Approaches to Studying Peptide Aggregation, Polymorphism and Thermodynamics

By

SRINIVAS CHAKRAVARTHY MUSHNOORI

A dissertation submitted to

The School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy in

Chemical and Biochemical Engineering

Written under the direction of Meenakshi Dutt

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER 2020

ABSTRACT OF THE DISSERTATION

Approaches to Studying Peptide Aggregation, Polymorphism and Thermodynamics

By

Srinivas Chakravarthy Mushnoori

Dissertation Director: Dr. Meenakshi Dutt

Harnessing the self-assembly of peptide sequences has demonstrated great promise in the domain of creating high precision shape-tunable biological materials. The unique properties of peptides allow for a building blocks approach to materials design. In this proposal, self-assembly of mixed systems encompassing two peptide sequences with identical hydrophobic regions and distinct polar segments is investigated. The two peptide sequences are diphenylalanine and phenylalanine-asparagine-phenylalanine. The dissertation aims to examine the impact of molecular composition (i.e, the total peptide concentration and the relative tripeptide concentration) on the morphology of the self-assembled hybrid biological material. We report a rich polymorphism in the assemblies of these peptides and explain the relationship between peptide sequence, concentration and the morphology of the supramolecular assembly. We discuss three techniques to explore the phase space of this morphological diversity: Classical Coarse Grained Molecular

Dynamics (CGMD), Replica Exchange Molecular Dynamics (REMD), and a hybrid Workflows/Machine Learning framework for Targeted Design of Supramolecular Assemblies, which we have called PACE: Pipelines for Automation of Compliance-based Elimination. Classical CGMD establishes the polymorphism in the phase space, REMD sheds light on the thermodynamics of assembly and helps determine the most likely thermodynamic states, and the PACE framework provides a workflow to automate the simulation, detection and screening of configurations that can target specific morphologies.

Acknowledgement

Over the course of my doctoral study, I have been extremely fortunate to have had many excellent mentors. I gratefully acknowledge Prof. Meenakshi Dutt, my advisor and dissertation director. She has been nothing but supportive and encouraging during my research. Her insightful comments and guidance have been critically important not only to the work presented in this dissertation, but my general approach to problem solving and research. My tenure in her group has taught me, beyond just the science, *how* to think, and solve problems with a careful, meticulous, and intuitive approach and an open mind.

I would like to extend my most sincere gratitude to Prof. Shantenu Jha. He has been a true ally to me, in every sense of the word, throughout my tenure as a student at Rutgers. I hung on to his every word, did my best to take his advice (and his rebukes) to heart and follow his example. I have an extremely long way to go in emulating him, and I'm still trying. Without his insight, support, encouragement, and patience, I simply would not be here today.

I would like to thank Prof. Matteo Turilli, Dr. Andre Merzky and Dr. Vivek Balasubramanian for their contributions to this dissertation. Without their support at the software and implementation levels, this work would not have been possible. My colleagues Kassandra Schmidt and Ethan Zang are gratefully acknowledged for their active contributions to this dissertation and their enthusiastic support. I thank Prof. Vikas Nanda for his co-authorship on our manuscript entitled *"Designing phenylalanine-based hybrid*

biological materials: controlling morphology via molecular composition", (Org. Biomol. Chem. 2018 Apr 4;16(14):2499-2507) as well as his valuable insight. This manuscript forms the bulk of the second chapter of this dissertation.

I extend my deepest gratitude to my family and friends: my parents, Padma and Shivkumar Mushnoori, my brother, Sriharsha, and my wife, Nimesha Tadepalle. They have been my rock, my grounding, and my anchor during the most strenuous and stressful periods of my time as a doctoral student. My friends Piyusha Gade and Piyush Deshmukh have been a constant source of inspiration, encouragement, and reassurance, and they have my sincere thanks.

Funding from the Molecular Science Software Institute (MolSSI, NSF grant ACI-1547580) fellowship is gratefully acknowledged. Funding from the National Science Foundation (DMR-1654325) is gratefully acknowledged.

I would like to extend my most sincere thanks to my committee members: Prof. Masanori Hara, Prof. Wilma Olson, Prof. Shantenu Jha, and Prof. Meenakshi Dutt for taking time out of their schedules to serve on my Dissertation committee.

Table of Contents

List of Figures

List of Tables

Chapter I: Introduction and Objectives

Recent advances in novel biological materials via peptide self-assembly has been driven by their potential use in a wide range of applications, such as neurodegenerative disease treatment [1-6], therapeutics[6-11], peptide-based electronics[12-15] and nanobiomaterials[16-24]. Earlier studies have reported peptides to assemble into nanofibers[25-29], nanotubes[30-37], vesicles[38-42] and nanosheets[41, 43]. Peptides are composed of a combination of any number of the 20 amino acids. This allows for a unique building blocks approach to biological materials design, providing a high degree of control over the morphology of supramolecular peptide assemblies.

Among the various peptide sequences employed for creating biological materials, diphenylalanine has been extensively studied using experimental approaches[30, 32, 37, 44-49] and computational techniques[48, 50, 51] . It is a small and stable peptide that has been reported to form supramolecular assemblies including vesicles and nanotubes[50-53] . More recently, diphenylalanine derivatives such as triphenylalanine (FFF) and diphenylalanine-fluorenylmethyloxycarbonyl chloride (FF-Fmoc) have been explored for their potential to self-assemble into nanostructures. These efforts have found a large diversity of assemblies including nanowires, ribbons and nanofibers[54].

The creation of nanostructures with specific control on their material characteristics can be fulfilled by using a variety of peptide sequences. The contribution of each sequence and the synergistic interplay between their molecular characteristics can generate materials with specific sequence-property relations. An earlier computational study reported mixtures of diphenylalanine and triphenylalanine to generate toroidal structures[54], which

were verified experimentally. Yet, there is a dearth in the understanding of sequence-property relations of biological materials encompassing distinct peptide sequences. In addition, very little is known about the mechanics of the assembly of distinct peptide sequences to form a hybrid biological material with a target property.

In this dissertation, we examine the impact of molecular composition on the properties (specifically, morphology) of self-assembled hybrid biological materials. For simplicity, we focus on diphenylalanine (FF) and its derivative phenylalanine-asparagine-phenylalanine (FNF). The derivative changes the sequence of diphenylalanine by the addition of asparagine between the two phenylalanine groups. In addition, these two sequences allow us to examine the impact of differences in the polar groups of the peptide species on the self-assembled hybrid material morphology. In **Chapter 2: Phase Space Exploration via Coarse Grained Molecular Dynamics**, we employ Coarse Grained Molecular Dynamics simulations to vary the molecular composition through the total concentration of the peptides and the relative concentration of the tripeptide (FNF). We find that the total concentration of the peptides along with the relative concentration of the tripeptides can control the morphology of the hybrid assemblies. Next, in **Chapter 3: Approaching Aggregation Polymorphism with Replica Exchange Molecular Dynamics**, we explore the applicability of Replica Exchange to predict the correct morphology for a given region of the phase space. Replica Exchange is an ensemble method that was first devised as an extension of Monte Carlo simulations of chemical systems, but quickly found use in MD simulations as well. Multicomponent systems are prone to kinetic traps due to the complexity in the inter-component interactions, making

them particularly well suited to study via ensemble-based approaches such as Replica Exchange. The method provides these systems with opportunities to overcome these kinetic traps by performing Monte Carlo moves between copies of the same molecular systems that have already escaped those specific traps. In **Chapter 4: Targeted design of supramolecular peptide assemblies**, we explore the use of a hybrid workflows/Machine Learning approach to target the desired morphologies by automating and concurrently executing multiple simulation-analysis pipelines. Machine Learning (ML) is a family of rapidly evolving techniques that utilizes "neural networks", a family of algorithms that vaguely mimic the human brain to perform pattern recognition. ML approaches combined with workflows are an extremely powerful approach for continuous, automated analyses of systems. In **Chapter 5: Conclusions and Future Directions,** we summarize our results and discuss ongoing work and immediate future plans primarily pertaining to the extension and further development of the framework discussed in Chapter 4.

CHAPTER II: Phase Space Exploration via Coarse Grained Molecular Dynamics

Methods

The assembly of diphenylalanine (FF) and phenylalanine-asparagine-phenylalanine (FNF) is promoted by the hydrophobic effect. We used the Molecular Dynamics (MD) simulation technique to capture the dynamics of the aggregation process along with the structure and morphology of the hybrid peptide materials. This required simulating the aggregation of several hundreds of molecules that were dispersed in an aqueous environment. An all-atom representation of the molecules would have been too demanding in terms of computational resources in resolving the temporal scales associated with the assembly process along with capturing the multiscale structural characteristics of the self-assembled material. To circumvent this difficulty, we employed coarse-grained representations of the peptides.

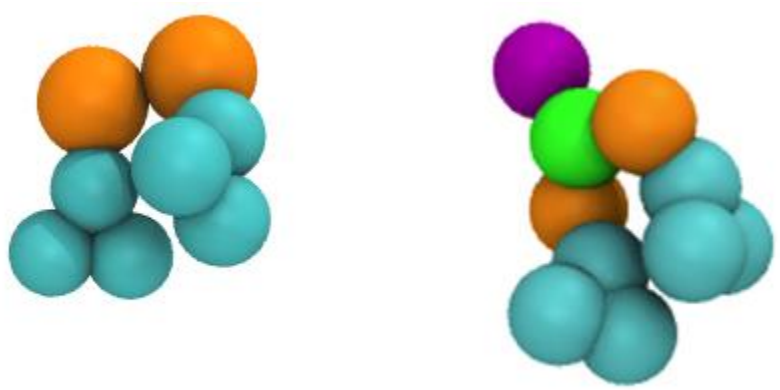

Figure 1: Coarse Grained representations of the two peptide sequences used. Left: Diphenylalanine (FF), right: Phenylalanine-Asparagine-Phenylalanine.

For this study, we used the Martini v2.2 coarse-grained (CG) force field[55] for biomolecules[56, 57], an established, generalizable force field that has been demonstrated to capture the behavior of a wide range of biomolecules. Atoms of the peptides were mapped using the Martini CG model[57] Figure 1 shows the coarse grained representations of both peptide sequences used. Particle type P5 was used for the backbone of each residue, as well as for the sidechain of asparagine. Three SC5 particles were used to represent the aromatic sidechain of phenylalanine. Since the peptides are assumed to be in a zwitterionic state, charges of +1e and -1e were assigned to the N and C termini respectively. The CG mapping details are represented graphically in Figure 2. Van der Waals forces were represented by a 12-6 Lennard-Jones approximation, while electrostatics were calculated using a shifted Coulomb potential. All cutoffs were maintained at 1.2nm in accordance with the Martini CG model.

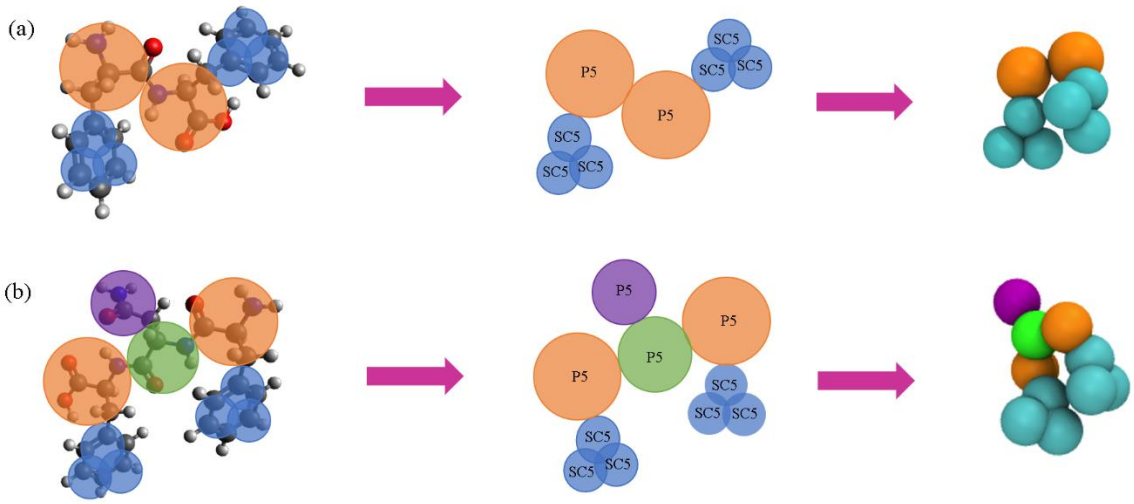

Figure 2: Coarse-grained mapping details for (a) diphenylalanine (FF) and (b) phenylalanine-asparagine-phenylalanine (FNF). Coarse-grained representations of the peptides in the Molecular Dynamics simulations shown on the right side. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

Simulations were performed using the GROMACS package (v5.2.1)[58-62] . The simulations were set up by using the inbuilt GROMACS functions to insert the FF and FNF molecules into the simulation box and solvate them. Since the system is charge neutral, no counterions are required. To prevent physically unrealistic freezing of the Martini water[55], a 0.1 mole fraction of antifreeze particles was maintained. The system was then energy minimized by the steepest descent integrator to remove any overlaps between particles. This was followed with a 500 ps (effective time) equilibration followed by a production simulation for 4 μs (effective time). The duration of the production simulation was selected so as to generate a single, stable self-assembled nanostructure.

All simulations were carried out in the isothermal-isobaric ensemble. The temperature was maintained at 310K[50, 54] using velocity rescaling with a stochastic term to ensure correct sampling[63, 64]. Pressure was maintained at 1 bar using the Parrinello-Rahman barostat[65-67]. The timestep was maintained at 25 fs and all simulations were run for an effective time of 4 μs. The neighbor lists were updated every 0.25 ps with a 1.2 nm cutoff. The particle trajectories were sampled every 0.25 ns for analysis and visualization. The LINCS algorithm was employed to constrain bond lengths in the aromatic phenylalanine sidechains.

