Protein Folding and Peptide Aggregation Simulations on the Face-Centered Cubic Lattice via the Replica-Exchange Wang-Landau Algorithm

by

Matthew S. Wilson

(Under the Direction of David P. Landau)

Abstract

Protein folding and amyloid protofibril formation of short peptides are studied through advanced Monte Carlo simulations of coarse-grained models on the face-centered cubic (fcc) and simple cubic (sc) lattices. Interacting self-avoiding walks (ISAW), hydrophobic-polar (HP), and hydrophobic-neutral-polar (H0P) models are used, where amino acids are classified into one, two, or three hydrophobicity groups, respectively. We extend the sc and fcc lattice simulations to handle multiple interacting H0P sequences, and apply them to model amyloid protofibril formation. The Wang-Landau, replica-exchange Wang-Landau, and multicanonical sampling algorithms are utilized to determine the density of states for each system with high precision. These methods are used to predict ground state energies and structures, along with average thermodynamic and structural properties for a set of biologically motivated HP model protein sequences on the fcc lattice. A comparison with analogous results for the sc lattice shows similar folding thermodynamics on both lattice geometries, with one longer HP sequence on the fcc lattice showing additional, subtle structural rearrangements after the initial hydrophobic collapse. Using the ISAW model as a testing ground, we characterize the thermodynamics of a four-body potential that success-

fully promotes the formation of helical bundles in fcc backbone models, which is then briefly tested for a H0P model of the Crambin protein to show a better agreement with experimental results. Finally, the formation of multi-layered protofibril structures is examined with the sc lattice model through calculation of the specific heat, cluster size distribution, and nucleation free energy barriers. We observe a two-step protofibril formation that involves both condensation and ordering processes, with a conformational rearrangement of peptide structures occurring after condensation in the case of weak intermolecular attraction. The aggregation model is used on the fcc lattice to make a comparison with the sc results for a small test system, and the predicted protofibril structures are presented for the fcc geometry.

INDEX WORDS: Monte Carlo simulation, Protein folding, Amyloid protofibril, Lattice protein model, Statistical mechanics, Replica-Exchange Wang-Landau

Protein Folding and Peptide Aggregation Simulations on the Face-Centered Cubic Lattice via the Replica-Exchange Wang-Landau Algorithm

by

Matthew S. Wilson

B.S., University of Georgia, 2015

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

Doctor of Philosophy

Athens, Georgia

2021

Protein Folding and Peptide Aggregation Simulations on the Face-Centered Cubic Lattice via the Replica-Exchange Wang-Landau Algorithm

by

Matthew S. Wilson

Major Professor: David P. Landau

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

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2021

Acknowledgments

I would like to sincerely thank my major advisor, Dr. David Landau, for his continuing support and patience throughout my time at the Center for Simulational Physics. He has taught me so much about the intriguing world of physics and computer simulation, and has given me every opportunity to grow as a scientist and person. I have had a great time under his supervision, and am ever grateful for all that he has taught me about science and life.

I would like to thank all of my committee members: Dr. Michael Bachmann, for his insights and for including me in the Soft Matter Research Group events, Dr. Steven Lewis, for the constant support and always being able to cheer me up, and Dr. Shan-Ho Tsai, for always being there to help when it is most needed. Thank you all for everything that you taught me, and for being in my committee.

I am extremely fortunate to have had great mentors, for which I am greatly thankful for a number of interesting scientific and travel experiences. Thank you, Dr. Friederike Schmid, for so graciously hosting me many times; I learned so many technical research skills that invaluable to me now, and also had some of the greatest adventures of my life while visiting beautiful Mainz. Thank you, Dr. Markus Eisenbach, for hosting me at Oak Ridge National Lab and teaching me so much about interesting, cutting-edge computational methods while giving me the opportunity to experience the lab setting. Thank you, Dr. Ying Wai Li and Dr. Kipton Barros, for having me at Los Alamos National Lab, and for the extreme level of support. I have learned so many new skills related to physics, computer science, and working

with others, and can not thank you enough for all of your help during my final graduate semester.

I would also like to thank a few more of my research role models that have aided me greatly: Dr. Guangjie Shi, for the mentorship and coding help, Dr. Thomas Vogel, for a number of helpful discussions and interesting scientific ideas, and Dr. Thomas Wüst, for the technical discussions, providing the code that my projects are based upon, and to whom I look up to a great deal.

Thank you to Dr. Craig Wiegert, Dr. Jean-Pierre Caillault, and Dr. Phillip Stancil for their support while I was teaching. Also, thank you Mr. Tom Barnello for coordinating the labs.

Thank you to all of the administrative assistants who put up with me and helped me for so long: Ms. Linda Lee, Ms. Stephanie Crowe, Ms. Traci McKinney, and all of the Physics & Astronomy staff. I would like to thank Dr. Shan-Ho Tsai and the rest of the GACRC staff for the computing resources and technical help. Also, thank you to Mr. Mike Caplinger and Mr. Jeff Deroshia for the support with department computing resources and poster printings.

I would like to thank all of my friends back home and abroad for being there for me. Thank you, Dr. Alfred Farris and Dr. Effrosyni Seitaridou, so much for your help with the scientific, teaching, and overall emotional support over the years. Thank you to all of the graduate students, visiting scientists, and workshop attendees that I met at the department; I had a blast spending time with you all.

Last, but certainly not least, I would like to thank my parents and family. My parents have given me so much love and support, and have been there for me always. I would not have made it this far without the frequent visits, emotional support, and wonderful food during my schooling, and I can not put into words how much I love you. Likewise for my awesome sister and brother, Anna and Jack. I am very grateful for the continuing support from everyone in my extended family.

Table of Contents

Page

Acknowledgments iv

List of Figures viii

List of Tables xviii

1 Introduction 1

2 Background 5

2.1 The composition of a protein 5

2.2 Hierarchy of protein structures 7

2.3 The protein folding problem 8

2.4 Amyloid fibrils 9

3 Coarse-Grained Models 10

3.1 Lattice models 10

3.2 Lattice models for aggregation 14

3.3 Additional energetic terms 17

3.4 Mappings of protein sequences 20

4 Monte Carlo Methodology 24

4.1 Equilibrium Monte Carlo sampling 24

4.2 Trial moves for fcc lattice models 25

4.3 Calculation of thermodynamic averages 33

4.4 Wang-Landau sampling . . . 36
4.5 Replica-exchange Wang-Landau sampling . . . 37
4.6 Multicanonical sampling . . . 41
4.7 Practical details for simulations . . . 42

5 HP Model Folding on the FCC Lattice . . . 43
5.1 Predicted ground states . . . 43
5.2 Thermodynamic comparison with sc results . . . 44
5.3 Average structural quantities . . . 48

6 Simulations with Helical Motifs on the fcc Lattice . . . 55
6.1 Helical bundles in fcc ISAWs . . . 56
6.2 H0P model with helical motifs on the fcc lattice . . . 66

7 Lattice Model Simulations of Amyloid Protofibrils . . . 75
7.1 Aggregation model on the sc lattice . . . 75
7.2 Aggregation model on the fcc lattice . . . 95

8 Conclusions . . . 101

Appendix A : Pseudocodes and Technical Considerations . . . 105
A1 Depth-first search . . . 105
A2 Number of clusters and cluster size distribution . . . 107
A3 Coordinate shifts for visualization . . . 108
A4 Internal bond-rebridging moves for the fcc lattice . . . 110
A5 Tree data structure for saving configurations . . . 112
A6 Replica exchange frequency in REWL . . . 114
A7 Initialization of REWL . . . 117

Bibliography . . . 118

List of Figures

2.1 Diagram of a single amino acid, where the side-chain is the purple -R group and the central carbon is denoted C_α. 6

3.1 A high-energy configuration for an ISAW with length $\ell = 46$ on the fcc lattice. 11

3.2 Example of the HP model mapping on the fcc lattice for the BPTI protein. Polar residues are colored orange and hydrophobic residues are white. A few HH contacts are shown with green, dashed lines. 12

3.3 Example of the H0P model mapping on the fcc lattice for the BPTI protein. Polar residues are colored orange, neutral are gray, and hydrophobic residues are white. A few HH and H0 contacts are shown with dashed green and blue lines, respectively. 14

3.4 A view of a cross section for a stacked protofibril with the sc lattice aggregation model with β-hairpin subunits. The different energetic contributions in Equation 3.5 are labeled as green dots(dashes) correspond to ε'_{PP}, red dots for ε'_{HH}, and blue dashes for ε_{HH}. 15

3.5 Diagrams showing the nearest-neighbor positions for the sc lattice (left) and the fcc lattice (right). 16

3.6 Diagram of the dihedral angle that leads to α-helices on the fcc lattice. The first three residues lie in the purple plane and the proceeding three residues the green plane. 19

3.7 Diagram of the β-segment configuration where $\delta_i = 1$ on the fcc lattice. . . . 20

4.1 Single site pull move for the fcc lattice. The left image shows the initial trial displacement with a blue arrow, and the kink in the chain that is eliminated as a result with a red arrow. The right image shows the new bond vectors with green dashes that are assigned after the move. 26
4.2 Diagonal move for the fcc lattice. The left image shows the initial trial displacement with a blue arrow. The right image shows the relabeled sequence and the new bond vectors with green dashes that are assigned after the move. 28
4.3 Reptation move for the fcc lattice. The left image shows the initial deletion with a red arrow and the trial displacement at the opposing end with a blue arrow. The right image shows the relabeled sequence and the new bond vectors with green dashes that are assigned after the move. 28
4.4 Pivot move for the fcc lattice. The left image shows the selected portion for rotation circled in blue. The right image shows the new bond vectors with green dashes that are assigned after the move is accepted. 29
4.5 Type 2 internal rebridging move for the fcc lattice. The left image shows deleted bonds with red x's and the new, proposed bonds with blue dashes. The right image shows the relabeled sequence and the new bond vectors with green dashes that are assigned after the move. 30
4.6 End rebridging move for the fcc lattice. The left image shows the deleted bond with a red x and the new, proposed bond with blue dashes. The right image shows the relabeled sequence and the new bond vector with green dashes that are assigned after the move. 31
4.7 Cluster translation move for the fcc aggregation model. The randomly chosen cluster is shown enclosed in a dashed box, and is displaced in the random direction shown by the arrow in the left image. After the move, the cluster in the right image has a different position relative to the other three clusters. . 32

4.8 Cluster rotation move for the fcc aggregation model. The chosen cluster is shown enclosed in the dashed box, and a random rotation operation about a random axis is applied; shown by the two arrows in the left image. After the move, the cluster in the right image has a different orientation relative to the other three clusters. 33

4.9 Schematic diagram of the REWL scheme for 4 energy windows. Ovals with dashed arrows denote random walks performed by replicas within windows. An exchange is shown between replicas i and j in the overlap region of the middle two windows (shaded in gray), where configurations A and B are swapped. 38

4.10 Un-merged density of states for a REWL simulation with 8 windows for HP88 on the fcc lattice. 40

5.1 Ground state structures and energies found for the benchmark HP sequences from Table 3.1 on the fcc lattice. 44

5.2 Comparison between the estimate $\ln[\hat{g}(E)]$ of the sequence HP58 for the fcc (black circles) and sc (red triangles) lattices. The values have been shifted by subtracting out $\max\{\ln[\hat{g}(E)]\}$ and normalizing the energy by $\mathbb{c}_n$. 45

5.3 Comparison between the fcc (black circles) and sc (red dashes) lattice folding thermodynamics for biologically inspired HP sequences. Error bars are shown and are smaller than the size of data points where not visible. 47

5.4 Comparison between the fcc (black circles) and sc (red dashes) lattice folding thermodynamics for HP sequences that are designed to have threefold (HP67) and non degenerate (HP88) ground state structures on the sc lattice. Error bars are shown and are smaller than the size of data points where not visible. 47

5.5 Specific heat (top) and thermal derivative of the relative shape anisotropy for all residues (blue curve, bottom) and just the H residues (orange curve, bottom) for HP103. Error bars shown in the bottom plot are smaller than the marker sizes. 49

5.6 Specific heat (top) and thermal derivative of the relative shape anisotropy for all residues (blue curve, bottom) and just the H residues (orange curve, bottom) for HP124. Error bars shown in the bottom plot are smaller than the marker sizes. 51

5.7 Additional structural quantities for HP124. Specific heat (black) and thermal derivative of the hydrophobic radius of gyration (pink) are shown in the top plot, and the thermal derivative of end-to-end distance (green) is shown in the bottom plot. 52

5.8 Specific heat (top) and thermal derivative of the relative shape anisotropy for all residues (blue curve, bottom) and just the H residues (orange curve, bottom) for HP136. Error bars shown in the bottom plot are smaller than the marker sizes. 53

5.9 Thermal derivative for the end-to-end distance of HP136. Error bars are shown and are smaller than the size of data points where not visible. 54

6.1 Three approximations for the α-helix on the fcc lattice. White structures show the lattice approximations, and green structures show a real helix backbone taken from the PDB. The three types are: repeating $90°, 120°$ bond angles (a), repeating $90°, 60°$ bond angles (b), and Pokarowski's helix (c). 56

6.2 Double helix structure for a 46-mer ISAW on the fcc lattice. Two α-helices wrap around one another to form a dense configuration contains a large number of nearest-neighbor contacts. 58

6.3 Diagram of the neighbors for the four closest distances to a site on the fcc lattice. The number of neighbors for each distance is given by the subscripted $\mathbb{c}$ (*e.g.* the third-nearest distance has 24 neighbors). 59

6.4 Contact maps and structures for the ground states containing 1-, 2-, 3-, and 4-helix bundles for the fcc lattice ISAW with 46 residues. Contact maps are shown on the bottom-left of each plot, where the axes show residue positions along the chain, and colored data in the map correspond to contact distances (red for nearest-neighbor and blue for third-nearest neighbor). Snapshots of the helical bundle structures are shown in the shaded region in the top-right of each plot. 60

6.5 Heatmap of the thermal derivative of the energy-weighted contact numbers at a range of temperatures (x-axis) and torsion penalty values (y-axis) for the fcc lattice ISAW with 46 residues. 62

6.6 Heatmap of the relative shape anisotropy at a range of temperatures (x-axis) and torsion penalty values (y-axis) for the fcc lattice ISAW with 46 residues. Structural regions are labeled with abbreviations: one helix (1H), two helices (2H), three helices (3H), four helices (4H), globular helix (GH), and random coil (RC). 64

6.7 Representative configurations for the fcc lattice ISAW with 46 residues taken at each of the structural regions labeled in Figure 6.6. Helical bundles are shown with a view along their elongated axes. 65

6.8 Comparison of C_V/ℓ for the HP (top) and H0P (bottom) models of Crambin on fcc (black points) and sc (red triangles) lattices. The temperature axis is normalized by the coordination $\mathbb{c}_n$ number for each lattice. Energetic couplings are set to ($\varepsilon_{HH} = 2$, $\varepsilon_{H0} = 1$) for the H0P model simulations. Data for sc lattice was provided by Alfred Farris. Error bars are shown and are smaller than the size of data points where not visible. 67

6.9 Comparison of C_V/ℓ for the standard H0P model (black curve), H0P model with contact distances set to d_3 (blue circles), and H0P model with nearest-neighbor repulsion and contact distances set to d_3 (green triangles). Error bars are shown and are smaller than the size of data points where not visible. 68

6.10 Ground state for the standard H0P model of Crambin. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} = 3.93 \times 10^5$ structures are used. 70

6.11 Ground state for H0P model of Crambin with contact distances set to d_3. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} = 2.49 \times 10^4$ structures are used. 70

6.12 Ground state for H0P model of Crambin with contact distances set to d_3 and nearest-neighbor repulsion. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} = 71$ structures are used. 71

6.13 Ground state for H0P model of Crambin with contact distances set to d_3, nearest-neighbor repulsion, and torsion penalties. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} =$ 1 structures are used. 71

6.14 Experimentally measured backbone structure from the PDB (1CRN) for Crambin. The H0P mapping is shown on top of the PDB structure. 74

7.1 The β-hairpin subunit on the sc lattice, where the left image is the optimal globular state with 6 HH contacts, and the right image is the planar β-hairpin structure with 5 HH contacts. Dashed, blue lines show intramolecular HH contacts. 76

7.2 Natural logarithm of the density of states for 24 β-hairpin subunits on the sc lattice where $\varepsilon_{HH} = 10$, $\varepsilon'_{HH} = 5$, and $\varepsilon'_{PP} = 4$. The inset shows a closeup view of $\ln[\hat{g}(E)]$ at high energies. 77

7.3 Internal energy for 24 β-hairpin subunits on the sc lattice where $\varepsilon_{HH} = 10$, $\varepsilon'_{HH} = 5$, and $\varepsilon'_{PP} = 4$. Error bars are shown and are smaller than the size of data points where not visible. 78

7.4 Thermodynamics for stacked protofibril formation of 16 β-hairpin subunits where $\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, and $\varepsilon'_{PP} = 1$. The specific heat (blacked, dashed curve) and contributions from thermal derivatives of average contact numbers (blue for intramolecular HH, red for intermolecular HH, and green for intermolecular PP) are plotted. Representative configurations are shown below the x-axis with arrows signifying the temperatures where they are recorded. The temperature for the structural transition of protofibril formation T_f is labeled with an arrow. 79

7.5 Thermodynamics for stacked protofibril formation of 24 β-hairpin subunits where $\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, and $\varepsilon'_{PP} = 1$. The specific heat (black, dashed curve) and contributions from thermal derivatives of average contact numbers (blue for intramolecular HH, red for intermolecular HH, and green for intermolecular PP) are plotted. Representative configurations are shown below the x-axis with arrows signifying the temperatures where they are recorded. The temperature for the structural transition of protofibril formation T_f is labeled with an arrow. 82

7.6 Alternate fibril cross sections that are observed in simulations with strong intermolecular HH coupling. Image (a) is equivalent in energy to the two-layer β-hairpin protofibril presented in Figure 3.4 when $\varepsilon_{HH} = \varepsilon'_{HH}$. 83

7.7 Specific heat for 24 β-hairpin subunits where the hydrophobicity scales for intermolecular interactions are chosen to promote protofibril stacking with 'one' (blue curve), two (green curve), and three (red curve) layers. Configurations of the stacked ground states are shown with arrows pointing to their respective data. 85

7.8 Free energy difference as a function of cluster size for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $< 10\%$ of $|\Delta F|$. 87

7.9 Free energy difference as a function of hydrophobic cluster size for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $<$ 10% of $|\Delta F|$. 88

7.10 Distribution of cluster sizes for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to cluster sizes observed with a fraction $\leq 10^{-6}$. 89

7.11 Distribution of hydrophobic cluster sizes for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to cluster sizes observed with a fraction $\leq 10^{-6}$. 90

7.12 Free energy difference as a function of cluster size for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $< 10\%$ of $|\Delta F|$. . 91

7.13 Free energy difference as a function of hydrophobic cluster size for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $< 10\%$ of $|\Delta F|$. 92

7.14 Distribution of cluster sizes for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to cluster sizes observed with a fraction $\leq 10^{-6}$. 93

7.15 Distribution of hydrophobic cluster sizes for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to temperatures where clusters were not recorded for the given size. . 94

7.16 The β-strand subunit on the sc (top) and fcc (bottom) lattices. The left images show disordered, non-interacting subunits, and the right images show pairs of subunits interacting as β-strands through intermolecular HH contacts. Dashed, red lines show intermolecular HH contacts. 95

7.17 Comparison of the specific heat for fcc (black circles) and sc (red triangles) lattice aggregation simulations with 16 β-strand subunits. Couplings are set to $\varepsilon'_{_{HH}} = 1$ for both lattices. The fcc lattice simulation has additional β-sheet energies $\varepsilon_{_{\beta}} = 1$ and angle penalties $\varepsilon_{(\theta=60^\circ)} = \varepsilon_{(\theta=180^\circ)} = -1$. Ground state configurations are shown below the x-axis for the fcc (left) and sc (right) lattices. 97

7.18 Series of aggregate states for 32 β-hairpin subunits, where black arrows point in the direction of decreasing energy, and the lowest energy structure found is shown at the bottom-right. Couplings are set to $\varepsilon'_{_{HH}} = 1$, $\varepsilon_{_{\beta}} = 1$, and $\varepsilon_{(\theta=60^\circ)} = \varepsilon_{(\theta=180^\circ)} = -1$. 98

7.19 Series of aggregate states for 32 β-hairpin subunits, where black arrows point in the direction of decreasing energy, and the lowest energy structure found is shown at the bottom-right. Couplings are set to $\varepsilon'_{_{HH}} = 1$, $\varepsilon'_{_{PP}} = 0.1$, $\varepsilon_{_{\beta}} = 1$, and $\varepsilon_{(\theta=60^\circ)} = \varepsilon_{(\theta=180^\circ)} = -1$. 99

A1 Two types of rebridging moves. The example configuration has residues arranged in two parallel planes, where the four larger, slightly shaded residues are in plane and in the foreground. The new (broken) bonds need not be parallel, as shown in steps (a). For clarity, the images show configurations without a sequence; otherwise, a trivial relabeling would be performed. . . . 111

A2 Tree data structure for storing unique configurations on the fcc lattice. The dots in the left- and right-most images are nodes that represent bond vectors in the configurations. Shown in the middle image, the added configuration has bond vectors that must be added as nodes (or a leaf) in the tree structure shown in red, and those that have been previously encountered shown in blue. 113

A3 Reference density of states ($\hat{g}_0(E)$, left plot) and ratio of converged REWL results ($\hat{g}(E)/\hat{g}_0(E)$, right plot) from the reference values. Two windows are used in a REWL simulation where replica exchanges are attempted every 10 MC sweeps. The x-axes show the magnitude of energy, where window 2 contains the ground state energy. 115

A4 Two-dimensional histogram of accepted replica exchanges between the two REWL windows from the simulations in Figure A3. The right image shows accepted exchanges for window 1 during the 1st REWL iteration, and the left image shows accepted exchanges for window 2 during the 17th REWL iteration. 116

A5 Schematic diagram of the initialization procedure employed for REWL and parallel MUCA simulations. The black line shows the trajectory for the temporary WL sampler, which is constrained to sample states according to some decreasing boundary (red line). An pad of some width (blue region) is assigned to prevent the sampler from being 'stuck' at a state when the boundary is decreased. The initialization completes for the replica when it reaches a state that is inside the assigned window bounds (green region). 118

List of Tables

2.1 20 genetically encoded amino acids with full names, three-letter abbreviations, and one-letter codes. 6

3.1 HP sequences for: [a] Bovine Pancreatic Trypsin Inhibitor (BPTI) [62], [b] benchmarks with low-degeneracy ground states on the sc lattice [39, 63], [c] Cytochrome C (apo form) [64], [d] Ribonuclease A [64], [e] 1-136 Staphylococcal nuclease fragment [64]. Subscripts in the specified sequences signify consecutively repeated residues. 21

3.2 HP and H0P mappings of the 46-residue Crambin protein. 22

6.1 Average $dRMSD$ values for H0P model variants of the Crambin protein. The first column gives parameter values of the models: HP model (row 1); H0P model (row 2); H0P model with contact distances set to d_3 (row 3); H0P model with contact distances set to d_3 and nearest-neighbor repulsion (row 4); H0P model with contact distances set to d_3, nearest-neighbor repulsion, and torsion penalty (rows 5 and 6). The last column gives the number of unique configurations used in each calculation. 73

Chapter 1

Introduction

The recent and rapid development of computational resources has established numerical simulation as an invaluable tool for the physical sciences, where all but the most idealized problems result in theoretically tractable solutions. Among the most intricate of systems, those involving biological macromolecules such as proteins and interacting polypeptides exhibit processes like folding and aggregation, and remain at the forefront of interdisciplinary research after many decades [1, 2]. Both molecules are chains of bonded amino acids, where polypeptides are short in length (< 50 amino acids), and proteins are typically much longer. Proteins are essential 'machinery' of cellular life, which along with many other roles [3], act as enzymes, govern the transport of chemicals through cellular membranes, and regulate biological functions as hormones. Such functions are closely related to the protein's folded structure, or native state, and are affected by interactions with the cellular environment, including other proteins. Aggregates of proteins and protein fragments can form alternative, non-functional structures called amyloid fibrils that are associated with a myriad of disorders, such as prion diseases, Alzheimer's disease, and type II diabetes [4]. The complexity of these systems is such that studies may aim to further the understanding of even basic aspects of their physical behavior.

The 'protein folding problem' [1] is an extensive research topic that attempts to address: identifying the balance of forces that result in folded or native structures, how to predict structures from an amino acid sequence alone, and determining what mechanism leads proteins to fold so quickly. Folding can have timescales ranging from microseconds to seconds [5], and typically involves several thousands of atoms with complex interactions that occur on the scale of a few angstroms ($1\text{Å} = 10^{-10}m$). Experimental study of protein folding with both adequate spatial and temporal resolution, therefore, becomes a daunting challenge. Theoretical descriptions are unfeasible due the extremely large number of degrees of freedom with complex interactions, and the inclusion of the *in vivo* folding environment only complicates matters further. Computer simulation is a useful tool to study this topic, but also has similar limitations, such as the problem of an unknown potential energy function in general, and also the long time scales with dynamical simulations.

Similar to the protein folding problem, the aggregation of proteins and polypeptides is another topic that has a plethora of challenging, unanswered questions with enormous implications. For example, large plaques of the Aβ peptide have been observed in the brain of those who suffer from Alzheimer's disease, and fibrils of the human islet amyloid polypeptide (amylin hormone) inside the pancreas in those with Type-II diabetes [4]. No direct link is known that connects the pathology to fibril structures. This mysterious hallmark of amyloidogenic diseases motivates the study of whether intermediate aggregate structures are responsible for the toxicity during the formation of mature fibrils [6]. Rather than being explicitly sequence-dependent like protein folding, it is believed that the amyloid state is a general propensity of polypeptides [7, 8], which stabilized through various intra- and intermolecular interactions, provides a thermodynamically stable alternative to a folded state. Many of the challenges present in protein folding arise when studying the formation of amyloid fibrils, though length scales range from 10Å to $> 1\mu m$, and characteristic time scales on the order of hours [9]. Much effort has been put towards studying the intermediate species

found during the formation of amyloid fibrils, but many aspects of the process remain unknown.

The use of numerical simulation is particularly advantageous for studying these problems, as one can probe the relevance of different physical features using models of various complexities. For example, if a very high-resolution, quantitative analysis of protein structure is desired, one could perform a dynamical simulation which accounts for all atomic interactions, but the length of feasible time integration becomes extremely limiting. Alternatively, if there is a question of large-scale, qualitative phenomena like folding and aggregation, a statistical analysis using reduced, or coarse-grained, simulation models may be the only viable option. Different levels of coarse-graining can be chosen, and protein models with discretized spatial coordinates to some rigid lattice geometry have been historically used in the physics community. Lattice models are advantageous to continuous coarse-grained models in terms of efficiency: energy calculations and self-avoidance checks can be performed quickly, and integer arithmetic can be used to store energies and coordinates. Used for its simplicity and computational accessibility, the simple cubic (sc) lattice is a popular choice for coarse-grained protein models, but suffers from a high limitation on allowed bond angles and possible interaction pairs between residues. The face-centered cubic (fcc) lattice, however, provides a good representation of the structures measured in real proteins [10–12] while retaining the convenience of simplicity when compared to other more sophisticated lattice types or continuum models.

In this work, we apply advanced Monte Carlo (MC) methods from statistical physics to study protein folding and peptide aggregation using fcc lattice models. More specifically, the thermodynamics of model protein folding are calculated and compared with results for the more geometrically constrained sc lattice. We then incorporate helical structural motifs into the protein folding simulations and examine the temperature-dependent behavior at various strengths of the potential. Additionally, the qualitative aspects of amyloid protofibril formation are examined using a sc lattice model, where the thermodynamics of aggregation

and fibril stacking are considered. An aggregation model is then implemented using the fcc lattice and tested with protofibril formation.

The arrangement of this dissertation is as follows: Chapter 2 provides a high-level overview of protein and polypeptide structures, protein folding, and the structure of amyloid fibrils. Chapter 3 details the coarse-grained models used in our studies and describes the geometries and energetic interactions that are considered. Chapter 4 presents the Monte Carlo algorithms employed in our simulations, the trial moves that are used to sample model configurations, and some relevant thermodynamic and structural quantities that are calculated for the systems. Chapter 5 contains simulation results for a set of benchmark model proteins on the fcc lattice and gives a thermodynamic comparison with existing results. Chapter 6 examines lattice polymer and protein folding simulations which incorporate additional interactions to promote structural motifs. Finally, Chapter 7 applies the methods to simulate amyloid protofibril formation and stacking using lattice models. A conclusion is presented in Chapter 8, followed by appendices that provide supplemental information in the form of technical details and pseudocodes.

Chapter 2

Background

2.1 The composition of a protein

Proteins are polypeptide macromolecules, which are linear chains of many bonded amino acids. Amino acids are organic compounds containing amino (-NH_2), carboxyl (-$COOH$), and alkyl (-R) functional groups bonded to a central carbon atom (C_α). Of these functional groups, the alkyl group, also known as the side chain, is what distinguishes the chemical properties of the 20 biologically relevant amino acids. Figure 2.1 shows a diagram of an amino acid with the different functional groups and C_α labeled. Properties such as hydrophobicity and electric charge are important classifiers for how amino acids affect a protein's behavior given the surrounding solvent conditions. Each amino acid contains 10-20 atoms, depending on the side chain. Peptide bonds, a type of covalent bond, form between the amino group and carboxyl group of neighboring amino acids to form what is called the 'backbone' of a protein molecule. The distance between two C_α of bonded amino acids is ≈ 3.8Å. Protein backbones are typically a few hundred of amino acids in length, meaning that a protein contains several thousands of atoms. The different amino acid types are listed in Table 2.1, with the full names, three-letter abbreviations, and one-letter codes shown.

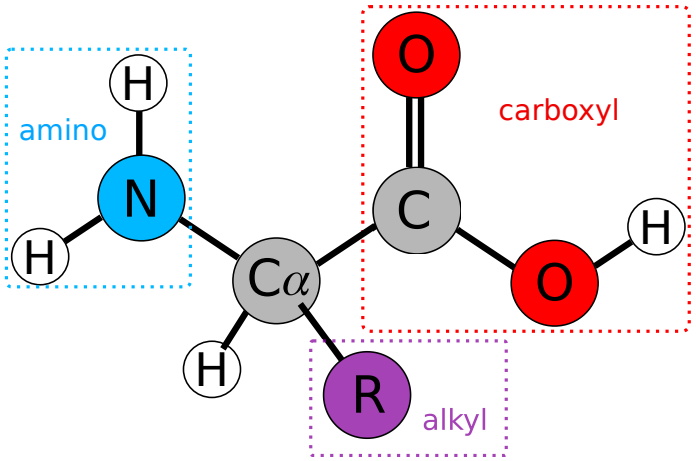

Figure 2.1: Diagram of a single amino acid, where the side-chain is the purple -R group and the central carbon is denoted C_α.

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

Table 2.1: 20 genetically encoded amino acids with full names, three-letter abbreviations, and one-letter codes.

2.2 Hierarchy of protein structures

A hierarchical nomenclature is used to describe the rich structural features that proteins contain across various length scales. At the lowest level, the primary structure is the genetically encoded sequence of amino acids which constitute a protein's backbone. This 1-dimensional primary structure is the information that gives a protein its identity, and ultimately determines its function. The primary structure for functional proteins can range from $\sim$50 to $\geq$ 1000 amino acids in length [13].

Secondary structures are the next level in the hierarchy, and contain local structural motifs that are largely stabilized by hydrogen bonds [14]. The predominant structures at this level are α-helices and β-strands, although there are other types such as the 3_{10}- and π- helices and β-hairpins. While not technically a structure, random coils are common as disordered regions between other secondary structural elements. α-helices are characterized as having a winding number of 3.6 amino acids per turn, and a pitch of 5.4Å. β-strands are extended structures that have a 'pleated' or 'zig-zag' appearance between consecutive amino acids, which have a pitch of 3.5Å. The emergence of these two secondary structures from steric constraints of polypeptides is well known through Ramachandran plots [15] of dihedral angles ϕ between backbone (C-C_α-N-C) and ψ between backbone (N-C-C_α-N).

When a protein assumes its folded or native configuration, the 3-dimensional structure is referred to as the tertiary structure. Any number of secondary structure elements that are present in the protein are organized globally according to the chemical properties of the amino acids and interactions with the solvent or environment. Because proteins are synthesized and active inside of cells, the surrounding solvent is aqueous, and amino acid hydrophobicity plays a large role in this global packing. Globular protein structures have a hydrophobic 'core', where hydrophilic residues interact with the polar solvent and form a surface layer that leaves hydrophobic residues in the interior. It is the tertiary structure which enables the function of a protein.

Finally, the quaternary structure is the higher-order compound of two or more folded proteins or polypeptides. Globular proteins often interact as subunits in a larger functional mechanism, such as the case with various enzymes and membrane proteins. Quaternary structures can manifest as various numbers of interacting polypeptides; from oligomers like dimers (two), trimers (three), etc., to higher-order structures. Amyloids are an aggregate state for polypeptide systems that are not specific to one type of sequence [7], and typically have a structure that differs from the quaternary structure of functional proteins.

2.3 The protein folding problem

In the mid 1960s, Anfinsen et. al. [16] hypothesized that a protein's native structure depends only on the properties of the solvent and its constituent sequence of amino acids, and that the native state is the most thermodynamically stable. This is known as the thermodynamic hypothesis of protein folding, and frames the problem as a minimization of the free energy F for the available states of the system. On a contrary view, the thought experiment known as Levinthal's paradox [17] regards the ability of a protein to find its native state so quickly seemingly paradoxical, and that a kinetic folding pathway occurs with a specific sequence of intermediate states. To expand on these viewpoints, the concept of a funnel-shaped free energy landscape, or folding funnel [18, 19], adopts a statistical interpretation of the problem. The free energy $F = U - TS$ has a term that includes the internal energy U and a temperature-dependent term that incorporates the entropy S of the system. As the protein approaches its native state at the bottom of the funnel, internal energy decreases but the number of accessible configurations, and therefore the entropy, decreases. There can be many local minima in the free energy landscape that signify macroscopic states, but the system evolves using an ensemble of many available microscopic states as the free energy is minimized. These are a few noteworthy postulations about the nature of protein folding,

but many aspects, including what a necessary set of interactions and whether there exists an ability to predict folded structures from a primary sequence alone, remain unsolved.

2.4 Amyloid fibrils

Amyloid fibrils are elongated, linear aggregate structures with polypeptides or proteins as subunits, or constituent molecules. Subunits typically orient perpendicular to the growth axis of the fibril with a β-sheet configuration to form what is referred to as the cross-β structure of fibrils. The radius of an amyloid fibril is $\sim 10nm$ and they can reach lengths of $\geq 1\mu m$ [9]. As fibril structures form, there can be an ensemble of intermediate oligomer species; both structurally disordered and with ordered helical and cross-β motifs. Protofibrils are ordered oligomers that form with approximate sizes of 15-40 subunits as fibrillar precursors [20]. Droplets, pores, and fiber-like oligomers are all examples of observed protofibril structures [20]. Mature amyloid fibrils can consist of many stacked β-sheet layers or several packed protofibrils that may exhibit a helical twist [9]. Large, insoluble plaques of fibrils are known to form and deposit in the extracellular matrix [6], as in case of amyloidogenic illnesses. Despite the observed presence of amyloid fibrils in affected organs and tissues, intermediate oligomers or protofibrils are suspected to be responsible for the cellular toxicity [20].

Regions of protein sequences that promote the formation of amyloid states appear to be short and non-specific, and fibril structures can be produced in systems of short polypeptides [21]. Unlike the native state and functional quaternary state of proteins, amyloids are thought to be a generic state of polypeptide systems, which provide an alternative free energy minimum to the functional states of their subunits. Many aspects about the nature of amyloid formation can be examined using coarse-grained computer simulations, including topics like multi-step nucleation [22] and conformational conversion [23], stacking and packing polymorphism [24, 25], and the adsorption of aggregates to surfaces [26] and membranes [27, 28].

Chapter 3

Coarse-Grained Models

3.1 Lattice models

3.1.1 Interacting self-avoiding walk (ISAW)

Polymeric systems, including proteins, have the underlying constraints of connectivity and self avoidance, meaning that residues are bonded linearly (excluding branched polymers) and can not overlap with one another spatially. These properties are captured by a model known as the self-avoiding walk (SAW) [29], which makes a 'walk' of consecutive steps in random lattice directions, so long as a given vertex on the lattice is not crossed more than once. Amino acids are modeled using a united-atom representation as point-like residues that lie on the vertices that the SAW passes through, as shown in Figure 3.1. Thermodynamics are incorporated by adding contact interactions (the 'I' in ISAW) between non-bonded residues that occupy adjacent vertices on the lattice.

$$\mathcal{H} = -n\varepsilon \tag{3.1}$$

Shown in Equation 3.1, the Hamiltonian for an ISAW counts the number n of energetic contacts in the configuration, and weights them by an energetic factor ε. For this model,

the ground state has the maximum number of contacts, and is a compact configuration that depends on the underlying lattice geometry and dimension.

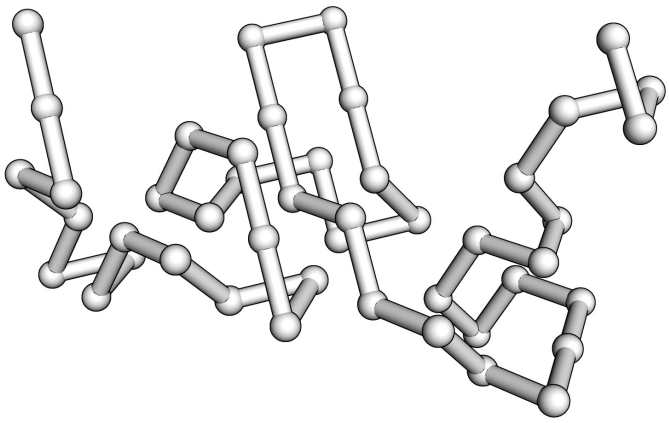

Figure 3.1: A high-energy configuration for an ISAW with length $\ell = 46$ on the fcc lattice.

The number C_ℓ of valid SAWs in 3D grows exponentially with the length ℓ of the walk, and depends on the entropic exponent $\gamma \approx 1.16$ and lattice-dependent growth constant μ as [30]:

$$C_\ell \;\propto\; \mu^\ell \ell^{\,\gamma-1}\,. \tag{3.2}$$

The enormous number of available states and entropic barriers associated with the polymeric backbone make adequately sampling phase space an extremely challenging task. SAWs and ISAWs have been used extensively to study various aspects of polymer and protein systems [29, 31–33], and have motivated the development of advanced MC algorithms [34]. We use ISAWs on the fcc lattice in simulations of helix bundles by augmenting the Hamiltonian with additional interactions described at the end of the chapter, and as a test for accuracy of the tree data structure described in Chapter 4.

3.1.2 The hydrophobic-polar (HP) model

The HP model [35, 36] greatly reduces the degrees of freedom of the represented protein while preserving the essential physics of folding. Amino acid types are classified as either hydrophobic (H) or polar (P) according to their side-chain properties, and the HP sequence mapped from the protein's primary structure is the defining input for the model. Like the ISAW, the model includes self-avoidance and contact energies defined by non-bonded united-atom residues that occupy adjacent vertices on a lattice, although only contacts between H residues (HH contacts) contribute to the energy. The model was first described using the sc and square lattices, but is also valid on other lattice geometries.

$$\mathcal{H} = -n_{HH}\varepsilon_{HH} \tag{3.3}$$

The Hamiltonian for the HP model is shown in Equation 3.3 and depends on the number of contacts between H residues, n_{HH}, and an energetic coupling constant ε_{HH}. The ground state conformation of this model maximizes the number of HH contacts and is determined by the HP sequence mapping for the protein in question. A key feature of this model is the

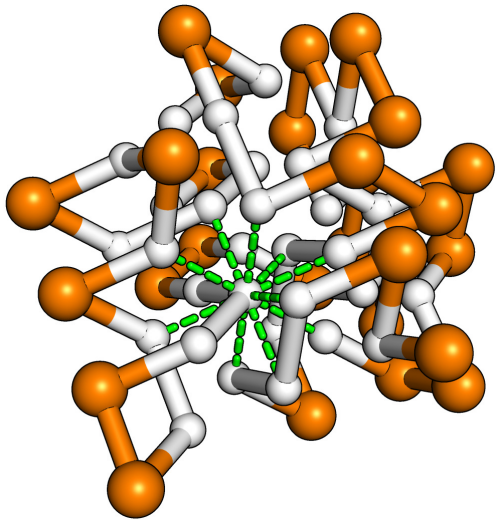

Figure 3.2: Example of the HP model mapping on the fcc lattice for the BPTI protein. Polar residues are colored orange and hydrophobic residues are white. A few HH contacts are shown with green, dashed lines.

formation of a hydrophobic core (H-core) through an implicit solvent, where hydrophobic amino acids group together in the interior of the protein's conformation. Though not a real force with a known potential, this behavior is a result of polar residues' interaction with the surrounding, aqueous solvent, and the tendency of hydrophobic residues to avoid contact with the solvent [35]. An example of a H-core containing state for the HP model on the fcc lattice is shown in Figure 3.2. While the HP model is conceptually very simple, the number of valid configurations for each energy (*e.g.* a large subset of states from Equation 3.2) are innumerable beyond short sequence lengths, and typically have a high level of degeneracy [37]. Identification of the ground state poses an NP-complete optimization problem [38], and has been used as a testing ground for advanced MC algorithms [39–45].

3.1.3 The H0P model

A simple extension of the HP model, called the H0P model [46, 47], involves the addition of a neutral (0) hydrophobicity group to the amino acid classification. The idea is motivated by studies with more complicated augmentations to the HP model on sc [48] and fcc lattices [49] that split the H group into separate hydrophobicity scales, and the P group based on positive, negative, and neutral charges. The Hamiltonian for the H0P model with HH and H0 interactions is shown in Equation 3.4, where the subscripted n represent contact numbers and the ε are energetic coupling constants for the corresponding quantities.

$$\mathcal{H} = -\Big(n_{HH}\varepsilon_{HH} + n_{H0}\varepsilon_{H0}\Big) \tag{3.4}$$

An additional term for 00 interactions can also be incorporated, but is not considered in our simulations. This model has been shown to be effective in reducing the degeneracy of ground states on the sc lattice, especially in combination with bending angle penalties [50,51]. Figure 3.3 shows an example of a folded state for the H0P model on the fcc lattice, which contains an H-core surrounded by a layer of H0 contacts.

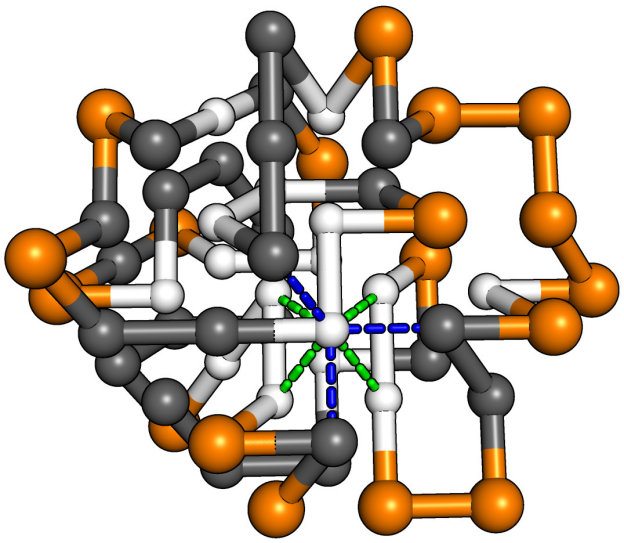

Figure 3.3: Example of the H0P model mapping on the fcc lattice for the BPTI protein. Polar residues are colored orange, neutral are gray, and hydrophobic residues are white. A few HH and H0 contacts are shown with dashed green and blue lines, respectively.

