COMPOSITION, THERMODYNAMICS, AND MORPHOLOGY: A MULTI-SCALE COMPUTATIONAL APPROACH FOR THE DESIGN OF SELF-ASSEMBLING PEPTIDES

BY

BRYCE A. THURSTON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Jun Song, Chair
Professor Andrew Ferguson, Director of Research
Professor Jianjun Cheng
Professor Lance Cooper

Abstract

Peptide self-assembly has generated significant interest as a means for the bottom-up fabrication of highly tunable biocompatible nanoaggregates. Individual peptides can be synthesized to include non-natural π-conjugated subunits, endowing assembled aggregates with a range of optical and electronic properties that render them useful in applications as biocompatible organic electronics. The immense number of possible peptides, however, causes the exhaustive traversal of sequence space to be intractable. This massive composition space lends itself toward the use of computer simulation and data science tools to understand molecular aggregation and guide experimental synthesis and design. In this dissertation, I present work employing a hierarchy of molecular modeling techniques to identify self-assembling peptides with specific photophysical properties by probing thermodynamic and structural characteristics of peptide aggregation. We employ classical molecular dynamics simulation to probe the key molecular forces governing the morphology and free energy of oligomerization, time dependent density functional theory to predict photophysical properties as a function of aggregate morphology, and data-driven quantitative structure property models to perform high-throughput virtual screening of chemical space to identify promising peptide chemistries. This work establishes a multi-scale framework for the principled computational design of self-assembling π-conjugated peptides with engineered photophysical properties.

Soli Deo Gloria

Acknowledgments

This work was funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award # DE-SC0004857 and by the National Science Foundation Grant No. DMR-1729011.

I am indebted to a multitude of people, without whom this work certainly could not have been done. I am deeply grateful for my advisor Andrew Ferguson, and the many interactions we have had in working together. Thank you very much for your mentorship, I could not have hoped for a more patient and supportive advisor. I certainly would not have made it to this point without you. I am also thankful for the other members of the Ferguson Group: Andy, Greg, Jiang, Wei, Yutao, and others whose input provided a number of useful contributions to this work. I especially would like to thank Rachael for the conversations we had in working together to better understand the peptides studied in this thesis.

There are a many collaborators we have worked with over the years on the projects here. I thank Lawrence Valverde, whose FCS work and initiative served as the impetus for Chapter 4. I thank J. D. Tovar for many helpful conversations we have had in preparing this work, particularly in Chapters 2 and 5. I would also like to thank Andre Schleife, Ethan Shapera, Bill Wilson, J. J. Cheng, and Charles Shcroeder for our interactions over the years.

I am also very grateful for my thesis committee: thank you for your time and willingness to serve in this manner.

I would also like to deeply thank the Physics Department at the University of Illinois. It took me no more than an hour on the visitation weekend to realize this department is top notch, both in quality of academic research as well as quality of mentorship of graduate students. Everything that has happened since has only served to confirm this, so I am very thankful for the time I have been able to spend here.

I am also grateful for the excellent experience I had as an undergraduate student at the Colorado School of Mines. In particular, I would like to thank Mark Lusk, Lincoln Carr, Jeffrey Squier, and Mark Coffey for their support. I also had an excellent experience for two summers as an undergraduate at NIST, from which I would like to thank Kartik Srinivasan and Albert Talin for taking me on as a summer student.

I have had many other great and inspirational teachers and coaches over the years. I would particularly

like to thank Damon Larson, Judy Murray, Brian Velarde, Rodney Padilla, and Mark Kelly.

No computational lab is complete without a good sys admin, and Gianni Pezzarossi is one of the best. His careful efforts kept our computer clusters running and our data servers safe, and his computer cluster performed a significant amount of the computational work reported here.

I am very grateful for my many friends at All Souls Presbyterian church, who are too numerous to exhaustively name. Thank you to Pastor Luke, Brian, and Scott for you ministry. I am especially thankful for my dear friends Mark, Emily, and Rachel. Without your support, encouragement, and prayers, I would not have made it to this point. I would also like to thank Nate and Rebekah for your friendship over the years.

I am deeply thankful for my whole family as well. In particular, thank you to my dear sisters for your love and friendship as we have grown up together. I would also like to thank my Uncle Greg for your kind discipleship, as well as my cousins Paul and David for your friendship.

Mom and Dad, thank you from the bottom of my heart for all that you have done for me. You have always been immensely supportive of me. In everything I have ever done you have always been behind me 100%. In so many different ways, I would not be where I am today or who I am today without your loving care.

Finally, I have an unpayable debt to my Lord and Savior Jesus Christ. 1 Timothy 1:15: "It is a trustworthy statement, deserving full acceptance, that Christ Jesus came into the world to save sinners, among whom I am foremost of all."

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A	Alanine (same as ALA)
AATS6s_wing	Average Broto-Moreau autocorrelation of lag 6 weighted by I-state of the peptide wing
ACE	Analytical Continuum Electrostatic approximation
ATSC4e_wing	Centered Broto-Moreau autocorrelation of lag 4 weighted by Sanderson electronegativities
ATSC4p_wing	Centered Broto-Moreau autocorrelation of lag 4 weighted by polarizabilities
AVP_5	Average Valance Path of order 5
BI	Boltzmann Inversion
C	Cysteine (same as CYS)
CD	Circular Dichroism
COM	Center Of Mass
CTMC	Continuous Time Markov Chain
D	Aspartic acid (same as ASP)
DFT	Density functional theory
DTMC	Discrete Time Markov Chain
E	Glutamic acid (same as GLU)
ESP	Electrostatic Potential
F	Phenylalanine (same as PHE)
G	Glycine (same as GLY)
GAFF	Generalized Amber Force Field
GATS2c	Geary autocorrelation of lag 2 weighted by charges
GATS2i	Geary autocorrelation of lag 2 weighted by first ionization potential
GATS6i_wing	Geary autocorrelation of lag 6 weighted by the first ionization potential of the peptide wing
GBSA	Generalized Born Solvent Accessible surface area approximation for implicit solvent
H	Histidine (same as HIS)
HD-LSS	High Dimensionality, Low Sample Size

h2t	Head to tail distance between terminal alpha carbons in a peptide
K	Lysine (same as LYS)
L	Leucine (same as LEU)
LINCS	LINear Constraint Solver (an algorithm for constraining bond lengths)
LOO-CV	Leave-One-Out Cross-Validation
LR-TDDFT	Linear-response time-dependent density functional theory
M	Methionine (same as MET)
MAE	Mean Average Error
MATS1c_wing	Moran autocorrelation of lag 1 weighted by charges of the peptide wing
MATS3c	Moran autocorrelation of lag 3 weighted by charges
MATS8s	Moran autocorrelation of lag 8 weighted by I-state.
maxHBint10	descriptor of strength for potential hydrogen bonds of path length 10
maxsssCH_wing	Maximum atom-type E-State for singly bonded carbons with one hydrogen of the peptide wing
MDEC-23	Molecular distance edge between all secondary and tertiary carbons
MD	Molecular dynamics
MLR	Multiple linear regression
N	Asparagine (same as ASN)
NDI	Naphthalenediimide
OBC	Onufriev, Bashford, Case model for implicit solvent
OPVx	oligo(p-phenylenevinylene) with x repeat units (eg. OPV3)
OTx	oligothiophene with x repeat units (eg. OT4)
P	Proline (same as PRO)
PDI	Perylenediimide
piPC3	Conventional bond order ID number of order 3
PME	Particle Mesh Ewald (an algorithm for simulation of long-range electrostatics)
PMF	Potential of Mean Force
Q	Glutamine (same as GLN)
QSAR/QSPR	Quantitative structure-activity relationship / Quantitative structure-property relationship
R	Arginine (same as ARG)
REDS	Restrained electrostatic potential/ Electrostatic potential charge Derive Server
RESP	Restrained Electrostatic Potential
RMSE	Root Mean Square Error

S	Serine (same as SER)
SPC	Simple Point Charge
SpMAD-Dzp	Spectral mean absolute deviation from Barysz matrix weighted by polarizabilities
SpMax6_Bhs	6th largest absolute eigenvalue of the Burden modified matrix weighted by the relative I-state
SpMin5-Bhm	5th smallest absolute eigenvalue of Burden modified matrix weighted by relative mass
T	Threonine (same as THR)
TDDFT	Time-dependent density functional theory
TIP3p	Transferable Intermolecular Potential with 3 Points
V	Valine (same as VAL)
W	Tryptophan (same as TRP)
WHAM	Weighted Histogram Analysis Method
X	Designates an arbitrary amino acid
Y	Tyrosine (same as TYR)

List of Symbols

a	Alignment metric
$c^{\ominus}$	standard concentration
D	Einstein diffusion coefficient
ΔF_n	Change in free energy in forming a peptide aggregate of size n from an aggregate of size $n-1$ and a peptide monomer
$\mathbf{I}$	Identity matrix
$K^{\ominus}$	Equilibrium constant at standard conditions
k_B	Boltzmann's constant
n	aggregate size
N	Number of atoms/molecules
n_a	Number of peptides in an aligned cluster
n_c	Number of peptides in a contact cluster
nn_i	ith nearest neighbor
P	Pressure
$\mathbf{p}(t)$	Mass weighted cluster size as a function of time
$\mathbf{Q}$	Transition rate matrix
q^2	Coefficient of determination for leave-one-out cross validation
q_i	Partial charge of atom i
Q_i	Steinhardt bond-order parameter of order i
R^2	Coefficient of determination
R^2_{adj}	Coefficient of determination adjusted to penalize more complex models
R_g	Radius of gyration
r_{ij}	distance between atom i and atom j
$\text{RMS}^{\text{core}}_{\text{plane}}$	RMS deviation of a peptide core from a plane
$\text{RMS}^{\text{ring}}_{\text{plane}}$	Mean RMS deviation of individual aromatic rings within a peptide core from a plane

ΔS	Change in entropy
T	Temperature
$\mathbf{T}(t)$	Time-dependent transition probability matrix
t	time
t_2	slowest relaxation time of markov matrix
ΔU	Change in potential energy
V	Volume
V_C	Coulombic potential
V_{LJ}	Lennard Jones potential
V^{NB}	Non-bonded potential energy
$[X]$	Concentration of species X
α	scaling fraction, or first free parameter in determining effective Born radii in OBC implicit solvent model
β	$\frac{1}{k_B T}$, or second free parameter in determining effective Born radii in OBC implicit solvent model
β-sheet	Secondary structure of protein consisting of lateral stacking of, and hydrogen bonding between, amino acid back bones
γ	friction constant for Langevin dynamics integration, or third free parameter in determining effective Born radii in OBC implicit solvent model
ϵ_0	Permittivity of free space
ϵ_{ij}	Lennard Jones potential well depth for potential between atoms i and j
ϵ_r	Relative permittivity
θ_i	Angle between the backbone of two peptide cores that are ith nearest neighbors
Π	Arbitrary π-conjugated core
ρ	Pearson correlation coefficient
σ_{AB}	symmetry number (2 if A=B, 1 otherwise)
σ_{ij}	Center of mass separation for Lennard Jones potential energy zero between atoms i and j
τ	Markov lag time
ϕ_i	Angle between the vectors normal to two cores that are ith nearest neighbors
ψ^{ring}	Mean angle between the vector normal to the planes of adjacent rings within a peptide core

Chapter 1

Introduction

1.1 Peptide self-assembly

The self-assembly of peptides is a promising methodology for the fabrication of novel macromolecular materials with desirable structural, functional, and biological properties [6–12]. Such materials have potential applications in drug and protein delivery, material templating of inorganic structures, regenerative medicine, and antimicrobials [7, 8, 12–19]. Assembly can be triggered by environmental variables such as primary structure, pH, temperature, and salt concentration [20–28] allowing for many degrees of freedom with which to influence peptide assembly. In addition to standard amino acids, synthetic oligopeptides can be functionalized with polymeric π-conjugated inserts [25, 29–32], endowing the self-assembled supramolecular aggregates with optoelectronic and photophysical activity. The resultant electrical and electronic properties – electron transport or exciton coupling, for example – provide the basis for a diverse array of organic electronic devices, such as light-emitting diodes, field-effect transistors, and solar cells [33–44]. Deterministic control of the structure, stability, and kinetics of self-assembled organic electronics by tuning monomer chemistry and environmental conditions presents a powerful route to the fabrication of "designer materials" possessing desirable structural and functional properties [45]. Conjugated peptides in particular offer a water-soluble and biocompatible medium to fabricate self-assembled aggregates with tunable biological and electronic properties [2, 4, 22, 39, 40, 46–49].

The chemical sequence space accessible to synthetic oligopeptides is vast, and it is of value to understand the microscopic molecular forces and mechanisms governing assembly in order to provide rational principles to guide experimental peptide design towards candidates with good assembly behavior. Such properties, however, can be difficult to access experimentally. As a result, computational studies provide an attractive method to elucidate these interactions [4, 50–57]. Specifically, molecular dynamics (MD) simulations provide a useful tool for probing the structure, dynamics, and thermodynamics of molecular systems [50–53, 58]. As a classical simulation, however, the excited state properties of a molecule are not accessible to MD simulations. Density functional theory is a common simulation technique used to take into account the quantum properties of a molecule due to its relatively high degree of accuracy without too large of a computational cost [59–64]. Incorporating both large length and time scale thermodynamic properties as well as quantum properties of self-assembling π-conjugated peptides will be important to further understand these systems from a computational standpoint. The goal of this work is to establish a procedure for the computational design of self-assembling peptides with tunable photophysical properties. In Chapter 2, we present work using MD simulations to probe the thermodynamic, kinetic, and structural properties of smaller scale aggregation of peptides having a DFAG-OPV3-GAFD composition. This chapter lays the groundwork for our use of MD simulations to study peptides of this nature. In Chapter 3 we expand our computations to include a variety of amino acid sequences and core composition in order to determine simple relationships between peptide composition and properties of interest in assembly. This chapter demonstrates the applicability of QSPR modeling to the prediction of thermodynamic properties underpinning peptide aggregation, and sheds light on the relationship between these properties and peptide alignment. In Chapter 4 we challenge the assumption of no aggregation in peptide systems at neutral pH, and predict the extent of prenucleation in untriggered peptide systems. This chapter provides a better understanding for the initial conditions which govern the kinetics of peptide assembly. In Chapter 5 we present ongoing work relating the geometries observed in MD simulation to excited state properties of peptide aggregates. This chapter lays the ground work for conducting *in silico* design of self-assembling π-conjugated peptides with specific excited state properties of interest. Finally, in Chapter 6 we present conclusions and future directions for this work.

Chapters 2-4 are based in full or in part on the following publications

- Thurston, B. A.; Tovar, J. D.; Ferguson, A. L. *Mol. Sim.* 42, 12, 955-975 (2016).

- Thurston, B. A.; Ferguson, A. L. *Mol. Sim.* 44, 11, 930-945 (2018).

- Valverde, L. R.; Thurston, B. A.; Ferguson, A. L.; Wilson, W. L. *Langmuir*, 34, 25, 7346-7354 (2018).

Chapter 2

Molecular Modeling of Assembly

2.1 Introduction

In this chapter, we present a theoretical study of the self-assembly of synthetic ASP-PHE-ALA-GLY-OPV3-GLY-ALA-PHE-ASP (DFAG) oligopeptides containing a π-conjugated oligophenylenevinylene (OPV) core (Figure 2.1). This prototypical peptide-Π-peptide triblock architecture presents a powerful and flexible architecture for self-assembling optoelectronic peptidic biomaterials [4, 56]. Experimental work has shown these peptides to exist as water-soluble monomers at neutral or basic pH, and under acidic conditions – due to protonation of the carboxylic acid termini that screens inter-peptide Coulombic repulsion – to self-assemble into 1D ribbons driven by inter-peptide hydrogen bonding and π-stacking of the conjugated cores [4, 23, 65, 66]. Experimentally, self-assembled 1-D nanomaterials formed from a wide variety of aromatic cores have been observed, including the phenylene vinylene subunit considered here, electron-rich oligothiophenes, diimide-based electron-deficient π-systems, and a series of polyaromatics of similar composition but increasing size (i.e., bi, ter tetra, quinque and sexithiophene) [67, 68]. Electronic delocalization along the aromatic core of this supramolecular construct imbues these aggregates with useful optoelectronic properties, making them putative candidates as biocompatible conductive ribbons at the biotic-abiotic interface, and as a triggerable, biocompatible, electro-conductive material with biosensing applications [22]. Previous experimental work has considered the impact of variations in the size of the conjugated core, peptide sequence, and N-to-C polarity upon the structural and photophysical properties of the peptide assemblies [4, 23, 30, 56, 69]. Theoretical work has probed the impact of peptide sequence N-to-C polarity and the impact of the number of OPV subunits upon the thermodynamic driving forces for dimerization and peptide sequence upon the morphology of peptide ribbons [4, 56], but the elementary kinetic steps and molecular mechanisms of self-assembly remain unknown. This chapter establishes

Most of this chapter is an excerpt from from ref. [55]. Used with permission from Thurston, B. A.; Tovar, J. D.; Ferguson, A. L. *Mol. Sim.* 42, 12, 955-975 (2016), available at https://doi.org/10.1080/08927022.2015.1125997. Copyright 2016 Taylor and Francis.

fundamental understanding of the early stages of DFAG-OPV3-GAFD self-assembly using molecular simulation.

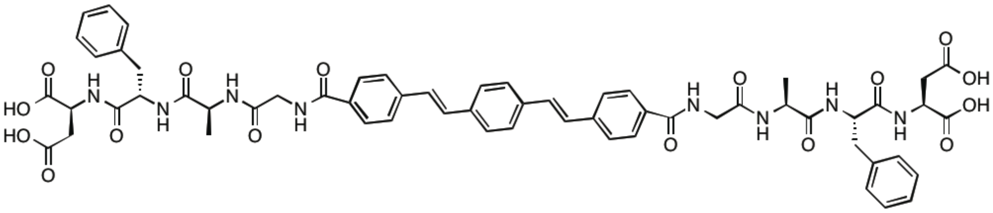

Figure 2.1: Chemical structure of the DFAG-OPV3-GAFD peptide studied in this chapter. The N-terminus to C-terminus directionality of each peptide wing proceeds away from the conjugated core, leading to an oligopeptide architecture possessing two C-termini. The pK_A of the carboxyl terminus and aspartic acid side chain residing at the C-termini are 2.09 and 3.86, respectively [3]. At pH $\sim$5 or greater, the oligopeptides are effectively fully deprotonated, carrying a formal charge of (−4) that precludes extended assembly due to Coulombic repulsion [4]. At pH $\sim$1 or less, the peptides are effectively fully protonated, and self-assemble into aggregates stabilized by inter-peptide hydrogen bonding and π-stacking of the conjugated cores.

Molecular dynamics simulations provide a means to study the atomistic structure of peptide assemblies and probe the thermodynamics and kinetics of peptide aggregation. In this chapter, we employ atomistic molecular dynamics simulations in explicit water to study the thermodynamic stability and configurational motions of isolated DFAG peptide monomers and the formation of peptide dimers. We then use the atomistic free energy landscapes to parameterize an implicit-solvent forcefield for the DFAG peptides that enables us to study the assembly of higher order aggregates (i.e., trimers, tetramers, pentamers) at length and time scales inaccessible to the explicit solvent model. As demonstrated by Mondal *et al.*, an understanding of higher-order aggregate formation, not just dimerization, is vital for a complete understanding of multi-body self-assembly [30]. Building upon a body of prior experimental and simulation work [4, 23, 66], this study advances the fundamental understanding of the molecular forces and mechanisms driving assembly of a prototypical peptide-Π-peptide triblock molecular to help guide and inform the design of self-assembly biocompatible and optoelectronic peptidic biomaterials.

2.2 Methods

2.2.1 Explicit solvent simulations

Molecular dynamics simulations were conducted using the GROMACS 4.6 simulation suite [70]. Initial DFAG-OPV3-GAFD peptide configurations were constructed with the assistance of the GlycoBioChem PRODRG2 Server [71]. Peptides were prepared in two protonation states, a high-pH (pH ≥ 5) deprotonated state carrying a formal charge of (−4) on the terminal ASP residues, and a low-pH (pH $\leq$ 1) protonated state that was electrically neutral. The protonated state is of primary interest as it is in this state that the peptides are observed to self-assemble into 1D ribbons [4]. Single peptides or pairs of peptides were placed in a 10×10×10 nm cubic simulation box with three-dimensional periodic boundary conditions and solvated with water molecules to a density of 0.994 g/cm^3. Where necessary, Na$^+$ counterions were added such that the system carried no net charge. Peptides and ions were modeled using the CHARMM27 force field [72], and water with the simple point charge (SPC) model [73]. Parameters for the OPV3 conjugated core do not exist natively within the CHARMM27 force field, but were straightforwardly derived by analogy with existing groups. Specifically, the aromatic rings were modeled using the parameters taken from the phenylalanine residue, the carbonyl linker from a component of a peptide bond, and the vinyl group treated as an alkene chain. In total, two additional bonds, 6 angles, and 11 proper dihedrals were added. All force field files are available upon request. The size of the simulation box was sufficiently large that with a 1.0 nm real space cutoff two peptides – each a maximum of 3.75 nm long in their fully extended configurations – could be drawn far enough apart in our umbrella sampling of their dimerization pathway to be non-interacting (cf. Section 2.2.3). High energy overlaps in the initial configurations were removed by steepest descent energy minimization to eliminate forces exceeding 1000 kJ/mol.nm. Simulations were conducted in the NVT ensemble – fixed number of particles N, volume V, and temperature T – at 298 K, employing a stochastic dynamics approach to maintain the temperature by integrating the Langevin equation with a friction constant of $\gamma = 2$ ps^{-1} [74, 75]. Initial atom velocities were randomly assigned from a Maxwell distribution at 298 K and the equations of motion numerically integrated using a leap-frog algorithm with a 2 fs time step [76]. Bond lengths were fixed using the LINCS algorithm to improve efficiency [77]. Electrostatic interactions were treated using particle mesh Ewald (PME) with a real-space cutoff of 1.0 nm and a 0.12 nm Fourier grid spacing that were optimized during runtime [78]. Lennard-Jones interactions were shifted smoothly to zero at 1.0 nm, and Lorentz-Berthelot combining rules used to determine interaction parameters between unlike atoms [79]. A 1 ns equilibration run was conducted for each system, at which time the temperature, pressure, energy, and peptide radius of gyration had reached steady values. This equilibrated state served as the initial state for the umbrella sampling simulations detailed below.

2.2.2 Implicit solvent simulations

Simulations were conducted using the GROMACS 4.6 simulation suite [70] in which solvent was modeled implicitly using the Generalized Born model with a relative dielectric constant of 78.3 [80–82]. Born radii are calculated using the Onufriev, Bashford, Case (OBC) model with the OBC(II) optimized parameter set of $\alpha = 1$, $\beta = 0.8$, and $\gamma = 4.85$ [83], and recalculated every time step with a cutoff of 3.4 nm and a dielectric offset of 0.009 nm. Non-polar interactions are treated using the solvent-accessible surface area model. Calculations were performed using an analytical continuum electrostatic (ACE) type model using the Born radius of each atom and a surface tension parameter of 2.26 kJ/mol.nm^2 [84]. The Generalized Born model has been known to overestimate the stability of inter-residue interactions [85–87], but has also been shown to accurately treat solvent effects in simulating proteins [88–90] even when such proteins are stabilized by solvent effects [91, 92]. Accordingly, we make use of this implicit solvent model in order to reach the requisite time and length scales to observe peptide self-assembly, but – as detailed in Section 2.3.3 – we employed a rescaled version of the CHARMM27 force field in which the Lennard-Jones and Coulomb interactions were scaled such that intramolecular free energy landscapes for both the peptide monomer and the dimerization free energy pathway in implicit solvent reproduced those computed under explicit solvation. Simulations were otherwise conducted in the same manner as the explicit-solvent systems except that Lennard-Jones interactions were shifted smoothly to zero at a cutoff of 3.4 nm, and electrostatics were also treated by shifting to zero at a cutoff of 3.4 nm. As detailed in Section 2.3.3, we employed a rescaled version of the CHARMM27 forcefield in which the Lennard-Jones and Coulomb interactions were scaled such that intramolecular free energy landscapes for both the peptide monomer and the dimerization free energy pathway in implicit solvent reproduced those computed under explicit solvation.

2.2.3 Umbrella sampling

We employed umbrella sampling to compute the intramolecular and intermolecular peptide free energy landscapes along a preselected order parameter by applying artificial biasing potentials to enforce good sampling [93]. This free energy along a reaction coordinate is also referred to in the literature as the potential of mean force (PMF). For single peptides we construct the free energy landscape for peptide collapse by performing umbrella sampling along the intramolecular head-to-tail distance ($h2t$) between the C$_\alpha$-atoms of the terminal aspartic acid residues. For pairs of peptides, we compute the dimerization free energy landscape along the center of mass separation (r_{COM}) between the peptides. For triplets, we compute the free energy of monomer addition along the center of mass separation between a preassembled dimer and a monomer. For quads we compute the free energy for (i) monomer addition to a trimer, and (ii) assembly of two preassembled dimers along the center of mass separation between the two aggregates. For quints, we compute the free energy of monomer addition along the center of mass separation between a preassembled tetramer and a monomer. In all cases, we ensure that the umbrella sampling simulations are conducted at sufficiently large separations that we reach a plateau in the free energy, indicating that we have reached the regime at which the two aggregates in the simulation are effectively non-interacting and the PMF ceases to be a function of separation.

Umbrella windows were initialized by nonequilibrium pulling of our system along the entire range of the order parameter of interest over the course of 1 ns. Frames of this trajectory were harvested every 0.2 nm along the order parameter to serve as initial configurations for each umbrella sampling window. Harmonic biasing potentials with a force constant of 1000 kJ/mol.nm^2 are applied to restrain the system within each window and biased umbrella simulations conducted for 10 ns. The first 1 ns of each simulation was used to let the system equilibrate and discarded prior to analysis. The weighted histogram analysis method (WHAM) [94] implemented in the g_wham module of GROMACS 4.6 [70, 95] was applied to the umbrella sampling data to obtain the (relative) free energy of the system as a function of the umbrella coordinate, also known as the potential of mean force (PMF). If a region is found to be poorly sampled – around 15 times fewer samples than the best sampled regions – additional simulations were conducted in the undersampled region.

2.3 Results

We now detail the results of our computational investigation of the early-stage self-assembly of the DFAG peptide monomers. First we describe the calculation of the potentials of mean force for the collapse of isolated peptide monomers and the dimerization of peptide pairs using computationally expensive explicit solvent simulations, and the use of these data to parametrize an implicit solvent model to access to 2-3 orders of magnitude longer time and length scales. We then describe our use of the implicit solvent model to probe the thermodynamics and morphology of small oligomeric aggregates ($n = 2$-5) and directly simulate the first 70 ns of self-assembly from an initial dispersion of monomers.

2.3.1 PMF of peptide collapse

The potential of mean force (PMF) for a single peptide parameterized by the intramolecular distance between the ASP C_α atoms quantifies the relative propensities for elongated versus collapsed conformations of an isolated peptide. We present the PMFs calculated for the protonated peptide in explicit and implicit solvent in Figure 2.2a. In its protonated (low-pH) state, the peptide PMF in explicit solvent is essentially flat for head-to-tail distances $h2t = 1.5$-2.5 nm. Unfavorable configurations at long extensions correspond to energetically unfavorable extension of the torsional angles within the amino acid residues, and at short extensions to unfavorable bending of the backbone. The PMF for peptide collapse is governed principally by intramolecular interactions, with the conjugated OPV3 core remaining rather rigid and extended, and rotation of the Φ and Ψ dihedral angles of the peptide wings mediating close approach of the peptide head and tail. The PMF computed in implicit solvent shows very good agreement with the explicit curve over the range $h2t$ = 0.5-3.0 nm, but poorer agreement at longer extensions.

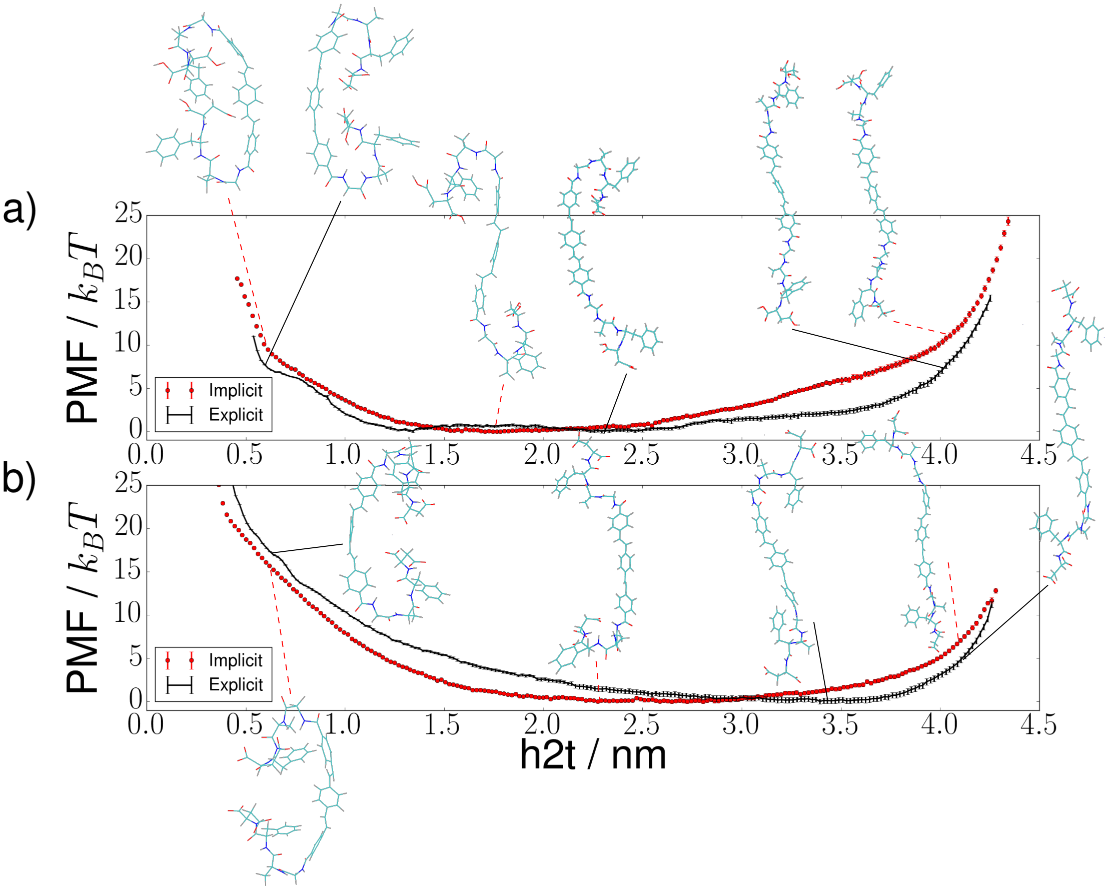

Figure 2.2: PMF curves for isolated DFAG-OPV3-GAFD peptides in explicit and implicit solvent in the (a) protonated (low-pH) and (b) deprotonated (high-pH) states parameterized by the intramolecular head-to-tail distance, $h2t$, between the C_α atoms of the terminal aspartic acid residues. Error bars in this and all subsequent PMF curves were computed by performing 100 bootstrap resamples of the data. Representative molecular structures extracted from our simulations and projected along the umbrella sampling coordinate in this and all subsequent figures were rendered in VMD [5].

The PMF curves for peptides in the deprotonated (high-pH) state in Figure 2.2b also illustrate relatively good agreement between the explicit and implicit solvent. The global free energy minimum in implicit solvent is at $h2t$ = 2.6 nm compared to $h2t$ = 3.4 nm for the explicit case, but the free energy difference is on the order of only ~1 k_BT. As anticipated, the most probable peptide configurations are displaced to longer extensions relative to the protonated state due to electrostatic repulsion between the doubly negatively charged termini.

2.3.2 PMF of peptide dimerization

The PMF for dimerization parameterized by the distance between the center of mass of the respective peptides estimates the changes in the free energy of a pair of peptides as they interact non-covalently through dispersion, Coulombic, and hydrophobic interactions. The PMFs for protonated peptides in implicit and explicit solvent are presented in Figure 2.3a. Five independent umbrella simulations were conducted in both implicit solvent and explicit solvent. As expected, dimerization is thermodynamically favorable. In implicit solvent the change in free energy for a pair of peptides from a non-interacting state to the minimum in free energy is $\Delta F_{dimer} = -(21.9 \pm 1.0)k_BT$, which differs significantly from the value of $\Delta F_{dimer} = -(15.7 \pm 0.7)k_BT$ computed in explicit solvent. Consistent with simulations of DNA base flipping in Ref. [96], we find simulations conducted in implicit solvent greatly overestimate the free energy computed in explicit solvent, in this case by 40%.

The PMF curves for deprotonated peptides are presented in Figure 2.3b. We compute the free energy of dimerization for deprotonated peptides in explicit solvent to be $\Delta F_{dimer} = -(4.7 \pm 0.9)k_BT$, indicating that formation of the contact pair is thermodynamically favored despite the electrostatic repulsion between the negatively charged termini. Inspection of the configurational ensemble in the global free energy reveals the associated pair to exist in an “I-shaped” configuration in which the OPV3 cores align yielding a favorable dispersion and π-stacking interactions between the hydrophobic conjugated cores while simultaneously orienting the negatively charged termini away from one another. As with the protonated peptides, the implicit solvent model prediction for the dimerization free energy, $\Delta F_{dimer} = -(11.4 \pm 1.7)k_BT$, overestimating that predicted by the explicit solvent model by 140%.

An understanding of higher-order aggregation is essential for a complete understanding of multi-body self-assembly [30], but the assembly of larger aggregates proceeds at length and time scales beyond those that can be reasonably obtained using explicit solvent. Implicit solvent simulations provide a means to explore the longer time assembly of larger aggregates, but the artifacts introduced by an implicit treatment of solvent observed in the dimerization PMFs indicate that the implicit forcefield requires reparameterization to match the atomistic results. We detail in Appendix B a systematic procedure to perform this reparameterization.

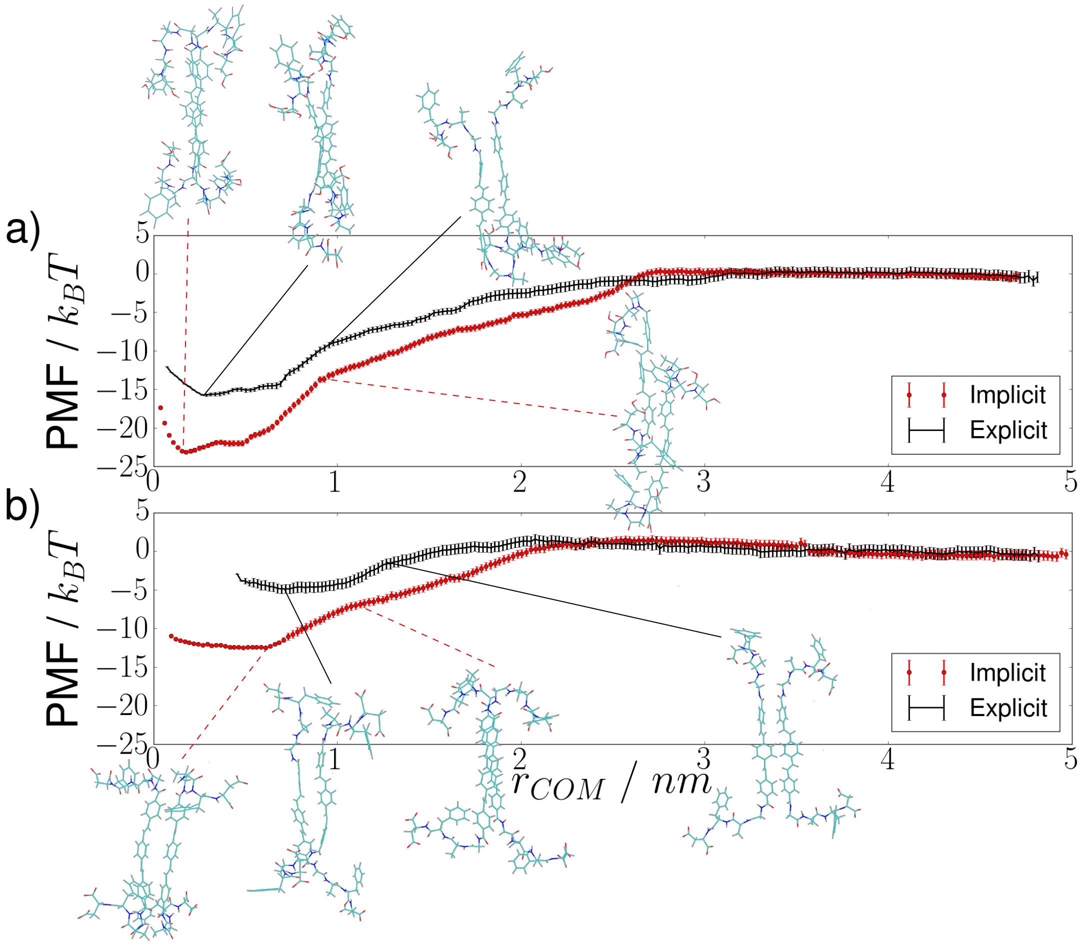

Figure 2.3: PMF curves for pairs of DFAG-OPV3-GAFD peptides in explicit and implicit solvent in the (a) protonated (low-pH) and (b) deprotonated (high-pH) states parameterized by peptide center of mass separation, r_{COM}. The dimerization free energy predicted by the explicit solvent model for the protonated (low-pH) peptides is strongly favorable at $\Delta F^{\mathrm{prot}}_{dimer} = -(15.7 \pm 0.7)\ k_BT$, whereas that for the deprotonated (high-pH) peptides is only weakly so at $\Delta F^{\mathrm{deprot}}_{dimer} = -(4.7 \pm 0.9)\ k_BT$.

2.3.3 Implicit solvent force field parameterization

We have observed that the use of an implicit solvent model introduces significant artifacts into the PMF curves for dimerization (Figure 2.3). To lend confidence to the predictions of the implicit solvent model at the length and time scales of multi-peptide assembly, we must reparameterize the CHARMM27 force field in order to match the quantities of interest [97–100]. In a similar manner to Zhang *et al.* [101], we adopt a minimally invasive strategy of rescaling interactions by a constant factor in order to optimally match the free energies calculated from implicit solvent simulations to those obtained in explicit solvent. Seeking to account for the absence of molecular water, we uniformly rescale the van der Waals and Coulomb interactions to reproduce the explicit solvent PMFs for peptide collapse and dimerization. This simple approach can be considered a form of the PMF matching approach to force field parameterization frequently used to optimize coarse-grained molecular potentials [96, 102, 103]. In this respect, our approach shares similarities with Boltzmann inversion (BI) wherein interaction potentials in the coarse-grained system are optimized to match distribution functions observed in all-atom simulations [100, 104]. In the case of pair potentials, this reduces to matching of the pairwise PMFs [105]. More sophisticated procedures to adjust the implicit solvent force field are possible [102, 104], but we demonstrate below that our simple strategy results in an implicit solvent model that quantitatively reproduces the explicit solvent results.

In our implicit solvent simulations, the Coulombic interaction energy between two atoms i and j separated by a distance r_{ij} is given by,

$$V_C(r_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_r r_{ij}}, \tag{2.1}$$

where q_i is the partial charge on atom i, ϵ_0 is the permittivity of free space, and $\epsilon_r = 78.3$ is the relative dielectric constant of liquid water [82]. The dispersion interaction between atoms i and j is given by the Lennard-Jones function,

$$V_{LJ} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{2.2}$$

where the ϵ_{ij} and σ_{ij} Lorentz-Berthelot combining rules were used to determine interaction parameters between unlike atom types: $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ and $\epsilon_{ij} = \sqrt{(\epsilon_i \epsilon_j)}$ [79]. We adopt a minimally perturbative approach to reparameterization of the CHARMM27 force field to account for the absence of molecular solvent by rescaling the non-bonded Lennard-Jones and Coulomb interactions by a single tuning parameter, α,

$$\epsilon \longrightarrow \alpha\epsilon \qquad q \longrightarrow \sqrt{\alpha}q, \tag{2.3}$$

that produces a uniform rescaling of the non-bonded interactions,

$$V^{nb} = V_C + V_{LJ} \longrightarrow \alpha V^{nb}, \tag{2.4}$$

Considering the protonated and deprotonated peptides independently, we sought the value of α resulting in an optimal match of the free energy well depth for dimerization ΔF_{dimer} measured by the absolute value of the difference between the implicit and explicit well depths,

$$\Delta\Delta F_{dimer}(\alpha) = |\Delta F^{\text{imp}}_{dimer}(\alpha) - \Delta F^{\text{exp}}_{dimer}|. \tag{2.5}$$

Implicit PMFs for the head-to-tail distance of an isolated monomer varied from those from the explicit simulations by an average of no more than 1.5 $k_B T$ for α = 0.6-0.9. Such variations are indistinguishable from thermal fluctuations, making the monomer PMF relatively insensitive to rescaling of the non-bonded interactions and a poor criterion by which to compute the optimal α.

We performed a series of umbrella sampling calculations of dimerization for both the protonated (low-pH) and deprotonated (high-pH) peptides at values of the scaling parameter α = {0.80, 0.82, 0.85, 0.87, 0.90, 1.00} for protonated peptides and α = {0.52, 0.60, 0.68, 0.70, 0.72, 0.76, 0.84, 0.92, 1.00} for deprotonated peptides. We illustrate in Figure 2.4 the corresponding values of $\Delta\Delta F_{dimer}(\alpha)$ calculated from these data. In the case of the protonated peptide, linear interpolation substantiated by direct implicit simulation reveals that a value of α^*_{prot} = 0.87 accurately matches the explicit solvent dimerization free energy. For the deprotonated peptide, we determine an optimal value of α^*_{deprot} = 0.72.