All the simulations were carried out using systems encompassing a total of 500 peptide molecules. The peptide molecules were assumed to exist in a zwitterionic state. The total peptide concentration was controlled via pre-selecting the box size and appropriately populating the box with solvent beads. The relative tripeptide concentrations were implemented by varying the ratio of dipeptides (FF) to tripeptides (FNF), while still maintaining the total number of peptide molecules to 500.

We examined 5 total peptide concentrations (0.1, 0.15, 0.2, 0.25 and 0.3 peptides/nm^3) and 11 relative tripeptide concentrations (ranging from 0% to 100%). The results for each system was based upon 10 independent particle trajectories. Hence, the study is based upon a total of 550 simulations. While interpreting timescales in simulations that use the Martini coarse-graining scheme, a four-fold speedup is assumed in the diffusion dynamics as compared to real systems, since due to the Martini force field's reduced frictional component from the removal of atomistic degrees of freedom leads to a

smoother energy landscape[55]. Therefore, unless explicitly stated, we used an "effective" time instead of the simulation time for the remainder of this discussion.

Simulation Results

All molecular compositions explored are observed to spontaneously assemble into nanostructures encompassing bilayers. The assembly is driven by the hydrophobic effect, wherein the hydrophobic aromatic rings of the phenylalanine side chains minimize their interface with the solvent by preferentially interacting with other aromatic side chains. In addition, the hydrophilic backbones and asparagine side chains preferentially interact with the solvent particles. These simultaneous processes drive peptide aggregation, which continues until a single, stable aggregate forms for the duration of the simulation[50, 54, 68-72].

The morphology of the hybrid nanostructures can be classified into three categories by their relationship in the three spatial dimensions. We classify a structure as "open" in a given spatial direction if one of the following criteria are met: (i) it extends infinitely (through periodic walls) along that direction, or (ii) if it does not extend infinitely in a given direction, two edges of the bilayer perpendicular to that direction are exposed to the solvent. Therefore, if the bilayer of the nanostructure is open in two dimensions, it is classified as a lamella, it may extend infinitely along two axes, or along one axis but have edges exposed to the solvent. If the bilayer of the nanostructure is closed in all three spatial dimensions (i.e. does not extend infinitely through periodic walls), it is classified as a

vesicle. In the event that the bilayer of the nanostructure is closed in two dimensions but extends infinitely in the third dimension, it is classified as a nanotube.

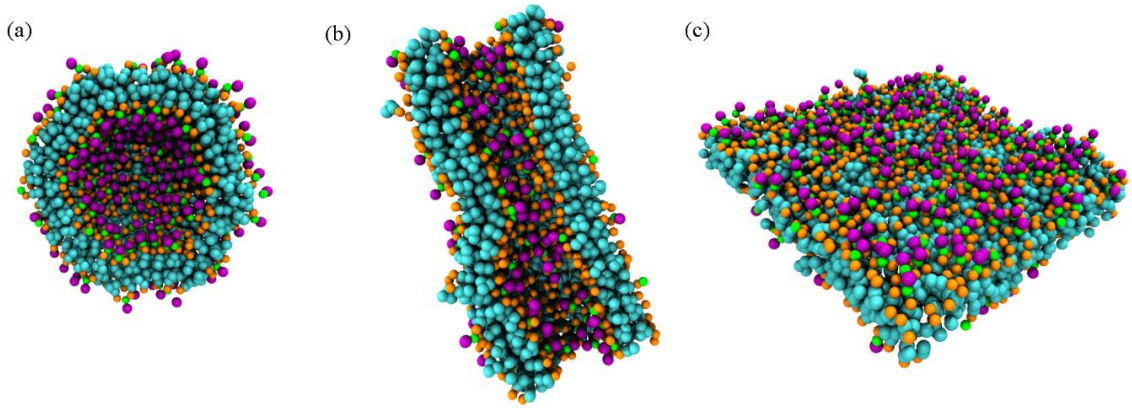

Figure 3: Self-assembled hybrid nanostructures: (a) diametrical cross-sectional view of a vesicle, (b) cylindrical axis cross-sectional view of a nanotube, and (c) lamellar bilayer. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

Representatives of these assemblies are shown in Figure 3a-c. The cross-sectional views of the vesicle (Figure 3a) and the nanotube (Figure 3b) indicate the presence of an internal cavity in the nanostructures. The internal cavity is stabilized by the interface between the solvent and the hydrophilic backbones and asparagine side chains.

There is a fourth category of self-assembled nanostructures, which is not representative of an organized morphology. These nanostructures are classified as “disordered” assemblies. Whereas the peptide species in these nanostructures are organized

into bilayers at a local scale, their morphology cannot be classified as either vesicles, nanotubes or lamellae. Disordered assemblies are handled separately at the end of this chapter. The morphologies of the ordered nanostructures for different molecular compositions are summarized in Table 1.

The phase space of the morphology of the nanostructures as a function of the total concentration of the peptides and the relative concentration of the tripeptides in Table 1 is further decomposed into four histograms, shown in Figures 4, 5 and 6. Each histogram summarizes the statistics for a given morphology of the hybrid nanostructure. Figure 4 shows the frequency of occurrence of vesicles in the ten independent trajectories for every molecular composition examined in this study. The molecular composition constitutes the total concentration of peptides and the relative concentration of the tripeptides. Vesicles are consistently observed for the lower total concentrations of the peptides (0.1-0.15 peptides/nm^3).

Total Peptide Conc. (peptides/nm^3)	0% Tripeptide				10% Tripeptide				20% Tripeptide				30% Tripeptide				40% Tripeptide				50% Tripeptide				60% Tripeptide				70% Tripeptide				80% Tripeptide				90% Tripeptide				100% Tripeptide			
	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL	U	EV	NT	BL
0.3			7	3		1	4	5			6	4		1	4	5			4	6	1		2	7				10			1	9	1		1	8	3			7	3			7
0.25		3	4	3		3	3	4		4	3	3		1	5	4			3	7		1	2	7	1	2		7		1	3	6				10			1	9	3			7
0.2		8	2			5		5		6	3	1		9	1			7	1	2		3	1	6	1	4		5		1	1	8		3	1	6		2		8	1	2	1	6
0.15		10				10				9	1			10				8	1	1		8	1	1		9		1		8		2		5		5	1	5		4		6	1	3
0.1		10				10				10				10				10				10				9		1		10					10			10			3	7		

Table 1: Summary of the morphologies for the hybrid nanostructures for different values of the total and relative concentration of the peptide molecules. Note: EV: elongated vesicle, NT: nanotube, BL: bilayer (lamellar) and U: Unstructured/Disordered. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

As the total concentration of the peptides increase, the propensity of the peptides to form vesicles decreases. This trend is consistent across all relative tripeptide concentrations, ranging from 0% (a pure FF system) to 100% (a pure FNF system). Increasing the total concentration of peptides, however, increases the propensity of the peptides to form lamellar bilayers. This effect is particularly pronounced at 0.25 and 0.3 peptides/nm^3, as shown in Figure 5. Lamellae are also observed at medium values of the total concentration of the peptides (0.2 peptides/nm^3) but only for high relative tripeptide concentrations (70% - 100%).

Nanotubes are a relatively infrequent occurrence, but can be observed for higher values of the total concentration of peptides (0.2-0.3 peptides/nm^3) and low-to-medium relative tripeptide concentrations (20%-50%), as demonstrated in Figure 6. Nanotubes are also observed when the net peptide concentrations are low and the relative tripeptide concentrations are high. For example, at 100% tripeptide with total concentrations of 0.1, 0.2, 0.25 and 0.3 peptides/nm^3. However, these are relatively rare occurrences.

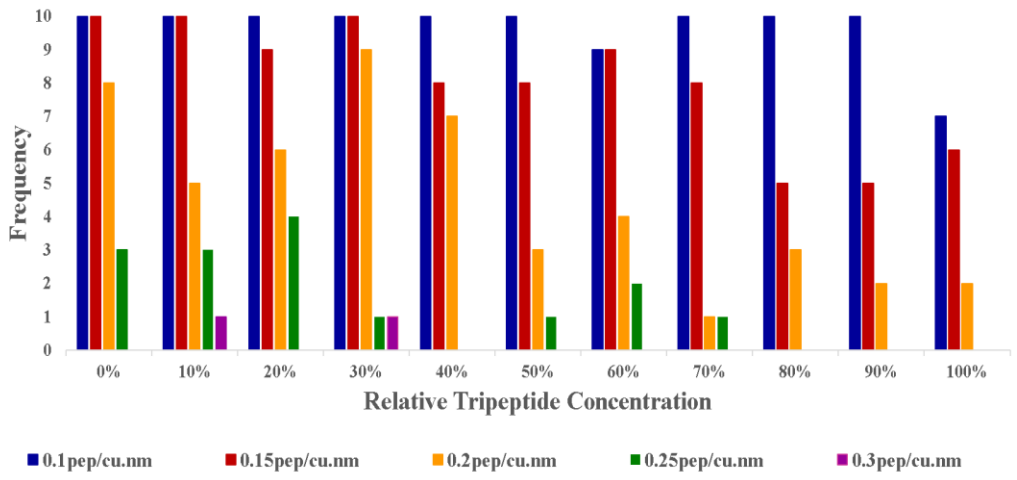

Figure 4: Histogram of the occurrence of vesicles in mixtures of FF and FNF for different molecular composition (that is, the values of the total peptide concentration and relative tripeptide concentration). These measurements used 10 independent particle trajectories for each molecular composition. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

The assembly pathways are observed to depend upon the relative concentration of tripeptides in the system. For FF alone, it has been shown experimentally and with exhaustive Monte Carlo simulations on discrete lattices that the peptide forms nanostructures in a concentration dependent fashion, from vesicles to nanotubes to lamella[73].

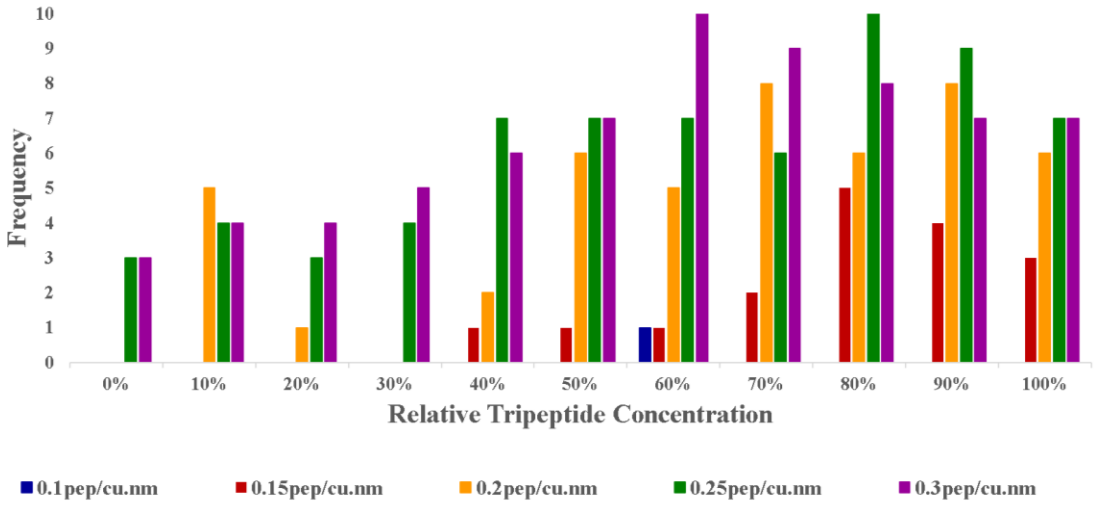

Figure 5: Histogram of the occurrence of lamellar bilayers in mixtures of FF and FNF for different molecular composition (that is, the values of the total peptide concentration and relative tripeptide concentration). These measurements used 10 independent particle trajectories for each molecular composition. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

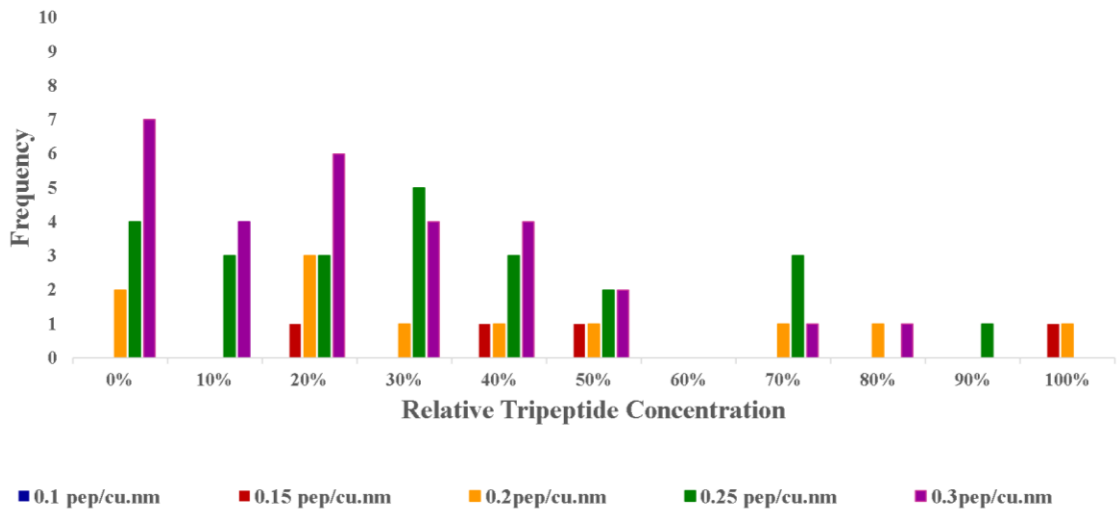

Figure 6: Histogram of the occurrence of nanotubes in mixtures of FF and FNF for different molecular composition (that is, the values of the total peptide concentration and relative tripeptide concentration). These measurements used 10 independent particle trajectories for each molecular composition. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

We observe the same transitions in other trajectories at similar dipeptide concentrations, supporting the hypothesis that these calculations are approaching equilibrium. Similarly, Guo et al[50] reported assembly pathways that included the fusion of smaller sized vesicles for pure FF systems with low concentration of peptides and the folding of bilayers into nanotubes for high concentration of peptides. This is thought to occur through a mechanism where, at low values of the total concentration of peptides (0.1-0.15 peptides/nm^3) (Figure 7), small vesicles are observed to diffuse in solution until they encounter other small vesicles and coalesce to form larger sized vesicles. At higher values of the relative tripeptide concentrations however, the peptides tend to assemble into a

bilayer that then either stabilizes itself via periodic boundary interactions, or bends and closes its free edges to form a vesicle or a nanotube.