3.2 Lattice models for aggregation

As a coarse-grained model for the aggregation of peptides and proteins with specified sequences, we employ multiple, simultaneously interacting instances of the H0P model. The aggregation and protofibril formation of short, identical peptides is considered, but the model can also be used for multiple interacting proteins with varying sequences and lengths. Although the H0P model was motivated for protein folding studies, we use the 0 residue type in the aggregation model to incorporate a placeholder that has no hydrophobic interactions and could be assigned as charged termini for the peptides.

N individual H0P peptide subunits, each with length ℓ, are allowed to interact in a periodically bounded, cubic simulation box with sides of length $\geq \ell + 2$. Both sc and fcc versions of the model are considered, with the same definition of the energy. The energy h_k of the k^{th} subunit is given by Equation 3.6, and incorporates three hydrophobicity levels for interactions.

$$h_k = -\left[n_{HH}\varepsilon_{HH} + \frac{1}{2}(n'_{HH}\varepsilon'_{HH} + n'_{PP}\varepsilon'_{PP})\right]_k \tag{3.5}$$

Intramolecular HH contact energy is denoted by the term $-n_{HH}\varepsilon_{HH}$, which has the same definition as the usual HP model. The other two terms that have 'primed' (superscripted with a tick mark) variables represent **inter**molecular HH and PP contact energies with nearest-neighbor residues from any of the other $N-1$ subunits. A factor of $\frac{1}{2}$ is included only in this notation to avoid double-counting, and the subscript $[...]_k$ signifies that the evaluation belongs to the k^{th} H0P subunit.

$$\mathcal{H} = \sum_{k=1}^{N} h_k \tag{3.6}$$

The Hamiltonian for the system sums over all subunit energies, as shown in Equation 3.6.

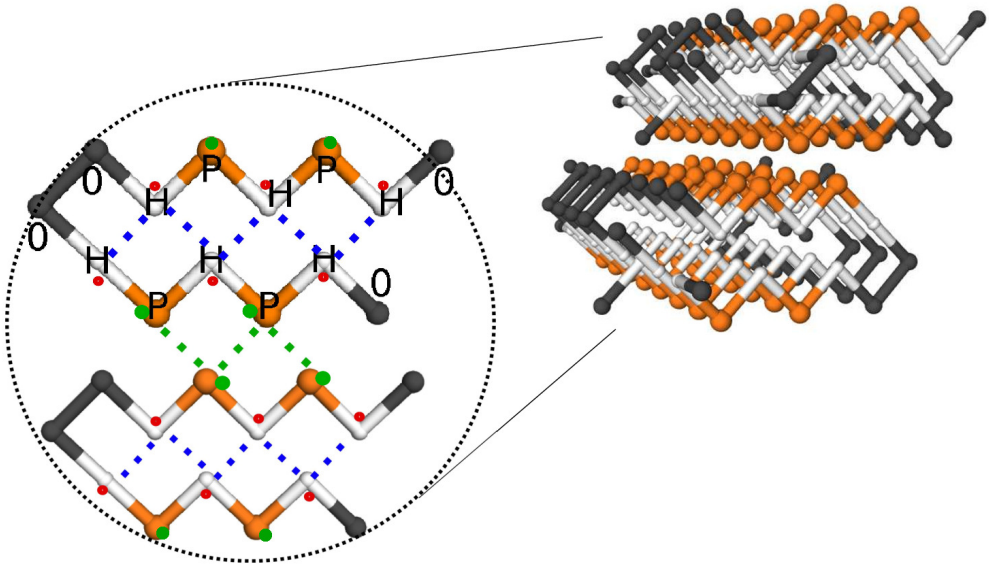

Figure 3.4: A view of a cross section for a stacked protofibril with the sc lattice aggregation model with β-hairpin subunits. The different energetic contributions in Equation 3.5 are labeled as green dots(dashes) correspond to ε'_{PP}, red dots for ε'_{HH}, and blue dashes for ε_{HH}.

Figure 3.4 shows an example of a sc model fibril with the interactions described in Equation 3.5 labeled with different colors. Red and green dots near H and P residues denote intermolecular HH and PP contacts along the fibril axis (out of page), respectively. Perpendicular to the fibril axis (in page), the dashed, blue lines signify intramolecular HH contacts, and dashed, green lines represent intermolecular PP contacts. Additional intramolecular interactions, like those described in Section 3.3, may also be incorporated into Equation 3.5.

3.2.1 Lattice geometries

The backbone models detailed in the previous section can be implemented on any discrete lattice geometry. Square and sc lattices have been historically used, especially in the case of the HP model and its variants, due to its simplicity and computational convenience. However, popular alternatives for the HP and other protein models include the tetrahedral, fcc, body-centered cubic, and 210 (or knight's walk) [52] lattices. More advanced lattice types with higher spatial resolution have also been successfully utilized [53,54]. We focus on using the fcc lattice to extend a powerful MC methodology that was applied to the listed sequences on the sc lattice.

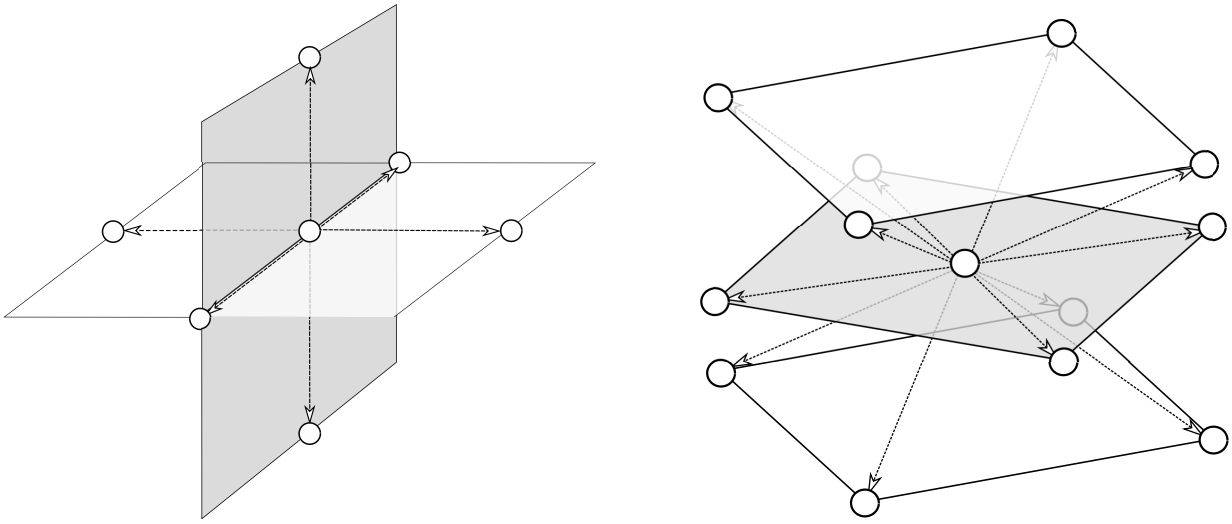

Figure 3.5: Diagrams showing the nearest-neighbor positions for the sc lattice (left) and the fcc lattice (right).

The sc lattice, shown on the left of Figure 3.5, has a cubic packing geometry with 6 nearest-neighbor sites at each site of the lattice. With a unit lattice spacing, the lattice vectors are $\pm\hat{x}$, $\pm\hat{y}$, and $\pm\hat{z}$. For backbone models on this lattice, there can only be angles of $\pm 90^\circ$ or 180° between any two consecutive backbone bonds. One severe constraint of the sc lattice is that for models with nearest-neighbor contact energies, only residues which are separated by an odd number of lattice sites can interact; called the 'parity problem'. The

growth factor for the sc lattice is $\mu = 4.684$, and SAWs with lengths of up to $\ell = 36$ have been enumerated at 2.9×10^{24} possible configurations [30].

The fcc lattice, shown on the right of Figure 3.5, has a cuboctahedral packing geometry [10] with 12 nearest-neighbor sites at each site of the lattice. This lattice has a close-packed geometry where identical spheres placed at the vertices have the highest packing density of 0.74 [55]. With a unit lattice spacing, the lattice vectors are $\pm\frac{1}{\sqrt{2}}(\hat{x}+\hat{y})$, $\pm\frac{1}{\sqrt{2}}(\hat{x}-\hat{y})$, $\pm\frac{1}{\sqrt{2}}(\hat{x}+\hat{z})$, $\pm\frac{1}{\sqrt{2}}(\hat{x}-\hat{z})$, $\pm\frac{1}{\sqrt{2}}(\hat{y}+\hat{z})$, and $\pm\frac{1}{\sqrt{2}}(\hat{y}-\hat{z})$. For backbone models on this lattice, there is a set of interior bond angles $\theta_i \in \{\pm 60°, \pm 90°, \pm 120°, 180°\}$ that are possible between any two consecutive backbone bonds. It is also meaningful to consider the set of dihedral angles $\phi_i \in \{0°, \pm 55°, \pm 70°, \pm 109°, \pm 125°, 180°\}$ that can be defined between any four consecutive residues. An important aspect of this lattice is that unlike the sc lattice, the considered models are not hindered by the parity problem on the fcc lattice. The growth factor for the fcc lattice is $\mu = 6.53$, and SAWs with lengths of up to $\ell = 24$ have been enumerated at 2.0×10^{24} possible configurations [56].

The coordination number $\mathbb{c}_n$ is used for normalization of temperature scales when comparing results between the fcc and sc lattices, where

$$\mathbb{c}_n = \begin{cases} 6, & \text{sc lattice} \\ 12, & \text{fcc lattice} \end{cases} . \tag{3.7}$$

3.3 Additional energetic terms

All of the interactions described in the model Hamiltonians so far have been two-body contacts between pairs of non-bonded, nearest-neighbor residues. Hydrogen bonding plays an important role in the secondary structural elements of proteins and polymers, and is not addressed by hydrophobicity based models like the HP and H0P models. The following potential energies are additional terms that we include in our fcc simulations of the ISAW, H0P, and aggregation models to promote the formation of structural motifs. One or more of the

following energetic contributions shown in Equations 3.8, 3.9 and 3.10 can be incorporated by directly adding them to Equations 3.1, 3.4, and 3.5, which will be further described in the relevant sections of this dissertation.

3.3.1 Angle energies

Energetic restraints of the angle interior to two consecutive backbone bonds, or between three consecutive residues, is used to introduce a level of semiflexibility to the models. Semiflexibility is a feature that has been previously studied for ISAWs [57] and lattice proteins [50], and involves adding a positive energy contribution for certain (or all) angles. For the sc lattice models, this simply involves a penalty term which depends on the total number of bends in a conformation, as only $\theta_i = 90°$ and $180°$ are possible.

$$E_\theta = -\sum_{i=1}^{4} n_{\theta_i} \varepsilon_{\theta_i} \tag{3.8}$$

For the fcc lattice, the angle energy restraint shown in Equation 3.8 is a sum over the four possible angles where n_{θ_i} is the total number of angles with a magnitude of $\theta_i \in \{60°, 90°, 120°, 180°\}$, each weighted by its own value of ε_{θ_i}. Angles can be pre-enumerated and referenced using a lookup table during simulations with the indices of adjacent bonds.

3.3.2 Torsional energies

One of the predominant secondary structures in proteins is the right-handed α-helix, which can be characterized by a repeating dihedral angle of $\sim 50°$ defined by four consecutive C_α residues in the backbone of real proteins. The diagram in Figure 3.6 shows the approximation for this α-helix dihedral angle on the fcc lattice, which is the angle measured between planes spanned by two consecutive triplets of residues.

$$E_\phi = -\sum_{i=1}^{6} n_{\phi_i} \varepsilon_{\phi_i} \tag{3.9}$$

Shown in Equation 3.9 the torsional energy for the fcc lattice is constructed as a weighted sum over the numbers n_{ϕ_i} of the six possible dihedral angles $\phi_i \in \{0°, 55°, 70°, 109°, 125°, 180°\}$, each with an assigned energetic weight ε_{ϕ_i}. The values ϕ_i can be chosen as a discrete

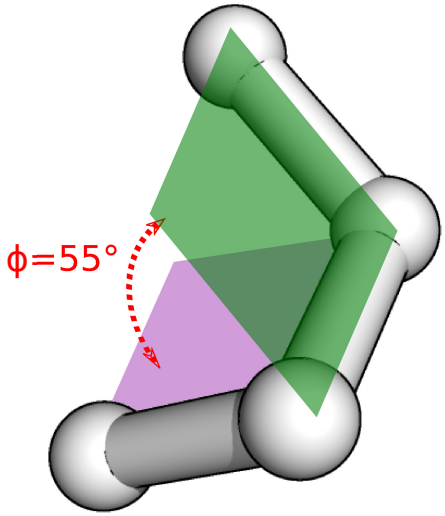

Figure 3.6: Diagram of the dihedral angle that leads to α-helices on the fcc lattice. The first three residues lie in the purple plane and the proceeding three residues the green plane.

approximation for a torsion potential of the form [58] $A\cos\left(\phi - \phi_0\right)$ with a reference angle $\phi_0 = 50°$ that promotes α-helices, or a simpler choice of $\varepsilon_{(\phi=55°)} = A$ with all other $\varepsilon_{\phi_i} = 0$.

3.3.3 β-strand energies

Along with helices, β motifs such as strands, sheets, and turns are important secondary structures found in proteins and peptides. In our fcc lattice aggregation simulations, we include a 4-body energy term from Pokarowski et.al. [59] which promotes the formation of consecutive $\theta = 120°$ bond angles in an extended conformation (with $\phi = 180°$) as a coarse-grained β-strand representation.

$$E_\beta = -\varepsilon_\beta \sum_{i}^{\ell-3} \delta_i \tag{3.10}$$

The β-strand energy, shown in Equation 3.10, simply counts the total number of β-segments in the configuration and weights them by some energetic constant ε_β. Four consecutive

residues are counted as a β-segment when $\delta_i = 1$ according to Equation 3.11; in which case an energetic contribution occurs. Note that the sum in Equation 3.10 has an upper limit of $\ell - 3$, as there are $\ell - 1$ bond vectors in a configuration.

$$\delta_i = \begin{cases} 1 & \text{if} \begin{cases} \vec{b}_i \cdot \vec{b}_{i+1} = \vec{b}_{i+1} \cdot \vec{b}_{i+2} = \frac{1}{2} \\ \vec{b}_i \cdot \vec{b}_{i+2} = 1 \end{cases} \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

The 4-residue arrangement of a β-segment is shown in Figure 3.7, with three consecutive bond vectors $\vec{b}_i$, $\vec{b}_{i+1}$, and $\vec{b}_{i+2}$ in a planar configuration, each with bond angles of $\theta = 120°$.

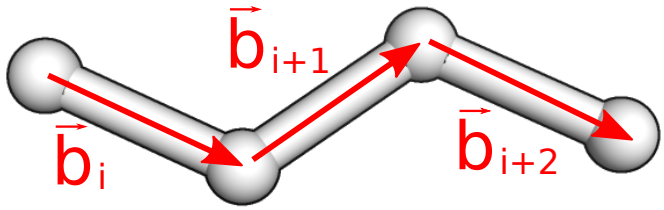

Figure 3.7: Diagram of the β-segment configuration where $\delta_i = 1$ on the fcc lattice.

3.4 Mappings of protein sequences

Any protein can be represented by the HP and H0P models by converting the sequence of amino acids that define the primary structure into a sequence of H's, 0's and P's of corresponding length. There is some variation on which hydrophobicity group each of the amino acids should be classified, but we use the scale reported by Kyte and Doolittle [60]. A multitude of HP sequences have been studied in the literature, both as mappings from real proteins, and for specifically designed benchmarks on the lattice that have a known ground state structure (and degeneracy). In order to make a direct comparison with the results for the fcc lattice, we choose a set of benchmark HP sequences that have been extensively studied on the sc lattice [61], most of which are mappings from real proteins. Table 3.1 lists

these benchmark HP sequences and their aliases, where subscripts in the sequence denote repeated residues (*e.g.* H_3 = HHH), and each alias ends with the length of the sequence (*e.g.* HP103 has 103 residues). The sequences HP67 and HP88 are designed for the sc lattice to

Name	Sequence
[a] HP58	$PHPH_3PH_3P_2H_2PHPH_2PH_3PHPHPH_2P_2H_3P_2HPHP_4HP_2HP_2H_2P_2HP_2H$
[b] HP67	$PHPH_2PH_2PHP_2H_3P_3HPH_2PH_2PHP_2H_3P_3HPH_2PH_2PHP_2H_3P_3HPH_2PH_2$- -$PHP_2H_3P$
HP88	$PHPH_2PH_2PHP_2H_2P_2HP_2HP_2HP_2HP_2HP_2H_2P_2H_3P_2H_3P_2H_3P_2H_3P_2HPH_2P$- -$H_2PHP_2HP_2HP_2H_2P$
[c] HP103	$P_2H_2P_5H_2P_2H_2PHP_2HP_7HP_3H_2PH_2P_6HP_2HPHP_2HP_5H_3P_4H_2PH_2P_5H_2P_4H_4$- -$PHP_8H_5P_2HP_2$
[d] HP124	$P_3H_3PHP_4HP_5H_2P_4H_2P_2H_2P_4HP_4HP_2HP_2H_2P_3H_2PHPH_3P_4H_3P_6H_2P_2HP_2H$- -$PHP_2HP_7HP_2H_3P_4HP_3H_5P_4H_2PHPHPHPH$
[e] HP136	$HP_5HP_4HPH_2PH_2P_4HPH_3P_4HPHPH_4P_{11}HP_2HP_3HPH_2P_3H_2P_2HP_2HPHP$- -$HP_8HP_3H_6P_3H_2P_2H_3P_3H_2PH_5P_9HP_4HPHP_4$

Table 3.1: HP sequences for: [a] Bovine Pancreatic Trypsin Inhibitor (BPTI) [62], [b]benchmarks with low-degeneracy ground states on the sc lattice [39,63], [c] Cytochrome C (apo form) [64], [d] Ribonuclease A [64], [e] 1-136 Staphylococcal nuclease fragment [64]. Subscripts in the specified sequences signify consecutively repeated residues.

have threefold degenerate and unique ground state structures, respectively. All other HP sequences are mapped from real proteins and protein fragments that are listed in the caption of Table 3.1.

Additionally, we use the HP and H0P models for the short, 46-residue Crambin protein. The H0P model for Crambin has been studied previously on the sc lattice [46], and the addition of semiflexibility through bond angle energies was able to recover a unique ground state structure [50]. We use this sequence with the fcc H0P model to compare folding thermodynamics with and without bond angle energies to those found on the sc lattice, and as a test for the addition of torsion energies described in Section 3.3. Table 3.2 shows the corresponding model sequences.

Name	Sequence
HP46	$P_2H_2P_2H_3P_3HPH_2PHP_5H_4PHP_2H_4P_2HPHP_3H_2P$
H0P46	$0_2H_20_2H_20P0PHPH_2PH0_4P0H_20_50_4H0_2P0_2P$

Table 3.2: HP and H0P mappings of the 46-residue Crambin protein.

3.4.1 Structural properties

In order to understand the structural transitions that occur during the folding simulations, several observables are calculated to supplement the measured energetic quantities. One useful measure of spatial fluctuations is the gyration tensor $S_{\alpha\beta}$, which is similar to the moment of inertia, but for all residues with equal mass.

$$S_{\alpha\beta} = \frac{1}{\ell}\sum_{i=1}^{\ell}(r_\alpha^{(i)} - r_\alpha^{cm})(r_\beta^{(i)} - r_\beta^{cm}) \tag{3.12}$$

$$\vec{r}^{\,cm} = \frac{1}{\ell}\sum_{i=1}^{\ell}\vec{r}^{\,(i)} \tag{3.13}$$

The calculation of $S_{\alpha\beta}$ is shown in Equation 3.12, where α and β are indices for x, y, z components and $\vec{r}^{\,cm}$ is the center of mass. The gyration tensor can be diagonalized as shown in Equation 3.14, such that the diagonal elements, or principal moments, are arranged as $\lambda_1 \geq \lambda_2 \geq \lambda_3$.

$$\mathbf{S} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \tag{3.14}$$

From the principal moments, various shape descriptors can be calculated [65]. The radius of gyration (R_g), shown in Equation 3.15, gives a measure of the standard deviation of residues about the center of mass.

$$R_g = \sqrt{\lambda_1 + \lambda_2 + \lambda_3} \tag{3.15}$$

The relative shape anisotropy (κ^2), shown in Equation 3.16, describes the symmetry and shape of a configuration with a value of 0 for a perfectly spherical configuration, and a value of 1 if the configuration is linear.

$$\kappa^2 = 1 - 3\left(\frac{\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3}{(\lambda_1 + \lambda_2 + \lambda_3)^2}\right) \tag{3.16}$$

Another simple structural observable, shown in Equation 3.17, is the end-to-end distance (R_{EE}), or the linear distance between first and last residues in the chain.

$$R_{EE} = \left|\vec{r}_1 - \vec{r}_\ell\right| \tag{3.17}$$

The number of contacts n_{HH}, n'_{HH}, n_{H0}, n_{PP} and n'_{PP} are also useful measurements. In aggregation simulations, the cluster sizes (m) and number of clusters (N_C) are calculated using a depth-first search (DFS) shown in Appendix A1-2. The calculation of thermodynamic averages and other quantities are discussed in Section 4.3.

Chapter 4

Monte Carlo Methodology

4.1 Equilibrium Monte Carlo sampling

MC methods work by generating random samples from a system's phase space and using them to calculate estimates of statistical quantities with an ensemble of states. For a probability distribution $P_i(t)$ that transitions between states $i \to j$ (corresponding to a discrete, fictitious MC time step Δt) with the transition rate $w_{i\to j}$, the condition for that distribution to be in equilibrium [66] is shown in Equation 4.1.

$$\frac{\Delta P_i(t)}{\Delta t} = \sum_j \Big(P_i(t) w_{i\to j} - P_j(t) w_{j\to i} \Big) = 0 \tag{4.1}$$

These transitions are implemented by a set of system-specific configuration updates, or trial moves, that should be efficient in traversing phase space while preserving stationarity of the distribution P_i (for which we can drop the time-dependence). The latter requirement is satisfied by the detailed balance condition that is shown in Equation 4.2, where the transition rates w become transition probabilities W.

$$P_i W_{i\to j} = P_j W_{j\to i} \tag{4.2}$$

Other requirements are that the trial moves are ergodic, meaning that every state of the system is able to be reached from any other state by a finite sequence of the moves, and the moves are reversible, meaning that the starting state i is also reachable by applying the move from the final state $j \ \forall \ (i \to j)$.

The transition probability $W_{i\to j}$ can be written as a product of the probability $\wp_{i\to j}$ for proposing the transition $i \to j$ with the probability $\alpha_{i\to j}$ for accepting the proposal, as shown in Equation 4.3.

$$W_{i\to j} = \wp_{i\to j}\alpha_{i\to j} \tag{4.3}$$

A famous solution for the acceptance probability is the Metropolis-Hastings [67, 68] step, shown in Equation 4.4, which combines Equations 4.3 and 4.2 while enforcing normalization.

$$\alpha_{i\to j} = \min\left(1, \ \frac{\wp_{j\to i}}{\wp_{i\to j}}\frac{P_i}{P_j}\right) \tag{4.4}$$

Under this scheme, transitions between states only depend on the current and proposed state, and sample the target probability distribution with what is called Markov Chain MC, or MCMC. Trial move sets that are ergodic and fulfill detailed balance are sufficient to avoid the introduction of systematic errors which may arise by erroneously altering the target distribution or 'locking out' regions of phase space.

4.2 Trial moves for fcc lattice models

The set of trial moves detailed here are described for polymer and protein backbone models on the fcc lattice, but analogous moves are used for the sc lattice simulations as well. Some differences in implementation between the two lattices will be discussed where applicable. In the following moves, detailed balance can be preserved by incorporating the ratio $\wp_{j\to i}/\wp_{i\to j}$ (see Equation 4.4) using the number of forward and backward transitions

between states, but we are able to satisfy this using an unbiased scheme that attempts all possible moves (whether valid or not) at a constant probability [61,69].

4.2.1 Pull moves

A MC move that is designed for entangled lattice backbone configurations, pull moves are powerful local moves which start with the displacement of one or more residues and 'pull' portions of the chain along previously occupied positions. For the fcc lattice, there are a few variants of the move described in the literature [70, 71], however, we use the original description [72] which displaces one residue at the start of the move. There are two types of pull moves: those which act at the first (last) residues in the chain, called end pull moves, and those which act on residues in $[2, \ell - 1]$, called internal pull moves. A move attempt begins by using a pseudo-random number generator to choose a residue j with uniform probability along the chain length. Shown by the blue arrow in Figure 4.1, an internal

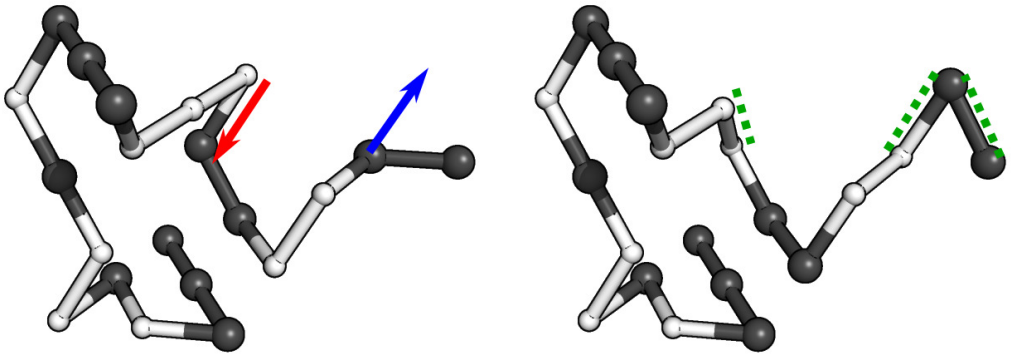

Figure 4.1: Single site pull move for the fcc lattice. The left image shows the initial trial displacement with a blue arrow, and the kink in the chain that is eliminated as a result with a red arrow. The right image shows the new bond vectors with green dashes that are assigned after the move.

pull move displaces a residue by one lattice site that is a mutual neighbor of the initial positions of j and $j \pm 1$. The displacement site and pull direction ($\pm$) are chosen at random with uniform probability. Next, the chain connectivity is checked, and if broken, the first disconnected residue $j \mp 1$ is set to the previous position of residue j. This procedure of

moving the next disconnected residue to the previous position of its proceeding (preceding) neighbor continues until the connectivity is recovered by destroying (signified by the red arrow in Figure 4.1) a kink in the chain, or the end of the chain is reached. It may be the case that a single residue displacement leaves the chain in a valid configuration, which is equivalent to the local moves known as kink-flip and crankshaft moves. When an end pull move is attempted, random lattice direction is again chosen with uniform probability, but only moves where the displaced position of j is not adjacent to residue $j \pm 1$ are accepted.

For the sc lattice, the implementation differs in that two consecutive residues must be displaced at the start of the move parallel to one another to preserve a linear chain [73]. Another difference which arises due to the different geometric constraints imposed by the sc lattice is that the end pull moves will not result in the first (last) residue being adjacent to $j \pm 1$. Also, there is one type of move on the sc which forms a 'hook' and is explicitly forbidden to assure reversibility [74]. This troublesome configuration does not occur for the fcc lattice.

4.2.2 Diagonal moves

A local MC move that is a subset of single-site pull moves is implemented separately to enhance sampling. Called 'diagonal moves' [75] on the fcc lattice, these are traditionally referred to as kink-flip or crankshaft moves in the literature. The moves work by displacing a residue, chosen with uniform probability, that remains connected to its preceding and succeeding residues in the chain after displacement. A sample diagonal move is illustrated in Figure 4.2, where the blue arrow in the left image shows a trial displacement, and the green, dotted lines show two bonds that are re-assigned after the move is accepted.

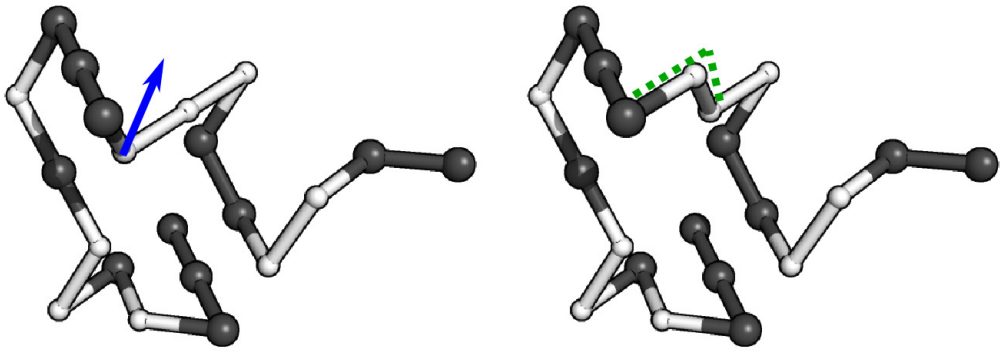

Figure 4.2: Diagonal move for the fcc lattice. The left image shows the initial trial displacement with a blue arrow. The right image shows the relabeled sequence and the new bond vectors with green dashes that are assigned after the move.

4.2.3 Reptation moves

To start a reptation move [76], a terminal residue (either 1 or ℓ) is selected randomly and deleted. Next, a single-site displacement is attempted at the opposite end of the chain, and the primary sequence is relabeled, if necessary. It is worth noting that these moves do not satisfy ergodicity alone [77], but are combined with the other MC trial moves described in this chapter to ensure ergodicity. This effect of broken ergodicity is less of an issue when dealing with high-coordination lattices which have fewer states that become 'trapped' as a result of the moves [77].

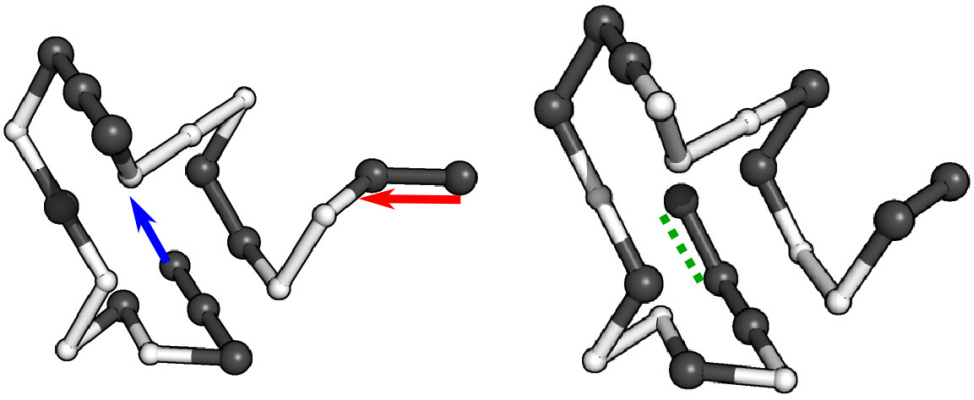

Figure 4.3: Reptation move for the fcc lattice. The left image shows the initial deletion with a red arrow and the trial displacement at the opposing end with a blue arrow. The right image shows the relabeled sequence and the new bond vectors with green dashes that are assigned after the move.

These moves appear to rapidly diminish in acceptance at lower energies when simulating HP and H0P models due to the additional energetic change associated with relabeling the sequence, but retain a high efficacy when simulating ISAWs. A sample reptation move is shown in Figure 4.3, where the red arrow in the left image shows the bond that is deleted initially, and the blue arrow shows a random trial displacement at the other end of the sequence. The new state, in the case of an accepted move, is shown in the right image, where the sequence has been relabeled and the green, dashed line shows the new bond that is accepted at the end involving a displacement.

4.2.4 Pivot moves

Besides local moves, it is beneficial to include global moves that result in concerted movements of large portions of the configuration at one time. One such move, called the pivot move [70, 78], simply consists of applying a random rotation to a portion of the configuration. The move starts by selecting a random residue which acts as the pivot point, where a portion of the chain before or after (also randomly selected) is rotated by some random angle that is allowed on the lattice. The application of a pivot move is shown in Figure 4.4, where

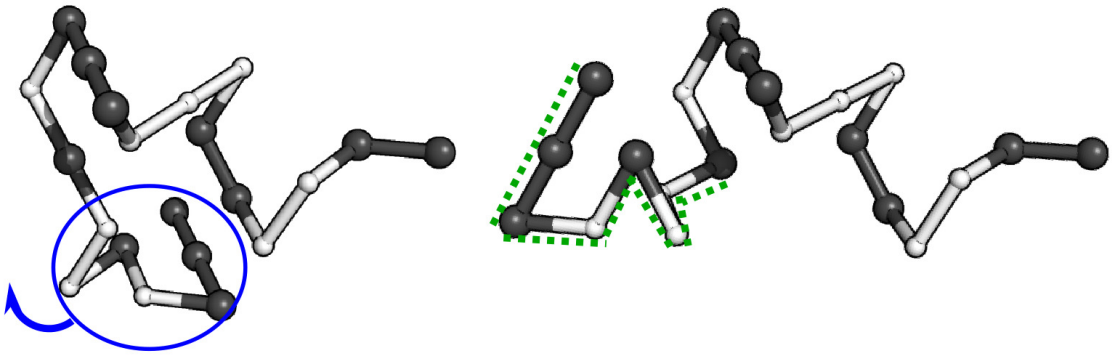

Figure 4.4: Pivot move for the fcc lattice. The left image shows the selected portion for rotation circled in blue. The right image shows the new bond vectors with green dashes that are assigned after the move is accepted.

a portion of the configuration, shown in the blue circle in the right image, is rotated by a random angle on the lattice. After self-avoidance is verified for the rotation, the bonds shown

with dashed, green lines in the right plot are re-assigned according to the rotation operation that was chosen. A major convenience of lattice simulations is that the rotation matrices can be easily enumerated before the start of the simulation and applied using a lookup table. While the implementation of pivot moves is kept very simple in our simulations, a more complex version has been described by Clisby et al. [79], which becomes very efficient when simulating extremely long SAWs.

4.2.5 Bond-rebridging moves

Another set of global moves, called bond-rebridging moves [80, 81], involves the rearrangement of bonds while leaving the coordinates of all residues fixed. Three types of bond-rebridging moves are used in our simulations: two types of internal rebridging moves, and end rebridging moves, a.k.a. 'backbite' [82] or Hamiltonian paths [83]. Figure 4.5 shows an example of a type 2 internal rebridging move, where the linear topology of the chain is preserved by one application of cutting and reforming pairs of bonds. The move starts by

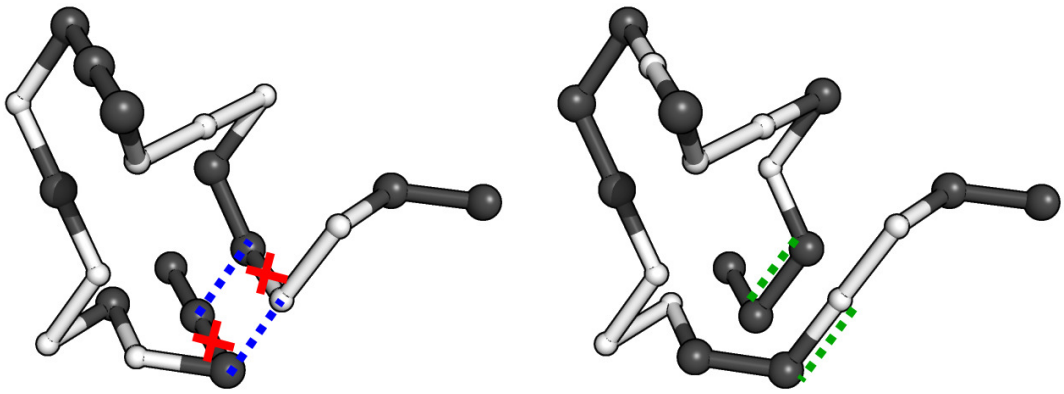

Figure 4.5: Type 2 internal rebridging move for the fcc lattice. The left image shows deleted bonds with red x's and the new, proposed bonds with blue dashes. The right image shows the relabeled sequence and the new bond vectors with green dashes that are assigned after the move.

choosing a random residue $n \in [2, \ell - 1]$, at which the first bond $\in [\vec{b}_2, \vec{b}_{\ell-2}]$ is deleted. Next, two random lattice directions $\vec{c}, \vec{c}\,'$ (shown as blue, dashed lines in the right of Figure 4.5) are chosen to query for any two consecutive residues $j, j \pm 1$ with bond $\vec{b}_j$ (or $\vec{b}_{j-1}$) that

are adjacent to residues $n, n \pm 1$, for which the bonds $\vec{b}_n$, $\vec{b}_j$ (or $\vec{b}_{j-1}$), $\vec{c}$, $\vec{c}\,'$ form a closed loop. If this condition holds, then bonds $\vec{b}_n$ and $\vec{b}_j$ (or $\vec{b}_{j-1}$) are deleted (shown by the red x's in the left of Figure 4.5) and replaced by $\vec{c}$ and $\vec{c}\,'$ (the green, dashed lines in the right of Figure 4.5). For type 1 internal rebridging moves, this initial procedure leaves the backbone as a closed loop and disconnected linear segment, and the described procedure must be applied again at a random section of the closed loop. See Appendix A4 for a description of such move. After the bonds are successfully manipulated, the sequence must be relabeled to preserve primary structure. One difference between this move on the fcc and sc lattices is that the vectors $\vec{c}, \vec{c}\,'$ not needing to be parallel for the fcc lattice. As a result, both type 1 and type 2 moves can be valid options simultaneously on the fcc lattice, in which case, one is chosen with uniform probability.

For end rebridging moves that act on residues $n = 1$ or ℓ, only one bond is deleted and reformed. The move starts by choosing one end of the sequence randomly and then picking a random lattice direction $\vec{c}$ (shown as the blue, dashed line in the right of Figure 4.6) to search for an adjacent residue j where $|n - j| > 2$. If a residue j is found, then the first bond

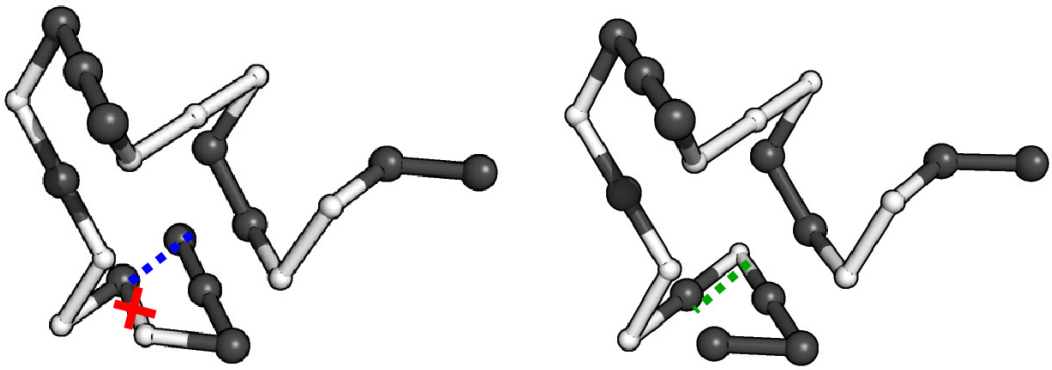

Figure 4.6: End rebridging move for the fcc lattice. The left image shows the deleted bond with a red x and the new, proposed bond with blue dashes. The right image shows the relabeled sequence and the new bond vector with green dashes that are assigned after the move.

$\vec{b}_j$ (or $\vec{b}_{j-1}$) between j and n is deleted and replaced by $\vec{c}$, as shown by the green, dashed line in the left of Figure 4.6. As with internal rebridging moves, the sequence is relabeled to preserve the primary structure.

4.2.6 Random translation moves

During aggregation simulations, it is helpful to introduce rigid body moves to sample translational degrees of freedom. To preserve long-range order that forms among groups of several interacting subunits, the random translation moves are performed for an entire cluster at one time, as shown in Figure 4.7. For the sake of simplicity, we implement cluster translation moves in a way that does not create or destroy new clusters - meaning that the energy remains constant. To start the move, a random cluster is chosen uniformly using

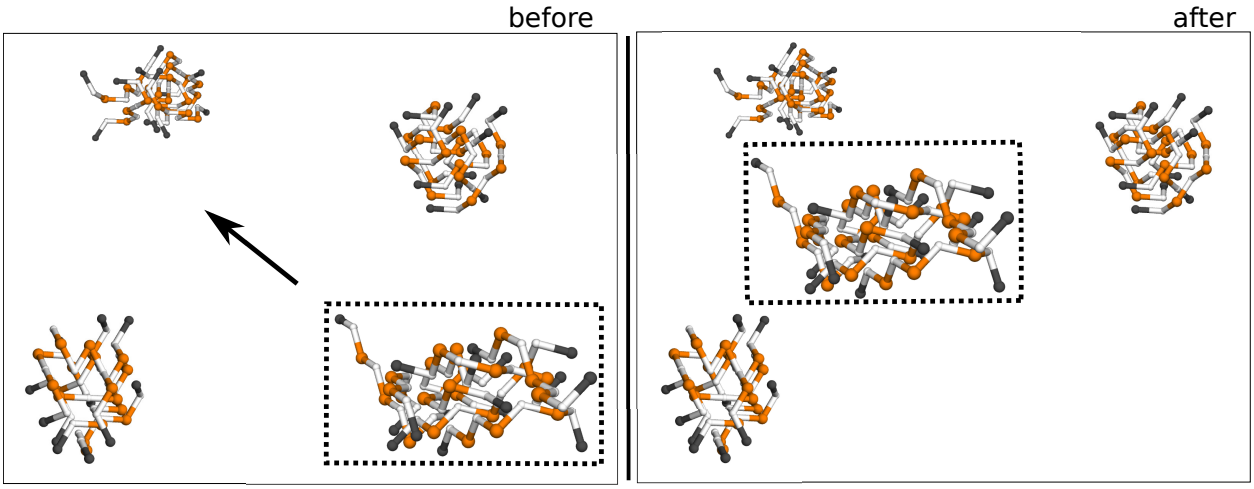

Figure 4.7: Cluster translation move for the fcc aggregation model. The randomly chosen cluster is shown enclosed in a dashed box, and is displaced in the random direction shown by the arrow in the left image. After the move, the cluster in the right image has a different position relative to the other three clusters.

a depth-first search (DFS) (see Appendix A1). A displacement vector is then chosen with random magnitude and direction, and the self-avoidance is queried for the displacement of the chosen cluster. If there is no steric clash and no clusters are created or destroyed (resulting in an energy change), then the move is accepted.