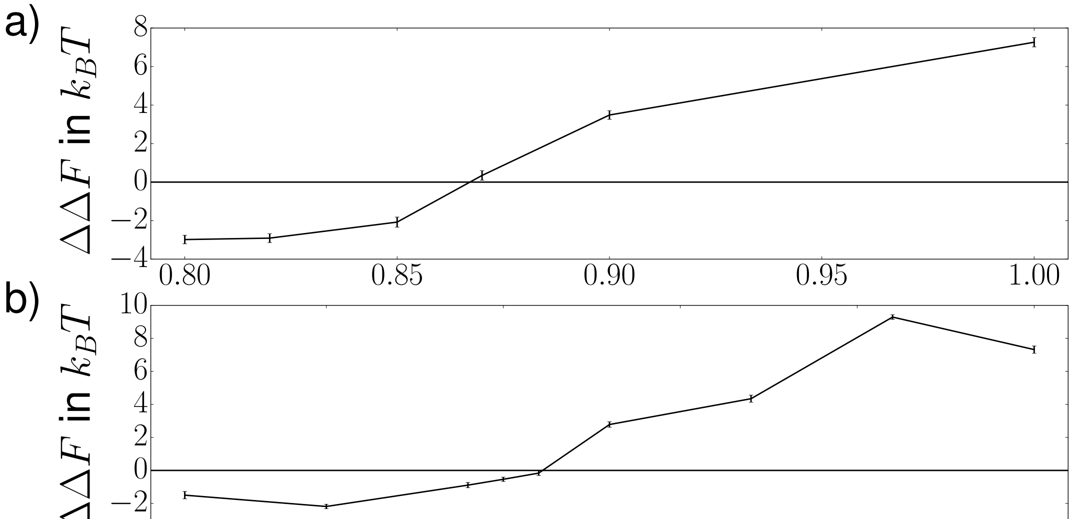

Figure 2.4: $\Delta\Delta F_{dimer}(\alpha) = |\Delta F^{\text{imp}}_{dimer}(\alpha) - \Delta F^{\text{exp}}_{dimer}|$ for (a) protonated peptides at values of the scaling parameter α = {0.80, 0.82, 0.85, 0.87, 0.90, 1.00}, and (b) deprotonated peptides at α = {0.52, 0.60, 0.68, 0.70, 0.72, 0.76, 0.84, 0.92, 1.00}.

Despite our choice of a simple single-parameter approach to rescale the non-bonded interactions in the CHARMM27 forcefield to account for the absence of molecular water, we achieved excellent agreement of the implicit solvent PMFs for peptide dimerization with those computed in explicit solvent. The explicit and rescaled implicit dimer PMFs for the protonated peptides are illustrated in Figure 2.5a, from which it is apparent that the rescaling of the non-bonded interactions has achieved quantitative agreement of PMF(r_{COM}) and ΔF_{dimer}. The analogous plots for the deprotonated peptides in Figure 2.5b show less good agreement, and while we have quantitatively matched the depth of the dimerization PMF, the minimum of the free energy well is shifted by (−0.10) nm from 0.74 nm in explicit solvent to 0.64 nm in implicit. We note that it is the protonated peptides that assemble into the 1D ribbons and which are therefore of primary interest in this chapter, so the slightly poorer quality of the deprotonated reparameterization is not of undue concern. The PMFs for peptide collapse (Figure 2.2) are governed principally by intramolecular interactions, and are minimally perturbed by the rescaling procedure (data not shown).

Colloquially, our results indicate that the standard (unscaled) CHARMM27 force field in implicit solvent makes the peptides too “sticky”, resulting in an artificial enhancement in the free energy of dimerization. Accordingly, it is vital that we perform rescaling of the non-bonded interactions to eliminate these artifacts as much as possible and lend credence to the predicted behavior of monomers, dimers, and higher order aggregates in our implicit solvent simulations.

2.3.4 Thermodynamics and morphology of higher order aggregation

Our reparameterized implicit solvent model permits us to efficiently perform umbrella simulations to compute the free energy of association for the formation of higher order aggregates and probe the elementary events at the early stages of assembly.

Protonated peptides

Considering first protonated (low-pH) peptides, we computed the association free energies for the formation of a pentamer along two different paths (i) successive monomeric addition, and (ii) dimer condensation followed by monomeric addition. We illustrate the two pathways and report the computed free energy changes in Figure 2.6. Uncertainties are computed by standard propagation of errors over five independent runs accounting for the variance between runs and the bootstrap uncertainties within each run. As illustrated in Figure 2.5, the dimerization free energy is highly favorable at $\Delta F_{1+1} = -(14.5 \pm 1.2)$ $k_B T$. The successive addition of two more monomers is also strongly spontaneous at $\Delta F_{2+1} = -(24.5 \pm 2.4)$ $k_B T$ and $\Delta F_{3+1} = -(23.6 \pm 3.0)$ $k_B T$, with the larger magnitude value reflecting favorable interactions between the approaching monomer and the multiple monomers in the cluster. Condensation of two dimers into a tetramer is also strongly favorable with $\Delta F_{2+2} = -(27.6 \pm 3.7)$ $k_B T$. The free energy for tetramer for-

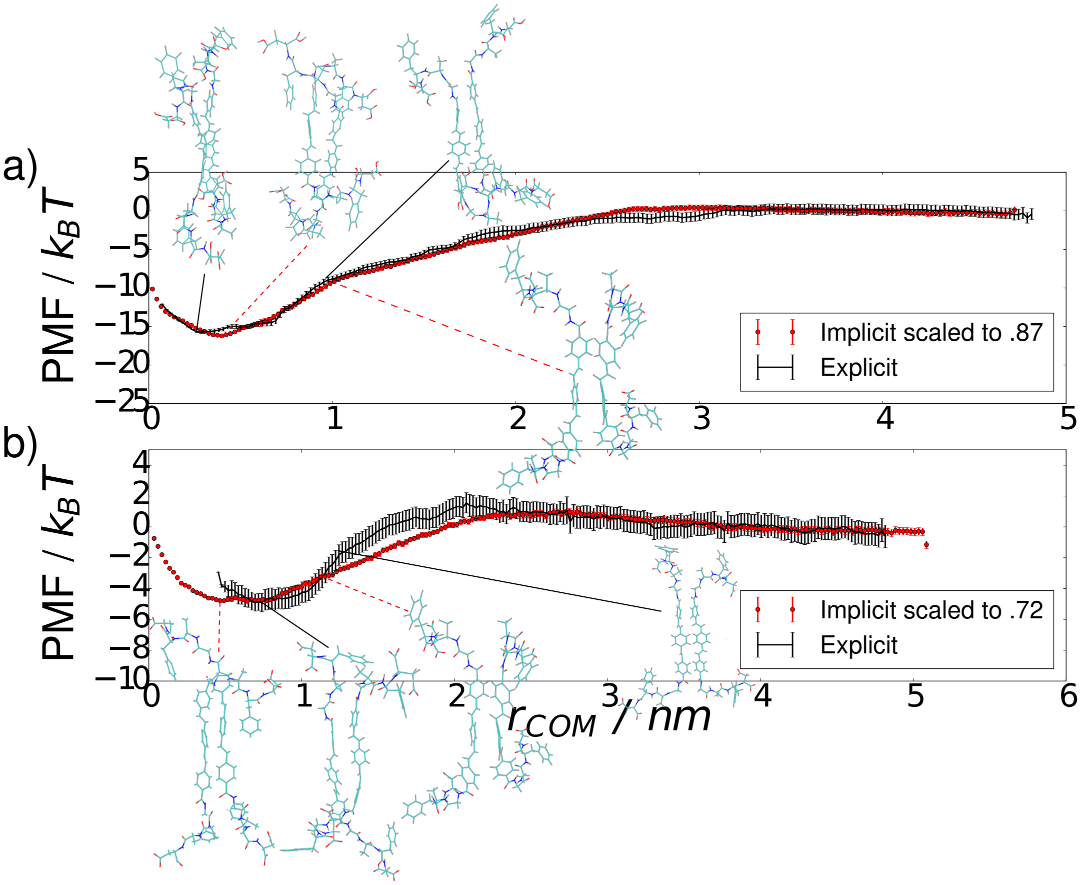

Figure 2.5: PMF curves for pairs of DFAG-OPV3-GAFD peptides in explicit and implicit solvent at the optimal value of (a) $\alpha^*_{\text{prot}} = 0.87$ in the protonated (low-pH) and (b) $\alpha^*_{\text{deprot}} = 0.72$ in the deprotonated (high-pH) states parameterized by peptide center of mass separation, r_{COM}. The dimerization free energy predicted by the rescaled implicit solvent model for the protonated (low-pH) peptides is strongly favorable at $\Delta F^{\text{prot}}_{dimer} = -(14.5 \pm 1.2)\ k_BT$, whereas that for the deprotonated (high-pH) peptides is only weakly so at $\Delta F^{\text{deprot}}_{dimer} = -(4.0 \pm 0.7)\ k_BT$. Both values are in excellent agreement with the explicit solvent predictions of $\Delta F^{\text{prot}}_{dimer} = -(15.7 \pm 0.7)\ k_BT$ and $\Delta F^{\text{deprot}}_{dimer} = -(4.7 \pm 0.9)\ k_BT$.

mation by successive monomeric addition is $\Delta F_{tetra} = (\Delta F_{1+1} + \Delta F_{2+1} + \Delta F_{3+1}) = -(62.5 \pm 4.1)\ k_BT$, and by dimer condensation is $\Delta F_{tetra} = (2 \times \Delta F_{1+1} + \Delta F_{2+2}) = -(56.5 \pm 4.1)\ k_BT$, where error bars are computed by standard propagation of uncertainties. Since free energy is a pathway independent state variable, it serves as an important internal validation of our methodology that the free energy for the assembly of a tetramer by these two paths is in agreement within error bars. The free energy to form a pentamer by monomeric addition to a tetramer is $\Delta F_{4+1} = -(25.0 \pm 1.9)\ k_BT$.

Representative images of the equilibrium structure of each oligomer are also presented in Figure 2.6, from which it is apparent that the preferred structural morphology of each oligomer is a β-sheet-like stack of monomers. Our umbrella sampling simulations constrained only the center of mass distance between the peptide clusters. That they spontaneously adopted this structure provides evidence that the early stages of self-assembly leads to the formation of short peptide stacks that serve as nuclei for subsequent monomeric addition and elongation into the experimentally observed 1D ribbons [4, 23, 65].

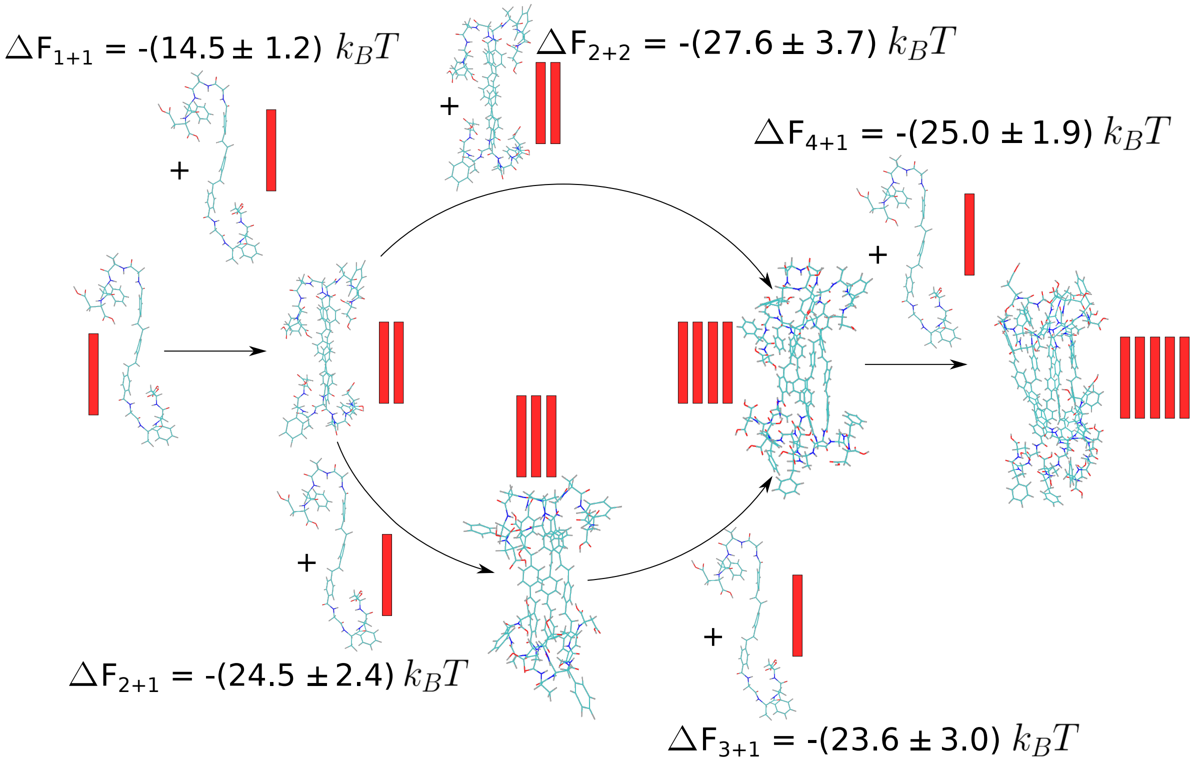
$\triangle F_{1+1}$ = -(14.5 ± 1.2) k_BT
$\triangle F_{2+2}$ = -(27.6 ± 3.7) k_BT
$\triangle F_{4+1}$ = -(25.0 ± 1.9) k_BT
+
+
+
+
+
$\triangle F_{2+1}$ = -(24.5 ± 2.4) k_BT
$\triangle F_{3+1}$ = -(23.6 ± 3.0) k_BT

Figure 2.6: Elementary free energy changes for pentamer formation by monomeric addition (lower path) and dimer condensation plus monomeric addition (upper path) for protonated (low-pH) peptides computed by umbrella sampling using the reparameterized implicit solvent model. Uncertainties are computed by standard propagation of errors over five independent runs accounting for the variance between runs and the bootstrap uncertainties within each run. The free energy change for formation of the tetramer from the dimer along the upper and lower paths are identical within error bars. For ease of comprehension, representative molecular structures of the equilibrium clusters are presented with cartoon diagrams in which each red oblong represents a peptide monomer.

In Figure 2.7 we plot the free energy for monomer addition to small aggregates of protonated peptides under low-pH conditions. It is expected that the free energy for the addition of monomer to a stack of n peptides should become increasingly favorable with increasing n, since there exist more atoms within the cluster to provide favorable dispersion interactions with the approaching monomer. At sufficiently large n, the free energy should saturate since the approaching monomer will not significantly interact with peptides located far from the end of the stack. Figure 2.7 shows the monomeric addition free energy plateauing surprisingly quickly at the trimer, with $\Delta F_{2+1} \approx \Delta F_{3+1} \approx \Delta F_{4+1} \approx -25 k_B T$ indistinguishable within error bars. This indicates that the free energy of monomeric addition in the early stages of assembly is both strongly favorable and essentially independent of oligomer size.

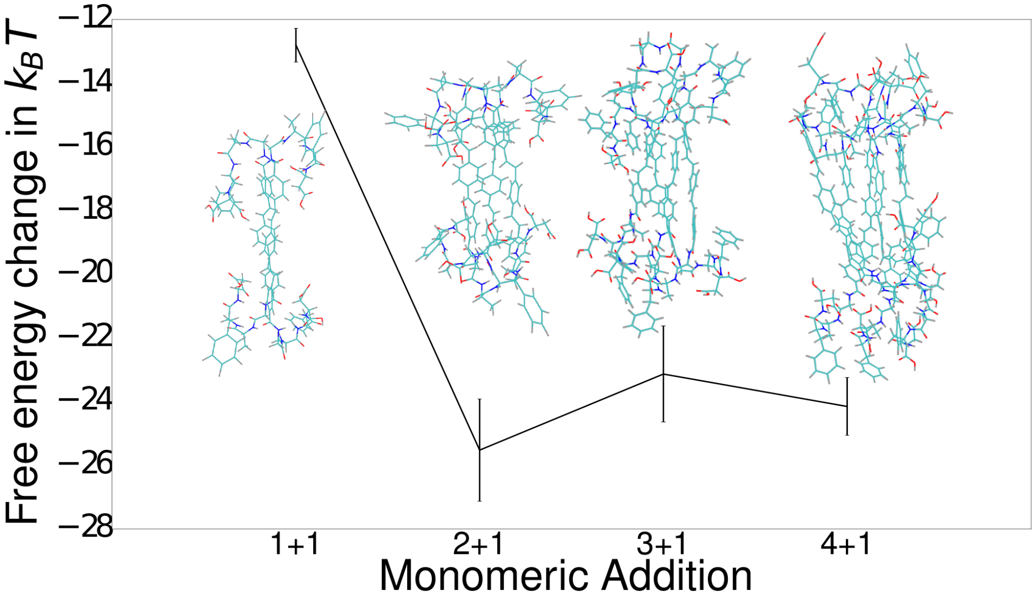

Figure 2.7: Free energy of monomeric addition to preassembled oligomers containing between one and four peptides under protonated (low-pH) conditions computed by umbrella sampling. Uncertainties are computed by standard propagation of errors over five independent runs accounting for the variance between runs and the bootstrap uncertainties within each run. The calculated free energy change for the addition of a monomer to a dimer, trimer, or tetramer are indistinguishable within error bars, suggesting that the free energy for monomer addition to the elongating β-sheet-like stack is strongly favorable and essentially independent of the size of the stack at $\Delta F_{n+1} \approx -25 k_B T$.

We can decompose the free energy of assembly of a peptide aggregate of size n from initially non-interacting peptide monomers into its various energetic and entropic contributions,

$$\Delta F_n = \Delta U_n^{\text{intrapeptide}} + \Delta U_n^{\text{peptide-peptide}} + \Delta U_n^{\text{peptide-water}} + \Delta U_n^{\text{water-water}} - T\Delta S_n + P\Delta V \tag{2.6}$$

where $\Delta U_n^{\text{intrapeptide}}$ is the change in the intramolecular peptide energy upon aggregation, $\Delta U_n^{\text{peptide-peptide}}$ accounts for the intermolecular dispersion and electrostatic interactions between peptides in the cluster, $\Delta U_N^{\text{peptide-water}}$ is the change in the peptide-water interaction energy upon association, T is the temperature, P the pressure, ΔS_n the entropy change upon association, and ΔV the volume change. The value for ΔF_n for n = 1-5 is straightforwardly computed by summing the monomeric addition free energies reported in Figures 2.6 and 2.7. By performing an energetic analysis of our equilibrated peptide monomers, dimers, trimers, tetramers, and pentamers residing at the global free energy minimum of the potential of mean force curves determined by umbrella sampling, we explicitly compute the values of $\Delta U_n^{\text{intrapeptide}}$

and $\Delta U_n^{\text{peptide-peptide}}$. Assuming that the peptide configurational entropy does not change significantly from the monomeric to oligomeric state, and neglecting pressure-volume work as a small contribution at standard temperature and pressure for such small aggregates [106] we define the solvent mediated contribution to the aggregation free energy as,

$$\Delta F_n^{\text{solvent}} = \Delta U_n^{\text{peptide-water}} + \Delta U_n^{\text{water-water}} - T\Delta S_n^{\text{water}} \approx \Delta F_n - \Delta U_n^{\text{intrapeptide}} - \Delta U_n^{\text{peptide-peptide}} \tag{2.7}$$

We plot in Figure 2.8a the free energy of aggregation for oligomers of size n = 1-5, and its decomposition into $\Delta U_n^{\text{intrapeptide}}$, $\Delta U_n^{\text{peptide-peptide}}$, and $\Delta F_n^{\text{solvent}}$. The plot reveals that the moderately favorable free energy of aggregation is the result of large competing energetic and entropic contributions. As anticipated, there is a large favorable peptide-peptide interaction on the order of ∼-100 k_BT per peptide that is balanced by a small unfavorable intrapeptide energy change of ∼10 k_BT per peptide. That the latter contribution is relatively small is consistent with the assumption that the configurational ensemble explored by the peptide is relatively similar in the monomeric and oligomeric states, and a small change in peptide entropy upon assembly. This contribution does grow with aggregate size, suggesting that peptides adopt increasingly unfavorable configurations in larger clusters. The solvent mediates a net large and unfavorable contribution to assembly on the order of ∼80 k_BT per peptide. This contribution accounts for the unfavorable free energy change associated with cavity formation and disruption of the hydrogen bonding network, loss of peptide-water interactions due to formation of the oligomeric aggregate, and the entropy change induced in the water due to oligomerization [106–108].

To unravel the relative contributions of the conjugated peptide core and peptide wings to the peptide-peptide interaction energy, we have further decomposed this term as,

$$\Delta U_n^{\text{peptide-peptide}} = \Delta U_n^{\text{wing-wing}} + \Delta U_n^{\text{core-core}} + \Delta U_n^{\text{wing-core}} \tag{2.8}$$

where $\Delta U_n^{\text{wing-wing}}$ accounts for intermolecular dispersion and electrostatic interactions between the amino acid wings of all of the peptides in the aggregate of size n, $\Delta U_n^{\text{core-core}}$ is the interaction between the conjugated cores, and $\Delta U_n^{\text{wing-core}}$ accounts for the cross interactions between the wings and cores. We plot this decomposition in Figure 2.8b. Interactions between the aromatic cores account for approximately 20% of the peptide-peptide interaction energy for all aggregate sizes considered. In small oligomers of size n = 2-3, the wing-core interactions dominate the intermolecular interactions, which can be understood from the structure of the dimers and trimers wherein the peptide wings double back on themselves to interact with the conjugated cores (cf. Figure 2.6).

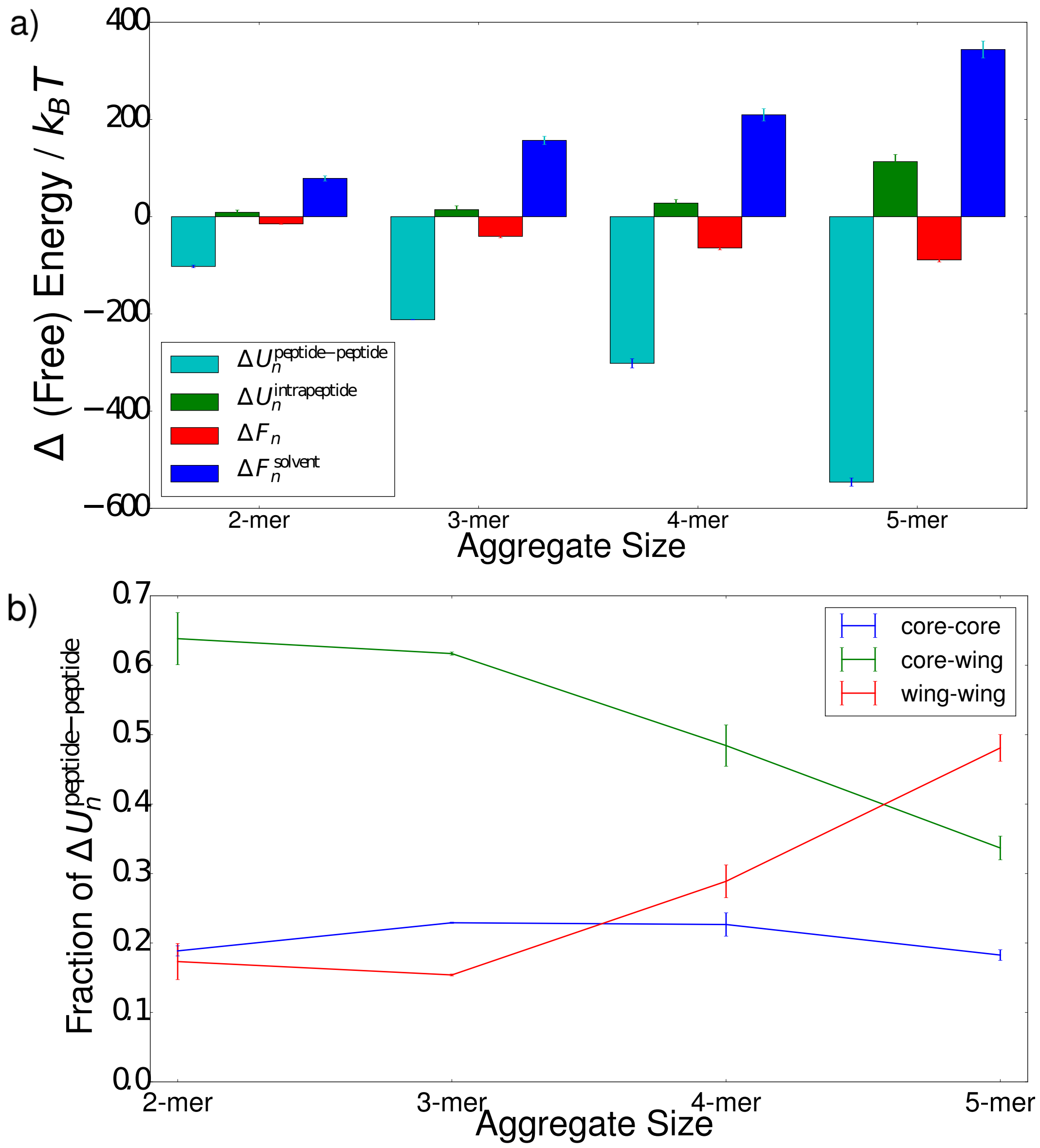

Figure 2.8: Decomposition of the energetic and entropic contributions to the free energy of association as a function of aggregate size for oligomers containing between two and five peptides. (a) The free energy of assembly of an aggregate of size n from non-interacting monomers ΔF_n and its decomposition into the change in the intramolecular, $\Delta U_n^{\mathrm{intrapeptide}}$, and intermolecular, $\Delta U_n^{\mathrm{peptide-peptide}}$, peptide energies, and the solvent mediated contribution to aggregation, $\Delta F_n^{\mathrm{solvent}} \approx \Delta U_n^{\mathrm{peptide-water}} + \Delta U_n^{\mathrm{water-water}} - T\Delta S_n^{\mathrm{water}}$, neglecting pressure-volume work and the entropy change of the peptides upon assembly. Intermolecular terms comprise dispersion and electrostatic interactions, while intramolecular terms comprise dispersion, electrostatic, and bonded interactions (b) Partitioning of $\Delta U_n^{\mathrm{peptide-peptide}}$ into interactions between the peptide wings, aromatic cores, and cross interactions between the wings and cores. Error bars are computed by the standard deviation of the energy computed over the course of the 10 ns simulation and standard propagation of uncertainties.

Finally, we characterize the structure of the peptides within oligomers of various sizes by analyzing the equilibrated peptide stacks containing n = 1-5 peptides in simulations conducted at the minimum of the assembly PMF determined by umbrella sampling. In Figure2.9a,b we report the mean center of mass spacing and angular offset between all pairs of nearest neighbors in the oligomeric aggregates. Consistent with prior work [4, 56], we find the spacing between neighboring peptides to be ~0.45 nm, and insensitive to aggregate size. The angular offset between neighboring peptide cores is determined by associating to each peptide a vector between the terminal carbon atoms of the OPV3 core, and computing the dot product between these vectors. Since the core is relatively rigid and linear (cf. Section 2.3.1), these vectors provide a good representation of the spatial orientation of each peptide in the stack. We observe that neighboring peptides in these small oligomers remain approximately parallel for all aggregate sizes studied, possessing a mean angular offset of $(10.9 \pm 7.0)^{\circ}$. These structural characterizations indicate that the π-conjugated cores of the peptides are closely stacked in well-aligned stacks. In Figure 2.9c we report the number of intra and intermolecular hydrogen bonds per peptide as a function of aggregate size. A hydrogen bond is defined between an eligible donor and acceptor according to a geometric criterion such that the donor-acceptor distance is less than 0.35 nm and the hydrogen atom lies not more than 30° offset from the vector connecting the donor and acceptor [109, 110]. We detect hydrogen bonds using the g_hbond program within the GROMACS 4.6 simulation suite [70]. The number of hydrogen bonds formed by each peptide in an aggregate is relatively insensitive to oligomer size, with each peptide forming approximately 1 and 3 intra and intermolecular hydrogen bonds, respectively. These results indicate the importance of inter-peptide hydrogen bonding in stabilizing the assembled oligomers. In sum, this structural characterization of the oligomers reveals that efficient stacking and alignment between peptides engenders favorable dispersion interactions and π-stacking between the aromatic cores and hydrogen bonding between the peptide wings, promoting self-assembly into β-sheet-like aggregates (cf. Figure 2.8).

Deprotonated peptides

Shifting our focus now to deprotonated peptides under high-pH conditions, we compute the free energy for the formation of a pentamer, this time only along the monomeric addition pathway. As shown in Figure 2.10 the formation of dimers is slightly favorable with $\Delta F_{1+1} = -(4.0 \pm 0.7)\ k_B T$. Surprisingly, further monomeric addition continues to be favorable with $\Delta F_{2+1} = -(6.3 \pm 0.9)\ k_B T$, $\Delta F_{3+1} = -(6.0 \pm 0.8)\ k_B T$, and $\Delta F_{4+1} = -(4.6 \pm 1.2)\ k_B T$. Despite the Coulombic repulsion between doubly negatively charged ASP termini, favorable dispersion interactions, hydrophobicity, and π-stacking between the aromatic peptide cores is sufficient to drive the assembly of small oligomers. The morphology of the oligomers is less or-

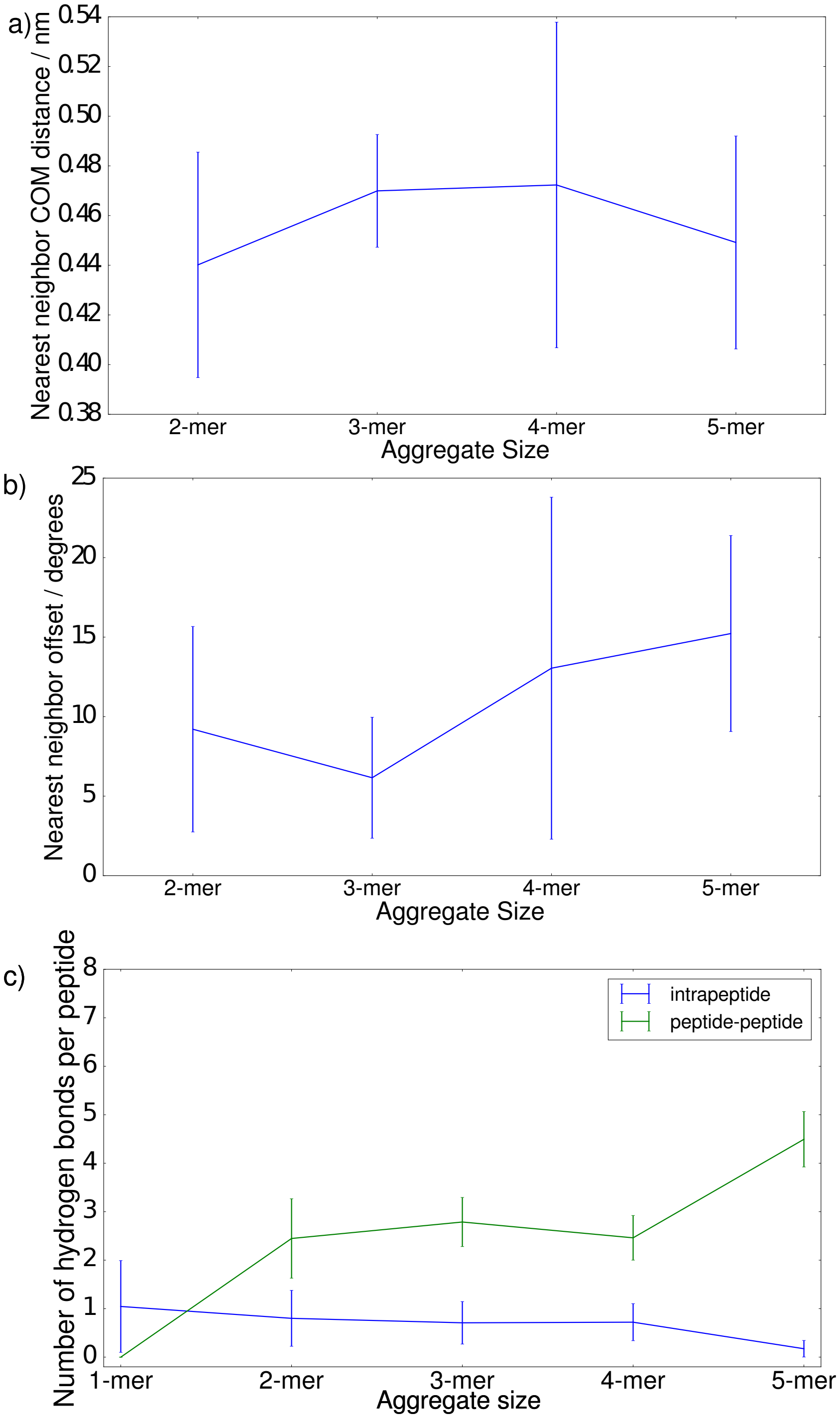

Figure 2.9: Structural characterization of oligomeric aggregates containing between two and five peptides. (a) Mean center of mass separation between nearest neighbor peptides in the aggregate. (b) Mean angular offset between nearest neighbor peptides computed by calculating the dot product between vectors oriented along the OPV3 cores of the peptides. (c) Mean number of intra and intermolecular hydrogen bonds per peptide. Error bars represent the standard deviation computed over the course of the 10 ns simulation.

dered than under low-pH conditions, with the peptide wings orienting away from one another to mitigate the unfavorable Coulombic repulsion while the cores associate into a weakly bound stack. The increasing concentration of negative charge on the growing aggregate makes the addition of subsequent monomers beyond the trimer successively less favorable and the aggregates more structurally disordered, suggesting that assembly will be ultimately self-limiting.

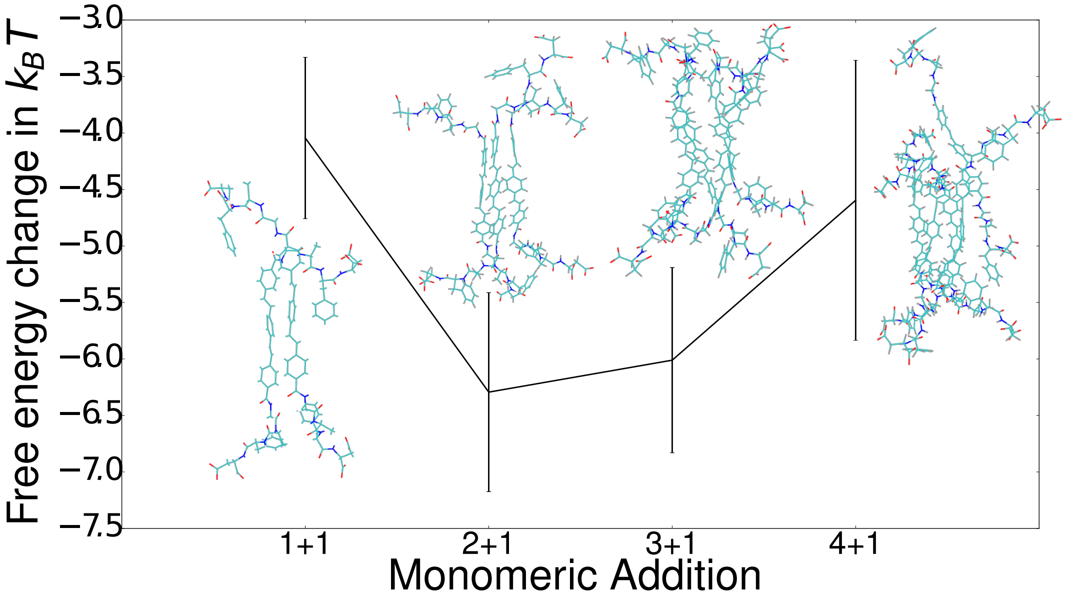

Figure 2.10: Free energy of monomeric addition to preassembled oligomers containing between one and four peptides under deprotonated (high-pH) conditions computed by umbrella sampling. Uncertainties are computed by standard propagation of errors over five independent runs accounting for the variance between runs and the bootstrap uncertainties within each run. Monomer addition is thermodynamically favored by dispersion interactions, hydrophobicity, and π-stacking between the aromatic cores, but significantly destabilized relative to the protonated (low-pH) conditions by unfavorable Coulombic repulsions between the doubly negatively charged ASP termini. The increasing concentration of negative charge on the growing oligomer makes monomer association increasingly unfavorable beyond the trimer.

2.3.5 Time scale correspondence in explicit and implicit solvent simulations

Compared to explicit solvent, the absence of molecular solvent in implicit solvent simulations may result in artificially accelerated dynamics [111]. In order to ascertain the time scale of the implicit solvent runs, we follow an approach commonly used to define the time scale in coarse-grained simulations by comparing the translational self-diffusion coefficient computed in the implicit calculations to that predicted from an explicit solvent calculation, which, possessing all atomic degrees of freedom, evolves with real-time dynamics [112–114]. We performed five independent 20 ns simulations of a single isolated peptide in implicit solvent using the parameters given in Section 2.2.2 then calculated for each run the self-diffusion coefficient by tracking the mean squared displacement of the peptide center of mass and applying the Einstein relation [115]. We take the mean of these five values as our estimate of the self-diffusivity, and report uncertainties in our estimate as the standard deviation over the estimates from the five independent runs, to determine a value of $D_{\mathrm{imp}} = (6.6 \pm 0.2) \times 10^{-5}$ cm^2/s. We compare this value to that of $D_{\mathrm{exp}} = (7.1 \pm 0.6) \times 10^{-5}$ cm^2/s, computed in explicit solvent from block averaging of a 10 ns simulation conducted using the parameters given in Section 2.2.1, with the exception that we instead of a Langevin thermostat we implement a Nosé-Hoover thermostat with a time constant of 0.5 ps, a value commonly used in the literature [116, 117], that has been shown to reliably approximate the true dynamical time scales of the atomic system [118]. The agreement of the implicit and explicit self-diffusion coefficients within error bars indicates that the Langevin thermostat implemented within our implicit runs with a time constant of 2 ps^{-1} accurately mimics the dynamics of random collisions of solvent molecules with our peptide [119–121] resulting in dynamical time scales in good agreement with explicit solvent simulations.

2.3.6 Aggregation and structural evolution in early-stage assembly

Our reparameterized implicit solvent model permits simulation of the self-assembly of hundreds of protonated peptides over the time scales of ~100 ns to probe in molecular detail the microscopic events in the early stages of self-assembly [4, 23, 56]. Using the simulation parameters detailed in Section 2.2.2 and rescaling time by the speed-up factor of 7.2 identified in Section 2.3.5, we performed 70 ns simulations of 64 and 125 protonated (low-pH) peptides in a 50×50×50 nm cubic simulation box to simulate peptide aggregation at concentrations of 0.85 mM and 1.66 mM. We selected these concentrations as experimentally realizable values at which self-assembly of π-conjugated peptide monomers has been observed [23]. (For reference, a concentration of 0.868 mM corresponds to 0.1 mg mL^{-1}.) We conducted three independent simulations at each concentration commencing from an initial monomeric dispersion in which the peptides were evenly spaced over a 3D grid and assigned random initial velocities drawn from a Maxwell distribution at 298 K. Each simulation required approximately 40 years of CPU time on high-performance parallel computing facilities. By rendering accessible time and length scales 2-3 orders of magnitude larger than those attainable by explicit solvent models, our implicit solvent model permits direct observation of peptide collision, dissociation, and structural relaxation events over the first several tens of nanoseconds of assembly.

We quantify the aggregation behavior by monitoring the formation of peptide clusters over the course of the simulation. A cluster is defined as a set of contiguously connected peptides, where a pair of peptides is deemed connected if any of their respective atoms are separated by less than $r_{\text{cut}} = 0.5$ nm. This cutoff was selected as a value close to the minimum of the Lennard-Jones potential for larger atoms, which is expected to be sufficiently small to mitigate the chance of false positives, while sufficiently large that peptides forming are part of the same aggregate will be recognized as such. Indeed, the location of the free energy minimum in the dimerization PMF lies very close to 0.5 nm (cf. Figure 2.5).

Low concentration

We present in Figure 2.11 the time evolution of the cluster size distribution observed in each of the three independent simulations at 0.85 mM concentration. In all three replicas we observe very similar trends in the mass fraction of monomers, dimers, and trimers. Tetramers also exhibited similar trends, but the time taken to form the first tetramer varies by 30 ns between runs. A single pentamer was observed at the end of run 1, but beyond this no higher order aggregates were observed. This suggests that our simulations are insufficiently large to furnish statistically robust data on clusters larger than trimers since the formation of such large aggregates over the course of a 70 ns simulation containing only 64 monomers is a very rare event. Accumulating better statistics on larger aggregates at this concentration would require that we simulate several times as many monomers in a correspondingly enlarged simulation box for a longer duration.

The good agreement in the trends for the smaller clusters permits us to draw statistical conclusions regarding their aggregation behavior. After an initial lag time of ~7 ns, the mass fraction of monomers monotonically decreases over the course of the simulation, falling to ~20% by the end of the 70 ns run. Recalling that the initial state of the system comprised of monomers evenly spaced over a 3D grid, this lag time corresponds to the characteristic time for monomer-monomer collisions commencing from this initial configuration. After the ~7 ns lag the mass fraction of dimers begins to grow from zero, increasing monotonically for the first 55 ns to ~50% but then decreasing to ~40% by the end of the simulation. We do observe a small number of trimer dissociation events, but the trimer mass fraction also grows nearly monotonically beginning at ~13 ns to occupy ~15% of the system mass. Aggregation is observed to proceed both through monomeric addition and the condensation of larger aggregates. Of the eight tetramers that formed, three assembled through dimer-dimer condensation (cf. Table 2.1).

In Figure 2.12 we present selected snapshots over the course of a trimerization event observed in one of the 0.85 mM simulations. We observe that smaller clusters of 2-4 peptides tend to relax to a well-aligned β-sheet-like stack on the time scale of <10 ns, regardless of the configuration in which the peptides initially associate.

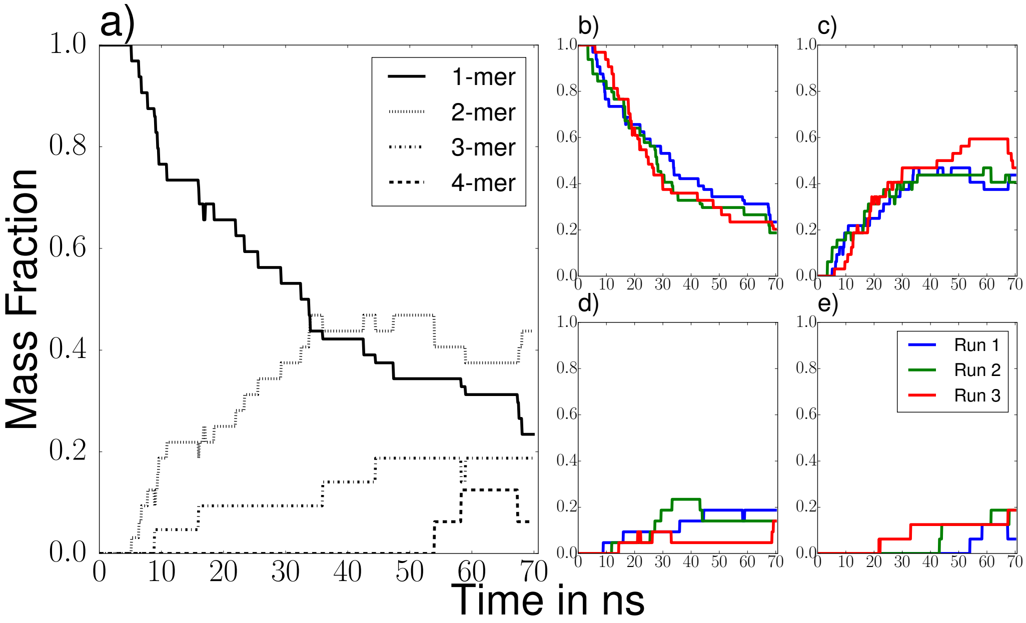

Figure 2.11: Mass fraction as a function of scaled simulation time for an unbiased simulation of 64 peptides at 0.85 mM for (a) one simulation, and (b-e) each of the three independent simulations tracking (b) 1-mers, (c) 2-mers, (d) 3-mers, and (e) 4-mers.

