ABSTRACT

NGUYEN, VAN THANH TRAN. Development of User-friendly DMD/PRIME20 Software Package and Its Application to Investigate Aggregation of α-Synuclein N-terminal Fragments. (Under the Direction of Dr. Carol K. Hall and Dr. Erik E. Santiso).

The misfolding and aggregation of specific native proteins in neuronal systems, which lead to their accumulation in amyloid plaques, causes various neurodegenerative diseases affecting millions of people worldwide. Scientists have been investigating amyloidosis using multiples approaches to gain insights into its causes and potential treatments. Computer simulations have proven to be essential tools for studying peptide aggregation at the molecular scale. In this dissertation, we present an open-source simulation package designed to simulate and evaluate the spontaneous self-assembly processes of short peptides starting from random coils. Additionally, we demonstrate the application of our software to advance understanding of the in vitro aggregation of N-terminal α-Synuclein fragments and their single-point mutations, as well as to assist in discovery of new peptides that can self-assemble into pre-determined fibrillar structures.

We developed a user-friendly version of the DMD/PRIME20 simulation package that is open for public use. The software utilized the discontinuous molecular dynamics (DMD) simulation method in conjunction with a coarse-grained (CG) force field to improve computational efficiency. The model has been used internally by the Hall group for twenty years to study the spontaneous self-assembly of various peptides associated with neurodegenerative diseases. Here, we release a new version of DMD/PRIME20, which was designed to be accessible and easy to operate, even for users with no computational experience. This version of DMD/PRIME20 allows users to have better control over the simulations, such as allowing for user-defined annealing temperatures and easy simulation replication or restarting. The package also includes an analysis module, which allows users to easily evaluate and visualize simulation results.

During the development and modification of DMD/PRIME20, we applied the software to study short peptide fragments from the N-terminal of α-Synuclein (αSyn). This protein is associated with various synucleinopathies such as Parkinson's disease, the second most common neurodegenerative disease worldwide. Inspired by the research from the group of Sheena Radford, we conducted simulations of the aggregation of four peptide fragments (P1, P2, P3, and P3Next) and three single-point mutations (L38M, L38A, and V40A); the mutation L38M is known to inhibit fibrillation in experiments on full-length αSyn. Our findings indicate that wild-type fragments tend to form fibrillation-prone structures, specifically parallel β-sheets and β-hairpins. We proposed an inhibition mechanism for the L38M mutation, suggesting that it prevents the formation of parallel β-sheets and β-hairpins. Additionally, L38M induces intra- and inter-molecular interactions with the C-terminal domain, thereby inhibiting the transition from oligomeric to fibrillar structures.

We next use DMD/PRIME20 simulations to uncover the kinetic fibril formation pathway for peptides that were designed using the Hall group's PepAD algorithm. The PepAD software was developed by Dr. Sudeep Sarma to design peptides that form amyloid-like structures. Our goal was to design peptides that form parallel amyloid-like structures, inspired by the Class 1 cross-β spine described by Sawaya et al. Using DMD/PRIME20, we refined and selected the best designed peptides to be used as prototypes for the subsequent round of sequence refinement. We finally identified eight 7-mer peptides that spontaneously form parallel amyloid fibrils in DMD/PRIME20 simulations, achieving a Class 1 β-sheet structure. Five out of eight of the designed peptides formed amyloid fibrils composed of parallel β-sheets in Fourier Transform Infrared Spectroscopy, Circular Dichroism, Electron Microscopy, and Thioflavin-T fluorescence spectroscopy experiments.

Development of User-friendly DMD/PRIME20 Software Package and Its Application to Investigate Aggregation of α-Synuclein N-terminal Fragments

by
Van Thanh Tran Nguyen

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Chemical Engineering

Raleigh, North Carolina
2024

APPROVED BY:

_____	_____
Dr. Carol K. Hall Committee Chair	Dr. Jason M. Haugh
_____	_____
Dr. Erik E. Santiso Committee Co-chair	Dr. Bala M. Rao

DEDICATION

This dissertation is dedicated to everyone in my big family, the giant tree that I know I can always run back to, and stay under, for love and support.

To my great grandparents, Tuu Tran and Dong Le, who put their roots in the U.S.A and made this country our new home.

To my two grandmothers, Mai Pham and Hanh Tran, two little ladies who are strong like the trunk of a tree. They carried their family through all of the hard times without a man by their side.

To my mother, Hue Tran, the flower that still blossomed after storms passed over. She taught me the importance of education and standing up after falling.

To my dad, Van Nguyen, the branch that was the "bridge" connecting and bringing me to the U.S.A.

To my other dad, Dat Nguyen, the branch that reached from outside, and intertwined with our tree, who supported and shaded me all these years.

To my uncles, Vu Nguyen and Thang Tran, the twigs that guided me up to the sky and reached higher education. I would have been lost when I first came to the U.S.A without their guidance.

To a distant "brother", Fu Shi Qi, the soothing wind that whispered to me in the burning days of COVID-19, "if flowers are in full bloom, breeze will blow naturally", meaning try your best and bloom, don't wait for the right time or luck.

BIOGRAPHY

Van Nguyen was born in 1992 in Phan Rang – Thap Cham, Ninh Thuan, a small city along the coast of Vietnam. She spent her childhood there and earned her high school diploma before moving to the U.S.A. in February 2011. Since then, she has steadily advanced her education. She had her first classes at Horizonte Instruction and Training Center, a high school for adult immigrants, from March to May 2011. She then went to Salt Lake Community College from May 2011 to December 2012. Starting in January 2013, she attended the University of Utah, obtaining her Bachelor of Science in Chemical Engineering in 2015. After graduation, she worked as a research assistant under the direction of Dr. John D. McLennan at the Energy and Geoscience Institute – a research institute at the University of Utah. In 2018, she started her graduate program at North Carolina State University in the Department of Chemical and Biomolecular Engineering. Soon after, she joined Dr. Carol K. Hall's group and was co-advised by Dr. Erik E. Santiso. She earned her Master of Science degree in Chemical and Biomolecular Engineering in 2021 and continued to pursue her Ph.D. under the direction of Professors Carol Hall and Eric Santiso.

ACKNOWLEDGMENTS

I give immense thanks to my advisor, Dr. Carol Hall, for her guidance, patience and mentorship. I have learned a lot from her, not only about research but also about personal development. Before starting my graduate program, public speaking and writing were my nightmare. Dr. Hall has given me advice to improve my skills and boost my confidence. I really appreciate her understanding and the way she respects the differences of each person. She taught me to move at my own pace as long as I kept moving forward instead of comparing myself to others. She supported me through hard times, helping me to find my motivation again after COVID-19. She is not only an advisor but a family member to me.

I also deeply thank my other advisor, Dr. Erik Santiso. He has always been there to help, always greeting me with a friendly and enthusiastic smile. He is creative and knowledgeable, often suggesting multiple solutions to a problem. I am lucky to have him as my advisor.

I would like to express my gratitude to my committee members: Professors Jason Haugh and Bala Rao, for their valuable advice and assistance throughout my graduate work.

I would like to thank all the Hall group members, both current and past. I thank our former post-docs, Dr, Xingqing Xiao and Dr. Yiming Wang, for their expertise and guidance at the beginning of my research. Dr. Wang is an expert in PRIME20 and I would not have understood the algorithm without his guidance. Additionally, I would like to thank Dr. Ryan Maloney and Dr. Amelia Chen for helping me improve my understanding of DMD. I also want to say “thank you” to my friends from my cohort: Dr. Sudeep Sarma, Dr. Rakshit Jain, and soon to be Dr. Xin Dong. Somehow, we were drawn to each other even before the first semester began. You helped me not only through hard times in research but also in life. Thanks to Matthew Dorsey, Micheal Bergman,

Corey Febo and Jake McKibbin for giving good suggestions every time I suddenly showed up in front of you and asked random questions.

Last but not least, I want to express my deep appreciation to my family and friends, who have always been supportive to me. I wouldn’t be able to complete my graduate program without knowing that they are always behind and around me.

TABLE OF CONTENTS

LIST OF TABLES ..... **ix**

LIST OF FIGURES ..... **x**

Chapter 1 Motivation and Overview **1**

1.1 Motivation 2

1.2 Overview 6

1.2.1 DMD/PRIME20 Simulation Package – A Discontinuous Molecular Dynamics Simulation Tool in Conjunction with Coarse-grained Force Field for Short Peptide Aggregation 6

1.2.2 In Silico Peptide Self-assembly Reveals the Importance of N-terminal Motifs and the Inhibition Mechanism of the Mutation L38M in α-Synuclein Fibrillation 6

1.2.3 Design of Parallel β-sheet Nanofibrils Using Monte-Carlo Search, Coarse-grained Simulations, and Experimental Testing 7

1.2.4 Future Work 8

1.3 Publications 9

1.4 References 10

Chapter 2 DMD/PRIME20 Simulation Package – A Discontinuous Molecular Dynamics Simulation Tool in Conjunction with Coarse-grained Force Field for Short Peptide Aggregation **14**

2.1 Introduction 15

2.2 Methods 19

2.2.1 Coarse-grained (CG) PRIME20 Force Field 19

2.2.2 Discontinuous Molecular Dynamics (DMD) Simulations 23

2.3 Software Installation and Implementation 25

2.3.1 Package Installation. 25

2.3.2 Input Parameters. 26

2.4 Simulation Procedure 28

2.4.1 Initial Configuration Generation 29

2.4.2 Annealing Process 29

2.4.3 DMD Simulation Production 31

2.5 Analysis Package 31

2.5.1 Trajectory Files for Visualizing Simulation Process 33

2.5.2 Energy vs Time 34

2.5.3 Hydrogen Bonds vs Time 34

2.5.4 Cluster Tracking During Simulation 35

2.5.5 Write PDB File at a Specified Time 35
2.6. Example: Simulation of α-Synuclein P1 Fragment ([G_{36}VLYVGS$_{42}$]) Using DMD/PRIM20 Package 36
2.7. Conclusion 37
2.8 References 38
Chapter 3 In silico peptide self-assembly reveals the importance of N-terminal motifs and the inhibition mechanism of the mutation L38M in α-synuclein fibrillation 52
3.1 Introduction 53
3.2 Results and Discussion 55
3.2.1 Wild-type αSyn Fragments Assemble into Fibrillar Structures in All Simulations. 55
3.2.2 Effect of Single-point Mutations L38A, V40A, and L38M on the Self-assembly of Fragments P1, P3, and P3Next. 59
3.2.3 L38M Facilitates the Interaction of P3Next with the C-terminal, Reducing Formation of Fibrillation-prone Conformations. 61
3.3 Conclusion 64
3.4 Materials and Methods 66
3.4.1 Discontinuous Molecular Dynamics (DMD) and PRIME20 Force Field 66
3.4.2 Relating Reduced Units to Real Units in DMD/PRIME20 Simulations 67
3.4.3 Simulation Procedure 67
3.7 References 69
3.8 Supplementary Materials 84
Chapter 4 Design of parallel β-sheet nanofibrils using Monte-Carlo search, coarse-grained simulations, and experimental testing 90
4.1 Introduction 91
4.2 Results 94
4.2.1 First Round of Design of Class 1 Cross β-spine Forming Peptides 94
4.2.2 Second Round of Design for Class 1 Cross β-spine Forming Peptides 97
4.2.3 Experiments Revealed Parallel Beta-sheets Nanofibers. 99
4.3 Discussion and Conclusion 102
4.4 Methods 103
4.4.1 Peptide Assembly Design Algorithm 103
4.4.2 Discontinuous Molecular Dynamics Simulation and PRIME20 Model 105
4.3 Experimental Assessment of Self-Assembly 106
4.4 Acknowledgements 106
4.5 Data Availability 106

4.6 Declaration of Competing Interests 107
4.7 References 108
Chapter 5 Conclusion and Future Work 132
5.1. Conclusion 133
5.2. Recommendation for Future Work 135

LIST OF TABLES

Table 2.1 Input parameters that are required for all DMD-PRIME20 simulations. 44
Table 2.1 (continued). 45
Table 2.2 Required folders must be created at the beginning of a new simulation. 46
Table 2.3 Command line arguments for using analysis package. 47
Table 3.1 Simulation conditions for all sequences. 74
Table 3.2 PRIME20 simulation results for all αSyn sequences. The term "aggregate" in the table means β-sheet formation. 75
Table 4.1 The sequences of the four in-silico discovered peptides in the first round of design with their corresponding $\boldsymbol{\Gamma}_{\text{score}}$, $\Delta G_{\text{binding}}$ and $\lambda \times P_{\text{aggregation}}$ values computed from PepAD. The %Parallel β-sheet content for each peptide is computed from the DMD/PRIME20 simulations. 114
Table 4.2 The sequences of the four *in-silico* discovered peptides obtained from the second round of design with their corresponding $\boldsymbol{\Gamma}_{score}$, $\Delta G_{binding}$ and $\lambda \times P_{aggregation}$ values computed from PepAD. The %Parallel β-sheet content for each peptide is computed from the DMD/PRIME20 simulations. 115
Table 4.3 Summary of computational analysis (DMD/PRIME20) and experimental measurements (FTIR, ThT, TEM and CD) of peptides PP1-PP8. 116
Table 4.4 Summary of secondary structures predicted by BeStSel and secondary structures determined by our characterization techniques 130

LIST OF FIGURES

Figure 2.1 Representation of a PRIME20 protein model for a peptide containing 3 amino acids. Each amino acid is represented by four beads: amine NH, alpha carbon CαH, carbonyl CO, and sidechain R; these are colored dark blue, cyan, red, and yellow, respectively. The central amino acid is indicated by a subscript i. The preceding and succeeding amino acids are numbered i-1 and i+1, respectively. Pseudo-bonds are dashed lines, covalent bonds and peptide bonds are solid lines. 48

Figure 2.2 NH_i-CO_j hydrogen bond (solid black arrow) formation/dissociation is restricted by the square-shoulder interactions of four auxiliary pairs (dashed line). 49

Figure 2.3 Discontinuous molecular dynamics simulation algorithm. 50

Figure 2.4 Results of analysis using the analysis package in DMD/PRIME20. 51

Figure 3.1 αSyn regions and the peptide fragments that were simulated herein. The amphipathic N-terminal (green) contains residues **αSyn[1-60]**. The central non-amyloid component (NAC) region (red) contains residues α**Syn[61-95]**. The acidic C-terminal region (orange) covers the last 45 residues. P1 and P2 are short motifs in the N-terminal region (light green). Cterm120-140 (light orange) was used in simulations of peptide mixtures to study potential inter-domain interactions. The sequences of P1, P2, P3, P3Next, and Cterm120-140 are listed. 76

Figure 3.2 Snapshots at different times during the early stage of fibril formation of P1-WT fragments from one simulation. See also Supplementary Movie 1. 77

Figure 3.3 (a) Simulation snapshot of final conformation and schematic representation of U-shaped and S-shaped conformation of P2 fibrils. (b) Simulation snapshots of P3 fragments at the end of each run show different fibril structures: flat mixed fibril (Run 1), S-shaped fibril (Run 2), and G-shaped fibril (Runs 3 and 4). As P3 contains both P1 and P2, the color code is used to display the position of P1 (red), P2 (green), and the in-between residues αSyn[43-44] (blue). (c) Schematic representations of peptide conformation for S-shaped (left) and G-shaped (right) fibrils. (d) Total interaction energy vs time for the four runs of P3-WT. The S-shaped and G-shaped structures aggregate faster and are more stable than the flat mixed fibril. 78

Figure 3.4 (a) Snapshots at the end of the simulations of P3Next wild-type fragments reveal that the number of β-hairpins varies in the P3Next-WT fibrils. The color code is used to display the position of P1 (red), P2 (green), the in-between residues αSyn[43-44] (blue), and the extended residues from P3 (orange). (b) Enlarged section from the snapshot of Run 1 shows two neighboring β-hairpins in the fibril. (c) The corresponding energy versus time profiles for all three runs display lag-times, with both large and small fluctuations in the interaction energy as a function of time. 79

Figure 3.5 (a) The DMD/PRIME20 simulations of P1 peptides with the sequence substitution L38M, L38A or V40A, as well as the wild-type P1 (P1-WT). Although all four sequences formed multi-layer fibrils, the number of parallel and antiparallel β-strands are different. (b) Average percentage of antiparallel residues within a fibril. 80

Figure 3.6 (a) Snapshot of β-hairpin formation in all P3Next variant fibrils. The number of β-hairpins in P3Next-L38M fibril is less than in the fibrils of the other variants. (b) Ratio of the number of β-hairpins to the total number of β-strands formed for each fibril in three runs plus the average..... 81
Figure 3.7 (a) DMD/PRIME20 final snapshots for mixtures that contain Cterm120-140 (red) and the P3Next variants (blue) at 300 billion collisions. (b) Snapshots at different times for the early stage of fibril formation for a P3Next-L38M and Cterm120-140 mixture. (c) Comparison of final simulation snapshots for pure P3Next-L38M and for the mixture of P3Next-L38M with C-term120-140 at 19.67μs (300 billion collisions). The mixture fibril contains significantly fewer β-hairpins than the pure fibril..... 82
Figure 3.8 (a) A possible mechanism for αSyn-FL-WT aggregation. (b) A possible mechanism for L38M inibition of αSyn-FL. L38M affects αSyn-FL by inducing intramolecular and intermolecular interaction between the P1 region and the C-terminal. The colored blocks indicate the regions that are simulated: P1 (red), P2 (green), and Cterm120-140 (blue). The rest of the αSyn-FL protein is gray..... 83
Figure 3.9 Final snapshots of DMD/PRIME20 simulations of C1ext. All simulation didn't aggregate at the end of 300 billion collisions (~20μs)..... 84
Figure 3.10 Snapshots at different times during the early stage of fibril formation of P1-L38A fragment from one simulation..... 85
Figure 3.11 Final snapshots of DMD/PRIME20 simulation of mutated **P3** variants. All snapshots were taken at the end of 500 billion collisions (~38μs) 86
Figure 3.12 Final snapshots of DMD/PRIME20 simulation of mutated **P3Next** variants. Final snapshots were taken at the end of each simulation. L38M variants formed stable fibrils within 300 billion collisions. Although L38A and V40A simulations were extended to 600 billion collisions to get stable structures, L38A Run 1 and V40A Run 2 didn't aggregate completely. 87
Figure 4.1 First round of design for Class 1 cross β-spine forming peptides: (A) Arrangement of peptides in the Class 1 cross β-spine, a 2-layer amyloid fibril consisting of parallel-oriented β-strands in each layer and antiparallel-oriented β-strands between the two layers. (B) Input fibril structure of peptide GNNQQNY for the PepAD algorithm (Conf-1) constructed using Discovery Studio 3.5 and Packmol packages and optimized by atomistic MD simulation in the AMBER14 package. (C) Fibril structure of peptide GNNQQNY obtained from PDB structure: 2omm (D) Input fibril structure of peptide GNNQQNY for the PepAD algorithm (Conf-2) constructed by performing a 25 ns MD simulation on structure in (C). (E) Plot of β-sheet content v/s simulation time describes the self-aggregation kinetics of peptide PP1 (sequence: GDIKIVV), PP2 (sequence: GNIVTFV) and PP3 (sequence: GAIDWVK) (left). The snapshot of the final simulation structures of peptides PP1, PP2, and PP3 indicating its fibrilization behavior is shown (right). Second 117
Figure 4.2 Second round of design of Class 1 cross β-spine forming peptides: **(A)** Input fibril structure of peptide PP2: GNIVTFV (with patched N- and C- termini) for the PepAD algorithm in Round 2 of design (**Conf-3**) (**B)** Plots of β-sheet content v/s simulation time describe the self-assembly kinetics of peptides PP7 and PP8.

Snapshots of the final structures of **(C)** PP7 and **(D)** PP8 were obtained from the DMD simulations. 118
Figure 4.3 TEM Imaging and FTIR spectra **(A)** TEM Image of PP1 (**B)** FTIR spectra comparing P12 from our previous work and PP3. Note that PP3 shows an absence of a peak at 1690cm^{-1}, which corresponds to antiparallel β-sheets..... 119
Figure 4.4 FTIR Spectra comparing peptides in our present study (left) with peptides from our previous study (right). Note that the peak at 1690 cm^{-1}, associated with antiparallel β-sheets, is absent in spectra on the left. 120
Figure 4.5 Thioflavin-T Fluorescence and Circular Dichroism **(A)** Fluorescence intensity plotted with time for peptides in our present study **(B)** Circular Dichroism experiments indicate that there is a dip at ~215nm characteristic of β-sheet secondary structure. 121
Figure 4.6 Class 1 cross-β spine containing 4 GNNQQNY peptides. 122
Figure 4.7 Plot of β-sheet content v/s simulation time describes the self-aggregation kinetics of peptide PP4: GAIDWVK and PP5: ADKVMFV. 123
Figure 4.8 Class 1 cross-β spine structure containing 16 GNIVTFV (PP2) peptides..... 124
Figure 4.9 Transmission electron micrographs of negatively stained samples prepared one day following peptide dissolution at 1 mg/ml..... 125
Figure 4.10 ThT fluorescence replicate curves for each of the five peptides predicted to aggregate. 126

Chapter 1 Motivation and Overview

1.1 Motivation

Do you like origami? This beautiful Japanese art form involves carefully folding a piece of paper into intricate shapes, following precise patterns. A small mistake can ruin the entire piece of art. Similarly, our bodies rely on the proper folding of numerous proteins to function correctly. A mistake in the folding process can lead to cellular dysfunction. Unlike a piece of paper that can be discarded and replaced, misfolded proteins are not easily unfolded or removed from our bodies. Instead, misfolded proteins can accumulate and form oligomers or fibrils which we call protein aggregation.

Protein aggregation is implicated in over 30 human neurodegenerative disorders, including Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease, amyotrophic lateral sclerosis (ALS), and prion diseases, affecting millions of people worldwide[1]. Various factors can influence the initiation and accumulation of misfolded proteins, such as increased protein concentration, complex intra- and inter-protein interactions, and particular cellular environments[2]. While scientists have identified the proteins associated with each disease (Aβ for AD[3], α-Synuclein (αSyn) for PD[4–6] or, Prion for prion diseases[7], etc.), the causes and mechanisms of protein aggregation are not fully known. Advancements in technologies have enabled scientists to capture high-resolution images of amyloid fibrils and to track their conformational changes, as well as their interactions during fibrillation[8–12]. Current research suggests that oligomers may be more toxic to brain cells than mature fibrils[13]. Therefore, improved understanding of the early stages of protein aggregation, transient interactions and conformations could provide insights into aggregation mechanisms and lead to effective medical treatments for these neurodegenerative diseases.

In recent years, computer simulation has become a valuable tool for studying proteins and protein aggregation. Many force fields, both atomistic and coarse-grained, and software packages have been introduced and made available to the research community such as AMBER[14], CHARMM[15], GROMACS[16], LAMMPS[17], and NAMD[18]. High-resolution models, such as atomistic-level protein models in explicit solvent, are powerful tools for investigating protein folding due to their detailed representation of the protein molecules. These models can predict secondary structure formation with high accuracy. However, high-resolution atomistic simulations are computationally expensive[19]. Even a few nanoseconds of simulation require extensive CPU time and often need to be run on supercomputers to accelerate the process. Intermediate-resolution models have been introduced to address the disadvantages of high-resolution models (long simulation time) and low-resolution models (lack of interaction details). Intermediate-resolution models adopt a coarse-grained (CG) approach to represent protein structure. They can be created by starting with atomistic high-resolution models and then reducing the number of geometric and energetic details to speed up the simulation. In CG models, a single interaction site represents a group of atoms. This helps reduce the number of interactions during the simulations. Consequently, the number of calculations decreases, and the simulations require less time to complete.

Alongside the development of coarse-grained force files, alternative simulation methods such as discontinuous molecular dynamics simulations (DMD) have emerged. DMD utilizes the simplification of discontinuous potential energy to accelerate the simulation time, offering a fast alternative to traditional molecular dynamics simulations (MD). MD simulation is a time-driven method where the simulated system is advanced at constant time steps. This approach allows detailed tracking of molecular changes but is very time-consuming, as new positions and velocities

of atoms are calculated at every step. In contrast, DMD is an event-driven method which advances the simulated system only when the distance between two particles reaches a discontinuity of the potential (an event). Until an event occurs, particles in the system are assumed to move at constant velocities as there are no changes in the force applied to the particles.

The Hall group leveraged the advantages of CG force field and DMD method to develop simulation software to investigate spontanous self-assembly of short peptides starting from their denatured state. The PRIME (intermediate-resolution protein model) force field was released in the early 2000s to study interactions and secondary structures of homogenous peptide sequences[20,21]. PRIME20, which was introduced in 2010, is an extended force field and enables examination of interactions between all 20 natural amino acids[22]. The PRIME and PRIME20 protein models use four beads to represent each amino acid; each side chain is represented by a single bead and the backbone is represented by three beads. The three-bead backbone model in PRIME20 embodies detailed information on bond angles and dihedral angles, imitating the geometry of proteins with reasonable accuracy. DMD/PRIME20 has significantly reduced the computational cost of simulating the fibrillation of small peptides and has enabled hundred-microsecond simulations[23]. The Hall group has used DMD/PRIME20 simulation to provide insights into amyloid formation by peptide fragments associated with neurodegenerative diseases including the Aβ[23–25], tau[26], prion[27–29] and α-Synuclein proteins. At the time of the writing this thesis, DMD/PRIME20 can be used to simulate amyloid aggregation in sizable systems containing nearly 200 short peptides over hundreds of μs without requiring a supercomputer system[24] making it a competitive computational tool as described in the recent review by Fatafta et al.[30]

The first focus of my research has been on developing the PRIME20 model into a user-friendly open-source software package. The package is now publicly available for download on

the Hall group's Github, complete with a detailed manual. New users, even those without experience in computational simulation, can easily install and run simulations using DMD/PRIME20 software. The package also includes analysis tools that allow users to evaluate simulation results both numerically and visually.

The second focus of my research is on applying DMD/PRIME20 to investigate the effects of point mutations on the aggregation of αSyn N-terminal fragments. Among the known neurogenerative disorders, Parkinson's disease, Lewy body dementia and other synucleinopathies are defined by the aggregation of αSyn – a presynaptic protein – into Lewy bodies. Our experimental collaborators, the Raford group, identified two short fragments, P1 (αSyn [36-42]) and P2 (αSyn[45-57]), in the N-terminal of full-length αSyn that can modulate αSyn fibrillation[31]. Inhibition was observed experimentally when deleting these fragments or when mutating leucine at residue 38 to methionine (L38M) on P1[32]. We applied DMD/PRIME20 to simulate the spontaneous assembly of fragments P1, P2 and extended fragments P3 (αSyn[36-57]) and P3Next (αSyn[27-57]) in both their wild-type and single point mutated states. From the simulation results, we identified key conformations that are prone to fibrillation. Additionally, the inhibition mechanism of the L38M mutation was also proposed based on these simulations.

DMD/PRIME20 is also applied to assist in the design of parallel β-sheet nanofibrils, conducted by Hall group member, Dr. Sudeep Sarma. Dr. Sarma developed a Peptide Assembly Design algorithm – PepAD - which enables the design of peptide sequences that self-assemble into specific supramolecular structures. We use DMD/PRIME20 simulations to refine the peptide sequences designed by PepAD, allowing us to select the most-promising sequences for in vitro verifications.

1.2 Overview

In this section we summarize the remainder of the dissertation, chapter 2-5. Each chapter contains a literature review and references.

1.2.1 DMD/PRIME20 Simulation Package – A Discontinuous Molecular Dynamics Simulation Tool in Conjunction with Coarse-grained Force Field for Short Peptide Aggregation

Chapter 2 describes the development of DMD/PRIME20 simulation package which was designed to investigate the spontaneous self-assembly of short peptides from their random denatured state. DMD/PRIME20 was built by leveraging the advantages of coarse-grained force field and discontinuous molecular dynamics (DMD) simulations which are fast alternative to classical all-atom molecular dynamics simulations. After 20 years being used in various research projects within our group, we introduced our open-source, user-friendly software package to the research community. The package includes comprehensive simulation and analysis tools. In this chapter, we explained in detail our software algorithms and architecture as well as instructions for installation and utilization of DMD/PRIME20 package. An illustrative example of simulation and analysis for peptide P1 ([G_{36}VLYVGS_{42}]) is also presented.

1.2.2 In Silico Peptide Self-assembly Reveals the Importance of N-terminal Motifs and the Inhibition Mechanism of the Mutation L38M in α-Synuclein Fibrillation

Chapter 3 describes the application of DMD/PRIME20 to reveal the importance of N-terminal motifs and the inhibition mechanism of the mutations L38M in α-synuclein (αSyn) fibrillation. αSyn is a presynaptic protein associated with several neurodegenerative diseases[5,6]. While the non-amyloid component (NAC) region of the αSyn sequence (residues 65-90) forms the core of all αSyn fibrils[33], recent findings suggest that the flanking regions play a key role in

initiating or preventing amyloid formation[31,34,35]. Two motifs in the N-terminal region, named P1 (αSyn[36-42]) and P2 (αSyn[45-57]), have been shown to be key modulators of fibril formation, with deletion of these regions, or single-point mutations in the P1 region inhibiting amyloid formation of full-length αSyn[31]. In this study, we use the coarse-grained molecular dynamics package DMD/PRIME20 to simulate the self-assembly of the P1 and P2 regions, alongside longer segments (P3 (αSyn[36-57])) and P3Next (αSyn[27-57])), and single-point mutations: L38A, V40A, and L38M, all of which have different effects on fibril formation in vitro[32]. The preliminary results show that P1, P2 and P3 have a high propensity to form parallel β-sheets while P3Next tends to form β-hairpins within fibrillar structures. The L38M substitution reduces the formation of both parallel β-sheets and β-hairpins, consistent with the inability of full-length αSyn containing L38M to form amyloid fibrils in vitro and in vivo in C. elegans[32]. In contrast, simulations of L38A and V40A show no such effect, consistent with their minimal impact on full-length αSyn fibrillation[32]. Simulations of P3Next mixed with the C-terminal segment, Cterm120-140, also show a reduced tendency for parallel β-sheet and β-hairpin formation. The DMD simulation results suggest that the presence of P1/P2 hairpins are required to unleash the amyloid potential of αSyn and offer a structural explanation of how mutations in this region protect the protein from amyloid formation. At the time of writing this thesis, additional simulations for all peptide variants are still ongoing.

1.2.3 Design of Parallel β-sheet Nanofibrils Using Monte-Carlo Search, Coarse-grained Simulations, and Experimental Testing

Chapter 4 describes our application of DMD/PRIME20 in conjunction with PepAD algorithm to design peptides that form parallel β-sheets. Peptide self-assembly into amyloid fibrils provides numerous applications in drug delivery and biomedical engineering applications[36]. We

augment our previously established computational screening technique along with experimental biophysical characterization to discover 7-mer peptides that self-assemble into "parallel β-sheets", i.e., β-sheets with N-terminus-to-C-terminus β-strand vectors oriented in parallel. To accomplish the desired β-strand organization, we applied the PepAD amino acid sequence design software to the Class-1 cross-β spine defined by Sawaya et al[37]. This molecular configuration includes two layers of parallel β-sheets stacked such that N-terminus-to-C-terminus vectors are oriented antiparallel for molecules on adjacent β-sheets. The first cohort of PepAD identified peptides were examined for their fibrillation behavior in DMD/PRIME20 simulations, and the top performing sequence was selected as a prototype for a subsequent round of sequence refinement. The two rounds of design resulted in a library of eight 7-mer peptides. In DMD/PRIME20 simulations, five of these peptides spontaneously formed fibril-like structures with a predominantly parallel β-sheet arrangement, two formed fibril-like structure with <50% in parallel β-sheet arrangement and one remained a random coil. Among the eight candidate peptides produced by PepAD and DMD/PRIME20, five were synthesized and purified. All five assembled into amyloid fibrils composed of parallel β-sheets based on Fourier Transform Infrared Spectroscopy, Circular Dichroism, Electron Microscopy, and Thioflavin-T fluorescence spectroscopy measurements.

1.2.4 Future Work

In Chapter 5, we propose future work to enhance the efficiency of the DMD/PRIME20 model. Although our software can achieve hundreds of microseconds for a system of nearly 200 peptides, there is still potential to extend simulation time to study long-time scale phenomena or to reduce the CPU time required for similar systems. We suggest parallelizing the simulation codes to maximize the use of hardware resource and adding the cis-configuration of proline to the force field to enhance the accuracy of the model.

1.3 Publications

Chapter 2-4 are based on the following publications:

Chapter 2: V. Nguyen, C. K. Hall. “DMD/PRIME20 Simulation Package – A Discontinuous Molecular Dynamics Simulation Tool in Conjunction with Coarse-grained Force Field for Short Peptide Aggregation”, *in preparation*.

Chapter 3: V. Nguyen, S. M. Ulamec, D. J. Brockwell, S. E. Radford, C. K. Hall. “In silico peptide self-assembly reveals the importance of N-terminal motifs and the inhibition mechanism of the mutation L38M in α-synuclein fibrillation”, in preparation.