4.2.7 Rigid rotation moves

Similar to cluster translation moves, rigid-body rotations of clusters are included in aggregation simulations. Again, rotation moves are only accepted in which the energy remains constant and no new clusters are created or destroyed. A random cluster is selected using

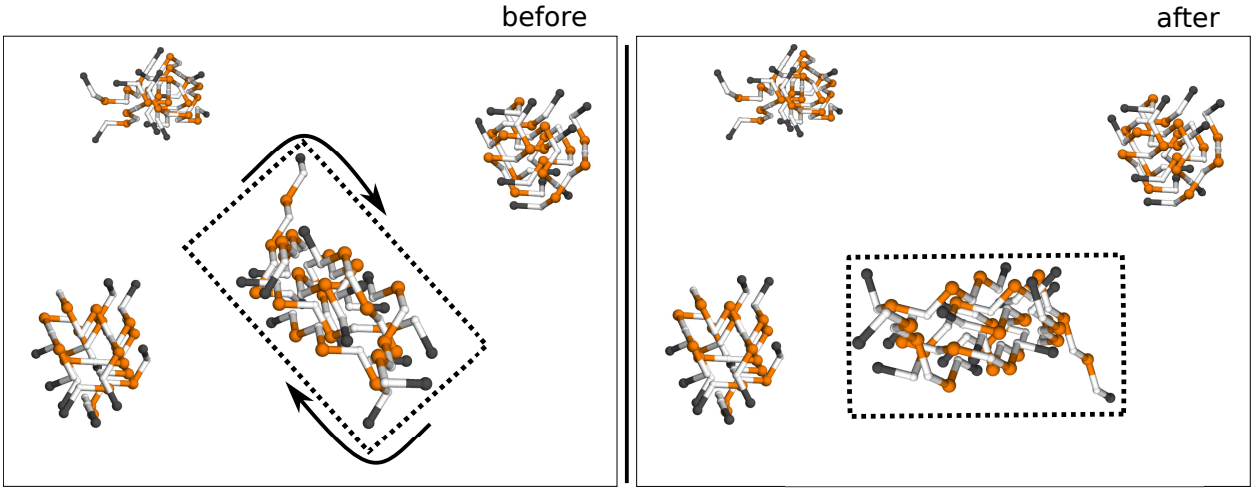

Figure 4.8: Cluster rotation move for the fcc aggregation model. The chosen cluster is shown enclosed in the dashed box, and a random rotation operation about a random axis is applied; shown by the two arrows in the left image. After the move, the cluster in the right image has a different orientation relative to the other three clusters.

a DFS, and a random rotation operation is applied to all of the cluster's subunits about a random axis. The same pre-enumerated rotation matrix used in pivot moves can be utilized, but all subunits in the chosen cluster retain their relative orientation with one another after the move, as shown in Figure 4.8.

4.3 Calculation of thermodynamic averages

The theory laid out in Section 4.1 details how samples can be generated for an equilibrium distribution in general. In statistical physics, the canonical ensemble describes a system in equilibrium at some fixed number of particles N, volume V, and temperature T, where the

probability for any microstate $\boldsymbol{X}$ is given by the distribution:

$$P(\boldsymbol{X}, T) = \frac{e^{-E(\boldsymbol{X})/(k_B T)}}{\mathcal{Z}(T)} \ . \tag{4.5}$$

This expression depends on the energy (E) for a specified microstate ($\boldsymbol{X}$), the Boltzmann constant (k_B), and is normalized by the partition function ($\mathcal{Z}$). Shown in left-most equality of Equation 4.6, the partition function is defined as a sum (for a discrete state space) over all possible microstates of the system, which is typically an intractable calculation due to the enormous number of possible microstates.

$$\mathcal{Z}(T) = \sum_{\boldsymbol{X}} e^{-E(\boldsymbol{X})/(k_B T)} = \sum_{E} g(E)\ e^{-E/(k_B T)} \tag{4.6}$$

It is beneficial to introduce the quantity $g(E') = \sum_{\boldsymbol{X}} \delta(E' - E(\boldsymbol{X}))$, or density of states, and instead calculate $\mathcal{Z}$ with a sum over energy states, as shown in the right-most equality of Equation 4.6. If $g(E)$ is known, from which $\mathcal{Z}$ becomes a tractable calculation, the probability for the system at any energy and temperature is given directly by:

$$P(E, T) = \frac{g(E)\ e^{-E/(k_B T)}}{\mathcal{Z}(T)} \ . \tag{4.7}$$

From the canonical probabilities shown in Equation 4.7, average thermodynamic properties, or ensemble averages, for any observable quantity ($\mathcal{O}$) can be calculated as

$$\langle \mathcal{O}(T) \rangle = \sum_{E} \overline{\mathcal{O}}(E)\ P(E, T) \ , \tag{4.8}$$

where angled brackets $\langle ... \rangle$ denote ensemble averages, and the quantity $\overline{\mathcal{O}}(E)$ in Equation 4.9 is the weighted average of the observable $\mathcal{O}$ at energy E. Equation 4.9 shows how the value of $\overline{\mathcal{O}}(E)$ is measured in the context of a numerical simulation which uses a two-dimensional

histogram $H(E, \mathcal{O})$ collected from a finite number of samples from the target distribution.

$$\overline{\mathcal{O}}(E) = \frac{\sum_{\mathcal{O}} \mathcal{O}\ H(E, \mathcal{O})}{\sum_{\mathcal{O}} H(E, \mathcal{O})} \tag{4.9}$$

Our goal for the MC simulations performed in this dissertation is to determine $g(E)$ for the model systems in question by sampling microstates and utilize the statistics outlined here to study average temperature-dependent behavior. Two thermodynamic quantities that can be calculated from the values in Equation 4.7 are the heat capacity C_V and the Helmholtz free energy F. Internal energy, $U = \langle E \rangle$, is defined as the average energy, and the heat capacity is related to its fluctuation:

$$C_V = \frac{1}{k_B T^2} \Big[\langle E^2 \rangle - \langle E \rangle^2 \Big] \ . \tag{4.10}$$

Signals in the curve of C_V vs. T, such as peaks and shoulders, are indicative of structural transitions that occur in the model systems. The free energy is an important quantity that describes the thermodynamic stability for a macrostate at some specified temperature, where the minimum of F corresponds to the most favorable state. Shown in Equation 4.11, the free energy has contributions from an energetic term U and an entropic term TS, where the entropy S is related to the microscopic properties of the system by $S = k_B \ln[g(E)]$.

$$F = U - TS \tag{4.11}$$

$$= -k_B T \ln(\mathcal{Z}) \tag{4.12}$$

There is a system-dependent and potentially complex competition between the minimization of U and maximization of S.

Free energy ($F(\mathcal{O})$) as a function of some additional observable quantity $\mathcal{O}$ is given by the following relation:

$$F(\mathcal{O}) = -k_B T \ln[\mathcal{Z}(\mathcal{O})] \ . \tag{4.13}$$

The partition function $\mathcal{Z}(\mathcal{O}) = \sum_E g(E,\mathcal{O})e^{-E/(k_BT)}$ involves a joint density of states $g(E,\mathcal{O})$ which is calculated using the reweighting $g(E,\mathcal{O}) = g(E)H(E,\mathcal{O})$.

4.4 Wang-Landau sampling

The Wang-Landau (WL) algorithm [84, 85] iteratively calculates an estimate for the density of states (or multiplicity) ($\hat{g}(E)$) using a random walk in energy space. Rather than sampling one or more temperature distributions, WL is an athermal procedure that samples a target distribution that approaches $g(E)$ for all energies of the system. Throughout each iteration, a histogram ($H(E)$) is kept to record a history of energy states visited by the random walker, with the goal of reaching a flat histogram - each energy state is visited equally regardless of potentially small probabilities. In practice, there are multiple options for determining what constitutes a 'flat' $H(E)$ [86], but a common method is to consider an average flatness criterion that is satisfied whenever ($H(E) \geq p \cdot \overline{H}(E) \quad \forall \quad E$), given some user-defined flatness measure p. The fulfillment of the flatness criterion signals the end of each iteration, at which point the histogram is reset $H(E) = 0 \ \forall \ E$.

To obtain uniform sampling over energy space, a walker must visit the various states with a transition acceptance probability that is inversely proportional to $g(E)$, as shown in Equation 4.14 from Algorithm 1. The instantaneous estimate $\hat{g}(E)$ guides the WL walker towards regions of less-visited state space, and is dynamically accumulated using a modification factor (f) that decreases after every iteration according to some schedule [87,88], where the original algorithm uses $f \leftarrow \sqrt{f}$. A simulation starts with an initial guess for $\hat{g}(E)$, that when no information is known *a priori*, is uniform for all energies $\ln[\hat{g}(E)] = 0 \quad \forall \quad E$. An adaptive scheme [61] is used that adds new energies to the simulation as they are found, where the new value of $\hat{g}(E)$ is set to the minimum existing value and $H(E)$ is reset for all energies. Sampling iterations repeat until the termination criterion ($f \leq f_{final}$) is satisfied. The procedure for WL sampling is summarized in Algorithm 1.

Algorithm 1 The WL algorithm

(1) Start with $f = e$ and $\hat{g}(E) = 1 \; \forall E$

(2) Using MC trial moves, randomly transition between states $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ with probability

$$\alpha^{WL}(\boldsymbol{X}_A \to \boldsymbol{X}_B) = \min\left(1, \; \frac{\hat{g}(E_A)}{\hat{g}(E_B)}\right) \tag{4.14}$$

(3) After each trial move, update $\hat{g}(E) \leftarrow f \cdot \hat{g}(E)$ and $H(E) \leftarrow H(E) + 1$

(4) If $H(E) \geq p \cdot \overline{H(E)} \; \forall E$ is satisfied, set $H(E) = 0 \; \forall E$ and let $f \leftarrow \sqrt{f}$

(5) Continue to iterate steps (2)-(4) until $f \approx 1$ within some user-defined threshold (*e.g.* $\ln f \leq 10^{-8}$)

Because the transition probability in Equation 4.14 depends on the current estimate $\hat{g}(E)$ which has a history of visited states, the process does not strictly satisfy detailed balance in the presented scheme. During the later iterations, however, the algorithm asymptotically converges towards an ensemble which does satisfy detailed balance, but with a systematic error that is controlled by the rate in which f is reduced [89] and the flatness measure p [90]. The WL algorithm can be used in tandem with the multicanonical sampling algorithm [91], described in Section 4.6, where the resulting $\hat{g}(E)$ can be verified or reweighted with a procedure that does satisfy detailed balance. From a single WL simulation, one can recover average properties at all temperatures by a statistical reweighting that is outlined in Section 4.3. The estimate $\hat{g}(E)$ obtained from WL proportional to the true $g(E)$ by some arbitrary normalization constant, and only gives relative information about the degeneracy at each state.

4.5 Replica-exchange Wang-Landau sampling

A parallel version of WL sampling, the replica-exchange WL (REWL) algorithm [92,93], can be used to determine $\hat{g}(E)$ for systems that would otherwise be prohibitively large with the serial implementation [94]. The method uses a distributed scheme where every 'replica'

is a WL instance with its own computing core that samples within an assigned 'window', or constrained region of state space. REWL is versatile in that it can be applied for state space of arbitrary dimension [94], but we use the one-dimensional setup in energy space. The energy range is subdivided by multiple overlapping windows, as shown by the rectangular boxes in Figure 4.9. Windows constrain one or more replicas to an energy range, where MC moves that put the replicas out of the assigned bounds are rejected. Each replica (i) has its own $\hat{g}_i(E)$ and $H_i(E)$ and must reach a flat histogram independently. For windows with multiple replicas, a single value of the modification factor f is used, and the average of $\hat{g}_i(E)$ between replicas is adopted by all replicas within the window at the end of each iteration. REWL has the advantage of converging the $\hat{g}_i(E)$ independently for much smaller regions of energy space compared to serial WL.

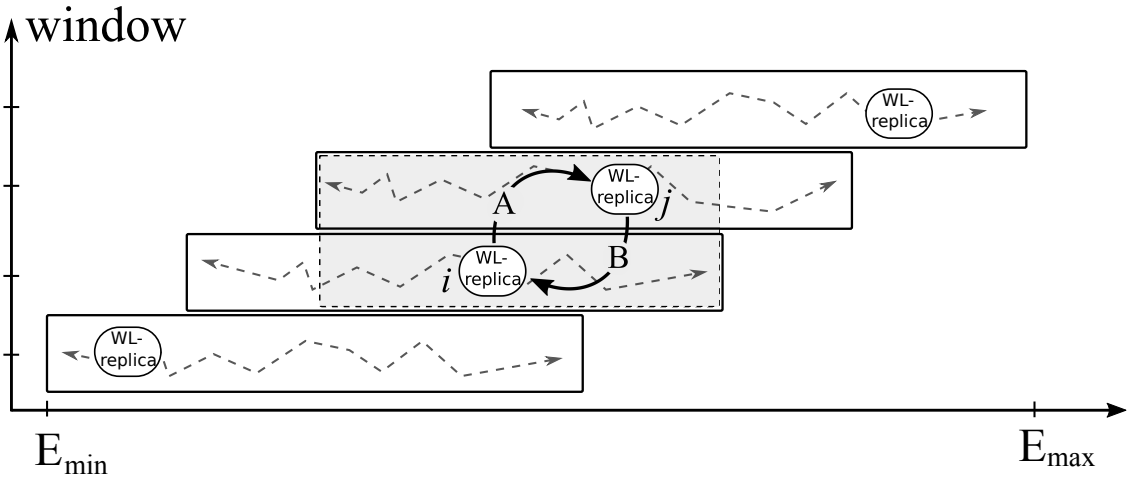

Figure 4.9: Schematic diagram of the REWL scheme for 4 energy windows. Ovals with dashed arrows denote random walks performed by replicas within windows. An exchange is shown between replicas i and j in the overlap region of the middle two windows (shaded in gray), where configurations A and B are swapped.

Configuration exchanges between two random replicas from neighboring windows are attempted throughout the REWL simulation with some fixed frequency. For any two replicas i, j from neighboring windows, the probability for accepting an exchange of their respective configurations $\boldsymbol{X}_A$, $\boldsymbol{X}_B$ is given by Equation 4.15, and depends only on the instantaneous

estimates $\hat{g}_i(E)$, $\hat{g}_j(E)$ of the two replicas at the energies E_A and E_B.

$$\alpha_{i,j}^{exch.}(\boldsymbol{X}_A \leftrightarrow \boldsymbol{X}_B) = \min\left(1,\ \frac{\hat{g}_i(E_A)}{\hat{g}_i(E_B)} \cdot \frac{\hat{g}_j(E_B)}{\hat{g}_j(E_A)}\right) . \tag{4.15}$$

Replica exchange attempts alternate between adjacent pairs of windows, and preserve ergodicity by allowing the replicas to swap configurations across state space. Additionally, replicas benefit from improved mixing within their respective windows as exchanges are accepted. The exchange frequency should ideally be chosen large enough so that the replicas have the opportunity to diffuse through the states in their windows and avoid a systematic effect. An example is provided in Appendix A6 to elaborate on this further. A REWL simulation terminates once all of the windows separately satisfy the criterion $f \leq f_{final}$ for some chosen f_{final}. Each of the $\hat{g}_i(E)$ are independently normalized and can be merged together after the simulation using the procedure described at the end of this section.

The initialization of REWL can be a little tricky. How can a valid state for a sequestered part of phase space be chosen to start of the WL instance inside each window, especially for those where $|E|$ is large? Appendix A7 gives a detailed description of two possible initialization methods: one that reads in an initial state from the user, and another that starts the simulation blindly using a scheme similar to WL sampling to enter each window bounds.

4.5.1 Merging REWL results

After each of the REWL windows converges with its own $\hat{g}_i(E)$, the results are independently normalized. A method for merging the $\hat{g}_i(E)$ is needed, and although there could be other options for such a procedure [95], a simple method based on enforcing the idea of continuity is employed. The procedure begins by calculating the absolute difference in the inverse microcanonical temperature $\Delta\beta_{i,i\pm1}(E) = \left|\frac{d\hat{g}_i(E)}{dE} - \frac{d\hat{g}_{i\pm1}(E)}{dE}\right|$ for two adjacent windows i, $i \pm 1$. When any two windows are merged, the point at which they are shifted together is

chosen to be $E_{merge} = \text{argmin}_E\Big(\Delta\beta_{i,i\pm1}(E)\Big)$. For example, the $\hat{g}(E)$ is constructed starting

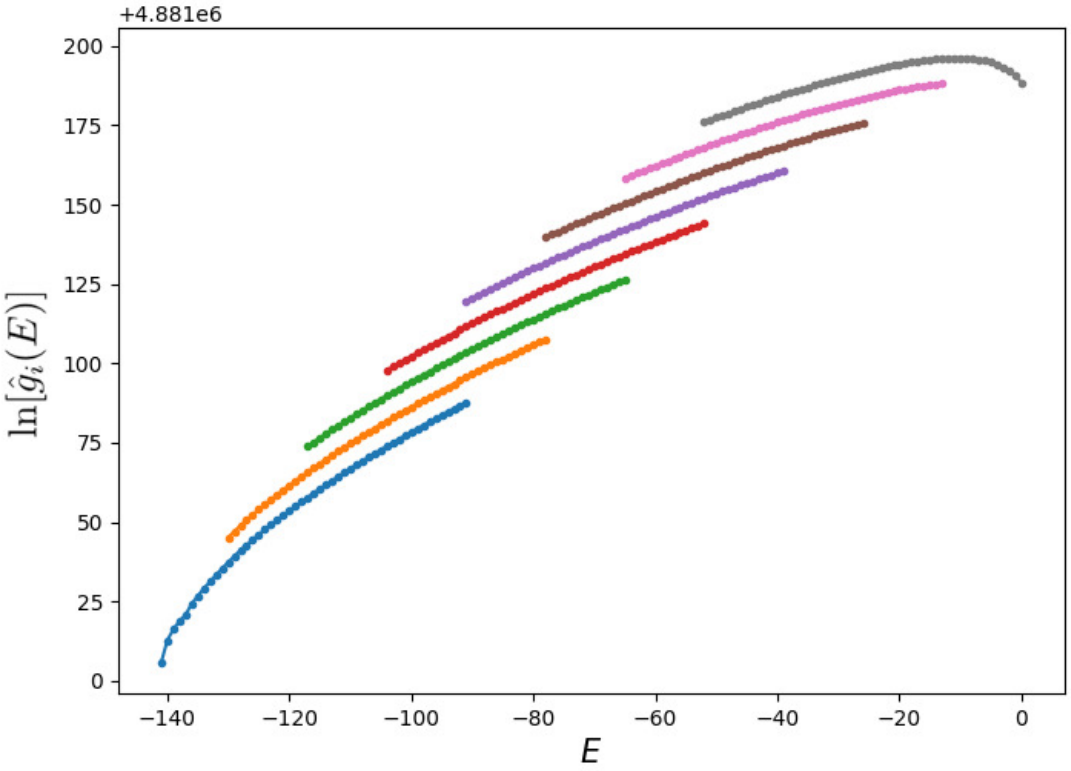

Figure 4.10: Un-merged density of states for a REWL simulation with 8 windows for HP88 on the fcc lattice.

from the first window using consecutive applications of:

$$\hat{g}(E) = \begin{cases} \hat{g}_i(E), & E < E_{merge} \\ \hat{g}_i(E)\left(\dfrac{\hat{g}_{i+1}(E_{merge})}{\hat{g}_i(E_{merge})}\right), & E \geq E_{merge} \end{cases} \tag{4.16}$$

for each pair $(i,\, i+1)$. Figure 4.10 shows an example of the $\ln[\hat{g}_i(E)]$ before the merging process is performed. In practice, the difference between each $\ln[\hat{g}_i(E)]$ can be much larger than what is shown, but this has no impact on the merging procedure.

4.6 Multicanonical sampling

The multicanonical method MUCA [96] is a powerful MC algorithm that has some similarities with, but predates, WL and REWL. Also a flat histogram method, MUCA uses a set of sampling weights W^{MUCA} that are ideally chosen to augment the canonical probability such that $W^{MUCA}(E,T)P(E,T) = Constant$. Temperature is not a requirement for the sampling, and can be set as $T \to \infty$, in which case $P(E,T) \propto g(E)$ and the ideal choice of weights is apparent as $W^{MUCA}(E) \propto 1/g(E)$ [13]. The major difference from WL is that MUCA strictly satisfies the detailed balance condition from Equation 4.2, as the weights are fixed during sampling. It is often beneficial to use WL or REWL to estimate W^{MUCA} and use MUCA to gather statistics from the target distribution with a large production run [91].

We utilize the MUCA algorithm and a parallel version, that uses the same windowing setup described for REWL, to perform a final reweighting of $\hat{g}(E)$, and to calculate ensemble averages of structural quantities using the methods described in Section 4.3. In our simulations, weights are held fixed as $W^{MUCA}(E) = 1/\hat{g}(E)$ using the $\hat{g}(E)$ obtained by REWL, and long production runs ($H(E) \geq 10^8 \ \forall \ E$) are performed with trial moves from Section 4.2.

The probability to accept a transition between states $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ is given by

$$\alpha^{MUCA}(\boldsymbol{X}_A \to \boldsymbol{X}_B) = \min\left(1, \ \frac{W^{MUCA}(E_A)}{W^{MUCA}(E_B)}\right) \ , \tag{4.17}$$

and if a replica-exchange scheme is utilized, then the exchange probability given in Equation 4.15 is used. After the samples are accumulated from a MUCA, the estimate $\hat{g}(E)$ is reweighted as

$$\hat{g}'(E) = \hat{g}(E)H(E) \tag{4.18}$$

to obtain a final estimate of the density of states $\hat{g}'(E)$. This reweighting can also be performed iteratively with batches of samples in a fashion that strictly satisfies detailed balance during each iteration [13], although the convergence of W^{MUCA} with this method can be

very difficult in practice. When calculating the ensemble average of a structural property $\mathcal{O}$, a two-dimensional histogram $H(E, \mathcal{O})$ is accumulated and used as in Equation 4.9.

4.7 Practical details for simulations

In all simulation, results for $\hat{g}(E)$ are obtained using the REWL algorithm with anywhere between 4-16 windows, and one replica per window. Typical values for the number of MC sweeps (a MC move attempted for each residue in the system) between checks for histogram flatness is around 100,000 sweeps. We typically find a flatness measure $p = 0.8$ and termination criterion $\ln f = 1 \times 10^{-8}$, but also find values of $p = 0.6$ and $\ln f = 1 \times 10^{-6}$ to provide results comparable to within statistical error. While system dependent, a typical value for the replica-exchange frequency is every 5,000 sweeps, although much lower frequencies are used in our aggregation simulations. After each simulation, a MUCA run is used to reweight $\hat{g}$ and also is used for the calculation of all presented structural properties, where the number of samples is at least $H(E) \geq 10^8$. A mixture of the listed MC trial moves is used with some user-specified ratios (*e.g.* 50% pull, 20% rebridging, 25% pivot, and 5% diagonal moves), with each window having its own set of ratios. Aggregation simulations typically have no more than 5% cluster translation and rotation moves. The Mersenne Twister [97] pseudo-random number generator from the C++ standard library is used for all simulations. Error bars are calculated using an omit-one jackknife procedure [98] for multiple (5-20) independently seeded simulations.

Chapter 5

HP Model Folding on the FCC Lattice

5.1 Predicted ground states

REWL simulations are employed for the set of benchmark HP proteins listed in Table 3.1 on the fcc lattice, where ground state energies and structures are successfully predicted without any prior knowledge of the energy ranges. A flatness measure of $p = 0.8$ and a termination criterion of $\ln f_{final} = 1 \times 10^{-6}$ are chosen, and simulations are performed with 8 windows for HP sequences with $\ell < 100$ and 16 windows when $\ell > 100$. Figure 5.1 shows representative configurations taken from the ground state energies for each of the HP sequences, with the value E_{min} shown below the structure. Energy values are verified as minimal using the HPStruct tool from the CPSP web server [99–102], which uses a constraint programming method to thread sequences through an optimal H-core to give a lower bound on the minimal possible energy. All predicted ground state energies are exactly matched with results from HPStruct, with the exception of $E_{min} = -141$ for HP88, for which HPStruct assures $E_{min} > -143$. Our identified E_{min} are also considerably lower than those predicted by a fcc lattice study [103] using the modified prune-enriched Rosenbluth method (nPERM) [41] chain-growth algorithm, with $E_{min} = -121$ vs -116 for HP103, $E_{min} = -164$ vs -154 for HP124, and $E_{min} = -174$ vs -168 for HP136. The ground states of HP lattice models

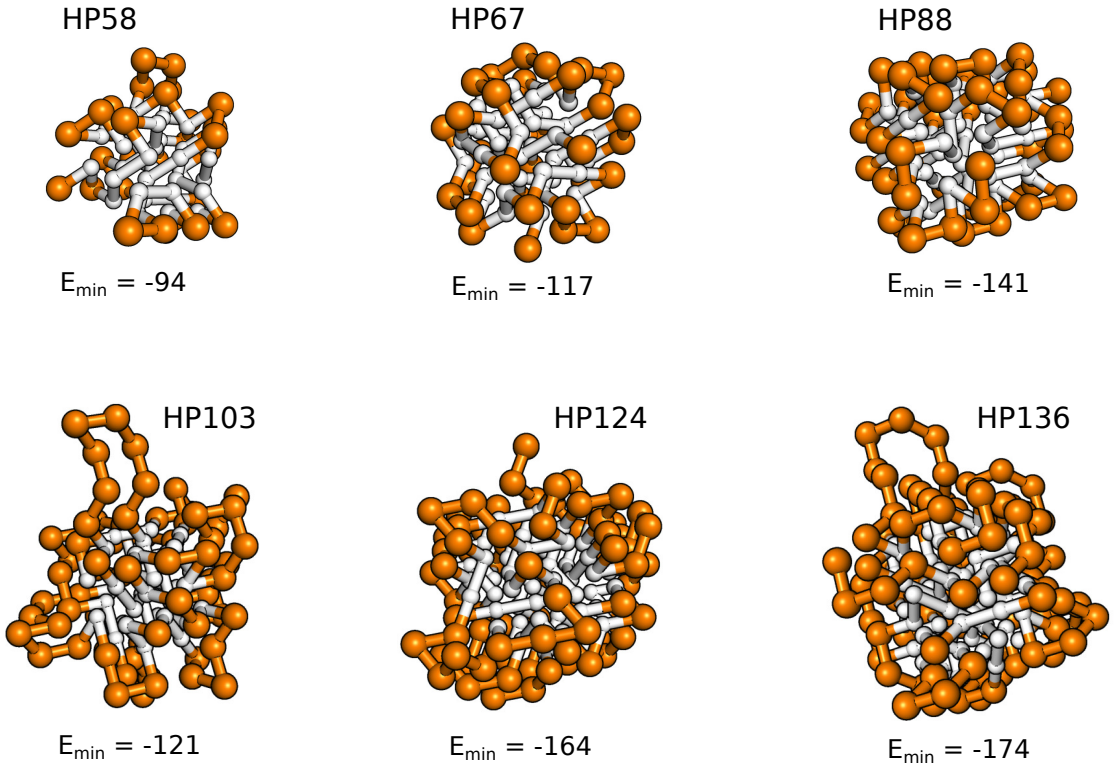

Figure 5.1: Ground state structures and energies found for the benchmark HP sequences from Table 3.1 on the fcc lattice.

are known to exhibit a high level of structural degeneracy, which is found to be the case for the benchmark sequences simulated here on the fcc lattice. An exact number was not determined, but on the order of 10^6 to 10^7 unique structures were found (using the tree data structure described in Appendix A5) within our REWL and subsequent MUCA simulations for the selected sequences.

5.2 Thermodynamic comparison with sc results

Results from previous studies by Wüst and Landau [61, 104] that successfully apply the WL algorithm with pull, pivot, and bond-rebridging trial moves for the sc lattice are used to compare with our REWL simulations with the fcc lattice models. The same HP sequences in Table 3.1 are also used in the sc lattice study, allowing a direct comparison with our fcc lattice results that differ only by lattice geometry.

Along with the full range of energy states, the density of states is obtained by the REWL simulations. Figure 5.2 shows $\ln[\hat{g}(E)]$ for HP58 on the fcc (black circles) and sc (red triangles) lattices. Results in this image are compared by shifting the values as $\ln[\hat{g}(E)] - \max\{\ln[\hat{g}(E)\}$ and normalizing the energy range by the coordination number of the respective lattices. The $\hat{g}(E)$ for HP58 on the fcc lattice spans about 35 orders of magnitude, but the

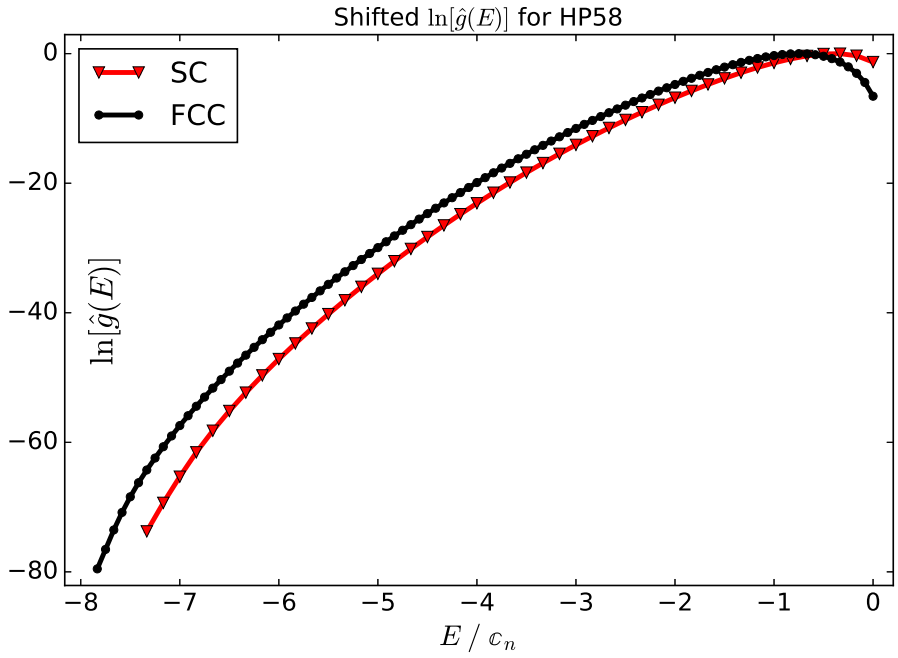

Figure 5.2: Comparison between the estimate $\ln[\hat{g}(E)]$ of the sequence HP58 for the fcc (black circles) and sc (red triangles) lattices. The values have been shifted by subtracting out $\max\{\ln[\hat{g}(E)]\}$ and normalizing the energy by $\mathfrak{c}_n$.

longer HP sequences (*e.g.* HP124) have a $\hat{g}(E)$ on the fcc lattice that spans more than twice this magnitude. It is helpful to remember that while the heat capacity C_V is calculated from the fluctuation of the average energy, it is equivalent to the thermal derivative of the internal energy (U), which for the HP model Hamiltonian, is equivalent to the thermal derivative of the average number of HH contacts weighted by $-\varepsilon_{HH}$. This relation is expressed as:

$$C_V = \frac{dU}{dT} = -\varepsilon_{HH} \frac{d\langle n_{HH} \rangle}{dT} \ . \tag{5.1}$$

Therefore, the signals (peaks and 'shoulders') in C_V/ℓ give information about the rate of H-core formation with respect to temperature.

To address whether the additional geometric freedom changes the folding 'transitions' exhibited by the HP model on the fcc lattice compared to the sc lattice, we compare the folding thermodynamics by plotting C_V/ℓ for both lattice types together. Figure 5.3 shows such data for the only the HP sequences that are mapped from real proteins and fragments of proteins, and Figure 5.4 shows the data for HP sequences that were designed specifically for ground states with low degeneracy on the sc lattice. The sc lattice data from the folding study of Wüst and Landau [61] are plotted as the red, dashed curves, and the fcc lattice data that are the results of our REWL simulations are plotted with the black, solid curves. Reduced temperature axes are normalized by $\mathbb{c}_n$ for the respective lattices to match up results for the comparison. For the remainder of the chapter, the shorthand $T^* = k_B T/(\mathbb{c}_n \varepsilon_{HH})$ will be used to refer to the reduced temperature.

In each plot, there is a distinct peak in C_V/ℓ between $0.08 < T^* < 0.125$ that is indicative of a hydrophobic collapse structural transition. The hydrophobic collapse involves a change of the model protein from a random, extended coil state to a disordered state that has one or more globular regions. A large increase in HH contacts signifies the initial formation of a disordered H-core or clusters of HH contacts. Both lattice geometries show comparable collapse temperatures (when shifted by $\mathbb{c}_n$), with the fcc lattice having a slightly larger collapse temperature and magnitude of C_V/ℓ. The observed difference in the fcc data at these temperatures is presumably due to a greater number of available globular states that can form on the lattice geometry, including effects from an absence of the parity problem.

At temperatures below the hydrophobic collapse, at least one additional structural rearrangement occurs, as suggested by the signals in C_V/ℓ near $T^* \leq 0.05$. Both the fcc and sc data for HP136 show a pronounced peak at $T^* \approx 0.05$, which shows some significant folding event involving an increase in HH contacts at a rate comparable to the initial collapse of the model protein. In contrast, the fcc data for HP124 show a small, additional shoulder at

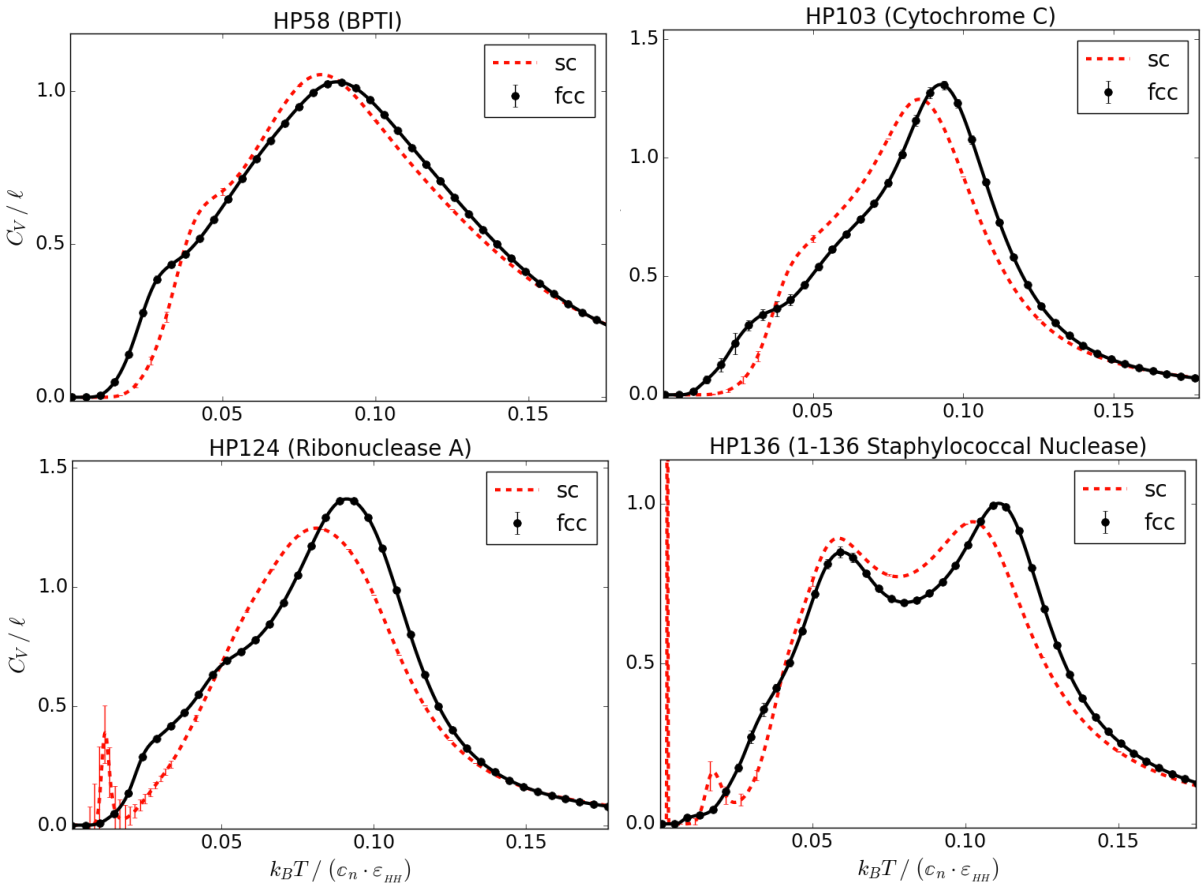

Figure 5.3: Comparison between the fcc (black circles) and sc (red dashes) lattice folding thermodynamics for biologically inspired HP sequences. Error bars are shown and are smaller than the size of data points where not visible.

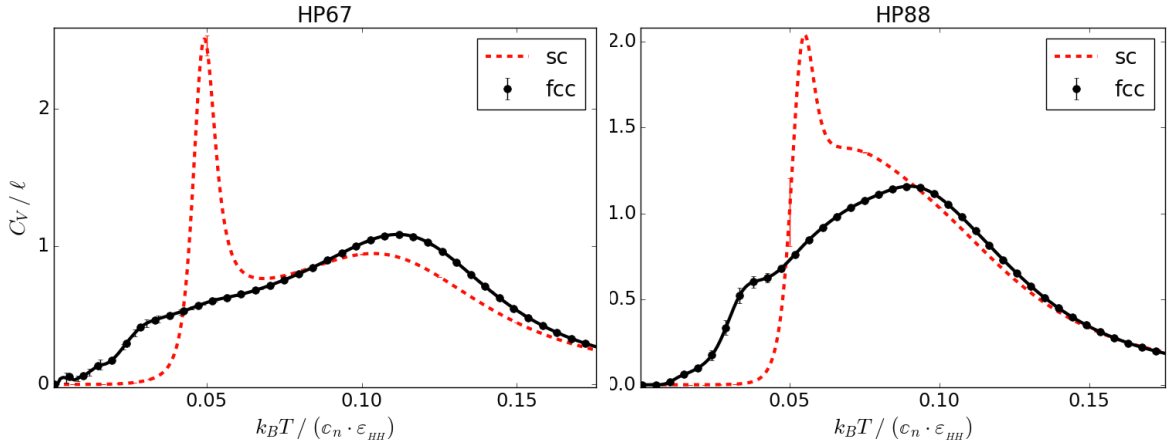

Figure 5.4: Comparison between the fcc (black circles) and sc (red dashes) lattice folding thermodynamics for HP sequences that are designed to have threefold (HP67) and non degenerate (HP88) ground state structures on the sc lattice. Error bars are shown and are smaller than the size of data points where not visible.

the same temperature, but no signal is present here for the sc lattice data. The behavior at these suspected folding events must be clarified using additional structural observables, as shown in Section 5.3. For the sc lattice data with HP124 and HP136 in Figure 5.3, the spikes and peaks in C_V/ℓ below $T^* < 0.05$ have a large degree of statistical error, and are likely a spurious result of under-sampled $\hat{g}(E)$ for the lowest energies.

For HP67 and HP88 in Figure 5.4, there is a stark contrast between the folding behavior between the two lattice types. In the sc lattice data for both sequences, there is a sharp maximum in C_V/ℓ at $T^* \approx 0.05$, where there are practically no additional HH contacts forming at temperature below this point. These HP sequences were specifically designed to have low-degeneracy ground state structures (threefold for HP67 and unique for HP88) on the sc lattice. It is unsurprising that this property is not preserved when the sequences are simulated on the fcc lattice (we observe a high structural degeneracy for these sequences as well using the tree data structure from Appendix A5), and instead of a first-order-like 'transition', there is a relatively steady increase in the number of HH contacts after the initial collapse.

One feature that is consistently observed in the fcc data for each sequence is the shoulder that persists to $T^* \approx 0.025$. This indicates a final rearrangement of the H-core as the ground state energy is acquired, and corresponds to the peaks near $T^* \approx 0.05$ for HP58, HP67, HP88, and HP103 in the sc lattice data.

5.3 Average structural quantities

To further resolve the physical behavior of the fcc HP model at the structural transitions indicated by the signals in C_V/ℓ, average structural properties are calculated from a large MUCA production run. Described in Section 3.4.1, we use the thermal derivative of the relative shape anisotropy for all residues ($d\langle\kappa^2\rangle/dT$) and for just the H residues ($d\langle\kappa^2_{HH}\rangle/dT$). κ^2 has values in the range $[0, 1]$, and gives a description of the model protein's overall shape.

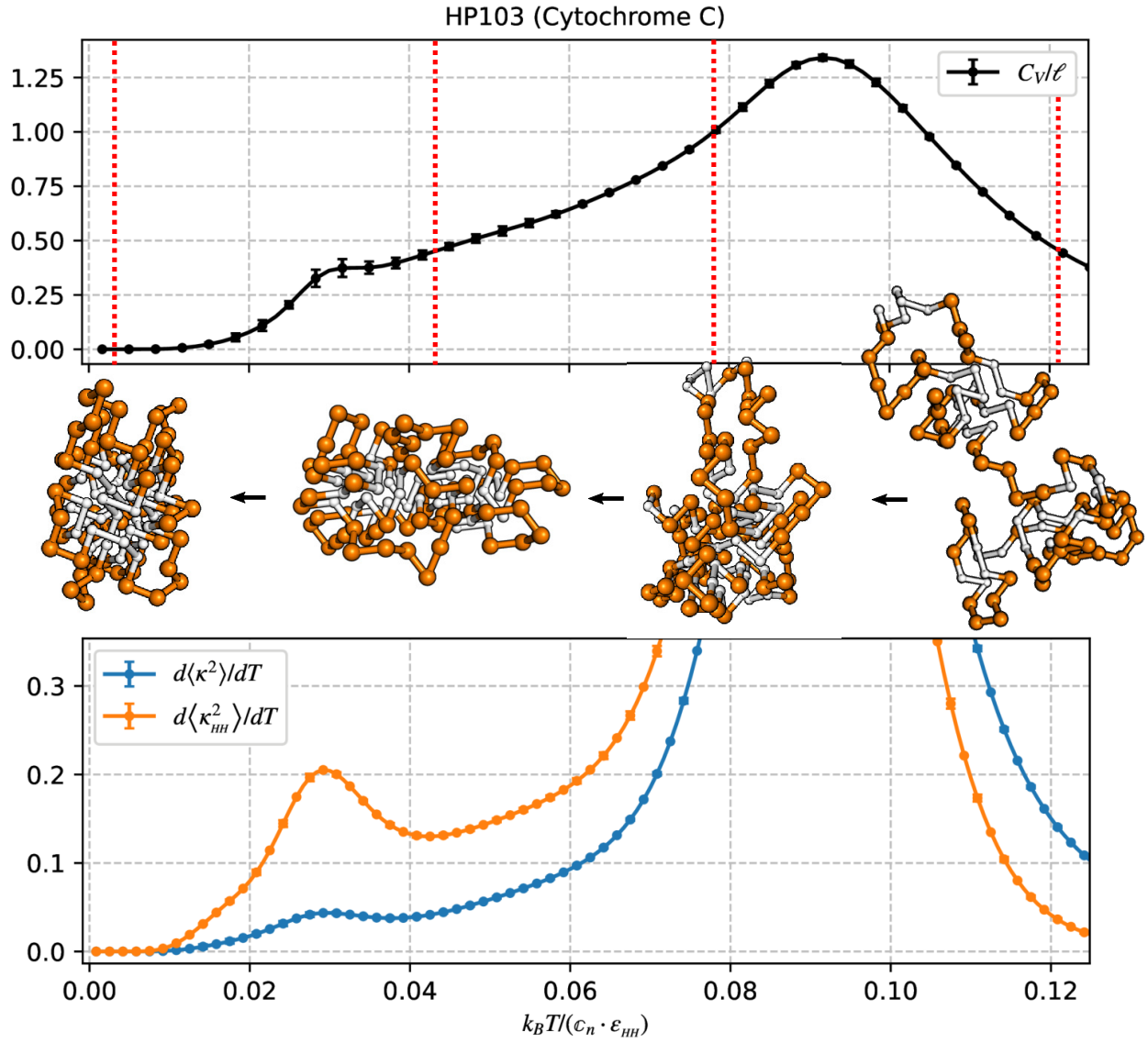

Figure 5.5: Specific heat (top) and thermal derivative of the relative shape anisotropy for all residues (blue curve, bottom) and just the H residues (orange curve, bottom) for HP103. Error bars shown in the bottom plot are smaller than the marker sizes.