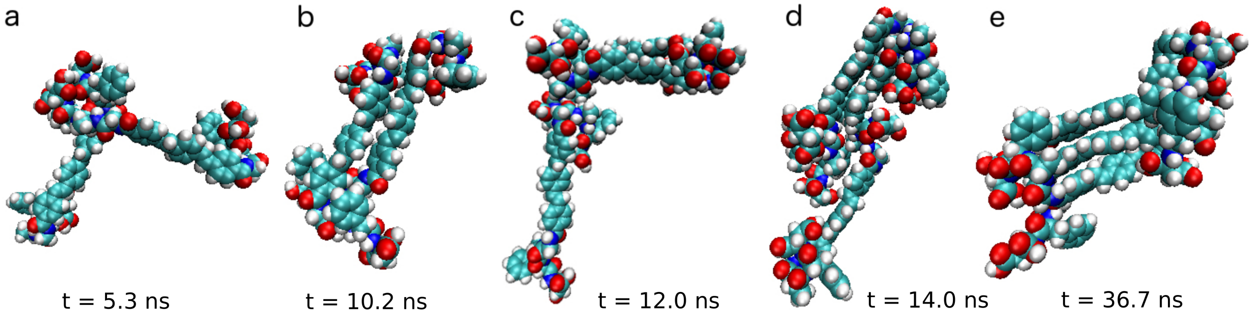

Figure 2.12: Representative snapshots of a trimerization event observed in Run 2 of the 0.85 mM concentration protonated (low-pH) peptide self-assembly simulations conducted over 70 ns using our reparameterized implicit solvent model. (a) Collision of two monomers at t = 5.3 ns induces rise to internal structural rearrangements until (b) the two conjugated cores are aligned to form the equilibrium dimer (cf. Figure 2.7) after ~5 ns. (c) Collision of the dimer with a third monomer at t = 12 ns results in (d) further structural reorganization and relaxation over 2 ns undergoes minor rearrangements to form (e) an equilibrium trimer β-sheet-like stack (shown here at t = 36.7 ns) (cf. Figure 2.7).

High concentration

We present in Figures 2.13 and 2.14 plots for the cluster size distribution for the 1.66 mM runs. Under these conditions, the largest cluster observed was an 11-mer, with the increased number of peptides and the elevated concentration favoring the more rapid formation of heavier aggregates. We observe good correspondence between the three independent simulations in the dynamical evolution of the mass fraction of monomers, dimers, and trimers, and tetramers Figure 2.13, but relatively poorer agreement in the case of the higher order aggregates for which far fewer assembly or disassembly events occur Figure 2.14. Again, larger and longer simulations would be required to furnish statistically robust data on the formation and dissociation of larger aggregates.

At this concentration we observe more interesting assembly behaviors than at the lower concentration. The mass fraction of monomers again falls essentially monotonically, but in this case drops to a terminal value of only ~10%. Dimer assembly commences much more rapidly after a lag period of only ~1 ns, leading to a peak in the dimer mass fraction of ~40% at ~15 ns and a subsequent drop off to ~20% by the end of the simulation as the dimers contribute to the assembly of heavier aggregates. Trimer assembly commences at ~5 ns and increases essentially monotonically to plateau at a terminal mass fraction of ~(15-35)%. Tetramer assembly commences at ~15 ns, reaching a mass fraction of ~(10-15)% by the end of the run. Finally, we observe the assembly of small numbers of pentamers, hexamers, heptamers, octamers, decamers, and undecamers, but no nonamers. Assembly proceeds not only by monomeric addition, but the agglomeration of higher order aggregates such as dimer-trimer and trimer-trimer association events (cf. Figure 2.15 and Table 2.2). We also observed dissociation of all cluster sizes into smaller aggregates, although only one such event was observed for the rarely observed heptamers and octamers.

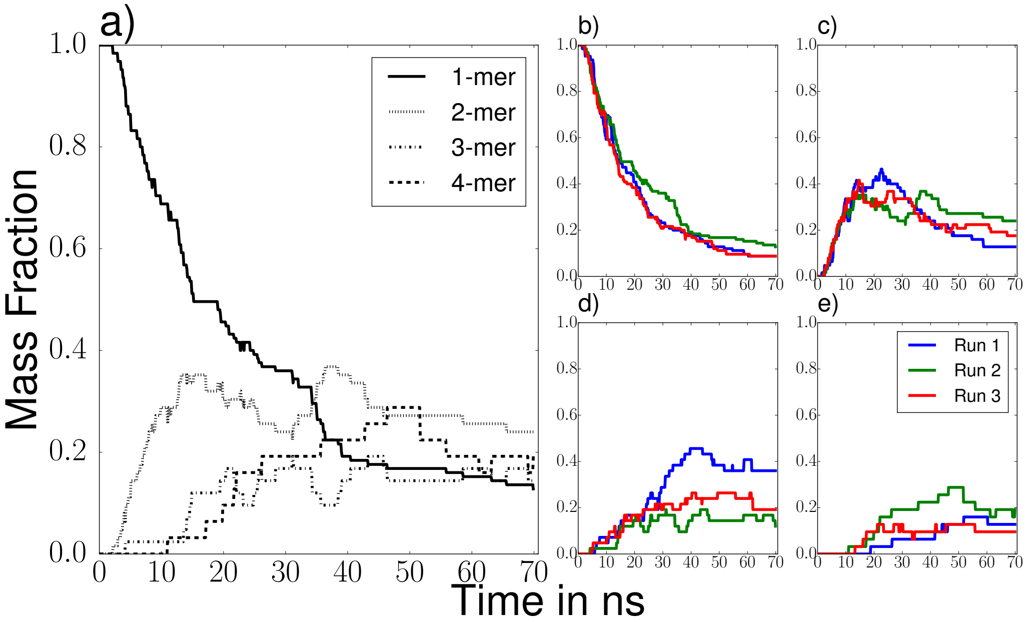

Figure 2.13: Mass fraction as a function of scaled simulation time for an unbiased simulation of 125 peptides at 1.66 mM for (a) one simulation, and (b-e) each of the three independent simulations tracking (b) 1-mers, (c) 2-mers, (d) 3-mers, and (e) 4-mers.

In Figure 2.15 we show snapshots of a sequence of aggregation events leading to the formation of a

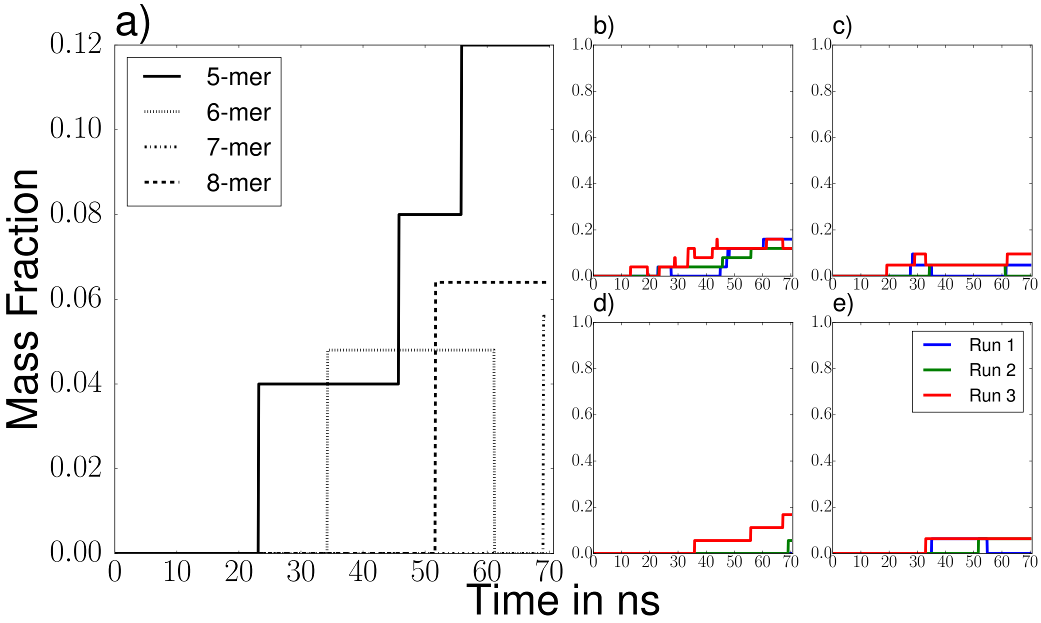

Figure 2.14: Mass fraction as a function of scaled simulation time for an unbiased simulation of 125 peptides at 1.66 mM for (a) one simulation, and (b-e) each of the three independent simulations tracking (b) 5-mers, (c) 6-mers, (d) 7-mers, and (e) 8-mers.

heptamer. As was observed in the 0.85 mM simulations, smaller aggregates of n=2-4 monomers tend to structurally relax into well-aligned β-sheet-like stacks, but those containing more $n > 4$ peptides tend to lack this structural ordering. This finding suggests that larger aggregates are kinetically trapped into disordered configurations on the time scale of our simulations, and the experimentally observed organization into 1D ribbons exhibiting π-stacking of the conjugated cores occurs on the time scale of hundreds of ns or longer [4, 23, 65, 66].

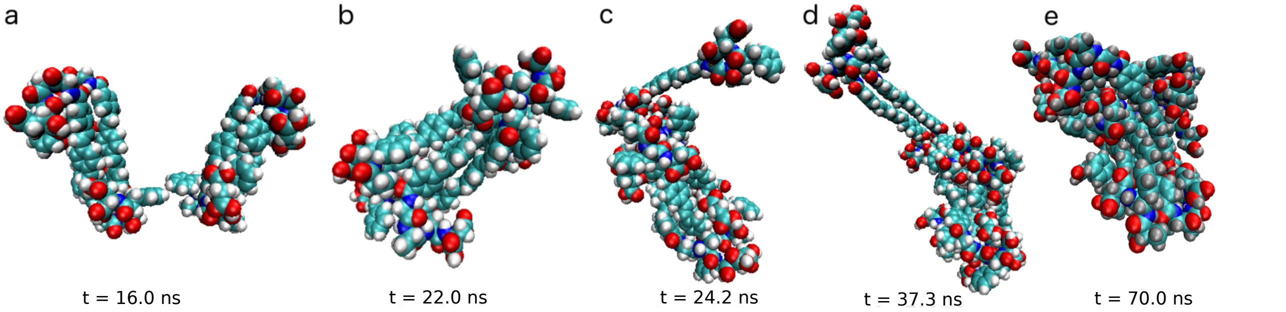

Figure 2.15: Representative snapshots of a heptamerization event observed in Run 3 of the 1.66 mM concentration protonated (low-pH) peptide self-assembly simulations conducted over 70 ns using our reparameterized implicit solvent model. (a) Collision of two pre-assembled dimers at t = 16 ns produces (b) a tetramer β-sheet after 4 ns that (c) interacts with a fifth monomer to produce a pentamer. (d) The pentamer collides with a dimer to generate (e) a disordered heptamer that has not fully undergone structural rearrangement into an ordered β-sheet-like stack by the termination of the simulation at t = 70 ns.

2.3.7 Kinetics of early-stage self-assembly

Given the time evolution of the cluster size distribution for the protonated (low-pH) peptides, we can gain insight into the microscopic assembly kinetics by extracting from these data effective rate constants for transitions between different cluster sizes during the early stages of assembly. We posit that we may model the aggregation process as a time homogeneous (i.e. stationary) continuous time Markov chain (CTMC) between different cluster sizes [122–129]. This modeling approach assumes that the rate at which an aggregate of a particular size transitions into an aggregate of a different size – when observed on sufficiently long time scales [124, 125, 130] – can be approximated as spatially invariant, time invariant, and memoryless. In other words, the transition rates do not depend upon the location of the aggregate or the time at which it is observed, and neither the remaining time that the system will spend in the current state nor the state to which it will transition next depend on its history. Since we initialize our system from a homogeneous mixture of monomers over a 3D grid, there should not – at least over relatively short time scales – exist large concentration gradients in the system, suggesting that we need not spatially resolve the system and may treat it as well-mixed. That time homogeneity should *a priori* be satisfied is less clear, since the rate at which one aggregate transforms into another certainly depends on the aggregation state of its neighbors with which it can collide. We provide below *a posteriori* validation that the rate constants extracted from our 70 ns simulations are not a function of the observation time. We select the lag time used to estimate the transition rates from our simulation trajectories as that which best reproduces the observed time evolution of the cluster size distribution, with good agreement between the predicted and observed distributions validating the memoryless assumption. Furthermore, we will demonstrate that this lag time exceeds the Markov time, verifying that the system possesses the Markov (i.e., memoryless) property on this time scale [124].

We define the states of our Markov process as aggregates of different sizes. This necessarily discards any information regarding the internal structure of the clusters, but we anticipate that omitting these details will – at least for the relatively small cluster sizes observed during early-stage assembly – admit a Markovian description provided that the internal cluster organization does not significantly affect the transition rates between different aggregate sizes or that clusters are typically able to relax into β-sheet-like configurations prior to undergoing further transitions. We show below that the transition rates for aggregates of a particular size are consistent over the multiple aggregates with different internal architectures observed in three independent simulations. The instantaneous state of the system as at a time instant t is represented as a probability row vector of length n holding the mass fraction of the system existing as clusters of size 1 to n, $\mathbf{p}(t)$, where n is the largest aggregate observed in any of the three independent simulations at the concentration of interest. By the definition of the mass fraction, the i^{th} element of this vector defines the probability with which a peptide monomer exists in an aggregate of size i [131]. Employing the same definition as above, a cluster is defined as a set of contiguously connected peptides, where pair of peptides are defined as connected if any of their respective atoms are separated by less than $r_{\text{cut}} = 0.5$ nm.

We denote as $\mathbf{Q}$ the *transition rate matrix* – or infinitesimal generator matrix – whose off-diagonal elements q_{ij} are transition rates of monomers from aggregates of size i to aggregates of size j, and whose diagonal elements are the negative sum of transition rates out of state i, $q_{ii} = -\sum_{j=1}^{n} q_{ij}$ [122]. By performing a Taylor expansion in time, the probability distribution of monomers among the various aggregate sizes at time $(t+\Delta t)$ given that it was distributed as $\mathbf{p}(t)$ at time t, is given by $\mathbf{p}(t+\Delta t) \approx \mathbf{p}(t)(\mathbf{I}+\Delta t\mathbf{Q})$, where Δt is an infinitesimal time increment [122]. Defined as such, the monomer distribution among the various cluster sizes at time t, given an initial distribution $\mathbf{p}(0)$, may be formally integrated to yield [122, 123, 130],

$$\mathbf{p}(t) = \mathbf{p}(0)e^{\mathbf{Q}t} = \mathbf{p}(0)\mathbf{T}(t), \tag{2.9}$$

where $\mathbf{T}(t) = e^{\mathbf{Q}t}$ is the *transition matrix* whose elements $t_{ij}(t)$ are the probabilities that a monomer will be found in an aggregate of size j after an elapsed time t given that it initially resided in an aggregate of size i [123, 124, 128]. We estimate the matrix exponential using the scaling squaring method of Lawson [132, 133],

$$\mathbf{T}(t) = e^{\mathbf{Q}t} \approx [r_m(\mathbf{Q}t/2^s)]^{2^s}, \tag{2.10}$$

where $r_m(x)$ is the $[m/m]$ Padé approximation for e^x and s is chosen so that $||\mathbf{Q}t/2^s||_\infty \approx 1$. We perform this estimation using the algorithm developed by Al-Mohy and Higham [134] and implemented in SciPy (`http://www.scipy.org`). We note that because our simulations are out of thermodynamic equilibrium, we do not enforce detailed balance in the estimation of our matrix elements [124].

We extract maximum likelihood estimator of the off-diagonal elements of $\mathbf{Q}$ from our simulation trajectories using the expression [130],

$$q_{ij} = \frac{N_{ij}}{R_i}, \tag{2.11}$$

where N_{ij} is the number of occasions a monomer was observed to transition from a cluster of size i to one of size j over the course of the simulation, and R_i is the total holding time in cluster size i, defined as the sum of time periods that a cluster of size i was observed summed over all such clusters. To define whether or not a transition has occurred, we adopt a lag time, τ such that N_{ij} is incremented by one if a peptide is observed to reside in a cluster of size i at time t and in one of size j at time $(t+\tau)$. Small values of τ are desirable in that they most closely reflect the assumption of a time continuous Markov process and provide fine time resolution of the dynamics. Cluster association and dissociation events are, however, associated with high frequency oscillations in the cluster size due to our definition of a hard cutoff in peptide proximity defining monomer connectivity. Accordingly, the value of τ should be sufficiently large such that the transition rate estimates are not artificially elevated due to the short time fluctuations inherent to cluster agglomeration and fragmentation events.

To select an appropriate value of the lag time, we compute the maximum likelihood estimates of the transition rate matrix $\mathbf{Q}$ at various choices of τ using Equation 2.11 to estimate its off-diagonal elements q_{ij} as an average over all time blocks of length τ over the three independent simulations at each concentration. Diagonal elements are computed as $q_{ii} = -\sum_{j=1}^{n} q_{ij}$ [122]. We then employ Equation 2.9 to predict the evolution of our system from its initial state in which it exists exclusively as monomers (i.e., $\mathbf{p}(0) = [1, 0, \ldots, 0, 0]$). By comparing the cluster size distribution predicted by our Markov model to that directly observed in our simulations, we select the lag time that best reproduces the simulation data. As anticipated, we observe a trade-off in accuracy as a function of the lag time: τ must be sufficiently large to average out the high frequency fluctuations in cluster association and dissociation, but sufficiently small to resolve the short-time dynamics of the system. We find $\tau = 100$ ps to be optimal, or near-optimal, for all six simulation trajectories as measured by root mean square deviation between the time evolution of the cluster size distribution predicted by Equation 2.9 and those measured directly from the simulation trajectory.

In Appendix A.1 we show that the estimates of q_{ij} extracted from our simulation trajectories are time invariant, supporting the assumption of a time homogeneous continuous time Markov chain. In Appendix A.2 we demonstrate that the time evolution of the cluster size distribution is also well modeled as a discrete time Markov chain (DTMC) employing a lag time $\tau = 100$ ps. We also show that this lag time exceeds that Markov time for the system, verifying that the discrete time Markov model constructed with this lag time possesses the Markov (i.e., memoryless) property [124].

Low concentration

In Table 2.1, we report the maximum likelihood estimates of the transition rates q_{ij} extracted from our 0.85 mM concentration simulations using a lag time of $\tau = 100$ ps. In Figure 2.16, we compare the time evolution of the cluster size distribution predicted from the transition rate matrix using Equation 2.9 to that measured directly from the simulation trajectories. The CTMC predictions show good agreement with the simulation results, reproducing the trends in the mass fraction observed over the course of the simulations. This good agreement validates that the system is well modeled as a temporally and spatially homogeneous Markov process. We can understand this agreement by comparing the characteristic transition time for an aggregate of a size i, $\frac{1}{q_{ii}}$, to the characteristic time for the structural relaxation of aggregates of this size subsequent to their formation (c.f., Figures 2.12 and 2.15 for illustrative examples). Our analysis reveals the characteristic relaxation time to be shorter than the transition time for aggregates of size < 5, validating our assumption that small aggregates should structurally relax into β-sheet-like aggregates before their next transition, and therefore that our Markov state decomposition should be approximately memoryless. For larger aggregates of size ≥ 5, we find these time scales to be approximately equal, indicating that structural relaxation occurs on a similar time scale to the transition time, and that a Markov state decomposition partitioning based on both aggregate size and geometry may be required to accurately describe the aggregation kinetics of much larger aggregates than those observed in this work.

The transition rates of this process provide insight into the microscopic kinetics of the early stage assembly process. In particular, the relaxation times of the system – also known as the implied time scales – t_i are related to the (magnitude) of the eigenvalues of the transition matrix **T** as [124, 135],

$$t_i = -\frac{\tau}{\ln(|\lambda_i|)}, \tag{2.12}$$

where τ is the lag time used to construct the transition matrix and λ_i is a (possibly complex) eigenvalue of **T** arranged in non-ascending order of magnitude $\{\lambda_2 \geq \lambda_3 \geq \ldots\}$ and neglecting the leading unit eigenvalue $\lambda_1 = 1$ associated with the steady state distribution. The slowest relaxation time t_2 provides an estimate of the characteristic relaxation time of any system observable [136], and may therefore be related to experimental measurements of structural relaxation kinetics [129, 137]. For this system, we calculate a value of $t_2 \approx 2.5$ μs with a 95% confidence interval of 0.5 μs to 9.3 μs. (We numerically estimate the confidence interval by generating 100,000 matrices with off-diagonal elements randomly drawn from a normal distribution with mean and standard deviation reported in Table 2.1.)

It is important to note that the absence of observations of higher order aggregates means that the transition rate matrix is a partial block of the full transition rate matrix between all possible cluster sizes of n = 1-64. As such, this relaxation time scale reflects only the subset of transitions that were actually observed in our simulation trajectories, and may be interpreted as the slowest relaxation time associated with early-stage assembly. Nevertheless, the estimation of a relaxation time from relatively short molecular simulations by modeling aggregation as a CTMC is a powerful feature of this analysis technique [128, 136], and suggests that simulations on the order of microseconds would be required to probe beyond early-stage assembly [136]. The large computational expense of such long simulations would likely require the use of even more highly coarse-grained models than those presented herein.

Table 2.1: Maximum likelihood estimate of transition rates between aggregate sizes in units of μs^{-1} in the continuous time Markov model for the 0.85 mM system employing a lag time of τ = 100 ps. Off-diagonal elements are estimated as an average over all time blocks of length τ over the three independent simulation trajectories, and uncertainties estimated as the standard error in the values constituting the mean. Diagonal elements are computed as $q_{ii} = -\sum_{j=1}^{n} q_{ij}$, and the associated uncertainly estimated by standard propagation of errors. Transition rates for which no such transitions were observed over the course of the simulation trajectories, or the total residence time was less than the lag time, are reported as zero.

	To 1-mer	To 2-mer	To 3-mer	To 4-mer	To 5-mer
From 1-mer	-26.4 ± 5.8	21.1 ± 3.0	2.9 ± 0.9	1.5 ± 0.8	1.0 ± 0.7
From 2-mer	13.5 ± 6.8	-25.4 ± 8.0	9.1 ± 2.6	2.8 ± 1.4	0.0 ± 0.0
From 3-mer	3.9 ± 2.8	7.8 ± 5.5	-30.7 ± 13.1	19.0 ± 10.8	0.0 ± 0.0
From 4-mer	6.2 ± 4.5	0.0 ± 0.0	18.7 ± 13.6	-51.0 ± 25.4	26.1 ± 20.1
From 5-mer	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	-0.0 ± 0.0

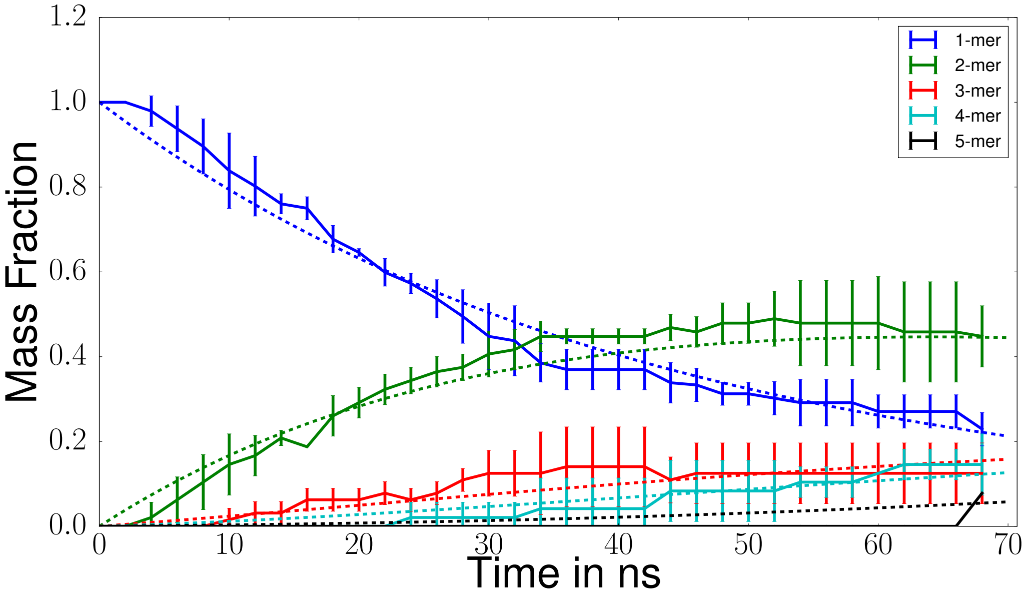

Figure 2.16: Comparison of the cluster size distribution in the 0.85 mM system predicted by the Markov model with a lag time of τ = 100 ps (dashed lines) to that directly observed in the simulations (sold lines with every 20th point plotted) for aggregates of size n = 1-5. The simulation data is plotted as the mean and standard deviation of the mass fraction over the three simulations. No error bars are reported when only a single observation was obtained at that particular time interval.

High concentration

In Table 2.2 and Figure 2.17 we present the inferred transition rates and predictions of the time evolution of the cluster size distribution for our 1.66 mM concentration simulations using a lag time of τ = 100 ps. We obtain good agreement between the simulation results and our model predictions for the lower order aggregates (n = 1-6), but poorer agreement for the larger aggregates (n = 7-11) for which our transition rate estimates are less statistically robust due to very few observations of the formation and dissociation of the larger clusters. Nevertheless, the qualitative trends in the time evolution in the mass fractions of all aggregates are adequately recapitulated by the CTMC. As with the 0.866 mM simulation, we find that typical structural relaxation occurs on time scales shorter than the characteristic transition time, validating our assumption that a Markov state decomposition based on aggregate size should be approximately memoryless. Encouragingly, we observe that nine of the twelve off-diagonal transition rates between the lower order aggregates (n = 1-4) are in agreement within error bars between the two concentrations studied (cf. Tables 2.1 and 2.2), suggesting that the transition rates inferred by our approach are not a strong function of concentration and lending confidence in the CTMC modeling approach. Of the three transition rates that fall outside of error bars, q_{14} differ by only 0.4 μs^{-1} after error, and q_{13}, and q_{23} by ~7 μs^{-1}. The absence of resolved transitions out of the heptamers, decamers, and undecamers over the time and length scales of our simulations reflects the behavior of these aggregates as sink states. At this concentration, we estimate the largest relaxation time for early-stage assembly to be $t_2 \approx$ 302 ns with a 95% confidence interval of 105 ns to 423 ns. This time scale is around an order of magnitude smaller than that computed at the lower concentration, but still suggests that simulations on the order of a microsecond would be required to observe events beyond the early stage assembly events studied in this chapter [136].

2.4 Conclusions

We have conducted molecular dynamics simulations to study the early-stage assembly of DFAG-OPV3-GAFD peptides in implicit and explicit solvent. We obtained the potential of mean force for the collapse of individual peptides and for pairwise dimerization using umbrella sampling in explicit solvent, and rescaled the non-bonded interactions of the CHARMM27 force field to reproduce these free energy landscapes with implicit solvation. Using this model, we employed biased sampling techniques to compute the equilibrium morphologies and association free energies for the formation of peptide dimers, trimers, tetramers, and pentamers under low-pH conditions where the four terminal carboxyl groups are protonated and the peptides are electrically neutral. All aggregates exist as β-sheet-like stacks mediated by π-stacking of the

Table 2.2: Maximum likelihood estimate of transition rates between aggregate sizes in units of μs^{-1} in the continuous time Markov model for the 1.66 mM system employing a lag time of $\tau = 100$ ps. Off-diagonal elements are estimated as an average over all time blocks of length τ over the three independent simulation trajectories, and uncertainties estimated as the standard error in the values constituting the mean. Diagonal elements are computed as $q_{ii} = -\sum_{j=1}^{n} q_{ij}$, and the associated uncertainly estimated by standard propagation of errors. Transition rates for which no such transitions were observed over the course of the simulation trajectories, or the total residence time was less than the lag time, are reported as zero.

	To 1-mer	To 2-mer	To 3-mer	To 4-mer	To 5-mer	To 6-mer	To 7-mer	To 8-mer	To 9-mer	To 10-mer	To 11-mer
From 1-mer	-41.3 ± 5.5	25.0 ± 4.4	11.7 ± 1.7	3.6 ± 0.9	2.1 ± 0.7	0.8 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
From 2-mer	25.0 ± 10.1	-56.3 ± 11.2	21.6 ± 2.9	4.6 ± 1.4	3.8 ± 1.4	0.0 ± 0.0	0.7 ± 0.5	0.5 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
From 3-mer	6.5 ± 4.5	6.8 ± 3.0	-42.9 ± 7.6	14.7 ± 3.5	8.4 ± 2.7	4.2 ± 2.2	1.2 ± 0.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.0 ± 1.1
From 4-mer	1.8 ± 1.1	0.7 ± 0.7	5.5 ± 3.3	-28.0 ± 7.0	14.3 ± 5.1	0.0 ± 0.0	2.6 ± 1.8	1.4 ± 1.4	0.0 ± 0.0	1.7 ± 1.7	0.0 ± 0.0
From 5-mer	5.2 ± 4.8	13.3 ± 6.8	19.9 ± 10.2	20.9 ± 19.3	-107.0 ± 34.2	42.5 ± 24.3	5.3 ± 3.8	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
From 6-mer	18.9 ± 18.2	0.0 ± 0.0	8.3 ± 8.4	0.0 ± 0.0	94.3 ± 91.1	-167.2 ± 102.3	0.0 ± 0.0	6.7 ± 6.8	0.0 ± 0.0	39.1 ± 40.3	0.0 ± 0.0
From 7-mer	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	-0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
From 8-mer	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	-75.8 ± 83.3	0.0 ± 0.0	0.0 ± 0.0	75.8 ± 83.0
From 9-mer	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
From 10-mer	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
From 11-mer	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

aromatic cores and hydrogen bonding between the peptidic wings. We predict a favorable dimerization free energy of $\Delta F \approx -15k_BT$, and compute a strongly favorable and approximately constant driving force for further monomer addition to the elongating stack of $\Delta F \approx -25k_BT$. Interestingly, under high-pH conditions where the peptides are deprotonated and carry a net -4 formal charge, we find dimerization to remain favorable with $\Delta F \approx -4k_BT$ driven by dispersion interactions, hydrophobicity, and π-stacking between the aromatic cores with the charged termini oriented away from one another in an "I-shaped" configuration. Successive monomer addition to form trimers, tetramers, and pentamers is also marginally favorable with $\Delta F \approx -5k_BT$, but becomes less so with increasing aggregate size due to higher concentrations of negative charge on the growing oligomer. These findings suggest that the peptides exist in smaller oligomeric clusters under high-pH conditions that form the fundamental units of assembly in pH-triggered assembly [4,22,23,56].

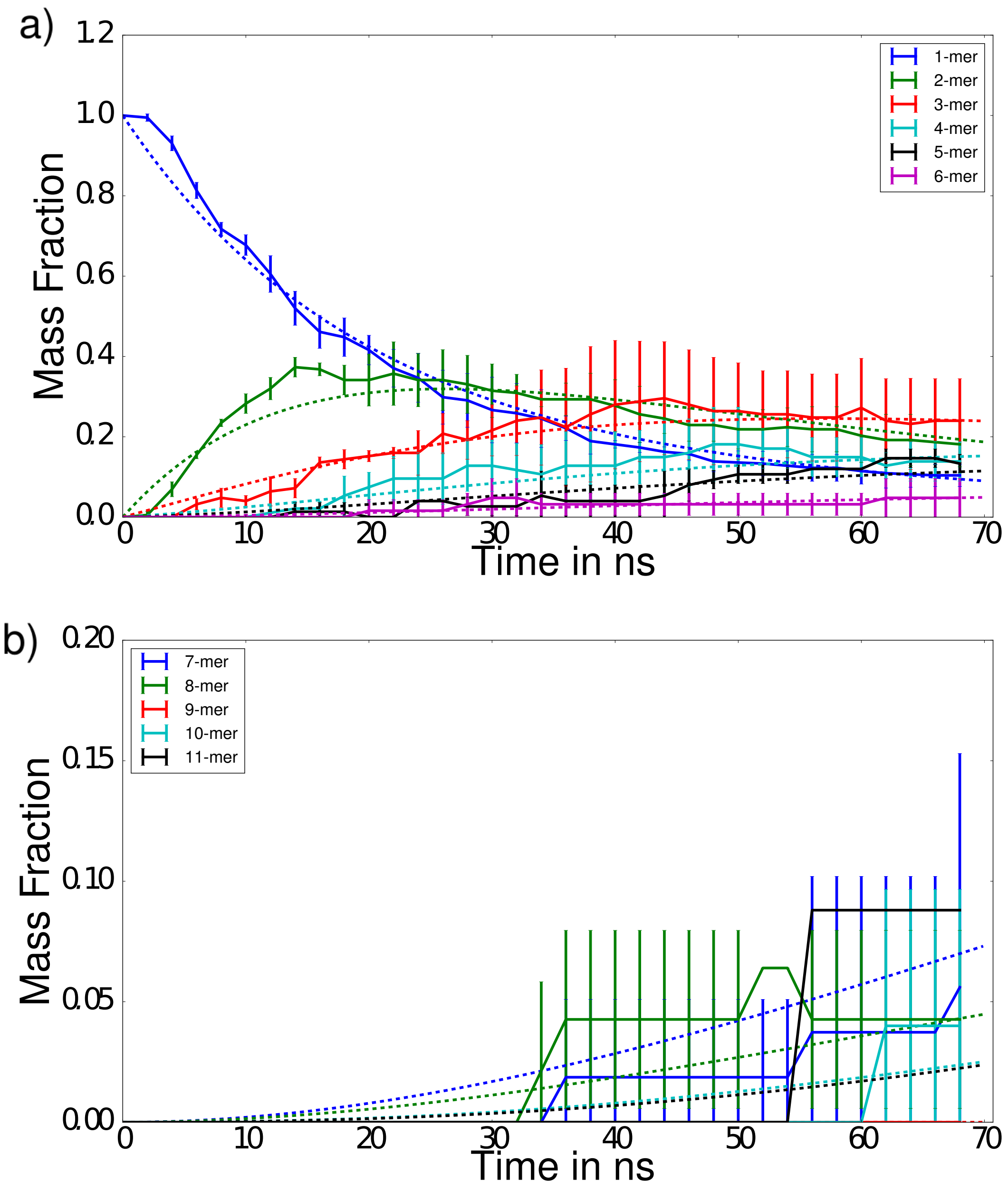

Figure 2.17: Comparison of the cluster size distribution in the 1.66 mM system predicted by the Markov model with a lag time of $\tau = 100$ ps (dashed lines) to that directly observed in the simulations (sold lines with every 20th point plotted) for aggregates of size (a) n = 1-6 and (b) n = 7-11. The simulation data is plotted as the mean and standard deviation of the mass fraction over the three simulations. No error bars are reported when only a single observation was obtained at that particular time interval.

The reparameterized implicit solvent model permitted us to access 2-3 orders of magnitude larger time and length scales than are accessible in explicit solvent, permitting us to directly simulate the early stages of self-assembly of the low-pH assembly of hundreds of protonated peptide monomers over 70 ns. At 0.85 mM concentration we see monotonic depletion of monomers leading to assembly of dimers, trimers, tetramers, and pentamers over the first 70 ns of assembly. These light aggregates are sufficiently small to undergo rapid internal structural relaxation to form well-aligned β-sheet-like stacks, but the assembly of heavier aggregates proceeds on time scales in excess of several hundreds of ns. At 1.66 mM concentration we see rapid depletion of monomers to form aggregates of size n=2-11 by both monomeric addition and the condensation of heavier aggregates. Morphologically, aggregates of size $n \leq 4$ undergo internal structural rearrangement into well-aligned β-sheet-like stacks that serve as nuclei for further elongation of the nascent ribbon. The larger aggregates appear to be kinetically trapped in disordered configurations on time scales of tens of ns. By modeling the assembly dynamics observed in our simulations as a continuous time Markov chain (CTMC), we extracted from our simulations transition rates between aggregates of different sizes providing insight into the microscopic kinetics of early-stage assembly. The predictions for the time evolution of the cluster size distribution from the CTMC are in good agreement with those extracted directly from our simulations, and predict early stage assembly to possess a slowest relaxation time on the order of several μs.

Our observations suggest a hierarchical model of early-stage assembly at experimentally relevant concentrations, in which light n=1-4 aggregates first rapidly assemble and reorganize into thermodynamically stable β-sheet-like stacks. These small subsequently agglomerate into larger disordered aggregates with internal structural relaxation times exceeding several tens of ns. These early stages of peptide assembly observed in our simulations resemble the initial stages of models of amyloid fibril formation in which amyloid peptides nucleate into roughly spherical micelle-like structures bound by primarily hydrophobicity, before ripening into beta-sheet-like structures with a defined backbone and linear stacking arrangement [138–143]. Moving beyond the 70 ns time horizon, our results suggest that the next phase of assembly to be structural ripening of these larger aggregates into well-aligned β-sheet-like stacks coupled with the formation of larger aggregates. The aggregation of small clusters with high mobilities into larger, less mobile clusters – we compute the self-diffusivity of monomers, tetramers, and octamers to be (6.6 ± 0.2), (1.64 ± 0.04), and (0.74 ± 0.02) $\times 10^{-6}$ cm^2/s respectively – suggests that higher order aggregation will likely be diffusion-limited. Further testing of these hypotheses would require the simulation of larger simulation boxes to gather sufficient data to draw statistically robust data on higher aggregate formation, and longer simulation times to probe higher-order assembly on microsecond time scales. Work has been done to push simulations to such time scales using high-level coarse-grained models [144].

This chapter lays the foundation for subsequent chapters in employing our reparameterized implicit solvent model to investigate the impact of modifications to peptide sequence and the conjugated core upon the thermodynamics, kinetics, and morphologies in early-stage assembly to make contact between monomer chemistry and assembly behavior. Specifically, we will study the DXXX-Π-XXXD peptide family where X = {G,A,I,V,F} [4,23,56,66,69], and Π corresponds to different conjugated cores of different characters and sizes, including oligophenylvinylenes, oligothiophenes, and rylene diimides [30,56,65] to identify amino acid sequences and cores predicted to possess desirable assembly characteristics. Peptide composition has a direct influence on both the structure of the assembled nanomaterials, such as the extent of fibrillization, as well as their functionality, such as the energy transport characteristics [2,4]. For example, different peptide chemistries have been observed to produce nanomaterials with excited state exciton outcomes spanning from the formation of "charge-trapped" excimer states that might be useful for light emission applications, to the formation of strong electronic coupling relevant for charge carrier transport [2]. Future computational studies can help to unravel the molecular-level morphologies underpinning this structure and function, and help guide the rational design of new biocompatible optoelectronic nanomaterials for energy transport and storage applications.

Chapter 3

QSPR modeling for high-throughput virtual screening

3.1 Introduction

We seek to expand the scope of our study to a variety of peptides of different composition in order to probe the relationships between primary structure and aggregate properties. While computer simulation can provide a great deal of insight, all-atom and even coarse-grained simulation at the length and time scales relevant to supramolecular assembly can be prohibitively expensive. This computational expense, coupled with the enormous palette of possible oligopeptide chemistries, makes it infeasible to directly evaluate the self-assembly behavior of every possible chemistry. Quantitative structure-activity relationship/quantitative structure-property relationship (QSAR/QSPR) models present a means to develop inexpensive predictive models of molecular behavior that can be used to perform high-throughput computational screening of chemical space to evaluate vastly more candidate peptides than would be possible by simulation and/or experimentation [146–149]. These models seek a predictive relationship between molecular behavior that is expensive to evaluate and a set of physicochemical molecular descriptors that are inexpensive to compute or measure. It is typically implicitly assumed that similar molecules behave similarly and that the input descriptors are sufficient to predict the desired molecular behavior [150]. The training of such models is a form of supervised learning, by which the relationship is extracted from a limited set of training data, validated against some test data, and then used to perform high-throughput "virtual screening". The use of QSPR models has a long history in chemometrics [151, 152], and has been applied to a diversity of peptidic systems in the context of structure, binding, drug loading, antimicrobial activity, and aggregation [150, 153–159].

In this chapter, we conduct molecular dynamics simulations of a limited number of synthetic oligopeptide chemistries, and use these data to train QSPR models to predict oligomerization thermodynamics

Most of this chapter is an excerpt from ref. [145]. Used with permission from: Thurston, B. A.; Ferguson, A. L. *Mol. Sim.* 44, 11, 930-945 (2018), available at https://doi.org/10.1080/08927022.2018.1469754. Copyright 2018 Taylor and Francis.

from molecular physicochemical descriptors. These models are then used to inform the important determinants of assembly behavior and perform high-throughput computational screening to identify peptide candidates with good predicted assembly behaviors. We focus our study on a class of synthetic oligopeptides with a peptide-Π-peptide symmetric triblock architecture of the form ASP-X_3-X_2-X_1-Π-X_1-X_2-X_3-ASP, where $\{X_1, X_2, X_3\}$ are amino acids from the set {ALA (A), PHE (F), GLY (G), ILE (I), VAL (V)} and the Π insert is either a naphthalenediimide (NDI) or a perylenediimide (PDI) conjugated core Figure 3.1). The peptide family represents a flexible archetype that may be readily synthesized by on-resin dimerization [65], and which has been previously shown in a number of prior computational [4, 24, 55, 56, 144] and experimental studies [2, 4, 23, 56, 65, 66] to possess a variety of desirable properties. Specifically, these biocompatible and water-soluble oligopeptides exist as dispersed small aggregates at neutral pH that are triggered to assemble into micron-sized pseudo-1D fibrils upon acidification due to protonation of the terminal ASP residues. This eliminates the electrostatic repulsion between the ASP residues, and promotes assembly by hydrophobic, hydrogen bonding, and π-stacking interactions. Delocalization of electrons between the π-conjugated cores provides the assemblies with functional electronic and photophysical properties, including electron transport and exciton migration, fluorescence, and gate voltage dependent current, which make such peptides viable materials to be used in biosensing, tracking molecular delivery to cells, energy transport and harvesting, imaging, field effect transistors, and other bioelectronic applications [2, 4, 48, 48, 49, 55, 65, 68, 69, 160]. Previous studies have probed the role of N-to-C polarity, peptide concentration, pH, and particular peptide sequences and core chemistries upon assembly [4, 24, 55, 56, 65, 161], but no work to date has sought to develop predictive physicochemical models of assembly to identify the important determinants promoting the formation of the ordered pseudo-1D assemblies required for good optoelectronic functionality and enable virtual screening of peptide sequence space. It is the principal motivation of this chapter to achieve these goals, and computationally test the model predictions by direct simulation of the assembly behavior of identified candidates in large-scale molecular simulations.