Chapter 4: S. Sarma, T. R. Sudarshan, V. Nguyen, A. S. Robang, X. Xiao, A. K. Paravastu, C. K. Hall. "Design of parallel β-sheet nanofibrils with Monte-Carlo search, coarse-grained simulations, and biophysical characterization”, *Protein Science*. 2024.

1.4 References

1. Baum, J. & Raleigh, D. Protein Aggregation. *Protein Science* **27**, 1149–1150 (2018).
2. Aguzzi, A. & O'Connor, T. Protein aggregation diseases: pathogenicity and therapeutic perspectives. *Nat Rev Drug Discov* **9**, 237–248 (2010).
3. Glenner, G. G. & Wong, C. W. Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and Biophysical Research Communications* **120**, 885–890 (1984).
4. Spillantini, M. G. *et al.* α-Synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).
5. Spillantini, M. G., Crowther, R. A., Jakes, R., Hasegawa, M. & Goedert, M. α-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6469–6473 (1998).
6. Grazia Spillantini, M. *et al.* Filamentous α-synuclein inclusions link multiple system atrophy with Parkinson's disease and dementia with Lewy bodies. *Neuroscience Letters* **251**, 205–208 (1998).
7. Prusiner, S. B. *et al.* Further purification and characterization of scrapie prions. *Biochemistry* **21**, 6942–6950 (1982).
8. Young, G. *et al.* Quantitative mass imaging of single biological macromolecules. *Science* **360**, 423–427 (2018).
9. Bacic, L., Sabantsev, A. & Deindl, S. Recent advances in single-molecule fluorescence microscopy render structural biology dynamic. *Current Opinion in Structural Biology* **65**, 61–68 (2020).
10. Shashkova, S. & Leake, M. C. Single-molecule fluorescence microscopy review: shedding new light on old problems. *Bioscience Reports* **37**, BSR20170031 (2017).

11. Van Oijen, A. M. Single-molecule approaches to characterizing kinetics of biomolecular interactions. *Current Opinion in Biotechnology* **22**, 75–80 (2011).
12. Clore, G. M. Practical Aspects of Paramagnetic Relaxation Enhancement in Biological Macromolecules. in *Methods in Enzymology* vol. 564 485–497 (Elsevier, 2015).
13. Verma, M., Vats, A. & Taneja, V. Toxic species in amyloid disorders: Oligomers or mature fibrils. *Ann Indian Acad Neurol* **18**, 138 (2015).
14. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).
15. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J Comput Chem* **30**, 1545–1614 (2009).
16. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
17. Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **117**, 1–19 (1995).
18. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J Comput Chem* **26**, 1781–1802 (2005).
19. Freddolino, P. L., Harrison, C. B., Liu, Y. & Schulten, K. Challenges in protein-folding simulations. *Nature Phys* **6**, 751–758 (2010).
20. Voegler-Smith, A. & Hall, C. K. α-Helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins* **44**, 344–360 (2001).
21. Nguyen, H. D. & Hall, C. K. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16180–16185 (2004).

22. Cheon, M., Chang, I. & Hall, C. K. Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins* **78**, 2950–2960 (2010).
23. Latshaw, D. C., Cheon, M. & Hall, C. K. Effects of Macromolecular Crowding on Amyloid Beta (16–22) Aggregation Using Coarse-Grained Simulations. *J. Phys. Chem. B* **118**, 13513–13526 (2014).
24. Latshaw, D. C., Cheon, M. & Hall, C. K. Effects of Macromolecular Crowding on Amyloid Beta (16–22) Aggregation Using Coarse-Grained Simulations. *J. Phys. Chem. B* **118**, 13513–13526 (2014).
25. Cheon, M., Hall, C. K. & Chang, I. Structural Conversion of Aβ17–42 Peptides from Disordered Oligomers to U-Shape Protofilaments via Multiple Kinetic Pathways. *PLoS Comput Biol* **11**, e1004258 (2015).
26. Cheon, M., Chang, I. & Hall, C. K. Influence of temperature on formation of perfect tau fragment fibrils using PRIME20/DMD simulations. *Protein Science* **21**, 1514–1527 (2012).
27. Wang, Y. & Hall, C. K. Seeding and cross-seeding fibrillation of N-terminal prion protein peptides PrP(120–144). *Protein Science* **27**, 1304–1313 (2018).
28. Wang, Y., Shao, Q. & Hall, C. K. N-terminal Prion Protein Peptides (PrP(120–144)) Form Parallel In-register β-Sheets via Multiple Nucleation-dependent Pathways. *Journal of Biological Chemistry* **291**, 22093–22105 (2016).
29. Wang, Y., Shao, Q. & Hall, C. K. N-terminal prion protein peptides (PrP(120–144)) form parallel in-register β-sheets via multiple nucleation-dependent pathways. *Journal of Biological Chemistry* **292**, 20655 (2017).
30. Fatafta, H., Khaled, M., Kav, B., Olubiyi, O. O. & Strodel, B. A brief history of amyloid aggregation simulations. *WIREs Comput Mol Sci* **14**, e1703 (2024).

31. Doherty, C. P. A. *et al.* A short motif in the N-terminal region of α-synuclein is critical for both aggregation and function. *Nat Struct Mol Biol* **27**, 249–259 (2020).
32. Ulamec, S. M. *et al.* Single residue modulators of amyloid formation in the N-terminal P1-region of α-synuclein. *Nat Commun* **13**, 4986 (2022).
33. Sawaya, M. R., Hughes, M. P., Rodriguez, J. A., Riek, R. & Eisenberg, D. S. The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell* **184**, 4857–4873 (2021).
34. Rodriguez, J. A. *et al.* Structure of the toxic core of α-synuclein from invisible crystals. *Nature* **525**, 486–490 (2015).
35. Kessler, J. C., Rochet, J.-C. & Lansbury, P. T. The N-Terminal Repeat Domain of α-Synuclein Inhibits β-Sheet and Amyloid Fibril Formation. *Biochemistry* **42**, 672–678 (2003).
36. Boyle, A. L. & Woolfson, D. N. De novo designed peptides for biological applications. *Chem. Soc. Rev.* **40**, 4295 (2011).
37. Sawaya, M. R. *et al.* Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature* **447**, 453–457 (2007).

Chapter 2 DMD/PRIME20 Simulation Package – A Discontinuous Molecular Dynamics Simulation Tool in Conjunction with Coarse-grained Force Field for Short Peptide Aggregation

Chapter 2 is essentially a manuscript by Van Nguyen, Erik Santiso, and Carol Hall soon to be submitted to Journal of Chemical Theory and Computation (JCTC)

2.1 Introduction

PRIME20 is an intermediate-resolution protein model that was designed by the Hall group to investigate the spontaneous assembly of a system of short random-coil peptides into α-helices, β-sheets, or amyloid fibrils. The combination of a coarse-grained (CG) resolution force field, PRIME20, and the discontinuous molecular dynamics (DMD) simulation method reduces computational costs and allows sampling of longer timescales than in classical atomistic molecular dynamics (MD) simulation. Introduced in 2001 by Voegler-Smith and Hall[1], the model was initially designed to investigate α-helix formation by a single polyalanine peptide. In 2004, Nguyen and Hall[2] further developed the CG model and called it PRIME (**Pr**otein **I**ntermediate-resolution **M**od**e**l). They successfully used it to simulate spontaneous fibril formation by systems of capped polyalanine peptides (Ac-KA_{14}K-NH_2) containing 12 to 96 random coils. In 2010, PRIME20, which contains pair-wise interaction parameters for all twenty natural amino acids, was introduced by Cheon et al.[3] Since then, the Hall group has used DMD/PRIME20 simulation to provide insights into amyloid formation by peptide fragments associated with neurodegenerative diseases including the Aβ[4–10], tau[11], prion[12,13] and α-Synuclein proteins. The longest simulation using DMD/PRIME20 traced the entire fibrillation pathway of 192 $A\beta_{16\text{-}22}$ peptides and crowders for approximately 657μs on a CPU cluster system[4]. The largest systems simulated by DMD/PRIME20 contained 200 CATCH peptides, half with positive charge (6K+) and half with negative charge (6E- or 6D-)[14]; these ran for approximately 16μs on the same CPU cluster system as the longest simulation. Besides being used by the Hall group, the force field has also been applied by other groups with modifications to suit their research needs. Sahimi and his coworkers used the PRIME force field model with DMD to investigate protein dynamics in nanopores[15–17], and dynamic scaling of protein aggregation in unbounded fluids[18] or confined media[19]. Boker and Paul applied

PRIME20 with Monte Carlo simulation to study the folding behavior of single polyalanine, polyserine, and polyglutamine chains[20]. At the time of the writing this article, DMD/PRIME20 can be used to simulate amyloid aggregation in sizable systems of nearly 200 short peptides over hundreds of μs without requiring a supercomputer system, making it a competitive computational tool as described in the recent review by Fatafta et al.[21]

In the last 20 years, the number of in silico studies of amyloid fibril formation has risen and increased significantly, contributing constructively to understanding of the mechanism underlying amyloid diseases and helping to develop therapeutics. Several human neurodegenerative disorders are associated with protein fibrillation including Alzheimer's disease[22], Parkinson's disease[23], amyotrophic lateral sclerosis[24], and prion diseases[25]. Despite improvements in experimental technology, and hence our ability to obtain detailed structures for mature fibrils and the transient interactions that occur during fibrillation, it is still challenging to build a complete picture of the protein aggregation process using only experimental approaches. Important information about the thermodynamic properties and structures of transient oligomers, keys to understanding fibrillation and toxicity, require access to molecular-level characterization. Computer simulations are a nice complement to experimental studies in this regard.

In recent years, MD simulation has become a popular method for studying protein aggregation. There are many MD simulation software packages (AMBER[26], CHARMM[27], GROMACS[28,29], LAMMPS[30], NAMD[31]) that allow application of both atomistic and CG force fields to model protein aggregation. Atomistic force fields are capable of modeling protein folding and aggregation with high accuracy, as their atomistic force fields are sufficiently detailed, and the solvent molecules can be simulated explicitly. However, all-atom atomistic models come with a high computational cost, especially when they are used with MD simulations. The classical MD

method advances the system of molecules at every timestep (1-2fs) by computing the forces on all atoms, solving Newton's equation of motion numerically, and then updating the entire system. All these steps make atomistic MD simulations expensive for simulating long timescale phenomena, such as full-length protein folding or the later stages of fibrillation. These simulations can only be completed in a reasonable time frame if they are run on parallel supercomputers or specially designed hardware (e.g. MDGRAPE[32] or ANTON[33]). It is very expensive for a research group to rent or build a supercomputing system; and supercomputer resource allocations have limits.

The modeling software package DMD/PRIME20, the focus of this paper, is an open-source package coupling the CG protein model PRIME20 and the DMD method to enable the simulation of the complete fibrillation process for a system of short peptides on a single CPU core. PRIME20 is a four-bead CG model in which the four major groups (amine - NH, alpha carbon Cα-H, carbonyl CO, and sidechain R) on a residue are each mapped into a united bead. This three-bead backbone model can reproduce the dihedral angles of a protein residue and backbone hydrogen bonding yet is simple enough to substantially reduce the simulation time. All the bonded and nonbonded pair-wise interactions are described using discontinuous potentials, such as hard-sphere or square-well potentials, making the force field very well suited for use with DMD, a fast alternative method to classical MD simulation. DMD advances the system only when an event occurs (hereafter referred to as a collision). An event occurs when the distance between two particles is at a discontinuity in the potential of interaction between them. As a result, DMD allows simulations to proceed for a much longer computational time than traditional MD method, and to access later stages of aggregation. DMD was first introduced in 1959 by Alder and Wainwright for simulations of hard spheres[34]. Later, it was used by Rapaport for simulations of polymer chains[35,36]. Smith et al. applied DMD to simulations of polymeric fluids in 1997[37]. Around the same time, Zhou[38] (1997) and

Dokholyan[39] (1998) introduced their application of DMD to study protein folding. The Hall group pioneered the application of DMD coupled with CG forcefield to model peptide aggregation in 2001[1,40]. The fundamental details of CG PRIME20 and DMD will be discussed in the methods section.

The entire DMD/PRIM20 package, which includes both simulation and analysis executables, is published on our GitHub account to give free access to the research community. It is designed to be friendly and accessible to users with limited experience in computation. The package is written in Fortran f90 and is compiled using Intel Fortran *ifort*. The simulation only requires a human-readable input file and a submission script to run. Data analysis can be done using interactive commands on the terminal command line. The analysis package enables users to: (a) visualize aggregation via trajectory and PDB files, and (b) analyze aggregation kinetics. For the latter, the software outputs simulation data (energies, numbers of hydrogen bonds, sidechain-sidechain interactions, numbers of parallel and antiparallel peptide strands, and oligomer formation) as a function of simulation time. Additional functions for data analysis can be easily added to the analysis package based on users' needs but require a good understanding of the software. In this paper, we describe how to install and operate the software and provide an illustrated example of α-Synuclein (αSyn) peptide fragment P3 (αSyn[36-57]) simulation using the DMD/PRIME20 package. The package is free to download at: https://github.com/CarolHall-NCSU-CBE/Serial-DMD-PRIME20.

2.2 Methods

2.2.1 Coarse-grained (CG) PRIME20 Force Field

PRIME20 geometry

PRIME20 is a four-bead intermediate-resolution CG model that contains geometric and energetic parameters for all twenty natural amino acids. Smith and Hall developed and applied the four-bead model to study α-helix formation by a single homopeptide[1]. This model, which was called PRIME[2], is the original version of our CG force field. The subsequent PRIME20 force field was extended from PRIME to study peptide aggregation[3]. The PRIME20 geometry is illustrated in **Figure 2.1**, which shows three neighboring residues along a peptide chain. Each individual residue is represented by four beads with different colors. The network of intra and inter-residue bonds is also indicated in the figure by solid and dashed lines.

The peptide backbone comprises three united beads, as depicted in **Figure 2.1**. The diameters and masses of the backbone beads are based on the actual size of the groups that they represent in the model. The center of the amine united bead is positioned at the center of its nitrogen atom, and the centers of the alpha carbon and the carbonyl united bead are positioned at the respective carbon atoms. The three-bead backbone model provides a nice balance between a relatively realistic backbone geometry and an efficient peptide model. Firstly, it explicitly includes hydrogen bonding in the backbone as we will explain. Secondly, it realistically represents bond angles and dihedral angles, important properties in determining protein structures. Details of the bond parameters and angles will be presented in the next section.

Each sidechain is mapped onto a single united bead positioned at the center of mass of the sidechain group (yellow bead in **Figure 2.1**). The mass of each sidechain bead equals the total mass of all the atoms in that sidechain, but the diameter of each sidechain is parameterized

specifically for each sidechain-sidechain interaction. Cheon et al.[3] calculated radial distribution function histograms between each pair of sidechains in 711 native states from the Protein Data Bank and selected the first non-zero value of the radial distribution function to be the "sphere diameter" of both interacting sidechains. If the sidechain-sidechain interaction is between amino acids of the same type, the "sphere diameter" is the diameter of that CG sidechain bead. If the sidechain-sidechain interaction is between different types of amino acids, the "sphere diameter" can be understood as the distance between the center of the two sidechain beads. As glycine doesn't have a sidechain, 190 "sphere diameters" for all possible interacting sidechain pairs can be found in the supplementary materials from Cheon et al[3].

PRIME20 maintains realistic protein geometry through covalent bonds and pseudo bonds. Three covalent bonds connect NH, CO, and R to CαH, and one covalent bond represents the peptide bond that links two amino acids together (solid lines in **Figure 2.1**). The bond lengths are set equal to the ideal bond length in real proteins. Six pseudo bonds (dashed lines in **Figure 2.1**) maintain the L-isomer conformation and keep dihedral bond angles in the allowed regions on the Ramachandran plot[41]. The pseudo bond connecting CαH to the neighboring $C_{\alpha}H$ maintains the peptide in the trans configuration. Currently, PRIME20 does not consider cis configurations, which means that it cannot model proline as accurately as the other amino acids. Values for the covalent and pseudo bond lengths were published in the work by Voegler-Smith and Hall[1].

PRIME20 interactions

All pair-wise interactions are represented by discontinuous hard-sphere and square-well potentials for compatibility with DMD simulations. There are three types of discontinuous potentials that are used to describe events in PRIME20: hard-sphere potentials for excluded volume between nonbonded beads, single infinitely deep square-well potentials for the covalent

bonds and pseudo bonds between nearby beads, and square-well potentials for backbone hydrogen bonds and for sidechain-sidechain interactions. The solvent, water, is modeled implicitly in the energy function as a potential of mean force.

The excluded volume effect between nonbonded beads is modeled as a hard-sphere potential. An infinite repulsion occurs when the distance between two beads is equal to the arithmetic mean of their diameters according to the Lorentz-Berthelot mixing rules[42]. This distance is used as the hard-sphere diameter of the interaction potential. Excluded volume events (hard sphere repulsions) also occur between beads that are separated by three covalent bonds. As united atoms are bulkier than atomistic atoms, excluded volume events are more likely to happen between the two united atoms than between the corresponding atomistic atoms. To reduce the likelihood of this excluded volume event so that the interaction frequency is more realistic, a "squeeze factor" is applied to reduce the hard-sphere diameters of these beads. Squeeze factors for backbone-backbone excluded volume interactions can be found in Voegler-Smith's dissertation and is published by Boker and Paul[47]. Squeeze factors for sidechain-backbone excluded volume interactions are published by Cheon et al.[8]

Covalent bonds and pseudo bonds (**Figure 2.1**) are modeled by single infinitely deep square-well potentials. The bond length can freely fluctuate within ± 2.375% of the bond length[43]. When the distance between the beads is at a maximum or a minimum, a collision event occurs (see section below on DMD). This event adds a computational cost to the simulation, because it happens with very high frequency. The advantage of this type of event is that it allows the chained beads to be decoupled. As a result, each chained bead can be treated as an individual bead when it participates in other types of colliding events.

The backbone hydrogen bonding interaction between amine and carbonyl groups is modeled by a square-well potential. Hydrogen bonding between intrapeptide NH and CO beads allows for α-helix and β-hairpin formation. Hydrogen bonding between interpeptide NH and CO allows for β-sheet formation. A hydrogen bond is formed only if the two beads satisfy three restrictions. First, the NH and CO beads must not participate in a hydrogen bond event with another bead. Second, the hydrogen bond angle constraints must be satisfied. Third, the NH and CO beads must be at least three residues apart from each other for intrapeptide hydrogen bonding to occur. The second restriction is maintained by the distances of four auxiliary pairs NH_i–$C\alpha H_j$, NH_i–NH_{j+1}, CO_j–$C\alpha H_i$, CO_j–CO_{i-1} neighboring to the hydrogen bonding pair NH_i-CO_j as shown in **Figure 2.2**. When NH_i and CO_j are captured in their square-well and have potential to form a hydrogen bond, the four auxiliary pairs are set to interact via square-shoulders. The distance of each auxiliary pair must not be shorter than its square-shoulder distance determined by Nguyen et al.[43] for the hydrogen bond to be formed and maintained. When the hydrogen bond between NH_i and CO_j brakes, the interactions between auxiliary pairs are reversed to their original collision types. Although the natural distance between the C atom and the N atom when forming hydrogen bond is 4.2 Angstrom, PRIME20 sets the hydrogen bond length to 4.5 Angstrom[11] to increase the possibility of forming hydrogen bonds when sidechain beads are large.

Sidechain-sidechain interactions are described by square-well potentials with a specific well-depth and well-width for each pair. Although Cheon et al.[3] derived a set of 171 bead diameters for all 19 sidechains (glycine doesn't have a sidechain), they reduced the number of energy parameters (well-depth and well-width) to 19 by grouping sidechains with similar characteristics. The 20 amino acids are classified into 14 groups based on their hydrophobicity, polarity, charge, hydrogen bonding potential, and size. A perceptron learning algorithm and a modified stochastic

learning algorithm were combined to search for parameters that can distinguish the native structure (lowest energy) from millions of decoys (larger energy than nature structure). The 19-energy parameter set was found to contain enough detail to realistically describe sidechain-sidechain interactions.

2.2.2 Discontinuous Molecular Dynamics (DMD) Simulations

DMD uses discontinuous potentials such as hard-sphere or square-well potentials to describe the interaction between CG beads. DMD is event-driven as opposed to molecular dynamics, which is time-driven. An event, also called a collision, occurs when the discontinuity in the interaction potential between two beads is reached. In this paper, three types of time units are used to describe simulation time: SI unit, reduced time unit, and number of collisions. DMD simulation of polymers and other chain-like molecules was first introduced by Rapaport[35,36,44] and later modified by Bellemans et al.[45] Since then, DMD simulation has been used to study protein folding and aggregation[46].

DMD speeds up the simulation process by reducing the number of calculations compared to MD. At each iteration, the algorithm assumes that all beads in the system move at constant velocities meaning that there is no force applied on the beads. Therefore, the DMD method excludes the computational cost of numerically solving Newton's equation of motion at small timesteps (1-2fs). Instead, the program calculates the collision times for each pair of beads when the distance between them is at a discontinuity in the interaction potential. Single-event scheduling is used to store collision times. This scheduling method only stores the soonest-to-happen event for each bead. A bucket sorting algorithm is used to reduce the cost of searching for the soonest-to-happen event for the whole system. A small portion of the collision times that are stored in the first bucket go through bubble sorting to reduce the number of sorted events. DMD then solves for

the new velocities and positions of the two colliding beads using conservation of energy and momentum. The system is then advanced to the next soonest-to-happen collision time. A flow chart describing the DMD simulation procedure is shown in **Figure 2.3**.

The types of events that occur are classified according to the types of colliding beads and their velocities. Excluded-volume events happen when nonbonded beads overlap. Details of this event are discussed in the previous section. Bond events occur when bonded beads move outward and try to break their assigned bonds. When this happens, infinitely strong attractive forces are applied so that the beads move back and forth within permissible bond lengths. Square-well events include capture, bounce and dissociation events that occur in hydrogen bonding and sidechain-sidechain interactions. A capture event occurs when a bead enters the square-well potential of another bead. A bounce event occurs when a bead attempts to leave the potential well but is unable to overcome the attractive potential and is reflected back into the well. A dissociation event occurs when a bead has sufficient kinetic energy to exit the square well, removing the interaction between the beads.

DMD simulations are performed in the canonical ensemble (NVT); the temperature of the system is maintained at a constant value using the Andersen thermostat[47]. The thermostat is realized by randomly selecting a bead to collide with a ghost bead after a certain time. This collision resets the velocity of the randomly selected bead and causes a change in the kinetic energy of the system. As a result, the temperature is maintained at the defined temperature. Energy and temperature in DMD are described in reduced units following the definitions by Allen and Tyldesley[48]. Reduced energy (ε^*) is defined by $\varepsilon^* = \varepsilon/\varepsilon_{HB}$ where ε is the energy in SI unit and ε_{HB} is hydrogen bond energy (12.47 kJ/mol). Reduced temperature (T^*) is defined by $T^* = k_B T/ \varepsilon_{HB}$ where k_B is Boltzmann's constant, and T is temperature in SI unit.

Unit conversion from reduced units to real units is based on the work of Wang et al.[13] They matched the self-diffusion coefficients predicted by DMD/PRIME20 to those predicted by MD simulations for Aβ(16-22) at 298K and found that a reduced time unit in DMD/PRIME20 is approximately 0.96 ns. Since Wang et al. found this correlation between reduced time units and real time units, all results from DMD/PRIME20 use this estimated correlation to better relate simulation time to real time. Users can derive their own correlation to produce more accurate conversions for their system of interest. This conversion works reasonably well as an indication of time evolution of other properties. The time conversion is implemented in the analysis package. Users can choose to present simulation time in terms of the number of collisions, reduced time or time in μs. Temperature conversion for DMD/PRIME20 simulations was introduced in the paper by Wang et al.[13] The equation for temperature conversion is T*= (T +115.79)/2288.46.

2.3 Software Installation and Implementation

2.3.1 Package Installation.

The DMD/PRIME20 package is available for download on the Hall GitHub repository at the link provided in the introduction. Detailed instructions for installation and usage are provided in the readme. The software is written in Fortran90 and is assembled using the Intel *ifort* compiler. An Intel compiler must be installed on the user's device before installing the DMD/PRIME20 software. The package includes two main directories: **/src/** and /**example**/. The **/src/** directory contains all of the simulation and analysis source codes. The /**example**/ directory contains a comprehensive example including the input.txt file, a submission script for a local computer system, and the simulation results. A new simulation goes through three steps: generating the initial configuration, running the annealing stages, and running the production stages. Initial configuration generation has a separate "makefile" – *genconfig.mk*, while annealing and

production simulation share the same "makefile" – *dmd.mk*. An extra "makefile" – *dmd_analysis.mk* is used to create the executable file for the analysis package. To compile on a command line interface, navigate to the **/src/** directory on your local device. and create the executables by entering the three commands below.

To create **initconfig** for generating initial configuration:

```
$ make -f genconfig.mk
```

To create **DMDPRIME20** for DMD simulations:

```
$ make -f dmd.mk
```

To create **dmdanalysis** for data analysis:

```
$ make -f dmd_analysis.mk
```

If no errors are returned during the previous step, the *initconfig*, *DMDPRIME20*, and *dmdanalysis* executable files have been successfully created in the **/src/** directory. The DMD/PRIME20 package can be installed anywhere on the user's local device separate from the directories in which users run their simulations. Paths to these executable files need to be specified in the job submission script so that the computer can correctly locate these files. An example of a job submission script is in the /**example**/ directory and will be discussed in detail in Section 4.

2.3.2 Input Parameters.

Simulations using DMD/PRIME20 require an input.txt file that contains the necessary information to generate a simulation system. The required parameters are listed in **Table 2.1** below. As our software is a CG model, it cannot take a molecular structure PDB file as an initial input as MD simulation software does. Users must specify all parameters listed in **Table 2.1** for DMD/PRIME20 to generate a complete initial configuration for their simulation. The parameters must be entered following the exact order and format listed below. We recommend that users start

with the *input.txt* file in the /**example**/ directory and modify the parameters for their desired system.

There are nine mandatory parameters, and six additional parameters required for a complete simulation with DMD/PRIME20. The nine mandatory parameters include: peptide sequence (*pep1 and pep2*), number of peptide chains for each type (*chain1 and chain2*), periodic boundary condition box length (*boxlength*), simulation temperature (*T*), number of collisions for each simulation round (*coll*), number of collisions for trajectory recording (*trajrecord*) and annealing option (*annealing*). The current DMD/PRIME20 allows simulation of either a pure system containing only one peptide or a mixed system containing two peptides. Each peptide has no longer than 30 residues. Two important properties that affect the simulation results are the concentration and the temperature. For our group's research, we usually choose high concentrations (10 or 20 mM) to speed up the simulation. However, in DMD/PRIME20, we don't specify concentration, instead we specify the length of the simulation box. The equation that relates the box length to concentration is given in **Table 2.1**. The simulation temperature is usually chosen so that it is high enough for the system to get out of its local minima and thereby allow formation of ordered structures. However, if the temperature is too high, the system will remain as a dissociated fluid phase[38]. The simulation temperature can be chosen by running a series of short simulations at different temperatures and selecting the one that results in ordered structures.

DMD/PRIME20 is designed to run, complete and record data in many cycles to avoid having a single large result file and to allow the simulation to restart if it crashes midway. Therefore, DMD/PRIME20 asks users to provide the number of collisions to run to complete a cycle. This will be specified in the submission script. The total simulation time will equal the number of collisions (*coll*) multiplied by the number of complete cycles (specified in submission

script). The *trajrecord* is the frequency at which to record coordinates of all beads in the system. Each set of coordinates is a frame that is collected to make a multi-frame trajectory that shows the time evolution of the system during a simulation. At the beginning of a new simulation, DMD/PRIME20 normally heats up the system to make sure that all peptides are denatured coils; then the temperature is reduced to the simulated temperature in an annealing process. Users can choose to use our default set of temperature for annealing process, or they can choose their own temperatures.

The six additional parameters are optional and only need to be specified if users want to run a simulation with non-default conditions. If users choose to run an annealing process with their choice of temperatures, they must specify the starting and ending temperatures, the temperature drop and how long they want to run at each temperature. The other two optional parameters are for replicating a simulation. This option allows the user to run an exact replicate of a simulation in case the results are lost, or to rerun a certain part of a simulation to record extra data. The *seedconfig* and *seedsim* are the seed numbers for the random number generator for initial configuration generation and simulation, respectively. These values can be found in the output logs that are located in the **/outputs/** directory listed in **Table 2.3** in the next section.

2.4 Simulation Procedure

The procedure for running a new simulation contains three steps: initial configuration generation, annealing to relax the system, and production. Before starting any new simulation, all folders listed in **Table 2.3** below must be created. These folders must be placed in the same running directory as the input.txt and the submission script. The names of these folders are specific, therefore misspelling folder names will cause error and job termination.

2.4.1 Initial Configuration Generation

Our CG PRIME20 software generates a system of peptides in random coil conformations based on the specified input parameters. The software cannot start a simulation with a premade configuration from other software because the CG force field is very strict about bead sizes, bond-lengths, and bond-angles. A cubic box with periodic boundary conditions is created first, with each side equal to the *boxlength*. Then a single generic 31 residue long peptide is placed in a random position within the box. The generic peptide is then modified to adopt the bond-lengths and sidechains of the user-specified peptide sequence. If the specified peptide is less than 31 residues, the rest of the generic peptide will be removed. After ensuring that there is no overlap between beads, copies of those peptides are placed at random locations in the box based on the number of peptide chains requested. Every time a chain is placed in the box, it is checked for overlapping with the chains that are already there. If there is an overlap, the new chain will be moved to another random position until no overlapping is detected. Velocities are then assigned to each bead randomly from a Maxwell-Boltzmann distribution at the simulation temperature. If restarting or resuming a simulation, this step is skipped as the initial configuration files produced are identical. The command to execute the initial configuration generation procedure is below. The **path_to_executable/** is the path to the directory where users saved the DMD/PRIME20 package, and the executable files compiled in Section 2.3.1.

```
# Generate initial configuration for new simulation:
path_to_executable/initconfig
```

2.4.2 Annealing Process

The newly generated system described above must undergo an annealing process before simulation production is executed to ensure that the peptides are in random coil conformations.

The annealing process starts at a very high temperature and then slowly cools down to near simulation temperature. DMD/PRIME20 has a default set of temperatures that are used for annealing. These temperatures have been used in previous simulations. When using the default set of annealing temperatures, the system is heated up to 1028K and cooled down to 388K. The annealing process consists of stepwise cooling to 9 constant temperatures. At every constant temperature step, a short simulation of 1 million collisions is run so that the system reaches a steady state before moving on to the next constant temperature step. The package also allows the user to define their own set of annealing stages; the user can assign a range of temperatures, temperature decrements and duration for the annealing process. We recommend that users check the temperature profile using the implemented analysis package (discussed in section 5) to ensure that the system reaches equilibrium at each temperature before the next temperature stage. If equilibrium is not reached at an annealing stage, the user should rerun the annealing process with an extended duration until equilibrium at each temperature stage is observed. Below is the command to execute the annealing process with default temperatures. If using default temperature, there are 9 cycles running from *{1..9}*. If using a user-defined temperature, the number of annealing cycles is found as *(endingtemp-startingtemp)/tempstep + 1*. For example, if the annealing process is set to run from 500K to 300K with the *tempstep* to be 50K, then the number of annealing cycles is (500K-300K)/50K + 1 which is 5. Thus, the loop is set as {1..5} in the submission script.

```
#Annealing:
# The numbers in the bracket are the number of time annealing temperature drop
for i in {1..9}
do path_to_executable/DMDPRIME20 < inputs/annealtemp_$i > outputs/out_annealtemp_$i
done
```

2.4.3 DMD Simulation Production

DMD simulation is performed using information that is specified in input.txt and in the submission script. The output data is written in the **/results/** and **/output/** folders. For the simulation to run, DMD/PRIME20 requires the user to define the simulation temperature (T), the number of collisions in each cycle (*coll*) and the number of simulation cycles. The first two parameters are specified in input.txt and are explained in **Table 2.1**. The number of simulation cycles specified in curly brackets in the submission script. For example, the below command shows that the simulation will be run for 300 cycles ({1..300}).

```
# DMD simulations
# The numbers in the bracket are the number of simulation cycles.
# If restarting or resuming the simulation, the first number is where the simulation is restarted
for i in {1..300}
do path_to_executable/DMDPRIME20 < inputs/simtemp > outputs/out_simtemp_$i
done
```

The numbers in the curly brackets can be adjusted, depending on how long users want their simulation to run. For a new simulation, the simulation starts at 1. For a resumed simulation, it should start from the most recently completed simulation (e.g. a simulation started from 101 and ending at 200 would be written as {101..200}, if the previous simulation completed and stopped after 100 cycles).