For $\kappa^2 = 1$, the system is in a completely linear configuration, and $\kappa^2 = 0$ corresponds to a perfectly spherical conformation. The quantity κ^2_{HH} is calculated to separate the changes of shape associated with just the H-core from that of the overall configuration. While other shape descriptors can be calculated from the gyration tensor, we find the relative shape anisotropies to be most informative for our particular simulations. Snapshots of the model structures saved at each energy during the REWL simulations are also valuable for a visual verification of what is represented in the data. Figures 5.5, 5.6, and 5.8 show these

quantities for HP103, HP124, and HP136, respectively. The top panels show C_V/ℓ with vertical, dashed lines indicating temperatures before and after each suspected folding event. Representative structures taken from these temperatures are shown below the corresponding dashed lines. In the bottom panels of these figures, the thermal derivatives of the relative shape anisotropies are plotted together.

The behavior of HP103 agrees with what is inferred from the thermodynamic comparison using C_V/ℓ, with the additional affirmation that there is no intermediate folding events occurring between the collapse and ground state acquisition. Both $\langle\kappa^2\rangle$ and $\langle\kappa^2_{HH}\rangle$ show a massive decrease during the hydrophobic collapse ($T^* > 0.08$), and another significant decrease near $T^* \approx 0.025$, where the final ordering of the H-core occurs. This is also the case at these temperatures for the other sequences with $\ell < 100$ on the fcc lattice.

What is more interesting is the behavior for HP124 and HP136 on the fcc lattice, which may have additional folding steps signaled by C_V/ℓ near $T^* \approx 0.05$ and 0.025. For HP124, the configuration snapshots at the temperatures indicated in the top panel of Figure 5.6, particularly the second from the right, give an indication of what is happening in the two subtle shoulders at $T^* \approx 0.05$ and 0.03. After the initial collapse of the model protein, the state is a disordered configuration with two clusters of HH contacts near $T^* \approx 0.065$, which are then consolidated to a single H-core at $T^* \approx 0.05$. Both anisotropy values show a noticeable decrease at this temperature, but $\langle\kappa^2_{HH}\rangle$ continues to decrease until $T^* \approx 0.025$, where the last shoulder in C_V/ℓ occurs. It becomes less clear what is happening near the lowest-temperature signal in C_V/ℓ, where the statistical error in both anisotropy values limit any interpretation near $T^* \approx 0.02$. However, the structures show a small rearrangement of the dense H-core, and $\langle\kappa^2_{HH}\rangle$ continues to decrease at this temperature.

To get a better understanding of the system's behavior in the region between $0.02 < T^* < 0.06$, where the decrease in $\langle\kappa^2_{HH}\rangle$ occurs, we look at the thermal derivatives of end-to-end distance and hydrophobic radius of gyration. These quantities are shown in Figure 5.7, where the top panel shows C_V/ℓ plotted with $d\langle R_{g,HH}\rangle/dT$ and the bottom panel shows

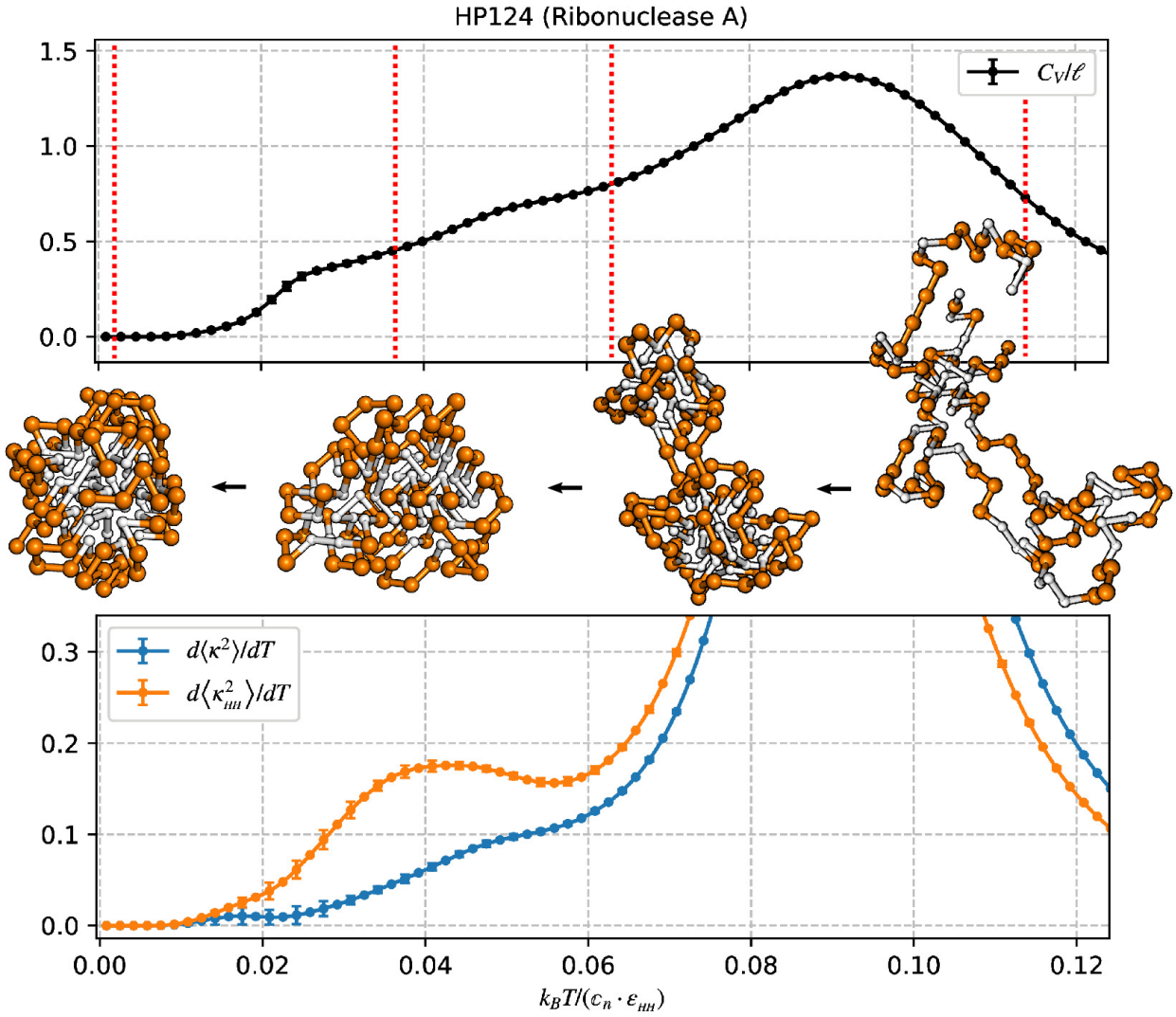

Figure 5.6: Specific heat (top) and thermal derivative of the relative shape anisotropy for all residues (blue curve, bottom) and just the H residues (orange curve, bottom) for HP124. Error bars shown in the bottom plot are smaller than the marker sizes.

$d\langle R_{EE}\rangle/dT$. There are two subtle shoulders in $d\langle R_{g,HH}\rangle/dT$ which have corresponding temperatures to those found in C_V/ℓ, but extremely small in magnitude. At the signal in C_V/ℓ near $T^* \approx 0.05$, the value of $\langle R_{EE}\rangle$ is at its global minimum (not explicitly shown here, but corresponding to where the derivative = 0). Below this temperature, $\langle R_{EE}\rangle$ increases, with a maximum thermal rate of increase corresponding to the second shoulder in C_V/ℓ where $T^* \approx 0.025$. These additional data for HP124 suggest that there is a small structural rearrangement upon the surface of a folded state during the acquisition of the ground state H-core near $T^* \approx 0.025$.

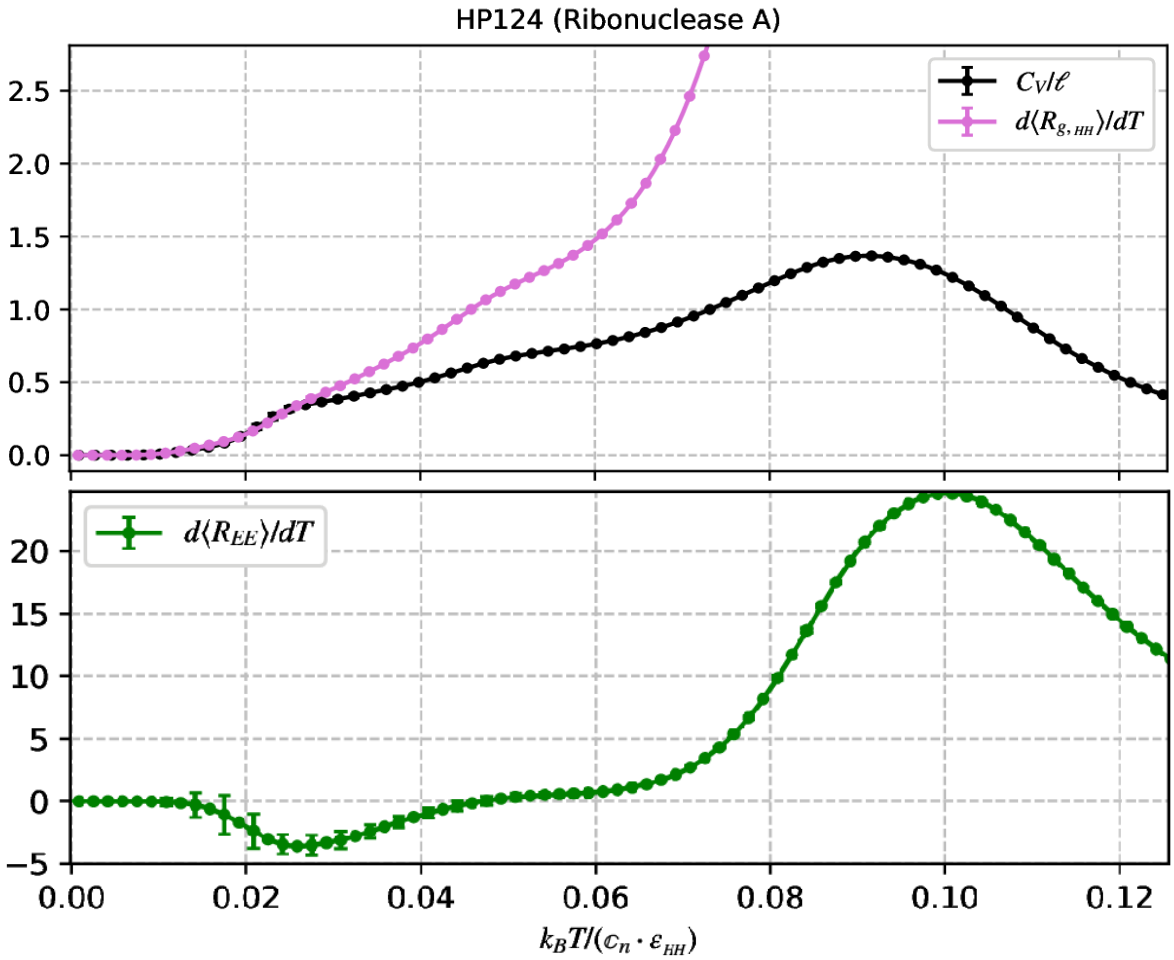

Figure 5.7: Additional structural quantities for HP124. Specific heat (black) and thermal derivative of the hydrophobic radius of gyration (pink) are shown in the top plot, and the thermal derivative of end-to-end distance (green) is shown in the bottom plot.

The fcc data for HP136 in Figure 5.8 is used in the same manner as before to look at some average structural properties for the signals in C_V/ℓ at $T^* \approx 0.06$ and 0.025. At $T^* \approx 0.08$, the representative configuration shows a state with three separate clusters of HH contacts rather than a single globular structure. The structural change at $T^* \approx 0.06$ involves a second step in hydrophobic collapse where the three separate clusters of HH contacts are merged to form a single disordered H-core, as shown in the second configuration from the left. This is supported by the substantial decrease in both the value of $\langle \kappa^2 \rangle$, as well as $\langle \kappa^2_{HH} \rangle$, at this temperature. Finally, the low temperature behavior ($T^* \leq 0.04$) is investigated using thermal derivatives of the anisotropy values and end-to-end distance.

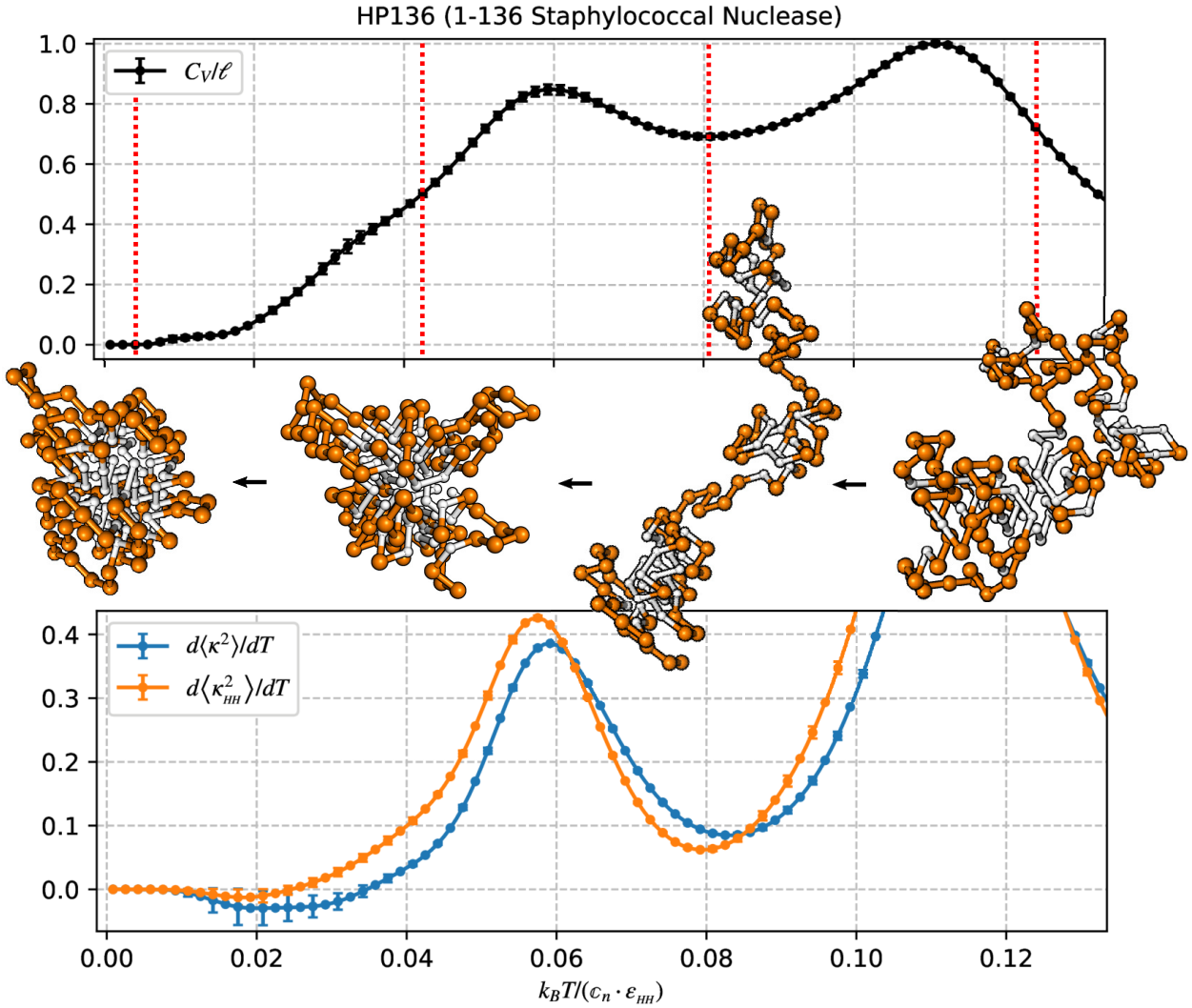

Figure 5.8: Specific heat (top) and thermal derivative of the relative shape anisotropy for all residues (blue curve, bottom) and just the H residues (orange curve, bottom) for HP136. Error bars shown in the bottom plot are smaller than the marker sizes.

The small undulations in C_V/ℓ are just within the range encompassed by statistical error, and any signals present in $d\langle\kappa^2\rangle/dT$, $d\langle\kappa^2_{HH}\rangle/dT$ and $d\langle R_{EE}\rangle/dT$ near $T^* \approx 0.02$ are certainly well within 2σ of statistical error. Structural data for HP136 do not give evidence of any additional structural rearrangements on the fcc lattice (at least given our calculated statistical precision), but rather, it is concluded that the hydrophobic collapse occurs in two steps before a nearly optimal H-core is obtained.

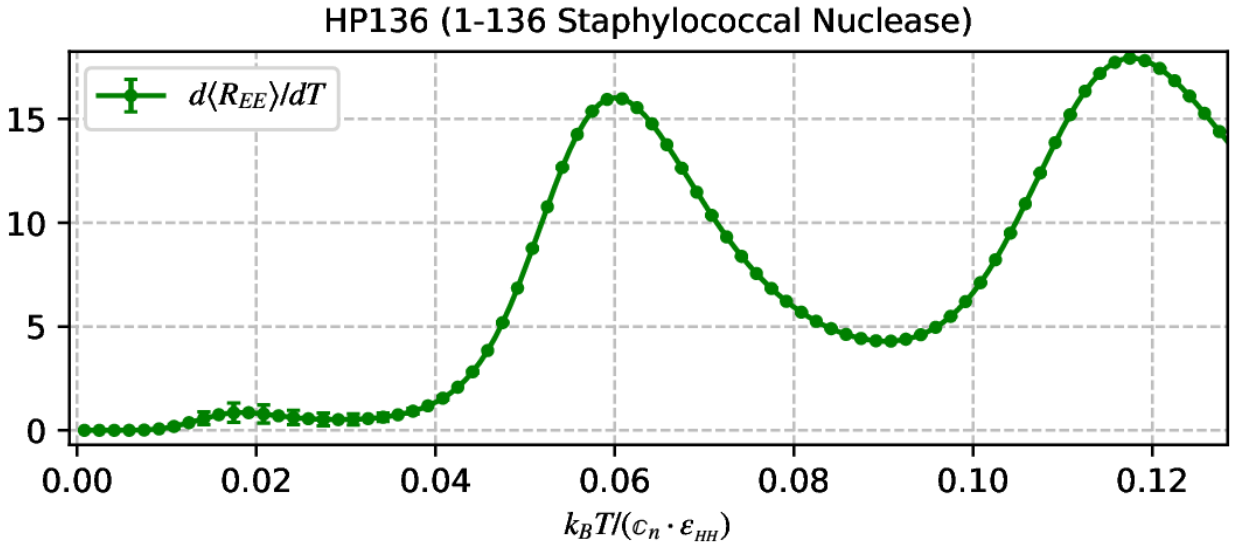

Figure 5.9: Thermal derivative for the end-to-end distance of HP136. Error bars are shown and are smaller than the size of data points where not visible.

Chapter 6

Simulations with Helical Motifs on the fcc Lattice

One of the major advantages of the fcc lattice over the sc lattice is the increased geometrical freedom for backbone structures, which includes the possibility of $\theta = 120°$ bond angles. While the previous section shows an analysis for the HP protein model that is motivated purely by hydrophobic effects, the ground state structures for this model with the unrestrained fcc lattice are often very compact, and include many acute angles and extended segments ($\theta = 60°,\ 180°$) which would not realistically be observed in protein structures. It has been posited [10, 11], through an analysis of many experimentally measured protein structures, that the cubo-octahedral subset (only the $\theta = 90°\ 120°$ bond angles) of the fcc lattice gives a minimal geometry that can reasonably approximate protein structures [105]. In this chapter, we explore a scheme for including the α-helix secondary structural motif in backbone-only models for fcc lattice.

6.1 Helical bundles in fcc ISAWs

6.1.1 fcc representations of α-helices

The α-helix is a predominant secondary structure that is found in polymer and protein structures. There have been some studies that describe the inclusion of helical motives for fcc lattice heteropolymers, but with modifications that deviate significantly from the HP model. We describe a few options for right-handed α-helices on the fcc lattice, which are shown as the white structures in Figure 6.1. The green structures in Figure 6.1 are an α-helix segment that is taken from the protein database (PDB) entry for Crambin.

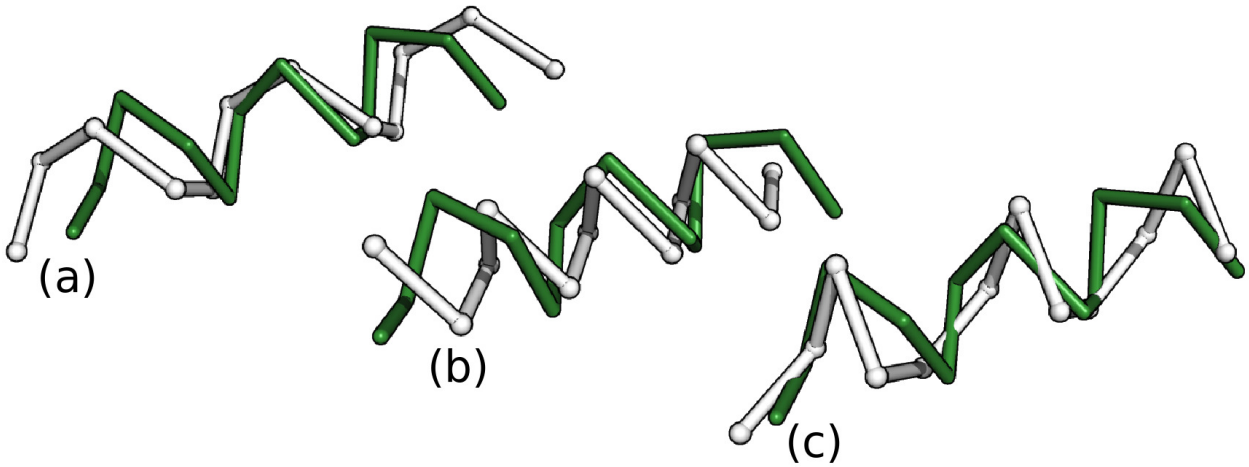

Figure 6.1: Three approximations for the α-helix on the fcc lattice. White structures show the lattice approximations, and green structures show a real helix backbone taken from the PDB. The three types are: repeating $90°, 120°$ bond angles (a), repeating $90°, 60°$ bond angles (b), and Pokarowski's helix (c).

Shown in image (c) of Figure 6.1, this helix representation is described by Pokarowski [106] for a heteropolymer on the fcc lattice. Their model adds an additional classification for residues that labels them as being in either flexible or helical regions, which must be provided along with the HP sequence. This version of the helical energy works by identifying four consecutive bond vectors of a specific order and weighting these with some energetic constant. Advantages of this helical representation are that the root mean square deviation (RMSD) is the lowest (1.66Å for the 12 residues shown) for each of the three fcc helices shown, and

that it does not form double helices like the one shown in Figure 6.2. Disadvantages which dissuade our use of this scheme are the requirement of additional residue information that specifies what the folded structure will be with regards to helical packing, and the lack of a regular dihedral angle that characterizes this helix, and the occurrence of 60° bond angles in each helical turn. Furthermore, the identification of each helical turn is a 5-body process.

Images (a) and (b) in Figure 6.1 are results of a torsional energy where the energetic weights in Equation 3.9 are chosen such that $E_{\phi} = A\cos(\phi_i - \phi_0)$, with a reference dihedral angle $\phi_0 = 50°$. Both of these helices occur when this torsional energy is assigned without any additional bond-angle preferences on the fcc lattice, where the helix in (b) consists of repeating $90°, 60°$ bond angles and the helix in (a) consists of repeating $90°, 120°$ bond angles. The helix in image (b) is densely packed, and accommodates a high number of energetic contacts within the structure. It does not form double helices, and has the highest RMSD (2.42Å for the 12 residues shown) of the three helix types. In our folding simulations, we adopt the helix type in image (a), which has an RMSD of 1.85Å for the 12 residues shown. It has the disadvantage of forming double helices, as shown in Figure 6.2, and does not form any nearest-neighbor contacts within the helix structure. Advantages are that the bond angles are most realistic for what is found in protein structures, and that it can be characterized by a single dihedral angle of 55° on the fcc lattice. A study by Toma and Toma [105] uses the cubo-octahedral lattice to simulate a version of the HP model that includes explicit side chains that are represented as residues separately from the backbone structure. In these simulations, they show that the backbone structure is equivalent to the helix shown in image (a). Furthermore, this is the same α-helix backbone structure that is predicted from constraint programming studies [107] on the fcc lattice. We use a simplified scheme for torsional energies that adds an energetic penalty whenever $\phi_i \neq 55°$ to recover the form: $E_{\phi} = -n_{\phi}\varepsilon_{\phi}$, where the shorthand $n_{\phi} = n_{(\phi \neq 55°)}$, $\varepsilon_{\phi} = \varepsilon_{(\phi \neq 55°)}$ is used (*e.g.* no penalty if all residues give $\phi = 55°$).

For backbone-only models like the ISAW and HP/H0P models, incorporating the helix motif from image (a) in Figure 6.1 along with nearest-neighbor contact energies gives rise to a predominance of double helix structures, as shown in Figure 6.2. These types of structures are valid under the models (and often have minimal energies), but are unrealistic or irrelevant for protein and polymer systems where a steric clash would occur between side chains.

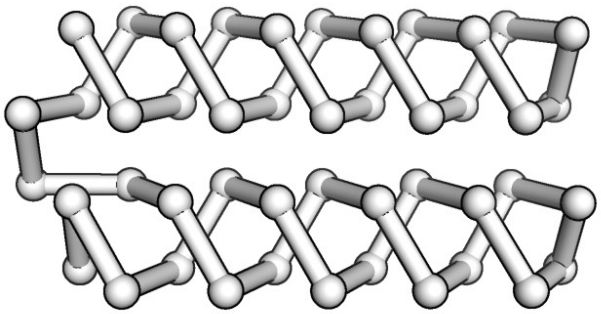

Figure 6.2: Double helix structure for a 46-mer ISAW on the fcc lattice. Two α-helices wrap around one another to form a dense configuration contains a large number of nearest-neighbor contacts.

To recover folded states on the fcc lattice which favor both contact energies and helical structures without the formation of double helices, we incorporate a simple distance-dependence into the contact energies. Extending the distance for which energetic contacts can form is motivated by our desire to include helix-containing structures that are most relevant to what would be found in real protein and peptide systems, and by the distribution of hydrophobic distances that are measured in such systems. In an analysis of many PDB entries [108], it is observed that the average distance between H residues in real proteins are typically greater than what would be the result of lattice simulations which only promote nearest-neighbor contacts corresponding to 3.8Å. Instead, hydrophobic interaction distances between 4.1 - 9.5Å could be more representative of those found in folded proteins [108]. Additionally, continuum models of proteins often utilize continuous, distance-dependent energies between non-bonded residues, such as the Lennard-Jones potential, which have a repulsive

term for distances below some specified bond length, and attractive term for distances above this. A step function approximation for this type of potential with contact energies has been previously explored [109]. Figure 6.3 shows the sites for the first 4 distances away from any

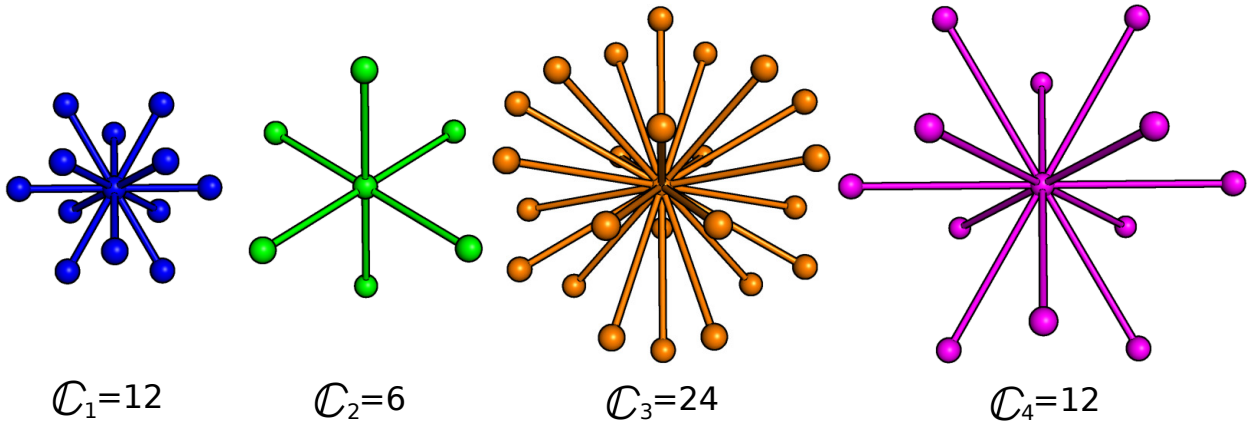

Figure 6.3: Diagram of the neighbors for the four closest distances to a site on the fcc lattice. The number of neighbors for each distance is given by the subscripted $\mathbb{c}$ (*e.g.* the third-nearest distance has 24 neighbors).

given point on the fcc lattice, where the subscripted $\mathbb{c}$ give the number of neighbors at those distances. For example, at $\mathbb{c}_3$, there are 24 sites that would each be equidistant from the central lattice position. Scaled to physical lengths, the distances for $\mathbb{c}_1, \mathbb{c}_2, \mathbb{c}_3, \mathbb{c}_4$ correspond to 3.8Å, 5.4Å, 6.6Å, and 7.6Å, respectively. We choose an interaction scheme that assigns energies to neighbors at the third such distance, so that each residue can have up to $\mathbb{c}_3 = 24$ energetic contacts. This set of neighbors has the highest occupation number out of the distances considered, and is in the middle of the short- to medium-range distances that occur in folded states from the PDB analysis. Shown in Equation 6.1, we use an ISAW model to demonstrate folding simulations that incorporate contact energies between nearest neighbors (ε_1), third-nearest neighbors (ε_3), and also a penalty term for non-helical dihedrals (ε_ϕ).

$$\mathcal{H} = -n_1\varepsilon_1 - n_3\varepsilon_3 - n_\phi\varepsilon_\phi \tag{6.1}$$

To recover ground states which have bundles of helices rather than double helices, we set $\varepsilon_1 < 0$ (repulsive), $\varepsilon_3 > 0$ (attractive), and $\varepsilon_\phi < 0$ (repulsive if 4 consecutive residues are

not a helical turn). It is worth noting that while calculation of the contact energies under this scheme takes 3× the computational effort than the original HP model, this is not a bottleneck in the simulation compared to the sampling of configurations for each energy.

6.1.2 Predicted ground state structures

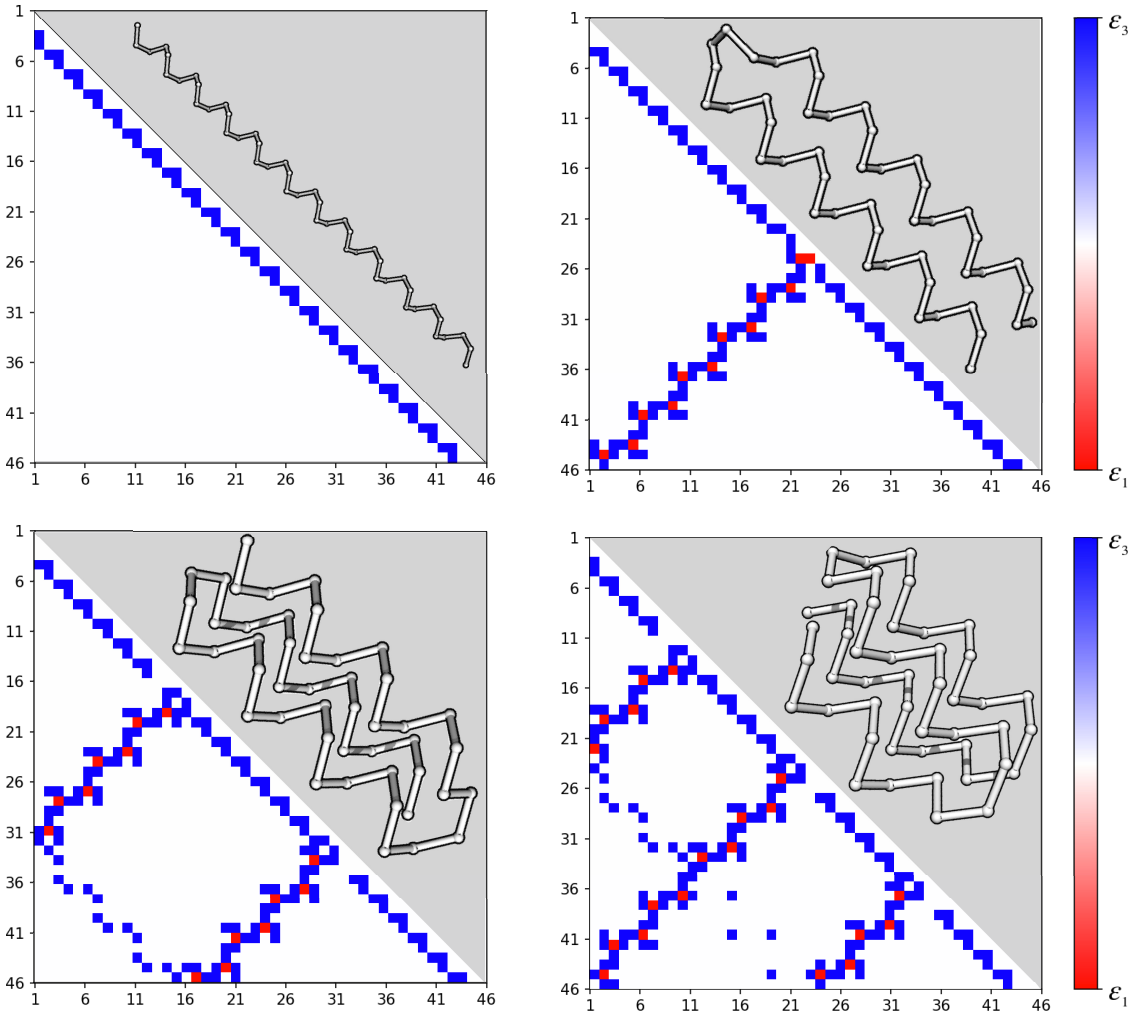

Figure 6.4: Contact maps and structures for the ground states containing 1-, 2-, 3-, and 4-helix bundles for the fcc lattice ISAW with 46 residues. Contact maps are shown on the bottom-left of each plot, where the axes show residue positions along the chain, and colored data in the map correspond to contact distances (red for nearest-neighbor and blue for third-nearest neighbor). Snapshots of the helical bundle structures are shown in the shaded region in the top-right of each plot.

REWL simulations were run for the modified ISAW model given in Equation 6.1, where a sweep over ε_ϕ was performed to search for the helix-containing structures. In each simulation, we set $\varepsilon_1 = -1$ and $\varepsilon_3 = 1$, although a smaller ratio of $\varepsilon_1/\varepsilon_3$ could be assigned. For this model, the simulations typically converged within ≤ 500 core-hours when using 4-8 REWL windows with a flatness measure of $p = 0.6$ and a termination criterion of $\ln f_{final} \leq 1 \times 10^{-6}$.

Highly ordered helical bundles containing the fcc helix type from image (a) in Figure 6.1 were found in all simulations where $\varepsilon_\phi \leq -2$. Figure 6.4 shows the contact maps and configurations for the ground states with helical bundles found in the REWL simulations of a ISAW with 46 residues on the fcc lattice. Contact maps are shown below the diagonal of each plot, where the x- and y-axes give the residue positions along the chain, and colored data in each map shows the contact distances between residues. Blue data points represent favorable energetic contacts between third-nearest neighbors, and red data points represent energetic penalties from nearest-neighbor contacts. The data points running parallel to the diagonals correspond to intra-helix contacts, while the bands of blue and red data points that run perpendicular to the diagonal correspond to inter-helix contacts that form as the helices are packed together. Ground state configurations are shown for each figure in the shaded regions above the diagonals, where alternate, top-down views can be found in Figure 6.7 at the end of the section.

6.1.3 Folding thermodynamics structural diagram

Using a long MUCA production run, we calculate the structural properties for the sweep over ε_ϕ. To get a better understanding of the transitions occurring in the folding simulations, thermal derivatives of the average contact numbers are examined, as shown in Figure 6.5.

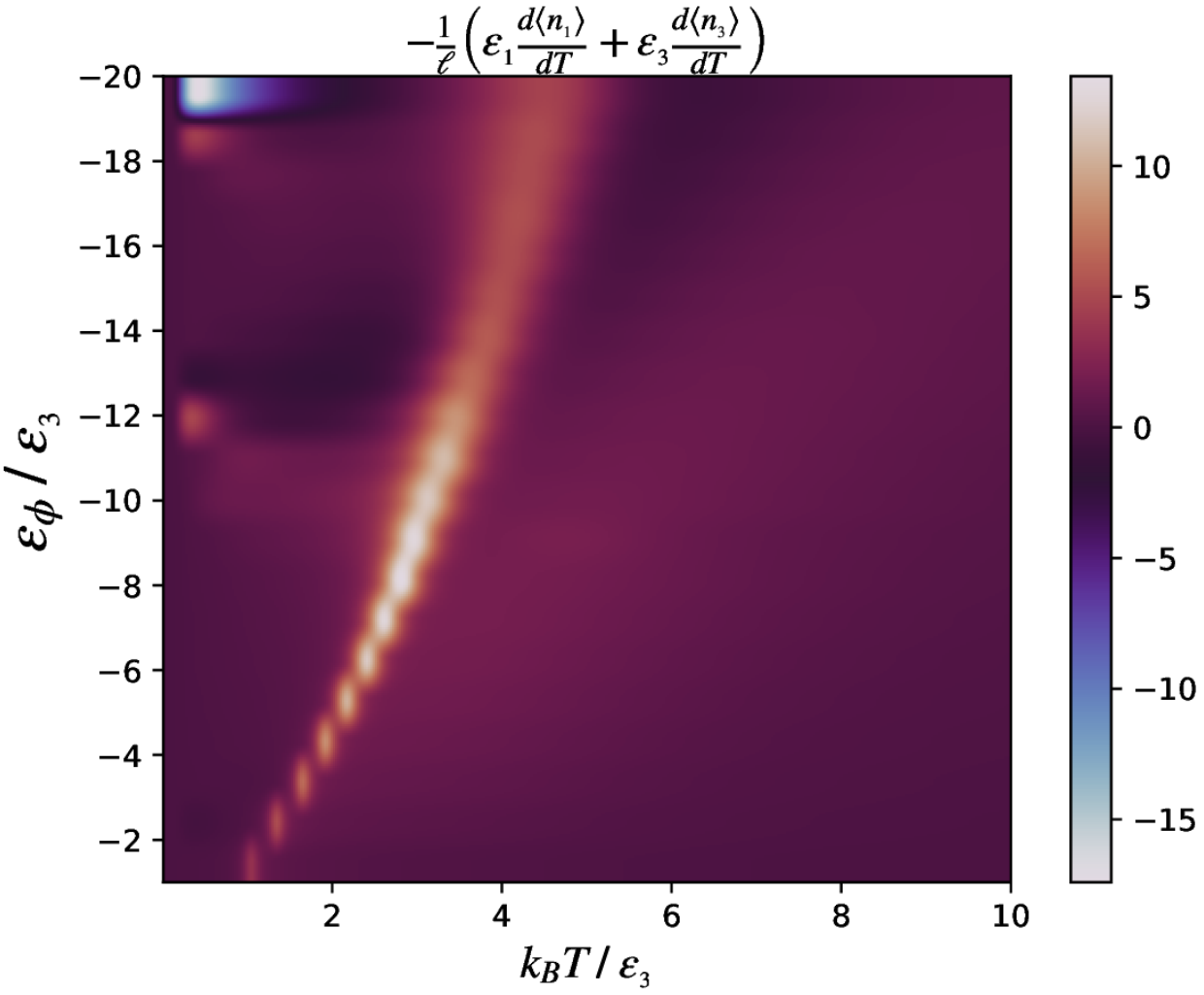

Figure 6.5: Heatmap of the thermal derivative of the energy-weighted contact numbers at a range of temperatures (x-axis) and torsion penalty values (y-axis) for the fcc lattice ISAW with 46 residues.

The quantity $-\frac{1}{\ell}\left(\varepsilon_1 \frac{d\langle n_1\rangle}{dT} + \varepsilon_3 \frac{d\langle n_3\rangle}{dT}\right)$ corresponds to the value of C_V/ℓ without the contribution from the fluctuation of torsional penalties. From the heatmap, it appears that there is a large structural transition in the region $2 \leq k_BT/\varepsilon_3 \leq 5$ for almost all parameter values, and that there are some other signals below this temperature for $\varepsilon_\phi/\varepsilon_3 \approx -12$ and -20.

To get a more clear picture of what is going on in the regions bounded by these transitions, the average relative shape anisotropy is calculated, as shown in Figure 6.6. These data show the distinction between structural 'phases' for the space of temperatures and torsion penalties much more clearly. $\langle\kappa^2\rangle$ assumes a value of 0 for a perfectly spherical structure, a value of 1.0 for a perfectly linear structure, and a value of 0.25 for planar symmetric structures. From the $\langle\kappa^2\rangle$ values and visual inspection of configuration snapshots taken during the simulations, regions of the diagram are labeled with abbreviations for their representative structures. For torsional penalties in $-11 \leq \varepsilon_\phi/\varepsilon_3 \leq -4$, there is a transition from a globular state with disordered helix content (GH) to a 4-helix bundle (4H) between approximately $2.0 \leq k_BT/\varepsilon_3 \leq 4.0$. At higher torsional penalty strengths where $\varepsilon_\phi/\varepsilon_3 < -12$, a transition from a random coil state (RC) to a 3-helix bundle (3H) occurs between temperatures $4.0 \leq k_BT/\varepsilon_3 \leq 6.0$ before another transition to a 2-helix bundle (2H) at a slightly lower temperature. Special values appear at $\varepsilon_\phi/\varepsilon_3 = -12$, where 3-helix structures are predominant below $k_BT/\varepsilon_3 \leq 4.0$, and $\varepsilon_\phi/\varepsilon_3 = -20$, where a single helix (1H) forms at low temperatures. The random coil structure is present in the high-temperature regions of all simulations.