We structure this project around two hypotheses. First, we propose that physicochemical molecular descriptors can be used to develop predictive models of oligomerization free energies. A prerequisite to predicting large-scale many-body aggregation is a proper understanding of the mechanisms and thermodynamics of oligomerization [57]. We demonstrate that QSPR models can ably predict oligopeptide dimerization and trimerization free energies for non-polar oligopeptides from small numbers of molecular descriptors, revealing the important physicochemical determinants of association and setting the stage for our second hypothesis. Second, we propose that the large-scale assembly behavior can be predicted from the thermodynamics of oligomerization. We ground this conjecture in the well-known prin-

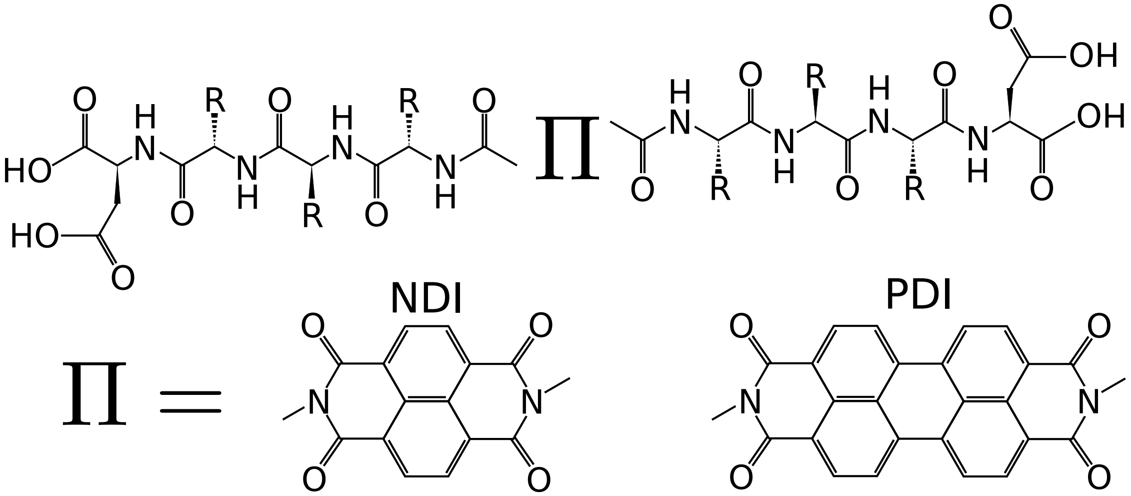

Figure 3.1: Chemical structure of the ASP-X_3-X_2-X_1-Π-X_1-X_2-X_3-ASP peptide family. {X_1, X_2, X_3} may be tuned to any one of the 20 natural amino acids, and the Π insert is a conjugated aromatic core, which in this chapter we restrict to be either a naphthalenediimide (NDI) or a perylenediimide (PDI) conjugated core. The N-to-C directionality of each peptide points away from the core, such that the oligopeptide sequence is antisymmetric and possesses two C-termini. The terminal aspartic acid residues and carboxyl termini are deprotonated at pH $\gtrsim 5$ such that the peptides carry a (-4) formal charge, and large-scale assembly is prohibited by electrostatic repulsion; at pH $\lesssim 1$, the termini protonate and assembly proceeds by π-π stacking, hydrogen bonding, and hydrophobic interactions [4].

ciple for self-assembling systems in general [162, 163], and experimental findings for this oligopeptide family in particular [4, 66, 69, 161], that interactions between self-associating building blocks should be sufficiently strong to mediate assembly, but not so strong as to prevent mutual rearrangements into ordered structures as opposed to kinetically trapped states. Our results provide good support that our QSPR model can accurately identify non-polar oligopeptides possessing intermediate oligomerization thermodynamics, and that these chemistries robustly assemble into aggregates with good in-register parallel stacking between neighboring molecules. We deploy our model to perform high-throughput screening of oligopeptide chemical space, and identify a number of novel candidate sequences that form well-ordered parallel-stacked nanoaggregates in large-scale molecular simulation. The structure of the remainder of this manuscript is as follows. In Section 3.2, we describe our simulation methodology and QSPR model development. In Section 3.3, we report the results of our free energy and alignment simulations, the implementation of a QSPR model to predict simulation results, and a high-throughput screening of chemistries based on this model. Finally, in Section 3.4, we close with our conclusions and outlook for future work.

3.2 Methods

3.2.1 Explicit solvent simulations

Molecular dynamics (MD) simulations of peptides in explicit solvent were conducted using GROMACS 4.6.7 [164, 165] in order to compute free energy profiles for peptide collapse and dimerization that we subsequently used to parametrise an implicit solvent model [55]. Peptide geometries were obtained using the GlycoBioChem PRODRG2 Server [71], and modelled using AMBER99SB [98, 166]. Terminal ASP residues were fully protonated to order simulate a low pH environment. The NDI and PDI cores are non-standard groups within AMBER99SB force field. Bonded parameters for the cores were determined using the parmchk2 method from Antechamber [167]. Native AMBER99SB parameters were unavailable for three bond angle interaction types, that instead were adopted from the Generalized Amber Force Field (GAFF) [168]. In keeping with methodology used for the derivation of partial charges for the AMBER force field [169], we compute partial charges on the core atoms by means of the Restrained Electrostatic Potential (RESP) method [170] using the RESP ESP charge Derive Server (REDS) [171]. Cores were parametrised as fragments by adding N-methylamide groups to either side and enforcing charge neutrality. The server computes charges utilizing a single configuration in two different orientations in each run [170, 172], and employs Gaussian09 [173] at the Hartree Fock/6-31G(d) level of theory to obtain the partial charges. Peptides were placed in a rhombic dodecahedral box with periodic boundary conditions and solvated in TIP3P water [174]. The box size was sufficiently large to accommodate the umbrella sampling calculations detailed in Section 3.2.4. The system was subjected to steepest descent energy minimization until the maximum force on any given atom was less than a threshold of 1000 kJ/mol.nm. Atomic velocities were initialized from a Maxwell distribution at 298 K and the system equilibrated for 100 ps in an NVT ensemble at a temperature of 298 K using a stochastic velocity rescaling thermostat [175] with a time constant of 0.5 ps, and finally for 100 ps in an NPT ensemble using the same thermostat and a Parrinello-Rahman barostat [176, 177] at a pressure of 1 atm with a time constant of 1 ps and a compressibility of 4.5×10^{-5} bar^{-1}. Production runs were conducted in the NPT ensemble using the same barostat and a Nosé-Hoover thermostat [178, 179] with a time constant of 0.5 ps to maintain a temperature of 298 K. The equations of motion were integrated using the leap-frog algorithm with a 2 fs time step [180]. Electrostatic interactions were treated using Particle Mesh Ewald (PME) with a cutoff of 1.0 nm and a 0.12 nm Fourier grid spacing that were optimized during runtime [78]. Lennard-Jones interactions were shifted smoothly to zero at 1.0 nm. Bond lengths were fixed using the LINCS algorithm [77], and Lorentz-Berthelot combining rules were used to determine interaction parameters between unlike atoms [79]. Execution speeds of 3.3 ns/day were achieved on one core of an Intel i7-4820K processor.

3.2.2 Implicit solvent simulations

To reach the long length and time scales necessary to observe peptide self-assembly and realize computational efficiency gains to enable us to simulate more molecular chemistries, we parametrised an implicit solvent model similar to that we previously employed in our study of the self-assembly of ASP-PHE-ALA-GLY-OPV3-GLY-ALA-PHE-ASP peptides [55]. Peptides were modelled using the AMBER99SB force field as described above [98, 166], but the water solvent is now represented implicitly using the Generalized Born model to treat polar interactions between peptide and solvent, and the solvent accessible surface area approximation to treat nonpolar interactions [80]. Nonpolar interactions were treated using an analytical continuum electrostatic (ACE) approximation [84] with a value of 2.259 kJ/mol.nm^2 for the surface tension [110]. Born radii were calculated using the method of Onufriev, Bashford, and Case with a relative dielectric constant of 78.3 and the standard parameter set of $\alpha = 1$, $\beta = 0.8$, and $\gamma = 4.85$ [83]. Coulombic interactions were treated using a cutoff of 3.4 nm and a dielectric offset of 0.009 nm. Lennard-Jones interactions were shifted smoothly to zero at 3.4 nm. All simulations were conducted in the NVT ensemble at 298 K by integrating the Langevin equation with a friction constant of 0.5 ps^{-1} [110]. Following our previous work [55], we rescale the non-bonded interactions within the AMBER99SB force field to compensate for the use of an implicit solvent model that overestimates inter-residue interaction strengths [85–87]. As detailed in Appendix B, we compute the optimal rescaling factor of $\alpha = 0.75$ from a best fit of the potential of mean force profiles for single peptide collapse (Section 3.2.3) and peptide dimerization (Section 3.2.4) to those computed in explicit solvent for a representative peptide chemistry ASP-PHE-ALA-GLY-NDI-GLY-ALA-PHE-ASP. Execution speeds of 33 ns/day were achieved on one core of an Intel i7-4820K processor, representing a ~10-fold speedup relative to the explicit solvent model.

3.2.3 Potential of mean force for peptide collapse

The potential of mean force (PMF) profile in the head-to-tail extent of a single oligopeptide specifies the free energy of the molecule as a function of its linear extent, quantifying the relative favorability of extended and collapsed configurations [93–95]. We determine PMF profiles for isolated oligopeptides in both explicit and implicit solvent and use these profiles to parametrise the implicit solvent model (Section 3.2.2). The PMF profiles were calculated by performing umbrella sampling in the molecular head-to-tail distance ($h2t$) defined as the distance between the C$_\alpha$ atoms in the terminal aspartic acid residues [93]. Umbrella windows were placed at evenly spaced 0.1 nm intervals over the range $h2t$ = 0-4.0 nm. Initial configurations for each umbrella window were obtained from non-equilibrium pulling of an initially fully extended peptide to induce collapse. Harmonic biasing potentials with a force constant of 1000 kJ/mol.nm^2 were applied in each umbrella window, and simulations run for 20 ns discarding the first 1 ns for equilibration. The unbiased PMF profile was estimated from the biased umbrella sampling data by solving the WHAM equations [94] using the g_wham program within GROMACS 4.6.7 [164, 165]. The PMFs resulting from each of the two independent umbrella sampling calculations are mutually aligned within the large-$h2t$ bond-stretching regions of the PMFs, and then averaged. Uncertainties in each individual PMF are estimated from 100 rounds of bootstrap resampling, and in the average by standard propagation of uncertainties.

3.2.4 Potential of mean force for peptide oligomerization

We also computed the PMF profiles for the formation of oligopeptide dimers ($n = 2$) and trimers ($n = 3$) as a function of the center of mass separation r_{COM} between a monomer and a preassembled $(n-1)$ oligopeptide stack. Initial stacks of n peptides were prepared by first stretching a monomer to its maximum head-to-tail extent, replicating it n times, and constructing parallel stacks of the copies with inter-monomer separations of 0.45 nm, corresponding approximately to the global free energy minimum of the dimerization PMF for stacked oligopeptides [55]. The system was allowed to equilibrate for 20 ps with the positions of the core atoms restrained, and then for another 20 ps with core restraints removed. This procedure allowed for peptides to relax into a well-stacked configuration. We then simulated the system for 1.5 ns, and used the system geometry at 0.5 ns, 1.0 ns, and 1.5 ns as the starting point for a non-equilibrium pulling runs. Nonequilibrium pulling is applied to the center of mass separation r_{COM} between a terminal monomer and the remaining $(n - 1)$ monomer stack. The configurations over the course of each pull are used to initialize umbrella sampling runs at evenly spaced 0.1 nm intervals over the range r_{COM} = 0-2.5 nm in the case of dimer aggregation and over the range r_{COM} = 0-3.0 nm in the case of trimer aggregation, where the upper bound of the range is specified to be sufficiently large that the two groups are effectively non-interacting. Harmonic biasing potentials with a force constant of 1000 kJ/mol.nm^2 were applied in each umbrella window, and simulations run for 20 ns discarding the first 1 ns for equilibration. The unbiased PMF profile was estimated from the biased umbrella sampling data by solving the WHAM equations [94] using the g_wham program within GROMACS 4.6.7 [164, 165]. The ΔF_{corr} = -2$k_B T$ ln(r_{COM}) correction is applied to each PMF to remove the purely entropic effects attributable to restraining the two groups to a particular separation [110, 181, 182]. The PMFs resulting from each of the three independent umbrella sampling calculations are mutually aligned within the large-r_{COM} plateau regions of the PMFs where the two groups are non-interacting, and then averaged. Uncertainties in each individual PMF are estimated from 100 rounds of bootstrap resampling, and in the average by standard propagation of uncertainties. The dimerization ΔF_2 and trimerization ΔF_3 free energies are defined as the difference in free energy between the large-r_{COM} non-interacting plateau of the PMF and the global free energy minimum containing the associated configurations.

3.2.5 Measurement of structural alignment

We ultimately seek to relate the ΔF_2 and ΔF_3 values predicted by our model to a measure of the quality of structural alignment within self-assembled aggregates formed by large numbers of peptides. Since we are interested in engineering peptides for optoelectronic functionality, our primary design objective is to establish good π-π stacking between neighboring oligopeptides within the self-assembled stacks. Essentially, we are using good parallel stacking of the π-conjugated aromatic cores as a classical proxy for good quantum delocalization of electrons over the backbone of the self-assembled stacks. We quantify the degree of structural alignment exhibited by a particular peptide chemistry by conducting 50 ns unbiased simulations of 64 peptides in a 50×50×50 nm^3 implicit solvent box, corresponding to a concentration of 0.85 mM. This concentration is both experimentally achievable and at which oligopeptide assembly has previously been observed [23]. The structure of aggregates is tracked as a function of time to monitor the formation of well-aligned parallel stacked clusters. The *association distance* between two oligopeptides is defined as [144, 183],

$$R_{a,b}^{\mathrm{assoc}} = \min_{i\in a} \min_{j\in b} r_{ij}, \tag{3.1}$$

where r_{ij} is the distance between atom i in oligopeptide a and atom j in oligopeptide b. Two oligopeptides are defined to be *associated* if $R_{a,b}^{\mathrm{assoc}} < 0.5$ nm. The *alignment distance* is defined as [144, 183],

$$R_{a,b}^{\mathrm{align}} = \max\left[\left(\max_{i\in(core\ a)} \min_{j\in(core\ b)} r_{ij}\right), \left(\max_{i'\in(core\ b)} \min_{j'\in(core\ a)} r_{i'j'}\right)\right]. \tag{3.2}$$

The first term in round brackets defines the minimum intermolecular distance from each atom i in the π-conjugated core of oligopeptide a to each atom j in the π-conjugated core of oligopeptide b, and selects the maximum of these. The second term in round brackets defines the reciprocal of this, computing the maximum minimum distance from any core atom i' in oligopeptide b to any core atom j' in oligopeptide a. In the square brackets we then take the maximum of the two terms. As has been previously observed, this distance is equivalent to the graph diameter [183, 184]. This measure presents a relatively strict definition of molecular association, with small alignment distances only reported that if all atoms within the cores of the two molecules are in close proximity. Accordingly, it presents a means to identify whether the cores of a pair of oligopeptides are in a parallel stacked configuration with in-register alignment between all of the fused aromatic cores. Two oligopeptides are defined to be *aligned* if $R_{a,b}^{\text{align}} < 0.5$ nm. We specify the two cutoffs based on the observed mean separation of two peptides in a 20 ns unbiased run starting from a well-aligned, π-stacked configuration. Based on these definitions, we define the alignment metric a for a particular snapshot of our molecular simulation as the ratio of the average size of aligned oligopeptide clusters to associated oligopeptide clusters,

$$a = \frac{\overline{n_a} - 1}{\overline{n_c} - 1}, \tag{3.3}$$

where $\overline{n_a}$ is the mean number of peptides in an aligned cluster and $\overline{n_c}$ is the mean number of peptides in an associated cluster. The subtraction of unity in the numerator and denominator assures that the metric does not spuriously assign high alignment scores to oligopeptide monomers, for which a is undefined. Averaging a over the equilibrated portion of our simulation trajectory provides a measure of the likelihood with which oligopeptides form well aligned clusters upon aggregation.

3.3 Results and Discussion

3.3.1 Dimerization and trimerization free energies computed by molecular simulation

Containing three independently mutable amino acid residues and NDI or PDI as potential π-conjugated cores, our ASP-X_3-X_2-X_1-Π-X_1-X_2-X_3-ASP peptide family comprises $20^3 \times 2$ = 16,000 members. Even with our implicit solvent model, exhaustive calculation of the dimerization and trimerization free energies from molecular simulation is computationally intractable. Accordingly, we instead perform these calculations over the restricted subset of oligopeptide chemistries DFAX-Π-XAFD, DFXG-Π-GXFD, and DXAG-Π-GAXD, where $X \in \{A, F, G, I, V\}$ and $\Pi \in \{NDI, PDI\}$. This choice of 26 different chemistries was motivated by experimental work showing good assembly behaviors of peptides belonging to these and similar families [2, 4, 65, 185], and the decision to avoid charged and/or polar residues that are expected to interfere with the triggerable low-pH association. We present in Table 3.1 the dimerization ΔF_2 and trimerization ΔF_3 free energies computed from the implicit solvent umbrella sampling simulations described in Section 3.2.4.

Analysis of the data reveals a significant difference between the free energies of interaction between

Table 3.1: Free energies of dimerization ΔF_2 and trimerization ΔF_3 computed from implicit solvent umbrella sampling calculations for the 26 chemistries in the families DFAX-Π-XAFD, DFXG-Π-GXFD, and DXAG-Π-GAXD, where $X \in \{A, F, G, I, V\}$ and $\Pi \in \{NDI, PDI\}$. Values are computed as the mean over the three independent runs, and uncertainties are estimated by propagation of uncertainties and bootstrap resampling. Eight of the 26 peptides were selected for large-scale simulations to assess the alignment quality of the self-assembled supramolecular assemblies (Section 3.3.4). We report for these eight chemistries the alignment metric a (Equation 3.3) averaged over the equilibrated portion of simulations of the assembly of 64 oligopeptides at a concentration of 0.85 mM as a measure of the probability oligopeptides will assemble into well-aligned stacks with in-register parallel stacking between the π-conjugated cores. Uncertainties were estimated by five-fold block averaging the equilibrated portion of the trajectory.

Chemistry	ΔF_2	ΔF_3	a
DAAG-NDI	-8.1±1.2	-13.8±4.8	–
DAAG-PDI	-22.3±1.6	-25.5±3.2	0.584 ± 0.039
DFAA-NDI	-7.7±2.6	-8.9±3.1	–
DFAA-PDI	-18.8±3.5	-22.5±3.1	–
DFAF-NDI	-11.8±2.4	-15.9±6.0	0.016 ± 0.007
DFAF-PDI	-24.0±3.2	-30.6±3.9	0.303 ± 0.039
DFAG-NDI	-8.1±2.1	-7.6±1.9	0.021 ± 0.016
DFAG-PDI	-22.1±2.8	-25.3±3.2	0.537 ± 0.042
DFAI-NDI	-9.4±3.1	-14.7±2.8	–
DFAI-PDI	-21.8±2.8	-34.9±4.0	0.268 ± 0.042
DFAV-NDI	-9.0±2.1	-17.8±3.6	0.020 ± 0.006
DFAV-PDI	-27.3±3.0	-21.2±2.9	0.348 ± 0.033
DFFG-NDI	-12.2±1.9	-16.6±4.9	–
DFFG-PDI	-29.2±3.1	-29.8±3.9	–
DFGG-NDI	-7.0±1.0	-9.1±2.7	–
DFGG-PDI	-26.9±2.7	-23.6±2.3	–
DFIG-NDI	-8.5±2.1	-14.6±2.6	–
DFIG-PDI	-27.6±4.7	-24.5±3.0	–
DFVG-NDI	-7.6±2.8	-12.9±3.2	–
DFVG-PDI	-22.9±2.6	-34.2±7.4	–
DGAG-NDI	-8.9±1.0	-5.7±2.5	–
DGAG-PDI	-26.9±1.7	-29.3±2.6	–
DIAG-NDI	-6.5±1.4	-8.0±3.1	–
DIAG-PDI	-24.4±3.3	-28.5±2.7	–
DVAG-NDI	-7.7±1.4	-4.6±1.2	–
DVAG-PDI	-20.3±1.6	-21.5±5.1	–

PDI and NDI cores: the most strongly interacting NDI peptides (DFFG-NDI for dimers and DFAV-NDI for trimers) possess more shallow free energy wells that the most weakly interacting PDI peptides (DFAA-PDI for dimers and DFAV-PDI for trimers). Interestingly, we observe that stronger free energy changes for the formation of a dimer do not necessarily imply stronger free energy wells in the formation of a trimer, indicating the importance of going beyond purely pairwise interactions in characterizing multi-body assembly [57]. For example, DFAV-PDI has one of the largest ΔF_2 values but one of the lowest ΔF_3 values. While it is clear that the larger PDI π-conjugated cores tend to elevate oligomerization free energies over that for NDI cores by a factor of 2-3, discerning more subtle trends based on peptide composition and sequence by inspection or intuition is challenging. In the following sections, we describe the development and interrogation of a QSPR model to assist in the discovery of the key determinants governing the thermodynamics of oligopeptide oligomerization.

3.3.2 QSPR modelling of oligomerization thermodynamics

We engage our hypothesis that oligomerization thermodynamics can be predicted from physicochemical molecular descriptors by training a QSPR model to predict oligopeptide dimerization and trimerization free energies computed by umbrella sampling. Training is conducted over data for the 26 chemistries reported in Table 3.1. Although these particular oligopeptides represent no more than a small sampling of the 16,000 possible ASP-X_3-X_2-X_1-Π-X_1-X_2-X_3-ASP chemistries, we show that it was sufficient to produce QSPR models capable of quantitatively predicting oligomerization free energies of a diversity of non-polar oligopeptides with diverse residue composition and sequence. Training of our QSPR model constitutes a form of supervised learning to regress a relationship of the form $\{\Delta F_2^i, \Delta F_3^i\} = f(\vec{d}_i)$, where ΔF_2 and ΔF_3 are the dimerization and trimerization free energies computed from molecular simulation, $\vec{d}$ is a vector of physicochemical molecular descriptors that can be inexpensively computed from the chemical sequence and/or three-dimensional structure of the peptide monomer, i indexes the particular peptide chemistry, and f is the functional mapping that is sought. Development of the QSPR model comprises four main steps: descriptor generation, descriptor cleaning, model construction, and model validation. An illustration of the QSPR training procedure is depicted in Figure 3.2.

Descriptor generation.

A molecular descriptor is a numerical quantity that can be computed directly from the molecular chemistry and/or structure [148,186]. The PaDEL software package [187] was used to compute a total of 1444 1D (dependent only on composition) and 2D (dependent on bond network) descriptors, and 431 3D (dependent on three-dimensional structure) descriptors. These descriptors correspond to a number of physical and chemical attributes, examples of which include numbers of atoms of various types, autocorrelations between atoms separated by particular numbers of bonds weighted by quantities such as electronegativity, and three-dimensional weighted radial distribution functions. Peptide structures required by PaDEL were generated by steepest descent energy minimization from an initially extended configuration until the maximum force in the system does not exceed 1000 kJ/mol.nm. The resulting file is converted to an MDL MOL format using Babel [188] and aromatic bonds manually assigned. When descriptors involve atomic partial charges, PaDEL computes them using the Gasteiger-Marsili method [189]. The resultant

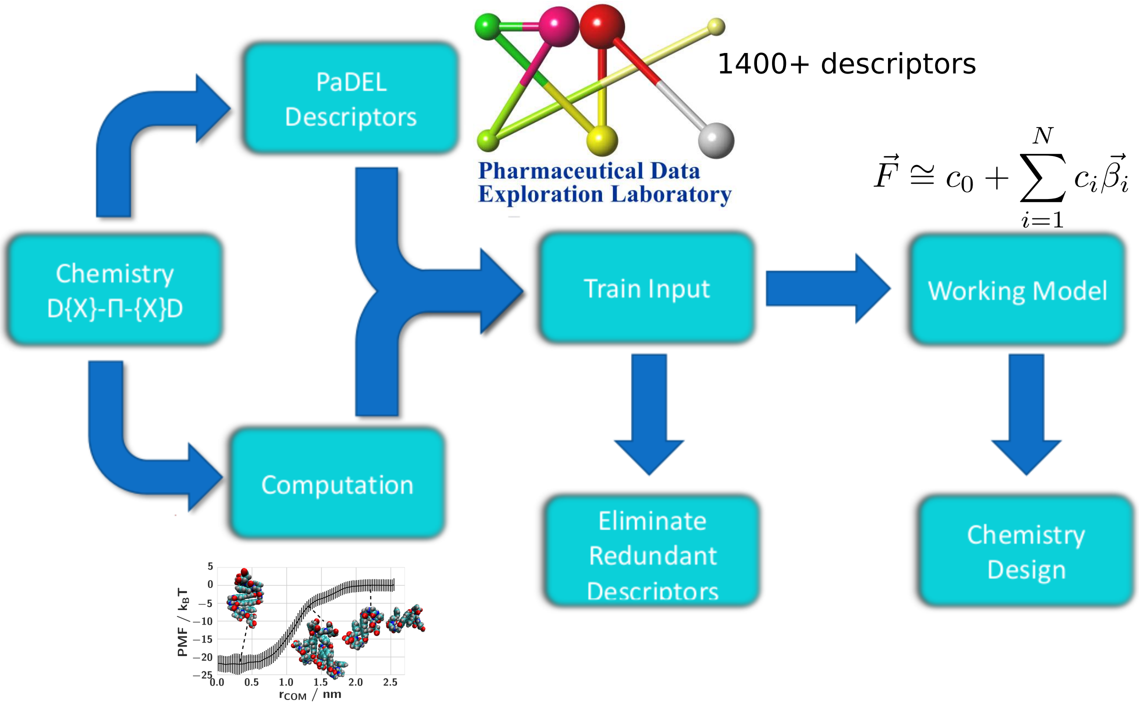

Figure 3.2: Schematic illustration of the QSPR model development protocol.

descriptors produced over the 26 chemistries are Z-scored such that they are centered, standardized, and de-dimensionalized [190]. To increase the diversity of descriptors and potentially improve their interpretability, we apply the descriptor generation protocol to the entire oligopeptide, the π-conjugated core alone, and the variable X_1-X_2-X_3 amino acid triplet. In this manner, we generate a total of 5625 descriptors for each of the 26 chemistries.

Descriptor cleaning.

We clean the ensemble of 5625 descriptors to eliminate unstable, uninformative or redundant descriptors [190,191]. First, we eliminated 1230 descriptors with a sensitive dependence on the three-dimensional structure of the oligopeptide. The oligopeptides are known to adopt a diversity of configurations both in isolation and within self-assembled aggregates, so we wished to discard descriptors that vary strongly with the peptide conformational state, and may be strongly influenced by the particulars of the methodology used to generate the initial peptide conformation. We compare descriptor values for all peptide chemistries in the training data set that are generated from our energy minimized peptide structure with those that are generated from the terminal configuration of a 20 ns simulation of an isolated peptide in implicit solvent, and select only those descriptors for which the root mean square deviation across all chemistries was less than a cutoff value of 0.15. Second, we eliminate 1686 descriptors that were found to be constant or nearly constant (defined as having a standard deviation less than 0.0001) over the 26 chemistries. Third, we removed 2462 highly correlated descriptors, identified from descriptor pairs possessing a Pearson correlation coefficient with magnitude in excess of $\rho = 0.90$. Highly correlated descriptors are removed in an iterative procedure. We first identify and retain the descriptor that is least correlated with all other descriptors, and then eliminate all descriptors with which it is highly correlated (i.e., $\rho \geq 0.90$). The next least correlated descriptor, of those that have not been selected or rejected, is then selected and those descriptors with which it is highly correlated are rejected, and so on. Together, these three cleaning protocols down-selected the number of descriptors from 5625 to 247.

Model construction.

We randomly partition the 26 chemistries into a training set consisting of 21 peptide chemistries (80% of the data) and a testing set comprising the remaining five chemistries. We train a QSPR model over the training data to regress a relationship of the form $\{\Delta F_2, \Delta F_3\} = f(\vec{d})$, where $\vec{d}$ is a vector of the 247 descriptors retained after cleaning. A number of choices of functional forms and machine learning approaches to determine f are possible, including artificial neural networks, support vector regression, Gaussian process regression, and nonlinear regression. In this chapter, we choose to employ simple multiple linear regression (MLR) for its simplicity, interpretability, and appropriateness for the high-dimensional low-sample size (HD-LSS) regime in which we are operating. As such, we seek a relationship of the form,

$$\Delta F^i = c_0 + \sum_{j=1}^{N} c_j d_j^i + \epsilon_i, \tag{3.4}$$

where i indexes the particular oligopeptide chemistry, j indexes the descriptor, ΔF^i is either the dimerization or trimerization free energy computed for oligopeptide i, d_j^i is the j^{th} descriptor associated with chemistry i, $\{c_i\}_{i=0}^{N}$ are the regression coefficients to be determined, and ϵ_i is the residual error associated with chemistry i. In principle, regression may be conducted over all $N = 247$ descriptors retained after the cleaning procedure. Typically, however, it is valuable to perform some form of descriptor selection in order to generate more simple, interpretable, and generalizable models in which a large number of the c_i are constrained to be zero [190, 191]. A number of means exist to perform wrapped and/or embedded descriptor selection, including genetic algorithms [192] and various flavors of regularization including ridge (L_2), LASSO (L_1), and elastic net (L_1 and L_2) [193–195]. In this chapter, we implement a combination of exhaustive and pseudo-greedy forward stepwise descriptor selection to regularize our models, and we avoid overfitting by dividing data into training and testing datasets and monitoring errors over the test set [196]. Specifically, we exhaustively compute all $\binom{247}{1} = 247$ univariate, $\binom{247}{2} = 30{,}381$ bivariate, and $\binom{247}{3} = 2{,}481{,}115$ trivariate MLR models by least squares fitting of Equation 3.4 over the 21 training chemistries. Exhaustive consideration of all $\binom{247}{4} = 151{,}348{,}015$ tetravariate models proved computationally expensive, so we instead greedily considered the 12,200 tetravariate models formed by adding one more descriptor to the top 50 trivariate models. Exhaustive computation of trivariate models proved to yield no significant benefit over trivariate models obtained using this greedy approach, so the final analysis was conducted using a greedy search approach for trivariate models as well and lending support for our use of this technique. Models were computed for 15 separate random divisions of data into training and testing sets in order to evaluate uncertainties in training and testing errors. In principle, we could have extended this stepwise selection procedure to models of arbitrary complexity, but in practice, we found tetravariate models or smaller were sufficient to predict dimerization and trimerization free energies with estimated uncertainties better that the accuracy with which these quantities were computed from our simulations. Furthermore, testing errors were observed to increase for both dimerization and trimerization free energies for higher than tetravariate models, indicating that the small size of our data set places us in a regime prone to overfitting. We avoid overfitting by controlling model complexity so as to minimize the testing error, but elect not to exacerbate the overfitting danger through the incorporation of nonlinear terms in our regression model. This also has the advantage of leading to more interpretable models formed from simple linear descriptor combinations.

Model validation.

At each order of model complexity $N = \{1,2,3,4\}$ the MLR models were validated and ranked according to their root mean squared error (RMSE) in leave-one-out cross-validation of ΔF_2 and ΔF_3 over the 21 chemistries constituting the training data. The performance of the top ranked $N = \{1,2,3,4\}$ models over the training and testing data for a random division of the data into training and testing datasets are illustrated in Figure 3.3.

In principle, we could have terminated our QSPR validation procedure here, and selected from the four top-ranked models with the smallest RMSE over the testing data as our terminal model. However, we seek to further improve our predictive accuracy by developing ensemble regressors that average over the top several models at each level of model complexity (i.e., the univariate, bivariate, trivariate, and tetravariate MLR models). It is well known that such ensemble models frequently exhibit better performance than any one of the constituent models alone [197–199]. We determine an appropriate number of top-ranked models over which to average at each level of model complexity by performing 15 rounds of shuffled cross-validation, in which we train the ensemble predictor on a randomly selected split of 80% of the training data and measure its prediction accuracy on the remaining 20%. We identify the optimal number of top-ranked models over which to average by identifying a knee in the curve of test RMSE against number of models participating in the average. This analysis identifies optimized ensemble predictors that average over the top one $N = 1$ order MLR models, top four $N = 2$ order models, top nine $N = 3$ order models, and top six $N = 4$ order models for ΔF_2, and the top three $N = 1$ order MLR models, top nine $N = 2$ order models, top four $N = 3$ order models, and top two $N = 4$ order models for ΔF_3. The performance of these ensemble models at each level of model complexity over the training and testing data is illustrated in Figure 3.4, from which we identify the $N = 2$ ensemble predictor to be optimal for prediction of ΔF_2, and the $N = 1$ ensemble predictor optimal for prediction of ΔF_3. The increase in testing error for models containing higher numbers of descriptors is a result of overfitting over our small dataset, and we avoid overfitting by selecting the model with the minimum testing error. The functional form of the best ΔF_2 ensemble model averaging over the four top-ranked $N = 2$ MLR models is,

$$\begin{aligned}\Delta F_2 = &\frac{1}{4}[(9.68 \times \text{MATS3c} + 4.58 \times \text{MATS1c\_wing} - 16.43)\\ &+ (9.85 \times \text{MATS3c} + 4.90 \times \text{ATSC4e\_wing} - 16.43)\\ &+ (-5.05 \times \text{SpMax6\_Bhs} - 10.73 \times \text{piPC3} - 16.43)\\ &+ (3.73 \times \text{AVP\_5} - 7.98 \times \text{piPC3} - 16.43)]\end{aligned} \tag{3.5}$$

and that of the best ΔF_3 ensemble model averaging over the three top-ranked $N = 1$ MLR models is,

$$\begin{aligned}\Delta F_3 = &\frac{1}{3}[(-7.80 \times \text{piPC3} - 19.30)\\ &+ (7.81 \times \text{maxHBint10} - 19.30)\\ &+ (7.83 \times \text{GATS2i} - 19.30)]\end{aligned} \tag{3.6}$$

An assessment of the statistical performance of these two ensemble models is presented in Table 3.2.

The particular descriptors resolved in our terminal ΔF_2 QSPR model are the Moran autocorrelation [200] of lag 3 weighted by partial charges (MATS3c) and Moran autocorrelation of lag 1 of the amino acids on one side of the core minus the ASP residue on the end weighted by partial charges (MATS1c_wing), and the centered Broto-Moreau autocorrelation [201] of lag 4 weighted by the Sanderson electronegativities [202] of the peptide wing (ATSC4e_wing), the 6^{th} largest eigenvalue of the modified Burden matrix [203] weighted by relative intrinsic state [204] (SpMax6_Bhs), conventional bond order ID number of order 3

Table 3.2: Statistical measures of our computed QSPR model for ΔF_2 and ΔF_3 for both training data and testing data. RMSE is the root mean square error of the model measured in $k_B T$, R^2 is the Pearson correlation coefficient, q^2 is the correlation coefficient over the leave one out cross validation of the training data, MAE is the mean average error of the model measured in $k_B T$, and R^2_{adj} is the adjusted correlation coefficient [1]. Error values are comparable to errors obtained in simulation. High values of the correlation coefficient and similar values of the adjusted correlation coefficient indicate the data are fit well by the model without overfitting.

	RMSE	R^2	q^2	MAE	R^2_{adj}
ΔF_2 training	1.9	0.95	0.90	1.5	0.91
ΔF_2 testing	3.0	0.83	-	2.6	0.75
ΔF_3 training	3.7	0.83	0.68	3.0	0.79
ΔF_3 testing	3.9	0.72	-	3.3	0.68

[186] (piPC3), and the average valance path of order 5 [205] (AVP_5). The particular descriptors resolved in our terminal ΔF_3 QSPR model are the conventional bond order ID number of order 3 [186] (piPC3), the maximum electrotopological state [204] descriptor of strength for potential hydrogen bonds of path length 10 (maxHBint10), and the Geary autocorrelation [206] of lag 2 weighted by first ionization potential (GATS2i).

Importantly, the simple ensemble MLR QSPR models defined in Equations 3.5 and 3.6 provide quantitatively accurate predictions of the dimerization and trimerization free energies to within calculation accuracy of ~4 $k_B T$ using just a handful of easily calculable molecular properties. Indeed, computation of the eight molecular descriptors required by these two expressions for a particular oligopeptide chemistry requires only about 4 s of computation on one core of an Intel i7-4820K processor, amounting to approximately a 3 million-fold speedup over direct calculation of ΔF_2 and ΔF_3 by molecular simulation. Accordingly, these models can be used to perform high-throughput virtual screening for oligopeptide chemistries possessing desirable oligomerization free energies.

3.3.3 Trained QSPR models provide molecular insight into determinants of oligomerization thermodynamics

We now proceed to interrogate the particular descriptors and associated regression coefficients appearing in the optimal MLR ensemble predictors in Equations 3.5 and 3.6. In doing so we pick apart the physicochemical properties reflected in each descriptor, and develop insight into the key molecular features governing the oligomerization thermodynamics.

MATS3c is the Moran autocorrelation [200] of lag 3 weighted by partial charges. Physically, this descriptor measures the correlation between atomic charges separated by three bonds. This descriptor appears twice with large positive regression coefficients in the expression for ΔF_2, indicating that large positive values of MATS3c favor large positive (i.e., unfavorable) values of ΔF_2. Bulkier aromatic residues and the larger PDI core tend to have lower values of MATS3c due to lower correlation between charges in atoms separated by three bonds in such residues compared to the higher correlation in the peptide backbone.

MATS1c_wing is the Moran autocorrelation weighted by partial charges between adjacent atoms in the peptide wing. This descriptor also appears with a positive coefficient in the expression for ΔF_2. This quantity is lowest for residues with multiple hydrogen atoms bonded to a single carbon atom in the peptide wings due to such a configuration resulting in the most polarized bonds in our training dataset. Aromatic residues, on the other hand, take on higher values. This term appears with and provides a correction to the MATS3c term by decreasing the magnitude of free energy wells for peptides containing wings with a higher fraction of aromatic atoms.

ATSC4e_wing is the centered Broto-Moreau autocorrelation [201] of lag 4 weighted by Sanderson electronegativities [202] of the peptide wing. This descriptor also has a positive correlation with ΔF_2. The Sanderson measure of electronegativity takes negative values for C and H atoms and positive values for N and O atoms. Due to spacing between atoms in the amino acid backbone, ALA and GLY residues tend to have lower values of this descriptor. Similar to MATS1c_wing, this descriptor appears with MATS3c and provides corrections to this term, in this case by predicting a larger free energy well for peptides containing more ALA and GLY residues.

piPC3 is the conventional bond order ID number of order 3 [186]. Physically, it measures the degree of branching in the bonded structure of the molecule. Aromatic bonds are weighted more heavily than single bonds so larger peptides, especially those containing large numbers of aromatic elements, with have larger values of this quantity. This descriptor appears twice in the expression for ΔF_2 and once in that for ΔF_3, in each case with large negative regression coefficients. PDI cores possess significantly larger values of piPC3 than NDI, with this descriptor reflecting the more favorable association free energies of the former relative to the latter. Less clear-cut correlations between residue size and lower free energies may play a role as well.

SpMax6_Bhs is the 6^{th} largest eigenvalue of the modified Burden matrix [203] weighted by relative intrinsic state [204]. It appears with piPC3 with a negative regression coefficient in the expression for ΔF_2. This descriptor does not have a simple physical interpretation, but is strongly negatively correlated $r = -0.922$ with the number of aromatic atoms. This descriptor appears to correct the piPC3 term by decreasing the free energy well for peptides containing a larger number of aromatic atoms, and increasing the free energy well for peptides containing fewer.

AVP_5 is the average valance path of order 5 [205]. It is positively correlated with ΔF_2 and appears with piPC3. AVP_5 can be thought of as a measure of molecular compactness: molecules with more paths of length 5 that contain heavier atoms with fewer valance electrons or atoms to which many hydrogen atoms are bound will have higher values for this descriptor. For the standard amino acids, this quantity will in general be larger when the ratio of hydrogen atoms to other atoms is larger. Accordingly, atoms containing PHE or ILE residues possess higher values, while NDI and PDI cores are not significantly differentiated from one another. This term appears to correct the piPC3 term that tends to overestimate the importance of residue size in determining ΔF_2.

maxHBint10 is the maximum electrotopological state [204] descriptor of strength for potential hydrogen bonds of path length 10. This descriptor is positively correlated with ΔF_3. Physically, the small size of the NDI core allows this path length to span across the core between two viable atoms, which is not possible for the PDI core. As a result, oligopeptides with NDI cores possess higher values for this descriptor than PDI peptides, resulting in less favorable trimerization free energies. Furthermore, the presence of atoms with lower Kier-Hall electronegativity in residues adjacent to the NDI core or residues one position away from the PDI core leads to lower values for this descriptor and deeper free energy wells for trimerization.

GATS2i is the Geary autocorrelation [206] of lag 2 weighted by first ionization potential. It has positive correlation with ΔF_3. This quantity is an inverse measure of autocorrelation (i.e., values > 1 indicate negative correlation and values < 1 indicate positive correlation) between the first ionization potential of atoms separated by two bonds. This quantity tends to reveal a weak negative correlation across all peptides, but the positive correlation between atoms in larger aromatic regions, such as PDI cores and PHE side chains, lowers the strength of this negative correlation and so the value of this descriptor. The second hydrogen in glycine residues also tends to decrease the strength of the negative correlation, so peptides containing more GLY will have lower values for this descriptor.

In sum, the QSPR model has identified a small number of physicochemcial properties that are the principal determinants of dimerization and trimerization free energies. Synthesizing the above analyses provides the following three physical insights into the molecular mechanisms governing assembly. First, the larger PDI cores lead to deeper free energy wells for dimerization and trimerization. Each set of descriptors for dimerization and each descriptor for trimerization has some way of drawing a sharp distinction between PDI and NDI cores. Specifically, this model predicts a $\sim 15 k_B T$ difference between PDI and NDI cores in both dimerization and trimerization free energies. Second, larger residues, especially PHE, are predicted to yield stronger free energies of aggregation, but that can easily be overestimated. Each pair of descriptors in the ΔF_2 model is characterized by the same general trend: one term distinguishes between

PDI and NDI cores and overestimates the stability of larger residues, and the second term corrects for this. Specifically, this model predicts an average increase in well depth over the range of simulated chemistries of $\sim 2.5 k_B T$ for replacing a given residue with PHE, with this effect larger for NDI than PDI cores. Third, larger residues with lower electronegativity nearer to the NDI core seem to play an important role in stabilizing the NDI trimer, while such residues are not as important in PDI trimers.