2.5 Analysis Package

DMD-PRIME20 allows for simulation timescales on the order of hundreds of microseconds. The average simulation for a system of 24 peptides of length 30 residues takes about 2 weeks to complete on a Dell PowerEdge R720xd Servers with an Intel Xeon E5-2670 32-core CPU. A complete simulation generates a large set of result files. To minimize the necessary storage space needed for a simulation, most of the data is binary. The analysis package allows users to

convert binary files into a human-readable format. It is under continuous development at the time of this writing, and new functions will be added based on user feedback. Users can also contribute their own functions to the current analysis package on the Github repository.

After the DMDAnalysis executable is created, it should be run in the directory containing input.txt and the submission script. It can take from a few seconds to a few minutes to complete this task depending on how big the results data set is. For longer simulations (execution times of about one month), the resulting data set will likely be larger than 300MB; in this case the analysis package would need extra time to extract the requested data. The recorded data is extracted. This contains essential information such as coordinates, velocities, energy, bonding partners, etc. Then the analysis algorithms conduct the necessary calculations to retrieve the information requested by users. Analysis outputs are in text *.txt* format with descriptive labels for use in further calculation or visualization. The data analysis is executed using the commands given below. Each type of analysis must be executed via an individual command that contains all of the required arguments. All options for input arguments are listed in

Table 2.4.

```
# General analysis command:
path_to_executable/DMDAnalysi argument1 argument 2 argument3
```

The arguments in

Table 2.4 are defined as follows. Argument 1 specifies what data the user wants to extract from the data set, for example energy (*evst*), hydrogen bonding (*hbvst*) or trajectory (*traj*). Argument 2 and argument 3 specify the *start_file* and *end_file* that the user is interested in; these are four-digit integers corresponding to the numbering of the result files at each cycle. Initial configuration files start as 0000. We skip these files when analyzing data. All results from the first

annealing cycle are labeled with 0001 and the number keeps ascending to the last annealing cycle. The simulation production cycle numbering continues after the annealing numbering. For example, all files from the last annealing cycles using default temperature set are labeled with 0009. The first simulation cycle is then 0010 and it will keep ascending up to the last cycle; for example, 0109 if the simulation is run for 100 cycles.

2.5.1 Trajectory Files for Visualizing Simulation Process

DMD/PRIME20 outputs the trajectory of the simulation process for the time frame defined by the user. During a simulation, or at the end of a simulation, the user can use the argument *traj* to combine these frames and make a multi-frame trajectory file in XTC format to create a movie that tracks the aggregation process. The file can be loaded to Visual Molecular Dynamics (VMD) software for visualization. The XTC format is a portable format for trajectories. It uses the XDR routines for writing and reading data which were created for the Unix NFS system. The XDR interface for Fortran was developed by Kai-Min Tu[49], and the XTC Library was developed by James 'Wes' Barnett[50,51]. As the XTC file only contains the coordinates, a PDB file for the same system needs to be loaded to VMD before loading the trajectory file to provide information about the peptide sequences in the system. All of the PDB files that are generated at the end of each cycle can be found in the **/results/** directory. We recommend that users use the last PDB file in which an ordered fibril has formed for the New Cartoon drawing method. VMD will use the secondary structure format obtained from this PDB file to display the peptide chains throughout the trajectory. With this format, peptide chains will consistently appear to be a ribbon that shows the head and tail of the peptide. VMD doesn't recalculate the secondary structure at every frame but instead uses the secondary structures of the PDB file to assign the shapes and colors of the peptides throughout the whole trajectory display. If the user opts not to use the last PDB file for the New

Cartoon drawing method, but instead uses the initial PDB file, VMD will adopt random coil format from the initial PDB file and apply this format throughout the trajectory, even after secondary structures form. The peptides and aggregates will appear to be fragmented due to the periodic boundary conditions, causing overlong bonds to stretch across the simulation box. The problem can be fixed using the '*pbc join*' command in VMD to rejoin the segments along the peptide chains. Additional information can be found in the user manual.

2.5.2 Energy vs Time

The energy profile can be used to indicate when peptide aggregation starts and when the system reaches a stable state. We recommend that the user generate energy vs time results as the first step of the analysis because the result from this step is required for all of the subsequent following analysis steps. A file is generated containing six columns: number of collisions, time (μs), temperature (K), total energy (Etotal), kinetic energy (KE) and potential energy (PE). If the user chooses to run the annealing process at temperatures of their choice, the user should generate and check for energy convergence as soon as the annealing completes. Based on the reported temperature, users will be able to tell if the system has reached the target temperature or if they need to run annealing for longer than they originally set. Aggregation typically starts when the potential energy drops significantly. In addition, the conversion between number of collisions and real time is also reported in the energy vs time text file. This information is needed for further analysis.

2.5.3 Hydrogen Bonds vs Time

The number of hydrogen bonds formed during the simulation can act as an indicator of peptide aggregation. Both interpeptide and intrapeptide hydrogen bonds are reported along with the corresponding number of collisions and time. The interpeptide hydrogen bond formation versus

time plot can tell us when aggregation starts (number of hydrogen bonds raise significantly) and when the systems reach steady state (number of hydrogen bonds fluctuates slightly around a constant value). The number of intrapeptide hydrogen bonds includes the number of hydrogen bonds that form into α-helix or β-hairpin. The result from running *hbvst* command is a text file that contains 6 columns: number of collisions, simulation time (μs), number of hydrogen bonds that form β-sheets, α-helices, β-hairpins, and amorphous structures.

2.5.4 Cluster Tracking During Simulation

The argument *cluster* is used to calculate the number of monomeric peptides and number of oligomers and fibrils over the course of the simulation. Users can learn if a fibril has formed by combining small oligomers into a larger fibril or by stringing monomers together to become a single oligomer. Plotting the number of clusters vs time is also a good tool to tell when important events occur. From this, users can generate a trajectory file to visualize the conformational state of the system at the time of interest. An oligomer is defined to be a cluster in which the total number of bonded peptides is larger than 3. A fibril is defined to be a cluster in which the total number of peptides on a β-sheet is larger than 5 and the cluster contains at least one β-sheet in that cluster.

2.5.5 Write PDB File at a Specified Time

Users can write a PDB file at any time at which the coordinates are recorded during the simulation, in addition to the final PDB file, which is automatically generated at the end of each simulation cycle. DMD/PRIME20 records data to binary files at a time step that is proportional to $\frac{1}{\sqrt{T*}}$ following the reduced time defined by Allen and Tildesley[52]. The higher the temperature, the less often the data is recorded. A simulation can last from a few to hundreds of microseconds; therefore, the results would be very large if data were recorded at every step. Another reason for

not recording data at every step is that bond vibrations happen much more often than the main events as we discussed in Section 2.1.2, therefore, there would not be is much to observe if it was recorded too often. The software searches for the nearest recorded configuration to the time (in billion collisions) that is requested and creates a PDB file after the correct recorded time. PDB files generated using the analysis command appear to be different from the PDB files generated at the end of each complete cycle. The end-of-cycle PDB files are created for users to quickly visualize the simulation process. In this case, the effect of periodic boundary condition is not removed. As a result, aggregate structures might appear as disrupted or segmented. The PDB generation tool in the analysis package applies the MDVWhole method developed by Bruininks et al.[53] to remove the periodic boundary condition so that the simulated structures appear as a whole.

2.6. Example: Simulation of α-Synuclein P1 Fragment ([G_{36}VLYVGS_{42}]) Using DMD/PRIM20 Package

In this section, we describe how to use the published DMD/PRIME20 package to run a simulation for a peptide named P1. P1 (GVLYVGS) is a seven-residue motif derived from residue 36 to 42 in the N-terminal region of the protein α-Synuclein. This motif was investigated by Doherty et al., who found experimentally that deleting P1 from the full-length α-Synuclein can inhibit peptide aggregation. We have performed computational studies on this peptide for the last few years. Details of this research can be found in our publication on ASyn, which we are about to submit. Here, we only use P1 to demonstrate the results that can be obtained from DMD/PRIME20. No further discussion on the aggregation kinetics of P1 will be presented.

Simulation of 48 P1 peptides at a concentration of 10mM and temperature of 310K is performed for 100 billion collisions, approximately 14μs simulation time. The trajectory file was

recorded every 100 million collisions. The simulation was run on the Hall group computer cluster which is 2 Dell PowerEdge R720xd Servers (32 cores), Dual QPI sockets, Intel Xeon E5-2670 processors with 96GB memory for nearly 8 days. Figure 3 shows the results from these DMD/PRIME20 simulations. Based on the plot of energy vs time, P1 aggregates very quickly at about 1μs. The fibril formation plot shows that 2 oligomers started to form at around 0.4539μs. One of the two oligomers grew and formed a fibril at 0.5597μs. Thereafter, the oligomer and fibril both elongated by having monomers attached to them. While the fibril grew steadily, the oligomer dissolved into monomers around 3.3228μs, reformed very shortly after that, and then dissolved again causing a rise in the number of monomers at 3.4362μs. Thereafter, the fibril kept elongating until only 1 monomer was seen in the system. These events are captured in the snapshots in **Figure 2.4** using VMD. The full trajectory video is in the supplementary information.

2.7. Conclusion

In this paper, we have described our DMD/PRIME20 simulation package and the architecture of the program. We also described the steps and commands for installation, implementation, and application of the package. The DMD/PRIME20 software combines the advantages of both the CG force field and the DMD method to simulate the complete aggregation process for short peptides. Simulations of hundreds of μs can be performed within a month as a serial job on a CPU computer workstation. This capability allows many simulations to be run at the same time on a CPU computer system. In addition, most of the aggregation processes can be completed in a shorter time than one month. Our software is easy to use as users only need two files to run a simulation and short commands to complete the data analysis. The analysis package attached to the software enables investigation of aggregation kinetics without any other analysis tools, except VMD for visualization.

2.8 References

1. Voegler-Smith, A. & Hall, C. K. α-Helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins* **44**, 344–360 (2001).
2. Nguyen, H. D. & Hall, C. K. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16180–16185 (2004).
3. Cheon, M., Chang, I. & Hall, C. K. Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins* **78**, 2950–2960 (2010).
4. Latshaw, D. C., Cheon, M. & Hall, C. K. Effects of Macromolecular Crowding on Amyloid Beta (16–22) Aggregation Using Coarse-Grained Simulations. *J. Phys. Chem. B* **118**, 13513–13526 (2014).
5. Latshaw, D. C. & Hall, C. K. Effects of Hydrophobic Macromolecular Crowders on Amyloid β (16–22) Aggregation. *Biophysical Journal* **109**, 124–134 (2015).
6. Cheon, M., Chang, I. & Hall, C. K. Spontaneous Formation of Twisted Aβ16-22 Fibrils in Large-Scale Molecular-Dynamics Simulations. *Biophysical Journal* **101**, 2493–2501 (2011).
7. Cheon, M., Kang, M. & Chang, I. Polymorphism of fibrillar structures depending on the size of assembled Aβ17-42 peptides. *Sci Rep* **6**, 38196 (2016).
8. Cheon, M., Hall, C. K. & Chang, I. Structural Conversion of Aβ17–42 Peptides from Disordered Oligomers to U-Shape Protofilaments via Multiple Kinetic Pathways. *PLoS Comput Biol* **11**, e1004258 (2015).
9. Wang, Y. *et al.* Thermodynamic phase diagram of amyloid-β (16–22) peptide. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2091–2096 (2019).

10. Wang, Y., Latshaw, D. C. & Hall, C. K. Aggregation of Aβ(17–36) in the Presence of Naturally Occurring Phenolic Inhibitors Using Coarse-Grained Simulations. *Journal of Molecular Biology* **429**, 3893–3908 (2017).
11. Cheon, M., Chang, I. & Hall, C. K. Influence of temperature on formation of perfect tau fragment fibrils using PRIME20/DMD simulations. *Protein Science* **21**, 1514–1527 (2012).
12. Wang, Y. & Hall, C. K. Seeding and cross-seeding fibrillation of N-terminal prion protein peptides PrP(120–144). *Protein Science* **27**, 1304–1313 (2018).
13. Wang, Y., Shao, Q. & Hall, C. K. N-terminal Prion Protein Peptides (PrP(120–144)) Form Parallel In-register β-Sheets via Multiple Nucleation-dependent Pathways. *Journal of Biological Chemistry* **291**, 22093–22105 (2016).
14. Dong, X. Y., Liu, R., Seroski, D. T., Hudalla, G. A. & Hall, C. K. Programming co-assembled peptide nanofiber morphology via anionic amino acid type: Insights from molecular dynamics simulations. *PLoS Comput Biol* **19**, e1011685 (2023).
15. Javidpour, L. & Sahimi, M. Confinement in nanopores can destabilize α-helix folding proteins and stabilize the β structures. *The Journal of Chemical Physics* **135**, 125101 (2011).
16. Javidpour, L., Tabar, M. R. R. & Sahimi, M. Molecular simulation of protein dynamics in nanopores. I. Stability and folding. *The Journal of Chemical Physics* **128**, 115105 (2008).
17. Javidpour, L., Tabar, M. R. R. & Sahimi, M. Molecular simulation of protein dynamics in nanopores. II. Diffusion. *The Journal of Chemical Physics* **130**, 085105 (2009).
18. Zheng, S., Javidpour, L., Shing, K. S. & Sahimi, M. Dynamics of proteins aggregation. I. Universal scaling in unbounded media. *The Journal of Chemical Physics* **145**, 134306 (2016).

19. Zheng, S., Shing, K. S. & Sahimi, M. Dynamics of proteins aggregation. II. Dynamic scaling in confined media. *The Journal of Chemical Physics* **148**, 104305 (2018).
20. Böker, A. & Paul, W. Thermodynamics and Conformations of Single Polyalanine, Polyserine, and Polyglutamine Chains within the PRIME20 Model. *J. Phys. Chem. B* **126**, 7286–7297 (2022).
21. Fatafta, H., Khaled, M., Kav, B., Olubiyi, O. O. & Strodel, B. A brief history of amyloid aggregation simulations. *WIREs Comput Mol Sci* **14**, e1703 (2024).
22. Glenner, G. G. & Wong, C. W. Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and Biophysical Research Communications* **120**, 885–890 (1984).
23. Grazia Spillantini, M. *et al.* Filamentous α-synuclein inclusions link multiple system atrophy with Parkinson's disease and dementia with Lewy bodies. *Neuroscience Letters* **251**, 205–208 (1998).
24. Rosen, D. R. *et al.* Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59–62 (1993).
25. Prusiner, S. B. *et al.* Further purification and characterization of scrapie prions. *Biochemistry* **21**, 6942–6950 (1982).
26. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).
27. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J Comput Chem* **30**, 1545–1614 (2009).
28. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).

29. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
30. Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **117**, 1–19 (1995).
31. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J Comput Chem* **26**, 1781–1802 (2005).
32. Ohmura, I., Morimoto, G., Ohno, Y., Hasegawa, A. & Taiji, M. MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations. *Phil. Trans. R. Soc. A.* **372**, 20130387 (2014).
33. Shaw, D. E. *et al.* Anton 3: twenty microseconds of molecular dynamics simulation before lunch. in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* 1–11 (ACM, St. Louis Missouri, 2021). doi:10.1145/3458817.3487397.
34. Alder, B. J. & Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics* **31**, 459–466 (1959).
35. Rapaport, D. C. Molecular dynamics simulation of polymer chains with excluded volume. *J. Phys. A: Math. Gen.* **11**, L213–L217 (1978).
36. Rapaport, D. C. Molecular dynamics study of a polymer chain in solution. *The Journal of Chemical Physics* **71**, 3299–3303 (1979).
37. Smith, S. W., Hall, C. K. & Freeman, B. D. Molecular Dynamics for Polymeric Fluids Using Discontinuous Potentials. *Journal of Computational Physics* **134**, 16–30 (1997).
38. Zhou, Y. & Karplus, M. Folding thermodynamics of a model three-helix-bundle protein. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 14429–14432 (1997).

39. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. Discrete molecular dynamics studies of the folding of a protein-like model. *Folding and Design* **3**, 577–587 (1998).

40. Voegler-Smith, A. & Hall, C. K. Assembly of a tetrameric α-helical bundle: Computer simulations on an intermediate-resolution protein model. *Proteins* **44**, 376–391 (2001).

41. Voet, D. & Voet, J. G. *Biochemistry*. (Wiley, New York, NY, 1990).

42. Lorentz, H. A. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Annalen der Physik* **248**, 127–136 (1881).

43. Nguyen, H. D., Marchut, A. J. & Hall, C. K. Solvent effects on the conformational transition of a model polyalanine peptide. *Protein Science* **13**, 2909–2924 (2004).

44. Rapaport, D. C. The event scheduling problem in molecular dynamic simulation. *Journal of Computational Physics* **34**, 184–201 (1980).

45. Bellemans, A., Orban, J. & Van Belle, D. Molecular dynamics of rigid and non-rigid necklaces of hard discs. *Molecular Physics* **39**, 781–782 (1980).

46. Proctor, E. A. & Dokholyan, N. V. Applications of Discrete Molecular Dynamics in biology and medicine. *Current Opinion in Structural Biology* **37**, 9–13 (2016).

47. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics* **72**, 2384–2393 (1980).

48. Allen, M. P., Tildesley, D. J. & Tildesley, D. J. *Computer Simulation of Liquids*. (Clarendon Pr, Oxford, 2009).

49. Tu, K. XDR Interface for Fortran.

50. Barnett, J. libxdrfile. (2016).

51. Barnett, J. libgmxfort. (2018).

52. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford university press, Oxford, 2017).

53. Bruininks, B. M. H., Wassenaar, T. A. & Vattulainen, I. Unbreaking Assemblies in Molecular Simulations with Periodic Boundaries. *J. Chem. Inf. Model.* **63**, 3448–3452 (2023).

Tables and Figures

Table 2.1 Input parameters that are required for all DMD-PRIME20 simulations.

Item	*Description*
pep1 and pep2	Sequences of the peptides that are simulated in abbreviated alphabetical format (e.g. pep1=GVLYVGS). If the system contains a single peptide sequence, then pep1 and pep2 must be the same as in the 'input.txt'
chain1 and chain2	The number of peptide chains for each peptide component in the simulation box.
boxlength	The length of the simulation box. DMD/PRIME20 uses a cubic box with periodic boundary conditions. boxlength is selected based on the number of peptide chains and the desired concentration: $$boxlength = \left(\frac{(chain1 + chain2) * 1000}{N_A * C}\right)^{\frac{1}{3}} * 10^9$$ where: - N_A is Avogadro's number. - C is peptide concentration in mM
T	Simulation temperature in Kelvin.
coll	The number of collisions for DMD/PRIME20 to finish a cycle and record simulation results.
trajrecord	the number of collisions for DMD/PRIME20 to record a set of coordinates of the whole system to form a frame in a multi-frame trajectory
annealing	The current version allows for an annealing simulation at a default set of temperatures (annealing = 0) or at user-defined temperatures (annealing = 1). If using user-defined temperature, additional parameters must be added to *input.txt*.

Table 2.2 (continued).

Item	*Description*
Additional parameters:	
startingtemp	Starting temperature for the annealing process (in Kelvin)
endingtemp	Ending temperature for the annealing process (in Kelvin)
tempstep	Temperature change for each annealing cycle (in Kelvin)
annealingcoll	Number of collisions for each temperature in the annealing stages. Recommended value is from 100 million to 250 million collisions
seedconfig	The seed for random number generation for initial configuration generation
seedsim	The seed for random number generation to start a simulation

Table 2.3 Required folders must be created at the beginning of a new simulation.

Folder	***Description***
inputs	Stores input data such as temperatures and sequence identities.
checks	Stores configurations for both single chains and for the entire system after initial configuration generation. Users can check these files to see if the simulated system is initiated correctly or not
parameters	Stores the identity of all peptide chains in the system, their sidechains, and the hydrophobicity of each sidechain.
outputs	Stores output log for each simulation cycle.
results	Stores all results data. Most of the data is binary and needs the analysis package in order to be read.
analysis	Stores results from using analysis package.

Table 2.4 Command line arguments for using analysis package.

Argument 1	***Argument 2***	***Argument 3***	***Description***
traj	start_file	end_file	Generate trajectory file
evst	start_file	end_file	Energy vs time
hbvst	start_file	end_file	Number of hydrogen bonds vs time
cluster	collision in billion		Return numbers of monomers, oligomers, and fibrils during simulation
pvsap	collision in billion (underdeveloped. I will write this part later)		Return the number of parallel and antiparallel strands within each β-sheet
pdb	collision in billion		Generate pdb file at any recorded point during the simulation

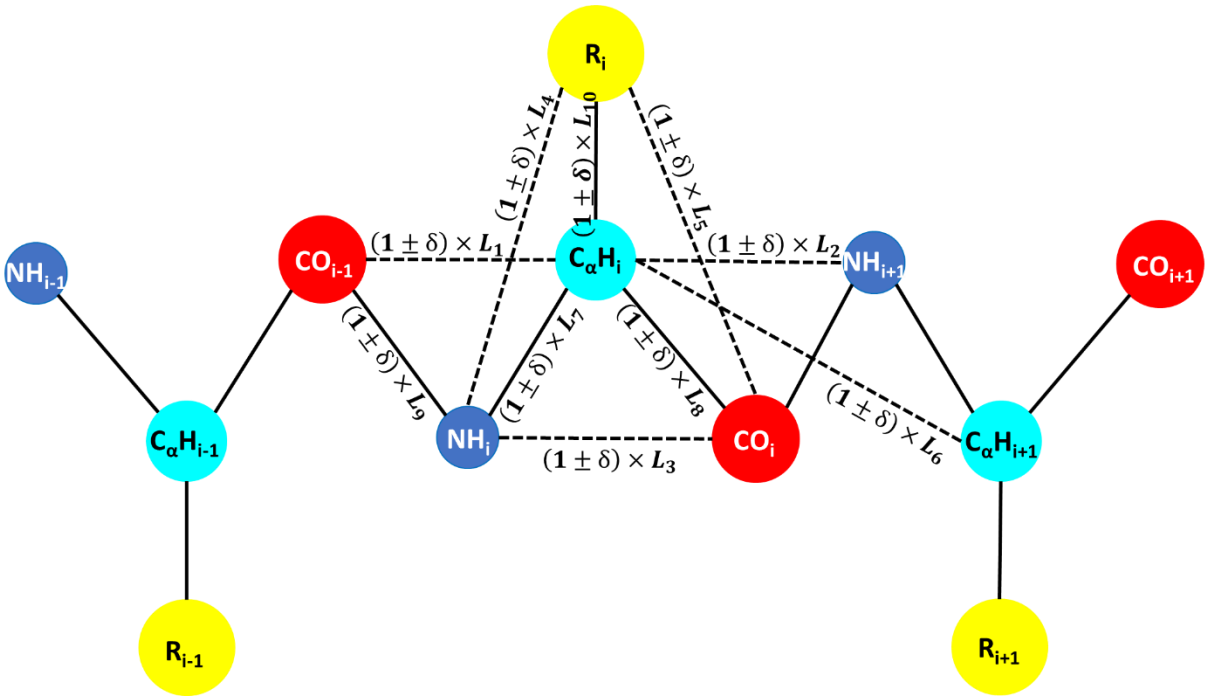


Figure 2.1 Representation of a PRIME20 protein model for a peptide containing 3 amino acids. Each amino acid is represented by four beads: amine NH, alpha carbon CαH, carbonyl CO, and sidechain R; these are colored dark blue, cyan, red, and yellow, respectively. The central amino acid is indicated by a subscript i. The preceding and succeeding amino acids are numbered i-1 and i+1, respectively. Pseudo-bonds are dashed lines, covalent bonds and peptide bonds are solid lines.

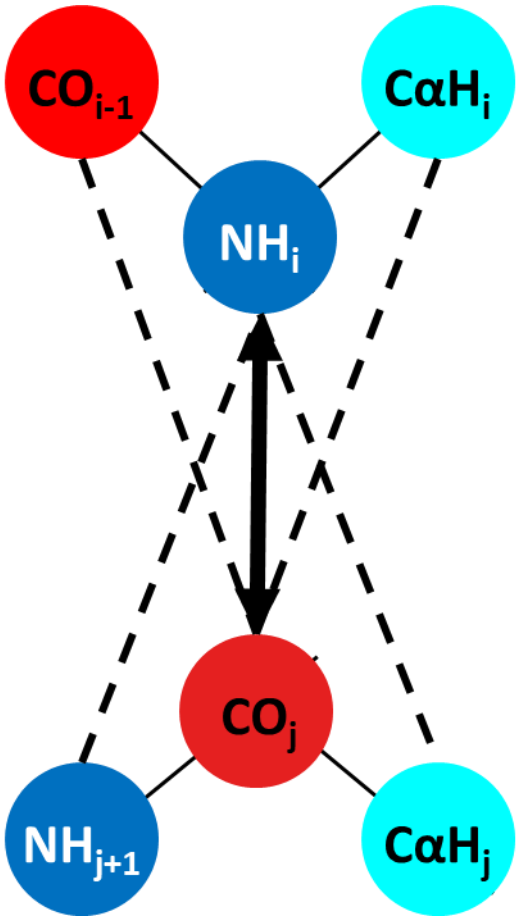


Figure 2.2 NH_i-CO_j hydrogen bond (solid black arrow) formation/dissociation is restricted by the square-shoulder interactions of four auxiliary pairs (dashed line).

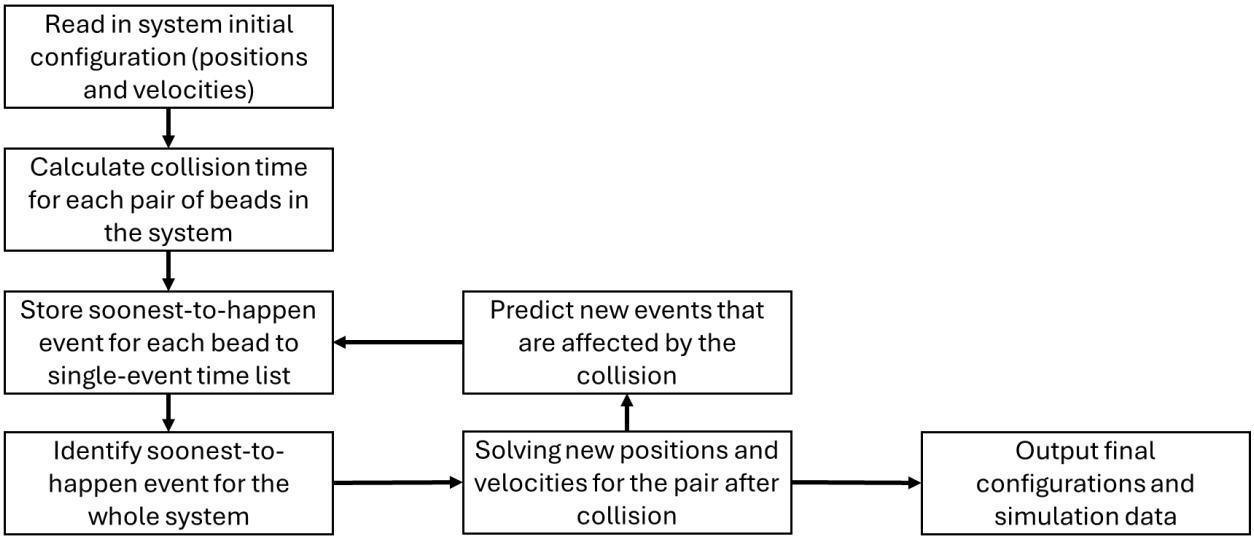


Figure 2.3 Discontinuous molecular dynamics simulation algorithm.

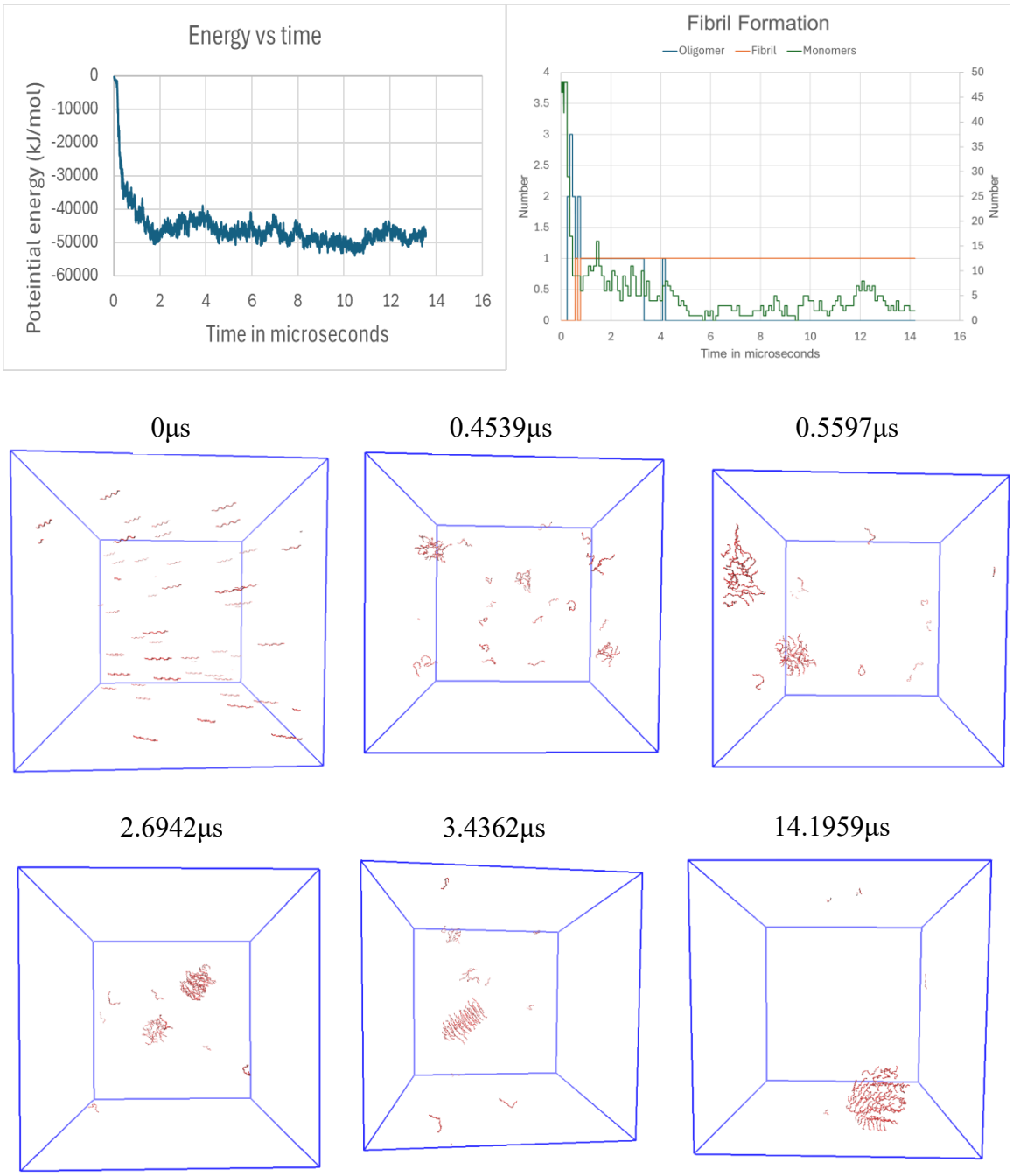


Figure 2.4 Results of analysis using the analysis package in DMD/PRIME20.

Chapter 3 In silico peptide self-assembly reveals the importance of N-terminal motifs and the inhibition mechanism of the mutation L38M in α-synuclein fibrillation

This is essentially a manuscript prepared by Van Nguyen, Sabine M. Ulamec, David J. Brockwell, Sheena E. Radford and Carol K. Hall soon to be submitted to Protein Science.

3.1 Introduction

Amyloid formation of α-Synuclein (αSyn), a presynaptic neuronal protein, is associated with several neurodegenerative diseases that afflict millions of people worldwide[1,2]. These diseases, which are classified as synucleinopathies, include Parkinson's disease (PD), Parkinson's disease dementia, dementia with Lewy bodies[3], multiple system atrophy[4], and others. αSyn fibrils are major constituents of Lewy bodies and Lewy neurites—the cytosolic filamentous inclusions found in neuronal cells of people with synucleinopathies[5]. Since the connection between αSyn amyloid formation and PD pathology was first recognized in 1997[6], researchers have been trying to understand the aggregation pathways, fibrillar structures, and toxicity of αSyn, with the aim of designing potential therapeutics.