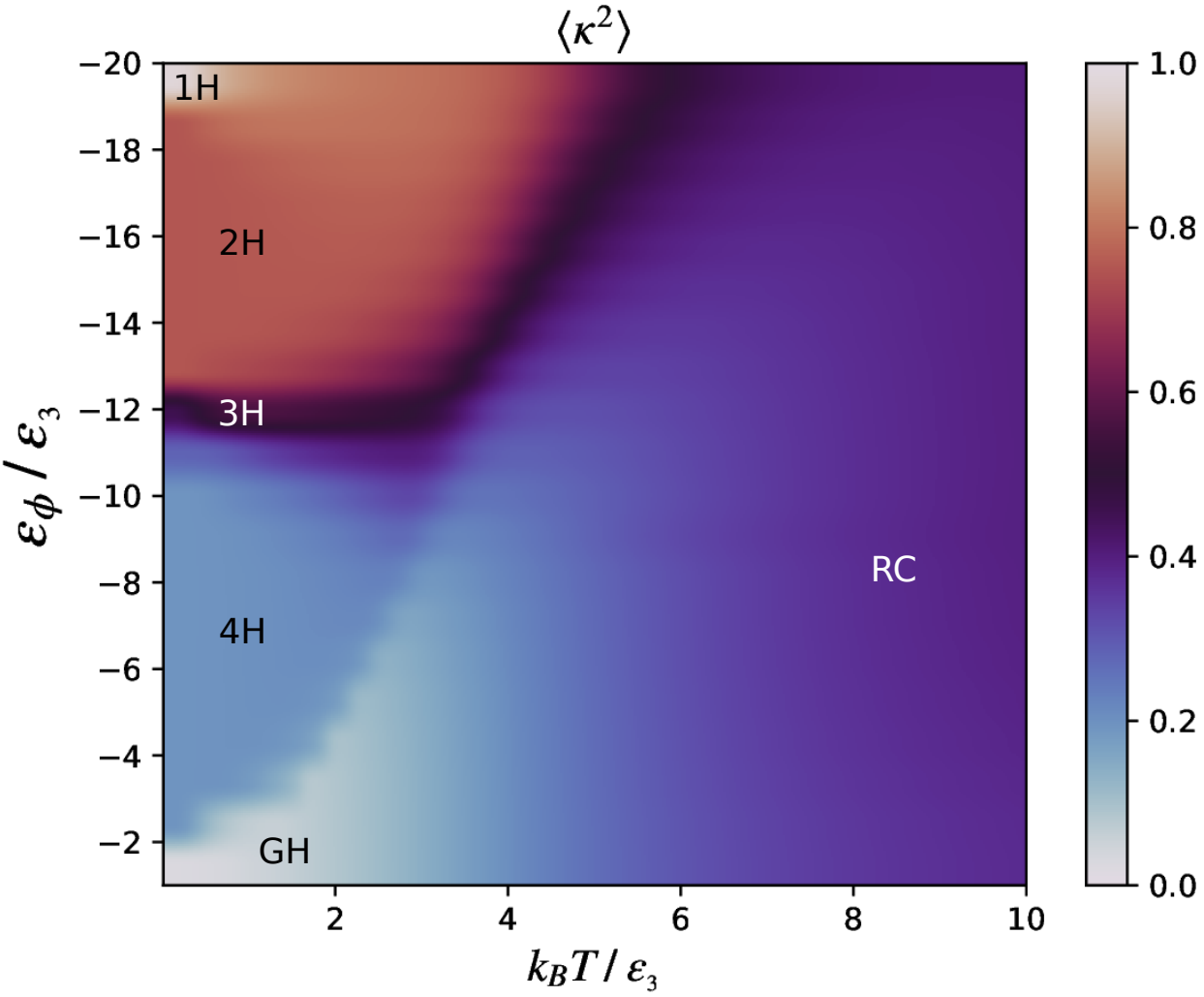

Figure 6.6: Heatmap of the relative shape anisotropy at a range of temperatures (x-axis) and torsion penalty values (y-axis) for the fcc lattice ISAW with 46 residues. Structural regions are labeled with abbreviations: one helix (1H), two helices (2H), three helices (3H), four helices (4H), globular helix (GH), and random coil (RC).

Figure 6.7 shows representative structures for each of the labeled regions in Figure 6.6. Helical bundles are shown in this image from a top-down view; along the axis of the helices. The helical bundle configurations shown in Figure 6.7 are the ground state structures from each labeled region, but they also occur with variations of helical packing and orientation, depending on the temperature.

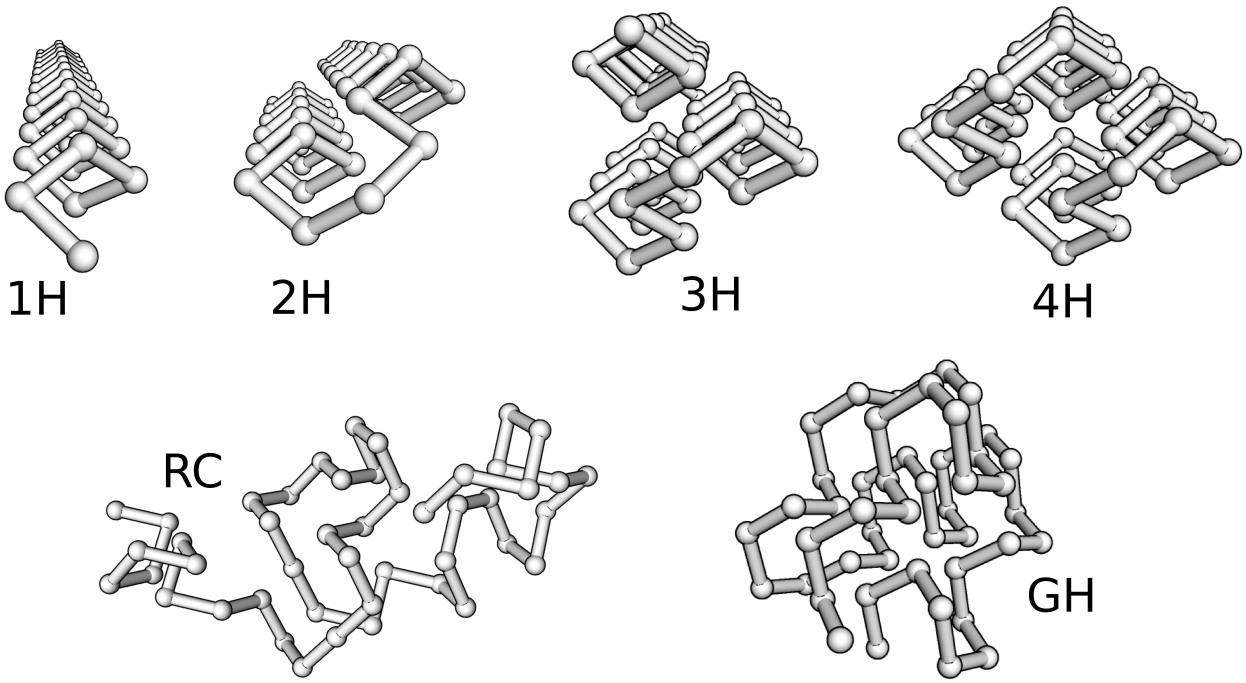

Figure 6.7: Representative configurations for the fcc lattice ISAW with 46 residues taken at each of the structural regions labeled in Figure 6.6. Helical bundles are shown with a view along their elongated axes.

6.2 H0P model with helical motifs on the fcc lattice

As a test for the model helices in the context of protein folding, we utilize the three-letter H0P model while introducing the distance-dependent contact energies and torsional penalties from the previous section. The H0P model is a simple extension of the HP model that has been shown to greatly reduce the ground state degeneracy for sc lattice proteins [47, 50], and is originally motivated by a model [48] which splits the classification of polar amino acids into three different groups based on electrostatic charge. We use the 46-residue Crambin protein for our tests, as its H0P mapping has been previously studied on the sc lattice, and it has a known folded state that contains two α-helix segments. The HP and H0P mappings for Crambin are listed in Table 3.2 from Section 3.4. In All of the H0P model simulations for the following discussions, the couplings for favorable energetic contacts are set as $\varepsilon_{_{HH}} = 2$ and $\varepsilon_{_{H0}} = 1$.

To give an example of the similarities and differences between folding simulations with the H0P and HP models, the C_V/ℓ is shown for each model in Figure 6.8, for both fcc and sc lattices. All of the C_V/ℓ data show a collapse transition at $k_BT/(\mathbb{c}_n\varepsilon_{_{HH}}) \approx 0.10$, but a notable difference between the H0P and HP models on both fcc and sc lattices, is that there are more distinct folding transitions with the H0P model near $k_BT/(\mathbb{c}_n\varepsilon_{_{HH}}) \approx 0.025$. While previous sc lattice studies with the H0P model for Crambin focused on the reduction of structural degeneracy, we instead examine effects of the distance dependence and torsional penalties that our lattice version for α-helices has on predicted low-energy structures.

In the following discussions, the quantities d_1 and d_3 will refer to nearest neighbor and third-nearest neighbor contact distances on the fcc lattice, respectively. These distances are associated with the $\mathbb{c}_1$ and $\mathbb{c}_3$ groups of neighbors that are shown in Figure 6.3, and when scaled to biological units, have magnitudes of $d_1 = 3.8$Å and $d_3 = 6.6$Å. Additional subscripts are added to the contact energy terms in subsequent variants of the H0P model Hamiltonian, where, for example, the standard H0P model Hamiltonian with nearest-neighbor contact energies would be written as: $\mathcal{H} = -(n_{_{1,HH}}\varepsilon_{_{1,HH}} + n_{_{1,H0}}\varepsilon_{_{1,H0}})$.

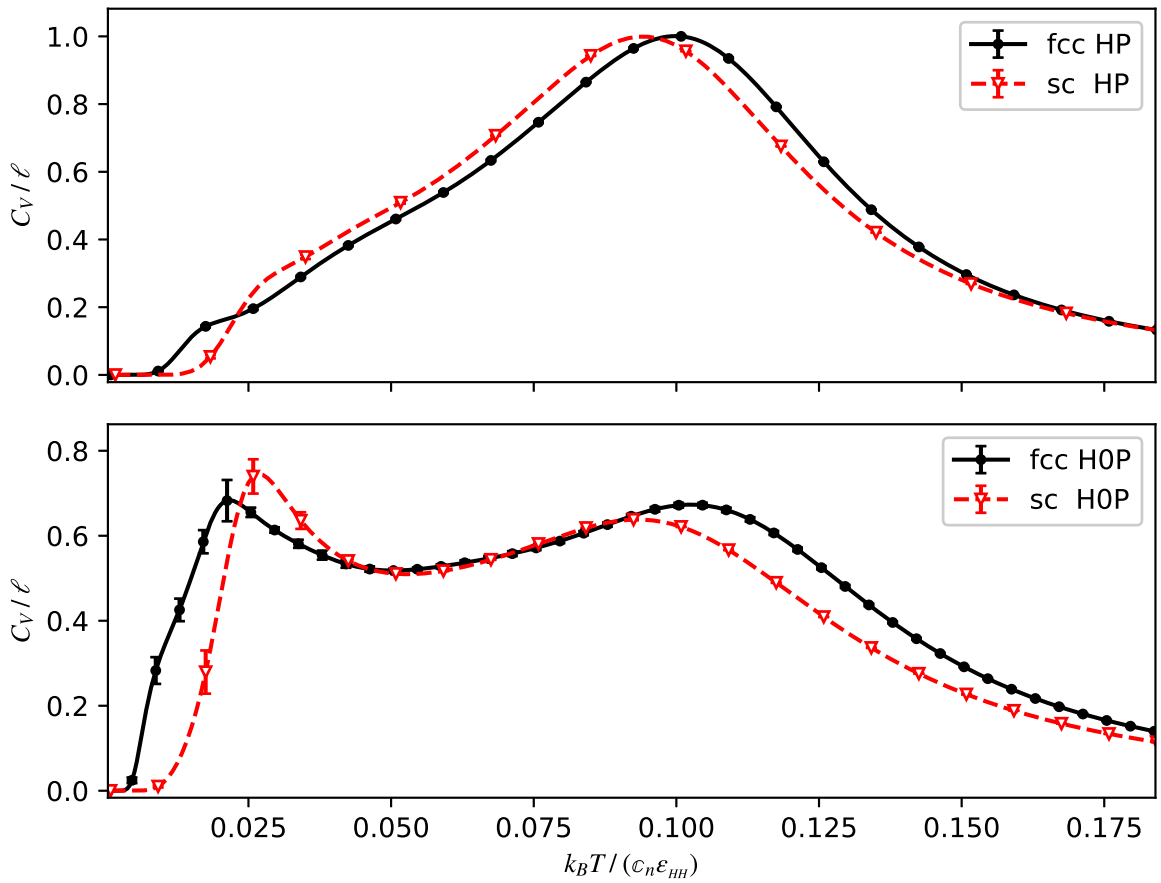

Figure 6.8: Comparison of C_V/ℓ for the HP (top) and H0P (bottom) models of Crambin on fcc (black points) and sc (red triangles) lattices. The temperature axis is normalized by the coordination $\mathfrak{c}_n$ number for each lattice. Energetic couplings are set to ($\varepsilon_{HH} = 2$, $\varepsilon_{H0} = 1$) for the H0P model simulations. Data for sc lattice was provided by Alfred Farris. Error bars are shown and are smaller than the size of data points where not visible.

When adding energetic penalties between nearest-neighbor contacts, the notation $-n_1\varepsilon_1$ is used to relate that all residue types are penalized equally at distance d_1. As with the previous section, torsional energy penalties are made whenever four consecutive residues form a dihedral angle $\phi \neq 55°$, and are incorporated with the notation $E_\phi = -n_\phi\varepsilon_\phi$.

Figure 6.9 compares C_V/ℓ for folding simulations of the H0P model for Crambin with the distance dependences associated with our α-helix motif. The black curve shows the standard H0P model, while the blue curve shows the H0P model with favorable energetic contact distances set to d_3. Both of these curves show two-step folding behavior, where the collapse transition for contacts distances set to d_3 occurs at a temperature that is nearly twice

that of the standard H0P model. Incorporating the nearest-neighbor repulsion ($\varepsilon_1 = -1$) for a H0P model with favorable energetic contact distances set to d_3, the green curve again shows the two-state folding behavior. In each model, a final folding transition before the ground state structure is attained occurs near $k_BT/\varepsilon_{HH} \approx 0.25$.

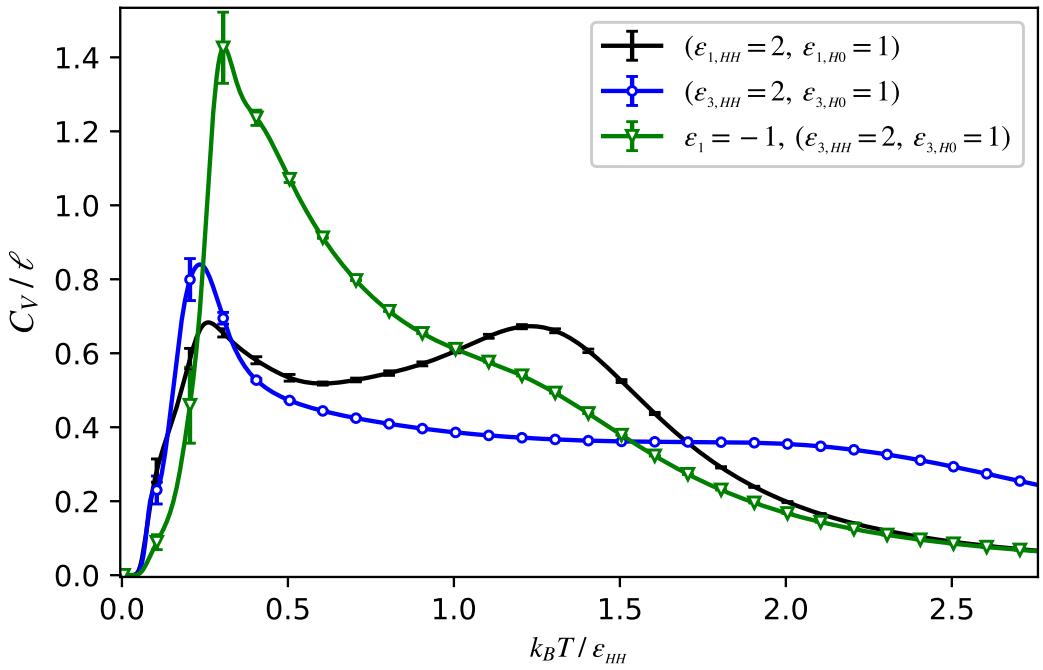

Figure 6.9: Comparison of C_V/ℓ for the standard H0P model (black curve), H0P model with contact distances set to d_3 (blue circles), and H0P model with nearest-neighbor repulsion and contact distances set to d_3 (green triangles). Error bars are shown and are smaller than the size of data points where not visible.

Unlike the fcc ISAW with helical motifs, where REWL simulations converged to a $\ln[\hat{g}(E)]$ quite rapidly, it is extremely challenging to sample low-energy states with the H0P model and the helix motif. This is likely due to the additional constraint that the H0P sequence puts on the acceptance of reptation and bond-rebriding trial moves, which would otherwise always be accepted if valid for the ISAW. For this reason, we focus on the prediction and structural analysis of helix-containing, low-energy states rather than a study of thermodynamics or structural degeneracy.

We use the tree data structure from Appendix A5 to save all unique configurations found at predicted minimal energy states for average structural calculations. Configurations are

recorded over the course of very long REWL simulations, which do not focus on converging $\ln[\hat{g}(E)]$, but rather have 10^7 MC sweeps per histogram check interval. These sets of configurations are almost certainly not complete (*i.e.* the true $g(E)$), but contain the largest number of unique conformations ($N_{confs.}$) found during the allotted simulations.

From the $N_{confs.}$ unique configurations found, the average contact maps are calculated to show the distribution of contact distances d_1 and d_3 under the various H0P model schemes. Figures 6.10, 6.11, 6.12, and 6.13, show representative configurations and contact maps for: the H0P model, the H0P model with contact distances set to d_3, the H0P model with contact distances set to d_3 and nearest-neighbor repulsion, and the H0P model with contact distances set to d_3, nearest-neighbor repulsion, and torsional penalties. The corresponding Hamiltonian and values of energetic couplings are shown next to the configuration snapshot in each figure. Contact maps are shown for both d_1 and d_3, even if the Hamiltonian does not have energetic interactions at those distances.

For the H0P model with interaction distance set to d_1 or d_3, the resulting ground state has a compact, globular structure with a dense H-core surrounded by a layer of 0 residues. Average contact maps are comparable for the two cases in Figures 6.10 and 6.11, where energetic interactions only between residues separated by d_3 still gives rise to a H-core with many nearest-neighbor contacts. This structural similarity is only loosely interpreted, as it could be due to 'smeared out' average due to the high degeneracy observed in these two cases, where $N_{confs.} \geq 2.49 \times 10^4$. With the addition of nearest-neighbor repulsion, the number of found unique configurations plummets by around three orders of magnitude. Unsurprisingly, the average number of HH contacts for d_1 nearly vanishes, where approximately 10 nearest-neighbor HH contacts are likely to form within the H-core for the case of Figure 6.12. Most of these d_1 contacts appear to be long-range, as the data is sparse near the contact map diagonal, and visually inspected configurations tend to have but a few 60° angles. Although relatively strong nearest-neighbor repulsions are considered here, one may wish to incorporate lower penalty strengths to tune the balance of contact and helical energies.

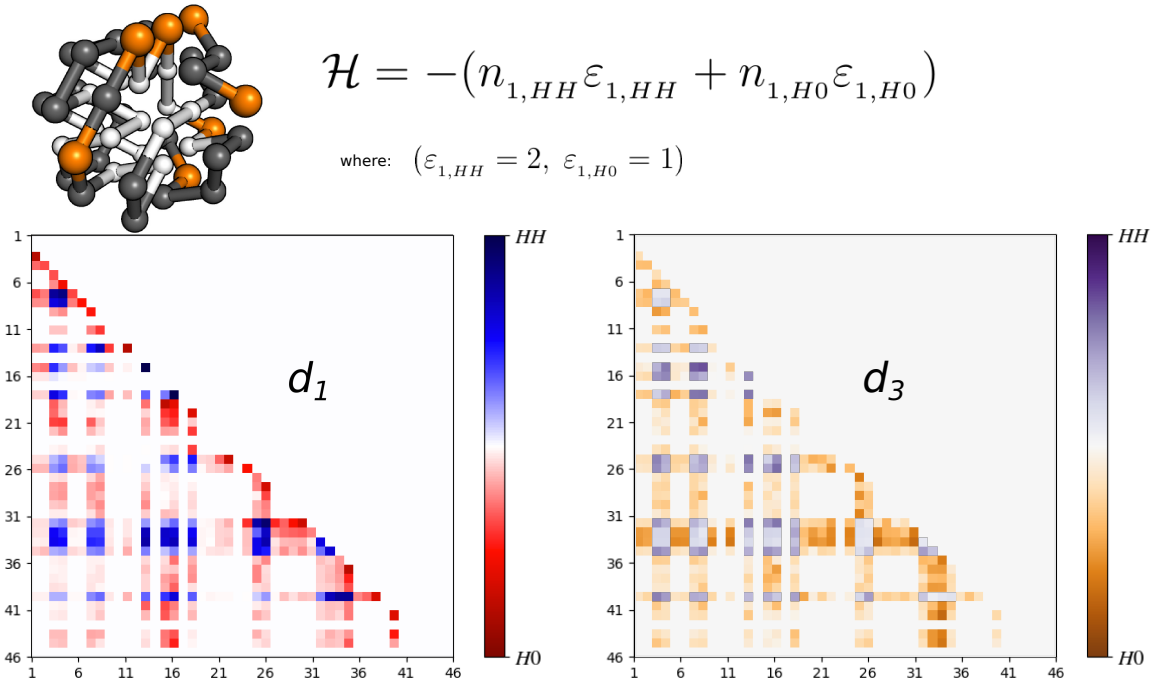

Figure 6.10: Ground state for the standard H0P model of Crambin. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} = 3.93 \times 10^5$ structures are used.

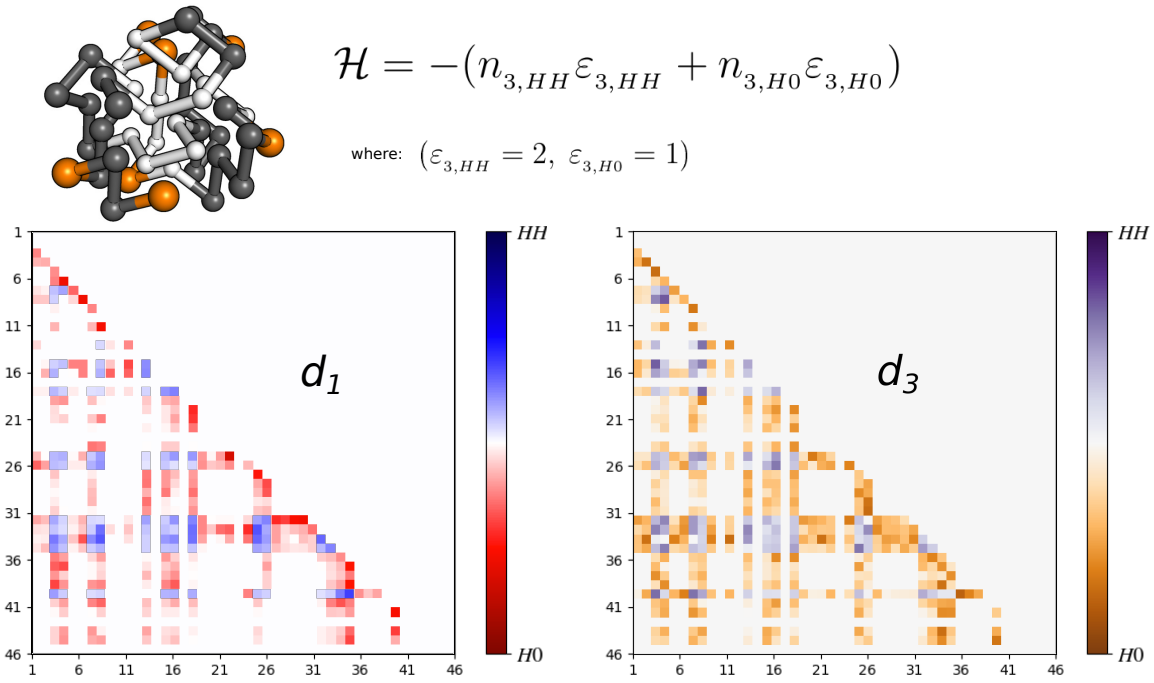

Figure 6.11: Ground state for H0P model of Crambin with contact distances set to d_3. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} = 2.49 \times 10^4$ structures are used.

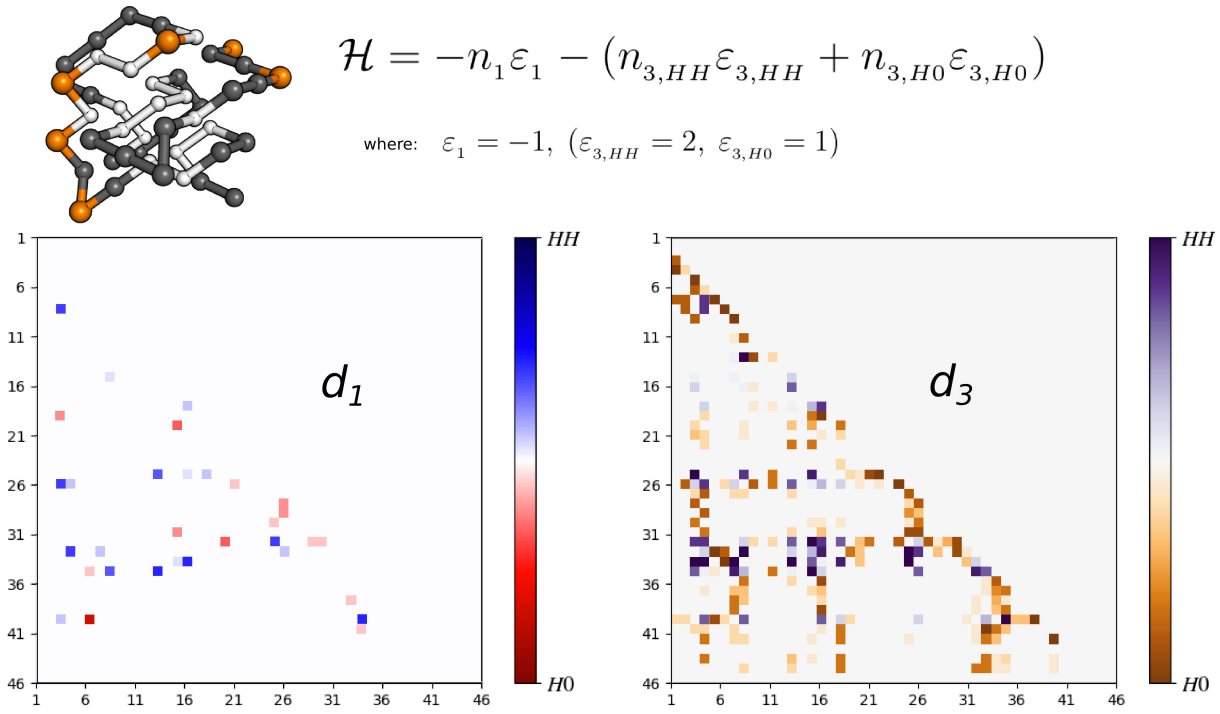

Figure 6.12: Ground state for H0P model of Crambin with contact distances set to d_3 and nearest-neighbor repulsion. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} = 71$ structures are used.

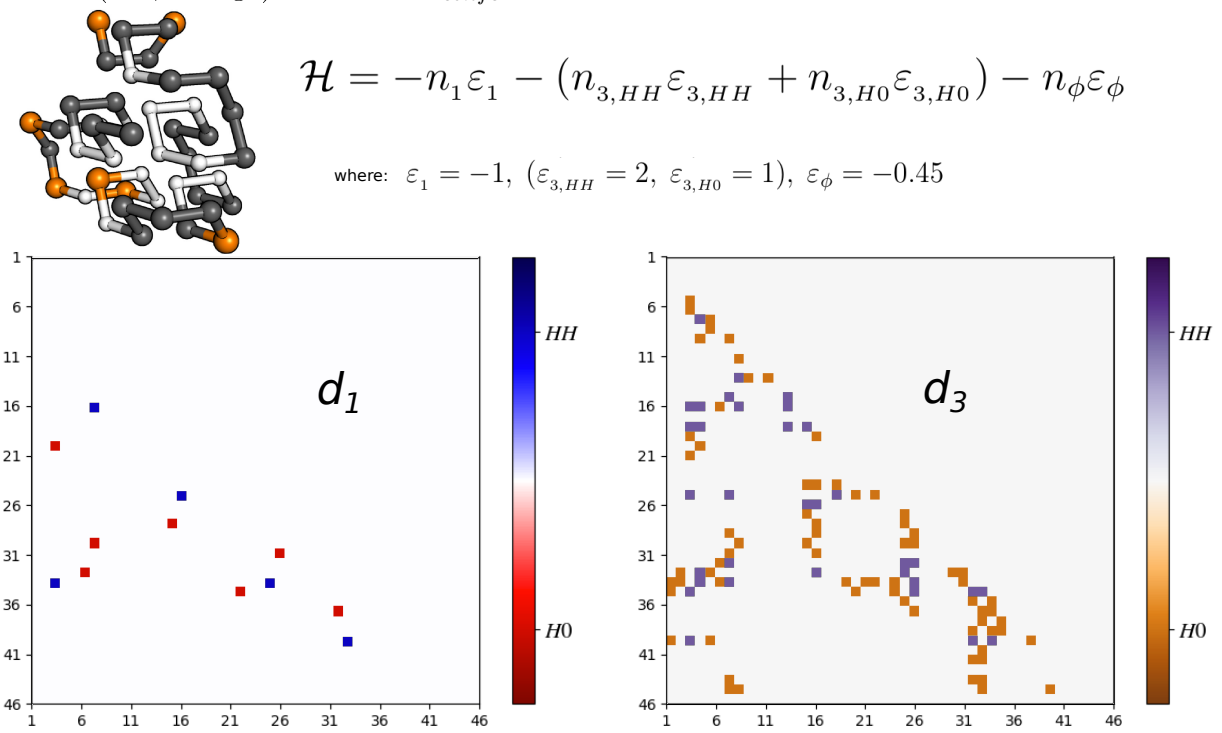

Figure 6.13: Ground state for H0P model of Crambin with contact distances set to d_3, nearest-neighbor repulsion, and torsion penalties. Contact maps for d_1 (left) and d_3 (right) show HH (blue, purple) and H0 (red, orange) contacts. $N_{confs.} = 1$ structures are used.

Finally, the H0P model with contact distances set to d_3, nearest-neighbor repulsion, and torsion penalties ($\varepsilon_{_\phi} = -0.45$) has only a single ground state structure that could be identified with our methodology in the given timeframe. Figure 6.13 shows that there are a total of 12 nearest-neighbor contacts in the structure, and helical content in the form of a disordered helix bundle. A helix that makes two full twists, and one with 1.5 twists, are packed next to one another in the configuration shown in Figure 6.13, where two more separate α occur in the more disordered region on the surface of these two helices.

A measure that is commonly used to compare model protein structures from simulations ($\{\vec{r}_{sim}\}$) to those measure in experiments ($\{\vec{r}_{exp}\}$), the distance root mean squared deviation ($dRMSD$) [110] is defined in Equation 6.2.

$$dRMSD = \sqrt{\frac{\sum_{i=1}^{\ell-1}\sum_{j=i+1}^{\ell}\left(|\vec{r}_{sim}^{(i)} - \vec{r}_{sim}^{(j)}| - |\vec{r}_{exp}^{(i)} - \vec{r}_{exp}^{(j)}|\right)^2}{\frac{1}{2}\ell(\ell-1)}} \tag{6.2}$$

Crambin has an experimentally measured structure [111] that is comprised of 46% helices, with the rest being β-sheets and disordered regions. For a more complete study, the H0P model with helical motifs should be tested with many PDB entries that have varying degrees of helical content.

Table 6.1 lists the average $dRMSD$ values in Angstroms and number of unique configurations used in the calculations for variants of the H0P model for Crambin on the fcc lattice. As a sense of scale, the average $dRMSD$ value for the fcc lattice reported by the LatFit [12,100] tool is ≈ 1.5Å . LatFit optimizes the best fitting backbone structure to match a desired PDB entry without considering any energy potential, and therefore has a $dRMSD$ that is much lower than our *ab initio* results. The first row in Table 6.1 lists data for the standard HP model of Crambin, which has an average $dRMSD$ of 6.48$\pm$0.06Å . This value for the HP model is $\sim$ 1Å larger than that reported for the fcc lattice HP model in another study [112], but we have a much larger sample size for the average calculation. Introducing

the standard H0P model for Crambin results in an increase of the $dRMSD$ by ~ 1.5Å . The last three rows have successively decreasing $dRMSD$ values as the H0P interaction distance is increased to d_3, the nearest-neighbor repulsion is added, and the torsion penalty is included. An H0P model with contact distances set to d_3 still retains a relatively large sample size (2.49×10^4), but results in a $dRMSD$ value that is ~ 0.3Å lower than its counterpart with an interaction distance of d_1. Adding the nearest-neighbor to the H0P model repulsion shows a significant decrease in $dRMSD$ standard deviation and the number of identifiable configuration for the averaging. Unlike the fcc ISAW model with helices, we observe that a much narrower range of ε_{ϕ} values result in low-energy configurations that have less than 4 helices.

Model	$\langle dRMSD \rangle$ (Å)	$N_{confs.}$
$\varepsilon_{1,HH} = 1$ (HP)	6.48 ± 0.06	3.93×10^5
$(\varepsilon_{1,HH} = 2,\ \varepsilon_{1,H0} = 1)$	7.04 ± 0.02	1.76×10^5
$(\varepsilon_{3,HH} = 2,\ \varepsilon_{3,H0} = 1)$	6.75 ± 0.02	2.49×10^4
$\varepsilon_1 = -1,\ (\varepsilon_{3,HH} = 2,\ \varepsilon_{3,H0} = 1)$	5.64 ± 0.003	71
$\varepsilon_1 = -1,\ (\varepsilon_{3,HH} = 2,\ \varepsilon_{3,H0} = 1),\ \varepsilon_{\phi} = -0.50$	5.55	1
$\varepsilon_1 = -1,\ (\varepsilon_{3,HH} = 2,\ \varepsilon_{3,H0} = 1),\ \varepsilon_{\phi} = -0.45$	$\boxed{4.99}$	1

Table 6.1: Average $dRMSD$ values for H0P model variants of the Crambin protein. The first column gives parameter values of the models: HP model (row 1); H0P model (row 2); H0P model with contact distances set to d_3 (row 3); H0P model with contact distances set to d_3 and nearest-neighbor repulsion (row 4); H0P model with contact distances set to d_3, nearest-neighbor repulsion, and torsion penalty (rows 5 and 6). The last column gives the number of unique configurations used in each calculation.

Narrowing in on a range $0.3 \leq \varepsilon_{\phi}/\varepsilon_{HH} \leq 0.5$, we note that most $dRMSD$ values for the lowest found energy states are ≈ 5.5Å . For the H0P model of Crambin with $\varepsilon_1 = -1,\ (\varepsilon_{3,HH} = 2,\ \varepsilon_{3,H0} = 1),\ \varepsilon_{\phi} = -0.45$, we find our best agreement with the experimental structure at $dRMSD = 4.99$Å . In our calculations, this shows an improvement of ~ 2.0Å to the $dRMSD$ when compared to the standard H0P model for Crambin.

For visual reference, the experimentally measured backbone structure for Crambin is shown in Figure 6.14 with the H0P mapping applied. Crambin has a mixed structure with two α-helices and β-sheet content.

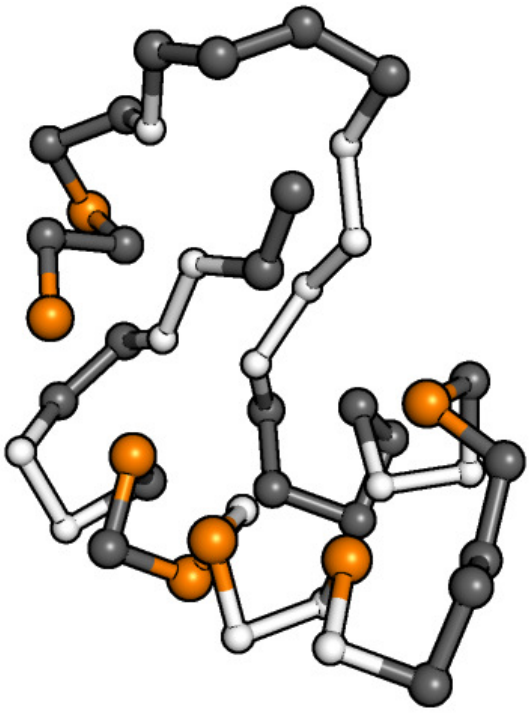

Figure 6.14: Experimentally measured backbone structure from the PDB (1CRN) for Crambin. The H0P mapping is shown on top of the PDB structure.

Chapter 7

Lattice Model Simulations of Amyloid Protofibrils

7.1 Aggregation model on the sc lattice

As a general feature of polypeptide systems, the amyloid state offers a thermodynamically favorable alternative to individually folded or dissolved, disordered states for constituent subunits. Amyloid formation can be studied by a variety of different computational models, where the choice of resolution and complexity must be made considering limitations like large time and length scales, along with large number of possible intra- and intermolecular orientations. A range of model resolutions have been employed; from coarse-grained representations using rectangular prisms [24, 25] and spherocylinders [28, 113, 114], mid-resolution lattice [26, 115–118] and continuum models [119–124], and high-resolution all-atom models [125]. Low- and mid-resolution models are often used to study the mechanism(s) in which intermediate oligomeric species form using thermodynamic, kinetic, and dynamical analyses. The goal of our simulations is to study a mid-resolution model for the qualitative thermodynamics of protofibril formation, including the initial aggregation of dissolved peptide subunits, the distribution of protofibril sizes and hydrophobic clusters, and the stacking

of β-sheets to form multi-layered protofibrils. By extending the methodology for the H0P model protein on the sc lattice [61] to handle simulations with multiple interacting sequences, we model aggregation and protofibril formation of short peptide subunits.

Motivated by generic peptide properties rather than any specific real peptide or protein fragments, we adopt short sequences of alternating H and P residues, which are known to form amyloid fibrils [21]. Neutral residues are incorporated as placeholder, non-interacting residues. The model incorporates energetic interactions between three hydrophobicity groups: intramolecular HH (ε_{HH}), intermolecular HH (ε'_{HH}), and intermolecular PP (ε'_{PP}). In each simulation, a periodically bounded box with sides of length $\max\{\ell, N\} + 2$ is used.

We focus on an aggregation model consisting of N identical β-hairpin subunits, with $\ell = 14$ and the H0P sequence: 0HPHPH00HPHPH0. This sequence length was chosen such that it is long enough to enable both globular states with intramolecular HH contacts and β-hairpin-type structures, but also short enough to allow the fast simulation of dozens or more subunits. Neutral residues are used to connect two segments of alternating H and P residues, along with capping the termini, where each subunit has a maximum of 6 HH contacts in a globular state, and a maximum of 5 HH contacts as a planar β-hairpin structure. The globular and planar subunit states that contain intramolecular HH contacts are shown in Figure 7.1.

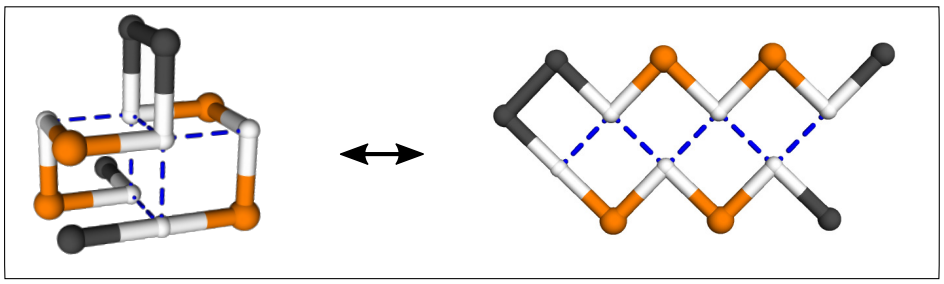

Figure 7.1: The β-hairpin subunit on the sc lattice, where the left image is the optimal globular state with 6 HH contacts, and the right image is the planar β-hairpin structure with 5 HH contacts. Dashed, blue lines show intramolecular HH contacts.

To give a sense of scale for the breadth of state space in the lattice aggregation model, the $\ln[\hat{g}(E)]$ from a REWL simulation of 24 β-hairpin subunits with weak intermolecular coupling (ε_{HH}=10, ε'_{HH}=5, ε'_{PP}=4) is shown in Figure 7.2. Spanning nearly 300 orders of magnitude, the $\ln[\hat{g}(E)]$ has a complex shape that is shown in the inlay of Figure 7.2, which persists until $E \approx -1100$. This 'zig-zag' shape in $\ln[\hat{g}(E)]$ has a period-doubled appearance resulting from discrete combinations of the energetic interactions (and therefore the discrete groupings of subunits into clusters), where repeating peaks in $\ln[\hat{g}(E)]$ occur with the two highest magnitudes at multiples of ε_{HH} and ε'_{HH}.

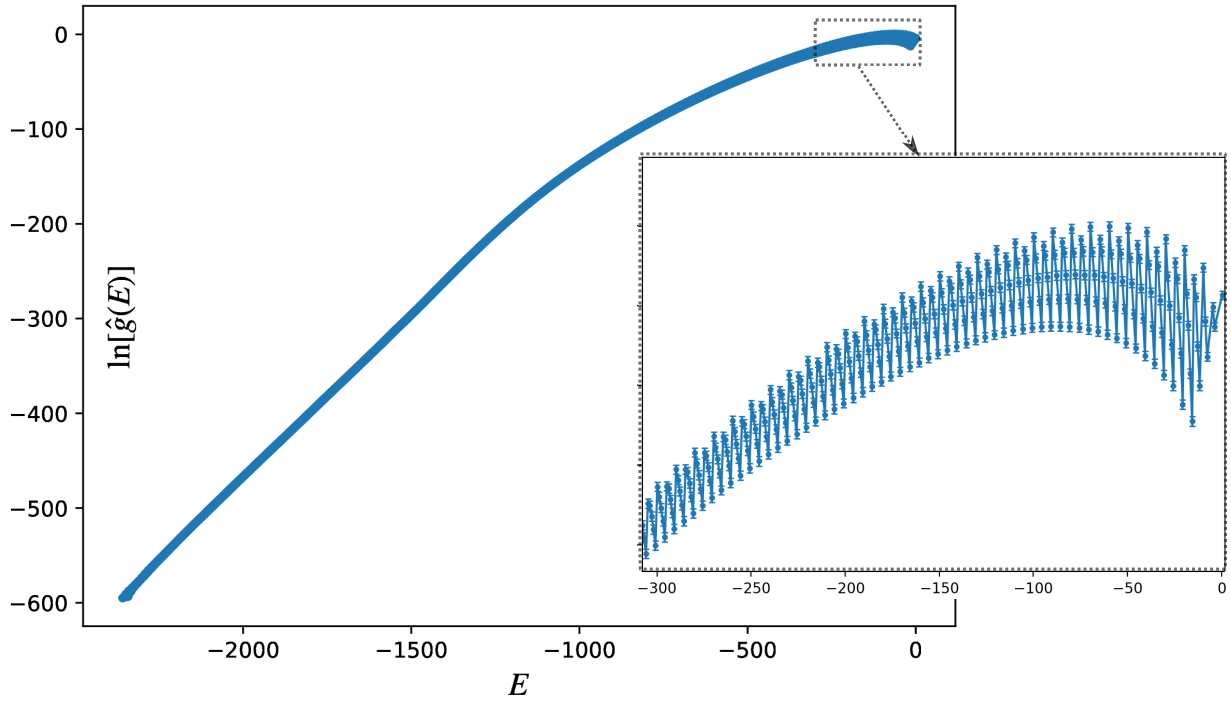

Figure 7.2: Natural logarithm of the density of states for 24 β-hairpin subunits on the sc lattice where $\varepsilon_{HH} = 10$, $\varepsilon'_{HH} = 5$, and $\varepsilon'_{PP} = 4$. The inset shows a closeup view of $\ln[\hat{g}(E)]$ at high energies.