3.3.4 Correlation of oligomerization thermodynamics with self-assembled alignment quality.

Having developed a predictive QSPR model of dimerization and trimerization free energies, we now move to test our second hypothesis that oligomerization thermodynamics can predict the large-scale self-assembly behavior. Following the precept that self-assembling building blocks should possess sufficiently strong interactions to stabilize self-assembled aggregates but not so strong as to impose kinetic trapping and prohibit mutual rearrangements and healing of defects to form ordered aggregates [162], it is our conjecture that peptide chemistries with intermediate ΔF_2 and ΔF_3 values should show the best assembly into well-ordered aggregates with in-register stacking of the π-conjugated cores. Experimental support for this assertion in the context of these oligopeptides comes from recent work showing significant differences in photophysical and conductive properties of assembled peptides resulting from variation of peptide amino acid sequence [2, 4, 185]. These results are hypothesized to be caused by kinetically trapped aggregates forming at early stages of assembly [4] and variations in local packing order [185]. While these results are rather qualitative and pertain to different π-conjugated cores than those studied here, they nevertheless support the hypothesis that the formation of kinetically trapped states, most likely resulting from overly attractive peptide interactions, have a negative impact on core alignment. To test this conjecture, we conduct large scale simulations of assembly during which we monitor the degree of in-register parallel stacking of the π-conjugated cores. The computational expense associated with these calculations precluded us from conducting these runs for all 26 chemistries, and so we judiciously selected DFAG-NDI, DFAG-PDI, DFAF-NDI, DFAF-PDI, DFAV-NDI, DFAV-PDI, DFAI-PDI, and DAAG-PDI as eight oligopeptides spanning a wide range of dimerization and trimerization free energies (cf. Table 3.1). We track alignment quality using the alignment metric a defined in Equation 3.3 as a measure of the probability that associated peptides will form well-aligned parallel stacks. The time evolution of this quantity over the 50 ns simulations are reported in Figure 3.5a, and the average a values over the equilibrated portion of the runs reported alongside ΔF_2 and ΔF_3 in Table 3.1.

Our calculations reveal that oligopeptides with the smaller NDI core tend not to form well-aligned aggregates regardless of peptide wing chemistry. Peptides possessing a PDI core show two distinct groupings: DFAG-PDI and DAAG-PDI align most readily, while DFAV-PDI, DFAF-PDI, and DFAI-PDI do not align as well, although still better than those with an NDI core (Figure 3.5a). We quantify this relationship by fitting a bivariate Gaussian relating alignment quality and dimerization and trimerization free energies,

$$a = a_0 \exp\left(-\frac{(\Delta F_2 - \mu_{\Delta F_2})^2}{2\sigma^2_{\Delta F_2}} - \frac{(\Delta F_3 - \mu_{\Delta F_3})^2}{2\sigma^2_{\Delta F_3}} \right), \tag{3.7}$$

where a_0 = 0.548, $\mu_{\Delta F_2}$ = -22.2 $k_B T$, $\mu_{\Delta F_3}$ = -25.3 $k_B T$, and $\sigma_{\Delta F_2} = \sigma_{\Delta F_3}$ = 6.6 $k_B T$ provide a good fit to the data (Figure 3.5b). This fit illuminates a “goldilocks” regime in which peptides possessing intermediate $\Delta F_2 \approx \Delta F_3 \approx$ -25 $k_B T$ exhibit the best alignment, pointing towards an optimal tradeoff between sufficiently strong interaction strength to mediate assembly, but not so strong as to result in kinetic trapping in poorly ordered clusters.

3.3.5 High-throughput virtual screening

The good fit of the alignment metric to the dimerization and trimerization free energies – although based on a relatively small training set of only eight peptides – gives confidence that we can use our QSPR model to perform high-throughput screening of chemical space to identify peptides with ΔF_2 and ΔF_3 values predicted produce well-aligned stacks. Computing the eight molecular descriptors required by the QSPR model takes only 4 s per oligopeptide on a single Intel i7-4820K core, enabling traversal of orders of magnitude more chemistries than would be possible by molecular simulation. We search over the $17^3 \times 2$ = 9,826 chemistries in the ASP-X_3-X_2-X_1-Π-X_1-X_2-X_3-ASP peptide family, where Π ∈ {NDI, PDI} and {X_1, X_2, X_3} take on all possible natural amino acids with the exception of Lys, His, and Arg. These three residues are neglected since they are positively charged at low pH, and would therefore disrupt the pH-triggered assembly mechanism due to electrostatic repulsion. We present in Table 3.3 the predicted ΔF_2 and ΔF_3 values from our QSPR model (Equations 3.5 and 3.6) and alignment metric a from our bivariate Gaussian fit (Equation 3.7) for a selected fraction of the 9,826 chemistries.

In order to test our model predictions, we select from our list four oligopeptide chemistries predicted to possess good alignment metrics, and also seven controls. We select DMPP-PDI and DAIA-PDI as the two highest ranked chemistries. We also select DAVG-PDI as the highest ranked chemistry possessing a GLY residue adjacent to the core, as experimental work has previously suggested this as an important factor in dictating good assembly [2, 4, 185]. We also select DWWW-NDI as the highest ranked NDI core chemistry. Finally we also select DWNN-PDI, DTCT-NDI, DWCG-PDI, DWYW-NDI, DSSW-PDI, DYGA-

Table 3.3: Dimerization ΔF_2 and trimerization ΔF_3 free energies predicted by Equations 3.5 and 3.6 and alignment metric a predicted by Equation 3.7 for a selected number of the 9,826 chemistries in the ASP-X$_3$-X$_2$-X$_1$-Π-X$_1$-X$_2$-X$_3$-ASP peptide family, where $\Pi \in$ {NDI, PDI} and {X$_1$, X$_2$, X$_3$} take on all possible natural amino acids with the exception of Lys, His, and Arg. Chemistries are ordered by the magnitude of the predicted alignment metric. Uncertainties in ΔF_2 and ΔF_3 are the mean error in prediction of the testing data over 15 rounds of shuffled cross validation. Uncertainties in a are estimated by applying Equation 3.7 to 10^5 {ΔF_2, ΔF_3} pairs generated by sampling from a Gaussian distribution with the specified mean and standard deviation and taking the standard deviation of the result. DMMP-PDI and DAIA-PDI are selected for further simulation as the chemistries with the highest predicted alignment, DAVG-PDI is selected as the chemistry having a GLY residue nearest the core with the highest predicted alignment, and DWWW-NDI is selected as the NDI core with the highest predicted alignment. Finally, DWCG-PDI, DWYW-NDI, DSSW-PDI, DYGA-PDI, DYGG-PDI, DTCT-NDI, and DWNN-PDI are all selected as controls having a wide variety of predicted free energies, alignments, and constituent amino acids.

Chemistry	Predicted ΔF_2	Predicted ΔF_3	Predicted alignment a
DMMP-PDI	-22.4 ± 3.0	-25.5 ± 3.9	0.430 ± 0.096
DAIA-PDI	-22.4 ± 3.0	-25.2 ± 3.9	0.430 ± 0.097
DAMI-PDI	-22.0 ± 3.0	-25.4 ± 3.9	0.430 ± 0.097
DCMV-PDI	-22.3 ± 3.0	-25.7 ± 3.9	0.429 ± 0.097
DVAV-PDI	-22.2 ± 3.0	-25.4 ± 3.9	0.429 ± 0.097
DMMI-PDI	-22.6 ± 3.0	-25.4 ± 3.9	0.429 ± 0.097
DMIM-PDI	-22.6 ± 3.0	-25.5 ± 3.9	0.429 ± 0.097
DMIA-PDI	-22.3 ± 3.0	-24.9 ± 3.9	0.429 ± 0.097
DAIM-PDI	-22.0 ± 3.0	-25.6 ± 3.9	0.429 ± 0.097
DCVM-PDI	-22.3 ± 3.0	-25.7 ± 3.9	0.429 ± 0.097
DAVG-PDI	-22.0 ± 3.0	-24.9 ± 3.9	0.429 ± 0.097
DAMP-PDI	-21.9 ± 3.0	-25.7 ± 3.9	0.429 ± 0.097
DAAP-PDI	-22.5 ± 3.0	-24.8 ± 3.9	0.429 ± 0.098
DVMM-PDI	-21.6 ± 3.0	-25.2 ± 3.9	0.429 ± 0.097
DWWW-NDI	-22.7 ± 3.0	-25.6 ± 3.9	0.428 ± 0.098
...			
DWCG-PDI	-26.1 ± 3.0	-22.7 ± 3.9	0.351 ± 0.123
DWYW-NDI	-33.6 ± 3.0	-23.8 ± 3.9	0.123 ± 0.088
DSSW-PDI	-35.7 ± 3.0	-23.2 ± 3.9	0.074 ± 0.065
DYGA-PDI	-38.4 ± 3.0	-24.6 ± 3.9	0.036 ± 0.039
DYGG-PDI	-38.8 ± 3.0	-24.5 ± 3.9	0.032 ± 0.036
DTCT-NDI	-37.6 ± 3.0	-6.6 ± 3.9	0.002 ± 0.006
DWNN-PDI	-72.6 ± 3.0	-28.1 ± 3.9	0.000 ± 0.000

PDI, and DYGG-PDI as controls possessing a range of predicted free energies and alignments, and constituent amino acids. Several of these oligopeptides contain polar amino acids, which were not contained in the training data set, and so allow us to assess the generalizability and transferability of our model. These 11 oligopeptides were then subjected to implicit solvent molecular simulation to evaluate their dimerization and trimerization free energies, and large-scale 40 ns simulations of 64 peptides at 0.85 mM to assess their alignment behaviors. We report the predicted and calculated values of ΔF_2, ΔF_3, and a in Table 3.4.

Comparison of the predicted and calculated values of ΔF_2 and ΔF_3 show that our QSPR model accurately predicts the oligomerization thermodynamics for both NDI and PDI oligopeptides containing non-

Table 3.4: Predicted and calculated dimerization free energy ΔF_2, trimerization free energy ΔF_3, and alignment metric a for the 11 oligopeptide chemistries selected from our high-throughput virtual screening. Chemistries above the horizontal line possess non-polar amino acid residues for which QSPR model predictions of the oligomerization thermodynamics and alignment quality are generally very good. The polar oligopeptide chemistries – defined as those containing a difference in partial charge between any two bonded atoms greater than 1.0 e – reside below the line, for which the model predictions are relatively poor. Chemistries are ordered by the magnitude of the predicted alignment metric.

Chemistry	ΔF_2 (pred)	ΔF_2 (sim)	ΔF_3 (pred)	ΔF_3 (sim)	a (pred)	a (sim)
DMMP-PDI	-22.4 ± 3.0	-21.3 ± 2.3	-25.5 ± 3.9	-22.8 ± 3.3	0.430 ± 0.096	0.245 ± 0.014
DAIA-PDI	-22.4 ± 3.0	-25.8 ± 2.1	-25.2 ± 3.9	-29.1 ± 5.2	0.430 ± 0.097	0.579 ± 0.048
DAVG-PDI	-22.0 ± 3.0	-23.2 ± 2.0	-24.9 ± 3.9	-25.4 ± 3.2	0.429 ± 0.097	0.640 ± 0.072
DWWW-NDI	-22.7 ± 3.0	-21.8 ± 2.3	-25.6 ± 3.9	-30.1 ± 4.9	0.428 ± 0.098	0.016 ± 0.005
DWCG-PDI	-26.1 ± 3.0	-32.0 ± 3.8	-22.7 ± 3.9	-27.0 ± 4.6	0.350 ± 0.123	0.272 ± 0.048
DWYW-NDI	-33.6 ± 3.0	-22.2 ± 7.1	-23.8 ± 3.9	-30.5 ± 3.9	0.123 ± 0.089	0.000 ± 0.000
DSSW-PDI	-35.7 ± 3.0	-51.2 ± 5.3	-23.2 ± 3.9	-103.5 ± 38.5	0.074 ± 0.064	0.269 ± 0.021
DYGA-PDI	-38.4 ± 3.0	-30.4 ± 2.6	-24.6 ± 3.9	-36.5 ± 7.9	0.036 ± 0.039	0.600 ± 0.085
DYGG-PDI	-38.8 ± 3.0	-26.1 ± 4.6	-24.5 ± 3.9	-40.3 ± 6.8	0.032 ± 0.036	0.468 ± 0.047
DTCT-NDI	-37.6 ± 3.0	-83.4 ± 9.2	-6.6 ± 3.9	-91.8 ± 28.9	0.002 ± 0.005	0.037 ± 0.010
DNWW-PDI	-72.6 ± 3.0	-31.2 ± 3.6	-28.1 ± 3.9	-49.6 ± 4.2	0.000 ± 0.000	0.139 ± 0.020

polar amino acid residues (Table 3.4, upper). For all five such chemistries, the predicted and calculated quantities li.e. well within the estimated uncertainties. Importantly, this group of chemistries contains MET (M), PRO (P), CYS (C), and TRP (W) residues that were not part of the training ensemble, but the model is sufficiently transferable to give good predictive performance. Further, the model correctly predicts that the six TRP residues in the peptide wings lead to strong associations in the DWWW-NDI chemistry, despite the fact that very few of the NDI training examples had dimerization and trimerization free energies that were even half as large (Table 3.1). This indicates the model is able to accurately estimate the impact bulkier aromatic regions have on the free energies of aggregation. Considering now the six polar chemistries (Table 3.4, lower), we see poor agreement of the predicted and calculated free energies. These chemistries all contain one or more polar residues SER (S), ASN (N), THR (T), or TYR (Y) containing OH or NH_2 polar moieties. The poor predictive performance of our QSPR model may be attributed to the fact that our training data contained only the non-polar residues ALA(A), PHE (F), GLY (G), ILE (I), and VAL (V), and clearly demonstrates that our current model cannot be reliably extrapolated to strongly polar molecules.

We observe similar trends in the QSPR model prediction of the alignment metric a. We see relatively good, although not quantitative, agreement between the predicted and calculated a values for the five non-polar chemistries, with the only outlier being DWWW-NDI. The large deviation for this chemistry may be attributed to the fact that although the dimerization and trimerization free energies li.e. within the identified optimal range, the association is mediated in large part through the aromatic groups in the peptide wings rather than the aromatic core. Accordingly, good core-core parallel stacking is compromised by core-wing π-π stacking interactions. It is a failure of our simple model predicting alignment quality from oligomerization free energies alone that we do not distinguish the structural locale of the π-interactions within the oligopeptide. Conversely, our model shows very poor performance in predicting the alignment quality of the polar oligopeptide chemistries.

Our results support our hypothesis that the ΔF_2 and ΔF_3 of non-polar peptide oligomers can be accurately predicted by our QSPR model, and these oligomerization free energies used to identify non-polar oligopeptide chemistries – excluding those possessing high aromatic residue contents – likely to possess good alignment ($a \gtrsim 25\%$) within the self-assembled aggregates. In particular, we identify a chemistry DAVG-PDI previously unstudied by either simulation or experiment showing very high structural alignment propensity of $a = 0.64$. A representative snapshot of the equilibrium aggregates formed by this oligopeptide chemistry is presented in Figure 3.6.

3.4 Conclusions

We conducted molecular dynamics simulations and developed QSPR models to understand and engineer self-assembling π-conjugated ASP-X$_3$-X$_2$-X$_1$-Π-X$_1$-X$_2$-X$_3$-ASP oligopeptides. These molecules exhibit pH-triggered assembly with in-register parallel π-π stacking between the conjugated aromatic cores leading to electronic delocalization along the nanoaggregate backbones and the emergence of desirable optical and electronic properties. Our study was founded on two hypotheses: that physicochemical properties of the oligopeptides can be used to accurately predict dimerization and trimerization thermodynamics, and that chemistries possessing moderate oligomerization free energies produce the best ordered nanoaggregates. To engage these hypotheses, we parametrised an implicit solvent molecular model against explicit solvent all-atom calculations, and used this efficient model to compute the dimerization ΔF_2 and trimerization ΔF_3 free energies for 26 oligopeptides generated from all ALA (A), PHE (F), GLY (G), ILE (I), and VAL (V) point mutants – excluding the distal ASP (D) residues required for pH-triggered assembly – of DFAG-Π-GAFD oligopeptides containing NDI and PDI cores. These results revealed the larger PDI cores to give rise to $\Delta F_2 \sim (-24)\ k_BT$ and $\Delta F_3 \sim (-27)\ k_BT$ compared to only $\Delta F_2 \sim (-9)\ k_BT$ and $\Delta F_3 \sim (-12)\ k_BT$ for NDI inserts. To parse more subtle trends based on the composition and sequence of the peptide wings, we parametrised a QSPR model based on eight molecular descriptors that was capable of quantitatively predicting the dimerization and trimerization of non-polar oligopeptides. The predictive performance for polar chemistries was poor, and attributable to the fact that the model was developed exclusively over non-polar training examples. The particular descriptors identified by the model are informative as to the underlying determinants of the oligomerization thermodynamics. It predicts oligomerization free energies to be ~15 k_BT larger for PDI cores as compared with NDI, and bulkier residues, especially PHE, to increase free energies of association by ~2.5 k_BT. Finally, we observe that amino acids having lower electronegativity near the peptide core may play an important role in stabilizing formation of the NDI trimer. In developing a qualitatively accurate QSPR model for non-polar oligopeptide dimerization and trimerization thermodynamics, we provide strong support for our first hypothesis. This result is weakened by the poor performance for polar chemistries, but we anticipate that expansion of the training set to encompass polar training examples can produce similarly accurate models for this class of molecules.

We then correlated the alignment quality of associated peptides with the computed oligomerization free energies to develop a model that supported the existence of optimal dimerization and trimerization free energies of $\Delta F_2 \approx \Delta F_3 \approx$ (-25) k_BT. Heartened by this support for our second hypothesis, we performed a high-throughput screen of oligopeptide chemical space to identify a number of novel candidate chemistries predicted to exhibit good alignment behavior alongside a number of controls. Direct large-scale simulation showed our QSPR model to be a good, but not quantitatively accurate, predictor of alignment quality for non-polar oligopeptides. Using this approach, we were able to computationally identify and validate DAVG-PDI-GVAD as a promising oligopeptide chemistry not previously studied by experiment or simulation that exhibits good ordering in its self-assembled pseudo-1D nanoaggregates, and is therefore disposed to desirable optical and electronic functionality.

In future work, we aim to expand the training data to incorporate polar oligopeptide chemistries in order to build a more general and transferable QSPR model. Moreover, we would like to expand the training set to incorporate side chains of differing lengths and a wider variety of Π cores, including oligophenylvinylenes, oligothiophenes, and other rylene diimides. We also propose to incorporate additional computational techniques, including deep learning techniques that obviate the need for descriptors [207–209], Markov state models parametrised by molecular simulation data to reach longer length and time scales [55, 125], and time-dependent density functional theory (TD-DFT) to explicitly engage the electronic properties of the self-assembled aggregates. Finally, we will work with experimental collaborators to explicitly test the optimal designs identified under our computational screening protocol thereby guiding and accelerating experimental discovery efforts, and also incorporate the experimental results into our modelling paradigm to refine and improve our computational screens. These endeavors will continue to pave the way for design and realization of self-assembling oligopeptides as novel biocompatible supramolecular optoelectronic materials.

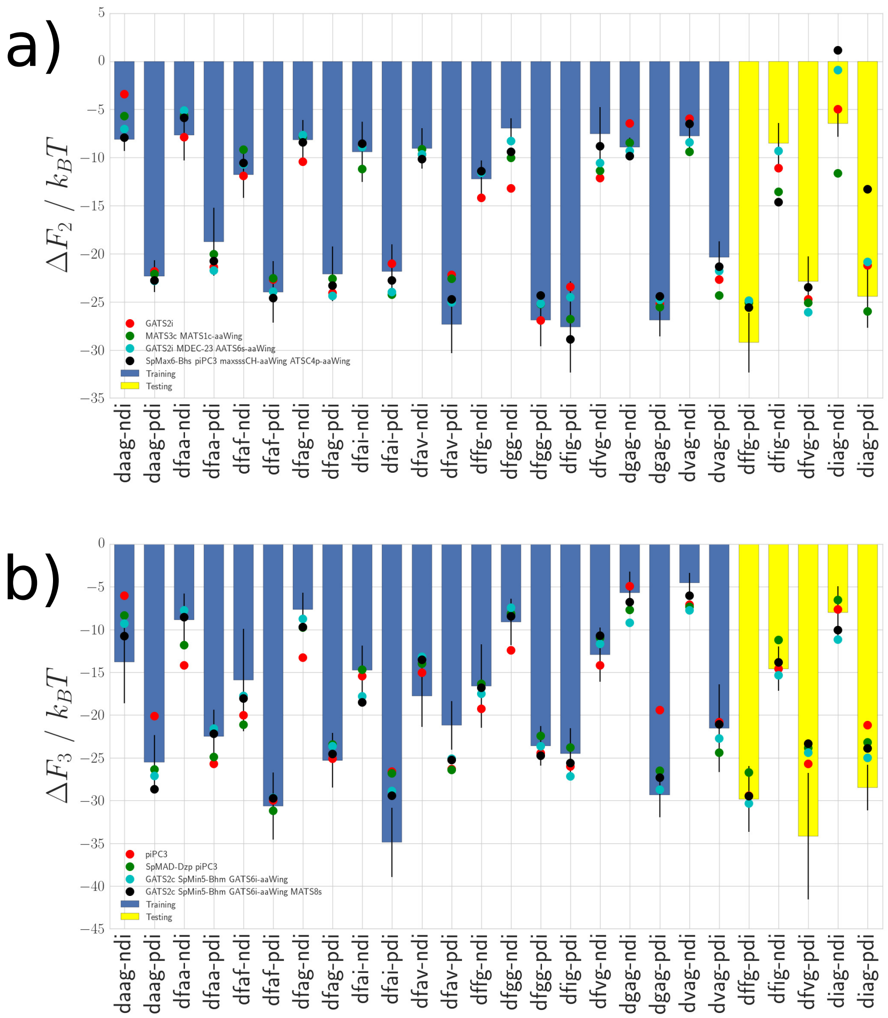

Figure 3.3: Performance of top-ranked MLR models comprising $N = 1$ (red dots), 2 (green dots), 3 (cyan dots), and 4 (black dots) molecular descriptors in predicting (a) dimerization free energy and (b) trimerization free energy computed in simulation. Free energies are reported in units of k_BT, where k_B is Boltzmann's constant and $T = 298$ K. The MLR models are fitted by least-squares fitting over the 21 training chemistries (blue bars), and their performance evaluated over the five testing chemistries (yellow bars). Black error bars indicate the estimated uncertainties in the ΔF_2 and ΔF_3 values computed from molecular simulation. The particular descriptors constituting the top ranked models are reported in the legends where GATS2i is the Geary autocorrelation of lag 2 weighted by first ionization potential, MATS3c is the Moran autocorrelation of lag 3 weighted by charges, MATS1c-aaWing is the Moran autocorrelation of lag 1 weighted by charges of the peptide wing, MDEC-23 is the molecular distance edge between all secondary and tertiary carbons, AATS6s-aaWing is the Average Broto-Moreau autocorrelation of lag 6 weighted by I-state of the peptide wing, SpMax6-Bhs is the sixth largest absolute eigenvalue of the Burden modified matrix weighted by the relative I-state, piPC3 is the conventional bond order ID number of order 3, maxsssCH-aaWing is the maximum atom-type E-State for singly bonded carbons with one hydrogen of the peptide wing, ATSC4p-aaWing is the centered Broto-Moreau autocorrelation of lag 4 weighted by polarizabilities, SpMAD-Dzp is the spectral mean absolute deviation from Barysz matrix weighted by polarizabilities, GATS2c is the Geary autocorrelation of lag 2 weighted by charges, SpMin5-Bhm is the fifth smallest absolute eigenvalue of Burden modified matrix weighted by relative mass, GATS6i-aaWing is the Geary autocorrelation of lag 6 weighted by the first ionization potential of the peptide wing, and MATS8s is the Moran autocorrelation of lag 8 weighted by I-state.

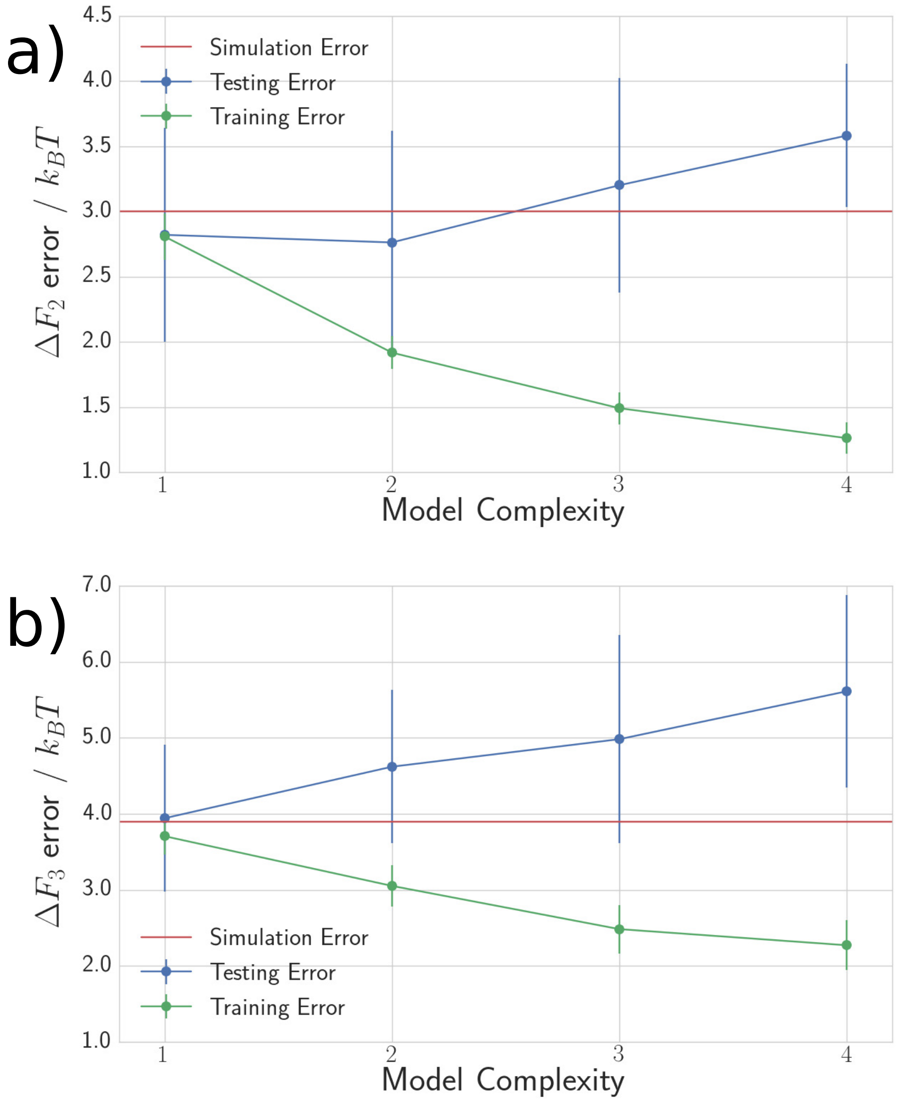

Figure 3.4: Performance of the optimal ensemble models at each level of model complexity over the training (green) and testing (blue) data in predicting the (a) dimerization free energy and (b) trimerization free energy computed in simulation. Free energies are reported in units of $k_B T$, where k_B is Boltzmann's constant and $T = 298$ K. For ΔF_2, the optimal ensemble model comprising $N = 1$ molecular descriptors averages over the single top-ranked MLR model, $N = 2$ over the top four, $N = 3$ over the top nine, and $N = 4$ over the top six. For ΔF_3, the optimal ensemble model comprising $N = 1$ molecular descriptors averages over the three top-ranked MLR models, $N = 2$ over the top nine, $N = 3$ over the top four, and $N = 4$ over the top two. The uncertainty in the ΔF_2 and ΔF_3 computed from simulation is depicted as a horizontal red line. Uncertainties in the model predictions are estimated from $K = 15$ rounds of shuffled cross-validation and depicted as error bars. This analysis reveals the $N = 2$ ensemble predictor to be optimal for prediction of ΔF_2, and the $N = 1$ ensemble predictor optimal for prediction of ΔF_3. The increase in testing error for models that utilize higher numbers of descriptors indicates that such models are overfitting the data.

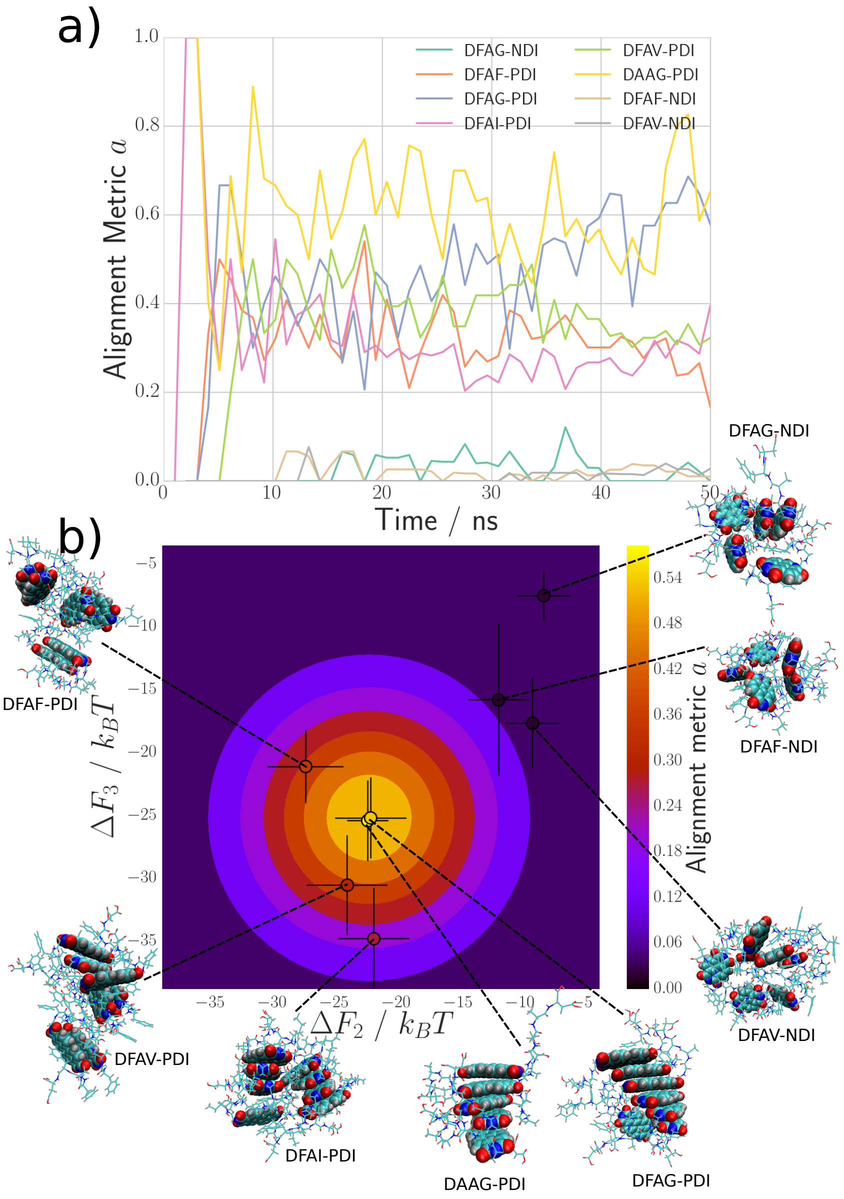

Figure 3.5: Alignment assessment of oligopeptide aggregates. (a) Time evolution of the alignment metric a (Equation 3.3) over the course of 50 ns runs of the self-assembly of 64 oligopeptides at a 0.85 mM initialized from randomly oriented monomers deposited over a grid. Values of a averaged over the equilibrated portion of the trajectory are reported in Table 3.1. (b) Scatter plot of the dimerization ΔF_2 and trimerization ΔF_3 free energies with points colored by the computed alignment metric a. Characteristic snapshots of the oligopeptide aggregates extracted from our molecular simulations show that DFAG-PDI and DAAG-PDI tend to form well-aligned stacks, DFAV-PDI, DFAF-PDI, and DFAI-PDI show a weaker propensity for good alignment, and DFAV-NDI, DFAF-NDI, and DFAG-NDI do not associate into well-formed stacks. The contour plot represents a best fit bivariate Gaussian with $\mu_{\Delta F_2}$ = -22.2 k_BT, $\mu_{\Delta F_3}$ = -25.3 k_BT, and $\sigma_{\Delta F_2} = \sigma_{\Delta F_3}$ = 6.6 k_BT, where k_B is Boltzmann's constant ant T = 298 K. The data and fit support the assertion that intermediate ΔF_2 and ΔF_3 values result in the optimal oligopeptide alignment.

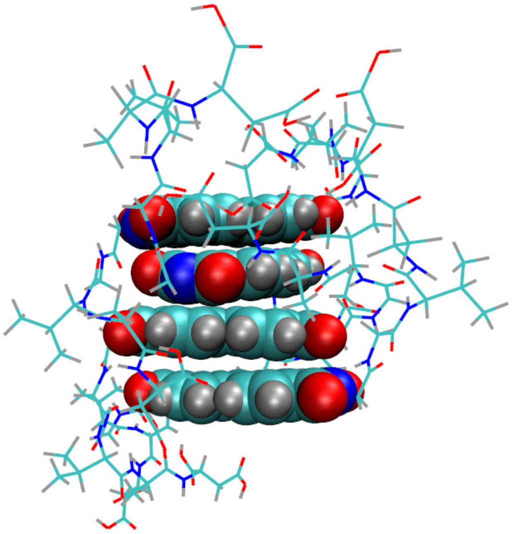

Figure 3.6: Representative snapshot of a self-assembled aggregate formed by the DAVG-PDI oligopeptide chemistry in 40 ns implicit solvent molecular dynamics simulations of 64 peptides at 0.85 mM.

Chapter 4

Prenucleation of pH-triggered assembly

4.1 Introduction

Utilizing the simulation and analysis of the techniques discussed in Chapter 2, we now turn to a study of the morphology and thermodynamics of π-conjugated oligopeptides in their high-pH charged state prior to acid-triggered assembly. This understanding is an important prerequisite to the principled control and manipulation of low-pH assembly [211]. This synthetic oligopeptide system consisting of a π-conjugated core flanked by symmetric sequences of amino acids has been explored in various permutations, [23, 30, 49, 66, 212] including recent work exploring the acid-mediated assembly of DFAG-Π-GAFD, where Π can be one of various π-conjugated systems including quaterthiophene (OT4), oligo(p-phenylenevinylene) (OPV3), and perylene-diimide (PDI). The π-conjugated cores of these materials can be tuned for the specific applications desired, such as making p-type semiconductors with OT4 and OPV3 cores and n-type with PDI. The pKa of the carboxyl terminus and C-terminal aspartic acid are 2.09 and 3.86, respectively [3]. Under the Henderson-Hasselbalch formalism, we can estimate that at pH 5 or higher the monomers are essentially completely deprotonated carrying a formal charge of (-4)e that precludes large-scale assembly by Coulombic repulsion. At pH 1 or lower, they are essentially completely protonated and electrically neutral, eliminating the Coulombic repulsion and favoring assembly through van der Waals, hydrophobic, hydrogen bonding, and π-π stacking interactions [55]. Upon introduction to an acidic environment, the titratable sites become fully protonated, and as a result peptide hydrophobicity, and peptide interaction via van der Waals forces, hydrogen bonding, hydrophobic interactions, and π-π stacking of the cores drive the peptides to assemble to form β-sheet-like aggregates [4, 56]. Ideally, these unassembled units would continue to stack in a ladderlike fashion with a helical twist (Figure 4.1 A and B) ultimately forming long fibers. Although these fibers can be seen experimentally (Figure 4.1D), very little is understood about

Most of this chapter is an excerpt from ref. [210]. Reprinted with permission from Valverde, L. R.; Thurston, B. A.; Ferguson, A. L.; Wilson, W. L. *Langmuir*, 34, 25, 7346-7354 (2018). Copyright 2018 American Chemical Society.

their assembly dynamics. Theoretical models suggest a variety of mechanisms by which these amyloid-like structures may form, with the classic example being an initial nucleation stage followed successive elongation stages [213, 214]. However, reaction speed upon introduction of acid to this system has made characterization of the initial stages difficult.

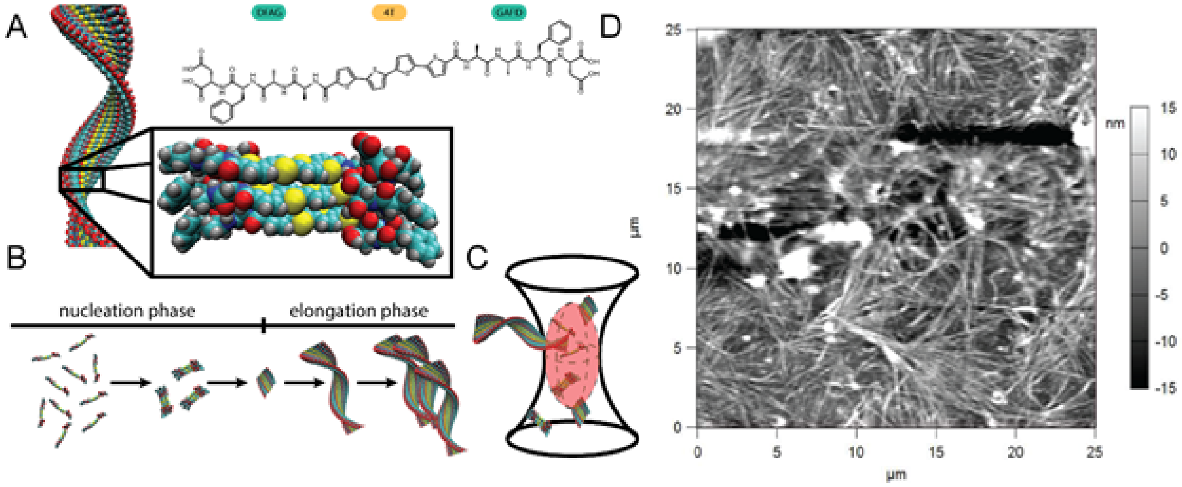

Figure 4.1: Illustration of DFAG-OT4 structure, aggregation, and FCS detection. A) Chemical structure and idealized stacking behavior of DFAG-OT4. B) Standard model for amyloid formation via nucleation-dependent aggregation. C) Confocal spot and observation volume (ellipsoid) used for FCS. As aggregates of various sizes pass into and out of the observation volume fluctuations in fluorescence intensity are detected. D) Atomic Force Microscope (AFM) image of DFAG-OT4 fibers deposited on Si.

We have previously employed molecular dynamics simulations to probe the smaller-scale early time assembly behaviors of DFAG-OPV-GAFD peptides. Simulations have analyzed the impact of peptide symmetry [56], concentration [144], pH, and fluid flow [24] on assembly thermodynamics, kinetics, and morphology. We observed a strongly favorable free energy well at ~15 k_BT for the dimerization of DFAG-OPV3-GAFD peptide in a low pH environment. Addition of each subsequent monomer was found to yield a further decrease in free energy of ~25 k_BT, indicating that monomeric addition of peptides beyond dimerization is increasingly favorable. Our simulations suggest that aggregates at the free energy minima exist as well-aligned stacks with significant π-stacking between cores. In addition, simulations probing the dynamics of assembly of protonated peptides beginning in the monomeric state indicated that peptides rapidly coalesce into spherical micelle-like structures, and then structurally ripen to form the well-ordered β-sheet stacks observed in free energy simulations on times scales larger than several tens of ns. The spontaneous formation of these self-assembled stacks in free energy simulations and the increasingly favorable changes in free energy upon further aggregation agree well with the suspected amyloid-like nucleation and monomeric addition and elongation into larger 1-D fibers that have been observed experimentally [23, 55, 66, 212].

Surprisingly, simulations of DFAG-OPV-GAFD under high-pH conditions also exhibited favorable dimerization with a change in free energy of ~4 k_BT due to hydrophobicity, π-stacking, and dispersion interactions [55]. Furthermore, the formation of higher order structures such as pentamers remain thermodynamically favorable with free energy change of ~5 k_BT [55]. These simulations suggest a paradigm in which early stage assembly consists of light aggregates which rapidly assemble and subsequently reorganize into more thermodynamically stable β-sheet-like structures, which in turn grow and elongate as further oligomeric units are added and structurally relaxed in a low-pH environment. The high-pH assembly predicted furthermore implies that when acid-mediated assembly is induced, the peptide precursor solution exists in a prenucleated state, significantly impacting how one should view the assembly kinetics in acid-mediated assembly experiments.

While most of the low-pH simulation observations support what has already been observed experimentally [23], other work has provided support for the possibility of spontaneous assembly at high-pH [215]. These recent microrheological observations demonstrate peptide assembly only down to concentrations of 0.1 mM with no evidence of assembly below that concentration. This lack of experimental evidence is in large part due to the length scales and numbers of molecules under consideration. Microrheology relies on large, brightly fluorescing probes to correlate observed fluorescence with material properties. The microrheologically observed critical fiber formation concentration is likely due to limitations of the technique rather than actual physical phenomena. Thus, we have worked with experimental collaborators Dr. Bill Wilson and Dr. Lawrence Valverde to employ fluorescence correlation spectroscopy (FCS), a single-molecule technique that allows us to directly detect peptide fluorescence to measure molecule size and to distinguish between low order aggregates of different sizes. We have conducted complementary simulation work to gain molecular-level insight into high-pH assembly and interpret the experimental data.