Human αSyn is a 140-amino-acid protein whose primary function is to modulate neurotransmitter release[7,8]. Under physiological conditions, αSyn can exist as disordered soluble monomers[9,10], membrane-bound (helical) monomers[11,12], and oligomers[13–15]. Full-length αSyn (αSyn-FL) consists of three distinct regions: the N-terminal (residues αSyn[1-60]), the non-amyloid component (NAC) - now named the central hydrophobic region (residues αSyn[61-95]), and the C-terminal (residues αSyn[96-160]) (**Figure 3.1**). While the highly hydrophobic NAC region forms the fibril core in all αSyn amyloid structures solved to date (see the Amyloid Atlas[16]), the N-terminal and C-terminal regions are known to be critical to controlling the rate of nucleation of αSyn amyloid formation[17–20]. A role for the N-terminal region in promoting αSyn self-assembly into amyloid is supported by the following observations. Of the seven-point mutations associated with familial PD, six are found in αSyn[46-53][18]. Deletion or insertion[21], of two imperfect repeats (KTKEGV) between residues 9 and 30 induces, or prevents, aggregation, respectively. The PreNAC region (αSyn[47-56]) fragment forms steric zipper protofilaments as an isolated peptide[22].

The formation of transient β-hairpins in residues αSyn[37-54] is suggested to be an important element in initiating dimerization; acting as intermolecular binding sites between two monomers[23]. Amyloid formation of αSyn can be prevented by adding the β-wrapin protein, AS69, to a solution of αSyn monomers[24]; this protein enhances the formation of β-hairpins in residues αSyn[37-54] and then forms stable complexes with them, making them inaccessible for fibrillation.

Two motifs in the N-terminal region were shown by Doherty et al.[18] to be important modulators of amyloid formation of full length αSyn (αSyn-FL): P1 (G_{36}VLYVGS_{42}), and P2 (K_{45}EGVVHGVATVAE_{57}). Deleting P1, or P1 and P2, or substituting these regions with alternating glycine-serine sequences, inhibits αSyn aggregation *in vitro*. Deleting P1 prevents amyloid formation of αSyn-FL at neutral pH, while at lower pH deletion of both P1 and P2 is required to prevent amyloid formation, at least on an experimentally accessible timescale. Mutational analysis of P1showed that L38M inhibits the nucleation phase of αSyn-FL amyloid formation starting from monomeric αSyn solutions, without perturbing the ability of the protein to elongate fibril seeds from the WT protein. More recent analyses have reinforced these conclusions, showing that P1 is required for oligomer to fibril conversion[15].

The work presented here aims to uncover the self-assembly mechanisms of the αSyn P1 and P2 motifs, to describe their potential role in amyloid formation of αSyn-FL, and to explain the mechanism by which L38M inhibits amyloid formation of αSyn-FL. We consider a series of αSyn fragments of increasing length: P1 (αSyn[36-42]), P2 (αSyn[45-57]), and P3 (αSyn[36-57]). P3 spans P1 and P2, and includes the residues that link these two regions ([G_{36}VLYVGS_{42}]KT[K_{45}EGVVHGVATVAE_{57}]). In addition to the variant, L38M, two other single-point mutations, L38A and V40A, are considered for comparison, as they do not inhibit αSyn-FL amyloid formation *in vitro*[25]. We also consider P3Next (αSyn[27-57]), which includes

P3 and the nine residues that precede it. This extended sequence was chosen for study because it allows us to investigate the effect of the flanking N-terminal residues on P1 and P2 assembly. All sequences are listed in **Figure 3.1**. Mixtures of P3Next (WT or sequence variants) with a fragment from the αSyn C-terminal region, Cterm120-140 (αSyn[120-140]), are also simulated to investigate interactions between the P1/P2 and C-terminal regions.

3.2 Results and Discussion

3.2.1 Wild-type αSyn Fragments Assemble into Fibrillar Structures in All Simulations.

The spontaneous self-assembly of αSyn peptide fragments was investigated using discontinuous molecular dynamics (DMD) simulation with the coarse-grained PRIME20 (Protein Intermediate-Resolution Model) force field. Systems containing 48 P1 peptides or 24 P2, P3, or P3Next peptides were simulated individually under conditions which allowed the fibrillation process to be completed within a reasonable time frame, resulting in well-ordered secondary structures. At least three independent simulations were performed for each peptide under each condition explored, hence a total of >60 simulations was performed. Simulation concentrations and temperatures are listed in **Table 3.1**. Each simulation started with a system of peptides in random coil configurations. The systems were equilibrated at fixed temperatures, chosen to have enough energy to escape local minima but not so much as to prevent fibrillation. In the DMD/PRIME20 simulations, all the wild-type fragments (P1, P2, P3, and P3Next) assembled into fibril-like structures with varying types of internal organization. In our simulations, a fragment is considered to be "aggregated" if 2/3 of the runs yield a well-ordered aassembly rich in β-sheet structure.

The simulations of P1-WT show that this sequence favors parallel arrangements of β-strands when forming a β-sheet, which then tend to stack in an antiparallel arrangement. Snapshots

of P1-WT fibrillation simulations are shown in **Figure 3.2**. At the beginning of the simulation, all of the P1-WT peptides were in random coil configurations. As time evolved, seven peptides quickly formed a small, disordered cluster (0.24μs) with two parallel layers stacking in antiparallel arrangement. This cluster then accumulated and formed a larger disordered oligomer (0.35μs). The peptides in this oligomer then rearranged themselves into a five-layer β-sheet-rich conformation (4.08μs). The five-layer structure quickly dissociated to a smaller four-layer structure (4.18μs). By the end of the simulation (~14μs), the P1-WT systems formed multi-layer fibrils containing an average of 86±0.54% parallel β-strands. Antiparallel strands appear randomly within the P1-WT β-sheets.

By contrast with P1 peptides, the P2 peptides formed a single-layer parallel β-sheet which curved into two conformations: a U-shaped structure with a hydrophobic core and a S-shaped structure (**Figure 3.3**a). The two conformations both prefer to elongate to form a long fibril instead of stacking into multiple layers. The P2 curved fibril can be explained by the glycine at residue 51 in the middle of the peptide, and the formation of a strong hydrophobic core toward its center, with charged residues facing outward toward the solvent. As a result, fragments that contain P2 have a high tendency to form bent structures (e.g. β-hairpin). In comparison, the PreNAC fragment (αSyn[47-56]) was shown previously to form a nanocrystal structure with antiparallel stacked β-sheets[22]. Each β-sheet contained in-register parallel β-strands with a slight bend at G51 on each strand. The DMD/PRIME20 simulations of P2 are consistent with these experimental results in that both predict an ordered parallel arrangement of β-sheets that bent at G51.

Preliminary simulation results show that the P3-WT systems formed three different types of assemblies: single flat β-sheet containing a mixture of parallel and antiparallel β-strands, S-shaped structure with two stacked parallel β-sheets connected by a single peptide that participates

in both sheets, and G-shaped β-fibril containing a single-layer parallel β-sheet. **Figure 3.3** shows snapshots of the three types of aggregates formed in the four simulations (**Figure 3.3**b), side views that show the conformations of the S-shaped and G-shaped peptide strands (**Figure 3.3**c), and plots of their total interaction energy versus the time in the four runs (**Figure 3.3**d). The parallel S-shape and G-shape structures have stronger interaction energies than the flat β-sheet with mixed structure, hence they are more stable. Although αSyn-FL forms polymorphic protofilaments in fibril assembly reactions *in vitro*, parallel β-strand arrangements within a β-sheet are dominant among all the identified fibril filaments reported in the Amyloid Atlas[16], suggesting that the parallel structures are more energetically stable. The lag-time of formation of the S-shape structure (2.64 μs) is similar to the G-shape structure (2.73μs). By comparison, the lag-time for P3-WT to form the flat β-sheet with mixed parallel and antiparallel β-strands is 6.3μs, approximately 2.3 times longer than that for the parallel structures. Videos demonstrating aggregation process during simulations are included in Supplementary Movies 2, 3 and 4 for mixed, S-shaped and G-shaped fibril respectively. At the time of rewriting this thesis, additional simulations are running for P3-WT to gain insight the peptide polymorphism. To ensure that we do not over-predict aggregation of P3, we also simulated a different 22-residue fragment from the N-terminal, C1ext (αSyn[14-35]**)** using the same conditions as for the P3 system. C1ext is an extension of fragment C1 that was predicted to have low amyloid propensity by the Zyggregator method[18]. C1ext did not aggregate in any of our simulations, consistent with the Zyggregator prediction. The simulation results for C1ext are shown in **Supplementary Figure 3.9**.

P3Next-WT preliminary simulations resulted in three different β-sheet structures, all of which contain β-hairpin motifs. At the end of run 1, P3Next-WT peptides formed a single-layer fibril that consists only of β-hairpins. This run had the shortest lag-time (~2μs) of the three

simulations. At the end of run 2, P3Next-WT formed a fibril comprised mainly of β-hairpins mixed with a couple of straight β-strands. With a lag-time of *ca.* 6μs, run 2 of P3Next-WT took longer to initiate aggregation than run 1. At the end of run 3, a double-layer fibrillar structure containing mainly straight β-strands was observed. There is only one β-hairpin on one layer and a curved strand connects the two layers. This run had a long lag phase, *ca*. 8.56μs. Snapshots of the final fibrillar structures of P3Next-WT are shown in **Figure 3.4**a Run 1 with the largest number of β-hairpins has the shortest lag-phase, suggesting that β-hairpins promote intermolecular interactions between monomers kinetically, thereby promoting the formation of oligomers. Run 1 of P3Next-WT simulation is recorded in Supplementary Movie 5. At the time of writing this thesis, additional simulations of P3Next-WT are also running.

The fibrils that are observed in our simulations of P1, P2, P3 and P3Next share a common structure with those of the P1/P2 region in some αSyn-FL fibrils seen *in vitro*[26–29]; peptides in the fibrils likely formed curved β-strands with bends likely at K43-K45 and G51-T54. Our observation of polymorphic fibril structures for P3 is also consistent with the polymorphism of αSyn-FLfibrils observed *in vitro*[30,31]. In both αSyn-FL experiments and DMD/PRIME20 simulations of P3, parallel fibrils are the most prevalent and the most stable of the structures observed. This is consistent with the finding of Wu and Baum[32] that fibrillation of αSyn-FL takes longer when antiparallel (as opposed to parallel) transient interactions are detected as well as the finding of Zhou and Kurouski[33] that the antiparallel oligomers need to rearrange to form stable parallel fibrils.

P3Next, which is the longest peptide that we simulated, shows that this region has a high tendency to form β-hairpins in all simulation runs (**Figure 3.4**a). A zoom-in picture of β-hairpin structures in a P3Next fibril is shown in **Figure 3.4**b. In addition, the rate of self-assembly into these structures in the simulation is faster when there are more β-hairpins (Run 1) as opposed to

few β-hairpins in the final fibril structure (Runs 2 and 3) (**Figure 3.4**c). This observation agrees with the finding by Yu et al.[23] that β-hairpins are formed in the P1/P2 region, and act as an intermolecular binding element to promote formation of on-pathway parallel oligomers.

3.2.2 Effect of Single-point Mutations L38A, V40A, and L38M on the Self-assembly of Fragments P1, P3, and P3Next.

We next investigated the effect of single point amino acid substitutions, L38A, V40A, and L38M on amyloid formation of αSyn. Previous experiments[25] showed that the substitutions L38A and V40A have no effect on the kinetics of amyloid formation of αSyn-FL, but that L38M inhibits αSyn-FL amyloid assembly. As no experiments have been performed on the equivalent fragments (substitutions in P1, P2, P3, and P3Next), DMD/PRIME20 simulations were performed on these fragments containing these amino acid changes (**Table 3.2**) under identical conditions to those used for the wild-type equivalents (**Table 3.1**).

The simulations of all the variant P1 systems (L38A, V40A, or L38M) resulted in multi-layer fibrils, but with different structures and total percentages of parallel or antiparallel strands. P1-L38A consistently formed three-layer fibrils in which each layer did not stack on top of each other. Instead, each P1-L38A β-sheet layer curved out at the mutated alanine and the first two layers connected at valine 40 in parallel arrangement. As both layers were curved, there was empty space between two layers that allowed the third layer to stack in via the hydrophobic bonds at valine 37 and 40. The aggregation process of P1-L38A is shown by a series of snapshots during simulation time in **Supplementary Figure 3.10**. P1-V40A consistently formed two-layer fibrils. The P1-V40A fibrils are highly ordered and contain two β-sheet layers stacked in an antiparallel fashion. P1-L38M also formed multi-layer fibrils, but the number of parallel strands is lower than in other P1 variants. Snapshots of the final configuration for each of the mutated sequences are

displayed in **Figure 3.5a**, with the final configuration of P1-WT is included for comparison. Plot of average percentages of antiparallel residues are shown in **Figure 3.5**b. Quantitative calculations shows that P1-L38M fibrils have the highest average antiparallel residues which is 35.26±3.22% comparing to 14.08±0.53% antiparallel residues in P1-WT fibrils. P1-L38A and P1-V40A fibrils both have lower antiparallel residues than in P1-WT fibrils which are 5.19±1.48 and 8.85±1.97, respectively.

We next investigated how the self-assembly of sequence P3 is affected by the sequence alterations (L38M, L38A, and V40A). In all four simulation runs, P3-L38M formed single-layer fibrils containing a mixture of parallel and antiparallel β-strands without any specific pattern, similar to P3-WT in run 1. The average lag-time for P3-L38M fibrillation was 12.5μs, which is *ca*. 3.5 times longer than in the P3-WT simulations. In three out of four runs, P3-L38A did not aggregate even though those simulations were extended to ~40μs. The final snapshots of P3-L38A simulations in **Supplementary Figure 3.11** show that P3-L38A stays mostly in random coils at ~38μs. The results obtained for P3-V40A were inconclusive as two runs formed mixed fibrils, while in the other two runs all peptides remained as random coils (**Supplementary Figure 3.11)**. Our simulation results on all sequence variants of P3 differ from the same variants in αSyn-FL experimental results by Ulamec *et al*.[25] In our simulations, the L38A and V40A amino acid substitutions significantly reduced fibrillation propensity of the P3 variants , but they had no/little effect in solution of αSyn-FL, as shown in **Table 3.2**. In addition, P3-L38M consistently aggregated in all simulations, whereas L38M inhibited αSyn-FL in vitro. This prompted us to consider a peptide that extends beyond P3 toward the N-terminal, P3Next (αSyn[27-57) (**Figure 3.1**). Note that extension at the C-terminus was not considered as this region incorporates NAC.

The increase in peptide length allows P3Next to take into account the effect of its N-terminal flanking residues on its behavior.

Simulations of the single-point amino acid substitutions in P3Next showed that L38M reduces the formation of β-hairpins in fibril structures. As the amyloid propensity of each sequence variant in our simulations is different (final snapshots are shown in **Supplementary Figure 3.12**), the ratio of the number of β-hairpins to the total number of β-strands in the fibril at the end of the simulation was used to quantitate the effect of the sequence changes on β-hairpin formation. In contrast to the inhibition seen in experiments of αSyn-FL with L38M, all simulations of P3Next-L38M result in fibril formation. However, the average ratio of the number of β-hairpins to the total number of β-strands is 0.24±0.05 for P3Next-L38M, smaller than that in the fibrils of P3Next-L38A (0.26±0.07), P3Next-V40A (0.33±0.04), and P3Next-WT (0.34±0.11) (**Figure 3.6**). The finding that P3Next-L38M has a lower propensity to form β-hairpins compared with P3Next-L38A and P3Next-is consistent with previous reports which have suggested β-hairpin formation involving P1/P2 to be an important factor in initiating self-assembly of αSyn-FL[23,24]. However, the relationship must be more complex, since P3Next-L38A also forms fewer β-hairpins, yet this substitution does not affect the rate of amyloid assembly[25].

3.2.3 L38M Facilitates the Interaction of P3Next with the C-terminal, Reducing Formation of Fibrillation-prone Conformations.

Long-range interactions between the basic N-terminal region and the acidic C-terminal region of αSyn are known to be critical for controlling its assembly into amyloid[34–36]. To examine how such interactions may alter the local structural propensities of P1/P2 region, DMD/PRIME20 simulations of mixtures of P3Next and the C-terminal 21 amino acids of aSyn (residues 120-140) were performed. Our goal was to gain insight into the possible impact of amino acid substitutions

in the P1 region on these long-range interactions in α*Syn-FL*. Each simulation contained 12 peptides of P3Next (WT or mutated) and 12 peptides of C-terminal region 120-140 (Cterm120-140). Simulations were performed at a *concentration of 10 mM and reduced temperature of 0.195,* to mimic the simulations of the P3Next variants alone.

In the presence of Cterm120-140, P3Next-L38M formed mixed fibrils that were mainly antiparallel and had a low ratio of the number of β-hairpins to the total number of β-strands. This peptide was the only sequence that formed a single-layer β-sheet with Cterm120-140 entwined (**Figure 3.7**a,b). In contrast, amyloid-like fibrils were not observed in mixtures that contained Ctem120-140 with P3Next-WT, -L38A, or -V40A. In the simulation of P3Next-L38M and Cterm120-140 at 7.44μs, all peptides are still in random coil configurations (**Figure 3.7**b). At 7.84μs, a small antiparallel β-sheet containing only P3Next-L38M is observed, consistent with the high tendency of L38M to form an antiparallel β-sheet as seen in simulations of this sequence in P2Next alone. At 7.97μs, a Cterm120-140 peptide binds to the outer strand on the L38M tetramer in a head-to-head orientation. The binding region involved residues 34-37 in P3Next-L38M, just before position 38. At 8.10μs, the β-strands formed by Cterm120-140 have zipped up and straightened out into a mixed β-sheet. At the end of the simulation (19.67μs (**Figure 3.7**a), the fibrils in the mixture mainly consist of straight β-strands. The ratio of the number of β-hairpins to the total number of β-strands is 0.062. **Figure 3.7**c shows the differences between fibrils that formed in a pure P3Next-L38M system (left) and in a mixture of P3Next-L38M and Cterm120-140 (right). The pure P3Next-L38M fibril contains multiple β-hairpins while the mixture fibril has only one β-hairpin in its structure. The aggregation process can be seen in Supplementary Movie 6.

DMD/PRIME20 simulations of mixtures of P3Next and the C-term120-140 peptides show that L38M increases the likelihood of interaction between the N-terminal and C-terminal domains and reduces the likelihood of β-hairpin formation in P1/P2 region, compared to the equivalent simulations involving P3Next-WT, -L38A, and -V40A. Although in our simulations the interactions between P3Next-L38M and the C-terminal 21 residues are necessarily intermolecular, such interactions could be intra- or inter-molecular in αSyn-FL (both such interactions have been observed experimentally[18,32,37]. Nonetheless, these contacts would be expected to contribute to a reduction in the fibrillation propensity of αSyn-FL as they both drive αSyn toward forming less fibrillation-prone structures (**Figure 3.8**). Such a mechanism has been proposed hitherto (**Figure 3.8**a)[23], with the new results presented here suggesting a possible mechanism whereby the L38M substitution could disfavor amyloid formation (**Figure 3.8**b). Such a mechanism is consistent with recent all-atom MD simulations of αSyn-FL by Onishi et al.[36] which showed that interactions within the P1/P2 region increase when the C-terminal is truncated, and those of Farzadfard et al.[20], which showed experimentally that deletion of the C-terminal enhances the rate of αSyn amyloid formation. In addition to preventing β-hairpin formation, our simulations suggest that L38M enhances long-range intramolecular interactions between the N- and C-terminal domains. This interaction was determined *in silico* to maintain αSyn-FL in a monomeric conformation by preventing the monomer from unfolding and hiding hydrophobic residues from solvent[36]. This interpretation is consistent with the conclusion by Ulamec et al.[25] that the L38M substitution interrupts the nucleation stage of αSyn-FL amyloid formation, trapping the protein in non-productive oligomers.

Finally, the mutation L38M in αSyn-FL appears to increase the likelihood of intermolecular interactions between P1/P2 and the C-terminal 21 residues. This could induce a

head-to-tail interaction between two monomers (**Figure 3.8**b), increasing the propensity to form antiparallel oligomers. Such antiparallel oligomers would be expected to reduce the rate of amyloid formation since they would require presumably slow rearrangement to form fibrillation-prone parallel oligomers (**Figure 3.8**b). They would also prevent fibrillation by forming unproductive oligomers trapped in an antiparallel state (**Figure 3.8**b). Such conclusions are consistent with experimental findings that antiparallel oligomers are often found in toxic prefibrillar oligomers rather than in αSyn-FL amyloid fibrils[17,32,38,39]. Santos et al.[15] recently reported that that off-pathway oligomers are stabilized when phenol-soluble modulin α3 (PSMα3) binds to the P1/P2 region of αSyn, and that the P1/P2 region is critical for αSyn-FL to transition from oligomers to fibrils. Taken together, therefore, we suggest that L38M might prevent the oligomer-to-fibril conversion by enhancing the P1/P2 interaction with the C-terminal region and reducing the accessibility and flexibility of the P1/P2 region necessary for conversion of oligomers into an amyloid state.

3.3 Conclusion

Although it is generally accepted that αSyn is involved in many neurodegenerative diseases, the mechanism of its assembly into amyloid and the role of the N- and C-terminal domains in enhancing or inhibiting fibrillation remain unclear. Combining experimental data with computational methods can provide new insights into the possible molecular interactions that drive or inhibit amyloid formation, including the effects of amino acid substitutions on the different steps in amyloid formation processes. For αSyn, a key structural change required to initiate the oligomerization process is thought to be the formation of β-hairpins within the PreNAC (P2) region[23]. Conversion/growth of the oligomer to the fibril is accelerated by the presence of parallel as opposed to antiparallel conformations[32,39]. We have shown that deleting P1 and P2[18], or

mutating Leu 38 to Met in P1[25], inhibits amyloid formation of αSyn-FL by a possible combination of different effects. In this paper, we use DMD/PRIME20 simulations to provide insight at the molecular level of the impact of these regions on the fibrillation of αSyn fragment variants (P1, P2, P3, P3Next, including both WT and mutated sequences). Our conclusions are based on the preliminary simulation results up to this point. At the time of writing this thesis, additional simulations for all peptides variants considered in this chapter are still running. We will have updates on our results when the manuscript is published.

Overall, we draw two conclusions from evaluating the fibrillar structures that formed in the simulations of the peptide variants (WT and mutated) in both homogeneous and mixed (P3Next variants with C-terminal fragment αSyn[120-140]) systems. First, WT fragments have a high propensity to form fibrillation-prone structures such as parallel β-sheets and β-hairpins. Second, the mechanism by which L38M prevents the aggregation of αSyn-FL is not by reducing the aggregation propensity of the sequence itself, but by altering its propensity to form the fibrillation-prone structures and the interactions that occur between the P1 region and more distant residues in the sequence in the early stages of amyloid formation. The homogeneous system simulations showed that the L38M substitution in the peptide fragment reduces the formation of amyloid-prone structures (parallel β-strands for P1 variants and β-hairpins for P3Next variants) compared to the WT fragments. In contrast to L38M, the other single-point substitutions, P1-L38A and P1-V40A, increased parallel β-fibril formation for P1 variants and had higher β-hairpin ratios in their P3Next fibrils. The mixed system simulations showed that L38M enhances interdomain interactions between P1/P2 region and the last 21 residues in the C-terminal. These interdomain interactions disrupt the formation of amyloid-prone structures, potentially altering the protein fibrillation

pathway by stabilizing monomeric conformations or promoting less amyloidogenic oligomer structures.

3.4 Materials and Methods

3.4.1 Discontinuous Molecular Dynamics (DMD) and PRIME20 Force Field

DMD simulation is a fast alternative to classical molecular dynamics (MD) that uses simple discontinuous potentials to describe the interaction between two particles. DMD is event-driven, as opposed to molecular dynamics, which is time-driven. DMD simulations are performed in the canonical ensemble; the system's temperature is kept constant using the Andersen thermostat[40]. PRIME20 is a coarse-grained (CG) implicit-solvent intermediate-resolution protein model that was developed by the Hall group[41–44]. The model was designed to be used with DMD simulations. PRIME20 contains geometric and energetic parameters that describe the sidechain-sidechain interactions of all 20 amino acids. All parameters, bond lengths, bead diameters, etc., were obtained by fitting the native state structures of 711 proteins from the protein data bank (PDB) using a knowledge-based approach. Each amino acid is represented by four beads: one each for the amino group (NH), the alpha carbon (CαH), the carbonyl group (CO), and the side chain (R). The diameter of each bead is chosen to be a reasonable estimate of the size of the group of atoms it represents. The model is designed for simulations of L-isomer polypeptides in the trans-configuration. Six pseudo-bonds are implemented to maintain this configuration and to keep the Φ-Ψ angles within the observed ranges on the Ramachandran plot. The bond length is allowed to fluctuate within 2.375% of its assigned value. All peptides generated and simulated by DMD/PRIME20 are capped by neutralizing the charges at both termini into the force field. Therefore, DMD/PRIME20 does not add physical capping groups such as acetyl (ACE) and N-methylamide (NME) to the termini as MD simulations do.

3.4.2 Relating Reduced Units to Real Units in DMD/PRIME20 Simulations

The PRIME20 model measures time and temperature in terms of reduced units. The real temperature (T, Kelvin) is a linear function of the reduced temperature in PRIME20. The correlation T* = (T+115.79)/2288.46 was obtained by matching the folding temperature of alanine-rich polypeptides in DMD/PRIME20 simulations to experimental values[45]. Conversion to real-time was performed by comparing self-diffusion coefficients of specific peptides in PRIME20 simulations to simulations of the same peptides computed in atomistic simulation at the same concentration. Wang et al. did this conversion using Aβ(16-22) and found that 1 unit of the reduced time in PRIME20 equals approximately 0.96ns in real-time[46].

3.4.3 Simulation Procedure

All DMD/PRIME20 simulations started from random coil conformations; the peptides aggregated spontaneously to form equilibrium structures[43,47,48]. The simulation concentrations and temperatures were high compared to those in the experiments. This can be attributed to two factors: 1. The run temperature for each peptide was selected so that the system formed ordered fibrillar structures in a reasonable computational time frame. 2. The temperature had to be high enough to prevent the system from being trapped in its local minima, but not so high that the system remained in the solution phase. Each simulation started at a high temperature of 1028K. Once all the peptides in the simulated system were relaxed into random coils, the temperature was reduced gradually to the desired simulation temperature. The production simulation then proceeded for at least 15μs. For systems that didn't aggregate or hadn't formed a stable structure, simulations were extended to over 40μs. P1, which is only 7 residues long, was simulated in systems with 48 peptide chains (twice the size of other peptide systems) so that the β-sheet alignment and stacking in P1 fibril could be better observed. Details of simulation systems are listed on **Table 3.1**.

3.5 Acknowledgements

This research was supported by funds from the National Science Foundation (Grants CBET 1835838 and CBET1743432). SMU was funded by Wellcome (215062/Z/18/Z). SER is a Royal Society Professorial Research Fellow (RSRP/R1/211057).

3.6 Data Availability

Data is available in the manuscript and/or supporting information. The PDB files of final conformations, the initial configurations, and the energy profiles of all peptides, as well as the Supplementary Movies 1-6 are available at:

https://github.com/VanNguyen23/Alpha-Synuclein-N-terminal-Fragments

The DMD/PRIME20 software is available for downloading at:

https://github.com/CarolHall-NCSU-CBE/Serial-DMD-PRIME20

3.7 References

1. Meade, R. M., Fairlie, D. P. & Mason, J. M. Alpha-synuclein structure and Parkinson's disease – lessons and emerging principles. *Mol Neurodegeneration* **14**, 29 (2019).
2. Oliveira, L. M. A. *et al.* Alpha-synuclein research: defining strategic moves in the battle against Parkinson's disease. *npj Parkinsons Dis.* **7**, 65 (2021).
3. McKeith, I. G. *et al.* Diagnosis and management of dementia with Lewy bodies: Third report of the DLB consortium. *Neurology* **65**, 1863–1872 (2005).
4. Grazia Spillantini, M. *et al.* Filamentous α-synuclein inclusions link multiple system atrophy with Parkinson's disease and dementia with Lewy bodies. *Neuroscience Letters* **251**, 205–208 (1998).
5. Spillantini, M. G., Crowther, R. A., Jakes, R., Hasegawa, M. & Goedert, M. α-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6469–6473 (1998).
6. Spillantini, M. G. *et al.* α-Synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).
7. Cabin, D. E. *et al.* Synaptic Vesicle Depletion Correlates with Attenuated Synaptic Responses to Prolonged Repetitive Stimulation in Mice Lacking α-Synuclein. *J. Neurosci.* **22**, 8797–8807 (2002).
8. Murphy, D. D. & Rueter, S. M. Synucleins Are Developmentally Expressed, and α-Synuclein Regulates the Size of the Presynaptic Vesicular Pool in Primary Hippocampal Neurons. *J. Neurosci.* **20**, 3214–3220 (2000).
9. Fauvet, B. *et al.* α-Synuclein in Central Nervous System and from Erythrocytes, Mammalian Cells, and Escherichia coli Exists Predominantly as Disordered Monomer. *Journal of Biological Chemistry* **287**, 15345–15364 (2012).

10. Burré, J. *et al.* Properties of native brain α-synuclein. *Nature* **498**, E4–E6 (2013).
11. Eliezer, D., Kutluay, E., Bussell, R. & Browne, G. Conformational properties of α-synuclein in its free and lipid-associated states 1 1Edited by P. E. Wright. *Journal of Molecular Biology* **307**, 1061–1073 (2001).
12. Ulmer, T. S., Bax, A., Cole, N. B. & Nussbaum, R. L. Structure and Dynamics of Micelle-bound Human α-Synuclein. *Journal of Biological Chemistry* **280**, 9595–9603 (2005).
13. Lorenzen, N. *et al.* The Role of Stable α-Synuclein Oligomers in the Molecular Events Underlying Amyloid Formation. *J. Am. Chem. Soc.* **136**, 3859–3868 (2014).
14. Dear, A. J. *et al.* Kinetic diversity of amyloid oligomers. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12087–12094 (2020).
15. Santos, J. *et al.* A Targetable N-Terminal Motif Orchestrates α-Synuclein Oligomer-to-Fibril Conversion. *J. Am. Chem. Soc.* jacs.4c02262 (2024) doi:10.1021/jacs.4c02262.
16. Sawaya, M. R., Hughes, M. P., Rodriguez, J. A., Riek, R. & Eisenberg, D. S. The expanding amyloid family: Structure, stability, function, and pathogenesis. *Cell* **184**, 4857–4873 (2021).
17. Celej, M. S. *et al.* Toxic prefibrillar α-synuclein amyloid oligomers adopt a distinctive antiparallel β-sheet structure. *Biochemical Journal* **443**, 719–726 (2012).
18. Doherty, C. P. A. *et al.* A short motif in the N-terminal region of α-synuclein is critical for both aggregation and function. *Nat Struct Mol Biol* **27**, 249–259 (2020).
19. Sorrentino, Z. A. & Giasson, B. I. The emerging role of α-synuclein truncation in aggregation and disease. *Journal of Biological Chemistry* **295**, 10224–10244 (2020).
20. Farzadfard, A. *et al.* The C-terminal tail of α-synuclein protects against aggregate replication but is critical for oligomerization. *Commun Biol* **5**, 123 (2022).

21. Kessler, J. C., Rochet, J.-C. & Lansbury, P. T. The N-Terminal Repeat Domain of α-Synuclein Inhibits β-Sheet and Amyloid Fibril Formation. *Biochemistry* **42**, 672–678 (2003).
22. Rodriguez, J. A. *et al.* Structure of the toxic core of α-synuclein from invisible crystals. *Nature* **525**, 486–490 (2015).
23. Yu, H., Han, W., Ma, W. & Schulten, K. Transient β -hairpin formation in α -synuclein monomer revealed by coarse-grained molecular dynamics simulation. *The Journal of Chemical Physics* **143**, 243142 (2015).
24. Mirecka, E. A. *et al.* Sequestration of a β-Hairpin for Control of α-Synuclein Aggregation. *Angew Chem Int Ed* **53**, 4227–4230 (2014).
25. Ulamec, S. M. *et al.* Single residue modulators of amyloid formation in the N-terminal P1-region of α-synuclein. *Nat Commun* **13**, 4986 (2022).
26. Guerrero-Ferreira, R. *et al.* Two new polymorphic structures of human full-length alpha-synuclein fibrils solved by cryo-electron microscopy. *eLife* **8**, e48907 (2019).
27. Zhao, K. *et al.* Parkinson's disease associated mutation E46K of α-synuclein triggers the formation of a distinct fibril structure. *Nat Commun* **11**, 2643 (2020).
28. Lövestam, S. *et al.* Seeded assembly *in vitro* does not replicate the structures of α-synuclein filaments from multiple system atrophy. *FEBS Open Bio* **11**, 999–1013 (2021).
29. Frieg, B. *et al.* Quaternary structure of patient-homogenate amplified α-synuclein fibrils modulates seeding of endogenous α-synuclein. *Commun Biol* **5**, 1040 (2022).
30. Guzzo, A. *et al.* Wild-Type α-Synuclein and Variants Occur in Different Disordered Dimers and Pre-Fibrillar Conformations in Early Stage of Aggregation. *Front. Mol. Biosci.* **9**, 910104 (2022).