7.1.1 Thermodynamics of protofibril formation

To study the formation of protofibrils that can have one or more stacked layers, systems with β-hairpin subunits are simulated on the sc lattice with ($\varepsilon_{HH} \geq \varepsilon'_{HH}$), and ε'_{PP} is tuned to control the number of layers in the protofibril ground state. The internal energy, shown in Figure 7.3, is calculated from the $\ln[\hat{g}(E)]$ in Figure 7.2, but shows typical behavior found in our aggregation simulations for all parameter sets. There are two regions where U decreases with respect to temperature, but the sharp, nearly discontinuous drop at around $k_BT/\varepsilon_{HH} \approx 0.3$ is the hallmark indicator of a 'first-order' transition (quotations are used because we are studying a finite system here, so there is technically no phase transition).

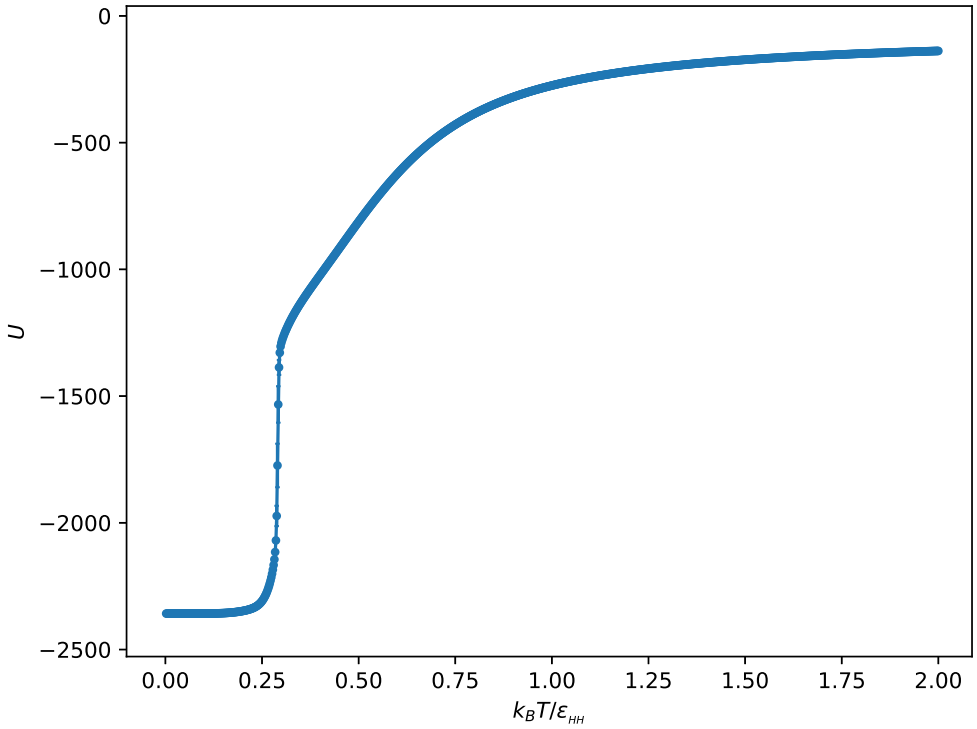

Figure 7.3: Internal energy for 24 β-hairpin subunits on the sc lattice where $\varepsilon_{HH} = 10$, $\varepsilon'_{HH} = 5$, and $\varepsilon'_{PP} = 4$. Error bars are shown and are smaller than the size of data points where not visible.

Energetic changes are clearly represented by the specific heat, which is plotted as the black, dashed curve in Figure 7.4 for a system of 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). Two distinct maxima are observed in $C_V/(N\ell)$

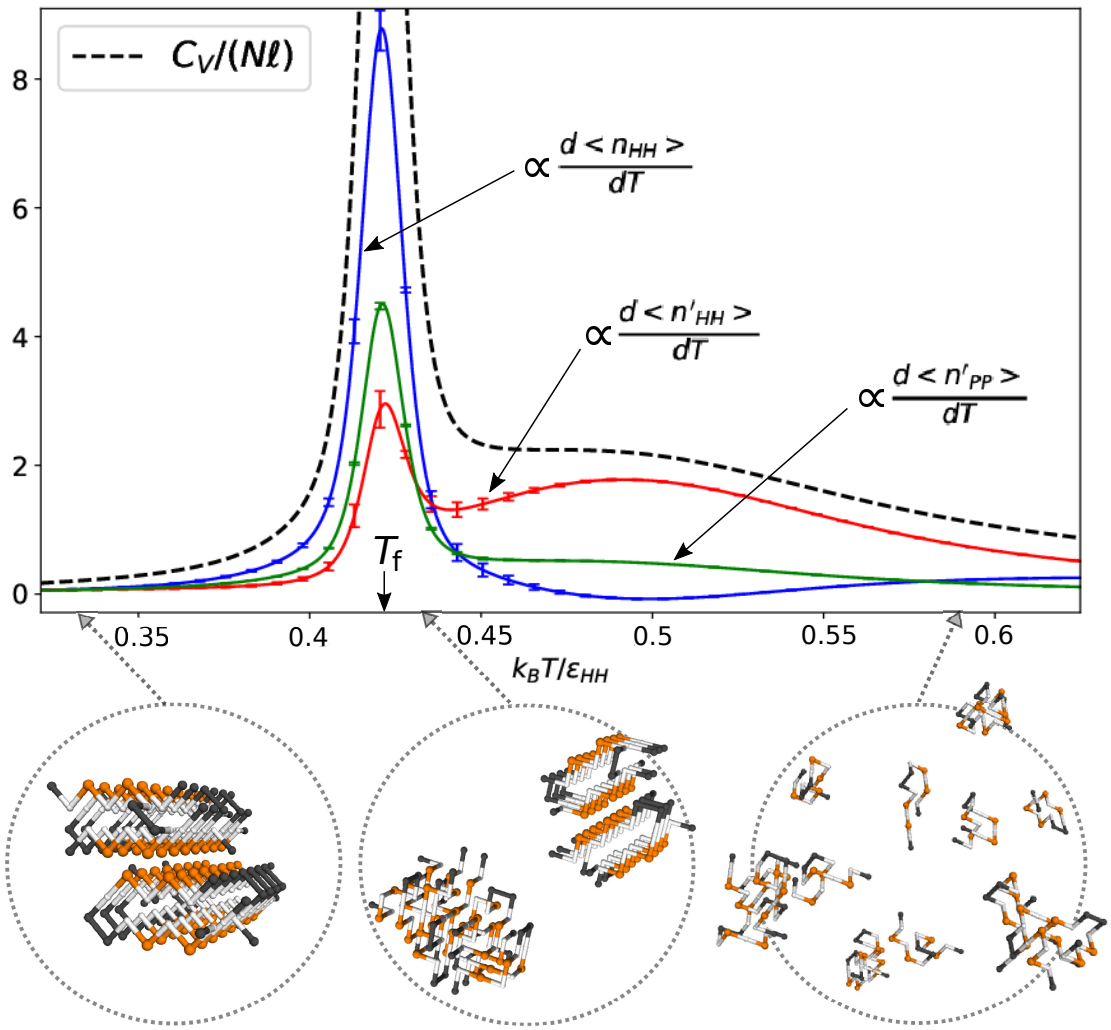

Figure 7.4: Thermodynamics for stacked protofibril formation of 16 β-hairpin subunits where $\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, and $\varepsilon'_{PP} = 1$. The specific heat (blacked, dashed curve) and contributions from thermal derivatives of average contact numbers (blue for intramolecular HH, red for intermolecular HH, and green for intermolecular PP) are plotted. Representative configurations are shown below the x-axis with arrows signifying the temperatures where they are recorded. The temperature for the structural transition of protofibril formation T_f is labeled with an arrow.

that signal structural transitions: one at $k_BT/\varepsilon_{HH} \approx 0.5$ and another at $k_BT/\varepsilon_{HH} \approx 0.425$. The peak in $C_V/(N\ell)$ at higher temperatures represents a condensation transition, where subunits are initially dispersed across a large volume in small clusters and individually folded configurations, but then aggregate to form a small number of disordered oligomers. At a temperature of $k_BT/\varepsilon_{HH} \approx 0.425$ (labeled as T_f in Figure 7.4), the peak in $C_V/(N\ell)$ corresponds to an abrupt drop in the internal energy, like the one shown in Figure 7.3. T_f will

be referred to as the fibrillization temperature. This is a structural transition where ordered β-hairpin content emerges, and any remaining clusters consolidate to form a protofibril with on or more stacked layers. In the presented data, the ground state is a two-layer, stacked protofibril structure. Configurations below the x-axis in Figure 7.4 show these structural transitions, with arrows pointing to the temperatures where each snapshot was taken.

Averages of the contact numbers, $\langle n_{HH} \rangle$, $\langle n'_{HH} \rangle$, and $\langle n'_{PP} \rangle$, and their thermal derivatives further clarify the structural transitions suggested by the peaks in $C_V/(N\ell)$. These quantities are plotted as colored curves in Figure 7.4, with arrows matching labels to their corresponding data. Each quantity is weighted by the negative value of their respective energetic couplings and normalized by $N\ell$ to show the contributions to $C_V/(N\ell)$. The presented form of these thermal derivatives should be interpreted with decreasing temperature (*e.g.* the red curve in Figure 7.4 is $\frac{-\varepsilon'_{HH}}{N\ell}\frac{d\langle n'_{HH}\rangle}{dT}$, which relates an increase in n'_{HH} as T decreases).

A decrease in $\langle n_{HH} \rangle$ is visible at the condensation transition, which occurs as subunits sacrifice some intramolecular HH contacts in favor of intermolecular contacts as they attach to a cluster. At the condensation transition, the number of intermolecular contacts ($\langle n'_{HH} \rangle$ in particular) is increasing, as would be expected during the formation of larger clusters. All of the contact numbers show a large increase at T_f, with the largest being an increase in $\langle n_{HH} \rangle$ that signifies the emergence of β-hairpin structures in each subunit. Furthermore, the simultaneous increase in all contact numbers, along with the larger increase of $\langle n'_{PP} \rangle$ compared to $\langle n'_{HH} \rangle$, suggests that the ordering happens in tandem with the formation of a second protofibril layer.

The same analysis is performed for a system of 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, and $\varepsilon'_{PP} = 1$). Shown in Figure 7.5, the $C_V/(N\ell)$ shows essentially the same behavior, but with a lower value of T_f. Thermal derivatives of the average contact numbers, however, show a contrast in the temperatures at which intramolecular and intermolecular contacts are formed, when compared to the previous case with strong intermolecular coupling. It is apparent that $\langle n_{HH} \rangle$ is increasing near $k_B T/\varepsilon_{HH} \approx 0.5$, and

an examination of the configurations reveals that subunits assume the structure in the left of Figure 7.1 while remaining dispersed. Just above T_f, an increase in $\langle n'_{HH} \rangle$ occurs at a significantly higher rate than in $\langle n'_{PP} \rangle$, which is likely due to the deposition of globular subunits onto the surface of any existing fibrillar structures, as shown in the middle configuration of Figure 7.5. At T_f, there is an abrupt decrease in $\langle n_{HH} \rangle$ corresponding to the conversion of subunits from a globular state with 6 HH contacts, to the β-hairpin structure with 5 HH contacts. Below this fibrillization temperature, the two-layer, stacked protofibril ground state is obtained. The key observation is that the order in which intramolecular and intermolecular contacts are formed is swapped when compared to the simulation with strong intermolecular coupling.

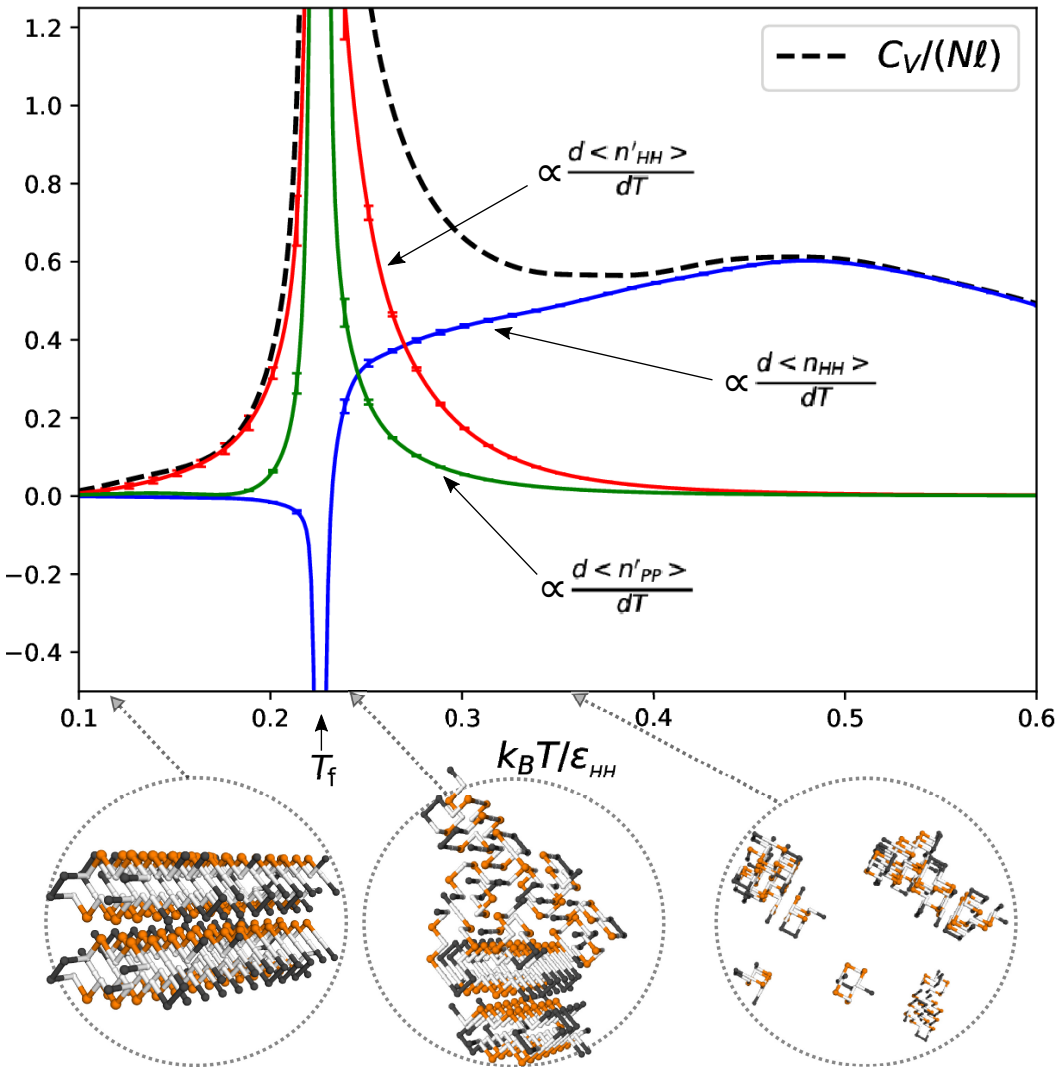

Figure 7.5: Thermodynamics for stacked protofibril formation of 24 β-hairpin subunits where $\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, and $\varepsilon'_{PP} = 1$. The specific heat (black, dashed curve) and contributions from thermal derivatives of average contact numbers (blue for intramolecular HH, red for intermolecular HH, and green for intermolecular PP) are plotted. Representative configurations are shown below the x-axis with arrows signifying the temperatures where they are recorded. The temperature for the structural transition of protofibril formation T_f is labeled with an arrow.

Before attempting to control the number of stacked β-hairpin layers through the tuning of ε'_{PP}, an extra consideration regarding degeneracy must be addressed in the case of strong intermolecular coupling ($\varepsilon_{HH} < \varepsilon'_{HH}$). Shown in Figure 7.6, two alternate cross sections for a two-layer fibril arise which differ in n'_{HH} compared to the purely β-hairpin cross section shown in Figure 3.4. Depending on the value of ε'_{HH} and number of subunits, the deformation

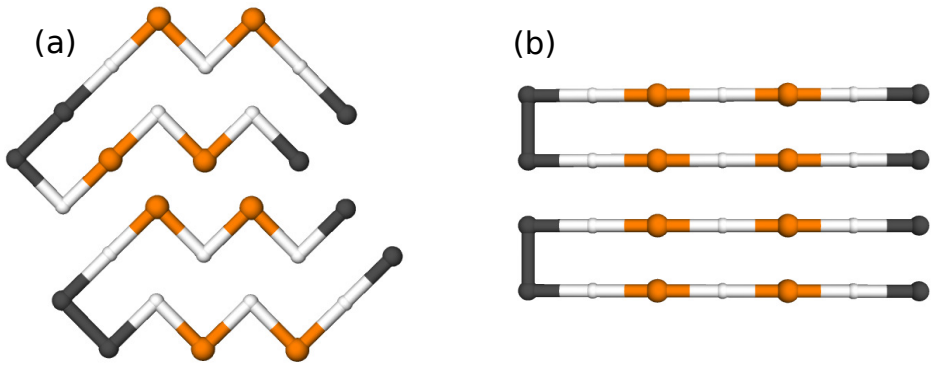

Figure 7.6: Alternate fibril cross sections that are observed in simulations with strong intermolecular HH coupling. Image (a) is equivalent in energy to the two-layer β-hairpin protofibril presented in Figure 3.4 when $\varepsilon_{HH} = \varepsilon'_{HH}$.

of additional layers to form an extra intermolecular HH contact per subunit can occur, as shown in image (a) of Figure 7.6. This structural motif is equivalent in energy to the two-layer β-hairpin when $\varepsilon_{HH} = \varepsilon'_{HH}$, and is observed in our simulations of 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). Although not shown, we also observe intermediate states as aggregated 'bundles' of the configuration in image (b) of Figure 7.6, where subunits orient such that there is no elongated axis of the protofibril.

In an attempt to explicitly control the ground state structure as a stacked β-hairpin protofibril with h layers, we write the total energy ($E_{tot}(h)$) in Equation 7.1 using energetic contributions from contacts that are perpendicular ($E_\perp$), and parallel ($E_\parallel$) to the long axis of the protofibril. Primed energy terms represent contributions from intermolecular contacts, where in this case, $E_\perp = -5\varepsilon_{HH}$, $E'_\perp = -3\varepsilon'_{PP}$, and $E'_\parallel = -(6\varepsilon'_{HH} + 4\varepsilon'_{PP})$. See Figure 3.4

for a corresponding diagram.

$$E_{tot}(h) = \underbrace{NE_{\perp}}_{\text{intramolecular}} + \underbrace{\frac{N}{h}(h-1)E'_{\perp}}_{\text{intermolecular btw. stacked layers}} + \underbrace{h\left(\frac{N}{h}-1\right)E'_{\parallel}}_{\text{intermolecular along fibril axis}} \tag{7.1}$$

From this expression, the energy change for stacked layers is simply

$$\Delta E_{tot}^{h\to h+1}(h) = \frac{N}{h(h+1)}E'_{\perp} - E'_{\parallel} \ , \tag{7.2}$$

which only depends on the intermolecular couplings. Using these set of equations, along with $E_{\perp}$, $E'_{\perp}$, and $E'_{\parallel}$ determined for the fibril types shown in Figure 7.6, a set of inequalities is constructed to choose integer values for couplings ($\varepsilon_{_{HH}}$, $\varepsilon'_{_{HH}}$, $\varepsilon'_{_{PP}}$) resulting in stacked β-hairpin protofibrils with h layers. The specific heat and ground states structure are shown for the couplings where $h = 1, 2, 3$. Displayed in the leftmost configuration of Figure 7.7, our simulation for $h = 1$ resulted in a protofibril that stopped elongation past a certain length, and oriented additional subunits along exposed patches of H residues (one position before/after the hairpin turn). The fibrillization temperature indicated by the sharp peak in $C_V/(N\ell)$ appears to decrease with the value of $\varepsilon'_{_{PP}}$.

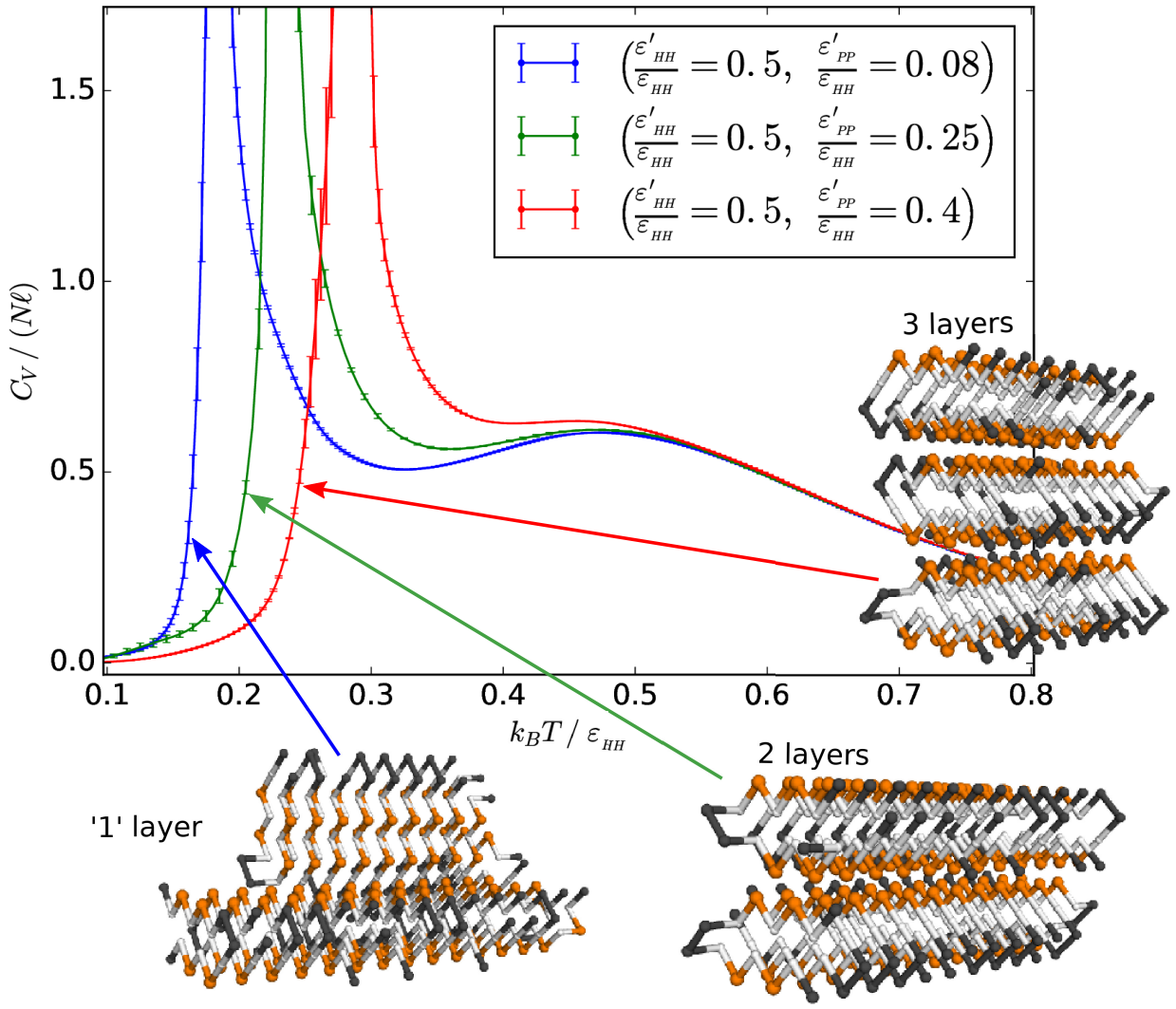

Figure 7.7: Specific heat for 24 β-hairpin subunits where the hydrophobicity scales for intermolecular interactions are chosen to promote protofibril stacking with 'one' (blue curve), two (green curve), and three (red curve) layers. Configurations of the stacked ground states are shown with arrows pointing to their respective data.

7.1.2 Free energy barriers and cluster size distributions

The prevalence of different oligomeric aggregates is valuable, yet often inaccessible, information for better understanding of amyloid fibril. Furthermore, the mechanism with which amyloid fibrils form is of long-standing interest, and is generally believed to be a nucleation and growth process [8, 24, 124]. Using the $\hat{g}(E)$ from our REWL simulations and structural data recorded in subsequent MUCA production runs, we calculate the cluster size distributions and nucleation barriers for the formation of two-layer protofibrils. The following results correspond to the two-layer β-subunit systems whose thermodynamics are presented in the previous subsection.

Free energy difference (ΔF) as a function of cluster size (m) gives a useful measure for the thermodynamic stability of oligomers observed in the protofibril formation simulations. Using F for a single subunit as the reference value, the curve for $\Delta F(m)$ gives information on whether it is more likely to add or remove subunits from a given oligomer at some specified temperature. To calculate $\Delta F(m)$, the method described in Section 4.3 is used, where $\Delta F(m) = F(m) - F(1) = -k_B T(\ln[\mathcal{Z}(m)] - \ln[\mathcal{Z}(1)])$. The value of m at the maximum of $\Delta F(m)$ is referred to as the critical cluster size (m^*), which is in an unstable equilibrium with the surrounding environment. For values of $m < m^*$, the removal of subunits reduces $\Delta F(m)$, whereas for $m > m^*$, the addition of subunits decreases $\Delta F(m)$. At some value of m, $\Delta F(m)$ can become negative, at which point the state is thermodynamically stable, as is the addition of subunits. Shown in Figure 7.8 for a system of 16 β-hairpin subunits with strong intermolecular coupling, $\Delta F(m)/k_B T$ is plotted at various temperatures. The red curve for $\Delta F(m)/k_B T$ at high temperature is monotonically increasing, suggesting that all subunits are dissolved in solution. As temperature is lowered, the curves begin to show a clear maximum at $m^* = 8$ and a value $\Delta F(m)/k_B T < 0$ for $m = 16$. This system has a ground state structure where $m = 16$ and has two layers; each containing 8 subunits. The slope of $\Delta F(m)$ increases in magnitude with decreasing temperature, which could signify

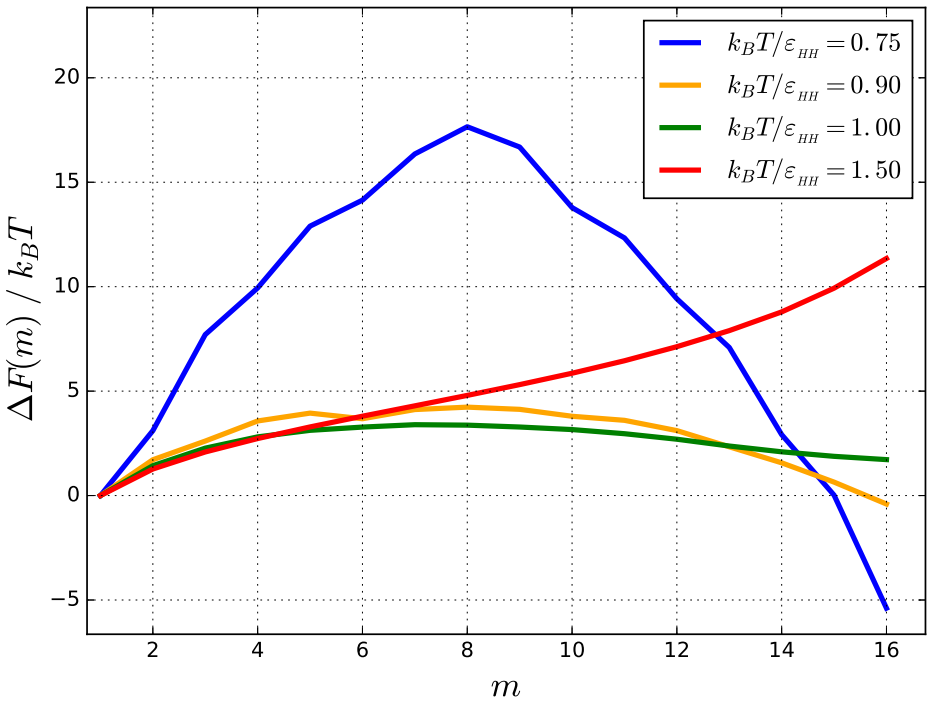

Figure 7.8: Free energy difference as a function of cluster size for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $< 10\%$ of $|\Delta F|$.

that smaller, preformed clusters are more likely to retain their sizes until they attach to an existing cluster where $m \geq 8$.

To further resolve the free energy barriers present as oligomers with various structures form, the calculation of ΔF is performed as a function of the H cluster size (m_{HH}), where the reference value is again taken to be $F(1)$. The system of 16 β-hairpin subunits has 96 total number of H residues (N_H). Figure 7.9 plots $\Delta F(m_{HH})/k_BT$ against the fraction of H residues in each cluster size. A reoccurring pattern in the plot of $\Delta F(m_{HH})$ are 'notches' that have local minima separated by barriers, and correspond to metastable states at various cluster sizes. The red curve for a temperature above the condensation transition is relatively smooth and has a critical cluster size at $m^*_{HH}/N_H \approx 0.5$. At the three lower temperatures, the curves for $\Delta F(m_{HH})/k_BT$ change shape dramatically, and contain many metastable states with intermediate free energy barriers. Near $m_{HH}/N_H \approx 0.1$ and 0.9, there are two large free energy barriers which respectively separate dissolved subunits from a two-layer

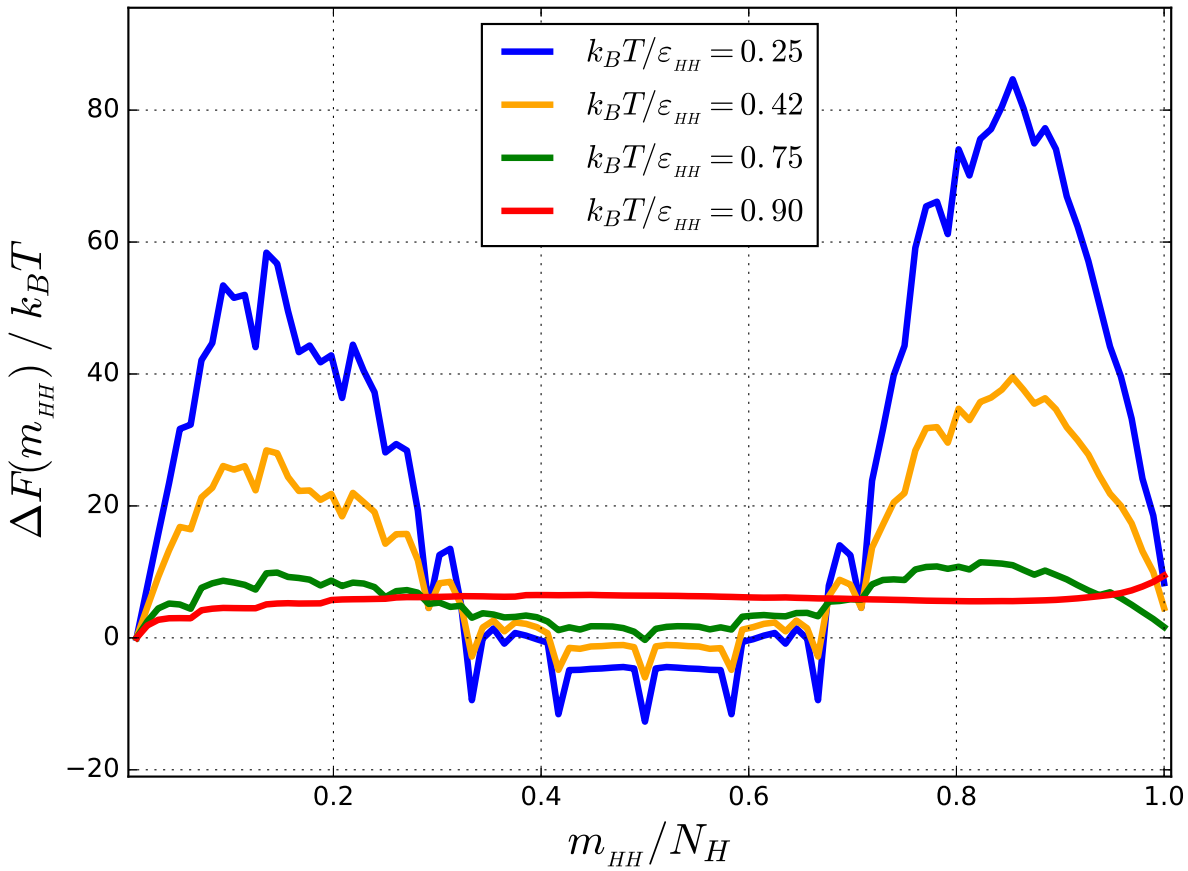

Figure 7.9: Free energy difference as a function of hydrophobic cluster size for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $< 10\%$ of $|\Delta F|$.

aggregate, and the two-layer structure from a protofibril with a single H cluster (one layer). The lower free energy barrier and presence of more local minima at $m_{HH}/N_H \approx 0.1$ compared to 0.9 suggests that, for these energetic couplings, a two-layer protofibril is more likely to assume disorder or have subunits dissociate and reattach than to enlarge or elongate. Most visible in the data for $k_BT/\varepsilon_{HH} = 0.25$, the free energy landscape within the region $0.3 \leq m_{HH}/N_H \leq 0.7$ appears symmetric. Minima in this region are all thermodynamically stable states (at least for the $k_BT/\varepsilon_{HH} \leq 0.42$ data), where the central three minima correspond to the two-layer β-hairpin protofibril ground state. Symmetry arises from complementary pairs of cluster sizes associated with the ground state degeneracy of the motif in image (a) of Figure 7.6.

Along with the free energy barriers, the distribution of cluster sizes is calculated to visualize the probability of oligomer species as a function of temperature. The heatmap in Figure 7.10 shows the ensemble average for the fraction of clusters with a given size for a range of temperatures. White regions have recorded fractions of $\leq 10^{-6}$ for those cluster sizes and temperatures. Oligomers with intermediate sizes occur predominantly at the condensation transition near $k_BT/\varepsilon_{HH} \approx 0.5$, with a fractions ≤ 0.20. Figure 7.11

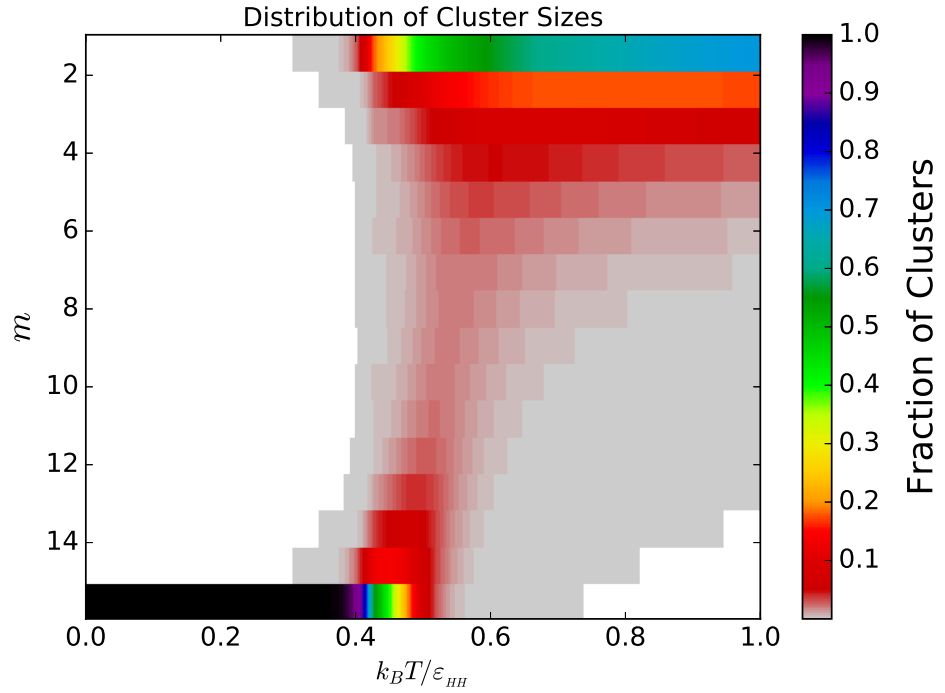

Figure 7.10: Distribution of cluster sizes for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to cluster sizes observed with a fraction $\leq 10^{-6}$.

shows the average fraction of H clusters with a given size for a range of temperatures. At $k_BT/\varepsilon_{HH} \geq 0.425$, the system is in a disperse state with a number of individual subunits and small clusters. Near the fibrillization temperature $T_f \approx 0.425$, larger H cluster sizes quickly emerge, with the maximal H cluster size occurring at a fraction of ≈ 0.05. The two-layer structure, signified by $m_{HH}/N_H = 0.5$, is predominant below T_f, with disorder becoming less prevalent as $m_{HH}/N_H \leq 0.2$. Low-energy states appear to occur with two-layer β-hairpin

structures at a fraction of ≈ 0.6, and the degenerate structure from image (a) of Figure 7.6 at a fraction of ≈ 0.4.

The same analyses are performed for the system of 24 β-hairpin subunits, but with a weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$) that is designed for a two-layer, purely β-hairpin ground state.

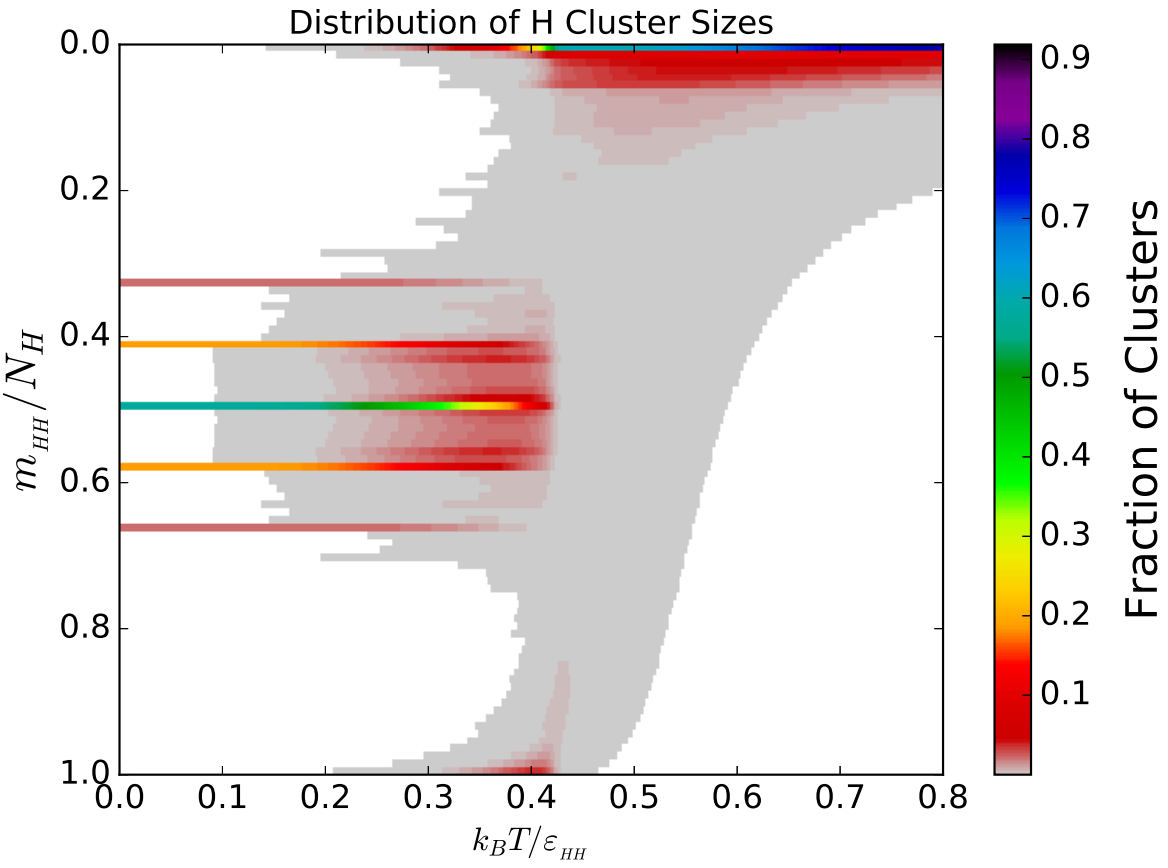

Figure 7.11: Distribution of hydrophobic cluster sizes for 16 β-hairpin subunits with strong intermolecular coupling ($\varepsilon_{HH} = 2$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to cluster sizes observed with a fraction $\leq 10^{-6}$.

Figure 7.12 shows $\Delta F(m)/k_BT$, which has similar behavior for this system as in Figure 7.8. With decreasing temperature, a critical cluster size of $m^* = 10$ appears, in contrast with $m^* = N/2$ in the case of a strong intermolecular coupling.

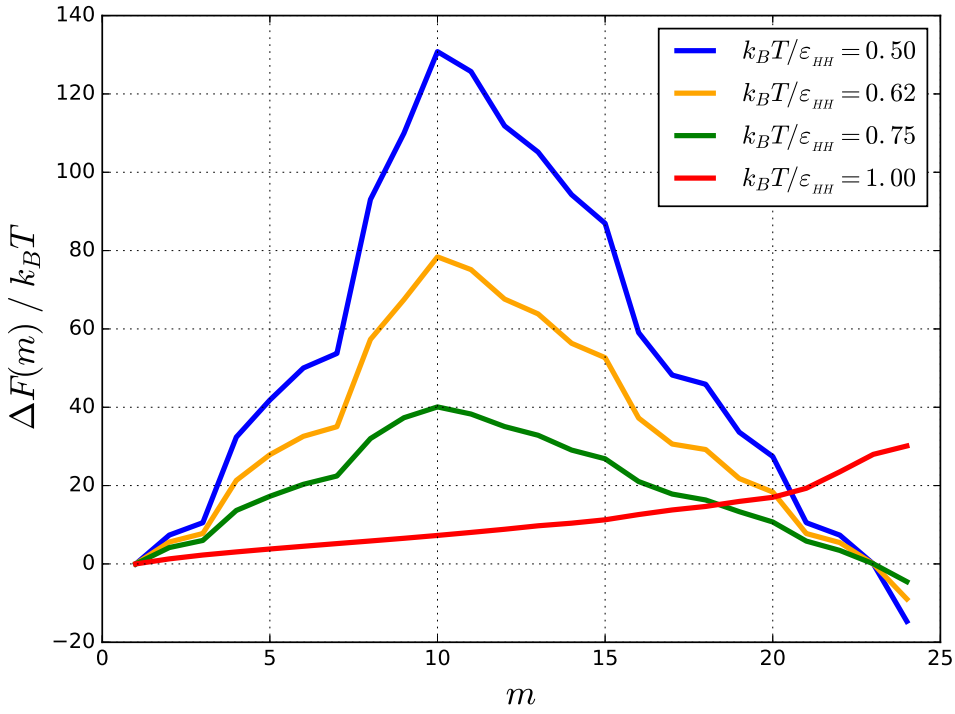

Figure 7.12: Free energy difference as a function of cluster size for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $< 10\%$ of $|\Delta F|$.

