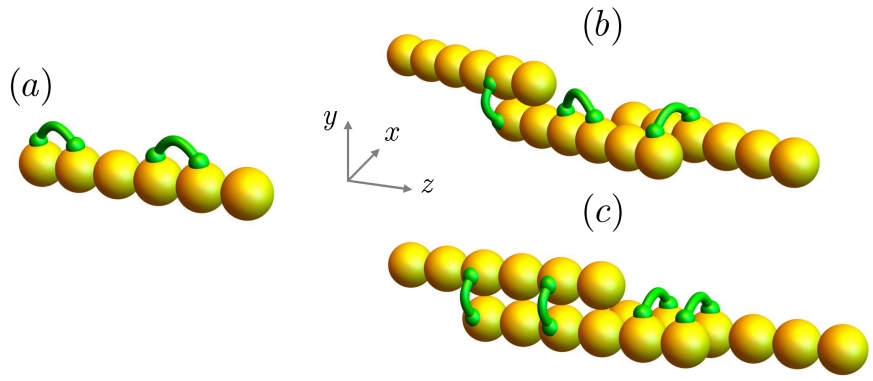

Figure 6.5: (a, b) show double bound molecules in SPOP/DAXX model in vapor-like phase and condensed phase and (c) presents the connectivity in Actin/Filamin model.

Accounting for the "missing bond" mechanism on the surface gives good approximation for the surface tension of many solid assemblies. However, a recent study[5] shows that this approach will not work for biomolecular condensates whose the driving force for condensation

is driven by the binding configuration *entropy*, with only a small change in the binding energy. Here we use two explicit cross-linking models to study the entropic contributions to the surface tension. The resulting structures in Actin/Filamin system studied in Weirich et al.[10] are morphologically similar to the SPOP/DAXX gel. However, they fundamentally have different driving force mechanisms. The latter is dominated by an entropy-driven mechanism since the cross-linkers can orient in all directions while the former is energy-driven as the cross-linkers are only allowed to extend in the directions perpendicular to the rods to connect with nearby neighbors (Fig. 6.5c). Fig. 6.5a,b show the double bound rods in vapor-like phase and in condensed phase. There is no change in binding energy in these two phases. The driving force, thus, comes from the number of ways of arranging cross-linkers on the rods. According to a recent study[5], the minimized free energy of the system with respect to the site occupancies is

$$f(z) = \ln(n_0) + n_2 \tag{6.8}$$

where z is the coordinate number describing the nearby binding sites that a cross-linker can reach, n_0 is the number of free binding sites and n_2 is the number of double bound molecules,

$$n_0 = \frac{-(c\mathrm{e}^{\epsilon_j}+1)+\sqrt{(c\mathrm{e}^{\epsilon_j}+1)^2+8zc\mathrm{e}^{2\epsilon_j}}}{4zc\mathrm{e}^{2\epsilon_j}}, \tag{6.9}$$

$$n_2 = zc\mathrm{e}^{2\epsilon_j}n_0^2. \tag{6.10}$$

Both n_0 and n_2 depend on the binding energy ϵ_j ($j = s$ for SPOP/DAXX and $j = a$ for Actin/Filamin) and concentration c of the cross-linkers. The free energy difference between condensed and gas phases comes from the increase entropy of organizing cross-linkers on the binding sites of rods and can be written (from Eq. 6.8) as

$$\Delta f = f(z_c) - f(z_v) = \ln\frac{n_0(z_c)}{n_0(z_v)} + n_2(z_c) - n_2(z_v), \tag{6.11}$$

where (z_c, z_v) are the number of nearby binding sites in the condensed phase and the va-

por phase respectively; $(z_g, z_v) = (6, 2)$ for SPOP/DAXX model and $(z_g, z_v) = (4, 0)$ for Actin/Filamin model. This free energy change from the increased entropy is equivalent to the bond density in the implicit cross-linking model.