31. Li, B. *et al.* Cryo-EM of full-length α-synuclein reveals fibril polymorphs with a common structural kernel. *Nat Commun* **9**, 3609 (2018).
32. Wu, K.-P. & Baum, J. Detection of Transient Interchain Interactions in the Intrinsically Disordered Protein α-Synuclein by NMR Paramagnetic Relaxation Enhancement. *J. Am. Chem. Soc.* **132**, 5546–5547 (2010).
33. Zhou, L. & Kurouski, D. Structural Characterization of Individual α-Synuclein Oligomers Formed at Different Stages of Protein Aggregation by Atomic Force Microscopy-Infrared Spectroscopy. *Anal. Chem.* **92**, 6806–6810 (2020).
34. Cho, M. *et al.* Structural characterization of α-synuclein in an aggregation prone state. *Protein Science* **18**, 1840–1846 (2009).
35. Stephens, A. D., Zacharopoulou, M. & Kaminski Schierle, G. S. The Cellular Environment Affects Monomeric α-Synuclein Structure. *Trends in Biochemical Sciences* **44**, 453–466 (2019).
36. Onishi, N. *et al.* Flanking Domains Modulate α-Synuclein Monomer Structure: A Molecular Dynamics Domain Deletion Study. Preprint at https://doi.org/10.1101/2024.03.23.586267 (2024).
37. Bertoncini, C. W. *et al.* Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1430–1435 (2005).
38. Chen, S. W. *et al.* Structural characterization of toxic oligomers that are kinetically trapped during α-synuclein fibril formation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, (2015).
39. Williams, J. K., Yang, X. & Baum, J. Interactions between the Intrinsically Disordered Proteins β-Synuclein and α-Synuclein. *Proteomics* **18**, 1800109 (2018).

40. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics* **72**, 2384–2393 (1980).
41. Cheon, M., Chang, I. & Hall, C. K. Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins* **78**, 2950–2960 (2010).
42. Voegler-Smith, A. & Hall, C. K. α-Helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins* **44**, 344–360 (2001).
43. Nguyen, H. D. & Hall, C. K. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16180–16185 (2004).
44. Nguyen, H. D., Marchut, A. J. & Hall, C. K. Solvent effects on the conformational transition of a model polyalanine peptide. *Protein Science* **13**, 2909–2924 (2004).
45. Wang, Y., Shao, Q. & Hall, C. K. N-terminal Prion Protein Peptides (PrP(120–144)) Form Parallel In-register β-Sheets via Multiple Nucleation-dependent Pathways. *Journal of Biological Chemistry* **291**, 22093–22105 (2016).
46. Wang, Y., Shao, Q. & Hall, C. K. N-terminal prion protein peptides (PrP(120–144)) form parallel in-register β-sheets via multiple nucleation-dependent pathways. *Journal of Biological Chemistry* **292**, 20655 (2017).
47. Cheon, M., Chang, I. & Hall, C. K. Spontaneous Formation of Twisted Aβ16-22 Fibrils in Large-Scale Molecular-Dynamics Simulations. *Biophysical Journal* **101**, 2493–2501 (2011).
48. Latshaw, D. C., Cheon, M. & Hall, C. K. Effects of Macromolecular Crowding on Amyloid Beta (16–22) Aggregation Using Coarse-Grained Simulations. *J. Phys. Chem. B* **118**, 13513–13526 (2014).

Tables and Figures

Table 3.1 Simulation conditions for all sequences.

Sequences	T (K)	Concentration	Number of peptides
P1	310	10mM	48
P2	307	10mM	24
P3	330	10mM	24
P3Next	330	10mM	24
Mixture (P3Next – Cterm120-140)	330	10mM	12 P3Next + 12 Cterm120-140

Table 3.2 PRIME20 simulation results for all αSyn sequences. The term "aggregate" in the table means β-sheet formation.

Sequences	PRIME20 Simulations	Experiments on αSyn-FL
P1-L38A	aggregate	aggregate
P1-V40A	aggregate	aggregate
P1-L38M	aggregate	no aggregate
P3-L38A	no aggregate	aggregate
P3-V40A	inconclusive	aggregate
P3-L38M	aggregate	no aggregate
P3Next-L38A	aggregate	aggregate
P3Next-V40A	aggregate	aggregate
P3Next-L38M	aggregate	no aggregate

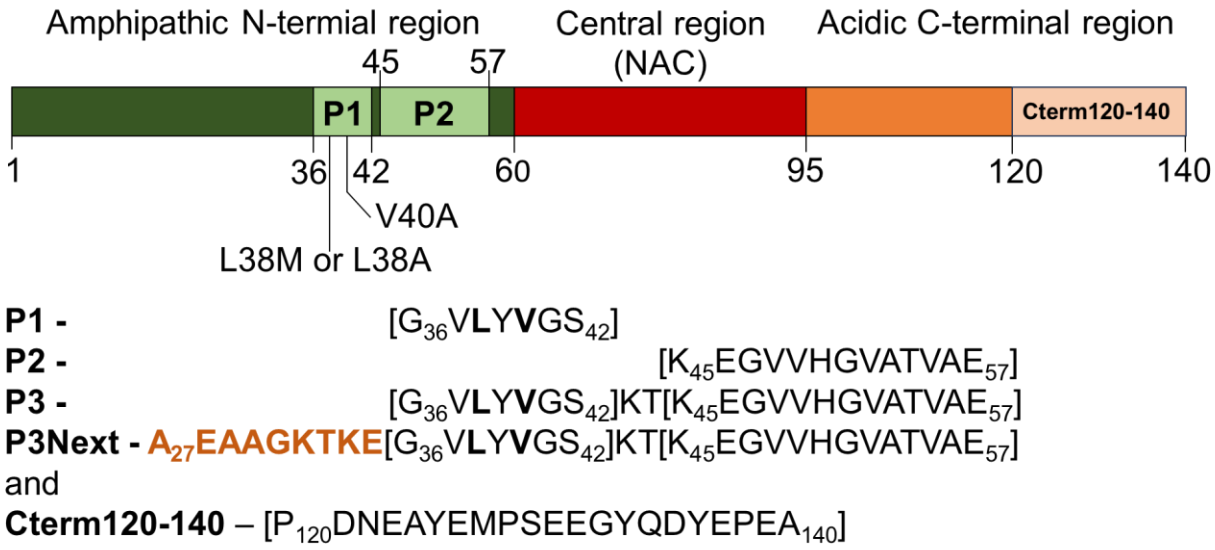


Figure 3.1 αSyn regions and the peptide fragments that were simulated herein. The amphipathic N-terminal (green) contains residues **αSyn[1-60]**. The central non-amyloid component (NAC) region (red) contains residues α**Syn[61-95]**. The acidic C-terminal region (orange) covers the last 45 residues. P1 and P2 are short motifs in the N-terminal region (light green). Cterm120-140 (light orange) was used in simulations of peptide mixtures to study potential inter-domain interactions. The sequences of P1, P2, P3, P3Next, and Cterm120-140 are listed.

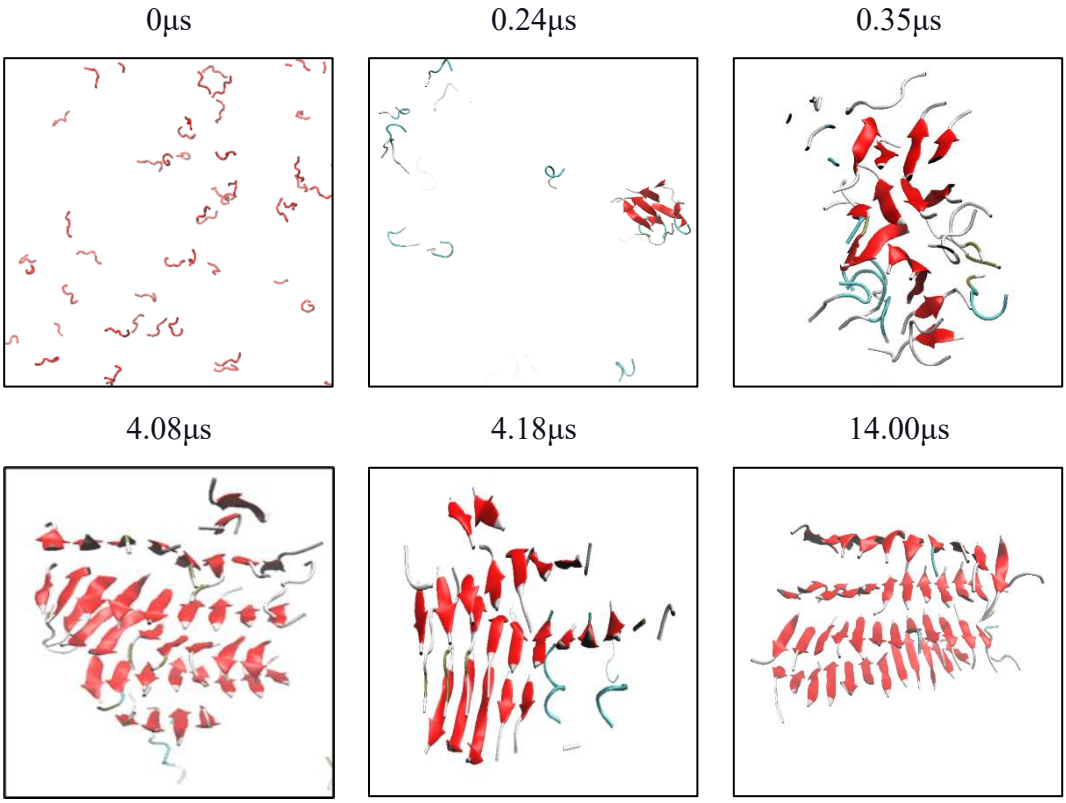


Figure 3.2 Snapshots at different times during the early stage of fibril formation of P1-WT fragments from one simulation. See also Supplementary Movie 1.

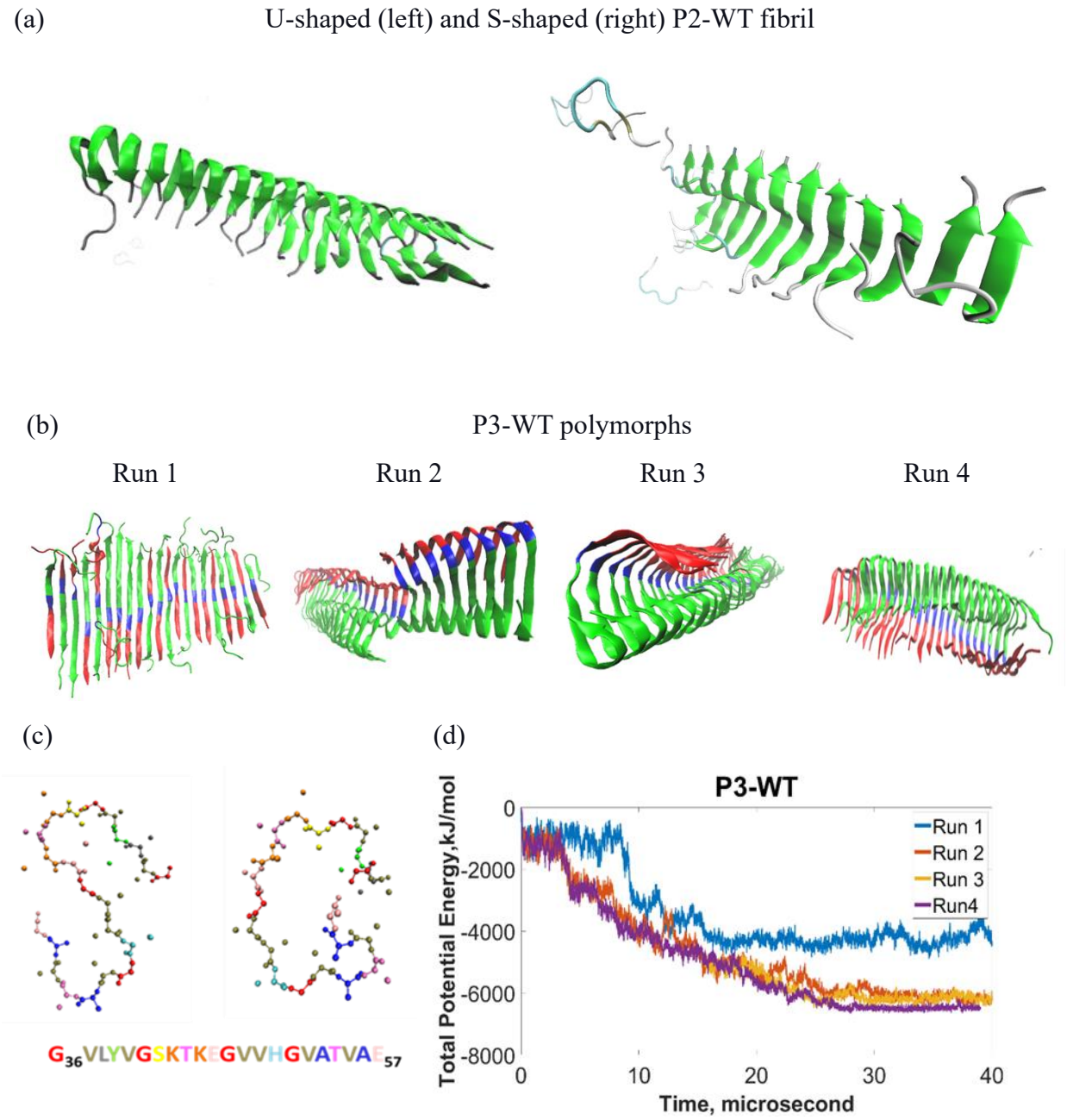


Figure 3.3 (a) Simulation snapshot of final conformation and schematic representation of U-shaped and S-shaped conformation of P2 fibrils. (b) Simulation snapshots of P3 fragments at the end of each run show different fibril structures: flat mixed fibril (Run 1), S-shaped fibril (Run 2), and G-shaped fibril (Runs 3 and 4). As P3 contains both P1 and P2, the color code is used to display the position of P1 (red), P2 (green), and the in-between residues αSyn[43-44] (blue). (c) Schematic representations of peptide conformation for S-shaped (left) and G-shaped (right) fibrils. (d) Total interaction energy vs time for the four runs of P3-WT. The S-shaped and G-shaped structures aggregate faster and are more stable than the flat mixed fibril.

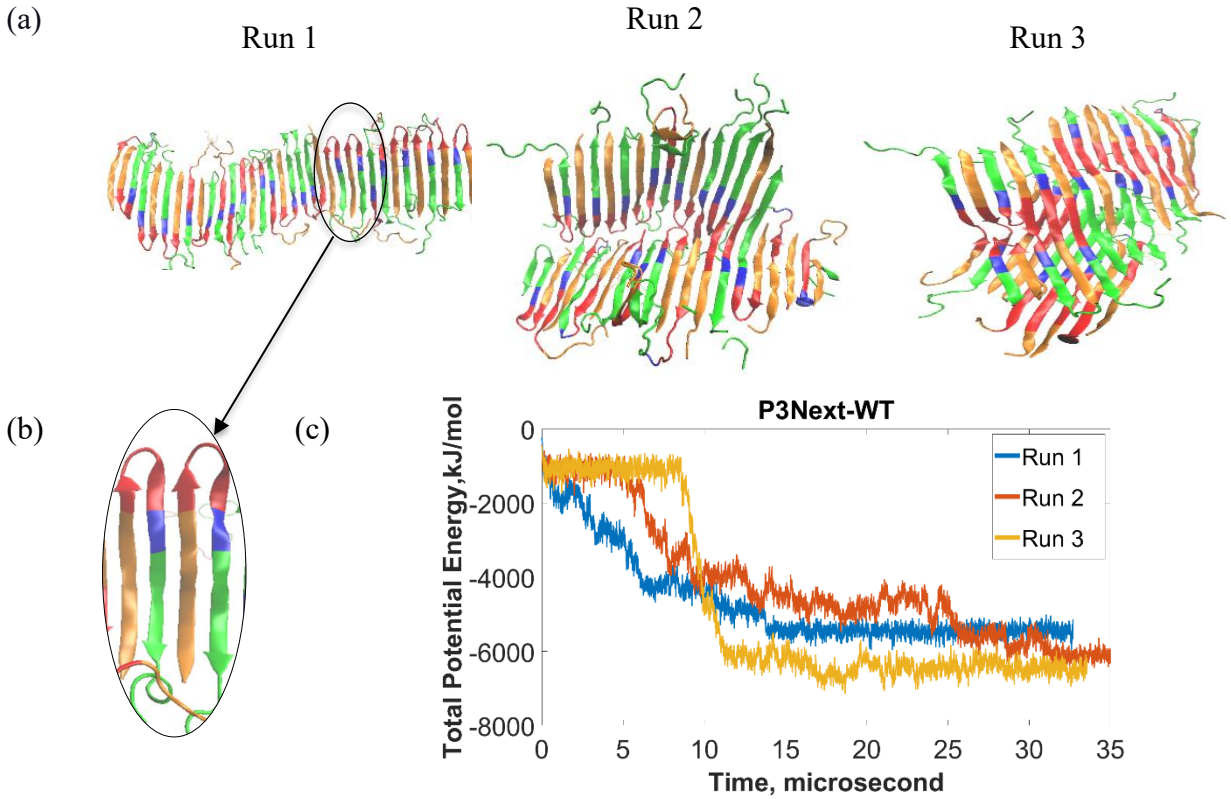
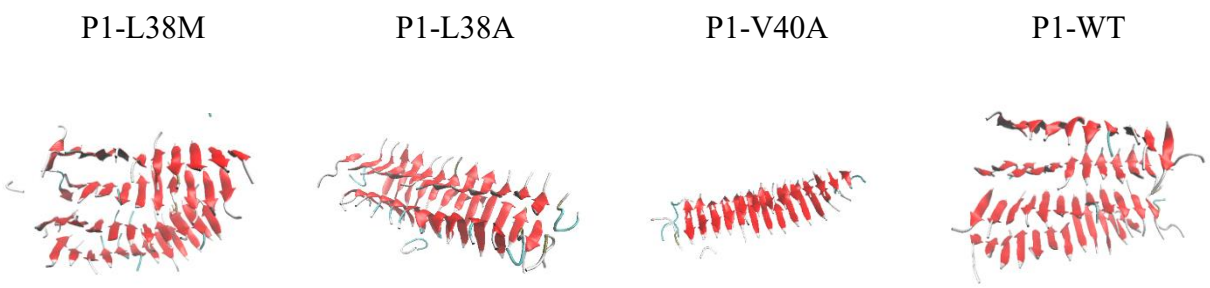


Figure 3.4 (a) Snapshots at the end of the simulations of P3Next wild-type fragments reveal that the number of β-hairpins varies in the P3Next-WT fibrils. The color code is used to display the position of P1 (red), P2 (green), the in-between residues αSyn[43-44] (blue), and the extended residues from P3 (orange). (b) Enlarged section from the snapshot of Run 1 shows two neighboring β-hairpins in the fibril. (c) The corresponding energy versus time profiles for all three runs display lag-times, with both large and small fluctuations in the interaction energy as a function of time.

(a)



(b)

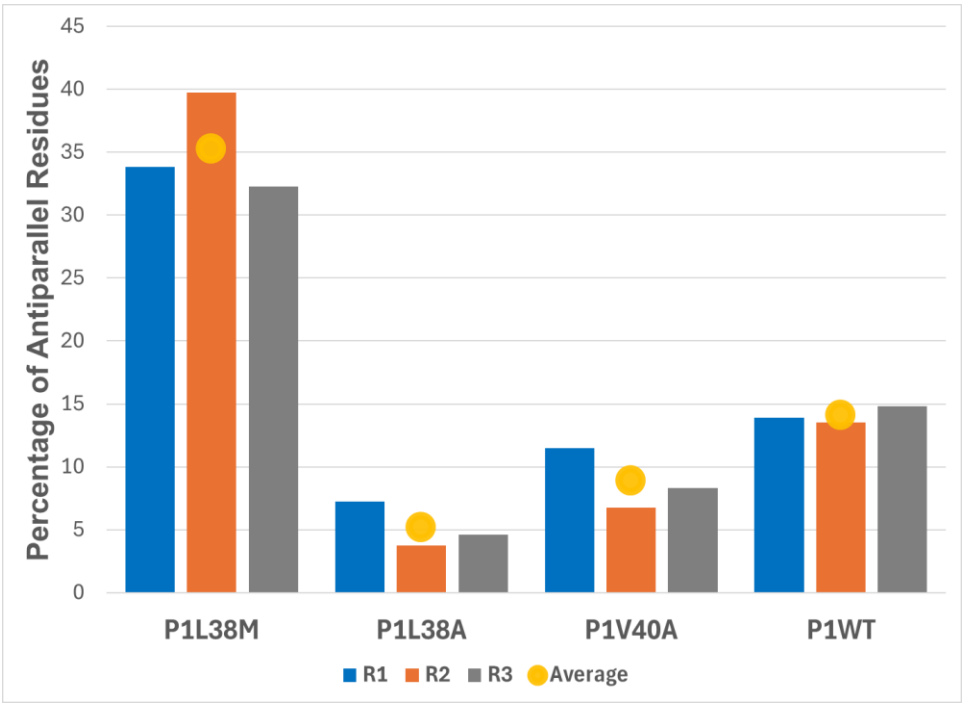
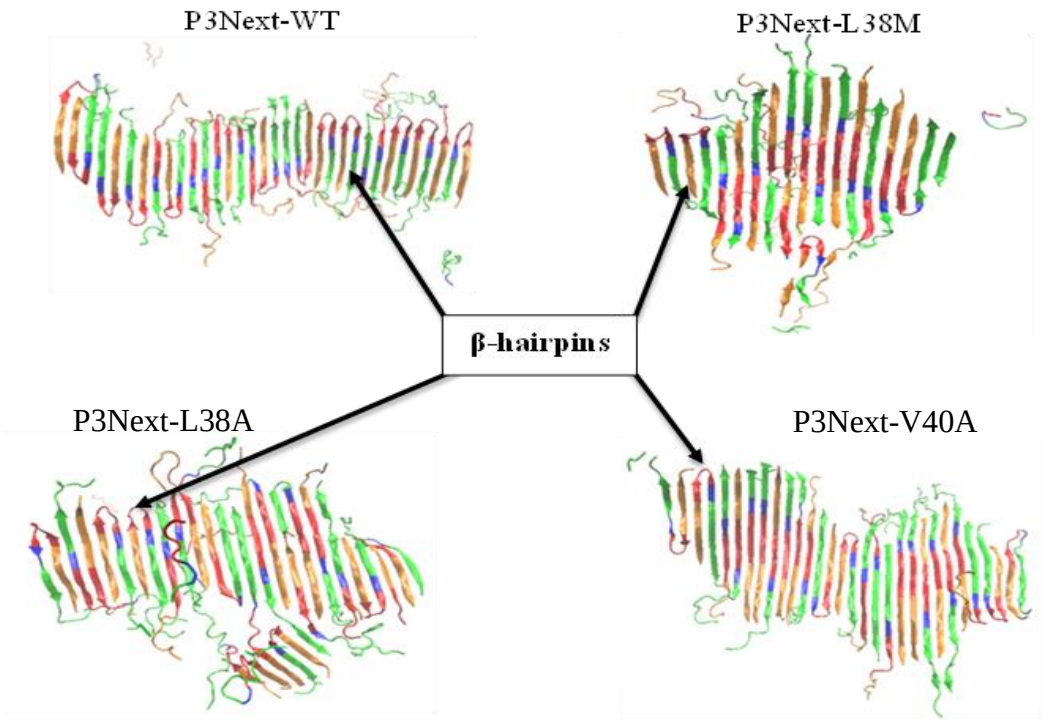


Figure 3.5 (a) The DMD/PRIME20 simulations of P1 peptides with the sequence substitution L38M, L38A or V40A, as well as the wild-type P1 (P1-WT). Although all four sequences formed multi-layer fibrils, the number of parallel and antiparallel β-strands are different. (b) Average percentage of antiparallel residues within a fibril.

(a)



(b)

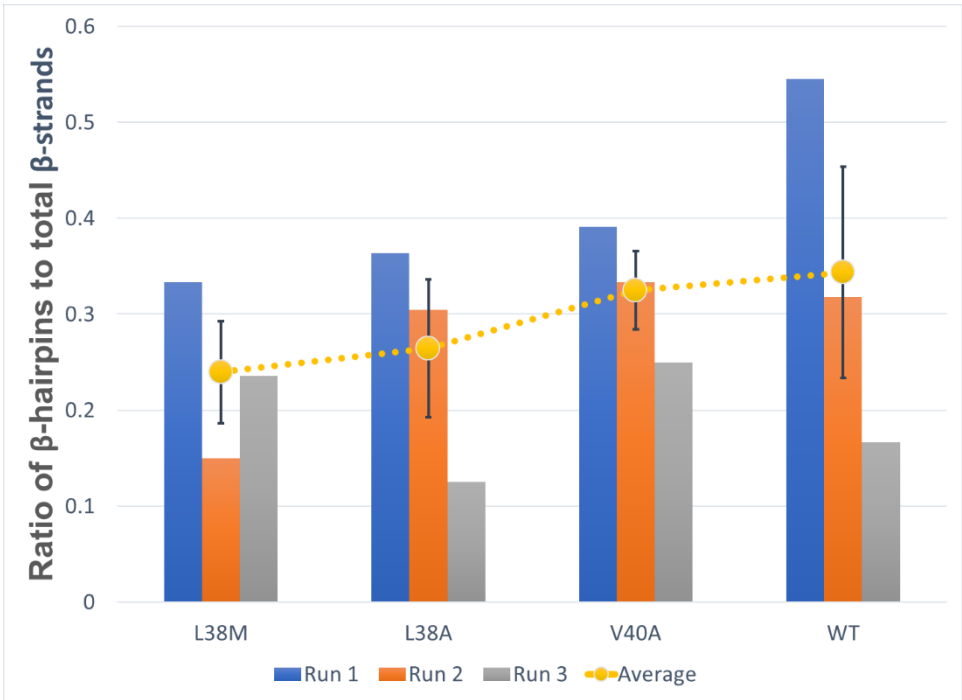


Figure 3.6 (a) Snapshot of β-hairpin formation in all P3Next variant fibrils. The number of β-hairpins in P3Next-L38M fibril is less than in the fibrils of the other variants. (b) Ratio of the number of β-hairpins to the total number of β-strands formed for each fibril in three runs plus the average.

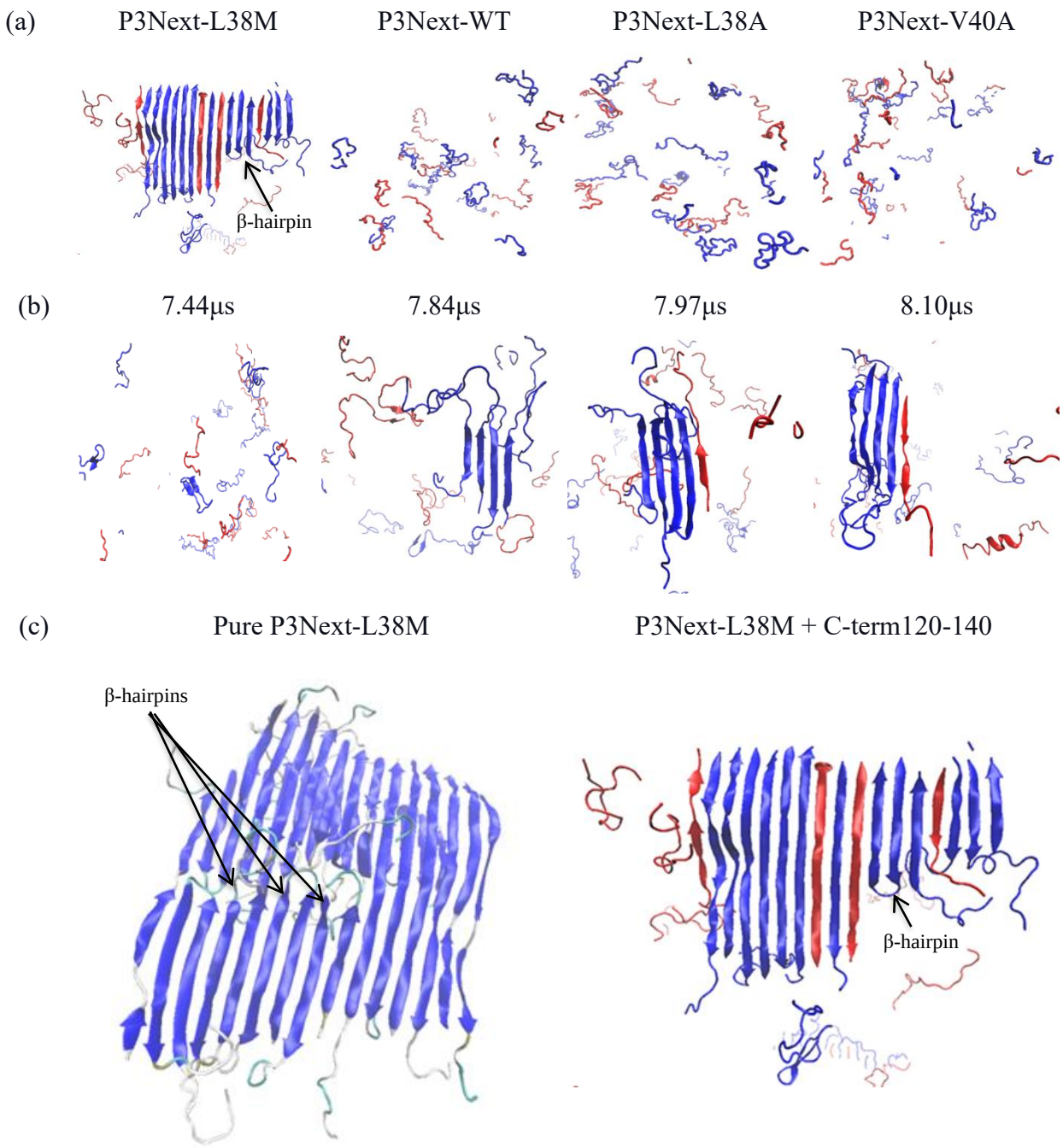


Figure 3.7 (a) DMD/PRIME20 final snapshots for mixtures that contain Cterm120-140 (red) and the P3Next variants (blue) at 300 billion collisions. (b) Snapshots at different times for the early stage of fibril formation for a P3Next-L38M and Cterm120-140 mixture. (c) Comparison of final simulation snapshots for pure P3Next-L38M and for the mixture of P3Next-L38M with C-term120-140 at 19.67μs (300 billion collisions). The mixture fibril contains significantly fewer β-hairpins than the pure fibril.

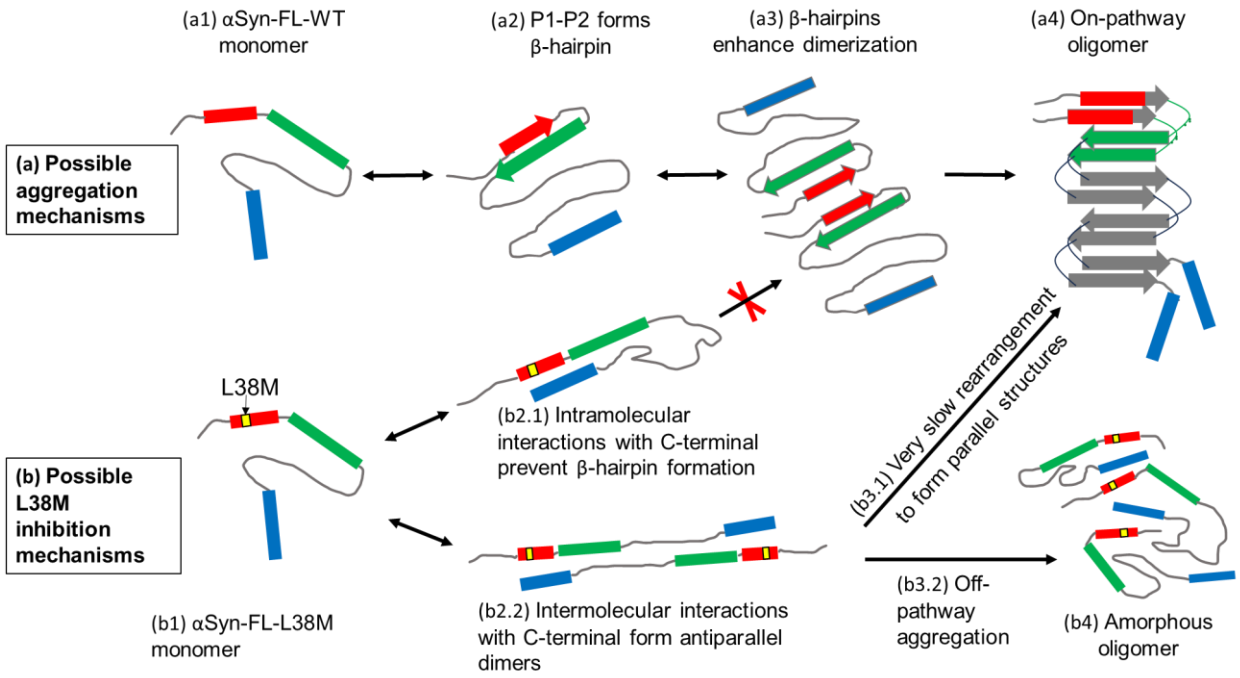


Figure 3.8 (a) A possible mechanism for αSyn-FL-WT aggregation. (b) A possible mechanism for L38M inibition of αSyn-FL. L38M affects αSyn-FL by inducing intramolecular and intermolecular interaction between the P1 region and the C-terminal. The colored blocks indicate the regions that are simulated: P1 (red), P2 (green), and Cterm120-140 (blue). The rest of the αSyn-FL protein is gray.