4.2 Methods

For the generic aggregation process in which molecular species A and B form a complex AB

$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{AB}$

the thermodynamic equilibrium constant can be estimated from molecular simulation as [182, 216]

$$K^{\ominus} = Kc^{\ominus} = \frac{1}{v^{\ominus}\sigma_{AB}} \int_0^{r_b} dr 4\pi r^2 e^{-\beta F(r)} \tag{4.1}$$

where σ_{AB} is the symmetry number (2 for A = B, 1 otherwise), r_b is the center of mass cutoff distance below which an aggregate is considered to have formed, $\beta = (k_B T)^{-1}$, k_B is Boltzmann's constant, T is the temperature, $F(r)$ is the calculated potential of mean force at a center of mass separation value of r, and $\mathrm{c}^{\ominus} = 1/\mathrm{v}^{\ominus}$ is the standard number concentration. The thermodynamic equilibrium constant may be related to the concentrations of the reactants and product as

$$K^{\ominus} = \frac{c^{\ominus}[AB]}{[A][B]} \tag{4.2}$$

where $[X]$ is the number concentration of species X, and it is assumed that the system is sufficiently dilute that concentrations may be used instead of activities. So, by combining Equations 4.1 and 4.2 and given the potential of mean force (PMF) for the aggregation of A and B to form complex AB, we can predict concentration equilibrium constants,

$$K = K^{\ominus} v^{\ominus} = \frac{[AB]}{[A][B]} = \frac{1}{\sigma_{AB}} \int_0^{r_b} dr 4\pi r^2 e^{-\beta F(r)} \tag{4.3}$$

For a system of monomeric self-assembly, by conservation of mass, the concentration of peptide in the system can be expressed as

$$[P] = [M] + 2[D] + 3[T] + \cdots \tag{4.4}$$

where $[P]$ is the total peptide concentration, $[M]$ is the concentration of peptides that exist as monomers, $[D]$ is the concentration of peptides that exist as dimers, $[T]$ is the concentration of peptides that exist as trimers, and so forth. From Equations 4.3 and 4.4, we then have

$$[P] = [M] + 2K_2[M]^2 + 3K_2K_3[M]^3 + \cdots \tag{4.5}$$

where K_2 is the equilibrium constant for the formation of dimers by $\mathrm{M} + \mathrm{M} \rightleftharpoons \mathrm{D}$, K_3 is the equilibrium constant for the formation of trimers by $\mathrm{D} + \mathrm{M} \rightleftharpoons \mathrm{T}$, and so forth. Equation 4.5 defines a polynomial in the peptide monomer concentration that can be solved for [M] and from which all higher aggregate concentrations can be computed using the calculated values of K_2, K_3, and so on [217]. For the peptides and concentrations investigated in this chapter, the equilibrium concentrations of aggregates heavier than six peptides are sufficiently low that the root of the polynomial is insensitive to truncation beyond the sixth term, so it is only necessary to compute equilibrium constants for the hexamers and lighter aggregates. We have verified the insensitivity of the polynomial solution by incorporating terms up to 200 employing extrapolated equilibrium constants, and find that the computed value of $[M]$ changes by less than 0.01%.

We use GROMACS 4.6.7 [164,165] to conduct all molecular dynamics simulations, with the AMBER99SB force field [166, 169], and used the GlycoBioChem PRODRG2 Server [71] to obtain initial peptide geometries. The terminal ASP residues and carboxyl termini were fully deprotonated to simulate a high pH (pH > 5) environment in which each peptide carries a formal (-2)e charge at each terminus [3, 55]. We conducted explicit solvent simulations in TIP3P water [174] with initial velocities generated from a Maxwell-Boltzmann distribution. Electrostatics were treated using the particle mesh Ewald scheme [78] with a cutoff of 1.0 nm and a 0.12 nm Fourier grid spacing. Lennard-Jones interactions were smoothly shifted to zero at a cutoff of 1.0 nm. Bond lengths were fixed using the LINCS algorithm [77], and Lorentz-Berthelot combining rules were used to determine interaction parameters between unlike atoms [79]. The system was integrated using the leapfrog algorithm with a 2 fs time step [180].

Energy minimization was conducted using the method of steepest descents until the maximum force on any atom was less than 1000 kJ/mol·nm. The system was then equilibrated in an NVT ensemble using a stochastic velocity rescaling thermostat [175] to a constant temperature of 298 K. Further simulations were conducted in an NVT ensemble using a NoseÌĄ-Hoover thermostat [178, 179] with a time constant of 0.5 ps.

Following our previous approach [55] we also conduct molecular dynamics simulations in implicit solvent with a modified model that rescales interactions to more accurately match explicit solvent. Polar interactions between solute and solvent are treated with the Generalized Born model while nonpolar interactions are implemented with a solvent accessible surface area approximation [80]. An analytical continuum electrostatic (ACE) type approximation [84] with a value of 2.259 kJ/mol·nm^2 for the surface tension [110] is made in treating nonpolar interactions. We calculated Born radii using the method of Onufriev, Bashford, and Case with a relative dielectric constant of 78.3 and with the standard parameter set of $\alpha = 1$, $\beta = 0.8$, and $\gamma = 4.85$ [83]. Since the peptides are not neutrally charged, implicit solvent simulations are conducted without the use of periodic boundary conditions. Coulombic and Lennard-Jones interactions are smoothly shifted to zero at the large cutoff value of 3.4 nm for the sake of stability.

We employed umbrella sampling [93] to compute the PMF in implicit solvent as a function of center of mass separation between peptide aggregates. To compute the PMF for the formation of an n-mer from an $(n - 1)$-mer and a monomer, the initial geometry of an n-mer aggregate is obtained by stacking n peptides at a core-core separation of 0.45 nm. The system was first equilibrated using the method of steepest descents until the maximum force on any given atom was less than 1000 kJ/mol·nm. Initial velocities of atoms were then drawn from a Maxwell-Boltzmann distribution and the system was equilibrated for 20 ps with the positions of the cores restrained in an NVT ensemble at a temperature of 298 K using a Langevin integrator as a thermostat with a friction constant of 0.5 ps^{-1}, [110] and for another 20 ps with unrestrained cores under the same conditions. The system was then simulated for 1.5 ns and the configuration at the end of each 0.5 ns served as the initial configuration for a series of three independent simulations. Each initial configuration was then pulled both closer together and farther apart at a rate of 0.04 nm/ps using a harmonic biasing potential with a spring constant of 1000 kJ/mol·nm2 between the center of mass of (n - 1) peptides and the center of mass of the remaining monomer. These simulations were run for a sufficiently long time to allow the monomer to reach a distance from the (n - 1)-mer at which the two were no longer able to interact. From these three separate pulling simulations, we then conducted three different umbrella sampling simulations by utilizing configurations over the course of each pulling simulation as the initial geometries for the restrained umbrella sampling. Windows were selected at evenly spaced intervals of 0.1 nm, were restrained using the same harmonic potential as the pulling simulation, and were run for 20 ns each. The first nanosecond of each simulation at each window was discarded to allow the system to equilibrate. We then used the weighted histogram analysis method (WHAM) [94, 95] to reconstruct the unbiased PMF. Statistical errors in each PMF were computed using 100 bootstrap resamples of the data, and sampling errors were computed as the standard of deviation between each of the three umbrella runs. In each case, the $-2k_BT\log(r)$ noninteracting entropic contribution to the PMF was removed in order to avoid double counting this entropic contribution which is already contained in the 4πr2 Jacobian of Equation 4.3 [110, 181].

4.3 Results and Discussion

Computing the PMF for the formation of aggregates of sizes 2-6 by means of monomeric addition at neutral pH (Figure 4.2), we observe free energy changes favoring aggregation on the order of 10 k_BT in each case. The dimerization of two peptides exhibits the largest free energy change at ΔF = (-15.2 ± 1.1) k_BT, while larger aggregates exhibit smaller free energy changes, although the formation of larger aggregates remains thermodynamically favorable. Despite repulsion between negatively charged termini, the minimum free energy configurations for each aggregate size exhibit a high degree of core-core stacking. Aggregates of 4 or fewer peptides also display a high degree of alignment in this stacking and frequently adopt linear stacks of parallel peptides. Aggregates of 5-6 peptides often favor configurations of 2-4 peptides existing in the same well aligned linear stacks with the remaining peptides stacking with one another. These results indicate that hydrophobic and π-π stacking interactions between the conjugated cores mean that it is favorable for peptides to form oligomeric aggregates even at neutral pH where the deprotonated ASP termini mediate substantial electrostatic repulsion.

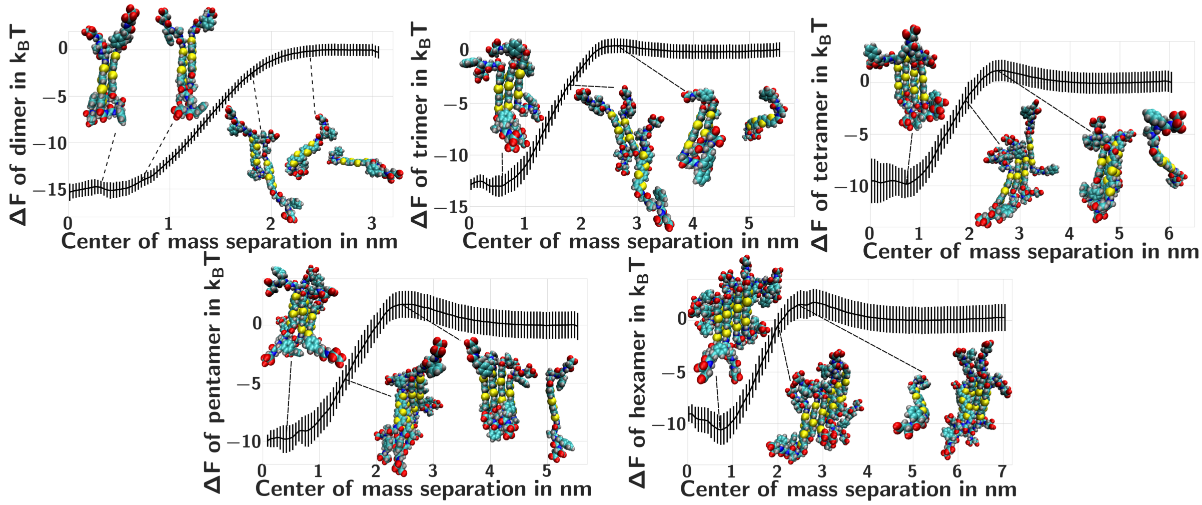

Figure 4.2: Computational prediction for π-π stacked association when deprotonated for (A) dimers, (B) trimers, (C) tetramers, (D) pentamers, and (E) hexamers with representative configurations of aggregates at various points along the reaction coordinate.

The thermodynamics of self-assembling systems involve a nontrivial interaction between competing interactions [211,218]. Different interactions including hydrogen-bonding, π-π interactions, hydrophobic interactions, and entropy all contribute to the thermodynamics governing peptide assembly. In order to more fully understand some of these contributions to the aggregation of our system, we follow a similar approach to Chapter 2 [55] and break the free energy of aggregation down into constituent components. In the implicit solvent systems studied, the change in free energy for the formation of an aggregate of size n may be written as

$$\Delta F_n = \Delta U_n^{\text{intrapeptide}} + \Delta U_n^{\text{peptide-peptide}} + \Delta U_n^{\text{peptide-water}} + \Delta U_n^{\text{water-water}} - T\Delta S_n \tag{4.6}$$

where $\Delta U_n^{\text{intrapeptide}}$ is the change in intramolecular peptide energy upon aggregation (including changes in intramolecular Lennard-Jones and Coulombic interactions, as well as angular stretching and dihedral torsions), $\Delta U_n^{\text{peptide-peptide}}$ is the change in intermolecular interactions between peptides upon peptide association, $\Delta U_n^{\text{peptide-water}}$ accounts for the change in dispersion and electrostatic interactions between peptide and solvent, $\Delta U_n^{\text{water-water}}$ is the change in energy due to solvent-solvent interactions, T is the temperature, and ΔS_n accounts for the change in entropy of the system on aggregation. To elucidate different contributing factors, we divide peptide-peptide interactions into their Lennard-Jones and Coulombic components. The entropic contribution may be divided into changes in solvent entropy and changes in peptide entropy. Grouping all solvent related terms together, we define

$$\Delta F_n^{\text{solvent}} \equiv \Delta U_n^{\text{peptide–water}} + \Delta U_n^{\text{water–water}} - T\Delta S_n^{\text{water}} \tag{4.7}$$

Assuming that the peptide configurational entropy does not change substantially upon aggregation allowing us to neglect the entropy change of the peptides [55], we then have

$$\Delta F_n \approx \Delta U_n^{\text{intrapeptide}} + \Delta U_n^{\text{peptide–peptide–LJ}} + \Delta U_n^{\text{peptide–peptide–Coulomb}} + \Delta F_n^{\text{solvent}} \tag{4.8}$$

The change in free energy ΔF_n on the left-hand side is precisely the well depth of the PMF computed by umbrella sampling. The three energetic terms on the right-hand side $\Delta U_n^{\text{intrapeptide}}$, $\Delta U_n^{\text{peptide–peptide–LJ}}$, and $\Delta U_n^{\text{peptide–peptide–Coulomb}}$ can be computed directly from our simulations from the energies of the various aggregate sizes averaged over 20 ns unbiased MD simulations. The solvent contributions ΔF_n solvent follow from the residual on the right-hand side of Equation 4.8.

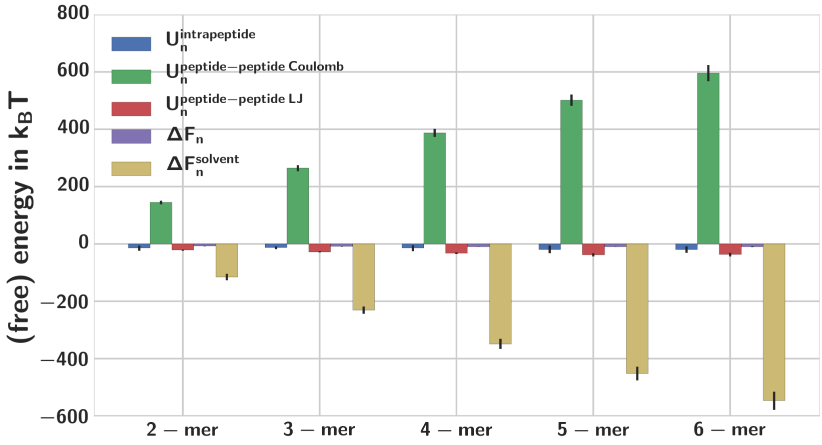

Figure 4.3: Decomposition of the free energy of association ΔF_n into energetic and solvent-mediated contributions for n = (1-6)-mers (Equations 4.6-4.8). The strongly unfavorable Coulombic repulsion $\Delta U_n^{\text{peptide–peptide–Coulomb}}$ is balanced by a favorable solvent-mediated term $\Delta F_n^{\text{solvent}}$.

We illustrate the results of this analysis for peptide aggregates ranging from two to six peptides in Figure 4.3. For each aggregate size we observe small favorable contributions in both intrapeptide interaction and LJ interaction between peptides. As anticipated, the most significant unfavorable contribution is due to Coulombic repulsion, but that this is balanced by a large favorable solvent contribution and smaller favorable dispersion and intrapeptide energetic contributions. We do observe that this decomposition was computed for an implicit solvent model, and that a more detailed analysis would employ a fully explicit solvent model with a polarizable force field.

From Equation 4.3, the equilibrium constants for the formation of aggregates are computed from these PMFs along with 90% confidence intervals (Table 4.1).

Table 4.1: Equilibrium Constants for the Formation of an Aggregate of Size n from a Tightly-Bound (n - 1)-mer and a Monomer

Aggregate size	Equilibrium Constant, K (M^{-1})	90 % CI, (M^{-1})
2	1.8×10^6	(5.1×10^5, 6.4×10^6)
3	6.6×10^5	(1.2×10^5, 3.9×10^6)
4	4.7×10^4	(4.9×10^3, 4.7×10^5)
5	2.8×10^4	(3.9×10^3, 2.1×10^5)
6	1.0×10^5	(9.7×10^3, 1.1×10^6)

Confidence intervals are estimated by randomly generating 10^6 PMFs by shifting each point on the PMF by the product of the bootstrap error at that point with a single number randomly generated from a Gaussian distribution with zero mean and unit standard deviation. Each PMF is then integrated over the binding region to obtain 10^6 different values for each equilibrium constant, the middle 90% of which defines the confidence interval.

Despite the favorable PMFs, low overall peptide concentrations favor light aggregate distributions. From Equations 4.3 and 4.5, we calculate the predicted distribution of aggregate sizes in deprotonated peptides from the computed PMFs based on the overall peptide concentration (Figure 4.4). Error estimates are obtained by randomly sampling equilibrium constants within the 90% confidence intervals. We predict that at a concentration of 10 nM the vast majority (~96%) of the peptides tend to exist as isolated monomers. When the peptide concentration is increased to 100 nM we observe a significant shift in the distribution of peptide sizes that indicates an appreciable amount of aggregation of peptide into larger aggregates, including dimers (22%) and trimers (2%). Such a transition is in qualitative agreement with what is observed experimentally [210].

Given the above observations, it is reasonable to generalize our findings to any acid-mediated system that also relies upon hydrophobic and/or π-π interactions for self-assembly. In any such case, the driving forces for assembly always exist and the role of protonation is to further shift the thermodynamic equilibrium in favor of assembly by eliminating electrostatic counterforces. Thus, we expect any acid-triggered system with synergistic avenues for self-assembly such as hydrophobic, π-π, or other van der Waals interactions to be in actuality an acid-mediated system beginning in a prenucleated state.

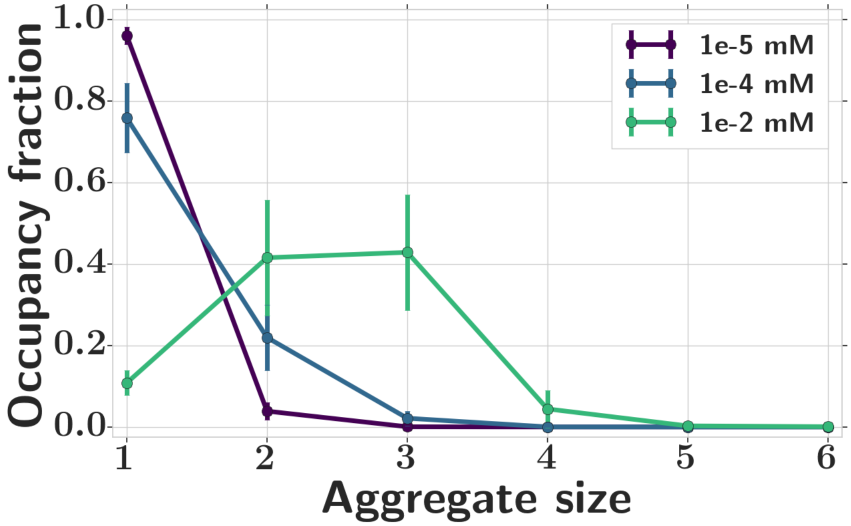

Figure 4.4: Predicted fraction of peptide existing in various aggregate sizes at peptide concentrations of 10 nM (green) and 100 nM (blue). Lines are drawn to guide the eyes. Error bars are estimated by random sampling of K values within the 90% confidence interval.

4.4 Conclusions

We have conducted molecular dynamics simulations to confirm the consistency of assembly behavior between previously computed DFAG-OPV3-GAFD peptides and their DFAG-OT4-GAFD cousins. These simulations demonstrate that, for both materials, not only is the macroscopically observable assembly triggered by lowering a solution's pH, but also even in nonprotonating environments some degree of aggregation is thermodynamically favorable. Single molecule measurements using fluorescence correlation spectroscopy provide experimental support for these computational predictions [210]. We find that aqueous solutions of peptides in concentrations as low as 100 nM will spontaneously aggregate to form heterogeneous solutions. However, below 100 nM, solutions appear to be homogeneous solutions of largely unassembled monomer. These results indicate that previously assumed paradigms of acid-triggered assembly in this system whereby monomer aggregates upon protonation were incomplete. In fact, the system only exists as pure monomer in very low concentrations, and under experimental conditions the high-pH un-triggered solution already exists in a prenucleated state.

Chapter 5

Design of spectroscopic properties

5.1 Introduction

In this chapter we present ongoing work looking at relating the structural properties of peptide aggregates to the excited state properties that can be measured in experiment. Our experimental collaborators – J.D. Tovar and Howard Katz at Johns Hopkins University – have recently conducted a study looking at alterations in excited state properties of assembled DXX-OT4-XXD peptides where OT4 is a quaterthiophene π-conjugated insert and X in {A, F, G, I, V} (Figure 5.1) [2]. Each peptide studied caused different shifts in the peak excitation wavelength of the absorption spectrum as compared with the unassembled system. The shifts for DGG and DAA were the most pronounced, leading to a blue-shift upon aggregation of ~50 nm as compared with ~10 nm for other peptides. The photoluminescence (PL) spectrum also showed similar differences wherein each peptide composition revealed shifting and quenching of the PL spectrum upon assembly, but DGG and DAA peptides led to an almost complete quenching of the spectrum as compared with the others. These significant differences seem to indicate that the size of the amino acid side chain could play a significant role in determining the photophysical properties of peptide aggregates. The precise dependence of these properties on the geometric configurations adopted by the peptides, however, remains unknown.

We conduct fully atomistic molecular simulations to probe both the geometric properties and spectroscopic properties of peptide aggregates as a function of chemistry. Due to the long relaxation times for peptide aggregates (see Chapter 2), we elect to study aggregate structure with purely classical MD simulations, and use the configurations obtained from them for time-dependent density functional theory (TDDFT) computation of the absorption spectrum. Such information is useful to understanding the geometric origin of excited state properties observed in experiment. Furthermore, the determination of simple relationships between geometric properties of peptides observed in MD simulation and the corresponding spectral properties enables the prediction of these properties directly from MD, allowing for prediction of spectral shifts for previously untested peptide systems. Coupling such results with our previous work

(Chapter 3) can ultimately enable the computational design of self-assembling peptides having specific properties.

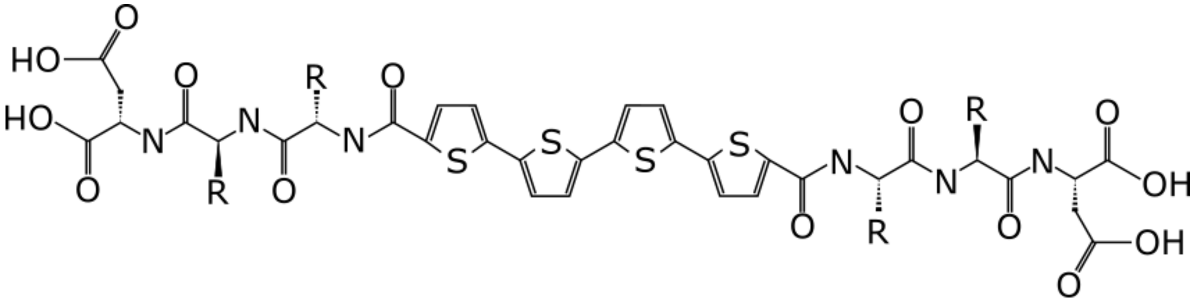

Figure 5.1: Peptide compositions studied in this chapter, where all R groups are the same and include A, F, G, I, and V, for a total of five different peptides.

5.2 Methods

5.2.1 Molecular dynamics simulations

We conduct molecular dynamics (MD) simulations for peptides of each composition, using fully protonated ASP residues (low pH environment). Based on our previous work, we also conduct simulations on fully deprotonated ASP residues (high pH environment) containing four peptides, since we expect the measured absorption spectra to contain some smaller pre-assembled peptide aggregates (Chapter 4) [210].

Simulations are conducted using GROMACS 4.6.7 [164, 165] using the AMBER99sb [98, 166] force field. Partial charges for OT4 residues were obtained using the RESP/ESP charge Derive Server (REDS) [171] which utilized Gaussian for Hartree-Fock calculations [173] (see refs [170, 172] for details on the method). Non-bonded parameters were obtained by analogy using the parmchk2 program from Antechamber [167] and the generalized Amber force field (GAFF) [168]. Due to the slow aggregation times for assembly of larger aggregates from disperse solution (Chapter 2), we initialize all simulations in a partially pre-assembled state. Peptides are stretched out to a maximum distance of separation between the α-carbons in the ASP residues on either side of the peptide. These monomers are then replicated to form a linear stack of peptide monomers with adjacent peptides separated by 0.45 nm, near the expected equilibrium distance between peptides in an aggregate [55]. Each peptide is rotated a small amount with respect to its nearest neighbors ($\frac{\pi}{20}$ rad) in order to help avoid steric clashes between adjacent peptides.

This system is then solvated in a rhombic dodecahedron with water using the TIP3P model [174]. The box is made to be large enough that the all peptides are at least 1.0 nm from the edge of the box. In high pH simulations, Na^+ counterions are added to make the whole system neutrally charged when necessary. The system is then minimized using the steepest descent algorithm until the maximum force on any atom was less than 1000 kJ/mol.nm. Initial velocities were obtained by sampling a Maxwell distribution at 298 K. The system is subject to a 100 ps NVT equilibration run in which the positions of atoms in the cores were restrained. Temperature was maintained by a stochastic velocity rescaling thermostat [175] with a time constant of 0.5 ps. The system then underwent a 100 ps NPT equilibration run with cores still restrained using the same thermostat. Pressure was equilibrated to 1.0 atm using the Berendsen pressure coupling scheme [219] with a compressibility of $4.6{\times}10^{-5}$ bar^{-1} and a time constant of 2.0 ps [110]. Electrostatic interactions were treated using Particle Mesh Ewald (PME) with a cutoff of 1.0 nm and a 0.12 nm Fourier grid spacing that were optimized during runtime [78]. Lennard-Jones interactions were shifted smoothly to zero at 1.0 nm. Bond lengths were fixed using the LINCS algorithm [77], and Lorentz-Berthelot combining rules were used to determine interaction parameters between unlike atoms [79]. A leap frog algorithm with a time step of 2 fs was used to integrate the system [180]. Initially, we run one system of 20 protonated peptides for each chemistry for 150 ns.

5.2.2 Core contributions

Even using less expensive methods for probing the absorption spectrum proves to be computationally infeasible for larger peptide aggregates. The peptide aggregates probed in Section 5.2.1 contain upwards of two thousand atoms, a single calculation of which is beyond our computational capacity. Furthermore, we wish to incorporate several different configurations observed across the MD runs for each of the five different peptides compositions. Accordingly, it is necessary to make approximations to reduce the system to a tractable size for study by TDDFT methods. We conduct this system size reduction using two fundamental approximations: (i) that the experimental trends observed in Reference [2] will be reproducible when observing systems containing fewer peptides (this is equivalent to assuming the absorption spectrum of a whole aggregate will be related to the absorption spectrum of the sum of its parts), and (ii) that the spectral response is governed primarily by the electronic transitions within the OT4 π-conjugated cores, and the influence of the peptide wings upon the spectra is only to modulate the morphological stacking of the cores. Under these two assumptions, we predict the spectral response of the oligopeptide aggregates by conducting TDDFT analyses of small OT4 stacks extracted from MD simulations of the full peptides. The first approximation is absolutely necessary; due to the expense of TDDFT calculations it is not feasible to perform many computations on thousands of atoms. The second approximation may be justified by several observations. First, the unassembled absorption spectrum is known from experimental work not to depend strongly on amino acid composition [2]. Therefore, the amino acid wings do not significantly alter the absorption spectrum on their own. Second, the experimental results appear to indicate that the differences in the absorption spectra are correlated with the size of the amino acids. This could be indicative of bulkier residues altering the stacking configurations of the peptide cores. Third, we expect the most significant contributions to the excited state properties of the system to come from the semiconducting quaterthiophene cores .

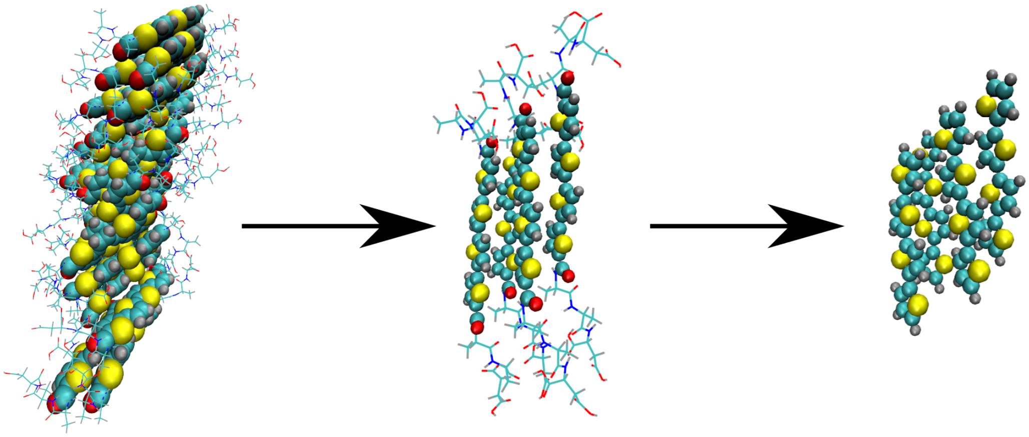

Figure 5.2: Pictorial summary of the two fundamental approximations made in determining the molecular configurations to study with LR-TDDFT. The first step involves the approximation that trends highlighting the differences between chemistries will still be visible when computing excitation wavelengths for a small subset of molecules. This amounts to the approximation that the absorption spectrum of the whole will be the sum of the absorption spectra of its parts. The second step involves the assumption that the most significant impact the peptide wings have on the spectroscopic properties of the system is the extent to which they influence the stacking geometries of the π-conjugated cores and may therefore be neglected in the TDDFT calculations.

5.2.3 Linear response time dependent density functional theory

Selected configurations explored by the system during MD simulation were extracted and subjected to electronic structure analysis using TDDFT. There are a number of methods commonly used for computing excited state properties of molecules a few of which include, the configuration interaction-singles (CIS) approach [220], time-dependent density functional theory (TDDFT) [64,221–223], the Bethe-Salpeter Green's function perturbation approach (BSE) [224–228], and the equation-of-motion coupled-cluster method (EOM CC) [229–231] (see References [232–234] for more details). Among the many available methods, TDDFT is a very commonly used method that provides a good tradeoff between accuracy and computational efficiency [232, 235–240]. Linear-response TDDFT (LR-TDDFT) is a common approach within the

TDDFT formalism for studying the absorption properties of a molecular system [223,238,239]. LR-TDDFT amounts to solving for the smallest eigenvalues of a matrix containing transitions between different states obtained in standard DFT. These eigenvalues give the excitation energies for the absoprtion spectrum. Their corresponding osccilator strengths may be determined from the eigenvectors [223, 238]. It is possible to improve upon the results obtained in LR-TDDFT by incorporrating transitions between different vibrational modes in the molecules [64,233,241–243]. This method, however, requires the optimization of the geometry of the excited state system, and the subsequent sampling of its normal modes in order to obtain their associated frequencies, both of which can be very computationally intensive for larger systems. As a result, this method is too expensive for our purposes, so we elect to utilize basic LR-TDDFT formalism to study the absorption spectrum of these peptides. Our approach is then to extract core configurations from MD simulation and then use LR-TDDFT to compute the roots and oscillator strengths of the excited states, from which we can estimate the wavelength at which the absorption spectrum is at a maximum.

We employ NWChem version 6.6 for all TDDFT calculations [244], using all default parameters unless otherwise noted. We use the hybrid B3LYP functional for all of our calculations [245–247] along with the 6-31G* basis set [248,249]. This functional and basis set have been shown to be quite accurate [235,239,246, 250,251]. The “fine” grid option is selected for DFT. All computations are conducted in the gas phase (as isolated molecules). The 40 lowest lying excited singlet-singlet transitions are computed in each computation. The peak excitation wavelength is obtained from by broadening the roots and oscillator strengths with gaussians having a full width half max of 50 nm.

5.2.4 Geometric model

In order to probe the geometric properties governing the computed TDDFT spectra, construct models to predict the wavelength of maximum absorbance λ_{max} in the computed LR-TDDFT spectra from structural properties of the oligopeptide aggregates. This approach may be considered to be a form of quantitative strcuture-property relationship (QSPR) modeling. We recall that we explicitly consider only the OT4 π-conjugated cores in our calculations of the absorbance spectra, so the spectral differences are attributable exclusively to the spatial arrangement of these cores within the assembled aggregate. The effect of the peptide wings–and therefore chemical differences between the peptides–is treated implicitly through their influence on the stacking of the cores. We construct QSPR models to predict λ_{max} from (a subset of) the following easily-computable and physically-interpretable structural features of the oligopeptide aggregates. In this manner we seek to construct a surrogate model that can be used to efficiently predict λ_{max} from the geometric structure of an aggregate without running expensive first principles calculations, and also provide mechanistic insight into the key structural determinants of λ_{max}. (i) The first, second, and third nearest neighbor distances between cores (nn_1, nn_2, nn_3). (ii) We define the alignment angle between two adjacent cores to be the angle between the vectors parallel to the peptide backbone, and utilize the angle between the first, second, and third nearest neighbors (θ_1, θ_2, θ_3). (iii) We define the tilt between two adjacent cores to be the angle between the vectors normal to the central thiophene rings and compute the tilt angle between first, second, and third nearest neighbors (ϕ_1, ϕ_2, ϕ_3). (iv) The radius of gyration (R_g), defined as

$$R_g = \left(\frac{\sum_i ||\mathbf{r_i}||^2 m_i}{\sum_i m_i}\right)^{\frac{1}{2}} \tag{5.1}$$

(v) We use lienar regression to fit the coordinates of the atoms to the equation of a plane and compute the RMS deviation of the atoms from that plane ($\text{RMS}_{\text{plane}}^{\text{core}}$). (vi) We follow the same approach to compute the RMS deviation of the atoms in each thiophene ring from a plane ($\text{RMS}_{\text{plane}}^{\text{ring}}$). (vii) We use the normal vectors of each of these planes of best fit for individual thiophene rings and compute the mean angle between the planes of adjacent cores (ψ^{ring}). (viii) We also consider the Steinhardt bond order parameters [252–254]. The Steinhardt bond order parameter of order j is given by

$$Q_l \equiv \left[\frac{4\pi}{2l+1}\sum_{m=-l}^{l}\left|\overline{Q_{lm}}\right|^2\right]^{1/2} \tag{5.2}$$

where $\overline{Q_{lm}}$ is

$$\overline{Q_{lm}} \equiv \frac{\sum_{i=1}^{N}\sum_{j=1}^{N_b(i)} Y_{lm}(\hat{r}_{ij})}{\sum_{i=1}^{N} N_b(i)} \tag{5.3}$$

N gives the number of particles in the system, i and j are used as indices indicating individual particles, $N_b(i)$ is the number of particles within a given cutoff distance of particle i, $\hat{r}_{ij}$ is the unit vector going from particle i to particle j, and Y_{lm} are the spherical harmonics. Accordingly, $\overline{Q_{lm}}$ is the average sum of the spherical harmonics of the unit vectors connecting a given particle to its nearest neighbors, averaged over all particles in the system. The Q_l define a rotationally invariant global parameter of the system which can be used to distinguish between different particle stacking schemes [253]. They are nonzero only for even values of $l > 2$. We find that higher orders of the Q_l become increasingly invariant over the different core configurations we observe. Accordingly, we use the parameters for $l = 4, 6$ for our model (Q_4, Q_6).

We utilize elastic net linear regression [195] to construct a model to fit the peak excitation wavelength from the geometric parameters we have defined. We expect that some of the geometric parameters that we proposed will not prove to be useful in determining peak excitation wavelength, so the regularization terms of elastic net regression will enable us to determine which parameters are and are not useful, and help us to avoid overfitting our data. Elastic net regression is a form of linear regression that employs regularization to penalize both the L1 and L2 norms of the coefficients of the by minimizing the objective function [255]

$$\min_{w}\left(\frac{1}{2n}\|Xw-y\|_2^2+\alpha\rho\|w\|_1+\frac{\alpha(1-\rho)}{2}\|w\|_2^2\right) \tag{5.4}$$

where w is the set of coefficients being optimized, X is the matrix of features used in the fit (here the structural descriptors listed above), y is the vector of data to be fit (here λ_{max}), n is the number of samples, and ρ and α are hyperparameters determining the regularization. The complexity and L1 ratio hyperparameters α and ρ control the absolute and relative strengths of the L1 and L2 regularization penalties. Elastic net regression subsumes both pure LASSO regression ($\rho = 1$, L1-norm only) and pure ridge regression ($\rho = 0$, L2-norm only). These hyperparameters are tuned using leave-one-out cross-validation (LOO-CV) over the training data set.

5.3 Results

5.3.1 Wing removal

In order to validate our assumption that the absorption spectra is governed largely by the optoelectronic response of the π-conjugated cores and that the peptide wings may be neglected in these calculations, we use LR-TDDFT to compute the absorption spectra of selected isolated peptides and peptide dimers with and without the peptide wings (Table 5.1). We determine removing the peptide wings does lead to a blue-shift in the absorption spectrum of about 30 nm. This shift, however, is revealed to be independent of peptide composition. This indicates that while this approximation may change the absolute position of the absorption peak, the relative differences caused by differences in composition will be preserved. This serves as *a posteriori* validation of our second approximation at the level peptide monomers and dimers where this comparative calculation is computationally tractable. If we were to apply this shift as a correction to all of the results we obtain for λ_{max} from TDDFT it would cause our results to be much further from the peak excitation wavelengths measured in experiemnt. This likely indcates there is cancellation between different sources of error that enables our computations to be as close to experiment as they are (see Table 5.3). Finally, since we are concerned only about the relative differences between absorption spectra of different peptides, we neglect any shift in the spectra caused by removing peptide wings.

Table 5.1: Shift in peak excitation wavelength of absorption spectrum of selected isolated peptides and peptide dimers with and without the peptide wings. Shifts are approximately independent of wing composition within error, indicating that the absolute position of the peak of the absorption spectrum will be affected by this approximation, but the relative differences between different compositions will be preserved.

Composition	Shift in peak excitation wavelength in nm
DAA-OT4-AAD (monomer)	31.3
DAA-OT4-AAD (dimer)	31.3
DFF-OT4-FFD (monomer)	32.5
DGG-OT4-GGD (monomer)	34.9
DII-OT4-IID (monomer)	30.7
DVV-OT4-VVD (monomer)	31.3

5.3.2 Peak excitation model

We adopt a hierarchical multi-scale approach in which classical MD simulations supply self-assembled peptide configurations for study by LR-TDDFT. MD simulations contain 20 peptides each and are run for 150 ns. We one run MD simulation for each of the five diffvery peptides identified in Figure 5.1. We harvest three different frames at 100, 110, and 120 ns from each MD run. Due to the computational cost of LR-TDDFT, we are unable to directly calculate the absorption spectrum of each 20-peptide stack directly, and so we partition the stack into five sets of contiguous clusters of four peptides, and compute the absorption for each stack. A four core system is the largest that we can comfortably study with the resources available to us. This procedure is motivated by the high computational cost of LR-TDDFT and assumes that the spectral response of the system may be adequately modeled by the mean response of data obtained from tetrapeptide stacks.

Elastic net regression is then conducted to fit all computed QSPR models of λ_{max} predicted from the ensemble of structural features detailed in Section 5.2.4. Data are divided into training and testing data sets by random selection of 80% of the data for the former and the remaining 20% for the latter. Values of the selected descriptors are normalized (Z-scored) by shifting the mean and standard deviation of each over the training data set to zero and one respectively. No descriptors were found to be highly correlated ($\rho > 0.9$) with other descriptors. Hyperparameters of the elastic net regression are tuned using LOO-CV on the training data set using the ElasticNetCV method of scikit-learn [255]. The performance of the model is then assessed by its ability to predict the peak excitation wavelengths of the absorption spectrum for the testing data set (Figure 5.3). We find that the model is able to accurately predict λ_{max} for both the training data set (RMSE = 3.9 nm) and the testing data set (RMSE = 5.4 nm).

The coefficients of each descriptor within the linear regression model indicate which play the most important role in relating structural properties to peak excitation wavelengths. In order to systematically

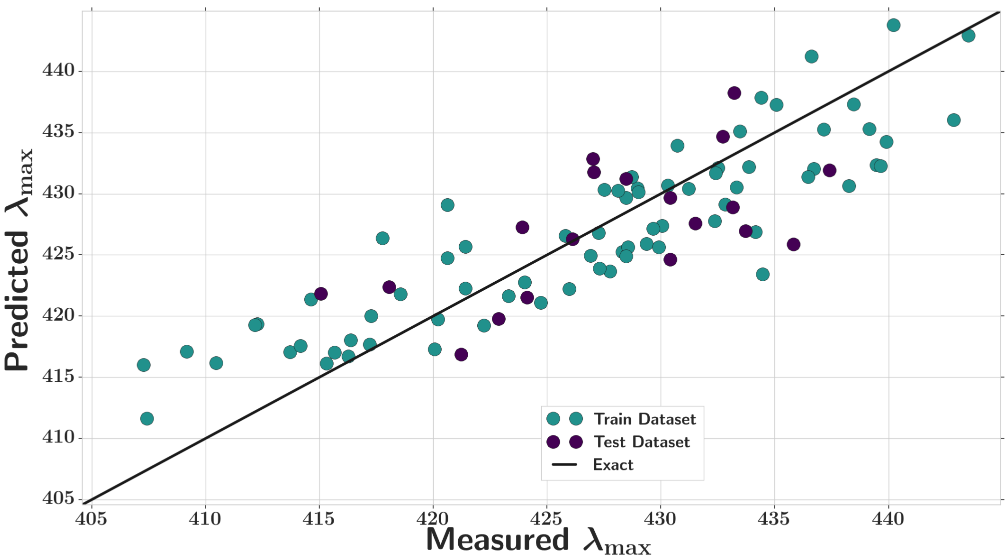

Figure 5.3: Elastic net regression fit applied to self-assembled aggregates containing four peptide cores. Despite the simplicity of the linear regression model, we find that it performs quite well in quantitatively predicting the value of λ_{max} computed by LR-TDDFT, possessing an RMS error of 4.3 nm in the fitting the training data set and 4.7 nm in fitting the testing data set.

quantify this, we conduct 50 random divisions of the data into training and testing sets and observe the variation in these coefficients (Figure 5.4). Over all 50 random divisions, we obtain a mean RMS error of *4.5 ± 0.2) nm over the training data set and a mean RMS error of (5.0 ± 0.7) nm over the testing data set. The complexity parameter α takes a mean value of (0.64 ± 0.15), while the L1 ratio parameter ρ takes a mean value of (0.92 ± 0.15), indicating the fit is close to pure LASSO regression.

Over this series of fits, the only parameters that are systematically nonzero are the first, second, and third nearest neighbor distances (nn_1, nn_2, nn_3), the angle between the axes of nearest neighbor peptides (θ_1), and the radii of gyration of the peptide cores (R_g). The signs of these results indicate that smaller distances between cores, smaller angle between nearest neighbors, and more extended peptide cores all lead to larger blue shifting of the spectrum, and these structural properties can be controlled through the amino acid sequence of the peptide wings. As a result of this analysis, we construct a least-squares linear regression model in these five features over the full training data to obtain our final QSAR model,

$$\lambda_{\text{max}} = 2.5nn_1 + 2.3nn_2 + 1.4nn_3 + 1.2\theta_1 - 2.3R_g + 428.0 \tag{5.5}$$

This model applies after Z-scoring the descriptors to fix their mean value to zero and standard deviation to one employing parameters detailed in Table 5.2.

Table 5.2: Values used to z-score the descriptors.