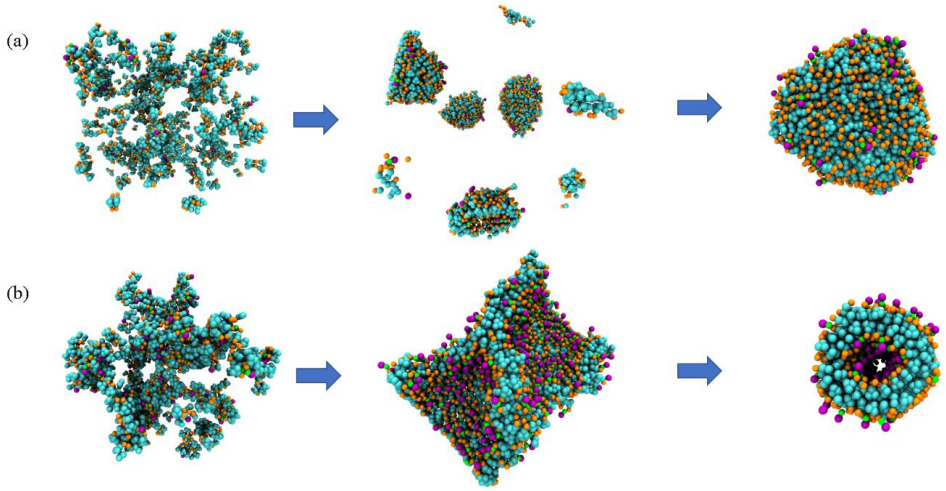

Figure 7: Assembly Pathways: (a) At low concentrations of both the total peptide as well as the relative tripeptide, smaller vesicles first form and then fuse into a larger vesicle. (b) At higher concentrations, Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130Hbilayer like structures first form and then fuse their free edges into nanotubes or vesicles (nanotube shown here).

Morphology also depends on the relative concentrations of di- and tri-peptides. For all peptide ratios, we found a preference for vesicles at the lowest total peptide concentration, and lamellar bilayers at the highest total concentrations. In addition, we have identified an additional degree of freedom, i.e. the relative tripeptide concentration. From the statistics gathered from our simulations, we have constructed a molecular-composition phase-space diagram (presented in Figure 7) that shows distinct molecular composition

parameters promoting the formation of vesicles, lamellar bilayers and nanotubes. At the bottom of the phase space (representing low total peptide concentrations), we find that the system is in the vesicle regime. We attribute this result to the fact that low total peptide concentrations do not form structures large enough to interact through periodic boundaries. Therefore, the most favorable configuration is a fully closed vesicle, as this morphology would minimize any interactions of the hydrophobic aromatic side chains with the water particles.

At higher total concentrations of peptides, as the size of the aggregates increase, the nanostructures in the systems have a higher tendency to interact through periodic walls and thereby, stabilizing themselves. The morphology of such nanostructures can be categorized by the number of period walls they interact through. Vesicles do not extend infinitely interact with themselves through periodic walls, nanotubes interact with themselves do so through two opposite periodic walls, and lamellar bilayers, interact through four periodic walls. To the upper left portion of the molecular composition phase space (shown in Figure 8), there is an overlap between the nanotube and bilayer regimes. This regime overlap is attributed to two competing effects: the tendency of the peptides to form bilayers at high concentrations, and the curvature-inducing effect of the tripeptides in low-to-medium relative tripeptide concentrations.

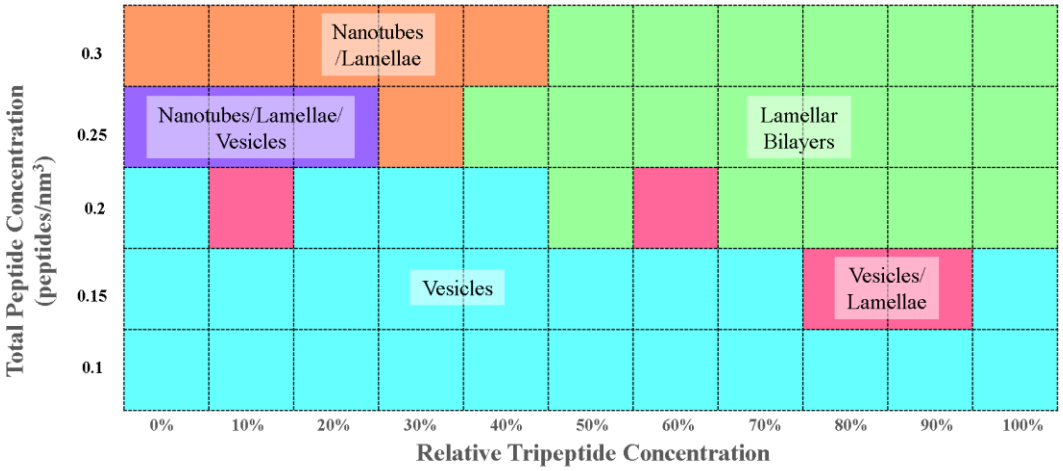

Figure 8: Phase space map showing regimes of Vesicles, Nanotubes and Lamellae. Of the ten trajectories, the most representative is shown here. In the cases where two or more structures are listed, it is because there were close to equal representations. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

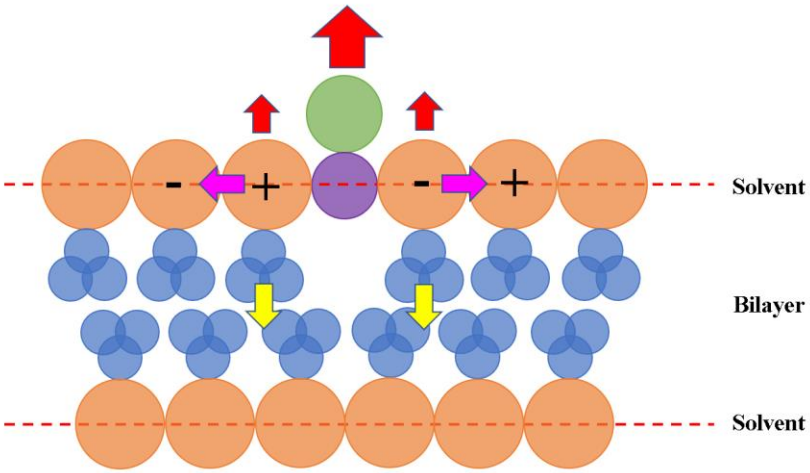

Figure 9: Tripeptide in a bilayer: yellow arrows represent hydrophobic attraction between aromatic side chains. Red arrows represent hydrophilic interactions between amino acid backbones and the asparagine side chains and water particles. Purple arrows indicate electrostatic attraction between

the C and N termini. The resultant force differential tends to induce curvature in a FF-FNF bilayer. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

Figure 9 shows the competing forces acting on a FNF tripeptide molecule that is embedded in a bilayer. The competing forces include the electrostatic attraction between adjacent C and N termini of the peptides, favorable hydrophilic interactions between the backbones and the asparagine side chain and the solvent particles, and favorable hydrophobic interactions between the aromatic phenylalanine side chains. These various forces create a force differential, which coupled to the hinge-like structure of the tripeptide, due to the presence of asparagine in the middle, makes this force differential manifest as a "pinching" effect that induces curvature. To quantify this effect, the average angle made by the FNF backbone was measured (see Figure 10).

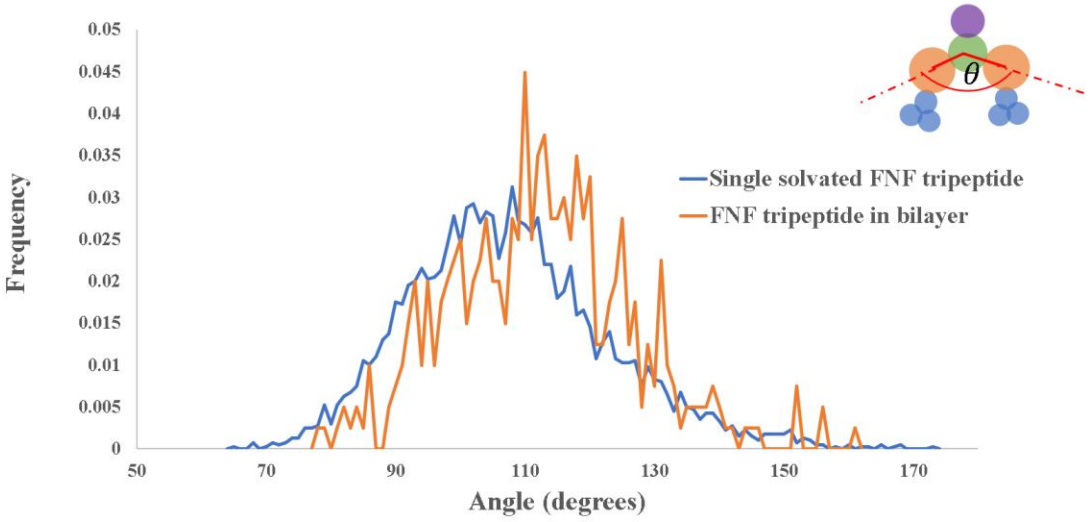

Figure 10: Distribution of the angle (theta) of the tripeptide (FNF). The average angle is 112.9˚, compared to the equilibrium angle of a single solvated FNF peptide pf 107.4˚. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

The angle between the backbone beads (~113˚) is significantly distorted as compared to the equilibrium angle (107˚). This indicates that the tripeptide stretches laterally while the asparagine side chain experiences an attractive force, pulling it into the solvent and keeping the hinge correctly aligned, thereby inducing a local curvature in the bilayer.

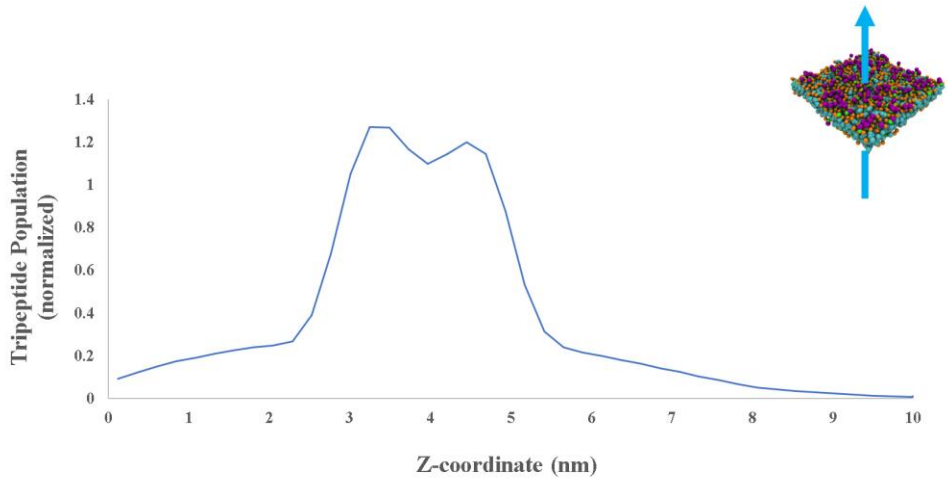

Figure 11: Population density of FNF tripeptides plotted against the Z axis coordinate for Total Peptide Concentration = 0.3 peptides/nm^3 and relative tripeptide concentration of 50%. Two closely spaced peaks show that the tripeptide distributes itself relatively evenly on both sides of the bilayer, leading to the lamella lacking a preference to fold in any specific direction Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

To the upper right of the phase map (see Figure 8), the lamellar bilayers are observed to dominate the molecular composition phase space. We attribute this result to the fact that in addition to the high overall peptide concentration, the high tripeptide concentration makes it far more likely that the tripeptides distributed uniformly on either side of the bilayer. This uniform distribution negates any preference to induce local curvature in a specific direction. A representative example of the bilayer tripeptide spatial

distribution is measured and shown in Figure 11. At lower tripeptide concentrations, a difference can be observed in the distribution of peptides on either side of the bilayer for different morphologies. This is demonstrated in Table 2.

	Nanotube	**Vesicle**	**Lamella**
Inside	13	19	24 (top)
Outside	37	31	26 (bottom)

Table 2: Distribution of tripeptides on either side of the bilayer for different morphologies. Total peptide concentration is 0.25 peptides/nm^3 at 10% tripeptide concentration. A representative system for each morphology was chosen and the number of peptides on either side of the bilayer were counted.

At high relative tripeptide concentrations (90%-100%), a small number of irregular or disordered nanostructures are observed (see Figure 12). This suggests that an overabundance of tripeptide “hinges” in the bilayer decreases its ability to attain a stable morphology. Additionally, overlaps are observed in the domains of the three main morphologies (that is, vesicles, nanotubes and lamellae). This indicates a transition between morphologies for specific values of the total concentration of peptides and relative concentrations of the tripeptides. It is important to note that nanotubes do not have an exclusive molecular composition parameter space like lamellae or vesicles.