3.8 Supplementary Materials

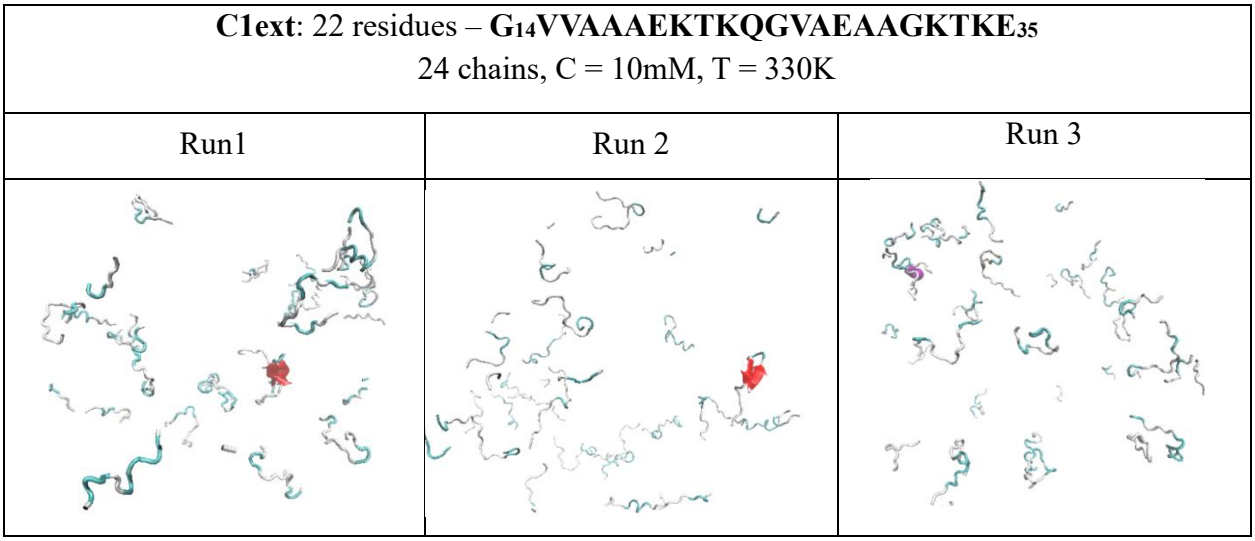


Figure 3.9 Final snapshots of DMD/PRIME20 simulations of C1ext. All simulation didn't aggregate at the end of 300 billion collisions (~20μs)

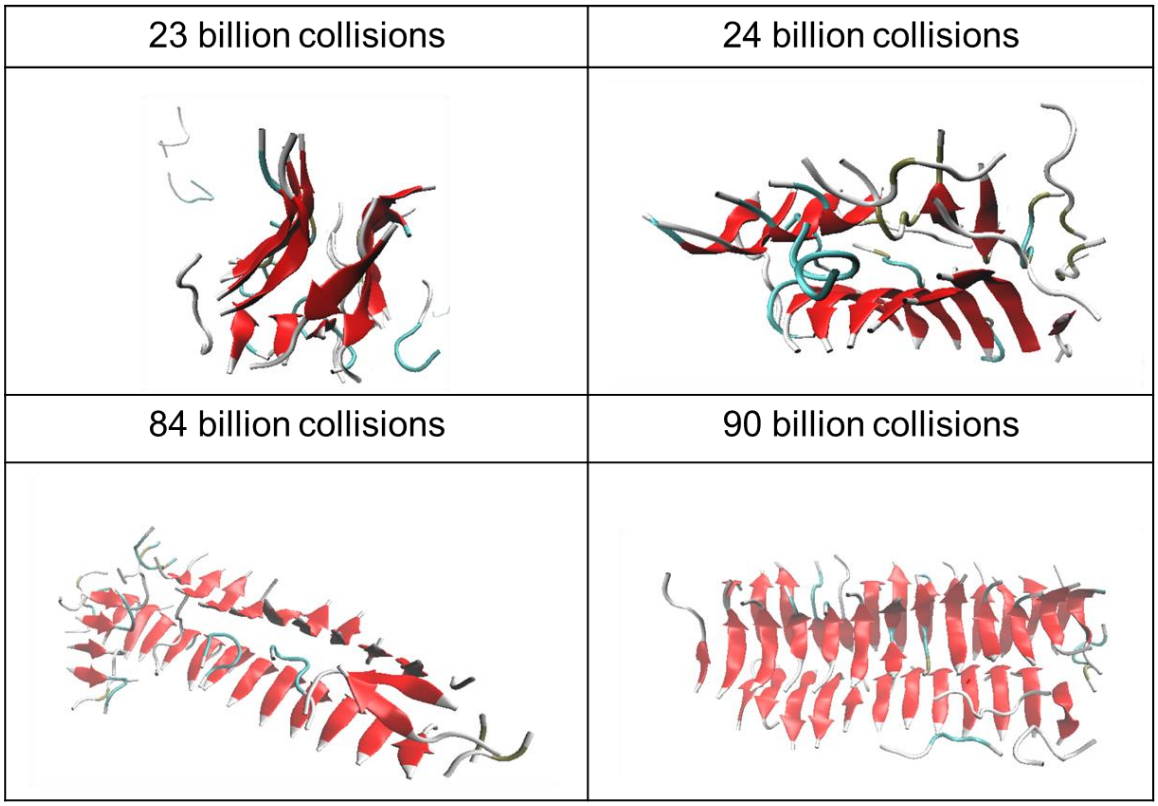


Figure 3.10 Snapshots at different times during the early stage of fibril formation of P1-L38A fragment from one simulation

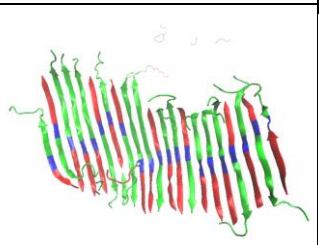
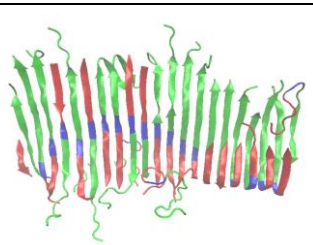
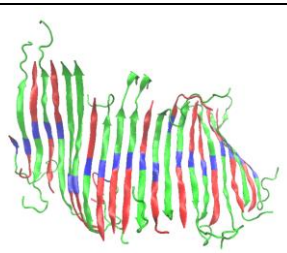
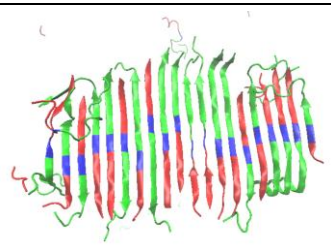
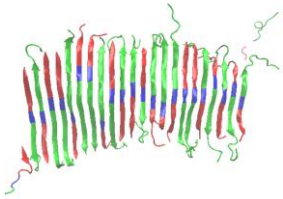
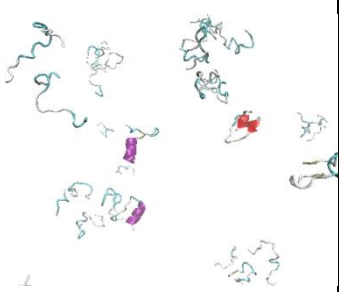
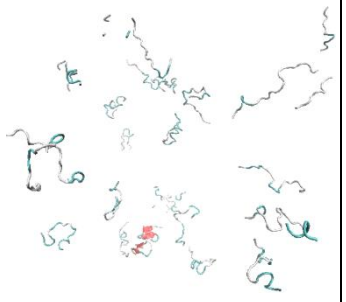
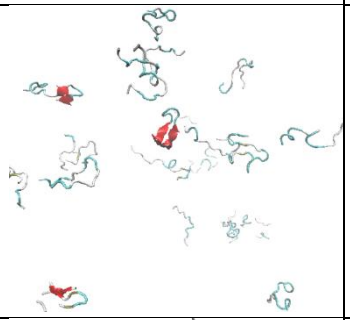
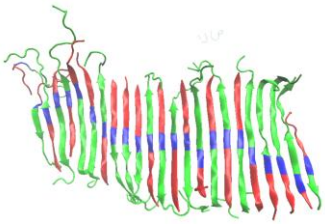
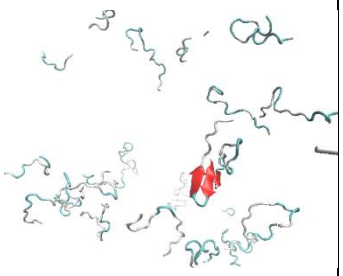
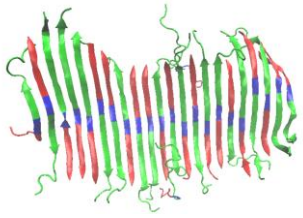
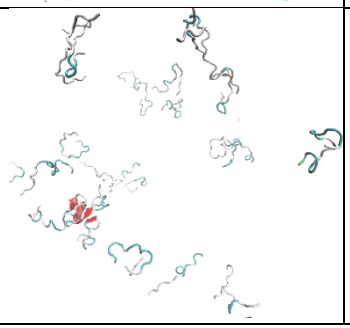
P3: 22 residues - G_{36}VLYVGSKTKEGVVHGVATVAE_{57} 24 chains, C = 10mM, T = 330K					
	Run1	Run 2	Run 3	Run 4	Experiment
L38M					Full length ASyn with L38M did not aggregate
L38A					Full length ASyn with L38A aggregated
V40A					Full length ASyn with V40A aggregated

Figure 3.11 Final snapshots of DMD/PRIME20 simulation of mutated **P3** variants. All snapshots were taken at the end of 500 billion collisions (~38μs)

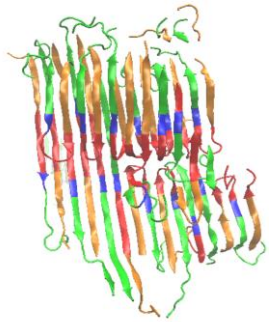
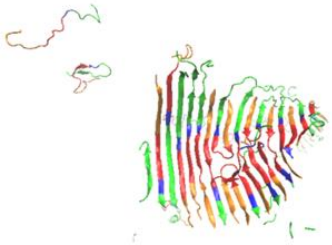
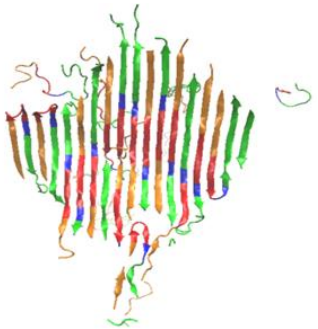
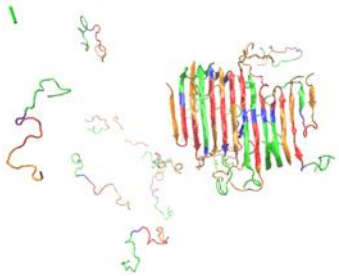
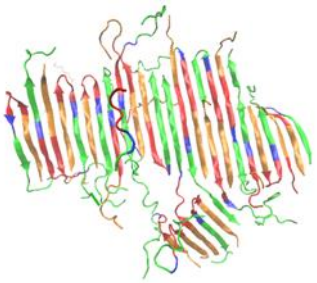
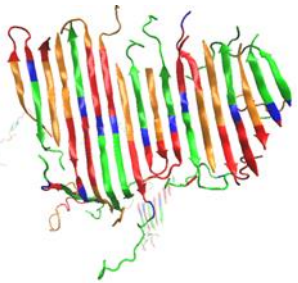
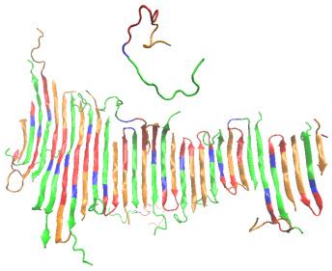
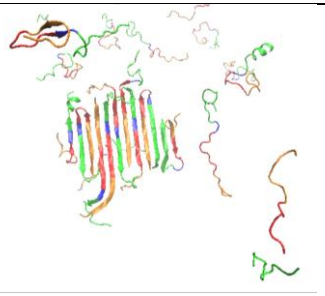
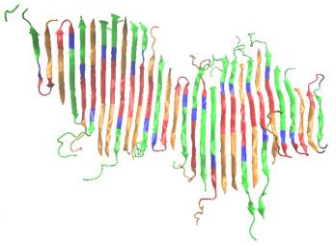
P3Next: 31 residues – A27EAAGKTKEGVLYVGSKTKEGVVHGVATVAE57 24 chains, C = 10mM, T = 330K				
	Run 1	Run 2	Run 3	Experiment
L38M				Full length ASyn with L38M did not aggregate
L38A				Full length ASyn with L38A aggregated
V40A				Full length ASyn with V40A aggregated

Figure 3.12 Final snapshots of DMD/PRIME20 simulation of mutated **P3Next** variants. Final snapshots were taken at the end of each simulation. L38M variants formed stable fibrils within 300 billion collisions. Although L38A and V40A simulations were extended to 600 billion collisions to get stable structures, L38A Run 1 and V40A Run 2 didn't aggregate completely.

Supplementary Movies

We also attached six Supplementary Movies for examples of aggregation process through the entire DMD/PRIME20 simulations. DMD/PRIME20 uses periodic boundary conditions. Therefore, the peptides would enter the opposite boundary when there was a boundary causing the peptides or fibrils to segment although they were still connected.

- Supplementary Movie 1: The aggregation of P1-WT from Run 1. Peptide P1 is red. This simulation Run 1 formed a five-layer fibril which contains parallel strand within a β-sheet and antiparallel β-sheet stacking.
- Supplementary Movie 2: The aggregation of P3-WT from Run 1. P3-WT is a 22 residues peptide that includes 7 residues of peptide P1 (red), 13 residues of peptide P2 (green) and the 2 residue KT in between P1 and P2 (blue). This simulation Run 1 formed a mixed fibril containing both parallel and antiparallel strands without a pattern.
- Supplementary Movie 3: The aggregation of P3-WT from Run 2. P3-WT is a 22 residues peptide that includes 7 residues of peptide P1 (red), 13 residues of peptide P2 (green) and the 2 residue KT in between P1 and P2 (blue). This simulation Run 2 formed an S-shaped parallel fibril.
- Supplementary Movie 4: The aggregation of P3-WT from Run 3. P3-WT is a 22 residues peptide that includes 7 residues of peptide P1 (red), 13 residues of peptide P2 (green) and the 2 residue KT in between P1 and P2 (blue). This simulation Run 3 formed a G-shaped parallel fibril.
- Supplementary Movie 5: The aggregation of P3Next-WT from Run 1. P3Next-WT is a 31 residues peptide that includes 7 residues of peptide P1 (red), 13 residues of peptide P2 (green),

the 2 residue KT in between P1 and P2 (blue) and the 9 N-terminal extending residues (orange). The simulation shows that β-hairpins formed in the fibril.

- Supplementary Movie 6: The simulation of P3Next-L38M and Cterm120-140. P3Next-L38M is blue and Cterm120-140 is red. The simulation shows the formation of mixed fibril that contains both P3Next-L38M and Cterm120-140 with only a β-hairpin observed in the fibril.

Chapter 4 Design of parallel β-sheet nanofibrils using Monte-Carlo search, coarse-grained simulations, and experimental testing

Chapter 4 is essentially a manuscript by Sudeep Sarma, Van Nguyen, Alicia S. Robang, Xingqing Xiao, Anant K. Paravastu, Carol K. Hall published on Protein Science

4.1 Introduction

We seek to establish a workflow to design previously-unknown amino acid sequences to produce peptides that assemble into specific desired structures. It is known that peptides can self-assemble into architectures like nanofibers[1,2], nanosheets[3], nanotubes[4], nanoparticles[5,6], but understanding of the relationship between amino acid sequence and structures of assemblies is limited. Our ability to engineer supramolecular structures at the nanoscale impacts a wide variety of potential applications[7,8] including as polymeric biomaterials[9], tissue-engineering scaffolds[10–13], hydrogels[2,8,14–17], drug release agents[18,19], and biomineralization components[20,21]. Short peptides are particularly desirable for biomaterials discovery because they can be easy to synthesize. Although peptide assemblies can be composed of molecules in various secondary structures, we focus here on β-sheet assemblies. "Bottom-up" strategies, in which the amino acid sequence length, composition and pattern are tailored, can be used to obtain supramolecular architectures with great structural variety and desired functional properties. The amino acid sequence composition of the peptides and their secondary structure drives the peptide self-assembly process to obtain peptide-based supramolecular assemblies.

In contrast to the "bottom up" search for amino acid sequences that we are aiming to establish here, previous designs for β-sheet peptide assemblies have been inspired by fragments from naturally-occurring amyloidogenic proteins, or have emphasized simple patterning of hydrophobic and polar amino acids. Examples of peptide fragments that self-assemble into β-sheet fibrils include the 7-mer peptide fragment Aβ (16-22) (sequence: *KLVFFAE*), which is associated with Alzheimer's disease, and the fibril-forming segment of the yeast prion protein Sup35 (sequence: GNNQQNY). Furthermore, Lynn et. al. chemically modified Aβ (16-22) to assemble into nano-sheets[4,7]. Examples of peptides designed with hydrophobic/polar patterning

include RADA16-I (sequence: Ac-RADARADARADARADA-NH_2)[1,22] and MAX1 (sequence: VKVKVKVKV^{D}PPTKVKVKVKV-NH_2)[2,23], where Ac- indicates an acetylated N-terminus, -NH_2 indicates an amidated C-terminus, ^{D}P indicates proline with D-chirality, -V^{D}PPT- corresponds to type-II' β-turn to promote β-hairpin formation.

There are eight possible classes of 2-layer β-sheet structures, called cross-β structures spines, that peptides can form according to a 2017 paper by Sawaya and Eisenberg[24]. Although there have been significant advances in statistical biophysics and bioinformatics-based tools to predict amyloidogenic regions in a peptide sequence, "bottom-up" computational design of novel peptide sequences not known to adopt β-sheet-rich supramolecular structures is still a challenge.

In our previous work, we developed a workflow for computational and experimental discovery of 7-amino-acid peptides for self-assembly into amyloid structures. The chosen target structure was the Class-8 cross-β spine structure described by Sawaya et al[24], with peptides arranged into a pair of stacked antiparallel β-sheets. The workflow started with PepAD; a Monte-Carlo based *pep*tide *a*ssembly *d*esign algorithm developed in the Hall lab. PepAD allows custom pre-settings for design parameters, such as the peptide length, amino acid sequence, backbone-scaffold, and hydration properties, to identify specific fibril-forming peptides. Additionally, PepAD uses atomistic force-fields rather than knowledge-based information and hence, enables the de novo design of peptides not known in nature. The self-assembling tendencies of the peptides identified by the PepAD algorithm are further evaluated using discontinuous molecular dynamics (DMD) simulations with the PRIME20 force field to examine their fibrilization kinetics. Eight of the twelve *in-silico* peptides identified by PepAD in our previous work successfully formed fibrils in the DMD simulations and self-assembled into anti-parallel β-sheets when tested experimentally[27]. Thioflavin-T (ThT) fluorescence measurements were used to monitor amyloid

fibril formation. Peptide secondary structure was probed using circular dichroism (CD) spectroscopy and Fourier-transform infrared spectroscopy (FTIR). FTIR also reported on β-strand organization within β-sheets. Finally, fibrils were imaged using negative-stain transmission electron microscopy (TEM). All peptides tested in that study exhibited nanofiber formation with FTIR signatures of antiparallel β-sheets. Since the relative alignment of peptides in neighboring sheets was not examined via DMD simulation or experiment, we could not claim that these peptides should form Class 8 structures.

In this work, we sought to test the ability of computational tools to design peptides with a different target β-strand organization, the Class-1 cross-β spine defined by Sawaya et. al [24]. This structure contains a 2-layer β-sheet structure, with parallel-oriented β-strands in each layer and antiparallel-oriented β-strands between the two layers. Sawaya et. al[24] reported 5 peptides that form this structure, including GNNQQNY of the prion protein Sup35. We know of no designer peptides and few naturally occurring peptides in this size range that assemble into parallel β-sheets. The energy associated with the hydrogen bond network of a parallel β-sheet is higher than that of an antiparallel β sheet[25], which may be why antiparallel β-sheets are favored for peptides in this size range. In contrast, larger amyloid peptides tend to favor parallel β-sheets, which maximize overlap of hydrophobic residues on adjacent β-strands. Furthermore, one can use the simple heuristic that oppositely charged sidechains near opposite termini can favor antiparallel organization[26], but we know of no analogous heuristic to favor parallel β-sheets. In the work described here**,** two rounds of designs were performed to obtain eight 7-mer peptides that self-assemble to form parallel β-sheets. In contrast to our workflow for antiparallel β-sheet structure, two rounds of PepAD design followed by DMD/PRIME20 simulations were needed to produce 8 candidate parallel β-sheet forming peptides for experimental testing. (As in our previous paper,

the relative alignment of peptides in neighboring sheets was not analyzed). Five of the peptides spontaneously formed fibril-like structures with a predominantly parallel β-sheet arrangement, two peptides formed fibril-like structures with <50% in parallel β-sheet arrangement and one peptide remained as a random coil in DMD/PRIME20 simulations. Fourier Transform Infrared Spectroscopy, Circular Dichroism, Electron Microscopy, and Thioflavin-T fluorescence spectroscopy measurements were conducted on 5 out of 8 peptides (commercially produced and received at >95% purity). These tests revealed that all 5 peptides self -assembled into parallel β-sheets.

4.2 Results

4.2.1 First Round of Design of Class 1 Cross β-spine Forming Peptides

PepAD is a Monte-Carlo based algorithm that searches for peptide sequences that can self-assemble to form supramolecular structures[27]. A score function, Γ_{score}, which considers (i) the binding free energy, $\Delta G_{binding}$, of the peptide chain with its neighboring peptides and (ii) the intrinsic self-aggregation propensity, $P_{aggregation}$, of the individual peptides, is used to evaluate new peptide sequences. Details are provided in *Section 4.1*.

The PepAD algorithm requires an initial backbone scaffold to design peptide sequences that can self-assemble into the Class 1 cross-β spine. As mentioned earlier, Sawaya et. al reported in their study[24] that the fibril forming segment GNNQQNY of the prion protein Sup35 forms a steric zipper characteristic of the Class 1 cross-β spine. Hence, GNNQQNY fibril was used as the reference peptide in the first round of design. This structure consists of a 2-layer amyloid fibril whose β strands are parallel within the β-sheet layer and antiparallel between the β-sheet layers (**Figure 4.1A**). Based on their study, we constructed two versions of the GNNQQNY structure

that forms the Class 1 cross-β spine; these were used as starting backbone scaffolds in two parallel design rounds. Hereafter, we refer to these two initial backbone scaffolds as **Conf-1** and **Conf-2**.

To obtain **Conf-1**, we constructed a 2-layer amyloid β-sheet structure using the Discovery Studio 3.5 and Packmol packages. Eight 7-mer GNNQQNY (PP0) peptide sequences were aligned in a parallel arrangement within each β sheet layer and an antiparallel arrangement between the β sheet layers with 4 peptides in each layer. The peptide distance within each sheet was set to be ~ 5.5 Å and the inter-sheet distance was specified as ~ 12 Å. A 5 ns explicit-solvent atomistic molecular dynamics simulation was conducted using the AMBER14 package to relax the two-layer β-sheet structure in the afore-mentioned parallel arrangement and to eliminate any atomic overlaps. The structure obtained following the 5ns simulation was used as the input backbone scaffold (**Conf-1**) for the PepAD algorithm (**Figure 4.1B**).

To build **Conf-2**, we used the crystal structure of peptide GNNQQNY (PDB ID: 2omm) reported in the Protein Data Bank. The primary coordinate file corresponding to PDB ID: 2omm contains the crystal asymmetric unit of a single GNNQQNY sequence and provides the information needed to generate the biological assembly of four GNNQQNY peptides into a Class 1 cross-β spine. This is a 2-layer β-sheet structure containing 2 parallel-oriented β-strands in each layer with antiparallel oriented β-strands between the two layers. An in-house python code was written to generate a configuration of the Class 1 cross-β spine with 4 peptides (**Figure 4.6**). Four replicas of this structure were generated to produce a 2-layer β-sheet structure with 8 parallel-oriented β-strands in each layer and an antiparallel arrangement between the two layers (**Figure 4.1C**) (Code is available at: https://github.com/CarolHall-NCSU-CBE/Parallel-self-assembling-peptides). We performed a 5 ns explicit solvent atomistic molecular dynamics simulation using the AMBER14 package to relax the aforementioned 2-layer sheet structure to eliminate any atomic

overlaps. The structure obtained following the 5ns simulation was used as a second input backbone scaffold (**Conf-2**) for the PepAD algorithm (**Figure 4.1D).**

Next, we specified the hydration properties for the designed amyloid-forming peptides as input parameters for the PepAD algorithm. Two cases were investigated with two different sets of hydration properties for the peptide chain. We classify the 20 natural amino acids into four residue types: hydrophobic residues (Leu, Val, Ile, Ala, Met, Phe, Tyr, Trp, Gly), polar residues (Ser, Thr, His, Asn, Gln), charged residues (Arg, Lys, Asp, Glu) and other residues (Cys, Pro). The two cases are as follows, Case 1: $N_{\text{hydrophobic}} = 5$, $N_{\text{polar}} = 0$, $N_{\text{charge}} = 2$, $N_{\text{other}} = 0$ and Case 2: $N_{\text{hydrophobic}} = 5$, $N_{\text{polar}} = 2$, $N_{\text{charge}} = 0$, $N_{\text{other}} = 0$. (We were interested in determining which hydration properties favor the parallel amyloid-β sheet formation.) For each case we performed the PepAD algorithm at two different values for the weighting factor λ, viz. λ=2.0 and λ=3.0 in equation (1). The weighting factors λ=2.0 and λ=3.0 were chosen to provide a good balance between optimizing the binding free energy ($\Delta G_{binding}$), and the aggregation propensity terms ($P_{aggregation}$) in the Γ_{score} of the amyloid-forming structure. All the searches start with random peptide sequences draped on the fixed backbone scaffold. By having random initial peptide sequences, we encourage our designs to proceed along different search pathways in sequence space and thereby sample peptides from a larger pool of peptide sequences than would otherwise be the case. The $\boldsymbol{\Gamma}_{score}$ profile fluctuates considerably as new amino acids are placed on the different sites of the peptide chain. By examining the $\boldsymbol{\Gamma}_{score}$ profiles over the sequence evolution, we identified four peptide sequences (PP1-PP4) with low scores for evaluation using DMD/PRIME20 simulations. PP1 and PP2 were identified with **Conf-1** as the starting structure, and PP3 and PP4 were identified with **Conf-2** as the starting structure.

We performed a preliminary screen to investigate the fibrilization kinetics of the four PepAD identified *in-silico* peptides (PP1-PP4) using DMD/PRIME20 simulations. DMD/PRIME20 is a fast alternative to traditional molecular dynamics simulations that uses discontinuous potentials to model peptide aggregation. The force field PRIME20, developed by the Hall group in 2010, is a coarse-grained model where each amino acid is represented by a 3-sphere backbone comprised of united atoms (NH, $C_\alpha H$ and CO) and a single-sphere side chain, R[28-33]. Details are provided in *Section 4.2*. The simulations were performed at temperatures ranging between 296 K to 310 K for 5 μs. The peptides were then extensively studied at the temperature at which they showed the highest fibrillation propensities by performing 12 μs DMD/PRIME20 simulations. The simulations predict that peptide PP1 (sequence: GDIKIVV), PP2 (sequence: GNIVTFV) and PP3 (sequence: GAIDWVK) spontaneously form amyloid-like fibrils and adopt a predominantly parallel β-sheet arrangement. Peptide PP4 (sequence: GGIDWKI) formed amyloid-like fibrils but exhibited less than 50% parallel β-sheet content in the DMD/PRIME20 simulations. **Table 4.1** contains the sequences of peptides PP1-PP4 with their associated scores, binding free energies, intrinsic self-aggregation propensities; the number of layers in the fibrils and the parallel β-sheet content percentage were estimated from DMD/PRIME20 simulations. The % β-sheet content v/s simulation time for peptides PP1-PP3 and PP4 are shown in **Figure 4.1E** (left) and **Figure 4.7**, respectively. (Snapshots of the final simulated structures of peptides PP1-PP3 are shown in **Figure 4.1E** (right).

4.2.2 Second Round of Design for Class 1 Cross β-spine Forming Peptides

Since peptide PP2 (sequence: GNIVTFV) was the most promising candidate in our first round of design of Class 1 cross-β spine forming peptides when studied via DMD simulations, it was selected as the reference sequence to create the initial backbone scaffold to perform a second

round of *in-silico* peptide design. (We liked that the β–sheets which assembled in the DMD simulations were 100% parallel.) To build the backbone scaffold we used the Pymol software to mutate the residues on the Class 1 cross-β spine structure of 16 GNNQQNY peptides (**Figure 4.1C)** to generate a Class 1 cross-β spine structure containing 16 GNIVTFV (PP2) peptides (**Figure 4.8**). For best comparison with the DMD/PRIME20 simulations and experimental biophysical characterization (s*ee Section 4.3*), the N-termini and C-termini were acetylated and amidated (N-cap and C-cap), respectively in this round of design using the PepAD algorithm. (The main effect of "patching" of termini is to eliminate charges at neutral pH). A 25ns explicit-solvent atomistic molecular dynamics simulation was conducted using the AMBER14 package to relax the parallel 2-layer PP2 β-sheet structure and eliminate any atomic overlaps. The structure obtained following the 25ns simulation (**Figure 4.2A)** was used as the input backbone scaffold for the second round of PepAD design. We refer to this structure as **Conf-3**.

We next performed the PepAD algorithm to generate a new cohort of parallel amyloid-forming peptides with Case 1 and Case 2 hydration properties using peptide PP2 draped on **Conf-3** as the reference peptide. Four peptides (PP5-PP8) obtained from this round were further investigated in DMD simulations to study their fibrillation kinetics. We again performed DMD simulations of the peptide systems for 5 μs to examine their self-aggregation kinetics at temperatures ranging from 296.1 K to 310K. These peptides were then extensively studied at the temperatures at which they showed the highest fibrillation propensity for 12 μs in DMD/PRIME20 simulations. Our simulations revealed that peptides PP5 (sequence: ADKVMFV) and PP7 (sequence: GNYTMFI) exhibited high parallel β-sheet content while peptide PP8 (sequence: ANMTVFV) exhibited < 50% parallel β-sheet content. Peptide PP6 (sequence: GDFVKFV) predominantly remained as random coil in our DMD/PRIME20 simulations. The sequence of

peptides PP5-PP8 with their associated scores, binding free energies and intrinsic self-aggregation propensities, and observations from DMD/PRIME20 simulations are reported in **Table 4.2**. The % β-sheet content v/s simulation time of PP7 and PP8 obtained from DMD/PRIME20 simulations are shown in **Figure 4.2B.** Snapshots of the final simulated structures of peptides PP7 and PP8 are shown in **Figure 4.2C and D** respectively.

4.2.3 Experiments Revealed Parallel Beta-sheets Nanofibers.

We experimentally evaluated the assembly of 5 out of the 8 candidates in **Table 4.1** and **Table 4.2**. We ordered commercial production of PP1, PP2, PP3, PP4, PP7 and PP8 (note that PP8 exhibited low fibrillar content in the DMD/PRIME20 simulations), but PP2 was not experimentally tested as it did not meet purity standards of 95%. Peptides PP5 and PP6 were not synthesized. All five final sequences formed parallel β-sheet nanofibers as evidenced by biophysical characterization techniques like negative-stain Transmission Electron Microscopy (TEM), Fourier-transform infrared spectroscopy (FTIR), Thioflavin-T Fluorescence (ThT) and Circular Dichroism (CD). All 5 sequences exhibited parallel β-sheet assembly.

The experimental techniques that we employed tend to be compatible with varying peptide concentrations. We performed Fourier-transform infrared spectroscopy at a concentration of 10mM, Transmission Electron Microscopy at a concentration of 1mM, Circular Dichroism at a concentration of 0.2mM, and Thioflavin-T Fluorescence at a concentration of 1mM. It is important to note that concentration of peptide solutions is a factor in the formation of β-strands. This concentration difference may result in varying levels of assembly between experimental techniques.