Interesting behavior is exhibited when looking at the free energy difference for H cluster formation in the case of weak intermolecular coupling. Figure 7.13 shows $\Delta F(m_{HH})/k_BT$ plotted against m_{HH}/N_H, where $N_H = 144$ for this system. Each temperature shows 24 clearly defined local minima in $\Delta F(m_{HH})/k_BT$, which correspond to the H cluster sizes associated with the N possible oligomer sizes. At $k_BT/\varepsilon_{HH} = 1.0$, the free energy barriers are of approximately constant magnitude between each consecutive local minimum, with a nearly linear increase of $\Delta F(m_{HH})/k_BT$ with m_{HH} for the minima. As temperature is decreased to $k_BT/\varepsilon_{HH} = 0.5$, free energy barriers emerge at $m_{HH}/N_H \approx 0.15$ and 0.9 that respectively separate dissolved subunits from stacked protofibrils, and the stacked structures from an aggregate with a single H cluster. Magnitudes of the free energy barriers increase with decreasing temperature, and in the data for $k_BT/\varepsilon_{HH} = 0.12$, the thermodynamically stable states ($\Delta F < 0$) are clearly visible. Free energy minima for protofibril states with 2 and 3 stacked layers are labeled with arrows, where both $h = 2$ and 3 states are ther-

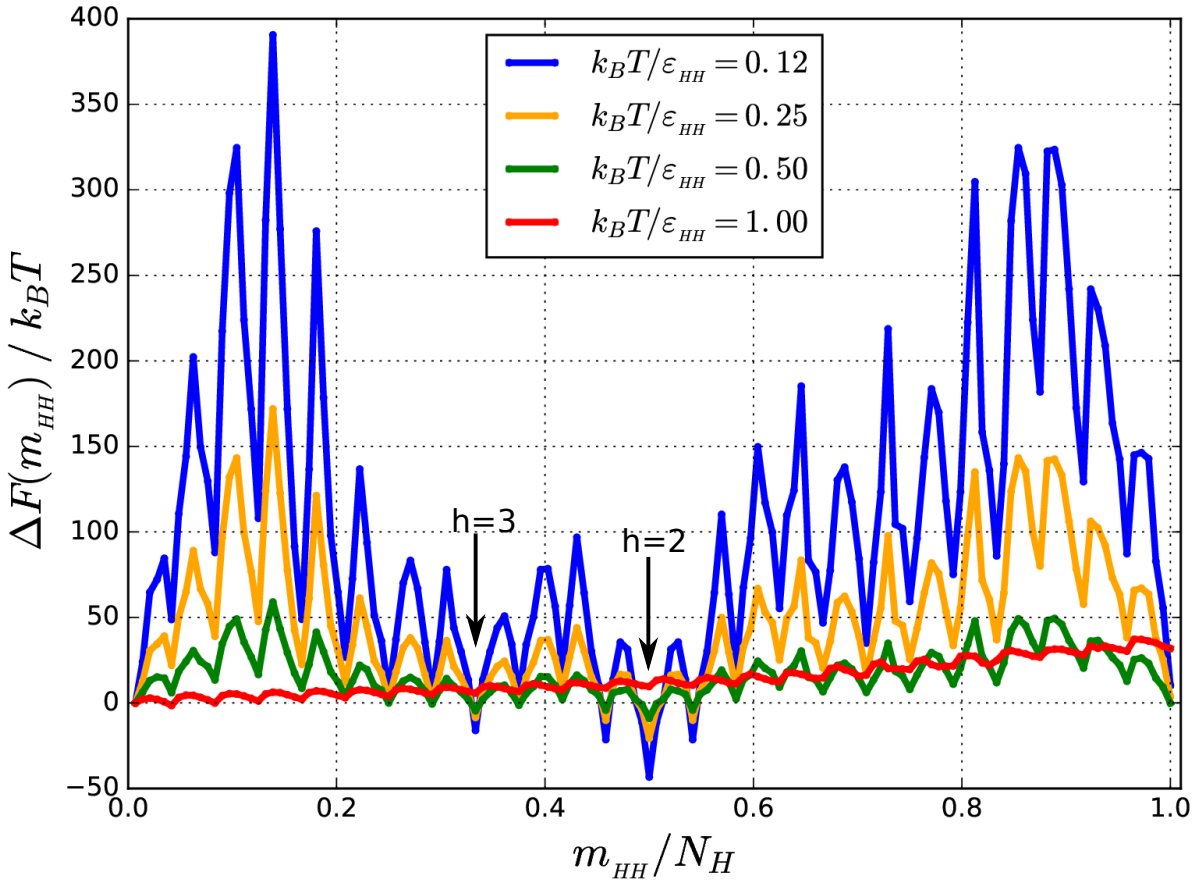

Figure 7.13: Free energy difference as a function of hydrophobic cluster size for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). The reference value is $F(1)$. For clarity, error bars are not shown, but are $< 10\%$ of $|\Delta F|$.

modynamically stable below the fibrillization temperature. The ground state structure is a β-hairpin protofibril with $h = 2$ that is non-degenerate (with the exception of stacking and packing polymorphisms and orientations of 0 residues), as signified by the global free energy minimum at $m_{HH}/N_H = 0.5$.

The heatmap in Figure 7.14 shows the ensemble average for the fraction of clusters with a given size for the system of 24 β-hairpin subunits with weak intermolecular coupling. Above k_BT/ε_{HH}, the system primarily consists of dissolved, disordered subunits and a small oligomers (dimers and trimers). The fibrillization with these energetic coupling strengths does not involve a significant number of intermediate oligomer sizes, as suggested by the suppressed fraction sizes between $8 < m < 16$ for $k_BT/\varepsilon_{HH} \approx 0.225$.

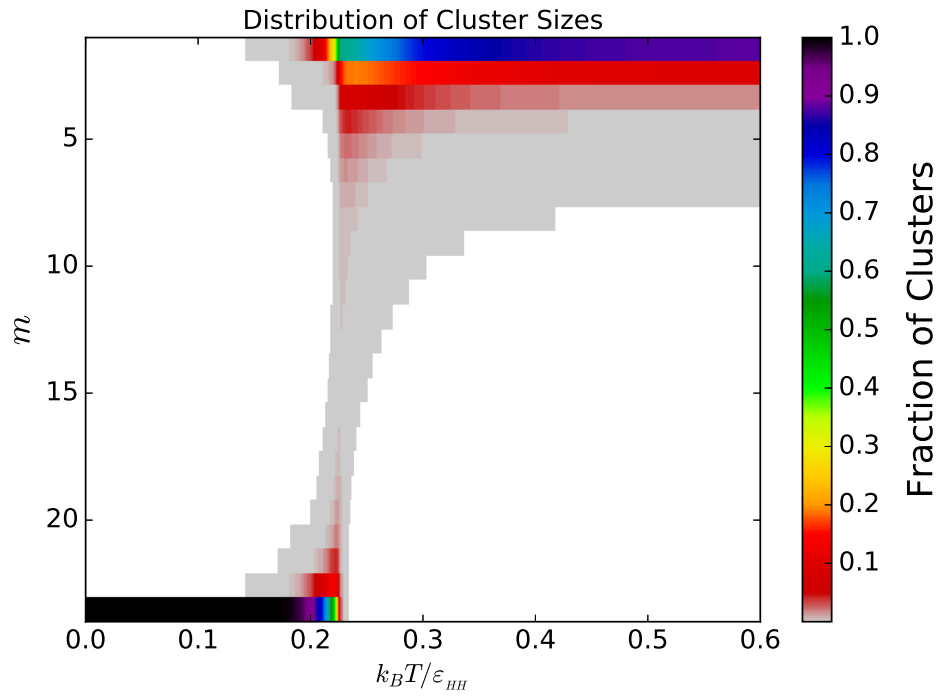

Figure 7.14: Distribution of cluster sizes for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to cluster sizes observed with a fraction $\leq 10^{-6}$.

Figure 7.15 shows the average fraction of H clusters with a given size for a range of temperatures. At higher temperatures, the system contains dispersed subunits that contain intramolecular HH contacts, where just above $k_BT/\varepsilon_{HH} \approx 0.225$, around half of the subunits have the configuration shown in the left image of Figure 7.1. When $k_BT/\varepsilon_{HH} \leq 0.225$, nearly all of the H cluster sizes $m_{HH}/N_H \leq 0.5$ occur with a fraction ≈ 0.1. Just below this temperature, the $h = 3$ protofibril structure occurs with a fraction of ≈ 0.2, until about $k_BT/\varepsilon_{HH} \approx 0.19$, where the $h = 2$ dominates.

In both cases presented here, we observe a two-step protofibril formation process involving condensation and ordering into stacked protofibril structures. With strong intermolecular coupling, a population of disordered oligomers emerges at intermediate sizes before the subunits transition directly to the ordered β-hairpin structure. For weak intermolecular coupling, there are only a small number of oligomers before the fibril ordering and stacking occurs, where a portion of the globular subunit peptides then have a configurational con-

version to adopt the ordered β-hairpin. There is an absence of intermediate oligomer sizes for the condensation step in this case. Calculated free energy differences show a nucleation-driven formation process with several metastable states containing state- and temperature-dependent intermediate barriers.

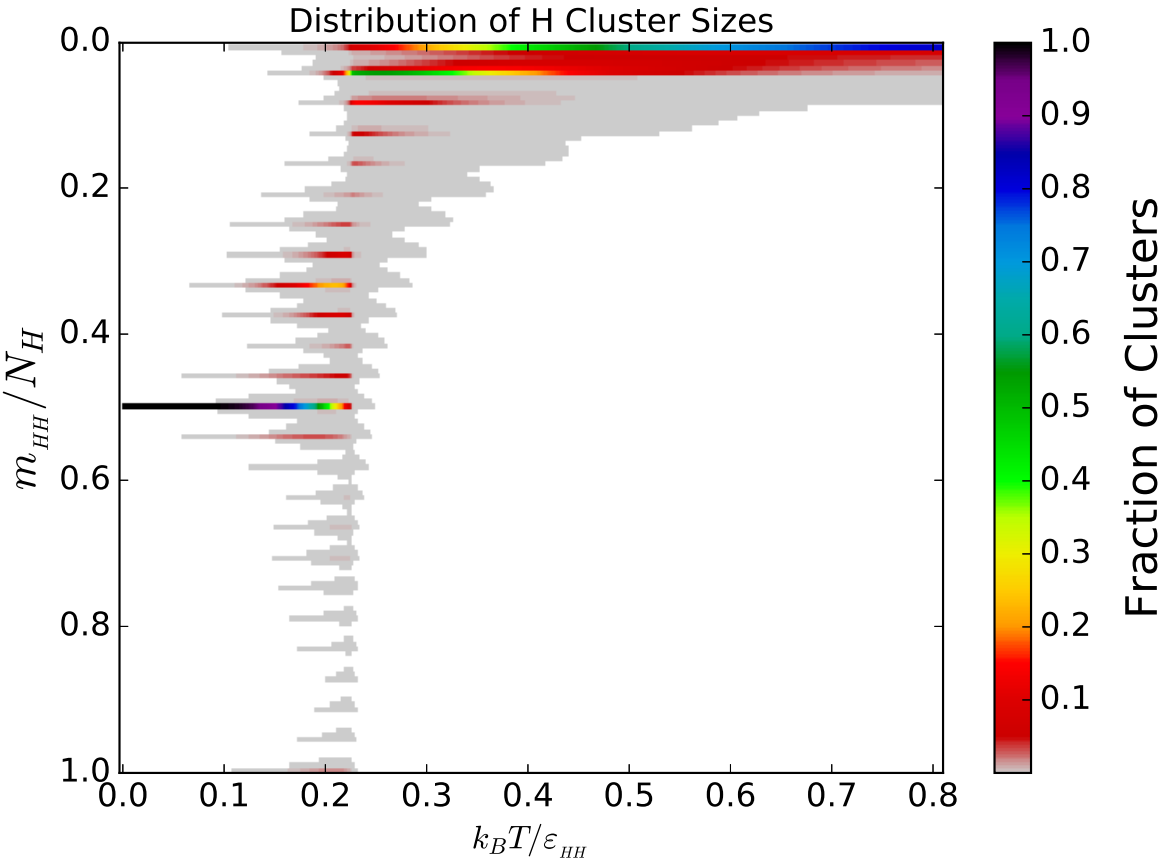

Figure 7.15: Distribution of hydrophobic cluster sizes for 24 β-hairpin subunits with weak intermolecular coupling ($\varepsilon_{HH} = 4$, $\varepsilon'_{HH} = 2$, $\varepsilon'_{PP} = 1$). White regions correspond to temperatures where clusters were not recorded for the given size.

7.2 Aggregation model on the fcc lattice

Protofibril simulations on the sc lattice have the disadvantage of heavy geometrical constraints on the representation of subunit structures, as well as the parity problem that limits the possibility of interactions. As an apparent example, subunits that have a β-strand structure with a repeating sequence of H and P residues, such as 0HPHPHPH0, cannot form intramolecular contacts on the sc lattice due to the parity problem. The fcc lattice does not have this issue, and offers twice as many bond angles as the sc lattice, and should therefore be able to accommodate a wider variety of intermediate oligomer structures than the sc lattice. Figure 7.16 shows an example of β-strand subunits on the sc and fcc lattices,

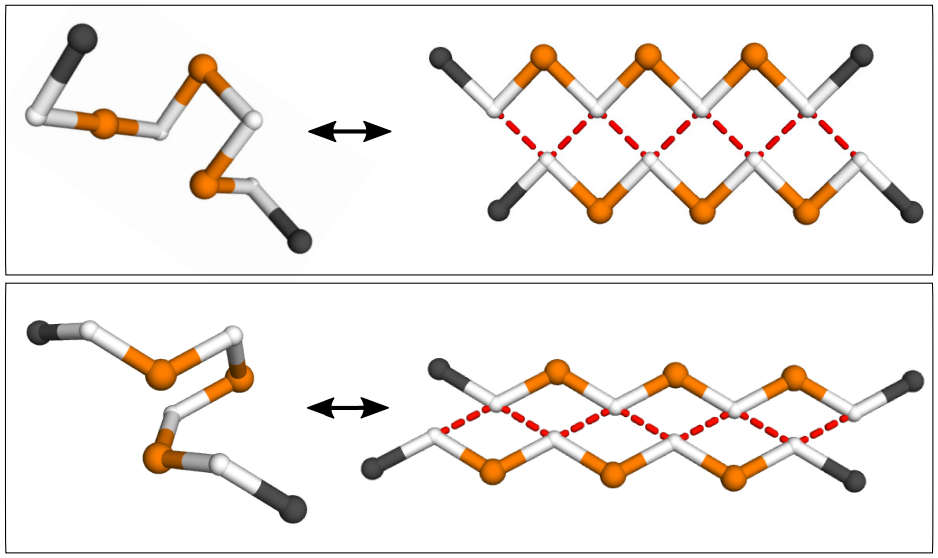

Figure 7.16: The β-strand subunit on the sc (top) and fcc (bottom) lattices. The left images show disordered, non-interacting subunits, and the right images show pairs of subunits interacting as β-strands through intermolecular HH contacts. Dashed, red lines show intermolecular HH contacts.

where images on the left are disordered subunits, and images on the right are two subunits that interact with intermolecular HH contacts. While not included in this figure, the fcc version of β-strand subunits are able to form intramolecular HH contacts. We implement the same aggregation model, now using the fcc lattice, as a testing ground for the simulation of protofibrils formed with β-strand peptide subunits.

As a simple comparison for the effect of lattice geometry on the aggregation of β-strand peptides, we simulate a system of 16 subunits on the fcc and sc lattices, with $\ell = 9$ and the sequence: 0HPHPHPH0. Because the sc lattice cannot incorporate intramolecular contacts, the energetic couplings are restricted to $\varepsilon'_{_{HH}} = 1$. The resulting ground state structure is a two-layer protofibril with H residues buried in the middle, and two surfaces composed of P residues, whose cross section is shown in the top-right image of Figure 7.16. For the fcc lattice, the intermolecular coupling is also set as $\varepsilon'_{_{HH}} = 1$, and angle energy penalties $\varepsilon_{(\theta=60^\circ)} = \varepsilon_{(\theta=180^\circ)} = -1$ are included to restrain sampling for a set of two relevant bond angles ($\theta = 90^\circ, 120^\circ$). These additional angle restraints are incorporated, as the $\ln[\hat{g}(E)]$ is extremely challenging to converge with the existing methodology when considering all angles. To obtain a fibrillar ground state structure on the fcc lattice, the β-strand energy from Section 3.3 is included with $\varepsilon_{_\beta} = 1$. These parameters for the fcc lattice simulation result in a protofbril with the cross section shown in the bottom-right of Figure 7.16.

Shown in Figure 7.17, the specific heat is compared for protofibril formation of 16 β-strand subunits on the fcc (black curve) and sc (red curve) lattices, with the resulting ground state structures shown below the x-axis. Reduced temperatures are normalized by the respective coordination number ($\mathbb{c}_n$) for each lattice geometry. The $C_V/(N\ell)$ for both lattices show two distinct maxima, which correspond to the condensation transition and fibrillization transitions. At $k_BT/(\mathbb{c}_n\varepsilon'_{_{HH}}) \approx 0.068$, the condensation transition temperature, where subunits begin to form disordered oligomers, agree quite well for the two lattices. For the fcc lattice, the fibrillization transition, where an ordered protofibril structure forms, occurs at a higher temperature than the sc lattice. One possible explanation for this difference is that each β-subunit on the fcc lattice can form intermolecular HH contacts with three subunits from the layer below(above), whereas the sc lattice only allows contacts between layers with the neighboring subunit directly above(below).

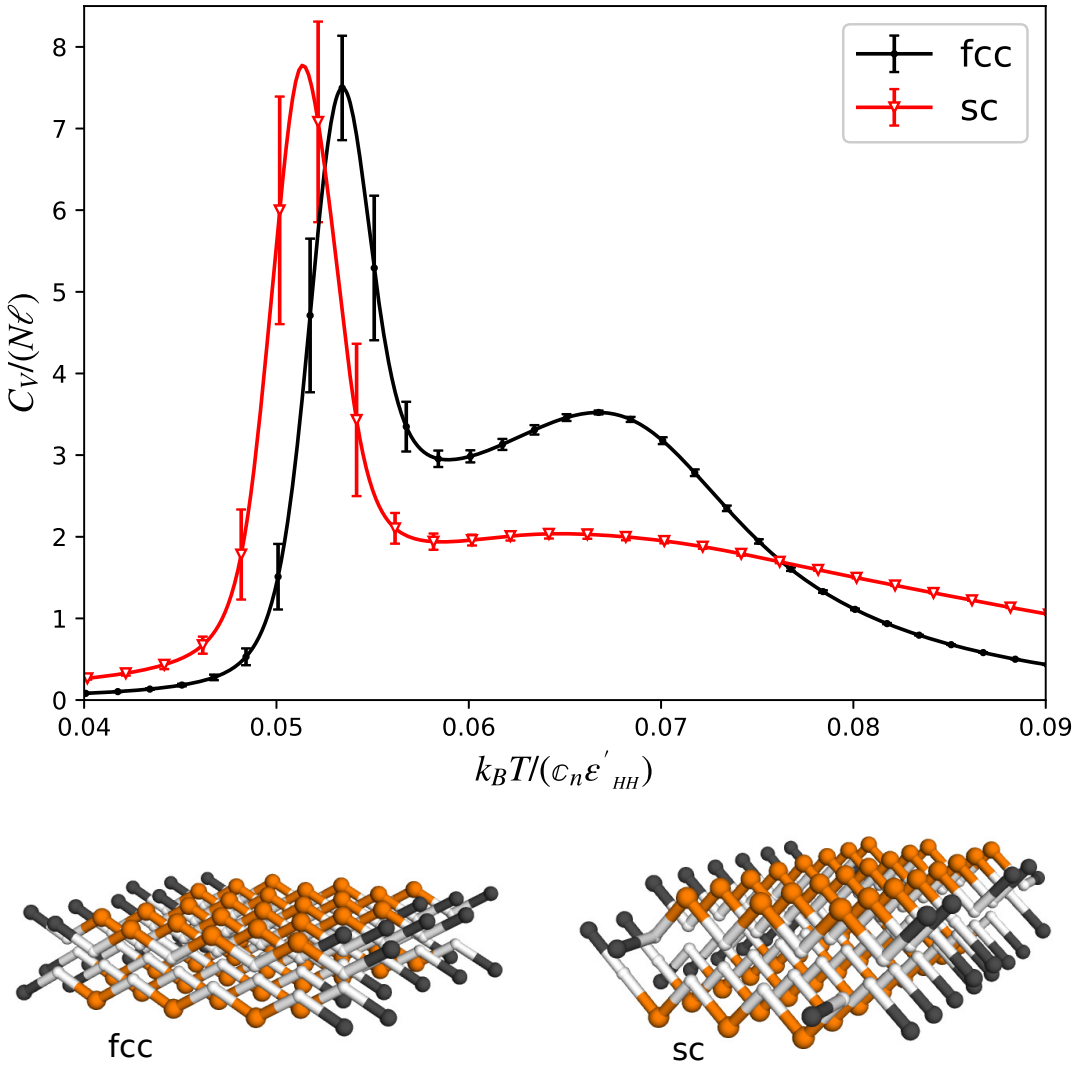

Figure 7.17: Comparison of the specific heat for fcc (black circles) and sc (red triangles) lattice aggregation simulations with 16 β-strand subunits. Couplings are set to $\varepsilon'_{HH} = 1$ for both lattices. The fcc lattice simulation has additional β-sheet energies $\varepsilon_\beta = 1$ and angle penalties $\varepsilon_{(\theta=60°)} = \varepsilon_{(\theta=180°)} = -1$. Ground state configurations are shown below the x-axis for the fcc (left) and sc (right) lattices.

Using REWL, the energy states and corresponding aggregate structures are predicted for a system of 32 β-strand subunits on fcc lattice with the energetic couplings $\varepsilon'_{HH} = 1$, $\varepsilon_{\beta} = 1$, and $\varepsilon_{(\theta=60^{\circ})} = \varepsilon_{(\theta=180^{\circ})} = -1$. Representative configurations are shown in Figure 7.18 with arrows pointing from high to low energy states. The first two images at the top-right show a pair of coalescing droplets that merge to form a single disordered oligomer. The next three images in the series of states show protofibrils that have a pore-like structure, with progressively increasing β-sheet content. Finally, the structure with the lowest found energy is a two-layer β-sheet fibril with $E_{min} = -515$. This is not the ground state, but has a couple of subunits that are attached between the two β-sheets and parallel to the fibril axis.

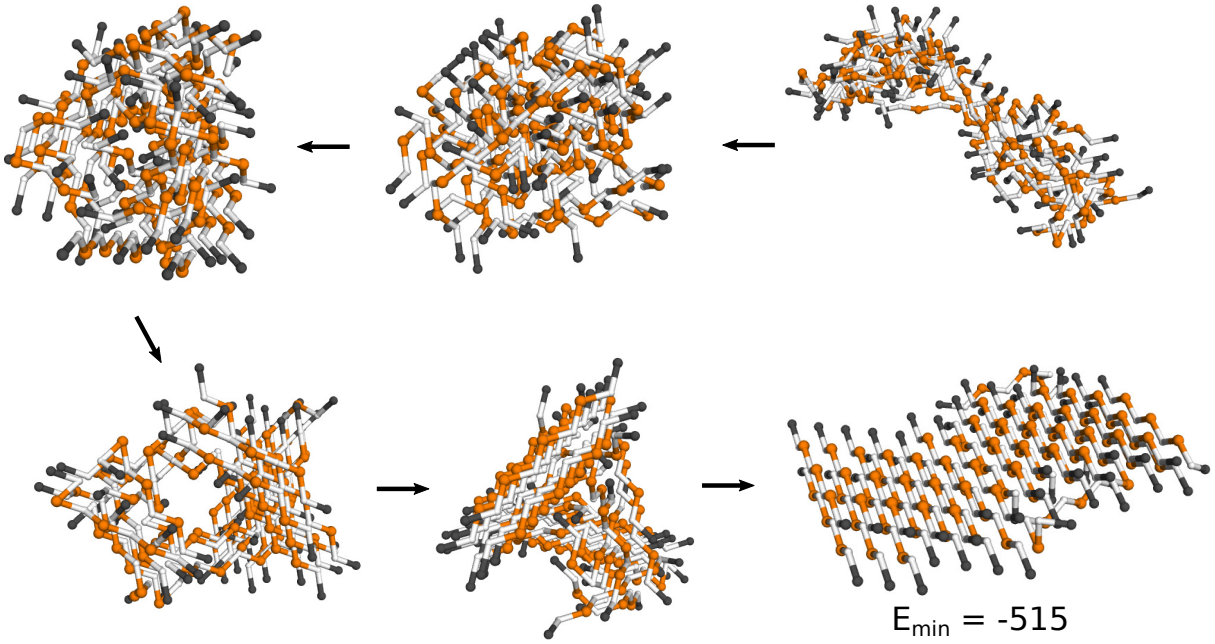

Figure 7.18: Series of aggregate states for 32 β-hairpin subunits, where black arrows point in the direction of decreasing energy, and the lowest energy structure found is shown at the bottom-right. Couplings are set to $\varepsilon'_{HH} = 1$, $\varepsilon_{\beta} = 1$, and $\varepsilon_{(\theta=60^{\circ})} = \varepsilon_{(\theta=180^{\circ})} = -1$.

As an attempt to simulate stacked protofibril structures, a system of 32 β-strand subunits on fcc lattice with the energetic couplings $\varepsilon'_{HH} = 1$, $\varepsilon'_{PP} = 0.1$, $\varepsilon_{\beta} = 1$, and $\varepsilon_{(\theta=60^\circ)} = \varepsilon_{(\theta=180^\circ)} = -1$. Energy states and corresponding aggregate structures are again predicted using REWL, and representative configurations are shown in Figure 7.19 with arrows pointing from high to low energy states. With the introduction of a weak coupling ε'_{PP}, rather than stacked protofibrils with clearly defined layers, we observe a prevalence of β-strand and β-sheet bundles that have no elongated direction. An interesting oligomer structure common in this simulation is the 'cross-hatched' bundle of β-strands, where layers of β-sheets are stacked with subunits oriented perpendicular to each previous layer.

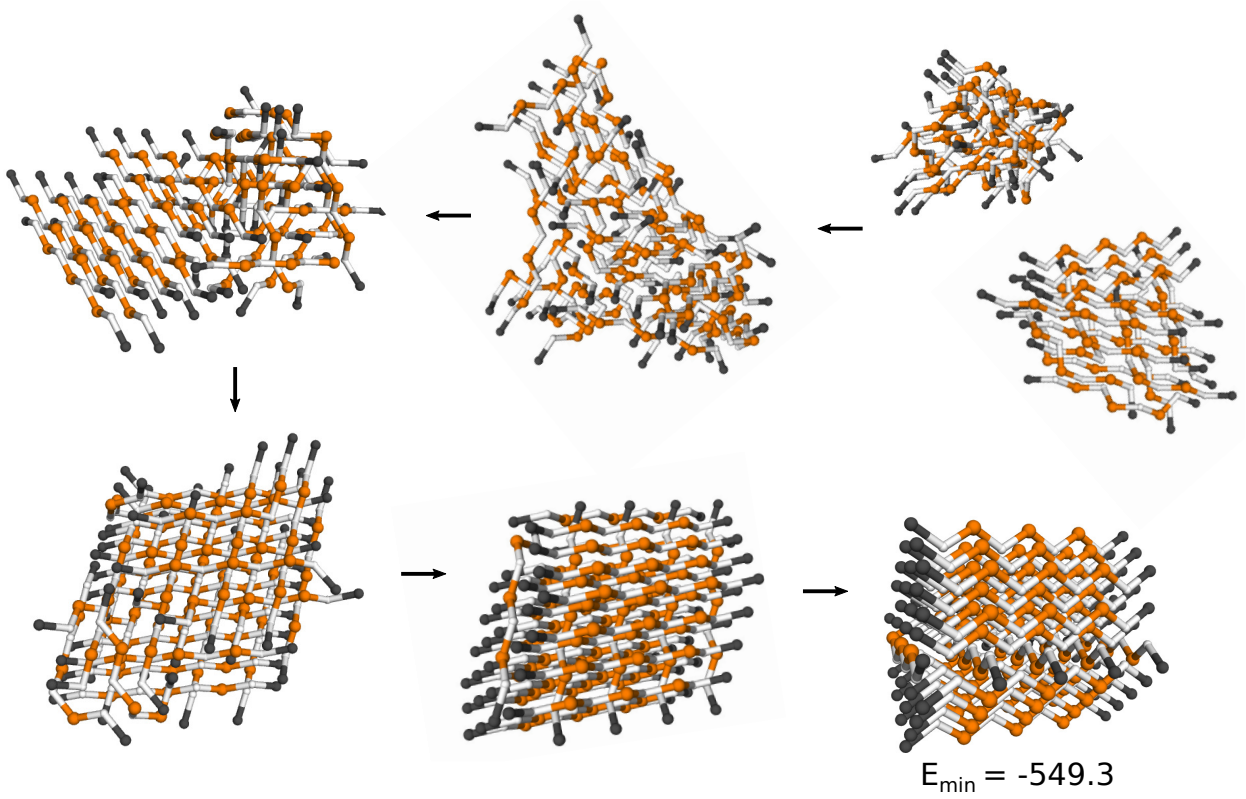

Figure 7.19: Series of aggregate states for 32 β-hairpin subunits, where black arrows point in the direction of decreasing energy, and the lowest energy structure found is shown at the bottom-right. Couplings are set to $\varepsilon'_{HH} = 1$, $\varepsilon'_{PP} = 0.1$, $\varepsilon_{\beta} = 1$, and $\varepsilon_{(\theta=60^\circ)} = \varepsilon_{(\theta=180^\circ)} = -1$.

The top-right image shows a state with two clusters; one is a disordered droplet, and the other has cross-hatched β-strands. At lower energies, a single disordered oligomer forms, which then develops some partial β-sheet content. Shown in the two bottom-left images, the system forms multiple cross-hatched β-sheet layers before rearranging to a bundle of aligned β-strands. With $E_{min} = -549.3$, the structure at the lowest found energy is a bundle of aligned β-strands in the shape of a parallelepiped, which is bisected by one β-sheet perpendicular to the bundle.

While the fcc lattice model for aggregation and protofibril formation of short H0P peptides shows a similar condensation transition behavior to the sc lattice, the intermediate oligomers and ordered states are more complex and, admittedly, difficult to sample with the presented methodology and interaction schemes. Further investigation is needed with the fcc system to tune REWL for the convergence of $\ln[\hat{g}(E)]$ and perform a detailed thermodynamic analysis, where additional considerations, such as more advanced MC trial moves [126] or an alternative statistical ensemble [127], may be helpful.

Chapter 8

Conclusions

In this work, we applied advanced Monte Carlo (MC) methodology, namely the replica-exchange Wang-Landau (REWL) and multicanonical sampling algorithms, to study protein folding and peptide aggregation with coarse-grained lattice models. We make a special focus on protein models that use the face-centered cubic (fcc) lattice, which offer a geometry more closely resembling real protein backbone structures when compared to the popular simple cubic (sc) lattice models. Application of the Wang-Landau methodology with an unbiased set of MC trial moves, which has proven successful in protein folding studies with sc lattice models, is extended to fcc lattice and used to examine average thermodynamic properties of folding for biologically inspired models. More specifically, the hydrophobic-polar (HP) model is used, and a direct comparison for the simulated folding behavior is made between the fcc version and existing benchmark results for the sc lattice. The capability of fcc lattice backbone models to represent the biological α-helix structure is investigated by designing a minimal scheme involving a four-body torsion term and simple, distance-dependent contact interactions. Using this scheme with the interacting self-avoiding walk (ISAW) model, an example set of parameters is determined that results in the formation of helix bundle structures. The α-helix motif is then tested in the context of protein folding by applying the interaction scheme to the hydrophobic-neutral-polar (H0P) model. Finally,

we extend the H0P lattice simulations to construct a model for peptide aggregation that leads to amyloid protofibril structures. With the sc lattice model, the thermodynamics of protofibril formation are studied using the specific heat, average cluster size distributions, and nucleation free energy barriers. Presented simulations focus on the case of stacked protofibril structures comprised of β-hairpin subunits. The model for aggregating H0P peptides is implemented for the fcc lattice and a brief comparison is made with the sc lattice for β-strand subunits. Protofibril structures are predicted for the fcc lattice model.

Our REWL simulations are successfully able to predict structures and energy values for the ground states of HP model mappings for real proteins and protein fragments. These predicted energies are lower than a previous MC study [103] using alternative methodologies, and have been verified as optimal using a popular constraint programming web server for lattice models [99–102]. Thermodynamic results for the HP model with the fcc lattice show that the overall folding behavior is comparable with that observed for the sc lattice, with the exception of HP sequences that are designed for low-degeneracy ground states on the sc lattice. For the two longest HP sequences studied here, average quantities like relative shape anisotropies and end-to-end distance show that while there may be some subtle rearrangement for HP124, this is not found to be the case for HP136.

Simulations that include α-helix motifs show the capability of forming ground state structures with one-, two-, three-, and four-helix bundles, depending on what strength of torsion energy penalty is used. In the 46-residue fcc ISAW model, the relative shape anisotropy shows consecutive structural transitions from random coil states, globular helical states, and four-helix bundles are observed for a range of torsion penalties. Higher values of torsion penalties show consecutive structural transitions between random coil states, three-helix-bundles, and two-helix bundles. Special values of the torsion penalty strength result in ground states with three-helix bundles, and a single helix state. When applying the α-helix motif to a H0P model mapping of the 46-residue Crambin protein, we find that the structural

similarity with experimental results is improved by $\sim 30\%$ when compared to the standard H0P model on the fcc lattice alone.

For sc lattice model simulations of peptide aggregation, we observe a two-step protofibril formation process involving condensation and ordering into stacked protofibril structures. With equal intramolecular and intermolecular hydrophobic interactions, a population of disordered oligomers emerges at intermediate sizes before the formation of an ordered fibrillar state. In this case, the peptide subunits transition from disordered, aggregated states directly to the ordered β-hairpin structure. When there is a weak intermolecular hydrophobic interaction, the subunit peptides form individual, globular states, and the number of oligomers remains small before the fibril ordering and stacking occurs. The condensation and protofibril ordering happen near the same temperature in this case, where a portion of the globular subunit peptides have a configurational conversion to adopt the ordered β-hairpin. Calculation of free energy differences for the two investigated cases of stacked protofibrils show a nucleation-driven formation process.

The peptide aggregation model extended to the fcc lattice shows a similar condensation-ordering protofibril formation process with β-strand peptide subunits when compared to the sc lattice case. A slightly higher ordering temperature is observed for the fcc lattice, which could arise from the availability of more possible intermolecular contacts between fibril sheets under the fcc geometry. The fcc lattice shows a variety of intermediate aggregate configurations, which include pore-like oligomers and cross-hatched β-sheet stacking. We find that the addition of even weak intermolecular polar interactions results in the formation of bundled aggregates rather than elongated fibrillar structures. While the aggregation model with the fcc lattice shows promise, a more powerful set of MC trial moves targeted at multi-sequence simulations is needed to perform an in-depth average structural analysis. The applied MC methodology requires typical values of $\sim$ 5,000 core-hours for a single simulation of the more challenging HP and H0P model simulations on the fcc lattice, and $\sim$ 2,500 core-hours for protofibril simulations with two dozen β-hairpin subunits.

In summary, the presented lattice models provide a rich testing ground for the simulation of protein folding and protofibrils, where the specific choice of lattice geometry has an impact on the represented backbone structures. While initial results are presented for α-helix motifs and protofibril models using the fcc lattice, these methods can be applied in the future to study various topics involving these features.

Appendix A

Pseudocodes and Technical Considerations

A1 Depth-first search

A depth-first search (DFS) is implemented for use with MC trial moves, calculating structural observables, and producing visualization coordinates in aggregation simulations. The method works by checking all lattice directions for the residues in each subunit, and keeping a list of subunits that have been visited until all subunits are visited. Each application of the DFS returns the indices of the subunits that are connected in a cluster. Cluster size distributions and the number of clusters are determined by successively applying DFS for each connected cluster, where the length of the search is between $O(N)$ and $O(N\ell)$ for N identical subunits of length ℓ. DFS is also used to simultaneously apply rigid-body MC moves to all subunits of a random cluster in the fcc and sc lattice aggregation simulations. Outlined below in Algorithm 2, it is important to note that this implementation (and pseudocode) uses function parameters that are passed by reference; meaning that the result is stored in the updated parameters rather than a return value. The pseudocode listing is implemented as a recursive function.

Algorithm 2 Depth-first search for identifying clusters
Input: pass by reference: ID, *count*, *visited*
Output: none: update input variables

```
1: procedure DFS(int& ID, int& count, vector<bool>& visited)
2:
3:     if ( visited[ID] ) then
4:         return
5:     end if
6:
7:     visited[ID] = true
8:
9:     if ( ++count > N ) then
10:        return
11:    end if
12:
13:    s = subunits[ID]
14:
15:    for i ∈ [1, ..., s → ℓ] do
16:        for Δr⃗ ∈ {lattice → displacement_vectors} do
17:
18:            j = lattice → query(s → r⃗_i + Δr⃗)
19:            ID_j = subunit_IDs[j]
20:
21:            if ( j == −1 or visited[ID_j] ) then
22:                continue
23:            end if
24:
25:            DFS(ID_j, count, visited)
26:
27:        end for
28:    end for
29:
30: end procedure
```

A2 Number of clusters and cluster size distribution

The following pseudocode listing shows how DFS is applied to count the number of clusters (N_C) and their sizes, which can be used to calculate the average cluster size distribution.

Algorithm 3 Count the number of clusters and their sizes
Input: none
Output: N_C, vector<int> *cluster_sizes*

```
1:  procedure COUNT_CLUSTERS( )
2:  |
3:  |   count = 0
4:  |   vector<bool> visited(N, false)
5:  |
6:  |   N_C = 0
7:  |   vector<int> cluster_sizes(N, 0)
8:  |   prev_count = 0
9:  |
10: |   for i ∈ [1, ..., N] do
11: |   |
12: |   |   if ( visited[i] ) then
13: |   |   |   continue
14: |   |   end if
15: |   |
16: |   |   DFS(i, count, visited)
17: |   |   N_C++
18: |   |
19: |   |   cluster_sizes[count − prev_count − 1]++
20: |   |   count_prev = count
21: |   |
22: |   end for
23: |
24: |   return (N_C, cluster_sizes)
25: |
26: end procedure
```

A slightly modified version is also used to identify cluster sizes ($m_{_{HH}}$) for just H residue types. In this case, the *visited* array has an entry for every H residue in the simulation instead of subunits.

A3 Coordinate shifts for visualization

During the aggregate simulations, the periodic boundary conditions are implemented in a way that uses modular arithmetic rather than restraining 3D coordinates to the simulation box. One challenge encountered when attempting to visualize configurations with this coordinate scheme is choosing at which position to apply a shift for each set of subunit coordinates, which contain multiple residues as an extended body. The workaround adopted in my code is to apply the DFS while cumulatively collective vector displacements that shift together the unrestrained coordinates for each isolated cluster.

Algorithm 4 Shift absolute coordinates to a periodic cell for visualization
Input: none
Output: none: update member variable vector<vector<T> > *shift*

```
1:  procedure SHIFT_COORDS( )
2:  |
3:  |   count = 0
4:  |   vector<bool> visited(N, false)
5:  |
6:  |   for i ∈ [1, ..., N] do
7:  |   |
8:  |   |   if ( visited[i] ) then
9:  |   |   |   continue
10: |   |   end if
11: |   |
12: |   |   c⃗_0 = subunits[i] ↠ r⃗_0
13: |   |   shift[i] = lattice ↠ PBC(c⃗_0) − c⃗_0
14: |   |
15: |   |   DFS(i, count, visited)
16: |   |
17: |   end for
18: |
19: end procedure
```

During coordinate shifts, the DFS function proceeds just as it is defined previously, but the following lines are added to the code listing to accumulate displacements for the shifts. The vector $shift$ is kept as a member variable in the class that represents the aggregate, where the calculated displacements are stored in the updated $shift$ as DFS executes.

---

procedure DFS(int& ID, int& $count$, vector<bool>& $visited$)

⋮

23: ...

$n = subunits[ID_j]$

$j \mathrel{-}= n \twoheadrightarrow start\_index$

$shift[ID_j] = shift[ID] + (s \twoheadrightarrow \vec{r}_i + \Delta\vec{r}) - (n \twoheadrightarrow \vec{r}_j)$

25: ...

⋮

end procedure

A4 Internal bond-rebridging moves for the fcc lattice

Internal (acting on residues in $[2, \ell - 1]$) bond-rebridging moves for backbone models on the fcc lattice closely follow the original procedures from Deutsch [80] (see also [81]), and involve moves of type 1 and type 2. Diagrams of the moves are given in Fig. A1, where type 2 moves result in a linear topology after one pair of bonds are deleted and re-assigned, and type 1 moves result in a disconnected loop that requires two successive pairs of bonds to be deleted and re-assigned before a valid configuration is found. The steps for internal rebridging moves on the fcc lattice are:

(1) Choose a random residue n, excluding end points.

(2) Choose a random lattice direction that is different from the two backbone bonds of n.

(3) If no residue j exists in the chosen direction, or if $|n - j| < 3$, the move is not possible and fails.

(4) Check if residue $n + 1$ is adjacent to residue $j - 1$ (type 1), residue $j + 1$ (type 2), or both. Otherwise, the move fails.

(5) Randomly choose between type 1 and type 2 moves if both are valid.

(6) If a type 1 move is chosen, then steps (1-4) are repeated on the disconnected loop (using variables n_2 and j_2) to recover a linear topology.

(7) Relabel the sequence information for the new configuration: for type 2, this is simply reversing the segment between $[j, n + 1]$, but extra steps must be taken for relabeling type 1 moves.

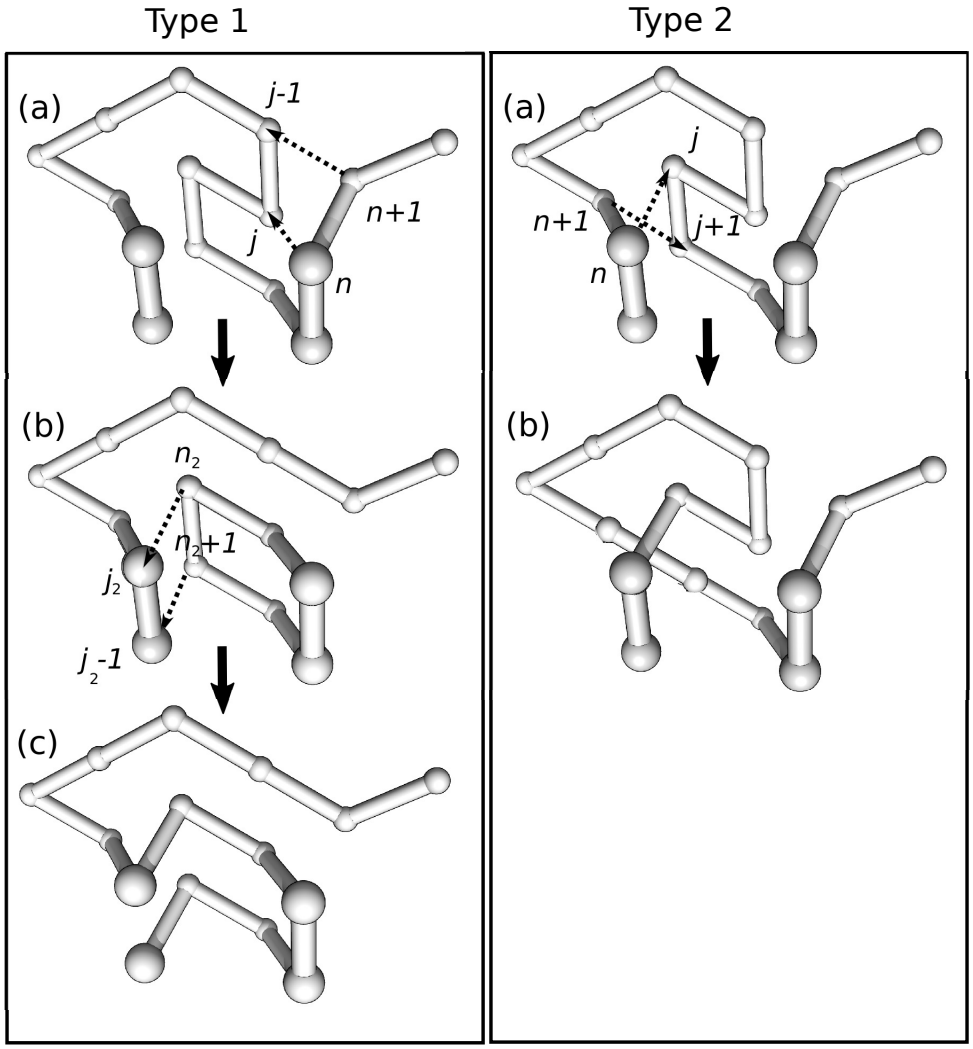

Figure A1: Two types of rebridging moves. The example configuration has residues arranged in two parallel planes, where the four larger, slightly shaded residues are in plane and in the foreground. The new (broken) bonds need not be parallel, as shown in steps (a). For clarity, the images show configurations without a sequence; otherwise, a trivial relabeling would be performed.