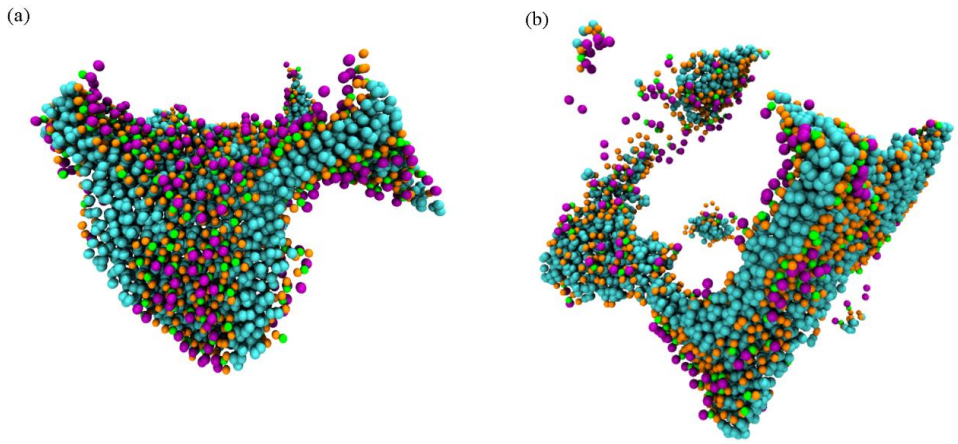

Figure 12: Disordered assembly occurring at 0.3 peptides/nm^3 and (a) 90% tripeptide, (b) 50% tripeptide. Copyright, Org. Biomol. Chem., 2018, 16, 2499. DOI: 10.1039/C8OB00130H

CHAPTER III: Approaching Aggregation Polymorphism with REMD

The phase space reported in Chapter 1 (Table 1 and Figure 8) has mostly well defined phase boundaries between the vesicular and lamellar regions. However, the vesicular/nanotube transition is less defined. This is particularly the case in the region of the phase diagram that corresponds to high total peptide concentrations and low relative tripeptide concentrations. In these regions, we find sizable frequency distributions of multiple morphologies, mainly vesicles and nanotubes. The predominance of curved structures can be explained by the mechanics of local tripeptide-rich zones. However, this strongly suggests the presence of a local minimum on the energy landscape, and does not sufficiently address which morphology corresponds to the local or the global thermodynamic minimum. This local minimum becomes a kinetic trap for the system that it cannot escape on its own.[74,75] To aid the system from escaping from said trap, a range of

enhanced sampling techniques can be applied, such as metadynamics, replica exchange, Gibbs Sampling, Expanded Ensemble methods, etc.

To this end, we apply the Replica-Exchange (RE) method: a family of simulation techniques used to enhance sampling and more thoroughly explore phase space of simulations. RE simulations involve the concurrent execution of independent simulations which interact and exchange information.

RE methods were devised as early as 1986 by Swendsen et. al.[76] in the domain of Monte Carlo simulations. Replica Exchange Molecular Dynamics (REMD) was first formulated in 1999 by Sugita and Okamoto[77]. Initially REMD was used to perform exchanges of temperatures, but was later extended to perform Hamiltonian Exchange[78], pH Exchange[79] and other exchange types. REMD can also be performed in two synchronization modalities: Synchronous and Asynchronous (discussed later). Over the years, REMD has been adopted by many scientific disciplines including chemistry, physics, biology and materials science, such as protein folding, self assembly and force field parameterization.

Theory

Replica Exchange is performed by concurrently executing MD timesteps on multiple “replicas”, i.e. copies of a molecular system, and periodically exchanging conformational information between them. This may be done across a wide range of

parameters as discussed before, such as temperature, pH, a Hamiltonian or restrained states (such as in Umbrella sampling). The exchange occurs by data movement across replica sandboxes (i.e. locations on disk where replica information resides) in case an exchange is accepted.

The criterion used to determine acceptance or rejection of an exchange of configurations between replicas is the Metropolis Criterion, the mathematical formalism for which is below:

$$A(x', x) = \min\left(1, \frac{P(x')}{P(x)} \frac{g(x \mid x')}{g(x' \mid x)}\right).$$

In the equation above, the Acceptance Ratio A between two states x' and x. The probability terms are the transition probabilities between the two states: $g(x|x')$ and $g(x'|x)$ are the distributions of probabilities of transitioning between the states x and x' and $P(x)$ and $P(x')$ are transition probabilities *to* those states drawn from the distributions. This is fundamentally a Monte Carlo move: the acceptance probability becomes

$$A = min\ [1, P(x')/P(x)]$$

Since $g(x|x') = g(x'|x)$, i.e. the probability of transitioning from state x to x' is the same as the probability of transitioning in the reverse direction.

In our case, we apply Temperature REMD, as formulated by Shirts and Chodera. Typically, REMD is performed over a temperature ladder where the replica of interest is the lowest. Replicas concurrently execute a fixed number MD steps equivalent to a defined simulation time, say, 500ps, and wait for all other replicas to reach the same state. The replicas then enter an exchange phase, where a random replica pair is chosen and an exchange of configurations is attempted between them via the application of the Metropolis criterion. Once the exchange is accepted or rejected, the replicas enter the next MD phase. This sequence of one MD phase followed by one Exchange Attempt is termed a ”cycle”. While this technique is useful, only one exchange is attempted every cycle, and this severely limits replica mobility across the temperature ladder. To address this, Chodera et al.[80] formulated Replica Exchange as a special case of Gibbs Sampling (also referred to as Independence Sampling[80,81]). This method applies the Infinite Swapping approach[80] to Replica Exchange: as replicas enter the exchange phase, a “swap matrix” is instantiated and populated with all *i-j* exchange probabilities (Figure 13). This matrix is effectively one Metropolis criterion, and allows for multiple swaps to be attempted in a single cycle, dramatically increasing replica mobility[80].

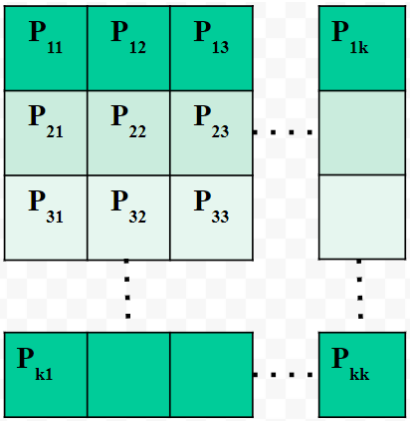

Figure 13: Swap Matrix of a system of K replicas. The matrix has K^2 elements. Element (*i*,*j*) corresponds to a unique exchange probability between replica *i* and replica *j*.

The exchanges may be performed synchronously or asynchronously, by defining a "Synchronicity Factor" as the ratio of replicas that must wait to perform an exchange to the total number of replicas. If this factor is 1, then all replicas must wait for all other replicas, making synchronous exchange a special case of asynchronous exchange.

All simulations were performed using RepEx, a scalable, flexible and user-extensible Replica Exchange Framework developed collaboratively between the Computational Hybrid Soft Materials Laboratory (http://sol.rutgers.edu/~md860/dutt.html) and the RADICAL group (https://radical.rutgers.edu/) at Rutgers. RepEx is a lightweight layer built on top of Ensemble Toolkit (RADICAL-EnTK), a workflows-management/runtime system. A detailed discussion of these packages is omitted here since the core focus of this document

is molecular science, rather than the software: that discussion will be handled in a separate manuscript (In Preparation).

Replica Exchange Setup

Typically, temperature ranges in REMD simulations (hence forth referred to as a "temperature ladder") are set up with the following considerations in mind: (i) high sampling rate in the temperature range of interest, (ii) minimizing sampling in temperature ranges that do not provide additional benefits. Secondly, combining REMD with coarse graining comes with its own challenges: coarse graining provides a very smooth energy landscape for individual replicas to reside in, making hops between adjacent temperatures less likely to occur due to the relative local stability. Nagai et al. and Stelter et. al. study phase transitions in lipid bilayers using REMD in conjunction with the coarse grained MARTINI model. While Nagai et al. employ the standard "wet" (i.e. explicitly represented solvent particles) model[82], Stelter et al. employ an implicit solvent "dry" MARTINI model[83]. Both studies successfully observe phase transition and reproduce the phase transition temperatures. Chebaro et al.[84] employ the coarse grained OPEP (Optimized Potential for Efficient protein structure Prediction) model with REMD to study protein systems in an implicit solvent.

Temperature selection

Temperature selection for replica exchange is a critical parametric consideration. Replicas with poorly selected temperatures may result in no exchanges occurring, or

inefficient resource utilization due to unnecessarily high sampling rates that are not necessarily beneficial. As pointed out above, combining REMD with CGMD is not always trivial: the CG force field is typically one which provides a very smooth energy landscape, making exchanges less likely due to the thermodynamic detail balance for exchange becoming very hard to satisfy. We address this by spacing replicas closer together (as demonstrated by Nagai and Okamoto[82]). To compute an appropriate ***ΔT*** value, simulations were set up as follows: 100-replica ensembles were set up to perform exactly one exchange, with the ***ΔT*** being varied from 10K down to less than 1K. All replicas performed 20000 MD timesteps before they attempted an exchange. Table 3 below lists the various ***ΔT*** values between adjacent replicas tested and the computed Exchange Acceptance Ratios (EAR).

ΔT(K)	***Computed EAR***
10	0.01
5	0.00
1.25	0.11
1	0.25
0.9	0.31
0.8	0.37
0.75	0.41

Table 3: Table showing the temperature differential between adjacent replicas and the computed exchange acceptance ratio (EAR).

We find that a 0.8-0.9K difference between adjacent replicas results in a good acceptance rate. This is, as described in Okamoto et al., dependent on potential energy

overlaps between replicas. Figure 14 shows the potential energy overlap for a 16 replica system. Each bell curve represents the energy states accessed by one replica. Therefore, the overlaps between curves represent the states accessible by more than one replica. The larger the overlap between two replicas, the greater the probability for the acceptance of an attempted exchange.

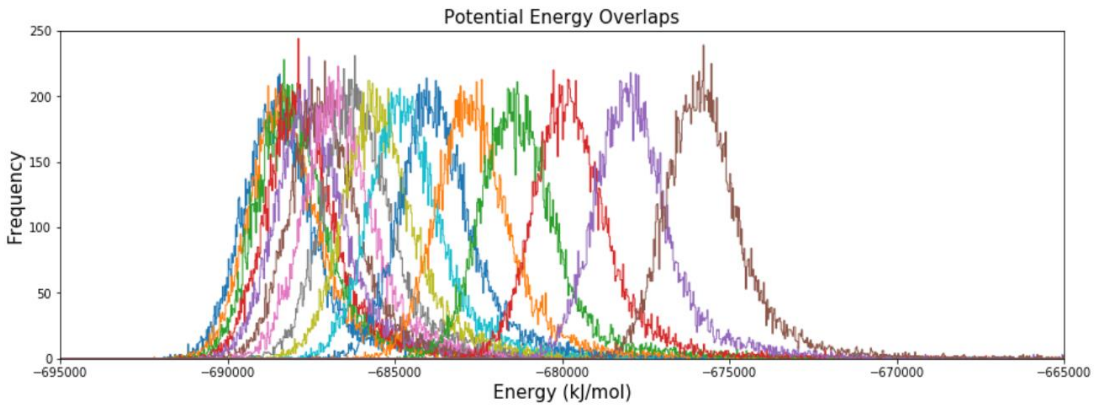

Figure 14: Potential energy overlaps between replicas. Each bell curve represents the energy states accessed by one replica. The larger the overlap between two replicas, the greater the probability for the acceptance of an attempted exchange.

However, replicas that are too high in temperature need not be sampled as frequently: They just serve as "helper" replicas that allow lower temperature replicas to overcome any kinetic barriers. T_{Max} is the temperature of the highest temperature replica in the ensemble. This must be chosen carefully, since if the maximum temperature is not high enough, the replicas of interest may not have access to higher energy states, but if it is too high, precious compute resource time is wasted on sampling energy states that do not add any additional information to the systems being studied. Further, having a high density of replicas at higher temperatures results in a compounding of this problem.

Therefore, in addition to carefully selecting the maximum thermostat, the distribution must not be linear: there need not be as many replicas at higher temperatures as there are near the temperature of interest. Another important consideration is the relative mobility of replica configurations up and down the temperature ladder. The downward mobility of configurations from higher temperatures is more desirable than the reverse: the target temperature (i.e. 310K in this case) is interested in sampling from temperatures higher than itself, whereas the higher temperatures, as discussed earlier, only exist as auxiliary "helper" replicas that aid the target replica in sampling a larger conformational space. For this reason, an exponential temperature distribution is chosen. This is in general agreement with conventional replica exchange simulations. The final temperature ladder selected is shown in Table 4 below.

T(K)	*ΔT(K)*
310.0	0
310.3099644507333	0.3099644507332755
310.7160068621849	0.40604241145160813
311.24790798667647	0.5319011244915828
311.9446795510655	0.6967715643890529
312.85742553069696	0.9127459796314383
314.05309031656384	1.1956647858668816
315.61936868104306	1.566278364479217
317.67113765846346	2.0517689774204086
320.3588820800015	2.6877444215380137
323.87973172487284	3.5208496448713618
328.4919195960311	4.612187871158255
334.53372174735154	6.041802151320439
342.4482677839449	7.914546036593379
352.81604173557395	10.367773951629033
366.3974570454462	13.581415309872227

Table 4: Final temperature ladder. An exponential distribution was chosen for the REMD simulations.

Results

We performed Temperature-Independence Sampling REMD (T-REMD-IS) on the FF-FNF system with a total peptide concentration of 0.25 peptides/cu.nm and a 20% relative tripeptide concentration. At these parameters, there are comparable counts for the

occurrences of nanotubes, lamellae and vesicles. This is due to the presence of kinetic traps, i.e. local minima, that the simulation is unable to escape. When these local minima are deep enough, the system collapses into these “final” states. To escape these local minima we must ensure that the replica has sampled from sufficiently high energy states. Therefore, we trace the path of a configuration that originates in the highest temperature replica to ensure that it has diffused to the target replica R0 at 310K. The trace of replica R15 is shown in Figure 15: we see that replica R15 has diffused its configurations all the way down to replica R0: ergo, R0 has “seen” the full temperature space up until ~367K.