We observed the most definitive evidence of assembly and parallel β-sheets using FTIR and TEM, which we performed on peptide solutions at the highest peptide concentration (10mM).

TEM can image the presence of nanofiber peptide assemblies. FTIR is sensitive to β-sheet secondary structure and can differentiate between parallel and antiparallel β-sheet structures. An FTIR peak near 1620cm^{-1} indicates β-strand secondary structures and an additional peak near 1690cm^{-1} is attributed to antiparallel organization of β-strands within the β-sheet. **Figure 4.3A** is a TEM image of fibrils of peptide PP3 showing assembly. **Figure 4.9** shows TEM images of all peptides. TEM Images reveal that all peptides form fibrils with thicknesses consistent with multi-layer β-sheets. **Figure 4.3B** compares the FTIR spectra from peptide PP3 to peptide P12 (sequence: ALRLELA) from our previous work[27]. P12 is a peptide from our previous successful effort to design peptides to form antiparallel β-sheets. The spectra of both peptides show a peak intensity near 1620cm^{-1} (1619cm^{-1} and 1625cm^{-1} respectively) indicating β-strand secondary structure, as expected. The additional peak at 1690cm-1 observed for P12 but not PP3 is indicative of anti-parallel organization of β-sheets. **Figure 4.4** compares FTIR spectra from all peptides tested in our present work (**Figure 4.4A**) to all peptides tested in our previous work (**Figure 4.4B)**[27]. The former group of peptides did not exhibit a peak at 1690cm^{-1}, whereas the latter group of peptides did. We interpret these results to indicate that the established design and characterization workflow was successful in producing self-assembling peptides with the target parallel β-strand organization.

We used Thioflavin-T fluorescence measurements to probe kinetics of β-sheet formation at peptide concentrations of 1mM. **Figure 4.5A** indicates that PP7 and PP8 assemble immediately, while PP3 and PP4 show low but increasing levels of fluorescence over 72h. PP1 has low fluorescence levels but shows evidence of assembly through other tested experimental methods. Note that absolute fluorescence level cannot be readily compared between different peptides: they are affected by factors such as binding of thioflavin-T to β-sheet surfaces and conformation of

bound ThT molecules, which both depend on amino acid sequence. There are differences in the kinetics of assembly among these peptides. PP3 and PP4 both show a gradual increase in fluorescence during Thioflavin-T fluorescence experiments indicating slower assembly than PP7 and PP8. In fact, definitive FTIR spectra for PP3 and PP4 were obtained only after 6 days post assembly. In contrast PP7 and PP8 assemble immediately during Thioflavin-T Fluorescence experiments. This variation in kinetics was not observed in our previous study of peptides that form antiparallel β-sheets.

We performed Circular Dichroism (CD) experiments on the peptides to characterize their secondary structure at 0.2mM. Assembled β-sheets are expected to show a single minimum near 220nm in CD spectra. We observed CD spectra consistent with heterogeneous secondary structure with partial β-strand signatures (**Figure 4.5B**). We used the BeStSel (Beta Sheet Selection) web server-based tool to predict the secondary structural distribution based on the CD spectra (**Table 4.4**)[34]. BeStSel interpreted the spectrum of PP1 to correspond to 100% β-sheets, whereas it associated the other CD curves to between 30 and 65% β-strands. BeStSel also attempts to distinguish between parallel and antiparallel β-sheets. In agreement with our FTIR results in **Figure 4.4**, BeStSel attributed parallel β-sheet structure of PP1. However, for peptides PP3, PP4, PP7, and PP8, BeStSel assigned antiparallel β-sheet structure to the partial β-strand signatures, which is inconsistent with our FTIR results. We suggest that BeStSel is unreliable in its ability to distinguish between parallel and antiparallel β-strand organization, especially when secondary structure is heterogeneous. **Table 4.3** summarizes the results of DMD/PRIME20 simulations and experimental measurements (FTIR, ThT, TEM and CD) of peptides PP1-PP8.

4.3 Discussion and Conclusion

The goal of this work was to identify peptides that self-assemble to form fibrils composed of parallel β-sheets. Inspiration for this work was the Class-1 cross-β spine structure described by Sawaya et al.[24] that contains two β-sheet layers parallel-oriented β-strands in each layer, and antiparallel-oriented β-sheets between the two layers. Thus far, only one 7-mer peptide, the fibril-forming segment GNNQQNY of the yeast prion protein Sup35, has been identified in the literature as forming amyloid fibrils of the 1st class in experiments. As a first step towards discovering new Class -1 peptides, we set out to design peptides that form fibrils with parallel β-sheets, regardless of the relative orientation of peptides in neighboring layers, by augmenting the workflow involving PepAD, DMD/PRIME20 simulations and experimental characterization. By using the PepAD algorithm coupled with DMD/PRIME20 simulations, we performed two rounds of designs with three different starting backbone scaffolds (**Conf-1**, **Conf-2** and **Conf-3**) to obtain a library of 7-mer amyloid-forming peptides that could potentially assemble into parallel β-sheets. This work complements our previous study where we identified peptides that assemble into anti-parallel β-sheets as are found in the cross-β spine of the 8th class. Discontinuous molecular dynamics simulations with the PRIME20 force field helped us computationally analyze the self-aggregation kinetics of the peptides identified by PepAD. Five out of the eight peptides were synthesized and experimentally tested, and all of them aggregated to form parallel β-sheets. Experimentally, aggregation was detected by observation of nanofibers in TEM images, measurement of CD curves consistent with β-stand secondary structure, positive ThT binding as detected by fluorescence, and FTIR spectra reporting parallel organization of β-sheets.

Although we were successful in showing that computational designs can control organization of β-strands within β-sheets, there are experimental observations that are not readily

interpretable based on outputs of PepAD and DMD/PRIME20. First, the ThT measurements indicate a large peptide-to-peptide variation in assembly kinetics, with PP7 and PP8 exhibiting maximal ThT fluorescence in ~1-2h and PP3 and PP4 showing continuous increase in ThT fluorescence intensity over the course of 72h. Based on DMD/PRIME20 simulations, we had not anticipated large variation in assembly kinetics between different peptides. Second, nanofiber thicknesses varied (**Figure 4.10**), but all observed thicknesses were far larger than the expected dimension based on 2 β-sheet layers (the interlayer distance between the two sheets from atomistic molecular dynamics simulations is ~ 4-5Å). Therefore, all the peptides observed to assemble form β-sheet interfaces that were not anticipated in the PepAD algorithm.

To summarize, we now have a computational workflow, PepAD algorithm & DMD/PRIME20 which can output novel sequences that form a desired organization of β-sheets. Thus far we have succeeded in designing peptides that control assembly into parallel or antiparallel β-sheet structures. The predicted structures have been experimentally tested in our previous and current work. For our future work, we could experimentally probe the stacking of β-sheets to validate the predictions of the computational workflow. Additionally, the computational workflow could be developed to discourage the multi-layer fibril formation observed experimentally.

4.4 Methods

4.4.1 Peptide Assembly Design Algorithm

The *Pep*tide *A*ssembly *D*esign (PepAD) algorithm is a Monte Carlo (MC)-based search procedure to discover peptides that can self-assemble to form supramolecular architectures. In this work, we have focused on designing peptides that self-assemble into the parallel peptides. The procedure is described briefly below.

(1) *Generate input peptide backbone scaffold:* A backbone-scaffold of a reference peptide is required to start the design process. In this work, a peptide scaffold corresponding to the Class 1 cross β-spine (2-layer β-sheet structure with 2 parallel-oriented β-strands in each layer and antiparallel oriented β-strands between the two layers) is generated.

(2) Compute score of initial peptide backbone scaffold: The tendency of the initial peptide backbone scaffold to self-aggregate into a well-organized amyloid-like structure is evaluated using, $\boldsymbol{\Gamma}_{score}$, a score function that considers the binding affinities between the neighboring chains on the peptide backbone scaffold ($\Delta G_{binding}$), and the intrinsic aggregation propensities of the individual peptides($P_{aggregation}$). The $\boldsymbol{\Gamma}_{score}$ is defined to be as follows:

$$\Gamma_{score} = \Delta G_{binding} - \lambda \times P_{aggregation} \qquad (1)$$

where λ is a weighting factor that adjusts the relative importance of the intrinsic aggregation propensity of the peptides during the sequence evolution.

(3) *Iteration of peptide sequence change moves:* The PepAD algorithm performs 10,000 evolution steps and generates variants of the reference peptide by performing two kinds of trial moves, *viz.* (i) *residue mutation* in which an old residue on all of the peptide chains is randomly chosen and replaced by a new one of the same residue type (hydrophobic, polar, charge, and other); (ii) *residue exchange* in which two residues on all of the peptide chains are randomly chosen and swapped, regardless of their residue type.

(4) *Evaluate score Γ_{score} of new peptide sequence:* The $\boldsymbol{\Gamma}_{score}$ for the newly generated peptide sequence draped on the backbone scaffold is evaluated.

(5) *Monte Carlo Metropolis Algorithm:* The Monte Carlo Metropolis algorithm is used to accept or reject new trial peptides.

More details regarding the PepAD algorithm and Γ_{score} can be found in our previous work[27]. The development of the PepAD algorithm has been inspired by our previous work on designing peptides that bind to biomolecular targets using a *Pep*tide *B*inding *D*esign (PepBD) algorithm[35-40].

4.4.2 Discontinuous Molecular Dynamics Simulation and PRIME20 Model

Discontinuous molecular dynamics (DMD) simulation with the PRIME20 force field have been used to study the fibrilization kinetics of designed peptides by the Hall group. DMD is a fast alternative to traditional molecular dynamics simulations in which the interaction between two particles is modeled with a discontinuous potential, such as hard-sphere, square-well, or square-shoulder potentials. The PRIME20 model is an implicit-solvent coarse-grained protein force field developed in the Hall group that was specifically designed for simulating peptide aggregation with DMD. In the PRIME20 model, each amino acid is represented by three backbone spheres (NH-, C_αH-, and CO-) and one side chain sphere (R-). Each side chain of the 20 natural amino acids is assigned a unique size, atomic mass, and C_α-R bond length. Details of the DMD simulations and PRIME20 model are described in our earlier work[28–33].

In this work, DMD/PRIME20 simulations of the PepAD generated peptides (PP1-PP8) were conducted at T=296K, 303K and 310K for 5 μs for the preliminary screen. The temperature at which a peptide showed highest fibril formation was studied extensively by simulating that peptide for three runs, each at 12 μs. For each *in-silico* peptide system, 48 peptides are placed into a cubic box with edge lengths of 200.0 Å, to achieve a peptide concentration ~10 mM. In each run, the peptide system starts from a random-coil state. The DMD simulations were carried out in the canonical ensemble. The Andersen thermostat is implemented to maintain the simulation system

at the desired temperature. Snapshots of the final simulated structures are obtained using the VMD 1.9.3 software.

4.3 Experimental Assessment of Self-Assembly

The sequences output by PepAD were experimentally tested by negative-stain transmission electron microscopy (TEM), Fourier-transform infrared spectroscopy (FTIR), Thioflavin-T Fluorescence and Circular Dichroism (CD) using the methods we detailed previously[23]. All peptides were imaged using negative-stain transmission electron microscopy at a peptide concentration of 1mg/ml in deionized water (DI water). FTIR measurements were conducted after a minimum of 72h post assembly at a concentration of 10mM in DI water. Additionally, we performed Circular Dichroism experiments at a concentration of 0.2mM in DI water. Thioflavin-T Fluorescence experiments were at a peptide concentration of 2mM over an assembly period of 72h.

4.4 Acknowledgements

The authors in this work would like to thank the National Science Foundation (Award No. OAC-1931430) for financial support. This work used the Extreme Science and Engineering Discovery Environment (XSEDE, Award no. ACI-1548562) and the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS, Award no. CHE110065) supported by the National Science Foundation. SS and CKH thank San Diego Supercomputer Center (SDSC) for providing computing time.

4.5 Data Availability

Data is available in the manuscript and/or supporting information. The PDB files of starting conformations (Conf-1, Conf-2, and Conf-3) and PepAD output files of PP1-PP8, PDB output files

and beta-sheet content from DMD/PRIME20 simulations of PP1-PP8 and analyis codes are available at: https://github.com/CarolHall-NCSU-CBE/Parallel-self-assembling-peptides.

4.6 Declaration of Competing Interests

The authors declare that they have no known competing interests that could have appeared to influence the work reported in this paper.

4.7 References

1. Cormier AR, Pang X, Zimmerman MI, Zhou HX, Paravastu AK. Molecular structure of RADA16-I designer self-assembling peptide nanofibers. *ACS Nano*. 2013;7(9). doi:10.1021/nn401562f

2. Nagy-Smith K, Moore E, Schneider J, Tycko R, DeGrado WF. Molecular structure of monomorphic peptide fibrils within a kinetically trapped hydrogel network. *Proc Natl Acad Sci U S A*. 2015;112(32). doi:10.1073/pnas.1509313112

3. Childers WS, Mehta AK, Ni R, Taylor J V., Lynn DG. Peptides organized as bilayer membranes. *Angew Chemie - Int Ed*. 2010;49(24):4104-4107. doi:10.1002/anie.201000212

4. Li S, Mehta AK, Sidorov AN, et al. Design of Asymmetric Peptide Bilayer Membranes. *J Am Chem Soc*. 2016;138(10). doi:10.1021/jacs.6b00977

5. Tian Y, Zhang HV, Kiick KL, Saven JG, Pochan DJ. Transition from disordered aggregates to ordered lattices: Kinetic control of the assembly of a computationally designed peptide. *Org Biomol Chem*. 2017;15(29). doi:10.1039/c7ob01197k

6. Villegas J, Sinha N, Teramoto N, Bargen C, Pochan D, Saven J. Computational design of single-peptide nanocages with nanoparticle templating. *Molecules*. 2022;27(1237):1-15.

7. Wilson CJ, Bommarius AS, Champion JA, et al. Biomolecular Assemblies: Moving from Observation to Predictive Design. *Chem Rev*. 2018;118(24). doi:10.1021/acs.chemrev.8b00038

8. Sinha NJ, Langenstein MG, Pochan DJ, Kloxin CJ, Saven JG. Peptide Design and Self-assembly into Targeted Nanostructure and Functional Materials. *Chem Rev*. 2021;121(22). doi:10.1021/acs.chemrev.1c00712

9. Katyal P, Meleties M, Montclare JK. Self-Assembled Protein-and Peptide-Based Nanomaterials. *ACS Biomater Sci Eng*. 2019;5(9). doi:10.1021/acsbiomaterials.9b00408

10. Deidda G, Jonnalagadda SVR, Spies JW, et al. Self-Assembled Amyloid Peptides with Arg-Gly-Asp (RGD) Motifs As Scaffolds for Tissue Engineering. *ACS Biomater Sci Eng*. 2017;3(7). doi:10.1021/acsbiomaterials.6b00570

11. Cunha C, Panseri S, Villa O, Silva D, Gelain F. 3D culture of adult mouse neural stem cells within functionalized self-assembling peptide scaffolds. *Int J Nanomedicine*. 2011;6. doi:10.2147/ijn.s17292

12. Jonnalagadda SVR, Ornithopoulou E, Orr AA, et al. Computational design of amyloid self-assembling peptides bearing aromatic residues and the cell adhesive motif Arg-Gly-Asp. *Mol Syst Des Eng*. 2017;2(3). doi:10.1039/c7me00016b

13. Zhang S. Fabrication of novel biomaterials through molecular self-assembly. *Nat Biotechnol*. 2003;21(10):1171-1178. doi:10.1038/nbt874

14. Jonker AM, Löwik DWPM, Van Hest JCM. Peptide- and protein-based hydrogels. *Chem Mater*. 2012;24(5). doi:10.1021/cm202640w

15. Sinthuvanich C, Haines-Butterick LA, Nagy KJ, Schneider JP. Iterative design of peptide-based hydrogels and the effect of network electrostatics on primary chondrocyte behavior.

Biomaterials. 2012;33(30). doi:10.1016/j.biomaterials.2012.06.097

16. Woolfson DN. Building fibrous biomaterials from alpha-helical and collagen-like coiled-coil peptides. *Biopolymers*. 2010;94(1). doi:10.1002/bip.21345

17. Boyle AL, Woolfson DN. De novo designed peptides for biological applications. *Chem Soc Rev*. 2011;40(8). doi:10.1039/c0cs00152j

18. Altunbas A, Lee SJ, Rajasekaran SA, Schneider JP, Pochan DJ. Encapsulation of curcumin in self-assembling peptide hydrogels as injectable drug delivery vehicles. *Biomaterials*. 2011;32(25). doi:10.1016/j.biomaterials.2011.04.069

19. Branco MC, Pochan DJ, Wagner NJ, Schneider JP. The effect of protein structure on their controlled release from an injectable peptide hydrogel. *Biomaterials*. 2010;31(36). doi:10.1016/j.biomaterials.2010.08.047

20. Mitchison TJ, Biol JC, Kelleher JF, et al. Self-Assembly and Mineralization of Peptide-Amphiphile Nanofibers. *Science (80-).* 2001;294:1684-1689.

21. Eby DM, Johnson GR, Farmer BL, Pandey RB. Supramolecular assembly of a biomineralizing antimicrobial peptide in coarse-grained Monte Carlo simulations. *Phys Chem Chem Phys*. 2011;13(3). doi:10.1039/c0cp01364a

22. Zhang S. Emerging biological materials through molecular self-assembly. *Biotechnol Adv*. 2002;20(5-6). doi:10.1016/S0734-9750(02)00026-5

23. Kretsinger JK, Haines LA, Ozbas B, Pochan DJ, Schneider JP. Cytocompatibility of self-assembled β-hairpin peptide hydrogel surfaces. *Biomaterials*. 2005;26(25).

doi:10.1016/j.biomaterials.2005.01.029

24. Sawaya MR, Sambashivan S, Nelson R, et al. Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature*. 2007;447(7143). doi:10.1038/nature05695

25. Zhao YL, Wu YD. A theoretical study of β-sheet models: Is the formation of hydrogen-bond networks cooperative? *J Am Chem Soc*. 2002;124(8). doi:10.1021/ja016230a

26. Collier JH, Messersmith PB. Enzymatic modification of self-assembled peptide structures with tissue transglutaminase. *Bioconjug Chem*. 2003;14(4). doi:10.1021/bc034017t

27. Xiao X, Robang AS, Sarma S, et al. Sequence patterns and signatures: Computational and experimental discovery of amyloid-forming peptides. *PNAS Nexus*. 2022;1(5). doi:10.1093/pnasnexus/pgac263

28. Wang Y, Gao Y, Hill SE, et al. Simulations and Experiments Delineate Amyloid Fibrilization by Peptides Derived from Glaucoma-Associated Myocilin. *J Phys Chem B.* 2018;122(22). doi:10.1021/acs.jpcb.8b03000

29. Wang Y, Latshaw DC, Hall CK. Aggregation of Aβ(17–36) in the Presence of Naturally Occurring Phenolic Inhibitors Using Coarse-Grained Simulations. *J Mol Biol*. 2017;429(24). doi:10.1016/j.jmb.2017.10.006

30. Wagoner VA, Cheon M, Chang I, Hall CK. Fibrillization propensity for short designed hexapeptides predicted by computer simulation. *J Mol Biol*. 2012;416(4). doi:10.1016/j.jmb.2011.12.038

31. Nguyen HD, Hall CK. Molecular dynamics simulations of spontaneous fibril formation by

random-coil peptides. *Proc Natl Acad Sci U S A*. 2004;101(46). doi:10.1073/pnas.0407273101

32. Wang Y, Shao Q, Hall CK. N-terminal prion protein peptides (PrP(120-144)) form parallel in-register β-sheets via multiple nucleation-dependent pathways. *J Biol Chem*. 2016;291(42). doi:10.1074/jbc.M116.744573

33. Cheon M, Chang I, Hall CK. Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins Struct Funct Bioinforma*. 2010;78(14). doi:10.1002/prot.22817

34. Micsonai A, Wien F, Bulyáki, E, et al. BeStSel: a web server for accurate protein structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res*. 2018;46(W1). doi:10.1093/nar/gky497

35. Xiao X, Hall CK, Agris PF. The design of a peptide sequence to inhibit HIV replication: A search algorithm combining Monte Carlo and self-consistent mean field techniques. *J Biomol Struct and Dyn*. 2014;32(10). doi: 10.1080/07391102.2013.825757

36. Xiao X, Wang Y, Leonard JN, Hall CK. Extended concerted rotation technique enhances the sampling efficiency of the computational peptidedesign algorithm. *J Chem Theory Comput.* 2017**;**13(11). doi: 10.1021/acs.jctc.7b00714

37. Xiao X, Sarma S, Menegatti S, et al. *In silico* identification and experimental validation of peptide-based inhibitors targeting *Clostrodium difficile* toxin A. *ACS Chem Biol*. 2022;17(1). doi: 10.1021/acschembio.1c00743

38. Xiao X, Kilgore R, Sarma S, et al. De novo discovery of peptide-based affinity ligands for the fab fragment of human immunoglobulin G. *J Chromatogr A*. 2022;1669. doi: 10.1016/j.chroma.2022.462941

39. Sarma S, Herrera SM, Xiao X, Hudalla GA, Hall CK. Computational design and experimental validation of ACE2-derived peptides as SARS-CoV-2 receptor binding domain inhibitors. *J Phys Chem B*. 2022;126(41). doi: 10.1021/acs.jpcb.2c03918

40. Sarma S, Catella C, San Pedro ET, et al. Design of 8-mer peptides that block *Clostridioides difficile* toxin A in intestinal cells. *Commun Biol*. 2023;6. doi: 10.1038/s42003-023-05242-x.

Tables and Figures

Table 4.1 The sequences of the four in-silico discovered peptides in the first round of design with their corresponding Γ_{score}, $\Delta G_{binding}$ and $\lambda \times P_{aggregation}$ values computed from PepAD. The %Parallel β-sheet content for each peptide is computed from the DMD/PRIME20 simulations.

Peptides	Case	Sequences	Γ_{score}	$\Delta G_{binding}$	$\lambda \times P_{aggregation}$	DMD/PRIME20
Starting with Conf-1						
PP1	1	GDIKIVV	-12.19	-12.37	-0.18	Two-layer fibril, ~77% parallel β-sheet content
PP2	2	GNIVTFV	-17.25	-12.39	4.85	Two-layer fibril, ~100% parallel β-sheet content
Starting with Conf-2						
PP3	1	GAIDWVK	-8.28	-16.20	-7.92	Multi-layer fibril, ~100% parallel β-sheet content
PP4	1	GGIDWKI	-6.55	-15.88	-9.33	Two-layer fibril, ~46% parallel β-sheet content

Table 4.2 The sequences of the four *in-silico* discovered peptides obtained from the second round of design with their corresponding Γ_{score}, $\Delta G_{binding}$ and $\lambda \times P_{aggregation}$ values computed from PepAD. The %Parallel β-sheet content for each peptide is computed from the DMD/PRIME20 simulations.

Peptides	Case	Sequences	Γ_{score}	$\Delta G_{binding}$	$\lambda \times P_{aggregation}$	DMD/PRIME20
Starting with Conf-3						
PP5	1	ADKVMFV	-23.50	-23.68	-0.17	Multilayer fibril, ~87% parallel β-sheet content
PP6	1	GDFVKFV	-26.61	-25.96	0.64	Low fibril content
PP7	2	GNYTMFI	-30.93	-27.23	3.70	Two two-layer fibrils, ~91% parallel β-sheet content
PP8	2	ANMTVFV	-30.15	-26.99	3.16	Two-layer fibril, ~47% parallel β-sheet content

Table 4.3 Summary of computational analysis (DMD/PRIME20) and experimental measurements (FTIR, ThT, TEM and CD) of peptides PP1-PP8.

Peptide	DMD/PRIME20 simulation	FTIR (peptide concentration 10 mM)	ThT(peptide concentration 1 mM)	TEM (peptide concentration 1 mM)	CD (peptide concentration 0.2 mM)
PP1	Two-layer fibril, ~77% parallel β-sheet content (Fast aggregation)	Parallel β-sheet content	Parallel β-sheet content (slow aggregation)	Multilayer fibrils	β-sheet
PP2	Two-layer fibril, ~100% parallel β-sheet content (Delayed aggregation)	Could not be synthesized at required purity	Could not be synthesized at required purity	Could not be synthesized at required purity	Could not be synthesized at required purity
PP3	Multi-layer fibril, ~100% parallel β-sheet content (Slow aggregation)	Parallel β-sheet content	Parallel β-sheet content (slow aggregation)	Multilayer fibrils	β-sheet + random coil
PP4	Two-layer fibril, ~46% parallel β-sheet content (Slow aggregation)	Parallel β-sheet content	Parallel β-sheet content (slow aggregation)	Multilayer fibrils	β-sheet + random coil
PP5	Multilayer fibril, ~87% parallel β-sheet content (Delayed aggregation)	Not synthesized	Not synthesized	Not synthesized	Not synthesized
PP6	Low fibril content	Not synthesized	Not synthesized	Not synthesized	Not synthesized
PP7	Two two-layer fibrils, ~91% parallel β-sheet content (Fast aggregation)	Parallel β-sheet content	Parallel β-sheet content (fast aggregation)	Multilayer fibrils	β-sheet + random coil
PP8	Two-layer fibril, ~47% parallel β-sheet content (Fast aggregation)	Parallel β-sheet content	Parallel β-sheet content (fast aggregation)	Multilayer fibrils	β-sheet + random coil

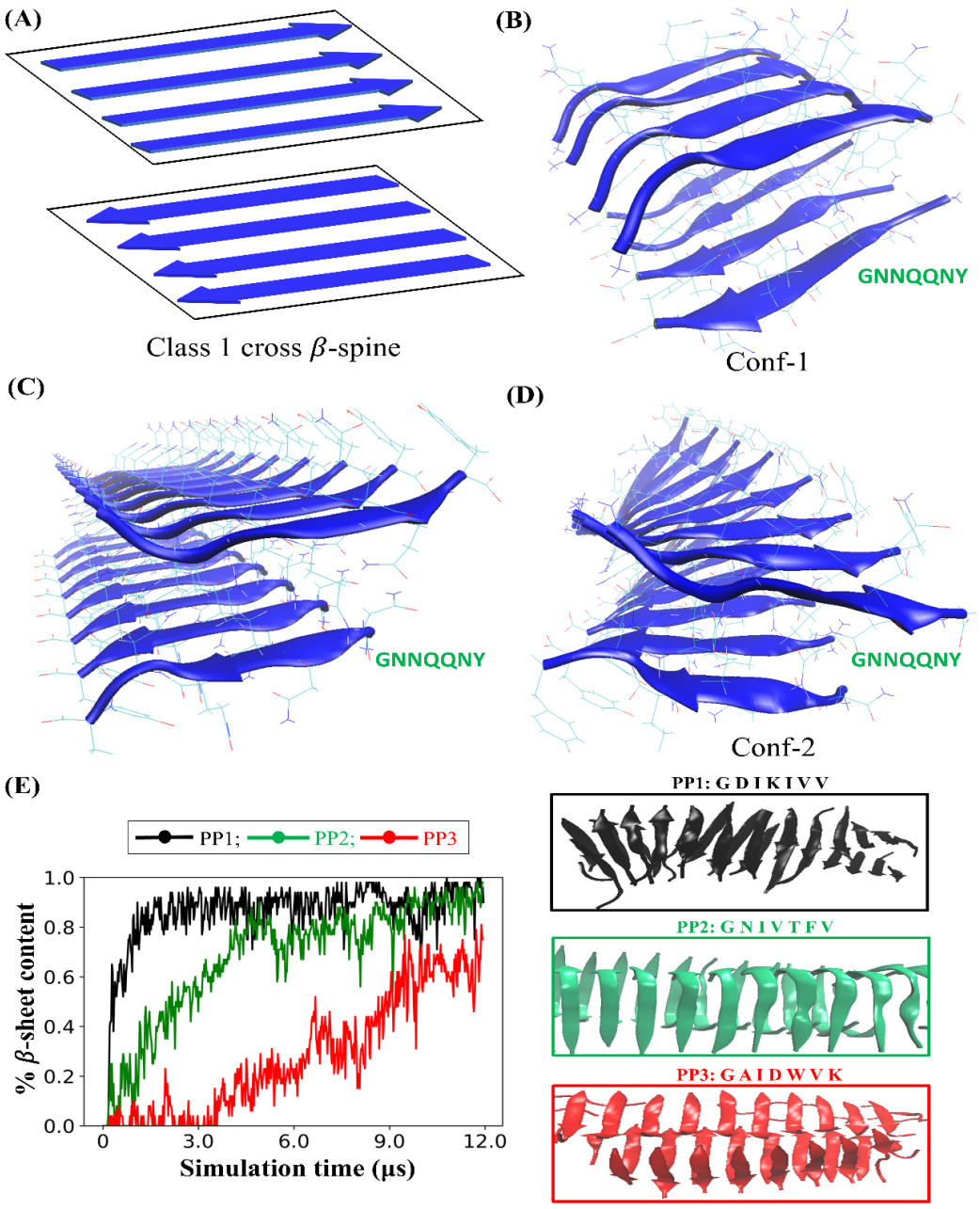


Figure 4.1 First round of design for Class 1 cross β-spine forming peptides: (A) Arrangement of peptides in the Class 1 cross β-spine, a 2-layer amyloid fibril consisting of parallel-oriented β-strands in each layer and antiparallel-oriented β-strands between the two layers. (B) Input fibril structure of peptide GNNQQNY for the PepAD algorithm (Conf-1) constructed using Discovery Studio 3.5 and Packmol packages and optimized by atomistic MD simulation in the AMBER14 package. (C) Fibril structure of peptide GNNQQNY obtained from PDB structure: 2omm (D) Input fibril structure of peptide GNNQQNY for the PepAD algorithm (Conf-2) constructed by performing a 25 ns MD simulation on structure in (C). (E) Plot of β-sheet content v/s simulation time describes the self-aggregation kinetics of peptide PP1 (sequence: GDIKIVV), PP2 (sequence: GNIVTFV) and PP3 (sequence: GAIDWVK) (left). The snapshot of the final simulation structures of peptides PP1, PP2, and PP3 indicating its fibrilization behavior is shown (right). Second

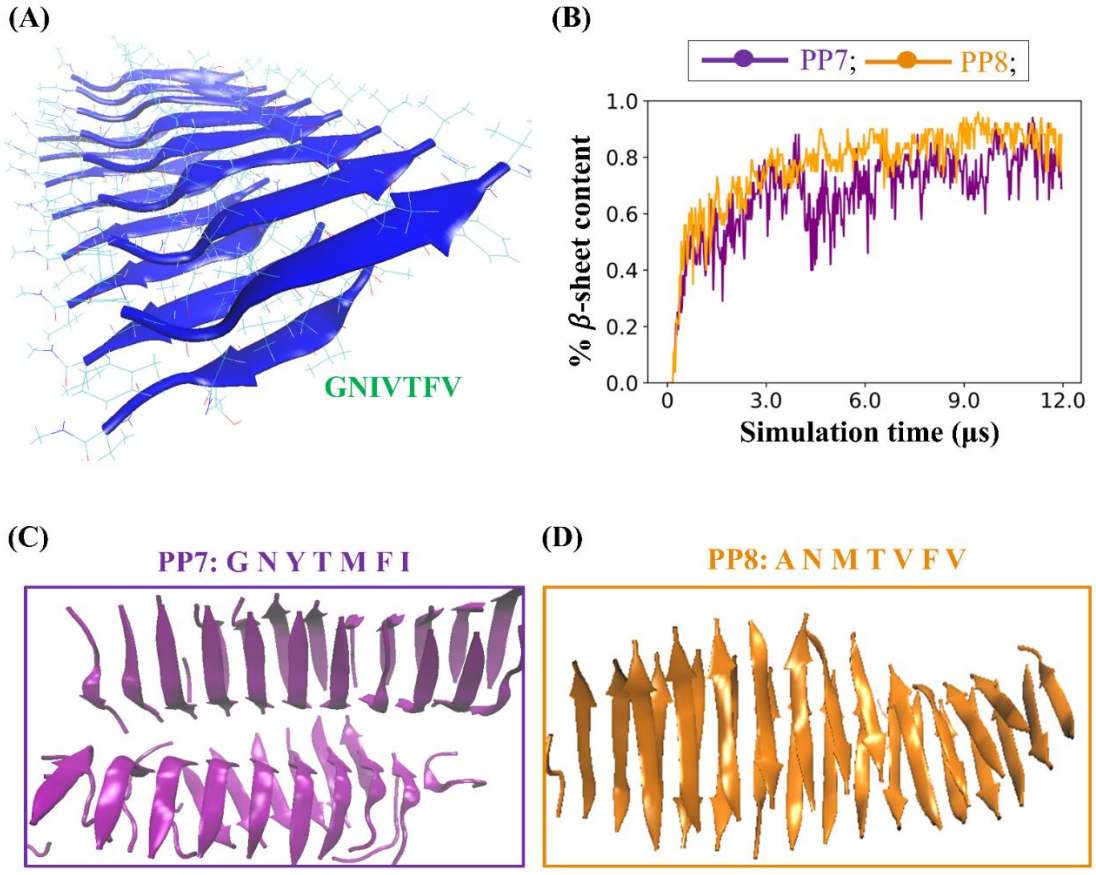


Figure 4.2 Second round of design of Class 1 cross β-spine forming peptides: **(A)** Input fibril structure of peptide PP2: GNIVTFV (with patched N- and C- termini) for the PepAD algorithm in Round 2 of design (**Conf-3**) (**B)** Plots of β-sheet content v/s simulation time describe the self-assembly kinetics of peptides PP7 and PP8. Snapshots of the final structures of **(C)** PP7 and **(D)** PP8 were obtained from the DMD simulations.