$$\Delta f \approx (z_c - z_v)\rho\frac{\epsilon_i}{2} \tag{6.12}$$

Eqs. 6.11, 6.12 can be used to find the binding energies ϵ_s and ϵ_a in the SPOP/DAXX and Actin/Filamin models, respectively, at a specific cross-linker concentration that corresponds to a particular implicit binding energy. Note that in Actin/Filamin model, the cross-linkers are not allowed to align along the orientation of the rods and thus the double bound molecules are excluded in Eq. 6.11 ($n_2 = 0$) in vapor phase. At a fixed corresponding implicit binding energy, the binding entropy in SPOP/DAXX model is larger than then binding energy in the Actin/Filamin model due to the extra entropy as shown in Fig. 6.6a. In explicit cross-linking models, the pinning effect contributes to the surface tension in the same way as in the implicit cross-linking model while the missing bonds on the surface boundary does not work the same. Instead, it depends on the arrangement of cross-linkers on the rods to make fully bound bonds with neighboring rods. In other words, the missing bond mechanism is replaced by the cross-linker binding entropy. Thus, the free energy can be either energy or entropy dominated.

Fig. 6.6b shows comparison of the aspect ratio, L/r, measured in the implicit and explicit cross-linking simulations at various implicit binding energies. The simulation box and the number of rods are the same for all simulations. In the explicit cross-linking simulations, the binding energies ϵ_s, ϵ_a are computed from the corresponding ϵ_i using Eqs. 6.11 and 6.12 at cross-linker concentration $c = 0.018$. Each data point was averaged from 10 replicas in the implicit simulation and 5 replicas in the explicit simulation since it is more expensive. There is a good correlation between the aspect ratios in the SPOP/DAXX and implicit cross-linking simulations. The Actin/Filamin model has weaker bind energies (Fig. 6.6a) and thus there are more binding sites on the surface, which leads to more elongated assemblies and longer aspect ratios.

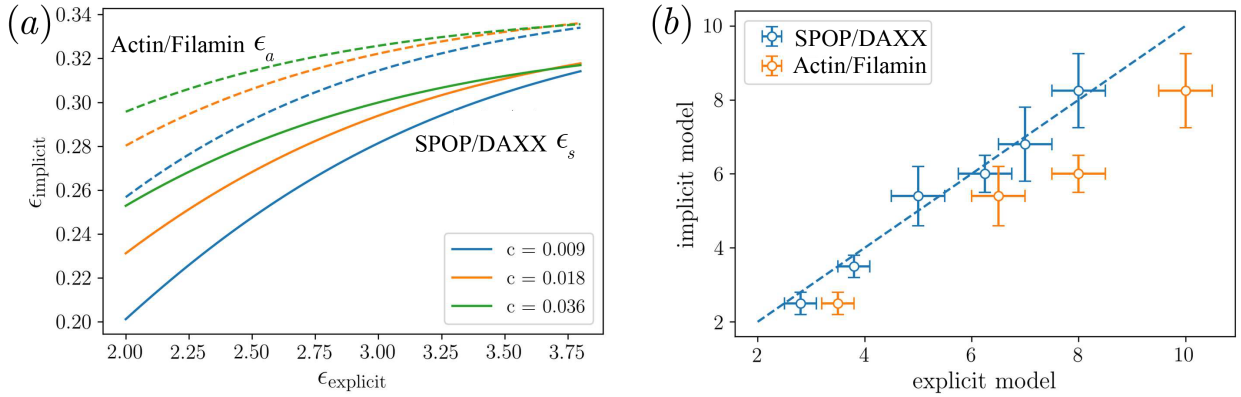

Figure 6.6: (a) Plot from Eqs. 6.11 and 6.12 showing the mapping of binding energies from implicit model to SPOP/DAXX (ϵ_s) and Actin/Filamin (ϵ_a) models. (b) Comparison of aspect ration, L/r, among all three models. All simulations were performed in the 3D cubic box of $200 \times 200 \times 500$ with the same rod length $\ell = 10$. The binding energies used in the implicit cross-linking simulations are $0.295, 0.3, 0.31, 0.32, 0.33, 0.34$, which were used to computed the binding energies for the explict cross-linking models at cross-linker concentration $c = 0.018$ via Eqs. 6.11 and 6.12

6.4 Conclusion

Model	Driving force	End surface tension	Side bonds	Side surface tension
Implicit	**Energy**	Rod pinning	Missing	**Energy**
SPOP/DAXX	**Entropy**	Rod pinning	Fully bound	**Entropy**
Actin/Filamin	**Energy**	Rod pinning	Fully bound	**Entropy**