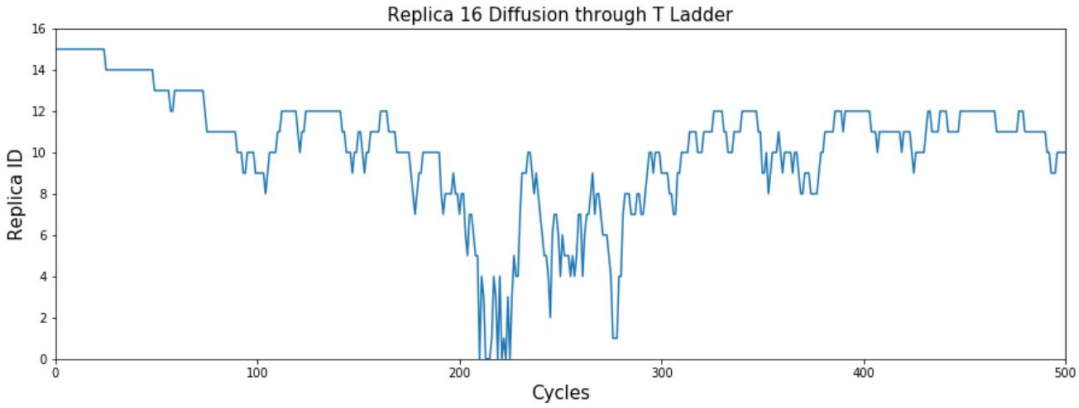

Figure 15: Trace of the highest temperature replica R15 diffusing through the Temperature ladder. R15 diffuses all the way to R0, the target replica at 310K.

We also compute the aggregate acceptance ratio. The exchanges maintain an average aggregate acceptance ratio of ~33.364%. This is shown in Figure 16.

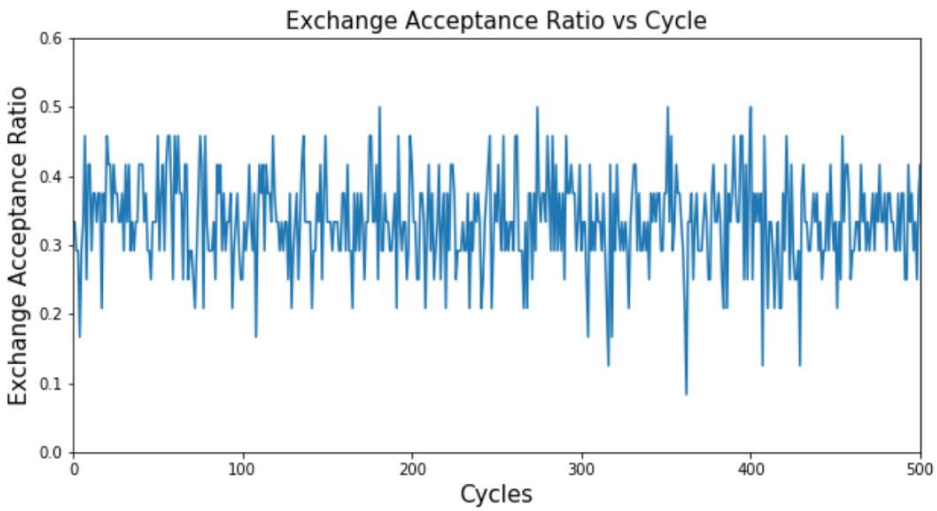

Figure 16: Aggregate exchange acceptance ratio vs Cycles. The EAR fluctuates about an average value of ~33.364%.

We validate our REMD simulation by generating a free energy plot (FEP). The plot is then compared with the FEP of a coarse grained brute force MD simulation (discussed in Chapter 1) of the same system that produced the same result. The FEP was generated by binned state histogramming over two reaction coordinates (Radius of Gyration and Solvent Accessible Surface Area) and computing Free Energy (FE) as:

$$FE = -RT \log P(x,y)$$

Where $P(x,y)$ is the joint probability density of the system existing in state (x,y), R is the universal gas constant in kJ/mol, and T is the temperature of R0 in K (i.e. 310K). Figure 17 shows the FEP's for the REMD run and the classical MD run respectively.

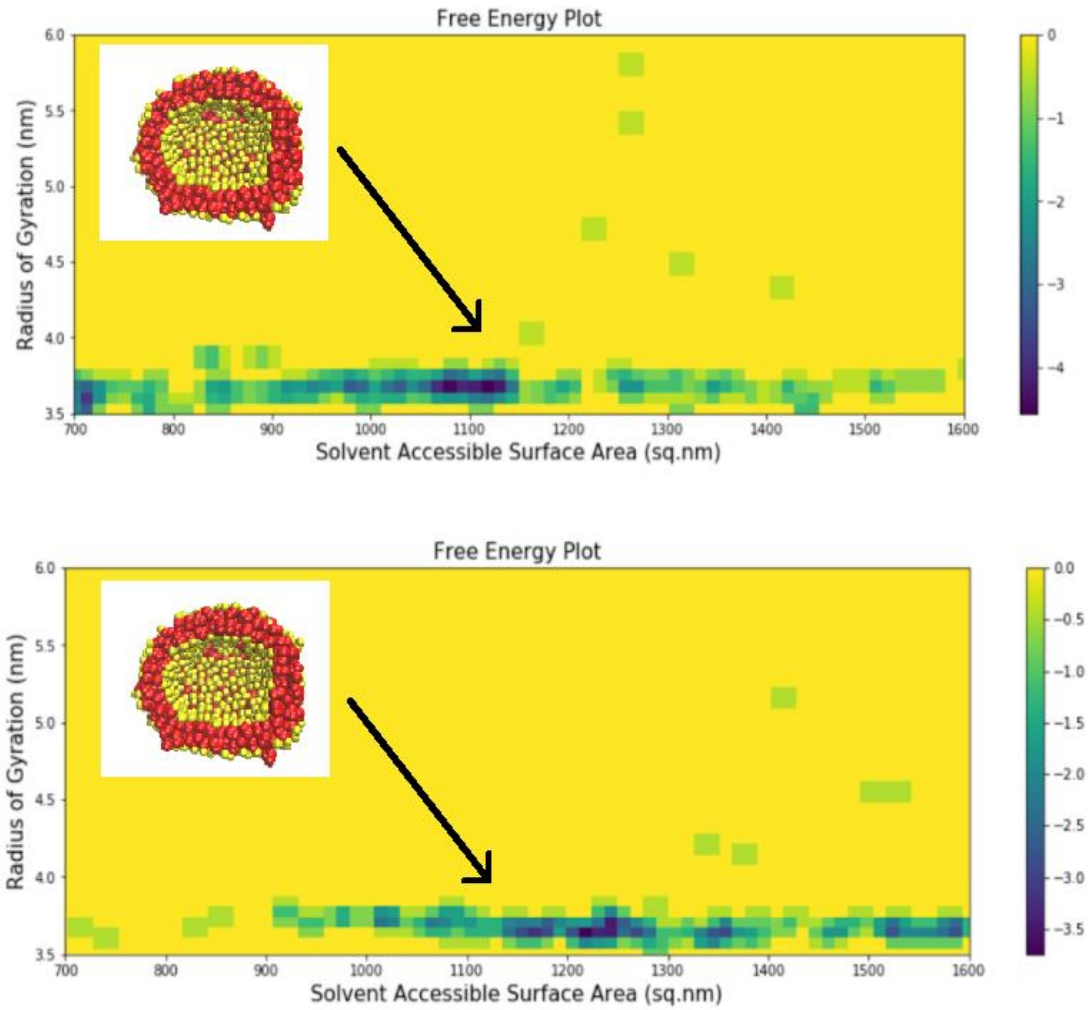

Figure 17: FEP's for the REMD run (top) and classical MD run (bottom). Our results indicate that the primary basin (i.e. lowest free energy region) lies around a SASA of ~1100-1300 sq. nm and an Rg of ~3.6nm. These results are in agreement with Guo et al.[50] The basin at (x,y)=(700 sq.nm and 3.6nm) in the top figure is an artifact of a transition state where peptides collapse into a non-hollow spherical structure with a lower SASA before stabilizing into a vesicle. The basin at (x,y)=(1550 sq.nm and 3.6nm) is a result of the classical MD simulation reporting states with high degrees of dispersion (and therefore, high Rg for the peptide component of the simulation box). The inset shows the vesicular configuration at the FE basins.

Replica exchange predicts that the preferred morphology at this concentration is vesicles, as shown in Figure 18. This system maintained an aggregate exchange acceptance rate of ~33%.

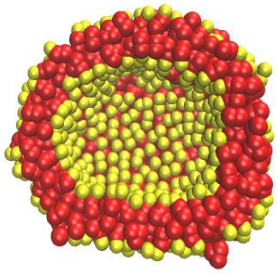

Figure 18: REMD simulation (final frame) of an FF-FNF system with a total peptide concentration of 0.25 peptides/cu.nm and a 20% relative tripeptide concentration.

Finally, we apply the same protocol for other regions of the phase space discussed in chapter 1 and shown in Figure 8. The partially REMD corrected phase diagram is shown in Figure 19.

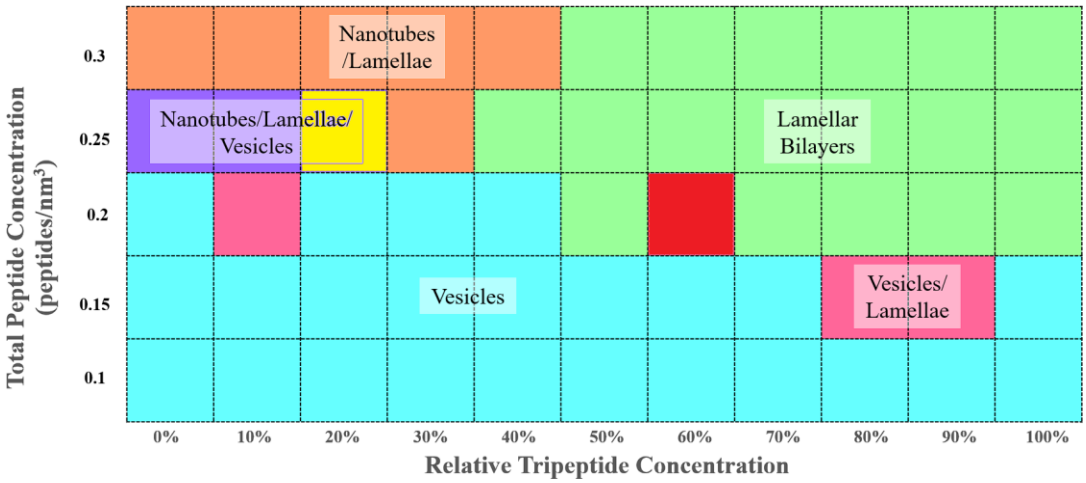

Figure 19: Partially corrected phase map of the mixed peptide phase space. The yellow block at 30% Relative Tripeptide Concentration and 0.25 peptides/cu.nm represents the part of the phase space that classical MD was unable to predict a clear most-likely-conformation between vesicles,

nanotubes and lamellae, but REMD predicts a Vesicle. The red block at 60% Relative Tripeptide Concentration and 0.20 peptides/cu.nm represents the part of the phase space that classical MD was unable to predict a clear most-likely-conformation between vesicles and lamellae, but REMD predicts a Vesicle.

CHAPTER IV: Targeted design of supramolecular peptide assemblies

The PACE Framework

As we see from Chapter 1, the morphological diversity of a single system with two components is vast: but there are 20 possible amino acid residues. Therefore there are millions of possible combinations of peptide sequences, providing the possibility of a vast range of supramolecular assemblies and nanomaterials. This chapter proposes a high throughput computational framework for a screening protocol for targeting specific morphologies. The computational framework will encompass a Simulation-Analysis pipeline.

The notion of a Simulation-Analysis pipeline is straightforward: A pipeline carries temporally separated simulation and analysis tasks, and a control flow decision is made based on the signal received from the analysis task. In other words, the first MD simulation runs to completion for a fixed number of timesteps and triggers an analysis task. The analysis task evaluates the resultant MD trajectory for a specific criterion. Compliant MD simulations are allowed to continue to the next MD phase with a high sample rate to allow further, detailed analysis. Non-compliant ones are terminated, and their computational resources are either allocated elsewhere or released: a consideration that is especially important due to the limited availability of high-performance computing resources. The pipeline can be replicated for concurrent processing of multiple candidates simultaneously.

This setup may be written in the form of shell scripts that explicitly assigns resources to all the simulations in question and resolves the necessary dependencies. While possible, this approach may prove to become extremely cumbersome, especially over large

candidate pools. A large candidate pool requires the explicit specification of command lines referring to each system. Further, the expression of the adaptivity required in a workflow such as this is extremely limited when done via shell scripting, underscoring the requirement of a framework that uses an API that allows for scalable and flexible expression of adaptive workflows. This framework, therefore, is written using the RADICAL-EnTK API[85]. As before, a detailed discussion of EnTK will be handled in a separate manuscript (in preparation).

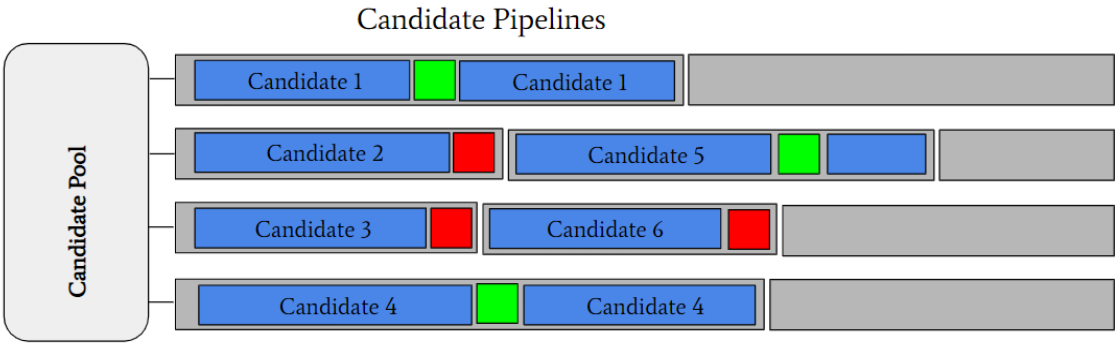

Figure 20: Screening Framework for peptide materials. Candidate systems set up *a priori* are drawn sequentially from a candidate pool and simulated in concurrent pipelines. Each pipeline performs an MD phase (blue) followed by screening. If the screening task returns a "compliant" signal (green), the pipeline propagates into an additional MD phase. A "non compliant" signal terminates the pipeline and spawns a new pipeline with the next candidate system.