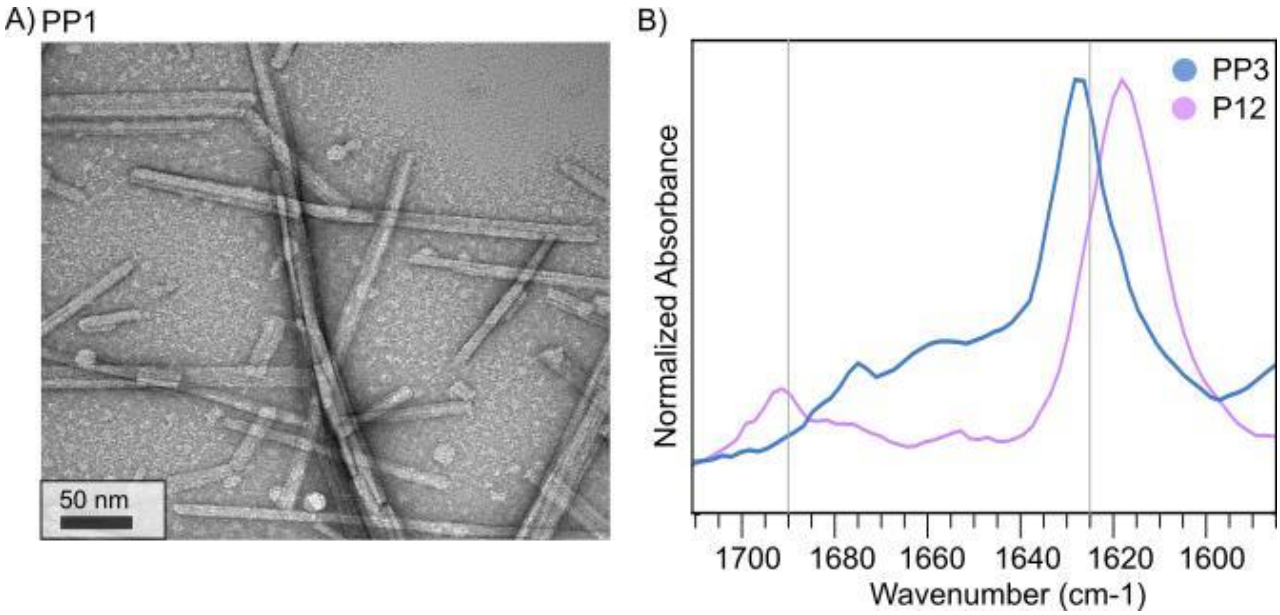


Figure 4.3 TEM Imaging and FTIR spectra **(A)** TEM Image of PP1 (**B)** FTIR spectra comparing P12 from our previous work and PP3. Note that PP3 shows an absence of a peak at 1690cm^{-1}, which corresponds to antiparallel β-sheets.

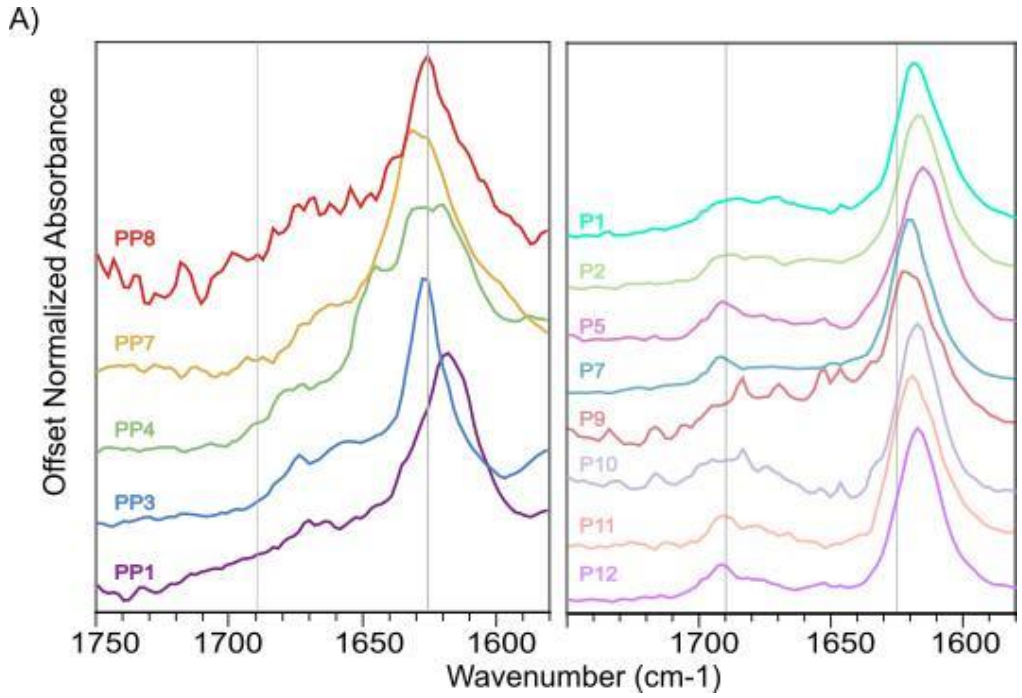


Figure 4.4 FTIR Spectra comparing peptides in our present study (left) with peptides from our previous study (right). Note that the peak at 1690 cm^{-1}, associated with antiparallel β-sheets, is absent in spectra on the left.

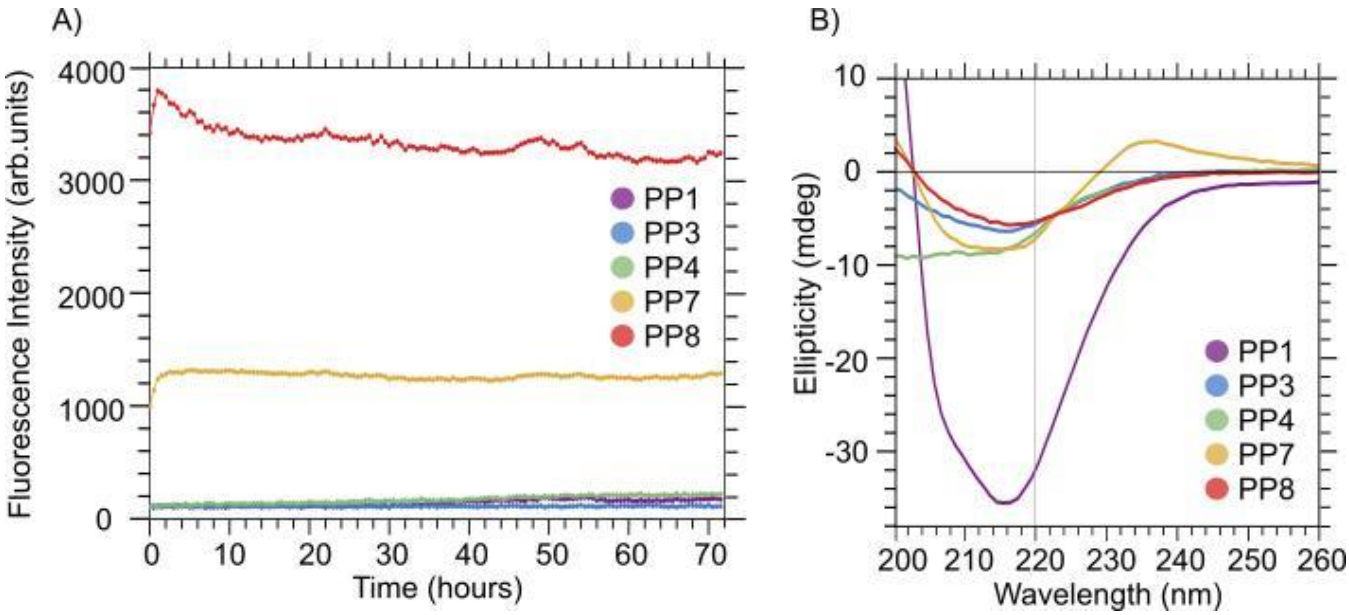


Figure 4.5 Thioflavin-T Fluorescence and Circular Dichroism **(A)** Fluorescence intensity plotted with time for peptides in our present study **(B)** Circular Dichroism experiments indicate that there is a dip at ~215nm characteristic of β-sheet secondary structure.

Supplementary Information

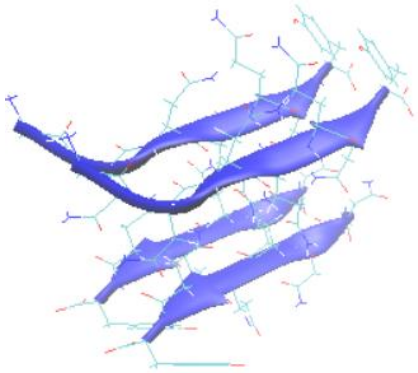


Figure 4.6 Class 1 cross-β spine containing 4 GNNQQNY peptides.

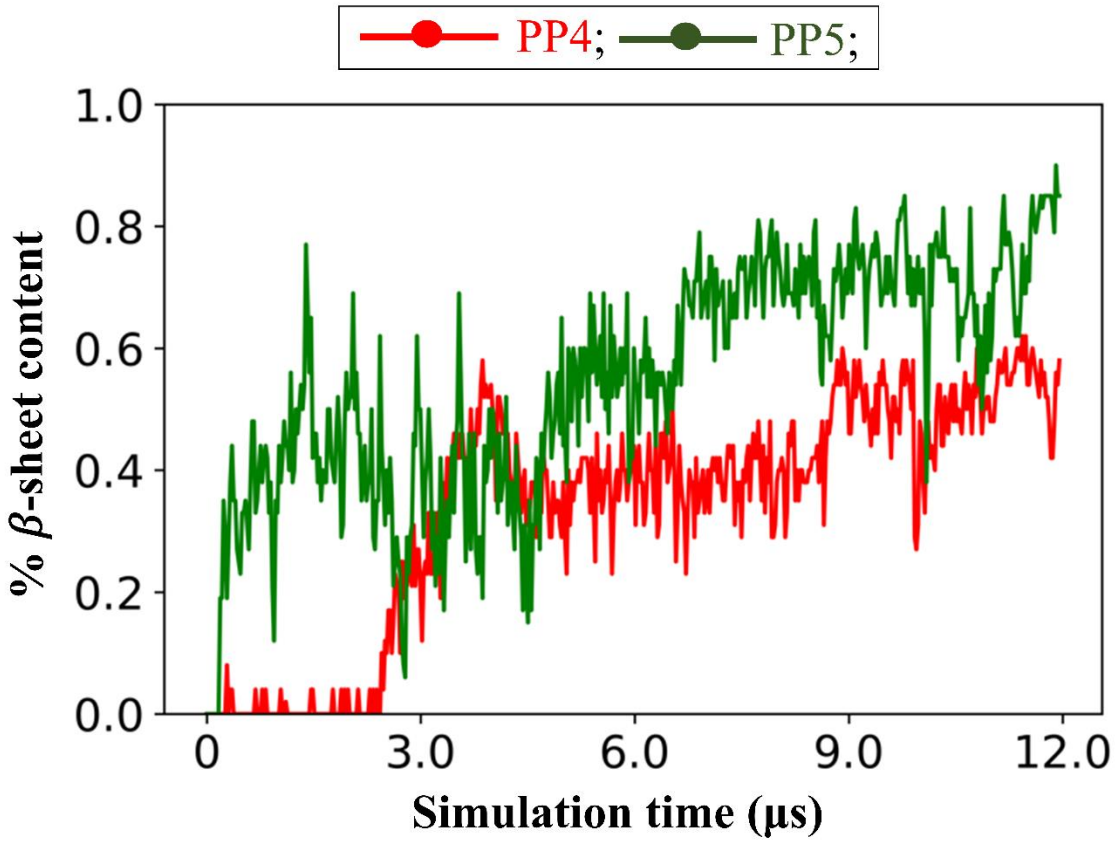


Figure 4.7 Plot of β-sheet content v/s simulation time describes the self-aggregation kinetics of peptide PP4: GAIDWVK and PP5: ADKVMFV.

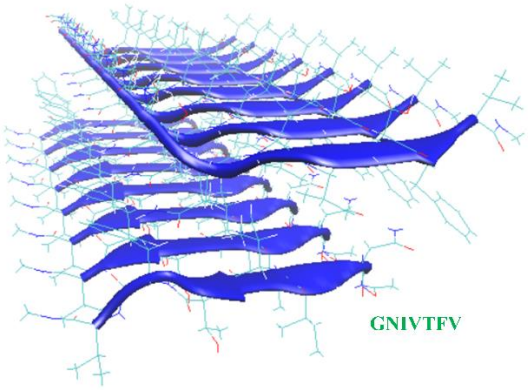


Figure 4.8 Class 1 cross-β spine structure containing 16 GNIVTFV (PP2) peptides.

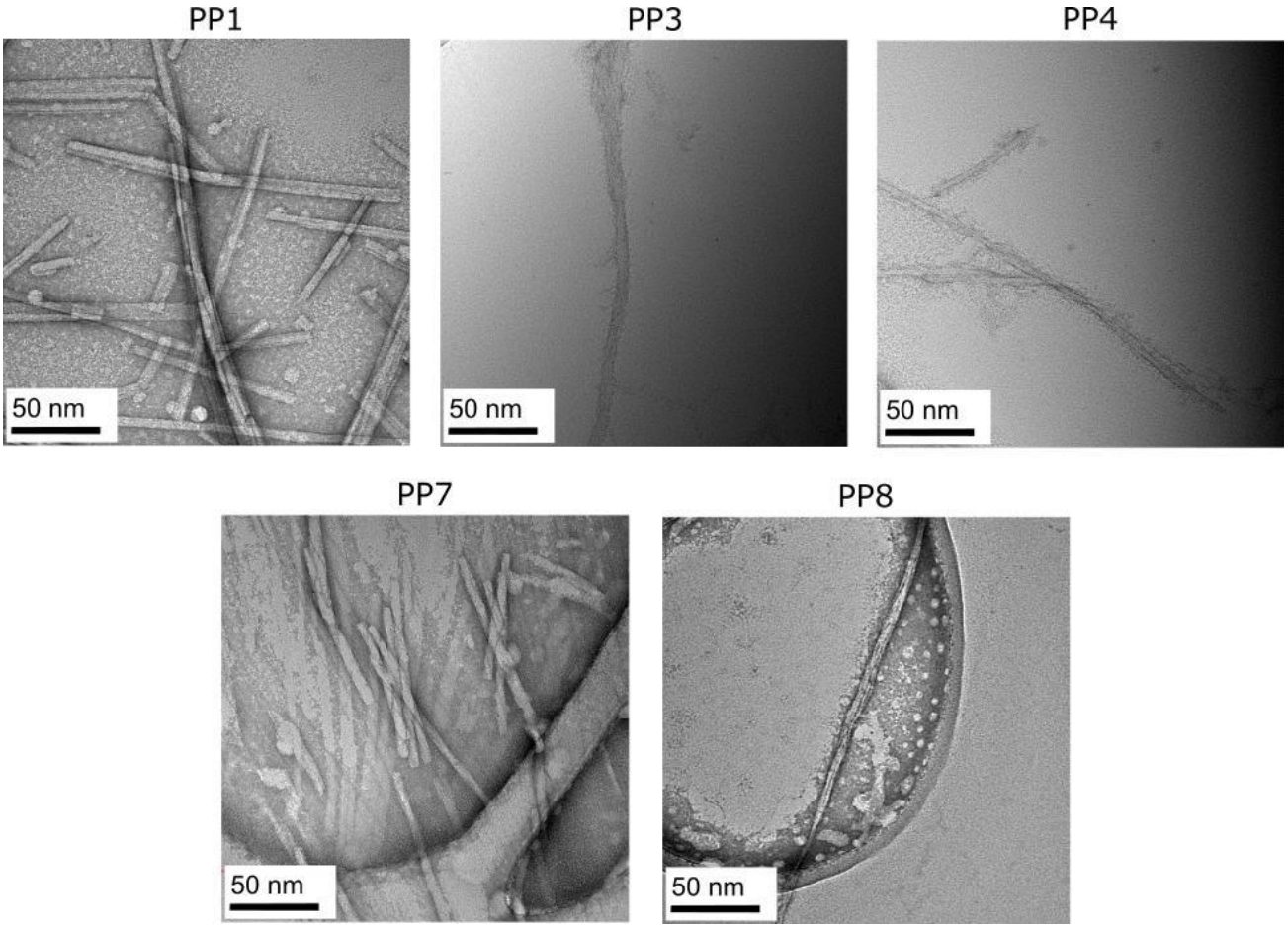


Figure 4.9 Transmission electron micrographs of negatively stained samples prepared one day following peptide dissolution at 1 mg/ml.

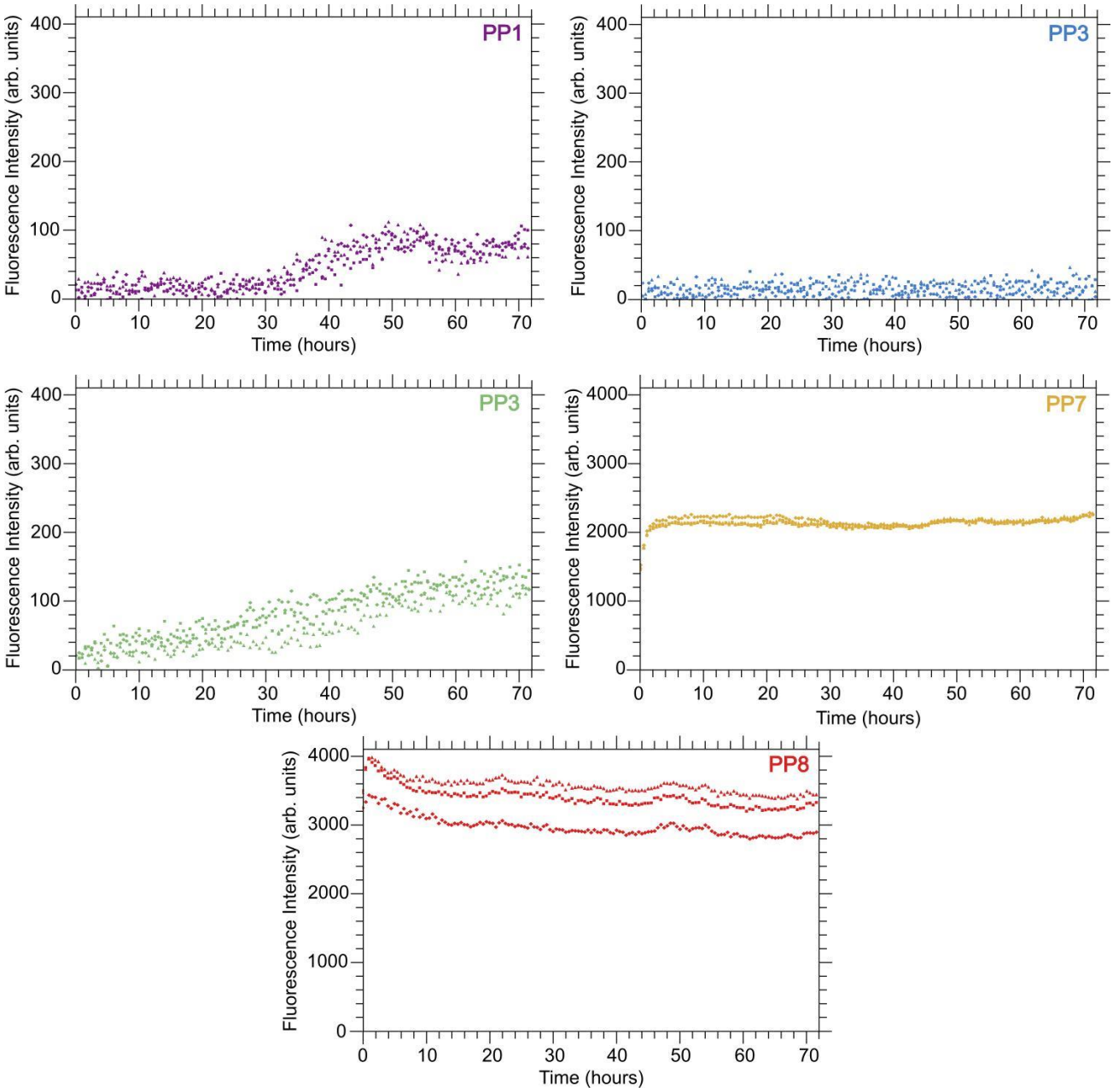


Figure 4.10 ThT fluorescence replicate curves for each of the five peptides predicted to aggregate.

Peptide Synthesis

Peptides (PP1, PP2, PP3, PP4, PP7 and PP8) were purchased from CPC Scientific, Inc. (Sunnyvale, CA). The peptides received at > 95% purity were used in experiments without modification.

Thioflavin-T fluorescence

Thioflavin-T (ThT) measurements were conducted using the procedure from our previous work[1]. Peptides were dissolved to a concentration of 2mM total peptide, 0.08mg/mL ThT, and DI water before being added to a black 96-well plate (Thermo Scientific Nunc). Peptide samples were analyzed using a BioTek Synergy H4 Microplate Reader (excitation 450nm, emission 482nm, slit bandwidth 9nm), and fluorescence intensity was recorded over 72h. We performed ThT fluorescence measurements in triplicate, with the means of the samples reported in the main text.

Fourier Transform Infrared Spectroscopy

Fourier transform infrared spectroscopy (FTIR) was performed using protocol developed in our previous study[1]. These measurements were conducted on Thermo Scientific Nicolet 6700 spectrometer with an attenuated total reflection (ATR) accessory. Peptide solutions prepared at 10mM were spotted onto the ATR accessory, and an average over 64 scans was collected after an assembly period of 72h. As PP3 and PP4 assembled slower as compared to rest of peptides, FTIR scans were collected after 6 days of assembly.

Transmission Electron Microscopy

Peptide solutions with DI water at 1mg/ml were prepared and assembled for minimum of 24h before measurements. Transmission Electron Microscopy (TEM) was conducted with a protocol similar to our previous work[1]. As PP3 and PP4 assembled slowly, an assembly period of 6 days was employed prior to TEM measurements.

Circular Dichroism Spectroscopy

Peptides were dissolved in water at a concentration of 0.2mM in DI water (De-ionized). Subsequent to dissolution, concentration was verified using UV/Vis absorbance. Following a 2h assembly period, CD was measured at room temperature with a Chirascan^{TM}-plus spectrometer (Applied Photophysics, Ltd.), following baseline correction with DI water without peptide. Quartz cuvettes with a 0.1mm path length were used. The protocol established for our previous study was employed to conduct measurements.

Comparison of secondary structure predicted by BeStSel and our current study:

BeStSel is a web server-based platform for the determination of secondary structure from circular dichroism spectra developed by Micsonai et al.[2] All tested peptides exhibit parallel β-sheet structures as evidenced by biophysical characterization techniques like negative-stain transmission electron microscopy (TEM), Fourier-transform infrared spectroscopy (FTIR), Thioflavin-T Fluorescence and Circular Dichroism (CD). We compared our interpretation of Circular Dichrosim experiments to the secondary structure predicted by BeStSel when data from

the experiment was input into it. We suggest that results are unreliable when the peptides tested have heterogenous structure.

Table 4.4 Summary of secondary structures predicted by BeStSel and secondary structures determined by our characterization techniques

Peptide	**BeStSel**	**Current Study: CD interpretation**	**Result from current study: All techniques**
PP1	100% parallel β-sheet	β-sheet	Parallel β-sheet
PP3	12.2% helix, 32.1% antiparallel β-sheet, 5.7% parallel β-sheet, 49.9% other	β-sheet + random coil	Parallel β-sheet (slow aggregation)
PP4	20.1% helix, 64.5% antiparallel β-sheet, 15.3% other	β-sheet + random coil	Parallel β-sheet (slow aggregation)
PP7	44.3% helix, 55.7% antiparallel β-sheet	β-sheet + random coil	Parallel β-sheet
PP8	9.1% helix, 9.1% antiparallel β-sheet, 18.3% parallel β-sheet, 57.2% other, 6.4% turn	β-sheet + random coil	Parallel β-sheet

References

1. Xiao X, Robang AS, Sarma S, et al. Sequence patterns and signatures: Computational and experimental discovery of amyloid-forming peptides. *PNAS Nexus*. 2022;1(5). doi:10.1093/pnasnexus/pgac263.
2. Micsonai A, Wien F, Bulyáki É, et al. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res.* 2018;46(W1). doi: 10.1093/nar/gky497.

Chapter 5 Conclusion and Future Work

5.1. Conclusion

Chapter 2. In this chapter, we described our DMD/PRIME20 simulation package: the program architecture, and the steps/commands for installation, implementation, and application. The DMD/PRIME20 software combines the advantages of both a simple but relatively-realistic coarse-grained force field and the DMD discontinuous molecular dynamics method to simulate the complete aggregation process for short peptides. Simulations of hundreds of μs can be performed within a month as a serial job on a CPU computer workstation. This capability allows many simulations to be run at the same time on a CPU computer system. In addition, most of the aggregation simulations can be completed in less than one month. Our software is easy to use, as users only need two files to run a simulation and short commands to complete the data analysis. The analysis package attached to the software enables investigation of aggregation kinetics without any other analysis tools, except VMD for visualization.

Chapter 3. In this chapter, we use DMD/PRIME20 simulations to provide insight at the molecular level into the fibrillation of the αSynuclein fragment variants (P1, P2, P3, P3Next, including both WT and mutated sequences). We have shown that deleting P1 and P2[1], or mutating Leu 38 to Met in P1[2], inhibits amyloid formation of αSyn-FL by a combination of different effects. Overall, we draw two conclusions from evaluating the fibrillar structures resulted from our preliminary simulations that formed in the simulations of the peptide variants (WT and mutated) in both homogeneous and mixed (P3Next variants with C-terminal fragment αSyn[120-140]) systems. First, WT fragments have a high propensity to form fibrillation-prone structures such as parallel β-sheets and β-hairpins. Second, the mechanism by which L38M prevents the aggregation of αSyn-FL is not by reducing the aggregation propensity of the sequence itself, but by altering its propensity to form fibrillation-prone structures and the interactions that occur between the P1

region and more distant residues in the sequence in the early stages of amyloid formation. The homogeneous system simulations show that L38M substitution in the peptide fragment reduces the formation of amyloid-prone structures (parallel β-strands for P1 variants and β-hairpins for P3Next variants) compared to the WT fragments. In contrast to L38M, the other single-point substitutions, P1-L38A and P1-V40A, increased parallel β-fibril formation for P1 variants and had higher β-hairpin ratios in their P3Next fibrils. The mixed system simulations showed that L38M enhances interdomain interactions between the P1/P2 region and the last 21 residues in the C-terminal. These interdomain interactions disrupt the formation of amyloid-prone structures, potentially altering the protein fibrillation pathway by stabilizing monomeric conformations or promoting less amyloidogenic oligomer structures. At the time of writing this thesis, additional simulations are still running and the update will be available when the manuscript is published.

Chapter 4. In this chapter, we identified peptides that self-assemble to form fibrils composed of parallel β-sheets inspired by the Class-1 cross-β spine structure described by Sawaya et al.[3] The model structure contains two β-sheet layers with parallel-oriented β-strands in each layer, and antiparallel-oriented β-sheets between the two layers. As a first step, we set out to design peptides that form fibrils with parallel β-sheets, regardless of the relative orientation of peptides in neighboring layers. This was accomplished by incorporating the PepAD workflow, DMD/PRIME20 simulations and experimental characterization. By using the PepAD algorithm coupled with DMD/PRIME20 simulations, we performed two rounds of designs with three different starting backbone scaffolds (**Conf-1**, **Conf-2** and **Conf-3**) to obtain a library of 7-mer amyloid-forming peptides that could potentially assemble into parallel β-sheets. Five out of the eight peptides were synthesized and experimentally tested, and all of them aggregated to form parallel β-sheets.

5.2. Recommendation for Future Work

Improving Efficiency and Accuracy for DMD/PRIM20

DMD/PRIME20 has been a useful tool with remarkably competitive computational efficiency. The longest simulation spanned 657μs for a system of 192 Aβ(16-22) peptides with crowders[4]. We believe that the efficiency of our software can be improved by taking advantage of code parallelization. This would allow us to reduce computer time, increase simulation speed to access long-time scale phenomena, and utilize hardware resources. Preliminary parallelization of DMD/PRIME20 was conducted by a previous student in the Hall group, Dr. Yiming Wang. He applied the event-decomposition parallelizing methods by Khan and Herbordt[5] and Shirvanyants et al.[6], in which collision events are distributed to slave nodes for calculation of collision dynamics. Each event was sent to a slave node following the time order on the event list. Unfortunately, this first approach to the parallelization of DMD/PRIME20 didn't increase the efficiency of our model. The inefficiency was caused by many factors. First, the master node had to be in idle model when the slave nodes did their tasks. If the slave nodes were functioning, the master node would have to stop and wait for the results from slave nodes. The reason for this waste of simulation time is because events in DMD simulation must happen in the correct time order as. results from an event can affect the next events. Therefore, we had to pause the workflow and check results returned from the slave nodes in time order to make sure there was not an invalid event being accepted. Second, we didn't target the most expensive part of DMD simulation, which is the whole system update. Consequently, the most expensive task was still conducted in serial by the master node.

A newer and better approach would be to parallelize the system update step. In this case, the event dynamics and partial system updates after each collision would still happen in serial on the master node. The parallelization would occur when there is a system update, which is when

the collision times need to be recalculated for all beads in the system. There are two ways to parallelize this step. The first approach is to distribute the task to slave nodes by bead id number. Groups of particles could be sent to slave nodes in id number from bead 1 to the last bead in the system. Then the slave nodes would calculate new collision times for the beads assigned to them. This approach is relatively simple, so it could be tried first to validate our suggestion that parallelizing system update step would improve efficiency. The second approach is to distribute the particles to slave nodes by spatial decomposition. The simulation space would be divided into small cells that contain a group of neighboring beads. Each group of neighboring beads in one cell would be sent to a slave node for new collision time calculations. A potential advantage of this approach is it allows us to search for new collision times between neighboring beads. This would reduce computational time and communication time as a slave node only needs to access the coordinates and velocities and perform calculations for the beads assigned to it.

Another improvement that we suggest for possible future work is to include the cis-configuration of proline to the force field. As we stated in Section 2.2 of Chapter 2, only the trans-configuration of proline is included in our force field. Therefore, the model of peptides containing proline is not as accurate as it is for the peptides without this residue. We suggest getting more information about the cis- and trans- configuration of proline. If the transition between the two configurations hardly happens then we can treat the cis-configuration of proline as an individual residue and derive an additional set of interaction parameters for the PRIME20 force field. If there is a transition between cis- and trans-configuration, the new potential energy of proline will need to be calculated to account for the transition, for example a square-shoulder or a double-well potential.

5.3 References

1. Doherty, C. P. A. *et al.* A short motif in the N-terminal region of α-synuclein is critical for both aggregation and function. *Nat Struct Mol Biol* **27**, 249–259 (2020).
2. Ulamec, S. M. *et al.* Single residue modulators of amyloid formation in the N-terminal P1-region of α-synuclein. *Nat Commun* **13**, 4986 (2022).
3. Sawaya, M. R. *et al.* Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature* **447**, 453–457 (2007).
4. Latshaw, D. C., Cheon, M. & Hall, C. K. Effects of Macromolecular Crowding on Amyloid Beta (16–22) Aggregation Using Coarse-Grained Simulations. *J. Phys. Chem. B* **118**, 13513–13526 (2014).
5. Khan, Md. A. & Herbordt, M. C. Parallel discrete molecular dynamics simulation with speculation and in-order commitment. *Journal of Computational Physics* **230**, 6563–6582 (2011).
6. Shirvanyants, D., Ding, F., Tsao, D., Ramachandran, S. & Dokholyan, N. V. Discrete Molecular Dynamics: An Efficient And Versatile Simulation Method For Fine Protein Characterization. *J. Phys. Chem. B* **116**, 8375–8382 (2012).