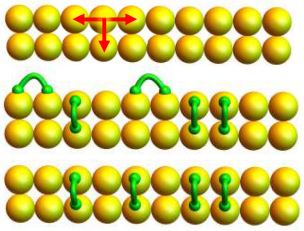

Figure 6.7: A summary shows the contributions to the surface tension in three model systems with different driving forces. All three models have the same pinning effect at the end of the assembly. For the side surface tension (at the surface boundary), in implicit cross-linking model the contribution is purely energetic (coming from the missing bonds) while in the explcit cross-linking models this contribution is dominated by the cofiguration entropy.

Our results highlight two contributions to the surface tension: (1) the missing bonds, which is an energy dominated mechanism appearing in purely energy-driven systems, (2) the entropy dominated mechanism arising from the reduced translational entropy of molecules at the surface boundary. We also find that these contributions do not depend on the driving force of the system as shown in the summary in Fig. 6.7.

Appendix

The Hoshen – Kopelman algorithm

In order to measure the major and minor axes lengths of the assembly, we need to identify the clusters and determine the largest cluster in the system. The Hoshen – Kopelman (HK) algorithm[13,14] is simple and very fast 'single-pass' routine for labeling cluster on a grid, which is a network of cells where each cell maybe either "occupied" or "unoccupied". The basic idea of this algorithm is scan through the grid looking for occupied cells and label them. If the cell is occupied, it needs to be labeled with a cluster label, which depends on its left and top neighbors. If the cell has no occupied neighbors, a new label is then assigned to it. If the cell has one occupied neighbor, then it is assigned to the same label as its occupied neighbor's. If the cell has more than one occupied neighbors, then we choose the lowest-numbered cluster label of the occupied neighbors to assign to the current cell. Because these neighboring cells have different labels, we add a note to ensure these labels correspond to the same cluster. For example, in Fig. 6.8 after the scanning on the third row has completed, the site with the question mark should be assigned to label **1**, and label **1** and **2** belong to the same cluster. The HK algorithm is a special application of the well-known Union-Find algorithm[15] used in computer science.

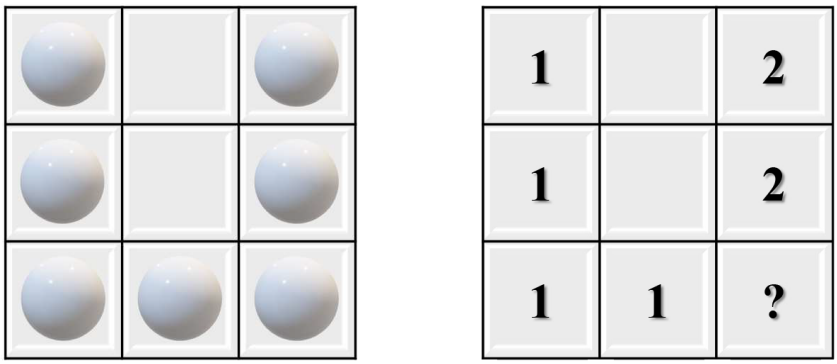

Figure 6.8: Labeling of cluster on a square lattice. The question mark shows the "conflict" when two occupied neighboring cells have different labels.

References

[1] S. F. Banani, A. M. Rice, W. B. Peeples, Y. Lin, S. Jain, R. Parker, and M. K. Rosen, Cell **166**, 651 (2016).

[2] J. Wang, J.-M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, et al., Cell **174**, 688 (2018).

[3] E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, and T. Mittag, Science **367**, 694 (2020).

[4] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, elife **6**, e30294 (2017).

[5] J. D. Schmit, J. J. Bouchard, E. W. Martin, and T. Mittag, Journal of the American Chemical Society **142**, 874 (2019).

[6] K. Bhandari, M. A. Cotten, J. Kim, M. K. Rosen, and J. D. Schmit, The Journal of Physical Chemistry B **125**, 467 (2021).

[7] M. Feric, N. Vaidya, T. S. Harmon, D. M. Mitrea, L. Zhu, T. M. Richardson, R. W. Kriwacki, R. V. Pappu, and C. P. Brangwynne, Cell **165**, 1686 (2016).