We therefore propose PACE: Pipeline for Automating Compliance-based Elimination. PACE is a workflows approach to simulating and studying biomolecular systems. A predefined number of "candidate" systems, such as various peptide sequences, peptide combinations, or even systems running at different physical conditions

(temperatures, concentrations, etc.) are prepared and set up in a Candidate Pool, a sandbox that contains all the necessary input files. As many pipelines as desired are instantiated (based on available resources: a single pipeline per node of an HPC platform is typical), say, k in number, and simultaneously spawn MD simulations of the first k candidate systems. These systems are then screened for the desired characteristics, and either allowed to propagate to the high resolution MD phase or terminated based on the screening analysis signal. This entire set of pipelines may then be replicated for statistical significance.

Figure 20 shows a schematic representation of the proposed PACE protocol. The blue blocks represent MD runs of candidates, whereas the green and red blocks show the analysis stage: green blocks represent an analysis stage that has returned a positive (Boolean "1") signal, and red blocks, a negative (Boolean "0"). The Analysis stage can in principle be anything: from an explicit determination of simulation parameters to using statistical analytical models to make decisions. For this project, we use a Machine Learning (ML) driven approach, discussed in the next section.

Background: Convolutional Neural Networks

For discriminating between the three possible nanostructures we employ a convolutional neural network (CNN)[86, 87] driven image recognition kernel. In this section, a cursory discussion of CNN's is provided for the sake of context and background. A detailed discussion of Neural Nets and Deep Learning is beyond the scope of this thesis.

A CNN is a “network” that vaguely mimics the connections in a human brain. They are designed for pattern recognition, much like their biological counterpart. It is composed of layers (or “nodes”) of units called “neurons”, which are basically switches that (i) assign weights to each input, (ii) combine the inputs into one signal, and (iii) apply an “activation function” to compute the signal to be propagated forward to the next layer and error to be propagated back to the previous layer.[86,87] Figure 21 shows a schematic of a hypothetical 4 layer Neural Network. There are 6 neurons in the input layer (i.e. it accepts six inputs), the two hidden layers have four neurons each, and the output layer has two neurons, i.e. it gives two outputs. In the context of image recognition, each input neuron corresponds to one pixel in the image being processed. The hidden convolutional layers do not follow a prescription for neuron count, as this depends on the requirements of the case in question. Image recognition CNN’s use a technique called “feature extraction”, and typically, the number of neurons in the hidden layers depends on the number of “features” to be extracted. Feature extraction is discussed in the section ahead.

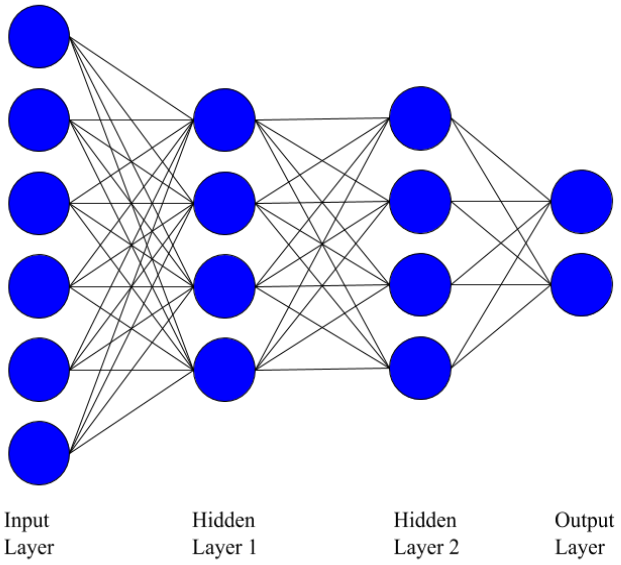

Figure 21: Schematic representation of a four layer CNN. The input layer has six input neurons, the hidden (convolution) layers have four neurons each, and the output layer has two output neurons.

The 4 layers interact by passing signals between each other. Each layer accepts inputs from the previous layer, computes the signal, and sends this signal to the next layer. This is known as "forward propagation". Simultaneously, each layer also computes the error from the expected result and communicates it with the previous layer. This is known as "back propagation" and is used to refine the model weights in a manner that minimizes said error.[87]

When the complete training data set completes one full pass of forward as well as back propagation through the entire depth of the neural network, it is termed as an "epoch." Figure 22 shows a schematic of the forward and back propagation steps between the various layers.

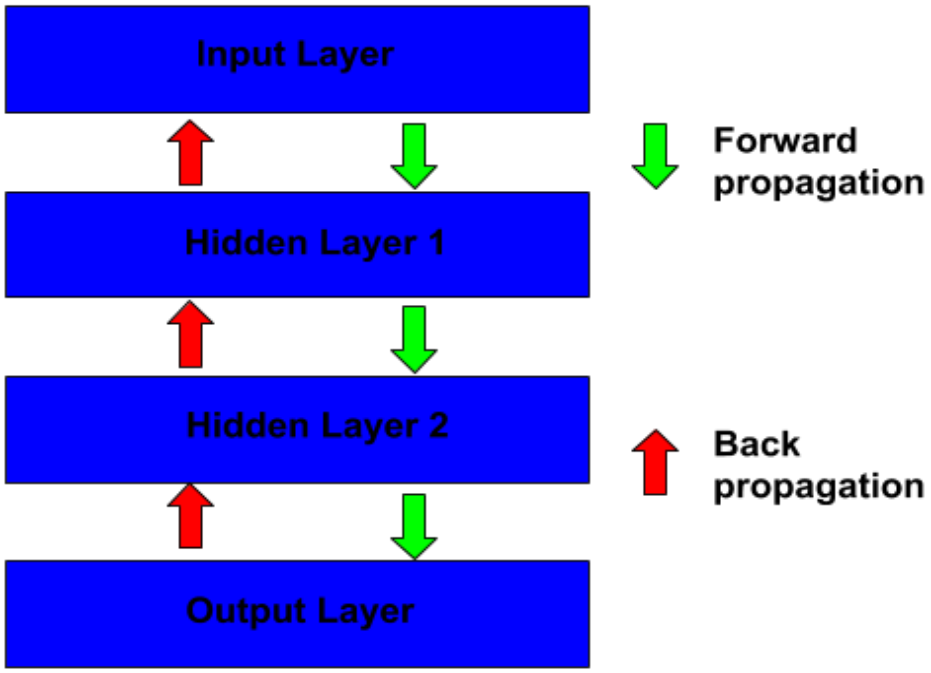

Figure 22: Back and Forward propagation through the CNN. A full pass of all training data through all four layers is termed as an "epoch".

Feature Extraction

"Feature extraction" refers to the process of recognizing the patterns in a set of images and reducing them to a set of "features" to allow for faster processing[88]. An image, fundamentally, is a matrix consisting of RGB (Red, Blue, Green) values. This makes images extremely amenable to be processed via matrix operations. Convolution and Pooling are the two primary operations we will be considering here. Convolution refers to the process of "filtering" a given image for features. These features are then "accumulated" through a process of Pooling.

Convolution is a method of "feature filtering" involving replacing a target group of pixels from the source image with a single matrix element in a "target matrix". This is done by reading the RGB values of the pixel at the center of the target group, computing the

weighted sum of the target pixel and all the neighboring pixels, and returning this value to the target matrix. This is then repeated over all the pixels in the image. Therefore, the convolution layer acts as a “sliding window” on the image.

The MaxPool() operation is a method of “accumulating” these filtered features by essentially picking a sub-matrix from the matrix generated by the Convolution step and choosing the maximum value in that sub matrix (i.e. picking the “most prominent feature”).

These processes put together comprise the “feature extraction” process.

Model and Training

The Image recognition kernel was set up with a converter script to generate .png image files 480 by 640 pixels in size from GROMACS output .gro files. This was done by simply plotting the *x*, *y* and *z* coordinates of the output coordinate file by fixing the perspective using Python’s matplotlib library. This image was then passed on to the CNN. Figure 23 shows the images that were fed to the CNN.

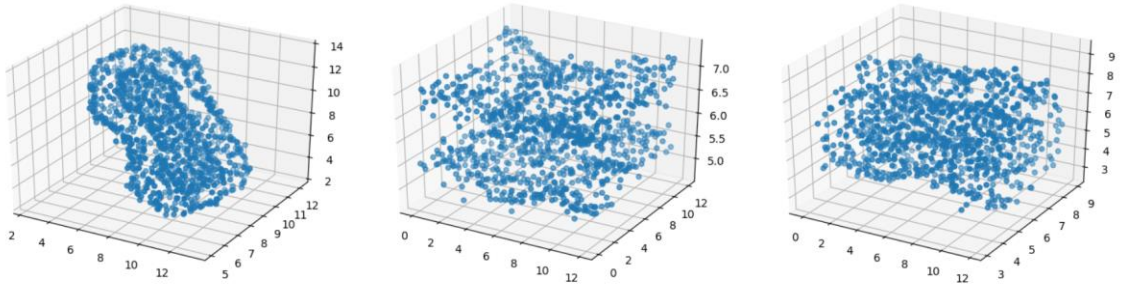

Figure 23: Images of Vesicle, Lamellae and Nanotubes (Left to Right) generated by the Pre-Processing module of the Neural Network.

The images generated by the matplotlib driven preprocessing module were then passed to the Neural network. In the case of training, a pre-generated set of images were supplied to the network, but for in-pipeline prediction within PACE, images were generated on the fly, as candidates completed their MD runs. Figure 24 shows the scheme for data flow from the MD step to the ML step until the shape of the structure is characterized.

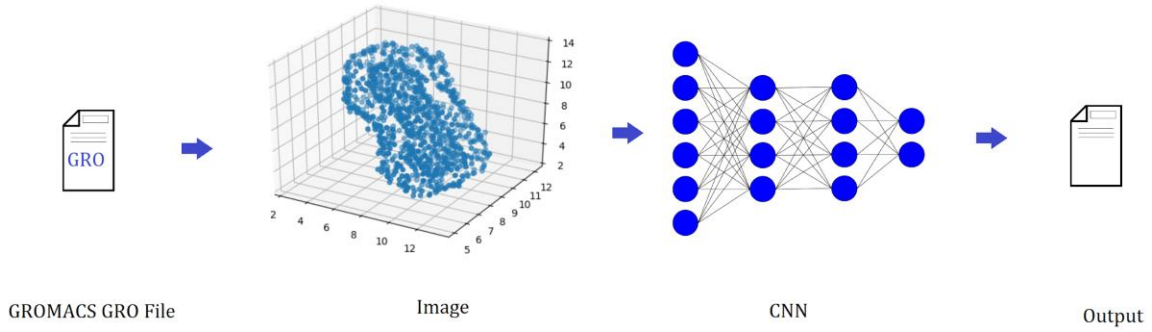

Figure 24: Setup of the CNN kernel. A GROMACS .gro file is first used to generate an image that is then fed to the CNN, which generates an output string and writes it to a file.

Our model was configured as a 4 layer CNN with one input layer, one output layer and two hidden layers. The input layer has 921600 Neurons, three for each pixel of the image (three to correctly parse the RGB values). This is followed by Convolve() and MaxPool() operations. The Convolution layers have a 3 by 3 convolution layer, therefore, the two convolution layers have 9 neurons each. Max pooling is performed by a 2 by 2 kernel, i.e. the sub-matrices picked from the input matrix are 2 by 2. These operations (the "layer") are then repeated for further filtering and extraction of features. The output is a

single string, that reports if the image initially fed to the network was a nanotube, vesicle or lamella.

The Model was trained using 300 images each of vesicles, lamellae and nanotubes, and validated using 100 images of each. Figure 25 shows the validation plot of the trained model. We found that after 7 epochs, the accuracy of prediction in the validation sets goes to ~1 in five independent training runs. Therefore, we concluded that 10 epochs is reasonable for good feature extraction.

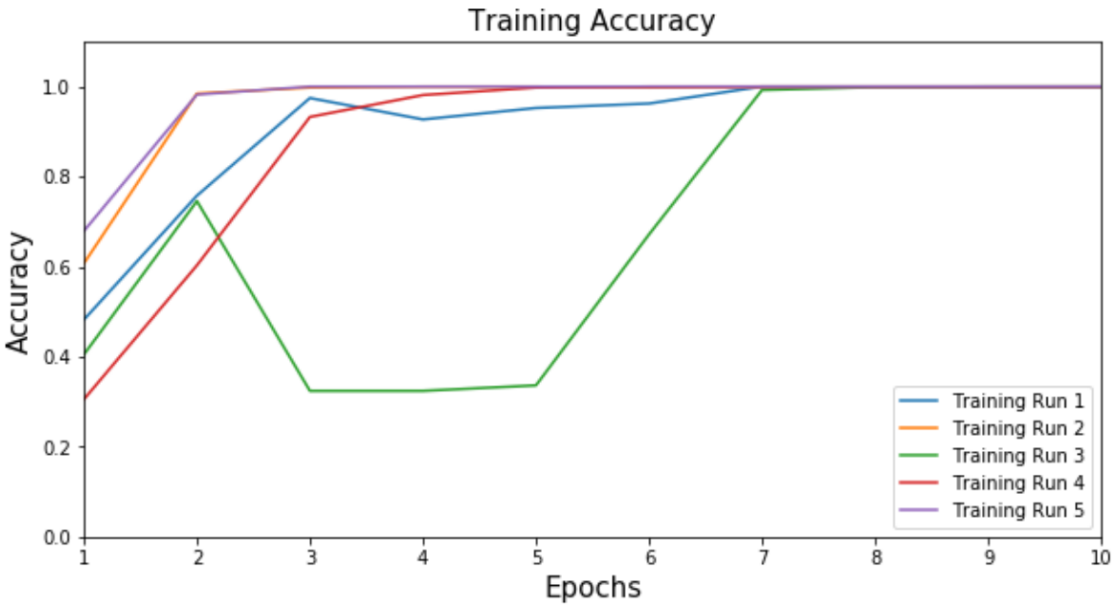

Figure 25: Training/validation of the model. The training set consisted of 300 images each of the three nanostructures. The remaining 100 images were used to validate the model after each epoch. Five independent training runs were performed. All runs converged to a ~1.0 accuracy after 7 epochs.