A5 Tree data structure for saving configurations

With so many valid backbone structures that are possible, the identification and storage of unique configurations becomes very expensive, both computationally and in terms of memory usage. We employ the following data structure to store configurations at energy states of interest for later visualization, calculations, and also to infrequently refresh replicas that have become 'stuck', or spend long times in one region of state space during a simulation. An efficient way of storing and counting unique configurations is through the use of a tree data structure. This technique can be applied to any lattice geometry (computer memory permitting), but we describe here the usage for a fcc lattice (for a description for sc see [128] [129]). In the tree structure, each node represents one of the 12 possible lattice directions (11 for residues $\in [2, \ell - 1]$) that bonds can take between residues in the chain, meaning the tree has a branching factor of $k = 11$. The tree structure is instantiated using a structure where the first three non-planar bonds $\vec{b}^{(0)}, \vec{b}^{(1)}, \vec{b}^{(2)}$ define a reference coordinate frame. When a configuration is queried or inserted in the tree structure, it is aligned with the reference coordinate frame, with care taken to eliminate all rotational symmetries. After the symmetry reduction is performed, the $\ell - 2$ bonds (omitting the first bond $\vec{b}_1$ which is identical for all configurations) are assigned characters ($\underline{\mathscr{a}}, \underline{\mathscr{b}}, \ldots, \underline{\mathscr{k}}, \underline{\mathscr{l}}$) corresponding to the 12 possible directions. The tree is then traversed, and if one of the bond characters results in a new node, then the remainder of the sequence is entered as a new branch. If the full tree is explored without a new node being created, then the configuration is not unique, and the tree structure is not altered. This scheme has an $\mathcal{O}(\ell)$ lookup complexity and assures that the minimal amount of memory is used to store all of the configurations. A schematic diagram of the entry of new structures into the tree is shown in Figure A2, where the left-most image shows the tree before the new structure is added and the right-most image shows the tree after. For the hypothetical example in the images, blue arrows represent bond vectors that are included in previously found configurations, and the red arrows represent bond vectors that distinguish the new structure and are added as a new branch in tree.

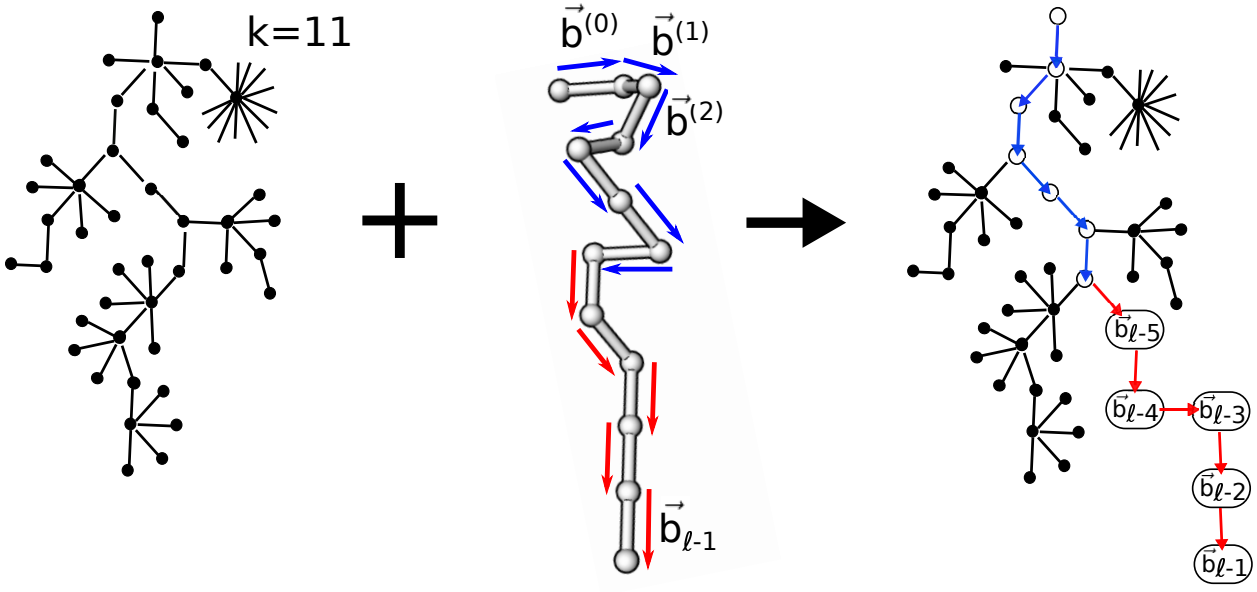

Figure A2: Tree data structure for storing unique configurations on the fcc lattice. The dots in the left- and right-most images are nodes that represent bond vectors in the configurations. Shown in the middle image, the added configuration has bond vectors that must be added as nodes (or a leaf) in the tree structure shown in red, and those that have been previously encountered shown in blue.

A6 Replica exchange frequency in REWL

During REWL simulations, replica exchanges take place with some fixed frequency as a two-way, blocking MPI communication between replicas in neighboring windows. Under this scheme, a constant number of exchanges are attempted within each histogram check interval, but the number of attempted exchanges during each window's complete REWL iteration can vary dramatically depending on how many histogram checks are required before the flatness criterion is satisfied. While the acceptance probability for replica exchanges shown in Equation 4.15 satisfies detailed balance, it is important to choose an exchange frequency that allows replicas to adequately explore energy states within their respective windows. Unfortunately, this consideration is system dependent, and a choice of exchange frequency that is too high has been observed to introduce systematic errors in our aggregation simulations. This effect is demonstrated with an example REWL simulation that has a replica-exchange frequency intentionally set very high.

The following example simulation was made early in the development of the codes for the sc lattice aggregation model, where a system of 8 identical H0P subunits (0HPHPHP0) were simulated with intermolecular HH interactions only (ε'_{HH}). A MC trial move set consisting of 75% pull moves, 20% pivot moves, and 5% single-subunit translations was used. A serial WL simulation with $\ln f_{final} \leq 1 \times 10^{-8}$ and $p = 0.8$ was used to obtain $\hat{g}(E)$, which is then used as a reference value $\hat{g}_0(E)$ to compare with the results from a REWL simulation. As an exaggerated example of a simulation with replica exchanges that occur too frequently, REWL was performed with 2 windows (one replica per window), where exchanges are attempted every 10 MC sweeps (640 MC moves). Figure A3 shows the reference value $\hat{g}_0(E)$ on the left, and the ratio of converged REWL results from this value $\hat{g}(E)/\hat{g}_0(E)$ in the plot on the right. The magnitude of energy is shown on the x-axes of this image, where REWL window 1 contains the ground state. A constant value for the ratio in the rightmost plot signifies that the results agree between REWL and serial WL simulations. It is immediately apparent that the $\hat{g}(E)$ for window 1 in the REWL simulation has inaccuracies in the region where the two

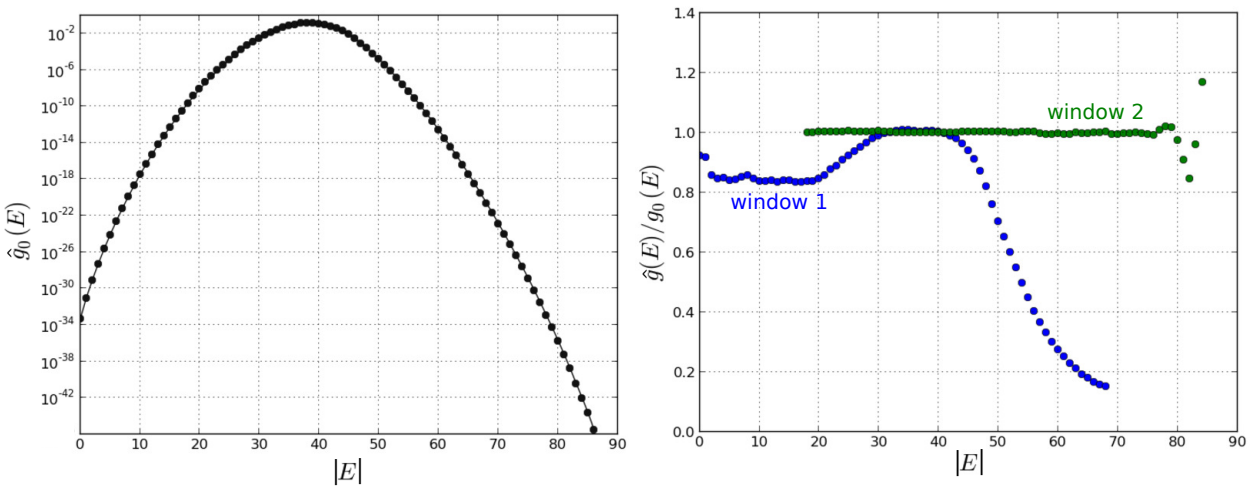

Figure A3: Reference density of states ($\hat{g}_0(E)$, left plot) and ratio of converged REWL results ($\hat{g}(E)/\hat{g}_0(E)$, right plot) from the reference values. Two windows are used in a REWL simulation where replica exchanges are attempted every 10 MC sweeps. The x-axes show the magnitude of energy, where window 2 contains the ground state energy.

windows overlap, whereas window 2 has converged towards the reference value for all energies (there are some statistical fluctuations near $|E| \leq 80$). For window 1, an overestimation of $\hat{g}(E)$ begins in the overlap region of the two windows until $|E| \approx 45$, where $\hat{g}(E)$ is dramatically underestimated. These systematic effects begin at energies corresponding to the maximum of $\hat{g}_0(E)$, which is the region of highest entropy for the system. This type of effect was also observed for similar tests for REWL simulations with 4 windows, which are not shown here.

To better understand what is occurring with regards to the replica exchanges, a two-dimensional histogram of accepted exchanges is recorded for the overlapping regions between the two REWL windows, shown in Figure A4. The x-axes show the value magnitude of energy before the replica exchanges are accepted ($|E_{before}|$), and the y-axes show the magnitude of energy after the replica exchanges are accepted ($|E_{after}|$). For every energy in the overlap region of window 1, replica exchanges can result in an $|E_{after}|$ where $\hat{g}_0(E)$ is at its maximum, as indicated by the dashed, white line. Exchanges received by window 1 that result in $|E_{after}|$ with a lower value of $\hat{g}_0(E)$ do not have a significant frequency of occurrence when $|E_{before}| \approx 40$. Successive samples that are recursively drawn from the exchange distribution

outlined by the histogram for window 1, as would be the case for a high exchange frequency where the replica cannot explore far from $|E_{after}|$, results in the replica being forced towards the energy where $\hat{g}_0(E)$. The opposite is the case in the exchange histogram for window 2; exchanges starting with $|E_{before}| \approx 40$ have equal probability to result in an $|E_{after}|$ for all available energies in the overlap region, as shown by the dashed, white line.

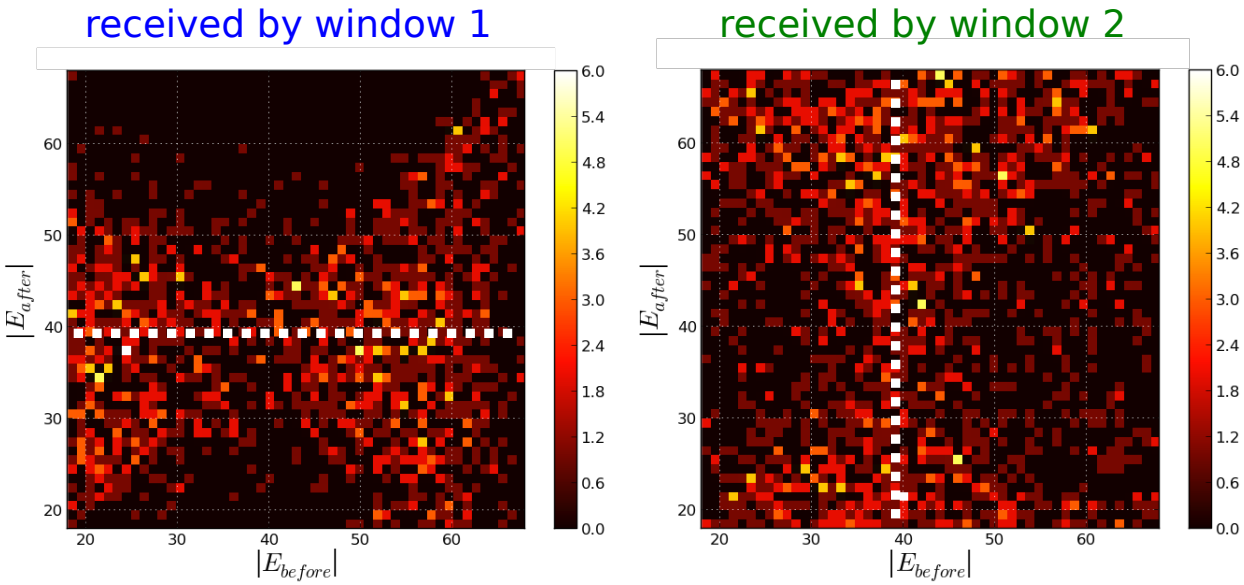

Figure A4: Two-dimensional histogram of accepted replica exchanges between the two REWL windows from the simulations in Figure A3. The right image shows accepted exchanges for window 1 during the 1st REWL iteration, and the left image shows accepted exchanges for window 2 during the 17th REWL iteration.

It should be stressed that window 2 converges much slower than window 1, as window 2 must sample low-energy states that are highly ordered. The histograms presented in Figure A4 are representative of what is observed in the first 19 iterations for window 1, and only the first 7 iterations of window 2. In practice, these effects can be ameliorated by choosing a large exchange frequency through trial and error or measurement of the round-trip times. A MUCA production run can also be performed to reweight the estimated $\hat{g}(E)$.

A7 Initialization of REWL

We use an ad-hoc initialization scheme for WL (and parallelized MUCA) replicas that can start with either a blind or informed initialization. If an informed initialization is desired, then one must already know a configuration at the start of the simulation to supply as a seed. In this case, the initial state provided is read in by one MPI rank and broadcasted to the ranks in every window. Each window where this received state is not already in bounds will proceed with a blind initialization, although ideally the walker with the initial state will be moving from low- to high-entropy states, resulting in rapid initialization. In a blind initialization, the communication scheme is chosen so that the windows wait until their neighboring window above (below) finishes initializing first (sequential order). A WL sampler with a temporary (discarded after initialization) $\hat{g}(E)$ and $H(E)$, and a modification factor fixed $f = 1$, is used to drive the replica to a state that is within the window bounds. During the initialization WL sampling, reflecting upper (lower) boundary is maintained based on the extremal value of the random walk. A user-specified 'pad', or distance in energy space, is maintained so that the reflecting boundary is decreased (increased) once the replica samples outside the padding distance from the current boundary. The purpose of the pad is to prevent a decrease of the reflecting boundary from trapping the WL sampler in state where only higher-energy states are easily accessible. This is shown visually in the schematic diagram in Figure A5, where the replica (shown as the open circle whose trajectory traces out the black line) is initializing from a high energy to a low energy (as is the common case in our simulations). The red region is off limits for the replica, according to the reflecting boundary which decreases with time (decreases signified by red, dashed arrows). Blue lines show the assigned width of the pad region, which is shaded in blue. Initialization completes once the replica finds any state that is within the window bounds, as shown by the shaded green region.

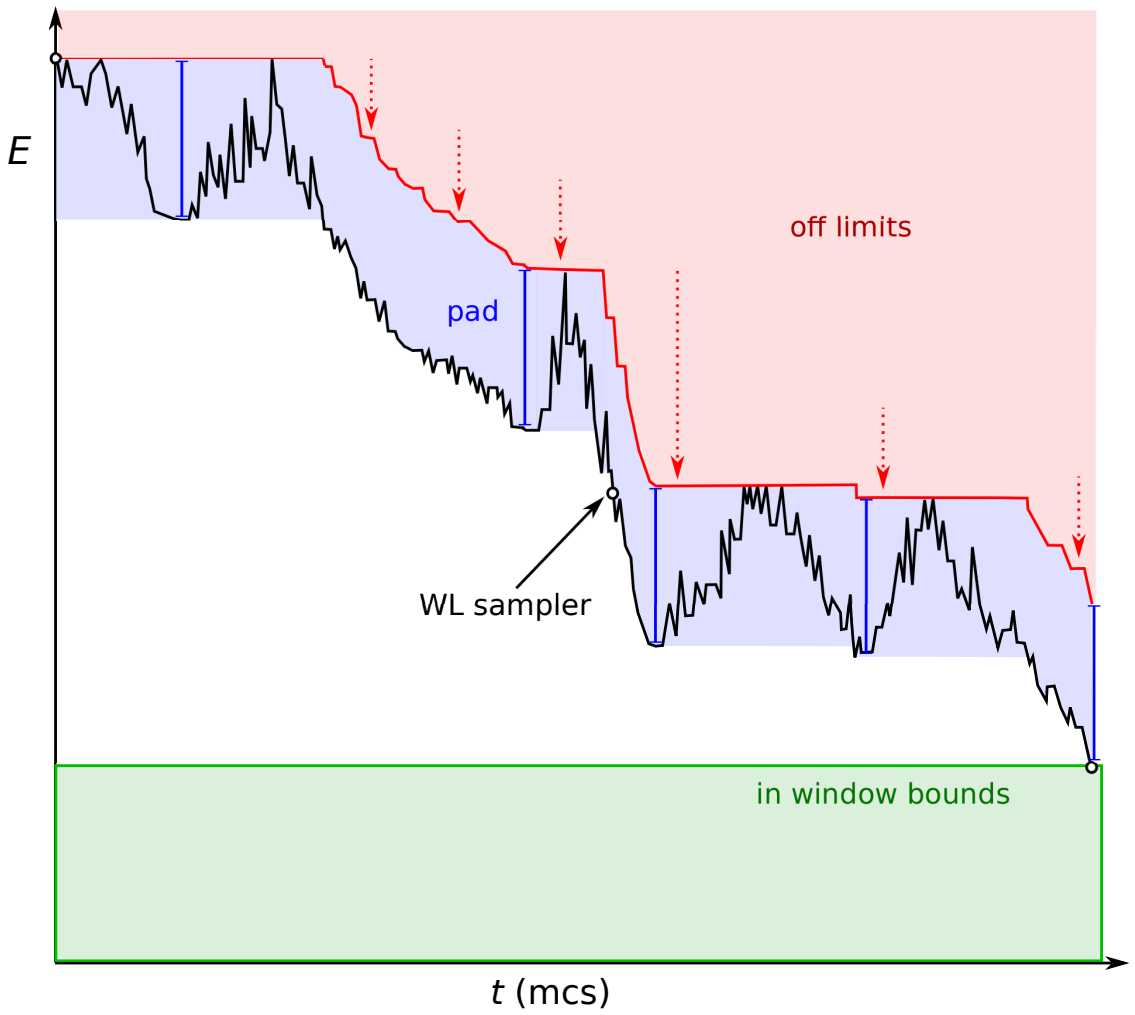

Figure A5: Schematic diagram of the initialization procedure employed for REWL and parallel MUCA simulations. The black line shows the trajectory for the temporary WL sampler, which is constrained to sample states according to some decreasing boundary (red line). An pad of some width (blue region) is assigned to prevent the sampler from being 'stuck' at a state when the boundary is decreased. The initialization completes for the replica when it reaches a state that is inside the assigned window bounds (green region).

Because a walker may get stuck as the reflecting boundaries contract, it is allowed to receive replica exchanges from the neighboring window above (below) that has previously finished initializing. This means that there are 2 windows active at a time, except when the very first window is initializing. All replicas experience a barrier and wait for the last window to initialize before starting the REWL or MUCA simulations.

Bibliography

[1] K. A. Dill and J. L. MacCallum. The Protein-Folding Problem, 50 Years On. *Science*, 338(6110):1042, 2012.

[2] J. Nasica-Labouze, P. H. Nguyen, F. Sterpone, O. Berthoumieu, N.V. Buchete, S. Coté, A. De Simone, A. J. Doig, P. Faller, A. Garcia, A. Laio, M. S. Li, S. Melchionna, N. Mousseau, Y. Mu, A. Paravastu, S. Pasquali, D. J. Rosenman, B. Strodel, B. Tarus, J. H. Viles, T. Zhang, C. Wang, and P. Derreumaux. Amyloid β Protein and Alzheimer's Disease: When Computer Simulations Complement Experimental Studies. *Chemical Reviews*, 115(9):3518, 2015.

[3] G. L. Gabor Miklos and R. Maleszka. Protein functions and biological contexts. *Proteomics*, 1(2):169, 2001.

[4] D. J. Selkoe. Folding proteins in fatal ways. *Nature*, 426(6968):900, 2003.

[5] G. R. Bowman, V. A. Voelz, and V. S. Pande. Taming the complexity of protein folding. *Current opinion in structural biology*, 21 1:4, 2011.

[6] M. Jackson and E. Hewitt. Why are Functional Amyloids Non-Toxic in Humans? *Biomolecules*, 7(4):71, 2017.

[7] C. M. Dobson. Protein misfolding, evolution and disease. *Trends in Biochemical Sciences*, 24(9):329, 1999.

[8] C. M. Dobson and M. Karplus. The fundamentals of protein folding: bringing together theory and experiment. *Current Opinion in Structural Biology*, 9(1):92, 1999.

[9] D. M. Marini, W. Hwang, D. A. Lauffenburger, S. Zhang, and R. D. Kamm. Left-Handed Helical Ribbon Intermediates in the Self-Assembly of a β-Sheet Peptide. *Nano Letters*, 2(4):295, 2002.

[10] G. Raghunathan and R. Jernigan. Ideal architecture of residue packing and its observation in protein structures. *Protein science : a publication of the Protein Society*, 6:2072, 10 2008.

[11] Z. Bagci, R. L. Jernigan, and I. Bahar. Residue packing in proteins: Uniform distribution on a coarse-grained scale. *The Journal of Chemical Physics*, 116(5):2269, 2002.

[12] M. Mann, R. Saunders, C. Smith, R. Backofen, and C. M. Deane. Producing High-Accuracy Lattice Models from Protein Atomic Coordinates Including Side Chains. *Advances in Bioinformatics*, 2012:1, 2012.

[13] M. Bachmann. *Thermodynamics and statistical mechanics of macromolecular systems*. Cambridge University Press, 2014.

[14] A.V. Efimov. Standard structures in proteins. *Progress in Biophysics and Molecular Biology*, 60(3):201, 1993.

[15] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95, 1963.

[16] C. B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science, New Series*, 181(4096):223, 1973.

[17] C. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique*, 65:44, 1968.

[18] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4(1):10, 1997.

[19] M. Karplus. Behind the folding funnel diagram. *Nature Chemical Biology*, 7(7):401, 2011.

[20] B. Caughey and P. T. Lansbury. Protofibrils, pores, fibrils, and neurodegeneration: Separating the responsible protein aggregates from the innocent bystanders. *Annual Review of Neuroscience*, 26(1):267, 2003.

[21] H. Xiong, B. L. Buckwalter, H. M. Shieh, and M. H. Hecht. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proceedings of the National Academy of Sciences*, 92(14):6349, 1995.

[22] R. Ni, S. Abeln, M. Schor, M. A. Cohen Stuart, and P. G. Bolhuis. Interplay between Folding and Assembly of Fibril-Forming Polypeptides. *Physical Review Letters*, 111(5), 2013.

[23] R. Pellarin, E. Guarnera, and A. Caflisch. Pathways and Intermediates of Amyloid Fibril Formation. *Journal of Molecular Biology*, 374(4):917, 2007.

[24] D. Kashchiev and S. Auer. Nucleation of amyloid fibrils. *The Journal of Chemical Physics*, 132(21):215101, 2010.

[25] S. Auer. Nucleation of Polymorphic Amyloid Fibrils. *Biophysical Journal*, 108(5):1176, 2015.

[26] N. Co and M. Li. Effect of Surface Roughness on Aggregation of Polypeptide Chains: A Monte Carlo Study. *Biomolecules*, 11(4):596, 2021.

[27] A. Morriss-Andrews, F. L. H. Brown, and J.E. Shea. A Coarse-Grained Model for Peptide Aggregation on a Membrane Surface. *The Journal of Physical Chemistry B*, 118(28):8420, 2014.

[28] J. Krausser, T. P. J. Knowles, and A. Šarić. Physical mechanisms of amyloid nucleation on fluid membranes. *Proceedings of the National Academy of Sciences*, 117(52), 2020.

[29] P. J Flory. *Principles of polymer chemistry*. Cornell University Press, 1953.

[30] R. D. Schram, G. T. Barkema, and R. H. Bisseling. Exact enumeration of self-avoiding walks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(06):P06019, 2011.

[31] T. Vogel, M. Bachmann, and W. Janke. Freezing and collapse of flexible polymers on regular lattices in three dimensions. *Physical Review E*, 76(6), 2007.

[32] R. D. Schram, G. T. Barkema, and H. Schiessel. On the stability of fractal globules. *The Journal of Chemical Physics*, 138(22):224901, 2013.

[33] J.C. Walter, M. Baiesi, G. T. Barkema, and E. Carlon. Unwinding Relaxation Dynamics of Polymers. *Physical Review Letters*, 110(6):068301, 2013.

[34] J. I. Siepmann and D. Frenkel. Conflgurational bias Monte Carlo: a new sampling scheme for flexible chains. *Molecular Physics*, 75(1):59, 1992.

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

[36] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986, 1989.

[37] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill. A test of lattice protein folding algorithms. *Proceedings of the National Academy of Sciences*, 92(1):325, 1995.

[38] B. Berger and T. Leighton. Protein Folding in the Hydrophobic-Hydrophilic (*HP*) Model is NP-Complete. *Journal of Computational Biology*, 5(1):27, 1998.

[39] T. C. Beutler and K. A. Dill. A fast conformational search strategy for finding low energy structures of model proteins. *Protein Science*, 5(10):2037, 1996.

[40] S. C. Kou, J. Oh, and W. H. Wong. A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling. *The Journal of Chemical Physics*, 124(24):244903, 2006.

[41] H.P. Hsu, V. Mehra, W. Nadler, and P. Grassberger. Growth-based optimization algorithm for lattice heteropolymers. *Physical Review E*, 68(2), 2003.

[42] M. Bachmann and W. Janke. Thermodynamics of lattice heteropolymers. *The Journal of Chemical Physics*, 120(14):6779, 2004.

[43] M. A. Rashid, S. Iqbal, F. Khatib, Md T. Hoque, and A. Sattar. Guided macro-mutation in a graded energy based genetic algorithm for protein structure prediction. *Computational Biology and Chemistry*, 61:162, 2016.

[44] J. Liu, G. Li, and J. Yu. Protein-folding simulations of the hydrophobic-hydrophilic model by combining pull moves with energy landscape paving. *Physical Review E*, 84(3), 2011.

[45] I. Dotu, M. Cebrian, P. Van Hentenryck, and P. Clote. On Lattice Protein Structure Prediction Revisited. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1620, 2011.

[46] G. Shi, T. Wüst, and D. P. Landau. Characterizing folding funnels with replica exchange Wang-Landau simulation of lattice proteins. *Physical Review E*, 94(5), 2016.

[47] G. Shi, T. Wüst, Y. W. Li, and D. P. Landau. Protein folding of the HOP model: A parallel Wang-Landau study. *Journal of Physics: Conference Series*, 640:012017, 2015.

[48] E. Bornberg-Bauer. Chain growth algorithms for HP-type lattice proteins. In *Proceedings of the first annual international conference on Computational molecular biology - RECOMB '97*, page 47, Santa Fe, New Mexico, United States, 1997. ACM Press.

[49] T. Hoque, M. Chetty, and A. Sattar. Extended HP Model for Protein Structure Prediction. *Journal of Computational Biology*, 16(1):85, 2009.

[50] A. C.K. Farris, G. Shi, T. Wüst, and D. P. Landau. The role of chain-stiffness in lattice protein models: A replica-exchange Wang-Landau study. *The Journal of Chemical Physics*, 149(12):125101, 2018.

[51] G. Shi, A. C.K. Farris, T. Wüst, and D. P. Landau. Folding in a semi-flexible lattice model for Crambin. *Journal of Physics: Conference Series*, 686:012001, 2016.

[52] J. Skolnick and A. Kolinski. Simulations of the Folding of a Globular Protein. *Science*, 250(4984):1121, 1990.

[53] A. Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51(2):349, 2004.

[54] C. L. Pierri, A. De Grassi, and A. Turi. Lattices for ab initio protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 73(2):351, 2008.

[55] T. C. Hales. A proof of the kepler conjecture. *Annals of Mathematics*, 162(3):1065, 2005.

[56] R. D. Schram, G. T. Barkema, R. H. Bisseling, and N. Clisby. Exact enumeration of self-avoiding walks on BCC and FCC lattices. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(8):083208, 2017.

[57] U. Bastolla and P. Grassberger. Phase Transitions of Single Semistiff Polymer Chains. *Journal of Statistical Physics*, 89(5-6):1061, 1997.

[58] D. C. Rapaport. Molecular dynamics simulation of polymer helix formation using rigid-link methods. *Physical Review E*, 66(1):011906, 2002.

[59] P. Pokarowski, A. Kolinski, and J. Skolnick. A Minimal Physically Realistic Protein-Like Lattice Model: Designing an Energy Landscape that Ensures All-Or-None Folding to a Unique Native State. *Biophysical Journal*, 84(3):1518, 2003.

[60] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105, 1982.

[61] T. Wüst and D. P. Landau. Optimized Wang-Landau sampling of lattice polymers: Ground state search and folding thermodynamics of HP model proteins. *The Journal of Chemical Physics*, 137(6):064903, 2012.

[62] K. A. Dill, K. M. Fiebig, and H. S. Chan. Cooperativity in protein-folding kinetics. *Proceedings of the National Academy of Sciences*, 90(5):1942, 1993.

[63] K. Yue and K. A. Dill. Forces of tertiary structural organization in globular proteins. *Proceedings of the National Academy of Sciences*, 92(1):146, 1995.

[64] E. E. Lattman, K. M. Fiebig, and K. A. Dill. Modeling Compact Denatured States of Proteins. *Biochemistry*, 33(20):6158, 1994.

[65] H. Arkin and W. Janke. Gyration tensor based analysis of the shapes of polymer chains in an attractive spherical cage. *The Journal of Chemical Physics*, 138(5):054904, 2013.

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

[67] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087, 1953.

[68] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[69] A. D. Swetnam and M. P. Allen. Improved simulations of lattice peptide adsorption. *Physical Chemistry Chemical Physics*, 11:2046–2055, 2009.

[70] J.J. Tsay and S.C. Su. An effective evolutionary algorithm for protein folding on 3D FCC HP model by lattice rotation and generalized move sets. *Proteome Science*, 11(Suppl 1):S19, 2013.

[71] J. Liu, B. Song, Y. Yao, Y. Xue, W. Liu, and Z. Liu. Wang-Landau sampling in face-centered-cubic hydrophobic-hydrophilic lattice model proteins. *Physical Review E*, 90(4), 2014.

[72] H.J. Böckenhauer, A. Z. M. Dayem Ullah, L. Kapsokalivas, and K. Steinhöfel. A local move set for protein folding in triangular lattice models. In Keith A. Crandall and Jens Lagergren, editors, *Algorithms in Bioinformatics*, page 369, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[73] N. Lesh, M. Mitzenmacher, and S. Whitesides. A complete and effective move set for simplified protein folding. In *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03*, page 188, Berlin, Germany, 2003. ACM Press.

[74] D. Gyorffy, P. Zavodszky, and A. Szilagyi. "Pull Moves" for Rectangular Lattice Polymer Models Are Not Fully Reversible. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(6):1847, 2012.

[75] M. A Rashid, MA H. Newton, Md T. Hoque, S. Shatabda, D. N. Pham, and A. Sattar. Spiral search: a hydrophobic-core directed local search for simplified PSP on 3D FCC lattice. *BMC Bioinformatics*, 14(S2), 2013.

[76] F. T. Wall and F. Mandel. Macromolecular dimensions obtained by an efficient monte carlo method without sample attrition. *The Journal of Chemical Physics*, 63(11):4592, 1975.

[77] J. Baschnagel, J. P. J. Wittmer, and H. Meyer. Monte carlo simulation of polymers: Coarse-grained models. *arXiv: Soft Condensed Matter*, 2004.

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

[79] N. Clisby. Efficient Implementation of the Pivot Algorithm for Self-avoiding Walks. *Journal of Statistical Physics*, 140(2):349, 2010.

[80] J. M. Deutsch. Long range moves for high density polymer simulations. *The Journal of Chemical Physics*, 106(21):8849, 1997.

[81] D. Reith and P. Virnau. Implementation and performance analysis of bridging Monte Carlo moves for off-lattice single chain polymers in globular states. *Computer Physics Communications*, 181(4):800, 2010.

[82] M. L. Mansfield. Monte Carlo studies of polymer chain dimensions in the melt. *The Journal of Chemical Physics*, 77(3):1554, 1982.

[83] R. Oberdorf, A. Ferguson, J. L. Jacobsen, and J. Kondev. Secondary structures in long compact polymers. *Physical Review E*, 74(5):051801, 2006.

[84] F. Wang and D. P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(5), 2001.

[85] F. Wang and D. P. Landau. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters*, 86(10):2050, 2001.

[86] C. Zhou and R. N. Bhatt. Understanding and improving the wang-landau algorithm. *Physical Review E*, 72:025701, 2005.

[87] A. D. Swetnam and M. P. Allen. Improving the Wang-Landau algorithm for polymers and proteins. *Journal of Computational Chemistry*, 32(5):816, 2011.

[88] R. E. Belardinelli and V. D. Pereyra. Fast algorithm to calculate density of states. *Physical Review E*, 75(4), 2007.

[89] R. E. Belardinelli and V. D. Pereyra. Wang-landau algorithm: A theoretical analysis of the saturation of the error. *The Journal of Chemical Physics*, 127(18):184105, 2007.

[90] H. K. Lee, Y. Okabe, and D.P. Landau. Convergence and refinement of the wang-landau algorithm. *Computer Physics Communications*, 175(1):36, 2006.

[91] T. Hayashi and Y. Okamoto. Efficient simulation protocol for determining the density of states: Combination of replica-exchange wang-landau method and multicanonical replica-exchange method. *Physical Review E*, 100:043304, 2019.

[92] T. Vogel, Y. W. Li, T. Wüst, and D. P. Landau. Scalable replica-exchange framework for Wang-Landau sampling. *Physical Review E*, 90(2), 2014.

[93] T. Vogel, Y. W. Li, and D. P. Landau. A practical guide to replica-exchange Wang-Landau simulations. *Journal of Physics: Conference Series*, 1012:012003, 2018.

[94] Y. W. Li, T. Vogel, T. Wüst, and D. P. Landau. A new paradigm for petascale Monte Carlo simulation: Replica exchange Wang-Landau sampling. *Journal of Physics: Conference Series*, 510:012012, 2014.

[95] K. A. Maerzke, L. Gai, P. T. Cummings, and C. McCabe. Incorporating configurational-bias Monte Carlo into the Wang-Landau algorithm for continuous molecular systems. *The Journal of Chemical Physics*, 137(20):204105, 2012.

[96] B. A. Berg and T. Neuhaus. Multicanonical algorithms for first order phase transitions. *Physics Letters B*, 267(2):249, 1991.

[97] M. Matsumoto and T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3, 1998.

[98] R. G. Miller. The jackknife–a review. *Biometrika*, 61(1):1, 1974.

[99] M. Mann, S. Will, and R. Backofen. CPSP-tools - Exact and complete algorithms for high-throughput 3D lattice protein studies. *BMC Bioinformatics*, 9(1):230, 2008.

[100] M. Mann, C. Smith, M. Rabbath, M. Edwards, S. Will, and R. Backofen. CPSP-web-tools: a server for 3D lattice protein studies. *Bioinformatics*, 25(5):676, 2009.

[101] M. Mann, R. Backofen, and Will S. Proceedings of WCB09 Workshop on Constraint Based Methods for Bioinformatics. page 43.

[102] R. Backofen and S. Will. A Constraint-Based Approach to Fast and Exact Structure Prediction in Three-Dimensional Protein Models. *Constraints*, 11(1):5, 2006.

[103] T. Vogel. HP-Proteine auf verallgemeinerten Gittern und Homopolymerkollaps. Diplom thesis, Leipzig University, 2004.

[104] T. Wüst and D. P. Landau. Versatile Approach to Access the Low Temperature Thermodynamics of Lattice Polymers and Proteins. *Physical Review Letters*, 102(17), 2009.

[105] L. Toma and S. Toma. Folding simulation of protein models on the structure-based cubo-octahedral lattice with the Contact Interactions algorithm. *Protein Science*, 8(1):196, 2008.

[106] P. Pokarowski, K. Droste, and A. Kolinski. A minimal proteinlike lattice model: An alpha-helix motif. *The Journal of Chemical Physics*, 122(21):214915, 2005.

[107] A. Dal Palú, A. Dovier, and E. Pontelli. A New Constraint Solver for 3D Lattices and Its Application to the Protein Folding Problem. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Geoff Sutcliffe, and Andrei Voronkov, editors, *Logic for Programming, Artificial Intelligence, and Reasoning*, volume 3835, page 48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[108] A. Onofrio, G. Parisi, G. Punzi, S. Todisco, M. A. Di Noia, F. Bossis, A. Turi, A. De Grassi, and C. L. Pierri. Distance-dependent hydrophobic-hydrophobic contacts in protein folding simulations. *Physical Chemistry Chemical Physics*, 16(35):18907, 2014.

[109] C. Clementi, M. Vendruscolo, A. Maritan, and E. Domany. Folding Lennard-Jones proteins by a contact potential. *Proteins: Structure, Function, and Genetics*, 37(4):544, 1999.

[110] R. Backofen, S. Will, and P. Clote. Algorithmic Approach to Quantifying the Hydrophobic Force Contribution in Protein Folding. In *Biocomputing 2000*, page 95, Honolulu, Hawaii, USA, 1999. World Scientific.

[111] M. M. Teeter. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. *Proceedings of the National Academy of Sciences*, 81(19):6014, 1984.

[112] J.J. Tsay, S.C. Su, and C.S. Yu. A Multi-Objective Approach for Protein Structure Prediction Based on an Energy Model and Backbone Angle Preferences. *International Journal of Molecular Sciences*, 16(12):15136, 2015.

[113] A. Šarić, Y. C. Chebaro, T. P. J. Knowles, and D. Frenkel. Crucial role of nonspecific interactions in amyloid nucleation. *Proceedings of the National Academy of Sciences*, 111(50):17869, 2014.

[114] A. Šarić, A. K. Buell, G. Meisl, T. C. T. Michaels, C. M. Dobson, S. Linse, T. P. J. Knowles, and D. Frenkel. Physical determinants of the self-replication of protein fibrils. *Nature Physics*, 12(9):874, 2016.

[115] J. H. M. van Gils, E. van Dijk, A. Peduzzo, A. Hofmann, N. Vettore, M. P. Schützmann, G. Groth, H. Mouhib, D. E. Otzen, A. K. Buell, and S. Abeln. The hydrophobic effect characterises the thermodynamic signature of amyloid fibril growth. *PLOS Computational Biology*, 16(5):e1007767, 2020.

[116] D. Bratko, T. Cellmer, J. M. Prausnitz, and H. W. Blanch. Molecular simulation of protein aggregation. *Biotechnology and Bioengineering*, 96(1):1, 2007.

[117] M. S. Li, D. K. Klimov, J. E. Straub, and D. Thirumalai. Probing the mechanisms of fibril formation using lattice models. *The Journal of Chemical Physics*, 129(17):175101, 2008.

[118] M. T. Oakley, J. M. Garibaldi, and J. D. Hirst. Lattice models of peptide aggregation: Evaluation of conformational search algorithms. *Journal of Computational Chemistry*, 26(15):1638, 2005.

[119] C. Junghans, M. Bachmann, and W. Janke. Microcanonical Analyses of Peptide Aggregation Processes. *Physical Review Letters*, 97(21), 2006.

[120] G. Bellesia and J.E. Shea. Self-assembly of β-sheet forming peptides into chiral fibrillar aggregates. *The Journal of Chemical Physics*, 126(24):245104, 2007.

[121] G. Bellesia and J.E. Shea. Effect of β-sheet propensity on peptide aggregation. *The Journal of Chemical Physics*, 130(14):145103, 2009.

[122] J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, and A. Trovato. Unified perspective on proteins: A physics approach. *Physical Review E*, 70(4), 2004.

[123] N. B. Hung, D.M. Le, and T. X. Hoang. Sequence dependent aggregation of peptides and fibril formation. *The Journal of Chemical Physics*, 147(10):105102, 2017.

[124] S. Auer, C. M. Dobson, and M. Vendruscolo. Characterization of the nucleation barriers for protein aggregation and amyloid formation. *HFSP Journal*, 1(2):137, 2007.

[125] D. J. Rosenman, C. R. Connors, W. Chen, C. Wang, and A. E. Garcia. Aβ monomers transiently sample oligomer and fibril-like configurations: Ensemble characterization using a combined md/nmr approach. *Journal of Molecular Biology*, 425(18):3338, 2013.

[126] S. Whitelam and P. L. Geissler. Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles. *The Journal of Chemical Physics*, 127(15):154101, 2007.

[127] C. Desgranges and J. Delhommelle. Evaluation of the grand-canonical partition function using expanded wang-landau simulations. i. thermodynamic properties in the bulk and at the liquid-vapor phase boundary. *The Journal of Chemical Physics*, 136(18):184107, 2012.

[128] G. Shi, T. Vogel, T. Wüst, Y. W. Li, and D. P. Landau. Effect of single-site mutations on hydrophobic-polar lattice proteins. *Physical Review E*, 90:033307, 2014.

[129] K. Yue and K. A. Dill. Sequence-structure relationships in proteins and copolymers. *Physical Review E*, 48(3):2267, 1993.