[8] Y. Shin and C. P. Brangwynne, Science **357** (2017).

[9] J. J. Bouchard, J. H. Otero, D. C. Scott, E. Szulc, E. W. Martin, N. Sabri, D. Granata, M. R. Marzahn, K. Lindorff-Larsen, X. Salvatella, et al., Molecular cell **72**, 19 (2018).

[10] K. L. Weirich, S. Banerjee, K. Dasbiswas, T. A. Witten, S. Vaikuntanathan, and M. L. Gardel, Proceedings of the National Academy of Sciences **114**, 2131 (2017).

[11] K. A. Dill, S. Bromberg, and D. Stigter, *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience* (Garland Science, 2010).

[12] P. Prinsen and P. van der Schoot, Physical Review E **68**, 021701 (2003).

[13] J. Hoshen and R. Kopelman, Physical Review B **14**, 3438 (1976).

[14] D. P. Landau and K. Binder, *A guide to Monte Carlo simulations in statistical physics* (Cambridge university press, 2014).

[15] R. Sedgewick and K. Wayne, *Algorithms* (Addison-wesley professional, 2011).

Chapter 7

Conclusions and Future Directions

This dissertation has revealed that computer simulation together with analytic theory can serve a role as a microscope to explore the spatiotemporal resolution of biomolecular systems that is often difficult to access experimentally. In chapters 4 and 5, we also showed that the self-assembly processes often share a fundamental characteristic, an energy barrier that the system needs to surmount to enter the thermodynamically favorable regime. Classical nucleation theory (CNT) has successfully captured the behaviors of these processes although it has some shortcomings as shown in Section 2.4. In chapter 4, our computer simulations and simple scaling arguments based on CNT described a good aspect of impurities in accelerating the growth layer-by-layer under conditions of low supersaturation and low temperature. This mechanism may provide a good explanation for protein crystallization using non-specific binding enhancers[1–3]. In chapter 5, we applied CNT to the nucleus that has anisotropic line tensions and used simple lattice simulations to show that the contribution of conformational entropy to the energy barrier can be reduced by limiting the extent of secondary structure in the cluster, which is highly dependent on the concentration of free protein. In chapter 6, we focused on the contributions of both energy and entropy to the surface tension, one of the two important terms in the free energy in CNT, in biomolecule condensates. We showed that the missing bond mechanism is dominated in energy-driven systems but it is replaced by an entropy-dominated mechanism when the system is mainly driven by binding entropy

and that these mechanisms do not depend on the driving force of the system.

There are also other directions that we can continue to explore the problems considered in this dissertation. In chapter 4, we consider proteins to sample just two states (binding and misbinding) but in protein crystal growth protein must sample an ensemble of 10^4 to 10^5 states to find a crystallographic state. Our simple scaling argument can easily apply to this regime but the simulations are expensive and may require longer CPU runtime to access the growth regimes. Pilot runs are needed to explore the parameter windows in which impurities are beneficial. In chapter 5, we generated the nucleation trajectories using a lattice Markov State Model developed on a single layer of β-sheet (2D). However, amyloid-like fibrils of different proteins have a common 3D structural cross-β spine[4]. In order to extend our current model to 3D model, we need the side chain interactions between beta-sheets. This can be achieved by adding another β-sheet layer to our current configurations (Figs. 5.1a,b) when measuring the rate constants of beta-sheet formation. The free energy difference between the two cases describe the effect of sidechain interactions in the 3D model. In chapter 6, the agreement when comparing the ratio γ_A/γ_I in the simulations to our analytic theory using the approximation in Eq. 6.1 (Section 6.3.2) is just applied for elongated assemblies with sharp tips. For general shapes, we need a more rigorous calculation involving functional free energy minimization with respect to the aspect ratio of assembly.

References

[1] P. G. Vekilov and J. I. D. Alexander, Chemical reviews **100**, 2061 (2000).

[2] R. Giegé, The FEBS journal **280**, 6456 (2013).

[3] S. D. Durbin and G. Feher, Annual Review of Physical Chemistry **47**, 171 (1996).

[4] J. D. Sipe and A. S. Cohen, Journal of structural biology **130**, 88 (2000).