Results

The trained model was then employed to run as the Analysis Stage of PACE. The model was loaded onto the working directory of an XSEDE HPC Platform (SDSC Comet) and hooks were set up so that the analysis kernels as shown in Figure 20 pointed towards this model. A system was set up as shown in Table 5:

System	Seeds	Total Peptide Concentration	Total Peptide Count	Di-Tripeptide Ratio
FF-FNF	10	0.2 peptides/cu.nm	500	4:1

Table 5: 10 seeds of the FF/FNF system were set up and run in PACE.

PACE was set up to generate statistics in order to compute the frequencies of each nanostructure generated. PACE reported the formation of a lamella in 1 seed out of the 10, nanotubes in 2 and vesicles in 7. These results are shown in Figure 26. In Chapter 1, we discussed the statistics generated by brute force MD followed by visual inspection (results tabulated in Table 1). The results generated by PACE are in excellent agreement with those results.

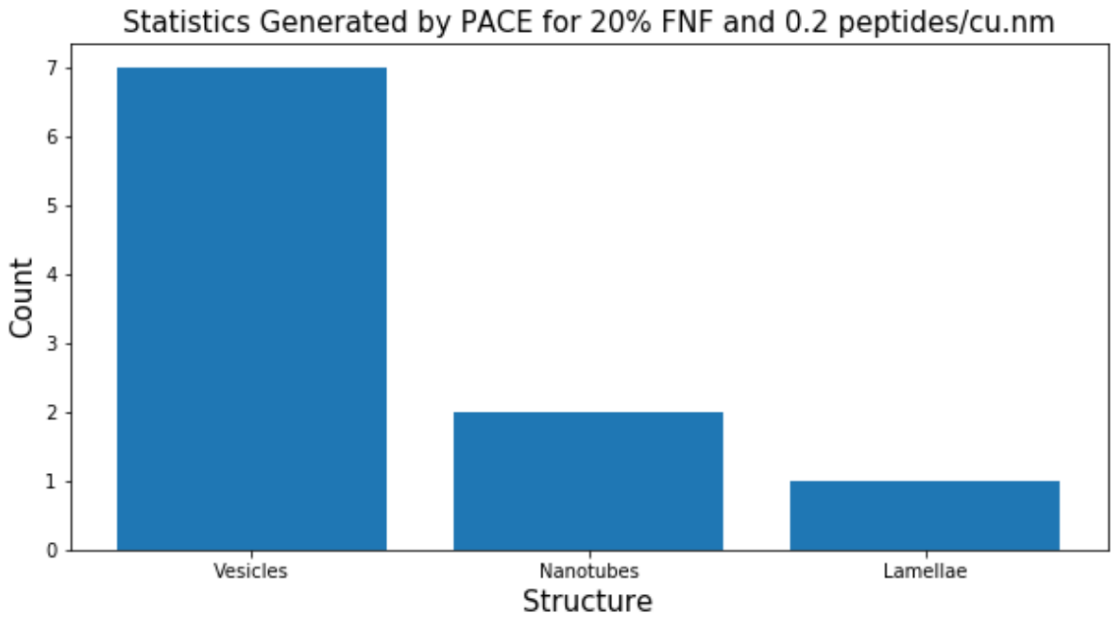

Figure 26: Statistics generated by PACE. PACE reports the formation of vesicles in 7 out of 10 seeds, nanotubes in 2, and a lamella in 1. This is in excellent agreement with results reported in Chapter 1, Table 1.

Chapter V: Conclusions and Future Directions

This dissertation covers a range of approaches to studying peptide polymorphism. We start with the use of Coarse Grained Molecular Dynamics simulations of a phase space consisting of varying mixed peptide systems. In Chapter II, We report rich polymorphism in these systems comprising two ultrashort peptides: Diphenylalanine (FF) and Phenylalanine-Asparagine-Phenylalanine (FNF). These peptide mixtures exhibit the formation of spherical and nonsperical vesicles, nanotubes and flat lamellar bilayers. We find that the structure of the supramolecular assemblies reported closely correlates with the molecular composition of the systems in question. The presence of tripeptides in the system biases the structure towards forming assemblies that have curvature, but there is a critical ratio beyond which the curvature begins to flatten out again. We establish this using measurements of bond angles in the peptides and looking at the intermolecular packing of the structures. Further, we find that the phase space has, for the most part, clear boundaries between the vesicle, lamellar and nanotube phases. However these boundaries in some instances are not entirely clear. We proceed to study this in Chapter III via Independence Sampling Replica Exchange (isREMD), and find that the method allows fast convergence to the thermodynamic minimum by allowing the system to escape kinetic traps. We thoroughly validate the results of these simulations by (i) ensuring that the target replica at 310K has traversed and sampled configurations spanning the entire temperature ladder via tracing the pseudotrajectory of the highest temperature replica, and (ii) ensuring that the free energy profile (FEP) derived from the REMD simulations agrees with both the FEP of a classical CGMD simulation that gives the same end state as well as FEP's of similar

structures from the literature. Finally, in Chapter IV, we propose, design and test a Machine Learning driven workflows approach to studying biomolecular systems to reproduce a part of the phase space discussed in Chapter II, which we have dubbed PACE. We train and apply a Convolutional Neural Network to determine the morphology of simulations, collect statistics and make adaptive decisions to either extend the workflow for further analysis or terminate the simulations that do not show promise. The code for these workflows are on Github, access will be available after publication of relevant research papers. Manuscripts are in preparation for the Replica Exchange and PACE components of this dissertation.

In the immediate future, we plan to explore the full phase space of the FF-FNF system by using both Replica Exchange as well as PACE: Replica Exchange to determine the thermodynamic minima of all regions of the phase space where phase boundaries are not clear, and PACE to (i) reproduce the full phase space by utilizing the Machine Learning approach and (ii) extend PACE to allow it to execute adaptive workflows with fully customizable analysis routines in addition to being able to extend, terminate or adapt molecular simulation-analysis pipelines based on the determinations of the Machine Learning kernel. This work will be carried out jointly with Akash Banerjee (PhD student) and Ethan Zang (Undergraduate Student). We will be handling generalizing the framework to adapt to a wide range of biomolecular simulation applications. Ethan Zang will be contributing to the development of further ML based algorithms and approaches for analysis and characterization of biomolecular simulations.

In conclusion, this dissertation has established the rich diversity of supramolecular assemblies in the peptide space, and has discussed three approaches to studying them. These approaches have tremendous potential to be expanded upon in order to glean new insights regarding biomolecular behavior and can have impacts on a wide range of fields including nanomaterials, biomaterials, nano/biomedical science, and bioelectronics.

BIBLIOGRAPHY

1. Marshall, K. E.; Vadukul, D. M.; Dahal, L.; Theisen, A.; Fowler, M. W.; Al-Hilaly, Y.; Ford, L.; Kemenes, G.; Day, I. J.; Staras, K.; Serpell, L. C., A critical role for the self-assembly of Amyloid-beta 1-42 in neurodegeneration. Sci Rep-Uk 2016, 6.
2. Al-Garawi, Z. S.; Morris, K. L.; Marshall, K. E.; Eichler, J.; Serpell, L. C., The diversity and utility of amyloid fibrils formed by short amyloidogenic peptides. Interface Focus 2017, 7, (6).
3. Deng, L.; Xu, H., Hierarchical processes in beta-sheet peptide self-assembly from the microscopic to the mesoscopic level. Chinese Phys B 2016, 25, (1).
4. Do, T. D.; de Almeida, N. E. C.; LaPointe, N. E.; Chamas, A.; Feinstein, S. C.; Bowers, M. T., Amino Acid Metaclusters: Implications of Growth Trends on Peptide Self-Assembly and Structure. Anal Chem 2016, 88, (1), 868-876.
5. Ke, P. C.; Sani, M. A.; Ding, F.; Kakinen, A.; Javed, I.; Separovic, F.; Davis, T. P.; Mezzenga, R., Implications of peptide assemblies in amyloid diseases. Chem Soc Rev 2017, 46, (21), 6492-6531.
6. Waqas, M.; Jeong, W. J.; Lee, Y. J.; Kim, D. H.; Ryou, C.; Lim, Y. B., pH-Dependent In-Cell Self-Assembly of Peptide Inhibitors Increases the Anti-Prion Activity While Decreasing the Cytotoxicity. Biomacromolecules 2017, 18, (3), 943-950.
7. Chen, J.; Zhang, B.; Xia, F.; Xie, Y. C.; Jiang, S. F.; Su, R.; Lu, Y.; Wu, W., Transmembrane delivery of anticancer drugs through self-assembly of cyclic peptide nanotubes. Nanoscale 2016, 8, (13), 7127-7136.
8. Hutchinson, J. A.; Burholt, S.; Hamley, I. W., Peptide hormones and lipopeptides: from self-assembly to therapeutic applications. J Pept Sci 2017, 23, (2), 82-94.
9. Jeena, M. T.; Palanikumar, L.; Go, E. M.; Kim, I.; Kang, M. G.; Lee, S.; Park, S.; Choi, H.; Kim, C.; Jin, S. M.; Bae, S. C.; Rhee, H. W.; Lee, E.; Kwak, S. K.; Ryu, J. H., Mitochondria localization induced self-assembly of peptide amphiphiles for cellular dysfunction. Nat Commun 2017, 8.
10. San, B. H.; Li, Y.; Tarbet, E. B.; Yu, S. M., Nanoparticle Assembly and Gelatin Binding Mediated by Triple Helical Collagen Mimetic Peptide. Acs Appl Mater Inter 2016, 8, (31), 19907-19915.
11. Zohrabi, T.; Habibi, N., Dendritic Peptide Nanostructures Formed from Self-Assembly of Di-L-phenylalanine Extracted from Alzheimer's beta-Amyloid Poly Peptides: Insights into Their Assembly Process. Int J Pept Res Ther 2015, 21, (4), 423-431.
12. Fan, J.; Liu, L.; Ding, S. M.; Fu, M. L.; Liu, F. F.; Zhang, C. C.; Wu, H.; Wang, J. H.; Qin, Y. Q., Studies on polymer capped quantum dots and peptide self-assembly using fluorescence coupled capillary electrophoresis. Int Nanoelectr Conf 2016.
13. Lim, J. M.; Ryu, M. Y.; Yun, J. W.; Park, T. J.; Park, J. P., Electrochemical peptide sensor for diagnosing adenoma-carcinoma transition in colon cancer. Biosens Bioelectron 2017, 98, 330-337.
14. Mansbach, R. A.; Ferguson, A. L., Coarse-Grained Molecular Simulation of the Hierarchical Self-Assembly of pi-Conjugated Optoelectronic Peptides. J Phys Chem B 2017, 121, (7), 1684-1706.

15. Nguyen, V.; Zhu, R.; Jenkins, K.; Yang, R. S., Self-assembly of diphenylalanine peptide with controlled polarization for power generation. Nat Commun 2016, 7.
16. Adibi-Motlagh, B.; Lotfi, A. S.; Rezaei, A.; Hashemi, E., Cell attachment evaluation of the immobilized bioactive peptide on a nanographene oxide composite. Mat Sci Eng C-Mater 2018, 82, 323-329.
17. Banerjee, J.; Azevedo, H. S., Crafting of functional biomaterials by directed molecular self-assembly of triple helical peptide building blocks. Interface Focus 2017, 7, (6).
18. Bartolami, E.; Knoops, J.; Bessin, Y.; Fossepre, M.; Chamieh, J.; Dumy, P.; Surin, M.; Ulrich, S., One-Pot Self-Assembly of Peptide-Based Cage-Type Nanostructures Using Orthogonal Ligations. Chem-Eur J 2017, 23, (57), 14323-14331.
19. Basak, S.; Singh, I.; Ferranco, A.; Syed, J.; Kraatz, H. B., On the Role of Chirality in Guiding the Self-Assembly of Peptides. Angew Chem Int Edit 2017, 56, (43), 13288-13292.
20. Castelletto, V.; Kaur, A.; Hamley, I. W.; Barnes, R. H.; Karatzas, K. A.; Hermida-Merino, D.; Swioklo, S.; Connon, C. J.; Stasiak, J.; Reza, M.; Ruokolainen, J., Hybrid membrane biomaterials from self-assembly in polysaccharide and peptide amphiphile mixtures: controllable structural and mechanical properties and antimicrobial activity. Rsc Adv 2017, 7, (14), 8366-8375.
21. Chang, C.; Liang, P. Q.; Chen, L. L.; Liu, J. F.; Chen, S. H.; Zheng, G. H.; Quan, C. Y., pH-responsive nanoparticle assembly from peptide amphiphiles for tumor targeting drug delivery. J Biomat Sci-Polym E 2017, 28, (13), 1338-1350.
22. Clerici, F.; Erba, E.; Gelmi, M. L.; Pellegrino, S., Non-standard amino acids and peptides: From self-assembly to nanomaterials. Tetrahedron Lett 2016, 57, (50), 5540-5550.
23. Dasgupta, A., Exploring architectures at the nanoscale: the interplay between hydrophobic twin lipid chains and head groups of designer peptide amphiphiles in the self-assembly process and application. Soft Matter 2016, 12, (19), 4352-4360.
24. Hsieh, M. C.; Lynn, D. G.; Grover, M. A., Kinetic Model for Two-Step Nucleation of Peptide Assembly. J Phys Chem B 2017, 121, (31), 7401-7411.
25. Han, H. Y.; Weinreb, P. H.; Lansbury, P. T., The Core Alzheimers Peptide Nac Forms Amyloid Fibrils Which Seed and Are Seeded by Beta-Amyloid - Is Nac a Common Trigger or Target in Neurodegenerative Disease. Chem Biol 1995, 2, (3), 163-169.
26. Gazit, E., Mechanistic studies of the process of amyloid fibrils formation by the use of peptide fragments and, analogues: Implications for the design of fibrillization inhibitors. Curr Med Chem 2002, 9, (19), 1725-1735.
27. Zanuy, D.; Porat, Y.; Gazit, E.; Nussinov, R., Peptide sequence and amyloid formation: Molecular simulations and experimental study of a human islet amyloid polypeptide fragment and its analogs. Structure 2004, 12, (3), 439-455.
28. Reches, M.; Gazit, E., Self-assembly of peptide nanotubes and amyloid-like structures by charged-termini-capped diphenylalanine peptide analogues. Isr J Chem 2005, 45, (3), 363-371.
29. Gilead, S.; Gazit, E., Self-organization of short peptide fragments: From amyloid fibrils to nanoscale supramolecular assemblies. Supramol Chem 2005, 17, (1-2), 87-92.
30. Reches, M.; Gazit, E., Casting metal nanowires within discrete self-assembled peptide nanotubes. Science 2003, 300, (5619), 625-627.