Parameter	Z-Score Mean	Z-Score Standard Deviation
nn_1	0.51	0.07
nn_2	0.78	0.15
nn_3	1.06	0.21
ang_1	17.6	7.6
r_g	0.455	0.001

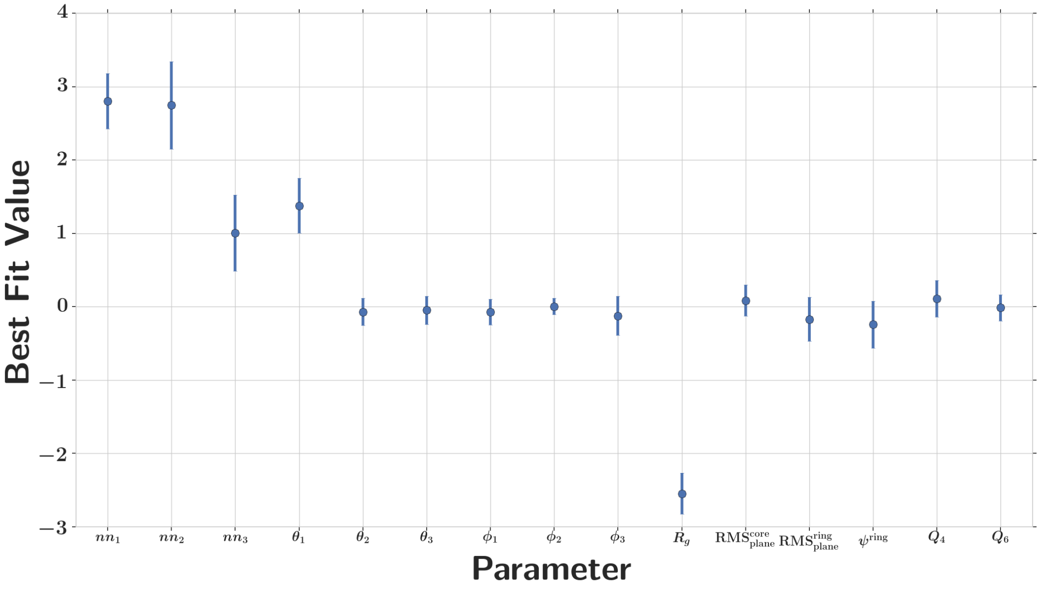

Figure 5.4: Mean elastic net coefficients for each geometric parameter over several random divisions of data into training and testing sets. Results indicate that nearest neighbor distances (nn_1, nn_2, nn_3), angle with nearest neighbor (θ_1), and radius of gyration (R_g) are the only consistently meaningful parameters in fitting the results.

5.3.3 Peak excitation model application and evaluation

We test out model by applying it to independent MD simulations of the self-assembly of each of the five peptide chemistries in systems containing 4, 10, 20, and 40 protonated peptides, and 4 deprotonated peptides. The former simulations model aggregation of increasingly larger systems under acidic conditions, and the latter the assembly of small oligomers under high-pH conditions (Chapter 4). Runs for four deprotonated peptides and for four and twenty protonated peptides are run for 300 ns while others are run for 150 ns due to time constraints. All simulation procedures are the same, except in the case of the 40 peptide runs, which are initialized as two parallel stacks of twenty peptides each rather than a single stack of forty peptides. This serves to significantly decrease the size of the box necessary to contain the peptides.

We can then apply our regression model to each MD run. The set of parameters used in the model takes a different set of values for each peptide contained in the system. Accordingly, each peptide within a run provides a different prediction for the value of the peak excitation wavelength. We calculate this value for each peptide in the system at a given time window and average over all peptides to obtain an estimate for $\lambda_{\max}$ at that time. We apply this for every step of the MD run to obtain estimates as a function of simulation time (Figure 5.5). We can the average over the final 50 ns of each MD run in order to obtain an estimate of the peak excitation wavelength for a system of peptides at that size (Table 5.3).

Smaller system sizes are difficult to distinguish from one another within error bars. This fits with experimental results in that unassembled, deprotonated peptides all exhibited the same absorption spectrum, independent of peptide composition. The exception is that DII-OT4 peptides are predicted to have a much higher peak excitation wavelength than any of the other peptides for both protonated and deprotonated tetramers. We expect that this is indicative of an unwaranted extrapolation of the model resulting from fundamentally different configuration observed in these runs than were seen in the training data set (see Figure 5.6). It is our immediate goal to conduct LR-TDDFT calculations for the DII-OT4 system and incorporate these data into our QSPR model to increase its predictive power.

Larger systems all place DAA and DGG at lower peak excitation wavelengths than the other three peptides, which also is in qualitative agreement with experiment, though the magnitude of the difference between these two sets is not nearly as large as in experiment. The model also consistently predicts the correct ordering of DVV and DII peaks relative to one another for larger runs.

Table 5.3: Mean predicted peak excitation wavelength $\lambda_{\max}$ in nm over the last 50 ns of simulation for aggregates of different sizes and compositions. Table rows are in order of increasing peak excitation wavelength as measured experimentally. Experimental values are taken from Figure 9 from Reference [2], and experimental errors are the estimated error in locating the peak value of the data.

	4 high pH	4 low pH	10 low pH	20 low pH	40 low pH	Experiment
DGG-OT4	422.9 ± 3.9	429.5 ± 3.4	416.7 ± 1.6	420.6 ± 1.1	417.5 ± 0.7	361 ± 3
DAA-OT4	421.2 ± 3.4	418.1 ± 2.4	422.5 ± 1.7	419.9 ± 1.0	419.9 ± 0.7	365 ± 3
DVV-OT4	421.6 ± 3.1	421.2 ± 2.5	428.6 ± 1.8	421.1 ± 1.0	423.8 ± 0.8	404 ± 3
DII-OT4	440.4 ± 4.5	439.9 ± 2.8	426.2 ± 1.8	424.9 ± 1.1	426.4 ± 0.9	408 ± 3
DFF-OT4	427.3 ± 5.5	424.8 ± 2.3	423.9 ± 1.7	425.0 ± 1.1	421.2 ± 0.7	411 ± 3

To further test the predictive capabilities of this model, we perform TDDFT calculations on sets of six peptides extracted from the same MD runs, as well as on sets of four peptides extracted from runs containing only four peptides (Figure 5.6). Our performs reasonably well for predictions involving six peptides (RMSE = 7.2 nm), though it is difficult to say anything definitive as a result of the small sample size due to the expense of such computations. Our model does tend to underestimate the extent of the blue shift in the spectrum at lower values of peak excitation wavelength in that it systematically predicts values of $\lambda_{\max}$ that are too high for lower computed values of $\lambda_{\max}$. This trend may be accentuated for systems containing

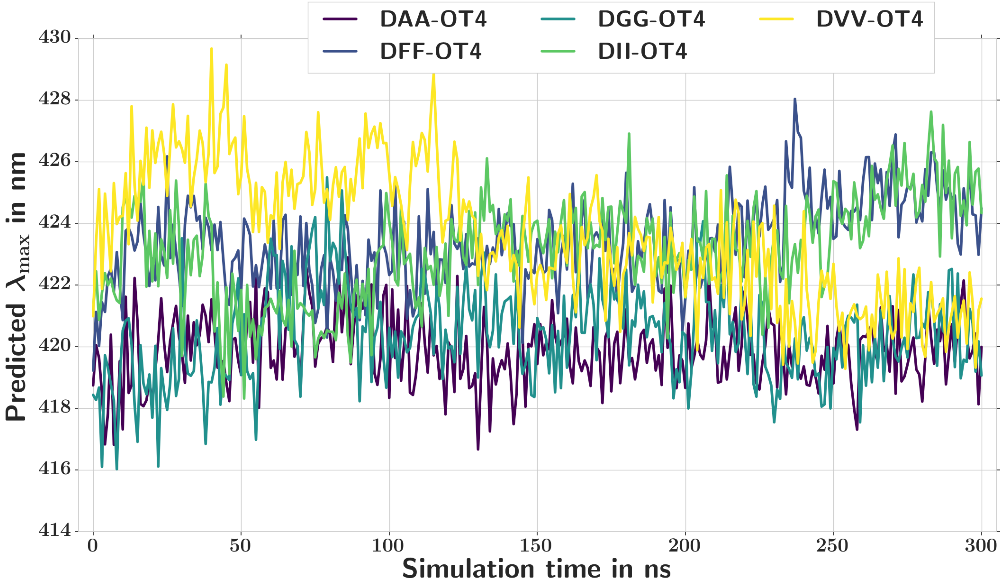

Figure 5.5: Application of the QSPR regression model to predict the peak excitation wavelength of peptides as a function of simulation time for simulations containing 20 fully protonated peptides of a given composition. Qualitative ordering of peaks is correct relative to experiment with DAA and DGG having the lowest peak excitation, followed by DVV, with DII and DFF having the highest values of $\lambda_{\max}$.

six peptides. The model's predictive accuracy falls significantly when applied to tetramers extracted from tetramer simulations (RMSE = 11.2 nm). This indicates that the configurations that tetramers adopt in a lone tetramer simulation are somewhat different from those that experience the forces applied by other peptides in a larger aggregate. This is an interesting result, but not necessarily a cause for concern since the intended application of our model is in predicting the absorption behavior in self-assembled aggregates of large numbers of peptides.

5.4 Conclusions and future work

We conducted MD simulations of peptides having five different compositions, from which we extracted configurations for LR-TDDFT calculations. We used these data to train a QSPR model to predict the wavelength of maximum absorption λ_{max} from simple structural properties of the oligopeptide aggregates to gain insight into the underlying relationship between aggregate structure and optoelectronic response. Encouragingly, we see significant agreement with qualitative trends and rank ordering observed in experiment suggesting that our model is robustly identifying some of the key factors determining the observed spectral responses. We also find evidence validating the assumption that primary impact alterations in amino sequence have on the absorption spectra of peptide aggregates results from the alterations induced

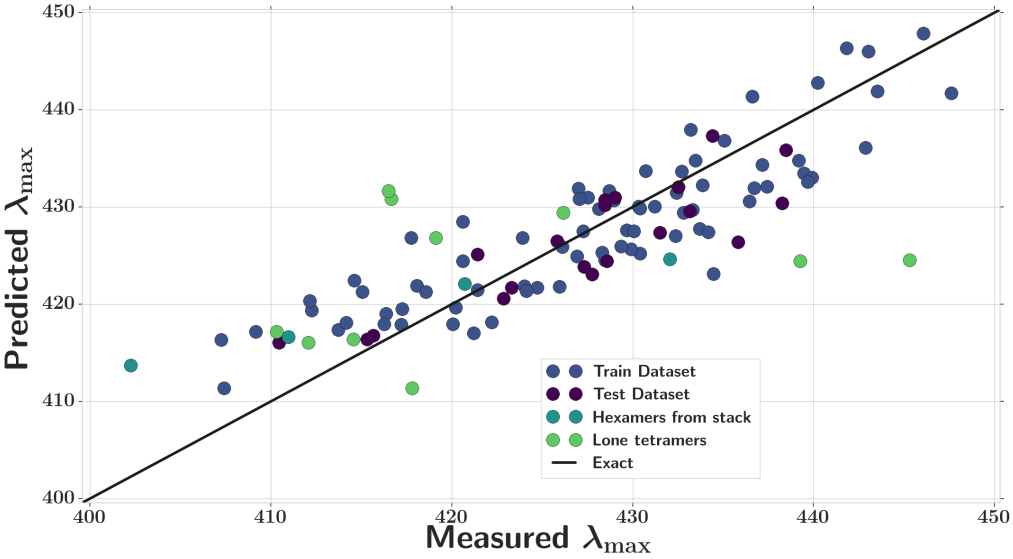

Figure 5.6: Application of the regression model in predicting data in the training and testing data sets, as well as predicting wavelengths for hexamers pulled from the same 20 peptide MD runs and tetramers pulled from runs containing four protonated peptides. The model does reasonably well in predicting the peak excitation wavelengths of the hexamers, but doesn't do quite as well at prediction of tetramers pulled from different MD runs. This is indicative of these runs exploring different core configurations than are observed in those extracted from 20 peptide simulations.

in core stacking configurations.

The principal result of our analysis is to show that some of the key structural properties of the peptide aggregates governing the wavelength of maximum absorption are the linear distances between first, second, and third nearest neighbors, the twist angle between nearest neighbors, and the radius of gyration of the constituent oligopeptides. These structural properties, and hence the absorbance spectra of the peptide aggregates, can be controlled by the sequence of the peptide wings providing a route to sequence-defined control of optoelectronic response.

Going forward, this model can be improved by incorporation of more data in the training process for a greater diversity of peptide chemistries, and, subject to the availability of more computer power, larger self-assembled aggregates. We are also working with collaborators in the Schleife group to uncover the exact atom to atom transitions causing the observed excited state properties. This work will enable the guided design of more accurate descriptors, which will improve the predictive capabilities of our model. It may also be possible to implement a semi-empirical approach to excited state calculations to increase the system size we are capable of sampling, allowing for more accurate excited state calculations [256–258]. With a sufficiently accurate model, we will conduct MD simulations on peptide compositions that have not yet been tested experimentally in order to identify peptides that optimize blue shift in assembly, enabling the design of assembling peptides with specific spectroscopic properties.

Chapter 6

Conclusion and future work

This thesis work utilized multi-physics and multi-scale computational methods to study the self-assembly of π-conjugated oligopeptides and their excited state properties. Peptide self-assembly provides a powerful approach for the design and creation of novel nano-materials that are both environmentally friendly and biocompatible. Such materials have application in the field of organic electronics as organic light-emitting diodes, field effect transistors, and photovoltaics. The use of molecular simulation and simple machine learning models provides the exciting possibility of *in silico* design of self-assembling materials with desired properties. Our work lays the foundation for this process.

In Chapter 2, we employed all-atom and implicit solvent MD simulations to study the thermodynamics, kinetics, and morphologies of a prototypical self-assembling π-conjugated oligopeptide. We uncovered a favorable dimerization free energy of $\Delta F \approx -15 k_B T$ and found the increase in free energy upon further aggregation by monomeric addition to be $\Delta F \approx -25 k_B T$ independent of aggregate size, indicating that the dimerization and trimerization free energies are of particular importance when studied thermodynamics of aggregation. We also studied the dynamics of assembly using Markov state modeling, and determined the growth of the system to occur by means of rapid aggregation and structural reorganization, followed by slower diffusion-limited aggregation. We determined a time scale on the order of microseconds that would be required to probe beyond the early time aggregation process. Finally we established the use of a rescaling procedure to improve accuracy of implicit solvent simulations for studying free energies of peptide aggregation. This study established the use of molecular simulation protocols for probing assembly of these π-conjugated peptides.

In Chapter 3, we expanded our study to include a variety of peptides with different compositions. We developed a QSPR model able to predict the changes in free energy upon formation of dimers and trimers directly from the peptide composition. This model was found to be able to predict dimerization and trimerization free energies within simulation error for non-polar peptides sufficiently similar to those contained in the training data. We related these free energies to peptide alignment and the existence of a window for dimerization and trimerization free energies enabling optimal alignment. It is also clear from this study that more than just the changes in free energy upon aggregation are necessary to fully characterize the extent of alignment in aggregates. These models enabled us to conduct a high-throughput scan of all peptide wings searching for those that exhibit optimal core alignment. Through this approach, we identified a previously untested peptide that is currently being examined by our experimental collaborators in the group of J.D. Tovar at Johns Hopkins University. This chapter established descriptor based-QSPR modeling as being useful for predicting peptide assembly thermodynamics from primary structure, and shed light on the properties underpinning aggregate alignment.

In Chapter 4, we studied the peptide system in a deprotonated environment. Through MD simulation computing the aggregation thermodynamics of this system, we were able to formulate equations to predict the extent of peptide alignment prior to acid-triggered self-assembly. This revealed that at concentrations of 100 nM or larger, the peptides system does not exist as a system of isolated monomers even in a neutral pH environment, contrary to prior assumption. Our predictions were corroborated by experimental evidence from fluorescence correlation spectroscopy collected by experimental collaborators Dr. Bill Wilson and Dr. Lawrence Valverde. This work constituted an important development in the understanding of the initial stages of the kinetics of the assembly process.

Finally, in Chapter 5 we probed the spectroscopic properties of peptide aggregates in order to determine relationships between structural conformations and the resultant absorption spectrum. We found evidence supporting our hypothesis that the primary impact peptide wings have on the excited state properties of the system is caused by the induced alterations in stacking of the π-conjugated cores. We also constructed a model able to accurately predict the peak absorption wavelength of a peptide aggregate from the geometry of its π-conjugated cores. Applying this model to MD simulations yielded predictions of the peak absorption wavelength that were qualitatively in agreement with experiment. This study constitutes a major piece of the process requisite for conducting high throughput scans of peptide composition to engineer peptides with desired optoelectronic responses.

There are several avenues for future progress on this work. The model obtained in Chapter 5 can be improved by incorporation of data for more diverse peptide chemistries and DFT-guided descriptors. The model could also be coupled with the use of extrapolation to more accurately predict the quantitative shifts in absorption that result from aggregation. Such a model could then be employed with several MD simulations of various sizes to estimate the minimum size and time of MD simulation necessary to be able to predict absorption properties of aggregates. With this knowledge in hand, minimal size MD simulations could be conducted for a wide variety of amino acid sequences in order to develop structure property relationships similar to Chapter 3. All of this would lead to the ability to conduct a high-throughput scan of peptides containing arbitrary π-conjugated units in order to design assembling peptides with controlled absorption properties.

Going further beyond the work conducted here, there are a number of variables that could be studied, including looking at non-standard amino acid groups [49], polydisperse peptide solutions [47], and multivalent peptide cores [259]. Other quantum mechanical approaches could also be employed to study properties going beyond the absorption spectrum, to study photoluminescence, charge transfer, CD spectra, and conductivity of peptide aggregates.

Chapter 7

References

[1] Kunal Roy, Supratik Kar, and Rudra Narayan Das. Statistical methods in QSAR/QSPR. In *A primer on QSAR/QSPR modeling*, pages 37–59. Springer, 2015.

[2] Herdeline Ann M. Ardoña, Kalpana Besar, Matteo Togninalli, Howard E. Katz, and John D. Tovar. Sequence-dependent mechanical, photophysical and electrical properties of pi-conjugated peptide hydrogelators. *J. Mater. Chem. C*, 3:6505–6514, 2015.

[3] Mary Campbell and Shawn Farrell. *Biochemistry*. Cengage Learning, 8th edition, 2014.

[4] Brian D. Wall, Ashley E. Zacca, Allix M. Sanders, William L. Wilson, Andrew L. Ferguson, and John D. Tovar. Supramolecular polymorphism: Tunable electronic interactions within π-conjugated peptide nanostructures dictated by primary amino acid sequence. *Langmuir*, 30(20):5946–5956, May 2014.

[5] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[6] Mazda Rad-Malekshahi, Ludwijn Lempsink, Maryam Amidi, Wim E. Hennink, and Enrico Mastrobattista. Biomedical applications of self-assembling peptides. *Bioconjugate Chem.*, 27(1):3–18, January 2016.

[7] Kristin M. French, Inthirai Somasuntharam, and Michael E. Davis. Self-assembling peptide-based delivery of therapeutics for myocardial infarction. *Advanced Drug Delivery Reviews*, 96:40–53, January 2016.

[8] Ashkan Dehsorkhi, Valeria Castelletto, and Ian W. Hamley. Self-assembling amphiphilic peptides. *J. Pept. Sci.*, 20(7):453–467, 2014.

[9] Mark P. Hendricks, Kohei Sato, Liam C. Palmer, and Samuel I. Stupp. Supramolecular assembly of peptide amphiphiles. *Acc. Chem. Res.*, 50(10):2440–2448, October 2017.

[10] Edsger C. P. Smits, Simon G. J. Mathijssen, Paul A. van Hal, Sepas Setayesh, Thomas C. T. Geuns, Kees A. H. A. Mutsaers, Eugenio Cantatore, Harry J. Wondergem, Oliver Werzer, Roland Resel, Martijn Kemerink, Stephan Kirchmeyer, Aziz M. Muzafarov, Sergei A. Ponomarenko, Bert de Boer, Paul W. M. Blom, and Dago M. de Leeuw. Bottom-up organic integrated circuits. *Nature*, 455:956–, October 2008.

[11] Charles M. Rubert Pérez, Nicholas Stephanopoulos, Shantanu Sur, Sungsoo S. Lee, Christina Newcomb, and Samuel I. Stupp. The powerful functions of peptide-based bioactive matrices for regenerative medicine. *Annals of Biomedical Engineering*, 43(3):501–514, 2015.

[12] Michael R. Caplan and Douglas A. Lauffenburger. Nature's complex copolymers: Engineering design of oligopeptide materials. *Ind. Eng. Chem. Res.*, 41(3):403–412, July 2001.

[13] Vincent F.M. Segers and Richard T. Lee. Local delivery of proteins and the use of self-assembling peptides. *Drug Discovery Today*, 12(13-14):561–568, July 2007.

[14] K. Subramani and W. Ahmed. Chapter 13 - self-assembly of proteins and peptides and their applications in bionanotechnology and dentistry. In Karthikeyan Subramani and Waqar Ahmed, editors, *Micro and Nano Technologies*, pages 209–224. William Andrew Publishing, Boston, 2012.

[15] Hossein Hosseinkhani, Po-Da Hong, and Dah-Shyong Yu. Self-assembled proteins and peptides for regenerative medicine. *Chem. Rev.*, 113(7):4837–4861, April 2013.

[16] Liam C. Palmer, Christina J. Newcomb, Stuart R. Kaltz, Erik D. Spoerke, and Samuel I. Stupp. Biomimetic systems for hydroxyapatite mineralization inspired by bone and enamel. *Chem. Rev.*, 108(11):4754–4783, 2008.

[17] Jeffrey D. Hartgerink, Elia Beniash, and Samuel I. Stupp. Self-assembly and mineralization of peptide-amphiphile nanofibers. *Science*, 294(5547):1684–1688, 2001.

[18] A. Aggeli, M. Bell, N. Boden, J. N. Keen, P. F. Knowles, T. C. B. McLeish, M. Pitkeathly, and S. E. Radford. Responsive gels formed by the spontaneous self-assembly of peptides into polymeric β-sheet tapes. *Nature*, 386(6622):259–262, March 1997.

[19] Ernest Y Lee, Benjamin M Fulan, Gerard C L Wong, and Andrew L Ferguson. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48):13588–13593, 2016.

[20] Marc A. Gauthier and Harm-Anton Klok. Peptide/protein-polymer conjugates: synthetic strategies and design concepts. *Chem. Commun.*, 23:2591–2611, 2008.

[21] Albertus P. H. J. Schenning and E. W. Meijer. Supramolecular electronics; nanowires from self-assembled π-conjugated systems. *Chem. Commun.*, pages 3245–3258, 2005.

[22] Mischa Zelzer and Rein V. Ulijn. Next-generation peptide nanomaterials: molecular networks, interfaces and supramolecular functionality. *Chem. Soc. Rev.*, 39:3351–3357, 2010.

[23] Amanda B. Marciel, Melikhan Tanyeri, Brian D. Wall, John D. Tovar, Charles M. Schroeder, and William L. Wilson. Fluidic-directed assembly of aligned oligopeptides with π-conjugated cores. *Adv. Mater.*, 25(44):6398–6404, 2013.

[24] Rachael A. Mansbach and Andrew L. Ferguson. Control of the hierarchical assembly of π-conjugated optoelectronic peptides by pH and flow. *Org. Biomol. Chem.*, 15:5484–5502, 2017.

[25] Miriam Mba, Alessandro Moretto, Lidia Armelao, Marco Crisma, Claudio Toniolo, and Michele Maggini. Synthesis and self-assembly of oligo(p-phenylenevinylene) peptide conjugates in water. *Chemistry – A European Journal*, 17(7):2044–2047, 2011.

[26] Jeffrey D. Hartgerink, Elia Beniash, and Samuel I. Stupp. Peptide-amphiphile nanofibers: A versatile scaffold for the preparation of self-assembling materials. *Proc. Natl. Acad. Sci. USA*, 99(8):5133–5138, 2002.

[27] Joseph K. Gallaher, Emma J. Aitken, Robert A. Keyzers, and Justin M. Hodgkiss. Controlled aggregation of peptide-substituted perylene-bisimides. *Chem. Commun.*, 48:7961–7963, 2012.

[28] Dennis W. P. M. L owik, E. H. P. Leunissen, M. van den Heuvel, M. B. Hansen, and Jan C. M. van Hest. Stimulus responsive peptide based materials. *Chem. Soc. Rev.*, 39:3394–3412, 2010.

[29] Harm-Anton Klok, Annette Rosler, Gunther Gotz, Elena Mena-Osteritz, and Peter Bauerle. Synthesis of a silk-inspired peptide-oligothiophene conjugate. *Org. Biomol. Chem.*, 2:3541–3544, 2004.

[30] Stephen R. Diegelmann, Justin M. Gorham, and John D. Tovar. One-dimensional optoelectronic nanostructures derived from the aqueous self-assembly of π-conjugated oligopeptides. *J. Am. Chem. Soc.*, 130(42):13840–13841, October 2008.

[31] Rachid Matmour, Inge De Cat, Subi J. George, Wencke Adriaens, Philippe Leclère, Paul H. H. Bomans, Nico A. J. M. Sommerdijk, Jeroen C. Gielen, Peter C. M. Christianen, Jeroen T. Heldens, Jan C. M. van Hest, Dennis W. P. M. Löwik, Steven De Feyter, E. W. Meijer, and Albertus P. H. J. Schenning. Oligo(p-phenylenevinylene)-peptide conjugates: Synthesis and self-assembly in solution and at the solid-liquid interface. *J. Am. Chem. Soc.*, 130(44):14576–14583, November 2008.

[32] David A. Stone, Lorraine Hsu, and Samuel I. Stupp. Self-assembling quinquethiophene-oligopeptide hydrogelators. *Soft Matter*, 5:1990–1993, 2009.

[33] J. H. Burroughes, D. D. C. Bradley, A. R. Brown, R. N. Marks, K. Mackay, R. H. Friend, P. L. Burns, and A. B. Holmes. Light-emitting diodes based on conjugated polymers. *Nature*, 347(6293):539–541, October 1990.

[34] Ullrich Mitschke and Peter Bauerle. The electroluminescence of organic materials. *J. Mater. Chem.*, 10:1471–1507, 2000.

[35] Jean Roncali. Conjugated poly(thiophenes): synthesis, functionalization, and applications. *Chem. Rev.*, 92(4):711–738, June 1992.

[36] Denis Fichou, editor. *Handbook of Oligo- and Polythiophenes.* Wiley-VCH, 1999.

[37] Linyi Bian, Enwei Zhu, Jian Tang, Weihua Tang, and Fujun Zhang. Recent progress in the design of narrow bandgap conjugated polymers for high-efficiency organic solar cells. *Progress in Polymer Science*, 37(9):1292–1331, September 2012.

[38] Xin Guo, Martin Baumgarten, and Klaus Müllen. Designing π-conjugated polymers for organic electronics. *Prog. Polym. Sci.*, 38(12):1832–1908, December 2013.

[39] Se Hye Kim and Jon R. Parquette. A model for the controlled assembly of semiconductor peptides. *Nanoscale*, 4:6940–6947, 2012.

[40] Freek J. M. Hoeben, Pascal Jonkheijm, E. W. Meijer, and Albertus P. H. J. Schenning. About supramolecular assemblies of π-conjugated systems. *Chem. Rev.*, 105(4):1491–1546, 2005.

[41] Christopher R. Newman, C. Daniel Frisbie, Demetrio A. da Silva Filho, Jean-Luc Brédas, Paul C. Ewbank, and Kent R. Mann. Introduction to organic thin film transistors and design of n-channel organic semiconductors. *Chem. Mater.*, 16(23):4436–4451, November 2004.

[42] Harald Hoppe and N. Serdar Sariciftci. Polymer solar cells. In Seth R. Marder and Kwang-Sup Lee, editors, *Photoresponsive Polymers II*, pages 1–86. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[43] Pierre M. Beaujuge and John R. Reynolds. Color control in π-conjugated organic polymers for use in electrochromic devices. *Chemical Reviews*, 110(1):268–320, 2010.

[44] Roman Marty, Ruth Szilluweit, Antoni Sánchez-Ferrer, Sreenath Bolisetty, Jozef Adamcik, Raffaele Mezzenga, Eike-Christian Spitzner, Martin Feifer, Stephan N. Steinmann, Clémence Corminboeuf, and Holger Frauenrath. Hierarchically structured microfibers of “single stack” perylene bisimide and quaterthiophene nanowires. *ACS Nano*, 7(10):8498–8508, October 2013.

[45] White House Office of Science and Technology Policy. Materials Genome Initiative for Global Competitiveness, 2011.

[46] Tejaswini S. Kale, Jeannette E. Marine, and John D. Tovar. Self-assembly and associated photophysics of dendron-appended peptide-π-peptide triblock macromolecules. *Macromolecules*, 50(14):5315–5322, July 2017.

[47] Herdeline Ann M. Ardoña, Emily R. Draper, Francesca Citossi, Matthew Wallace, Louise C. Serpell, Dave J. Adams, and John D. Tovar. Kinetically controlled coassembly of multichromophoric peptide hydrogelators and the impacts on energy transport. *J. Am. Chem. Soc.*, 139(25):8685–8692, June 2017.

[48] Allix M. Sanders, Timothy J. Magnanelli, Arthur E. Bragg, and John D. Tovar. Photoinduced electron transfer within supramolecular donor–acceptor peptide nanostructures under aqueous conditions. *J. Am. Chem. Soc.*, 138(10):3362–3370, March 2016.

[49] Herdeline Ann M. Ardoña and John D. Tovar. Energy transfer within responsive pi-conjugated coassembled peptide-based nanostructures in aqueous environments. *Chem. Sci.*, 6:1474–1484, 2015.

[50] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):6679–6685, 2005.

[51] Tamar Schlick, Rosana Collepardo-Guevara, Leif Arthur Halvorsen, Segun Jung, and Xia Xiao. Biomolecular modeling and simulation: a field coming of age. *Quarterly Reviews of Biophysics*, 44(2):191–228, 001 2011.

[52] Jérémie Mortier, Christin Rakers, Marcel Bermudez, Manuela S. Murgueitio, Sereina Riniker, and Gerhard Wolber. The impact of molecular dynamics on drug design: applications for the characterization of ligand–macromolecule complexes. *Drug Discovery Today*, 20(6):686–702, June 2015.

[53] Stefan Tsonchev, George C. Schatz, and Mark A. Ratner. Electrostatically-directed self-assembly of cylindrical peptide amphiphile nanostructures. *J. Phys. Chem. B*, 108(26):8817–8822, May 2004.

[54] Olga A. Gus'kova, Pavel G. Khalatur, Peter Bäuerle, and Alexei R. Khokhlov. Silk-inspired 'molecular chimeras': Atomistic simulation of nanoarchitectures based on thiophene–peptide copolymers. *Chemical Physics Letters*, 461(1–3):64–70, August 2008.

[55] Bryce A. Thurston, John D. Tovar, and Andrew L. Ferguson. Thermodynamics, morphology, and kinetics of early-stage self-assembly of π-conjugated oligopeptides. *Molecular Simulation*, 42(12):955–975, 2016.

[56] Brian D. Wall, Yuecheng Zhou, Shao Mei, Herdeline Ann M. Ardoña, Andrew L. Ferguson, and John D. Tovar. Variation of formal hydrogen-bonding networks within electronically delocalized π-conjugated oligopeptide nanostructures. *Langmuir*, 30(38):11375–11385, September 2014.

[57] Jagannath Mondal, Xiao Zhu, Qiang Cui, and Arun Yethiraj. Self-assembly of β-peptides: Insight from the pair and many-body free energy of association. *J. Phys. Chem. C*, 114(32):13551–13556, July 2010.

[58] G. Hummer, J. C. Rasaiah, and J. P. Noworyta. Water conduction through the hydrophobic channel of a carbon nanotube. *Nature*, 414:188, November 2001.

[59] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, November 1964.

[60] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, November 1965.

[61] A. J. Freeman and E. Wimmer. Density functional theory as a major tool in computational materials science. *Annu. Rev. Mater. Sci.*, 25(1):7–36, October 1995.

[62] Lars Goerigk and Stefan Grimme. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.*, 13:6670–6688, 2011.

[63] Kieron Burke. Perspective on density functional theory. *The Journal of Chemical Physics*, 136(15):150901, October 2012.

[64] Carlo Adamo and Denis Jacquemin. The calculations of excited-state properties with time-dependent density functional theory. *Chem. Soc. Rev.*, 42:845–856, 2013.

[65] Geeta S. Vadehra, Brian D. Wall, Stephen R. Diegelmann, and John D. Tovar. On-resin dimerization incorporates a diverse array of π-conjugated functionality within aqueous self-assembling peptide backbones. *Chem. Commun.*, 46:3947–3949, 2010.

[66] Brian D. Wall and John D. Tovar. Synthesis and characterization of π-conjugated peptide-based supramolecular materials. *Pure Appl. Chem.*, 84:1039–1045, 2012.

[67] Allix M. Sanders, Thomas J. Dawidczyk, Howard E. Katz, and John D. Tovar. Peptide-based supramolecular semiconductor nanomaterials via Pd-Catalyzed solid-phase "dimerizations". *ACS Macro Lett.*, 1(11):1326–1329, November 2012.

[68] Herdeline Ann M. Ardoña and John D. Tovar. Peptide π-electron conjugates: Organic electronics for biology? *Bioconjugate Chem.*, 26(12):2290–2302, December 2015.

[69] Brian D. Wall, Stephen R. Diegelmann, Shuming Zhang, Thomas J. Dawidczyk, William L. Wilson, Howard E. Katz, Hai-Quan Mao, and John D. Tovar. Aligned macroscopic domains of optoelectronic nanostructures prepared via shear-flow assembly of peptide hydrogels. *Adv. Mater.*, 23(43):5009–5014, 2011.

[70] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, February 2008.

[71] Alexander W. Schüttelkopf and Daan M. F. van Aalten. *PRODRG*: a tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallogr., Sect D: Biol. Crystallogr.*, 60(8):1355–1363, Aug 2004.

[72] Alexander D. MacKerell and Nilesh K. Banavali. All-atom empirical force field for nucleic acids: II. application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.*, 21(2):105–120, 2000.

[73] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. *Intermolecular Forces*. Reidel, Dordrecht, 1981.

[74] W. F. Van Gunsteren and H. J. C. Berendsen. A leap-frog algorithm for stochastic dynamics. *Mol. Simulat.*, 1(3):173–185, March 1988.

[75] N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink, and H. J. C. Berendsen. Efficient algorithms for langevin and DPD dynamics. *J. Chem. Theory Comput.*, 8(10):3637–3649, June 2012.

[76] R. W. Hockney and J. W. Eastwood. *Computer Simulation using Particles*. CRC Press, 2010.

[77] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[78] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8593, 1995.

[79] M. P. Allen and D. J. Tildesley. *Computer Simulations of Liquids*. Oxford University Press, 1989.

[80] W. Clark Still, Anna Tempczyk, Ronald C. Hawley, and Thomas Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112(16):6127–6129, August 1990.

[81] Raymond Constanciel and Renato Contreras. Self consistent field theory of solvent effects representation by continuum models: Introduction of desolvation contribution. *Theor. Chim. Acta*, 65(1):1–11, 1984.

[82] D. P. Fernández, Y. Mulev, A. R. H. Goodwin, and J. M. H. Levelt Sengers. A database for the static dielectric constant of water and steam. *J. Phys. Chem. Ref. Data*, 24(1):33–70, 1995.

[83] Alexey Onufriev, Donald Bashford, and David A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.*, 55(2):383–394, 2004.

[84] Michael Schaefer, Christian Bartels, and Martin Karplus. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.*, 284(3):835–848, December 1998.

[85] R. G. Endres. Accelerating all-atom protein folding simulations through reduced dihedral barriers. *Molecular Simulation*, 31(11):773–777, September 2005.

[86] Youngshang Pak, Eunae Kim, and Soonmin Jang. Misfolded free energy surface of a peptide with $\alpha\beta\beta$ motif (1PSV) using the generalized born solvation model. *The Journal of Chemical Physics*, 121(18):9184–9185, 2004.

[87] Daniel R. Roe, Asim Okur, Lauren Wickstrom, Viktor Hornak, and Carlos Simmerling. Secondary structure bias in generalized born solvent models: Comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J. Phys. Chem. B*, 111(7):1846–1857, February 2007.

[88] Thu Zar Lwin, Ruhong Zhou, and Ray Luo. Is Poisson-Boltzmann theory insufficient for protein folding simulations? *The Journal of Chemical Physics*, 124(3):–, 2006.

[89] Fatih Yaşar, Ping Jiang, and Ulrich HE Hansmann. Multicanonical molecular dynamics simulations of the n-terminal domain of protein l9. *Europhysics Letters*, 105(3):30008, 2014.

[90] Giuseppina Andreotti, Israel Cabeza de Vaca, Angelita Poziello, Maria Chiara Monti, Victor Guallar, and Maria Vittoria Cubellis. Conformational response to ligand binding in phosphomannomutase2: Insights into inborn glycosylation disorder. *Journal of Biological Chemistry*, 289(50):34900–34910, 2014.

[91] M. Scott Shell, Ryan Ritterson, and Ken A. Dill. A test on peptide stability of AMBER force fields with implicit solvation. *J. Phys. Chem. B*, 112(22):6878–6886, June 2008.

[92] Hailey R Bureau, Dale R Merz, Eli Hershkovits, Stephen Quirk, and Rigoberto Hernandez. Constrained unfolding of a helical peptide: Implicit versus explicit solvents. *PLoS ONE*, 10(5):e0127034–, April 2015.

[93] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, February 1977.

[94] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[95] Jochen S. Hub, Bert L. de Groot, and David van der Spoel. g_wham–a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.*, 6(12):3713–3720, November 2010.

[96] A. R. Brice and B. N. Dominy. Examining electrostatic influences on base-flipping: A comparison of TIP3P and GB solvent models. *Commun. Comput. Phys.*, 13(1):223–237, 2013.

[97] Olof Allnér, Lennart Nilsson, and Alessandra Villa. Magnesium ion–water coordination and exchange in biomolecular simulations. *J. Chem. Theory Comput.*, 8(4):1493–1502, 2012.

[98] Lihua Wang, Brian E. Hingerty, A.R. Srinivasan, Wilma K. Olson, and Suse Broyde. Accurate representation of b-DNA double helical structure with implicit solvent and counterions. *Biophys. J.*, 83(1):382–406, July 2002.

[99] Justin R. Spaeth, Ioannis G. Kevrekidis, and Athanassios Z. Panagiotopoulos. A comparison of implicit- and explicit-solvent simulations of self-assembly in block copolymer and solute systems. *J. Chem. Phys.*, 134(16):–, 2011.

[100] Christopher Maffeo, Thuy T. M. Ngo, Taekjip Ha, and Aleksei Aksimentiev. A coarse-grained model of unstructured single-stranded DNA derived from atomistic simulation and single-molecule experiment. *J. Chem. Theory Comput.*, 10(8):2891–2896, 2014.

[101] Linda Yu Zhang, Emilio Gallicchio, Richard A. Friesner, and Ronald M. Levy. Solvent models for protein–ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. *J. Comput. Chem.*, 22(6):591–607, 2001.

[102] Alessandra Villa, Christine Peter, and Nico F. A. van der Vegt. Self-assembling dipeptides: conformational sampling in solvent-free coarse-grained simulation. *Phys. Chem. Chem. Phys.*, 11:2077–2086, 2009.

[103] T. Sanghi and N. R. Aluru. Coarse-grained potential models for structural prediction of carbon dioxide (CO_2) in confined environments. *J. Chem. Phys.*, 136(2):–, 2012.

[104] W G Noid. Perspective: coarse-grained models for biomolecular systems. *J. Chem. Phys.*, 139(9):090901, 2013.

[105] Jayendran C. Rasaiah. *Encyclopedia of Chemical Physics and Physical Chemistry: Fundamentals*, chapter A2.3 Statistical mechanics of strongly interacting systems: liquids and solids. Taylor & Francis, 2001.

[106] David Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–647, September 2005.

[107] Ka Lum, David Chandler, and John D. Weeks. Hydrophobicity at small and large length scales. *J. Phys. Chem. B*, 103(22):4570–4577, June 1999.

[108] Manoj V. Athawale, Gaurav Goel, Tuhin Ghosh, Thomas M. Truskett, and Shekhar Garde. Effects of lengthscales and attractions on the collapse of hydrophobic polymers in water. *Proceedings of the National Academy of Sciences*, 104(3):733–738, 2007.

[109] R. Kumar, J. R. Schmidt, and J. L. Skinner. Hydrogen bonding definitions and dynamics in liquid water. *The Journal of Chemical Physics*, 126(20):–, 2007.

[110] David van der Spoel, Erik Lindahl, Berk Hess, and the GROMACS developement team. *GROMACS User Manual.* Royal Institute of Technology and Uppsala Univeristy, 4.6.7 edition, 2014.

[111] Nathan A. Baker, Donald Bashford, and David A. Case. Implicit solvent electrostatics in biomolecular simulation. In Benedict Leimkuhler, Christophe Chipot, Ron Elber, Aatto Laaksonen, Alan Mark, Tamar Schlick, Christoph Schütte, and Robert Skeel, editors, *New Algorithms for Macromolecular Simulation*, volume 49 of *Lecture Notes in Computational Science and Engineering*, pages 263–295. Springer Berlin Heidelberg, 2006.

[112] R. D. Groot. Electrostatic interactions in dissipative particle dynamics—simulation of polyelectrolytes and anionic surfactants. *J. Chem. Phys.*, 118(24):11265–11277, 2003.

[113] R.D. Groot and K.L. Rabone. Mesoscopic simulation of cell membrane damage, morphology change and rupture by nonionic surfactants. *Biophys. J.*, 81(2):725–736, 2001.

[114] Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B*, 108(2):750–760, January 2004.

[115] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications.* Academic Press, 2nd edition, 2001.

[116] Yao-Feng Hu, Wen-Jie Lv, Shuangliang Zhao, Ya-Zhuo Shang, Hua-Lin Wang, and Hong-Lai Liu. Effect of surfactant SDS on DMSO transport across water/hexane interface by molecular dynamics simulation. *Chemical Engineering Science*, 134:813–822, September 2015.

[117] Orsolya Gereben and László Pusztai. Investigation of the structure of ethanol-water mixtures by molecular dynamics simulation I: analyses concerning the hydrogen-bonded pairs. *J. Phys. Chem. B*, 119(7):3070–3084, February 2015.

[118] Joseph E. Basconi and Michael R. Shirts. Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations. *J. Chem. Theory Comput.*, 9(7):2887–2899, July 2013.

[119] Jason Swails, Darrin M York, and Adrian E Roitberg. Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: Implementation, testing, and validation. *Journal of Chemical Theory and Computation*, 10(3):1341–1352, December 2013.

[120] Orkid Coskuner and Olivia Wise-Scira. Arginine and disordered amyloid-β peptide structures: Molecular level insights into the toxicity in alzheimer's disease. *ACS Chemical Neuroscience*, 4(12):1549–1558, September 2013.

[121] Alessio Atzori, Audrey E Baker, Mark Chiu, Richard A Bryce, and Pascal Bonnet. Effect of sequence and stereochemistry reversal on p53 peptide mimicry. *PLoS ONE*, 8(7):e68723–, June 2013.

[122] Samuel Karlin. *A First Course in Stochastic Processes*. Academic Press, revised edition, 2014.

[123] Saravanapriyan Sriraman, Ioannis G. Kevrekidis, and Gerhard Hummer. Coarse master equation from bayesian analysis of replica molecular dynamics simulations. *J. Phys. Chem. B*, 109(14):6479–6484, April 2005.