31. Yemini, M.; Reches, M.; Gazit, E.; Rishpon, J., Peptide nanotube-modified electrodes for enzyme-biosensor applications. Anal Chem 2005, 77, (16), 5155-5159.
32. Reches, M.; Gazit, E., Self-assembly of peptide nanotubes and amyloid-like structures by charged-termini capped diphenylalanine peptide analogues. Febs J 2005, 272, 388-388.
33. Kol, N.; Adler-Abramovich, L.; Barlam, D.; Shneck, R. Z.; Gazit, E.; Rousso, I., Self-assembled peptide nanotubes are uniquely rigid bioinspired supramolecular structures. Nano Lett 2005, 5, (7), 1343-1346.
34. Gazit, E.; Reches, M., Self-assembly of peptide nanotubes by charged-termini capped diphenylailanine peptide analogue. Abstr Pap Am Chem S 2005, 229, U1154-U1154.
35. Reches, M.; Gazit, E., Controlled patterning of aligned self-assembled peptide nanotubes. Nat Nanotechnol 2006, 1, (3), 195-200.
36. Carny, O.; Shalev, D. E.; Gazit, E., Fabrication of coaxial metal nanocables using a self-assembled peptide nanotube scaffold. Nano Lett 2006, 6, (8), 1594-1597.
37. Reches, M.; Gazit, E., Molecular self-assembly of peptide nanostructures: mechanism of association and potential uses. Curr Nanosci 2006, 2, (2), 105-111.
38. Gazit, E.; Boman, A.; Boman, H. G.; Shai, Y., Interaction of the Mammalian Antibacterial Peptide Cecropin P1 with Phospholipid-Vesicles. Biochemistry-Us 1995, 34, (36), 11479-11488.
39. Ghosh, S.; Reches, M.; Gazit, E.; Verma, S., Bioinspired design of nanocages by self-assembling triskelion peptide elements. Angew Chem Int Edit 2007, 46, (12), 2002-2004.
40. Kornmueller, K.; Lehofer, B.; Meindl, C.; Frohlich, E.; Leitinger, G.; Amenitsch, H.; Prassl, R., Peptides at the Interface: Self-Assembly of Amphiphilic Designer Peptides and Their Membrane Interaction Propensity. Biomacromolecules 2016, 17, (11), 3591-3601.
41. Hamley, I. W.; Hutchinson, J.; Kirkham, S.; Castelletto, V.; Kaur, A.; Reza, M.; Ruokolainen, J., Nanosheet Formation by an Anionic Surfactant-like Peptide and Modulation of Self-Assembly through Ionic Complexation. Langmuir 2016, 32, (40), 10387-10393.
42. Zhang, S. G., Discovery and design of self-assembling peptides. Interface Focus 2017, 7, (6).
43. Magnotti, E. L.; Hughes, S. A.; Dillard, R. S.; Wang, S. Y.; Hough, L.; Karumbamkandathil, A.; Lian, T. Q.; Wall, J. S.; Zuo, X. B.; Wright, E. R.; Conticello, V. P., Self-Assembly of an alpha-Helical Peptide into a Crystalline Two-Dimensional Nanoporous Framework. J Am Chem Soc 2016, 138, (50), 16274-16282.
44. Adler-Abramovich, L.; Gazit, E., Controlled patterning of peptide nanotubes and nanospheres using inkjet printing technology. J Pept Sci 2008, 14, (2), 217-223.
45. Reches, M.; Gazit, E., Biological and chemical decoration of peptide nanostructures via biotin-avidin interactions. J Nanosci Nanotechno 2007, 7, (7), 2239-2245.
46. Gazit, E., Self-assembled peptide nanostructures: the design of molecular building blocks and their technological utilization. Chem Soc Rev 2007, 36, (8), 1263-1269.

47. Krysmann, M. J.; Castelletto, V.; McKendrick, J. E.; Clifton, L. A.; Hamley, I. W.; Harris, P. J. F.; King, S. A., Self-assembly of peptide nanotubes in an organic solvent. Langmuir 2008, 24, (15), 8158-8162.
48. Frederix, P. W. J. M.; Ulijn, R. V.; Hunt, N. T.; Tuttle, T., Virtual Screening for Dipeptide Aggregation: Toward Predictive Tools for Peptide Self-Assembly. J Phys Chem Lett 2011, 2, (19), 2380-2384.
49. Huang, R. L.; Su, R. X.; Qi, W.; Zhao, J.; He, Z. M., Hierarchical, interface-induced self-assembly of diphenylalanine: formation of peptide nanofibers and microvesicles. Nanotechnology 2011, 22, (24).
50. Guo, C.; Luo, Y.; Zhou, R. H.; Wei, G. H., Probing the Self-Assembly Mechanism of Diphenylalanine-Based Peptide Nanovesicles and Nanotubes. Acs Nano 2012, 6, (5), 3907-3918.
51. Rissanou, A. N.; Georgilis, E.; Kasotaids, E.; Mitraki, A.; Harmandaris, V., Effect of Solvent on the Self-Assembly of Dialanine and Diphenylalanine Peptides. J Phys Chem B 2013, 117, (15), 3962-3975.
52. Lu, Y. Q.; Wang, M. F.; Qi, W.; Su, R. X.; He, Z. M., Artificial Hydrolase Based on Short Peptides Self- and Co-assembly Nanofiber. Chem J Chinese U 2015, 36, (7), 1304-1309.
53. Adler-Abramovich, L.; Marco, P.; Amon, Z. A.; Creasey, R. C. G.; Michaels, T. C. T.; Levin, A.; Scurr, D. J.; Roberts, C. J.; Knowles, T. P. J.; Tendler, S. J. B.; Gazit, E., Controlling the Physical Dimensions of Peptide Nanotubes by Supramolecular Polymer Coassembly. Acs Nano 2016, 10, (8), 7436-7442.
54. Guo, C.; Amon, Z. A.; Qi, R. X.; Zhang, Q. W.; Adler-Abramovich, L.; Gazit, E.; Wei, G. H., Expanding the Nanoarchitectural Diversity Through Aromatic Di- and Tri-Peptide Coassembly: Nanostructures and Molecular Mechanisms. Acs Nano 2016, 10, (9), 8316-8324.
55. Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H., The MARTINI force field: Coarse grained model for biomolecular simulations. J Phys Chem B 2007, 111, (27), 7812-7824.
56. Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J., The MARTINI coarse-grained force field: Extension to proteins. J Chem Theory Comput 2008, 4, (5), 819-834.
57. de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schafer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J., Improved Parameters for the Martini Coarse-Grained Protein Force Field. J Chem Theory Comput 2013, 9, (1), 687-697.
58. Bekker, H.; Berendsen, H. J. C.; Dijkstra, E. J.; Achterop, S.; Vondrumen, R.; Vanderspoel, D.; Sijbers, A.; Keegstra, H.; Reitsma, B.; Renardus, M. K. R., Gromacs - a Parallel Computer for Molecular-Dynamics Simulations. Physics Computing '92 1993, 252-256.
59. Berendsen, H. J. C.; Vanderspoel, D.; Vandrunen, R., Gromacs - a Message-Passing Parallel Molecular-Dynamics Implementation. Comput Phys Commun 1995, 91, (1-3), 43-56.
60. Vandrunen, R.; Vanderspoel, D.; Berendsen, H. J. C., Gromacs - a Software Package and a Parallel Computer for Molecular-Dynamics. Abstr Pap Am Chem S 1995, 209, 49-Comp.

61. Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C., GROMACS: Fast, flexible, and free. J Comput Chem 2005, 26, (16), 1701-1718.
62. Goga, N.; Marin, I.; Vasilateanu, A.; Pavaloiu, I. B.; Kadiri, K. O.; Awodele, O., Improved GROMACS Algorithms using the MPI Parallelization. 2015 E-Health and Bioengineering Conference (Ehb) 2015.
63. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. J Chem Phys 2007, 126, (1).
64. Bussi, G.; Zykova-Timan, T.; Parrinello, M., Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. J Chem Phys 2009, 130, (7).
65. Parrinello, M.; Rahman, A., Crystal-Structure and Pair Potentials - a Molecular-Dynamics Study. Phys Rev Lett 1980, 45, (14), 1196-1199.
66. Parrinello, M.; Rahman, A., A Molecular-Dynamics Study of Crystal-Structure Transformations. B Am Phys Soc 1981, 26, (3), 380-380.
67. Parrinello, M.; Rahman, A., Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method. J Appl Phys 1981, 52, (12), 7182-7190.
68. Aydin, F.; Ludford, P.; Dutt, M., Phase segregation in bio-inspired multi-component vesicles encompassing double tail phospholipid species. Soft Matter 2014, 10, (32), 6096-6108.
69. Aydin, F.; Uppaladadium, G.; Dutt, M., Harnessing Nanoscale Confinement to Design Sterically Stable Vesicles of Specific Shapes via Self-Assembly. J Phys Chem B 2015, 119, (32), 10207-10215.
70. Aydin, F.; Uppaladadium, G.; Dutt, M., The design of shape-tunable hairy vesicles. Colloid Surface B 2015, 128, 268-275.
71. Walker, S. A.; Chiruvolu, S.; Zasadzinski, J. A.; Schmitt, F. J.; Israelachvili, J. N., Controlled multi-stage self-assembly of vesicles. Hollow and Solid Spheres and Microspheres: Science and Technology Associated with Their Fabrication and Application 1995, 372, 95-100.
72. Israelachvili, J. N., Thermodynamic Principles of Self-Assembly. Intermolecular and Surface Forces, 3rd Edition 2011, 503-534.
73. Song, Y. J.; Challa, S. R.; Medforth, C. J.; Qiu, Y.; Watt, R. K.; Pena, D.; Miller, J. E.; van Swol, F.; Shelnutt, J. A., Synthesis of peptide-nanotube platinum-nanoparticle composites. Chem Commun 2004, (9), 1044-1045.

74. Schwierz, N., Frost, C.V., Geissler, P.L., Zacharias,M., Dynamics of Seeded Aβ40-Fibril Growth from Atomistic Molecular Dynamics Simulations: Kinetic Trapping and Reduced Water Mobility in the Locking Step, J. Am. Chem. Soc. 2016, 138, 2, 527–539

75. Wiehe, K.; Schmidler, S. C. *Monitoring Convergence of Molecular Simulations in the Presence of Kinetic Trapping*; Citeseer, 2011

76. Swendsen, R.H., Replica Monte Carlo Simulation of Spin-Glasses ,Phys. Rev. Lett. 57, 2607 75. Sugita, Y., Okamoto, Y., Replica-exchange molecular dynamics method for protein folding, Chemical Physics Letters 314 1999 141–151

77. Meli, M., Colombo, G., A Hamiltonian Replica Exchange Molecular Dynamics (MD) Method for the Study of Folding, Based on the Analysis of the Stabilization Determinants of Proteins, Int J Mol Sci. 2013 Jun; 14(6): 12157–12169.

78. Swails, J.M., York, D.M., Roitberg, A.E., Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation, J. Chem. Theory Comput. 2014, 10, 3, 1341-1352

79. Chodera, J.D., Shirts, M.R., Replica exchange and expanded ensemble simulations as Gibbs sampling: Simple improvements for enhanced mixing, J. Chem. Phys. 135, 194110 (2011)

80. Lee, K.H., Chen, J., Efficacy of Independence Sampling in Replica Exchange Simulations of Ordered and Disordered Proteins, J Comput Chem. 2017 Nov 15; 38(30): 2632–2640.

81. Plattner, N., Doll, J.D., Meuwly, M., Overcoming the Rare Event Sampling Problem in Biological Systems with Infinite Swapping J. Chem. Theory Comput. 2013, 9, 9, 4215-4224

82. Nagai, T. & Yuko Okamoto, Y. Replica-exchange molecular dynamics simulation of a lipid bilayer system with a coarse-grained model, Molecular Simulation, 38:5, 437-441(2012)

83. Stelter, D., Keyes, T., nhanced Sampling of Phase Transitions in Coarse-Grained Lipid Bilayers, J. Phys. Chem. B 2017, 121, 23, 5770–5780

84. Chebaro, Y., Dong, X., Laghaei,R., Derreumaux, P., and Mousseau, N.,Replica Exchange Molecular Dynamics Simulations of Coarse-grained Proteins in Implicit Solvent, J. Phys. Chem. B 2009, 113, 1, 267–274

85. Balasubramanian, Vivekanandan, Antons Treikalis, Ole Weidner, and Shantenu Jha. “Ensemble toolkit: Scalable and flexible execution of ensembles of tasks.” In Parallel Processing (ICPP), 2016 45th International Conference on, pp. 458-463. IEEE, 2016

86. Ian Goodfellow and Yoshua Bengio and Aaron Courville (2016). Deep Learning. MIT Press.

87. LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). "Deep learning". *Nature*. **521** (7553): 436–444

88. Feffer, S., “It’s All About The Features” https://reality.ai/it-is-all-about-the-features/