[124] Gregory R Bowman. An overview and practical guide to building Markov State Models. In Gregory R Bowman, Vijay S. Pande, and Frank Noe, editors, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, chapter 2, pages 7–22. Springer, 2014.

[125] Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, September 2010.

[126] William C Swope, Jed W Pitera, and Frank Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B*, 108(21):6571–6581, 2004.

[127] William C Swope, Jed W Pitera, Frank Suits, Mike Pitman, Maria Eleftheriou, Blake G Fitch, Robert S Germain, Aleksandr Rayshubski, TJ Christopher Ward, Yuriy Zhestkov, et al. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β-hairpin peptide. *J. Phys. Chem. B*, 108(21):6582–6594, 2004.

[128] J. Chodera, W. Swope, J. Pitera, and K. Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.*, 5(4):1214–1226, January 2006.

[129] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, 25:135–144, 2014.

[130] Yasunari Inamura. "Estimating continuous time transition matrices from discretely observed data". *"Estimating Continuous Time Transition Matrices From Discretely Observed Data," Bank of Japan Working Paper Series*, No.06 – E07, April 2006.

[131] Michael Rubinstein and Ralph H. Colby. *Polymer Physics*. OUP Oxford, 2003.

[132] J. Lawson. Generalized Runge-Kutta processes for stable systems with large lipschitz constants. *SIAM J. Numer. Anal.*, 4(3):372–380, September 1967.

[133] N. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM. J. Matrix Anal. & Appl.*, 26(4):1179–1193, January 2005.

[134] A. Al-Mohy and N. Higham. A new scaling and squaring algorithm for the matrix exponential. *SIAM. J. Matrix Anal. & Appl.*, 31(3):970–989, August 2009.

[135] Masaaki Kijima. *Markov Processes for Stochastic Modeling*, volume Volume 6 of Stochastic Modeling Series, chapter 2, page 64. CRC Press, 1997.

[136] Marco Sarich, Jan-Hendrik Prinz, and Christof Schütte. Markov model theory. In Gregory R Bowman, Vijay S. Pande, and Frank Noe, editors, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, chapter 3, pages 23–44. Springer, 2014.

[137] Gregory R Bowman, Vijay S. Pande, and Frank Noe. Introduction and overview of this book. In Gregory R Bowman, Vijay S. Pande, and Frank Noe, editors, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, chapter 1, pages 1–6. Springer, 2014.

[138] Jeremy D. Schmit, Kingshuk Ghosh, and Ken Dill. What drives amyloid molecules to assemble into oligomers and fibrils? *Biophys. J.*, 100(2):450–458, 2011.

[139] Winnie Yong, Aleksey Lomakin, Marina D. Kirkitadze, David B. Teplow, Sow-Hsin Chen, and George B. Benedek. Structure determination of micelle-like intermediates in amyloid β-protein fibril assembly by using small angle neutron scattering. *Proc. Natl. Acad. Sci. USA*, 99(1):150–154, 2002.

[140] J E Gillam and C E MacPhee. Modelling amyloid fibril formation kinetics: mechanisms of nucleation and growth. *J. Phys. Condens. Matter*, 25(37):373101, 2013.

[141] Jennifer M. Andrews and Christopher J. Roberts. A Lumry-Eyring nucleated polymerization model of protein aggregation kinetics: 1. aggregation with pre-equilibrated unfolding. *J. Phys. Chem. B*, 111(27):7897–7913, July 2007.

[142] Yi Li and Christopher J. Roberts. Lumry-Eyring nucleated-polymerization model of protein aggregation kinetics. 2. competing growth via condensation and chain polymerization. *J. Phys. Chem. B*, 113(19):7020–7032, May 2009.

[143] Victoria Wagoner, Mookyung Cheon, Iksoo Chang, and Carol Hall. Computer simulation study of amyloid fibril formation by palindromic sequences in prion peptides. *Proteins*, 79(7):2132–2145, May 2011.

[144] Rachael A. Mansbach and Andrew L. Ferguson. Coarse-grained molecular simulation of the hierarchical self-assembly of π-conjugated optoelectronic peptides. *J. Phys. Chem. B*, 121(7):1684–1706, February 2017.

[145] Bryce A. Thurston and Andrew L. Ferguson. Machine learning and molecular design of self-assembling -conjugated oligopeptides. *Molecular Simulation*, 44(11):930–945, July 2018.

[146] Corwin Hansch, Albert Leo, and D. H. Hoekman. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, volume 1. American Chemical Society, 1995.

[147] Alan R. Katritzky, Victor S. Lobanov, and Mati Karelson. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.*, 24:279–287, 1995.

[148] Mati Karelson, Victor S. Lobanov, and Alan R. Katritzky. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.*, 96(3):1027–1044, January 1996.

[149] Saeed Yousefinejad and Bahram Hemmateenejad. Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, 149(Part B):177–204, 2015.

[150] Ernest Y Lee, Gerard CL Wong, and Andrew L Ferguson. Machine learning-enabled discovery and design of membrane-active peptides. *Bioorganic & Medicinal Chemistry*, 2017.

[151] Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.*, 57(12):4977–5010, June 2014.

[152] Tu Le, V. Chandana Epa, Frank R. Burden, and David A. Winkler. Quantitative structure–property relationship modeling of diverse materials properties. *Chem. Rev.*, 112(5):2889–2919, May 2012.

[153] HÅvard Jenssen, Christopher D. Fjell, Artem Cherkasov, and Robert E. W. Hancock. QSAR modeling and computer-aided design of antimicrobial peptides. *Journal of Peptide Science*, 14(1):110–114, 2008.

[154] Mariya A. Toropova, Aleksandar M. Veselinović, Jovana B. Veselinović, Dušica B. Stojanović, and Andrey A. Toropov. QSAR modeling of the antimicrobial activity of peptides as a mathematical function of a sequence of amino acids. *Computational Biology and Chemistry*, 59, Part A:126–130, December 2015.

[155] Xuan Xiao, Pu Wang, Wei-Zhong Lin, Jian-Hua Jia, and Kuo-Chen Chou. iAMP-2L: a two-level multilabel classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry*, 436(2):168–177, May 2013.

[156] Heng Luo, Hao Ye, Hui Wen Ng, Sugunadevi Sakkiah, Donna L. Mendrick, and Huixiao Hong. sNebula, a network-based algorithm to predict binding between human leukocyte antigens and peptides. *Scientific Reports*, 6:32115–, August 2016.

[157] Chao Ji, Sujun Li, James P. Reilly, Predrag Radivojac, and Haixu Tang. XLSearch: a probabilistic database search algorithm for identifying cross-linked peptides. *J. Proteome Res.*, 15(6):1830–1841, June 2016.

[158] Wensheng Wu, Canyang Zhang, Wenjing Lin, Quan Chen, Xindong Guo, Yu Qian, and Lijuan Zhang. Quantitative structure-property relationship (QSPR) modeling of drug-loaded polymeric micelles via genetic function approximation. *PLOS ONE*, 10(3):e0119575–, March 2015.

[159] Chen Chen, Yonglan Liu, Jin Zhang, Mingzhen Zhang, Jie Zheng, Yong Teng, and Guizhao Liang. A quantitative sequence–aggregation relationship predictor applied as identification of self-assembled hexapeptides. *Chemometrics and Intelligent Laboratory Systems*, 145:7–16, July 2015.

[160] Allix M. Sanders and John D. Tovar. Solid-phase Pd-catalysed cross-coupling methods for the construction of π-conjugated peptide nanomaterials. *Supramolecular Chemistry*, 26(3-4):259–266, March 2014.

[161] Bo Li, Songsong Li, Yuecheng Zhou, Herdeline Ann M. Ardoña, Lawrence R. Valverde, William L. Wilson, John D. Tovar, and Charles M. Schroeder. Nonequilibrium self-assembly of π-conjugated oligopeptides in solution. *ACS Appl. Mater. Interfaces*, 9(4):3977–3984, February 2017.

[162] George M Whitesides and Bartosz Grzybowski. Self-assembly at all scales. *Science*, 295(5564):2418–2421, 2002.

[163] Faifan Tantakitti, Job Boekhoven, Xin Wang, Roman V. Kazantsev, Tao Yu, Jiahe Li, Ellen Zhuang, Roya Zandi, Julia H. Ortony, Christina J. Newcomb, Liam C. Palmer, Gajendra S. Shekhawat, Monica Olvera de la Cruz, George C. Schatz, and Samuel I. Stupp. Energy landscapes and functions of supramolecular systems. *Nature Materials*, 15:469–, January 2016.

[164] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: a message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1–3):43 – 56, 1995.

[165] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. GROMACS: fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

[166] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[167] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247 – 260, 2006.

[168] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[169] Junmei Wang, Piotr Cieplak, and Peter A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000.

[170] Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, October 1993.

[171] Enguerran Vanquelef, Sabrina Simon, Gaelle Marquant, Elodie Garcia, Geoffroy Klimerak, Jean Charles Delepine, Piotr Cieplak, and François-Yves Dupradeau. R.E.D. server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Research*, 39(Web Server issue):W511–W517, April 2011.

[172] Christopher A. Reynolds, Jonathan W. Essex, and W. Graham Richards. Atomic charges for variable molecular conformations. *J. Am. Chem. Soc.*, 114(23):9075–9079, November 1992.

[173] Michael J. Frisch, G. W. Trucks, H. Bernhard Schlegel, Gustavo E. Scuseria, Michael A. Robb, James R. Cheeseman, Giovanni Scalmani, Vincenzo Barone, Benedetta Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, Xiaosong Li, H. P. Hratchian, Artur F. Izmaylov, Julien Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, François Ogliaro, Michael J. Bearpark, Jochen Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, Rika Kobayashi, J. Normand, Krishnan Raghavachari, Alistair P. Rendell, J. C. Burant, S. S. Iyengar, Jacopo Tomasi, M. Cossi, N. Rega, N. J. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ödön Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and Douglas J. Fox. Gaussian 09, 2009.

[174] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[175] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):–, 2007.

[176] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[177] Shuichi Nosé and M. L. Klein. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, 50(5):1055–1076, December 1983.

[178] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, June 1984.

[179] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, Mar 1985.

[180] R.W Hockney, S.P Goel, and J.W Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 1974.

[181] Richard M. Neumann. Entropic approach to brownian movement. *American Journal of Physics*, 48(5):354–357, May 1980.

[182] Joohyun Jeon and M Scott Shell. Charge effects on the fibril-forming peptide KTVIIE: a two-dimensional replica exchange simulation study. *Biophysical Journal*, 102(8):1952–1960, March 2012.

[183] Jiang Wang and Andrew L. Ferguson. Mesoscale simulation of asphaltene aggregation. *J. Phys. Chem. B*, 120(32):8016–8035, August 2016.

[184] Frank Harary. *Graph Theory*. Addison-Wesley, Reading, MA, 1969.

[185] Kalpana Besar, Herdeline Ann M. Ardoña, John D. Tovar, and Howard E. Katz. Demonstration of hole transport and voltage equilibration in self-assembled π-conjugated peptide nanostructures using field-effect transistor architectures. *ACS Nano*, 9(12):12401–12409, December 2015.

[186] Roberto Todeschini and Viviana Consonni. *Molecular descriptors for chemoinformatics*. Weinheim: Wiley VCH, 2010.

[187] Chun Wei Yap. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474, 2011.

[188] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: an open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011.

[189] Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980.

[190] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[191] Josef Kittler. Feature selection and extraction. *Handbook of pattern recognition and image processing*, pages 59–83, 1986.

[192] David Rogers and A. J. Hopfinger. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.*, 34(4):854–866, July 1994.

[193] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, February 1970.

[194] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.

[195] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[196] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, New York, 2001.

[197] Manoj Bhasin and G.P.S. Raghava. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine*, 22(23):3195–3204, 2004.

[198] Irini A. Doytchinova and Darren R. Flower. Predicting class I major histocompatibility complex (MHC) binders using multivariate statistics: Comparison of discriminant analysis and multiple linear regression. *J. Chem. Inf. Model.*, 47(1):234–238, January 2007.

[199] Johannes Söllner. Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins. *Journal of Molecular Recognition*, 19(3):209–214, 2006.

[200] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[201] G. Moreau and P. Broto. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv J Chim*, 4:359–360, 1980.

[202] R. T. Sanderson. Principles of electronegativity part II. applications. *J. Chem. Educ.*, 65(3):227–, March 1988.

[203] Frank R. Burden. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.*, 29(3):225–227, August 1989.

[204] Lemont B. Kier and Lowell H. Hall. An electrotopological-state index for atoms in molecules. *Pharmaceutical Research*, 7(8):801–807, 1990.

[205] Lowell H. Hall and Lemont B. Kier. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling*, pages 367–422. John Wiley & Sons, Inc., 2007.

[206] R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115–146, 1954.

[207] Daniel Veltri, Uday Kamath, and Amarda Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 1:8, 2018.

[208] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and E Weinan. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical Review Letters*, 120(14):143001, 2018.

[209] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and E Weinan. DeePCG: constructing coarse-grained models via deep neural networks. *arXiv preprint arXiv:1802.08549*, 2018.

[210] Lawrence Valverde, Bryce A. Thurston, Andrew L. Ferguson, and William L. Wilson. Evidence for prenucleated fibrilogenesis of acid-mediated self-assembling oligopeptides via molecular simulation and fluorescence correlation spectroscopy. *Langmuir*, 34(25):7346–7354, June 2018.

[211] Juan Wang, Kai Liu, Ruirui Xing, and Xuehai Yan. Peptide self-assembly: thermodynamics and kinetics. *Chem. Soc. Rev.*, 45(20):5589–5604, 2016.

[212] John D. Tovar. Supramolecular construction of optoelectronic biomaterials. *Acc. Chem. Res.*, 46(7):1527–1537, July 2013.

[213] Thomas C. T. Michaels and Tuomas P. J. Knowles. Mean-field master equation formalism for biofilament growth. *American Journal of Physics*, 82(5):476–483, September 2018.

[214] Samuel I. A. Cohen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. *The Kinetics and Mechanisms of Amyloid Formation*, chapter 10, pages 183–209. Wiley-Blackwell, 2013.

[215] Yuecheng Zhou, Bo Li, Songsong Li, Herdeline Ann M. Ardoña, William L. Wilson, John D. Tovar, and Charles M. Schroeder. Concentration-driven assembly and sol-gel transition of π-conjugated oligopeptides. *ACS Cent. Sci.*, 3(9):986–994, September 2017.

[216] M K Gilson, J A Given, B L Bush, and J A McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*, 72(3):1047–1069, March 1997.

[217] Ken Dill and Sarina Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience.* Garland Science, 2nd edition, 2010.

[218] Chengqian Yuan, Shukun Li, Qianli Zou, Ying Ren, and Xuehai Yan. Multiscale simulations for understanding the evolution and mechanism of hierarchical peptide self-assembly. *Phys. Chem. Chem. Phys.*, 19(35):23614–23631, 2017.

[219] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[220] Janet E. Del Bene, R. Ditchfield, and J. A. Pople. Self-consistent molecular orbital methods. X. molecular orbital studies of excited states with minimal and extended basis sets. *The Journal of Chemical Physics*, 55(5):2236–2241, October 1971.

[221] Erich Runge and E. K. U. Gross. Density-functional theory for time-dependent systems. *PRL*, 52(12):997–1000, March 1984.

[222] E. K. U. Gross and Walter Kohn. Local density-functional theory of frequency-dependent linear response. *PRL*, 55(26):2850–2852, December 1985.

[223] Mark E. Casida. Time-dependent density functional response theory for molecules. In D. P. Chong, editor, *Recent Advances in Density Functional Methods, Part I*, volume Volume 1 of *Recent Advances in Computational Chemistry*, pages 155–192. World Scientific, October 1995.

[224] E. E. Salpeter and H. A. Bethe. A relativistic equation for bound-state problems. *Phys. Rev.*, 84:1232–1242, Dec 1951.

[225] L. J. Sham and T. M. Rice. Many-particle derivation of the effective-mass equation for the Wannier exciton. *Phys. Rev.*, 144:708–714, Apr 1966.

[226] W. Hanke and L. J. Sham. Many-particle effects in the optical excitations of a semiconductor. *Phys. Rev. Lett.*, 43:387–390, Jul 1979.

[227] G. Strinati. Dynamical shift and broadening of core excitons in semiconductors. *Phys. Rev. Lett.*, 49:1519–1522, Nov 1982.

[228] Yuan Ping, Dario Rocca, and Giulia Galli. Electronic excitations in light absorbers for photoelectrochemical energy conversion: first principles calculations based on many body perturbation theory. *Chem. Soc. Rev.*, 42:2437–2469, 2013.

[229] K. Emrich. An extension of the coupled cluster formalism to excited states (I). *Nuclear Physics A*, 351(3):379–396, January 1981.

[230] Hideo Sekino and Rodney J. Bartlett. A linear response, coupled-cluster theory for excitation energy. *Int. J. Quantum Chem.*, 26(18):255–265, October 1984.

[231] Jan Geertsen, Magnus Rittby, and Rodney J. Bartlett. The equation-of-motion coupled-cluster method: Excitation energies of Be and CO. *Chemical Physics Letters*, 164(1):57–62, December 1989.

[232] Andreas Dreuw and Martin Head-Gordon. Single-reference ab initio methods for the calculation of excited states of large molecules. *Chem. Rev.*, 105(11):4009–4037, November 2005.

[233] Adèle D. Laurent, Carlo Adamo, and Denis Jacquemin. Dye chemistry with time-dependent density functional theory. *Phys. Chem. Chem. Phys.*, 16(28):14334–14356, 2014.

[234] Leticia González, Daniel Escudero, and Luis Serrano-Andrés. Progress and challenges in the calculation of electronic excited states. *ChemPhysChem*, 13(1):28–51, October 2011.

[235] Denis Jacquemin, Eric A. Perpète, Ilaria Ciofini, and Carlo Adamo. Accurate simulation of optical properties in dyes. *Acc. Chem. Res.*, 42(2):326–334, February 2009.

[236] Denis Jacquemin, Valérie Wathelet, Eric A. Perpète, and Carlo Adamo. Extensive TD-DFT benchmark: Singlet-excited states of organic molecules. *J. Chem. Theory Comput.*, 5(9):2420–2435, September 2009.

[237] Adèle D. Laurent and Denis Jacquemin. TD-DFT benchmarks: A review. *Int. J. Quantum Chem*, 113(17):2019–2039, October 2013.

[238] M. A. Marques, C. A. Ullrich, F. Nogueira, A. Rubio, K. Burke, and E. K. U. Gross. *Time-Dependent Density Functional Theory*. Springer, 2006.

[239] Denis Jacquemin, Valérie Wathelet, Julien Preat, and Eric A. Perpète. Ab initio tools for the accurate prediction of the visible spectra of anthraquinones. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 67(2):334–341, June 2007.

[240] Vincenzo Barone and Antonino Polimeno. Integrated computational strategies for uv/vis spectra of large molecules in solution. *Chem. Soc. Rev.*, 36:1724–1731, 2007.

[241] Marc Dierksen and Stefan Grimme. Density functional calculations of the vibronic structure of electronic absorption spectra. *The Journal of Chemical Physics*, 120(8):3544–3554, October 2004.

[242] Fabrizio Santoro, Roberto Improta, Alessandro Lami, Julien Bloino, and Vincenzo Barone. Effective method to compute Franck-Condon integrals for optical spectra of large molecules in solution. *The Journal of Chemical Physics*, 126(8):084509, October 2007.

[243] Fabrizio Santoro, Alessandro Lami, Roberto Improta, and Vincenzo Barone. Effective method to compute vibrationally resolved optical spectra of large molecules at finite temperature in the gas phase and in solution. *The Journal of Chemical Physics*, 126(18):184102, October 2007.

[244] M. Valiev, E.J. Bylaska, N. Govind, K. Kowalski, T.P. Straatsma, H.J.J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T.L. Windus, and W.A. de Jong. NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications*, 181(9):1477–1489, September 2010.

[245] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, Sep 1988.

[246] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.*, 98(45):11623–11627, November 1994.

[247] Axel D. Becke. Density-functional thermochemistry. III. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, October 1993.

[248] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople. Self-consistent molecular orbital methods. XX. a basis set for correlated wave functions. *The Journal of Chemical Physics*, 72(1):650–654, October 1980.

[249] W. J. Hehre, L. Radom, P. R. Schleyer, and J. A. Pople. *Ab Initio Molecular Orbital Theory*. Wiley, 1986.

[250] Wei-Lu Ding, Dong-Mei Wang, Zhi-Yuan Geng, Xiao-Ling Zhao, and Yun-Feng Yan. Molecular engineering of indoline-based D-A-π-A organic sensitizers toward high efficiency performance from first-principles calculations. *The Journal of Physical Chemistry C*, 117(34):17382–17398, 2013.

[251] Denis Jacquemin, Julien Preat, Valérie Wathelet, and Eric A. Perpète. Substitution and chemical environment effects on the absorption spectrum of indigo. *The Journal of Chemical Physics*, 124(7):074104, 2006.

[252] Paul J. Steinhardt, David R. Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *PRB*, 28(2):784–805, July 1983.

[253] Pieter Rein ten Wolde, Maria J. Ruiz-Montero, and Daan Frenkel. Numerical calculation of the rate of crystal nucleation in a Lennard-Jones system at moderate undercooling. *The Journal of Chemical Physics*, 104(24):9932–9947, October 1996.

[254] Stefan Auer and Daan Frenkel. Numerical simulation of crystal nucleation in colloids. In Christian Dr. Holm and Kurt Prof. Dr. Kremer, editors, *Advanced Computer Simulation: Approaches for Soft Matter Sciences I*, pages 149–208. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[255] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[256] J. Ridley and Michael Zerner. An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines. *Theoretica chimica acta*, 32(2):111–134, June 1973.

[257] Joan E. Ridley and Michael C. Zerner. Triplet states via intermediate neglect of differential overlap: Benzene, pyridine and the diazines. *Theoretica chimica acta*, 42(3):223–236, September 1976.

[258] Marco Caricato, Benedetta Mennucci, and Jacopo Tomasi. Solvent effects on the electronic spectra: An extension of the polarizable continuum model to the ZINDO method. *J. Phys. Chem. A*, 108(29):6248–6256, July 2004.

[259] Allix M. Sanders, Tejaswini S. Kale, Howard E. Katz, and John D. Tovar. Solid-phase synthesis of self-assembling multivalent π-conjugated peptides. *ACS Omega*, 2(2):409–419, February 2017.

Appendix A

Markov chain time dependence

A.1 Time dependence of transition rate matrix

By modeling the aggregation of DFAG-OPV-GAFD peptide as a time homogeneous continuous time Markov chain, we implicitly assume that the elements of the transition rate matrix **Q** are time invariant. We present *a posteriori* validation of this assumption by dividing each 70 ns simulation trajectory into contiguous 3.5 ns blocks, and computing the maximum likelihood estimates for the off-diagonal elements of the matrix q_{ij} in each block as an average over all time intervals of length equal to the lag time of $\tau = 100$ ps over the three independent simulations at each concentration. Uncertainties are estimated as the standard error in the values constituting the mean. Diagonal elements are computed as $q_{ii} = -\sum_{j=1}^{n} q_{ij}$, and the associated uncertainly estimated by standard propagation of errors.

In Figure A.1, we present a time resolved plot of the estimates of the transition rate matrix elements for the 0.85 mM concentration and in Figure A.2 analogous plots for the 1.66 mM concentration. Rarely observed transitions precluded the reporting of reliable transition rate estimates for all time blocks, but the calculable transition rates are time invariant within estimated uncertainties. By conducting additional simulations, it is possible that these uncertainties may be reduced to the point that statistically meaningful time dependencies in the transition rates may be extracted, but the observed insensitivity of the transition rates to the observation time coupled with the good reproduction of the time evolution of the cluster size distribution (cf. Figures 2.11, 2.13 and 2.14) supports the assertion that the dynamical evolution is adequately modeled as a time homogeneous Markov process.

A.2 Discrete time Markov chain (DTMC)

Our molecular simulation trajectories are continuous in time, making it possible to estimate the transition rate matrix **Q** directly from the data and model the system evolution as a continuous time Markov chain (CTMC). We may then predict the time evolution of the cluster size distribution at any future time t by

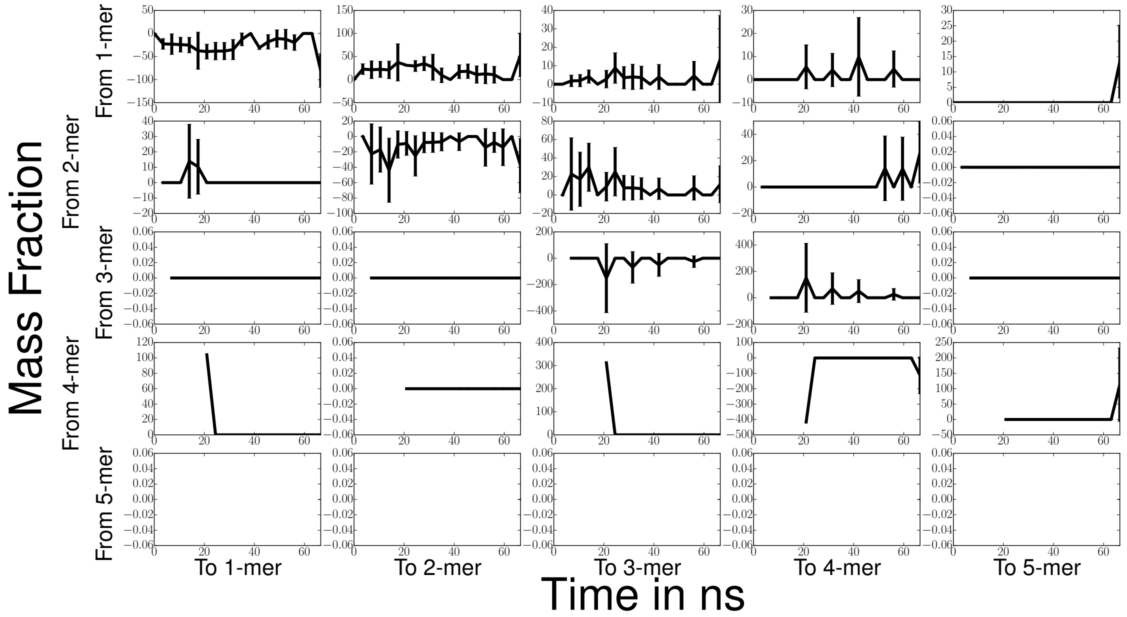

Figure A.1: Time dependence of the transition rate matrix elements extracted from three 70 ns simulations of the aggregation of 64 protonated (low-pH) peptides at a concentration of 0.85 mM using our reparameterized implicit solvent model. Off-diagonal elements q_{ij} are reported over 3.5 ns time blocks as an average over all time intervals of length equal to the lag time of $\tau = 100$ ps over the three independent simulations. Uncertainties are estimated as the standard error in the values constituting the mean. Diagonal elements are computed as $q_{ii} = -\sum_{j=1}^{n} q_{ij}$, and the associated uncertainly estimated by standard propagation of errors. The absence of an error bar implies that the transition was sufficiently rare that fewer than three data points were obtained for that time interval. The absence of data at a given time implies that no peptides of that size existed in any of the simulations at that time so the transition rate out of such a state cannot be quantified.

forming the matrix exponential $e^{\mathbf{Q}t}$ and applying Equation 2.9 [122]. We favor the continuous time formulation for its attractive capacity to predict the cluster size distribution at an arbitrary future time point, but it is also possible to model the system as a discrete time Markov chain (DTMC). In the discrete-time formulation, the transition matrix $\mathbf{T}(\tau)$ containing the transition probabilities between the various cluster sizes over a particular observation interval, or lag time, τ is estimated directly from the data, rather than as the exponential of the transition rate matrix $\mathbf{Q}$ (cf. Equation 2.9). Similar to the CTMC approach, the discrete time treatment assumes that the transition probabilities possess neither temporal nor spatial dependencies, and depend only on the current state of the system and not its past history. The cluster size distribution at some integer multiple of the lag time can be estimated from repeated applications of the transition matrix as [124, 128],

$$\mathbf{p}(n\tau) = \mathbf{p}(0)\,[\mathbf{T}(\tau)]^{n}\,. \tag{A.1}$$

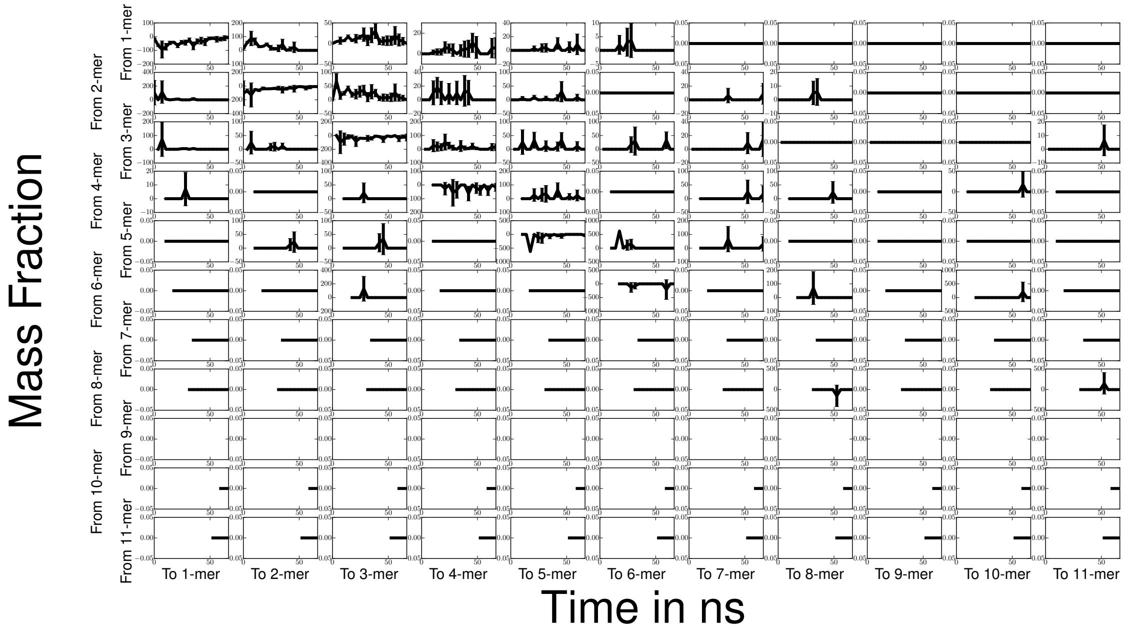

Figure A.2: Time dependence of the transition rate matrix elements extracted from three 70 ns simulations of the aggregation of 125 protonated (low-pH) peptides at a concentration of 1.66 mM using our reparameterized implicit solvent model. Off-diagonal elements q_{ij} are reported over 3.5 ns time blocks as an average over all time intervals of length equal to the lag time of $\tau = 100$ ps over the three independent simulations. Uncertainties are estimated as the standard error in the values constituting the mean. Diagonal elements are computed as $q_{ii} = -\sum_{j=1}^{n} q_{ij}$, and the associated uncertainly estimated by standard propagation of errors. The absence of an error bar implies that the transition was sufficiently rare that fewer than three data points were obtained for that time interval. The absence of data at a given time implies that no peptides of that size existed in any of the simulations at that time so the transition rate out of such a state cannot be quantified.

This relationship makes clear that the discrete time formulation is restricted to predict the cluster size distribution at discrete time intervals. The element $t_{ij}(\tau)$ of the transition matrix $\mathbf{T}$ represents the probability that a monomer will be found in an aggregate of size j after a lag time τ given that it initially resided in an aggregate of size i. Maximum likelihood estimates for the transition probabilities are given by [124,128,130],

$$t_{ij}(\tau) = \frac{N_{ij}(\tau)}{\sum_{m=1}^{M} N_{im}(\tau)}, \tag{A.2}$$

where N_{ij} is the number of transitions observed from state i to state j over the observation interval τ. We estimate transition probabilities as an average over all time blocks of length τ over the three independent simulations conducted at each concentration. For the DTMC to possess the Markov (i.e., memoryless) property, the lag time must exceed the Markov time for the system [124]. The Chapman-Kolmogorov applied to a time homogeneous DTMC possessing the memoryless property states that [124],

$$\mathbf{T}(n\tau) = [\mathbf{T}(\tau)]^n \,, \tag{A.3}$$

providing a mathematical statement that n repeated application of a transition matrix constructed with a lag time of τ should be equivalent to a single application of a transition matrix constructed with a lag time of $n\tau$. Testing this condition provides a commonly used validation that the lag time τ is sufficiently large that the system is Markovian [124].

In Figure A.3, we compare the observed time evolution of the cluster size distribution measured directly from our simulations at 0.85 mM concentration to that predicted from the CTMC using Equation 2.9 and employing a lag time of 100 ps, and two DTMCs using Equation A.1, and employing lag times of 100 ps and 400 ps. In Figure A.4 we present the analogous plot for the 1.66 mM concentration system. The predictions of the three Markov models are in excellent agreement, demonstrating that the time evolution of the system can be equally well formulated as a CTMC or DTMC, and that a lag time of $\tau = 100$ ps is sufficiently high to assure Markovian behavior.

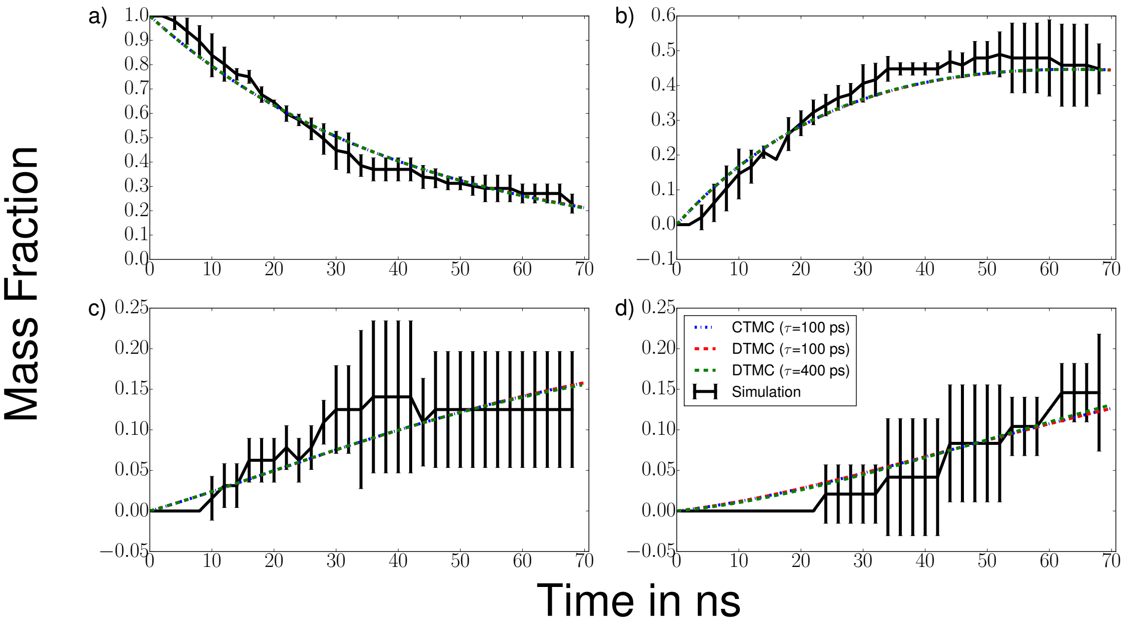

Figure A.3: Comparison of the cluster size distribution in the 0.85 mM system predicted by the (i) CTMC with τ = 100 ps (blue dot-dash lines), (ii) DTMC with a lag time of τ = 100 ps (red dashed lines), (iii) DTMC with a lag time of 4τ = 400 ps (green dashed lines) to that directly observed in the simulations (solid black lines, every 20th point plotted). The simulation data is plotted as the mean and standard deviation of the mass fraction over the three simulations. We present in each panel the results for (a) monomers, (b) dimers, (c) trimers, and (d) tetramers. No error bars are reported when only a single observation was obtained at that particular time interval.

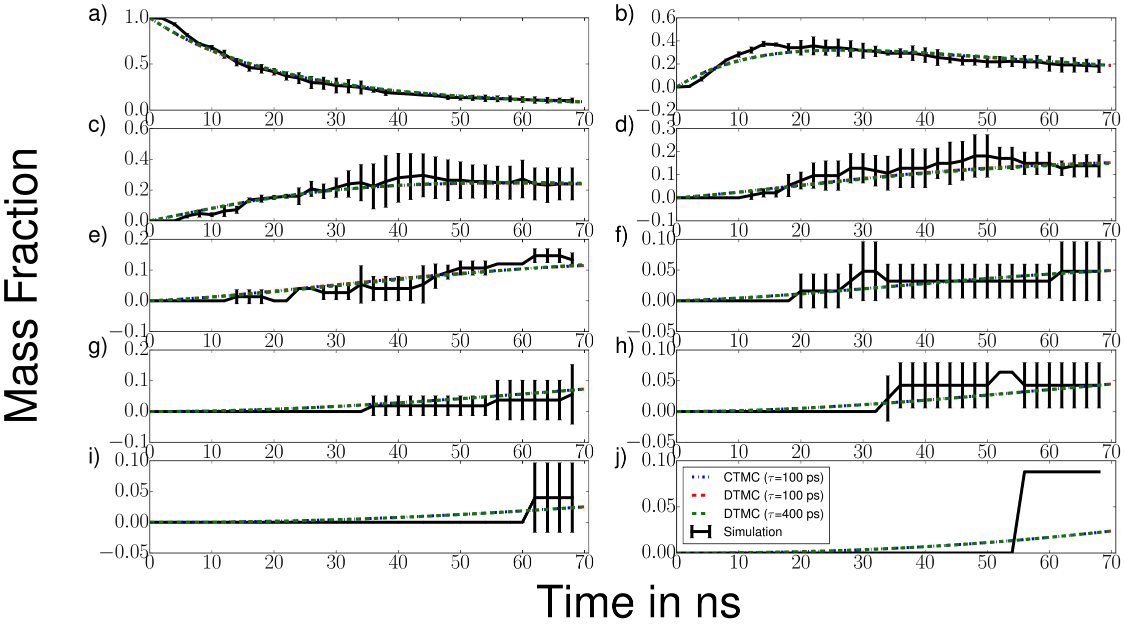

Figure A.4: Comparison of the cluster size distribution in the 1.66 mM system predicted by the (i) CTMC with τ = 100 ps (blue dot-dash lines), (ii) DTMC with a lag time of τ = 100 ps (red dashed lines), (iii) DTMC with a lag time of 4τ = 400 ps (green dashed lines) to that directly observed in the simulations (solid black lines, every 20th point plotted). The simulation data is plotted as the mean and standard deviation of the mass fraction over the three simulations. We present in each panel the results for (a) monomers, (b) dimers, (c) trimers, (d) tetramers, (e) pentamers, (f) hexamers, (g) heptamers, (h) octamers, (i) decamers, and (j) undecamers. No nonamers were formed during our simulations. No error bars are reported when only a single observation was obtained at that particular time interval.

Appendix B

AMBER scaling

We demonstrated in our previous work that the GBSA implicit solvent model substantially overestimates the strength of non-bonded interactions between peptides, and that it is necessary to rescale these interactions in order to reliably reproduce the thermodynamics of peptide aggregation [55]. Following our previous protocol, we adopt the minimally invasive strategy of uniformly rescaling the non-bonded interactions within the peptide force field in implicit solvent to best reproduce the potential of mean force (PMF) profiles for single peptide collapse (Section 3.2.3) and peptide dimerization (Section 3.2.4) computed in explicit solvent [55, 101]. This rescaling protocol can be considered a form of PMF matching [96, 102, 103]. Despite its simplicity, we previously showed the approach to produce satisfactory performance in reproducing explicit solvent results [55]. We adopt DFAG-NDI-GAFD as a prototypical oligopeptide for which to perform the fitting procedure and ascertain the optimal value of the rescaling factor.

The non-bonded interactions comprise Coulombic and Lennard Jones interactions $V^{nb}(r) = V_C(r) + V_{LJ}(r)$ that are each pairwise decomposable functions of interatomic separation r. As detailed in Ref. [55], a uniform rescaling of these pairwise interactions $V^{nb} \to \alpha V^{nb}$ amounts to rescaling the Lennard-Jones interaction parameter by a factor of α and the partial charges by a factor of $\sqrt{\alpha}$. We define the optimal scaling factor as that which minimizes the error function,

$$\text{RMSE}(\alpha) = \text{RMSE}_1(\alpha) + \text{RMSE}_2(\alpha), \tag{B.1}$$

where RMSE_1 and RMSE_2 are, respectively, the root mean squared error between the PMF for single peptide collapse and peptide dimerization computed in explicit and implicit solvent. We computed the implicit solvent PMF curves at values of α = [0.50, 0.60, 0.70, 0.72, 0.75, 0.78, 0.80, 0.90, 1.00], and report in Figure B.1 the values of RMSE, RMSE_1, and RMSE_2 as a function of α. From these results, we discern α = 0.75 to be the optimal value of the scaling factor at which the PMF profiles for peptide collapse agree to within a root mean squared error of 0.7 k_BT and for peptide dimerization within 1.3 k_BT.

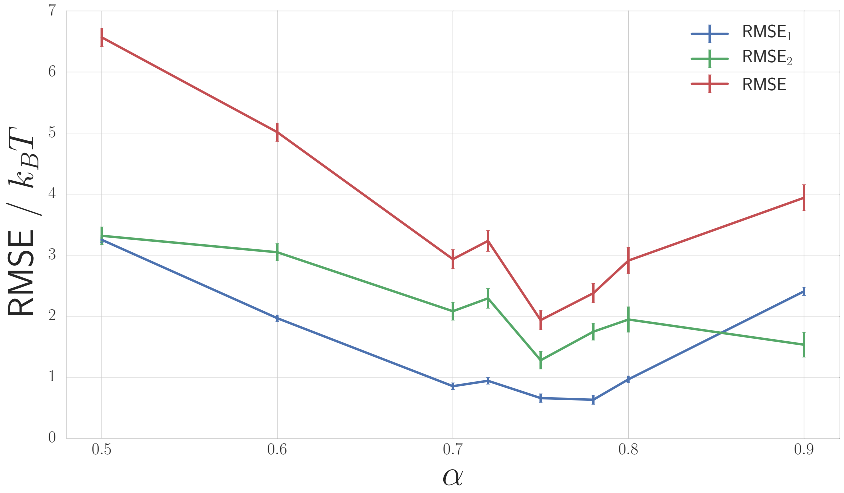

Figure B.1: Root mean squared error between the PMF profiles for single peptide collapse and peptide dimerization in implicit and explicit solvent as a function of the scaling factor for the implicit solvent non-bonded interactions. The agreement for single peptide collapse RMSE$_1$ and peptide dimerization RMSE$_2$ both attain their optima at a scaling factor of $\alpha = 0.75$. Uncertainties are estimated by 100 bootstrap re-samples of the simulation data used to compute the PMF profiles.