To appear in *Molecular Physics*
Vol. 00, No. 00, Month 200x, 1–43

RESEARCH ARTICLE

Thermodynamics, morphology, and kinetics of early-stage self-assembly of π-conjugated oligopeptides

Bryce A. Thurston[a], John D. Tovar[b], and Andrew L. Ferguson[c*]

[a]*Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Email: thursto2@illinois.edu. Phone: (217) 244-8487.* ; [b]*Department of Chemistry, Institute of NanoBioTechnology, and Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD 21218, USA. Email: tovar@jhu.edu. Phone: (410) 516-4358*;[c]*Department of Materials Science and Engineering and Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Email: alf@illinois.edu. Phone: (217) 300-2354. Fax: (217) 333-2736.*

Acknowledgments: This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award # DE-SC0004857. Molecular simulations were partially conducted on University of Illinois Computational Science and Engineering Program parallel computing resources, which are supported by the University of Illinois.
(*Received 00 Month 200x; final version received 00 Month 200x*)

*Corresponding author.

ISSN: 0040-5167 print/ISSN 1754-2278 online
© 200x Taylor & Francis
DOI: 10.1080/0040516YYxxxxxxxx
http://www.informaworld.com

Abstract
Synthetic oligopeptides containing π-conjugated cores self-assemble novel materials with attractive electronic and photophysical properties. All-atom, explicit solvent molecular dynamics simulations of Asp-Phe-Ala-Gly-OPV3-Gly-Ala-Phe-Asp peptides were used to parameterize an implicit solvent model to simulate early-stage self-assembly. Under low-pH conditions, peptides assemble into β-sheet-like stacks with strongly favorable monomer association free energies of $\Delta F \approx -25k_BT$. Aggregation at high-pH produces disordered aggregates destabilized by Coulombic repulsion between negatively charged Asp termini ($\Delta F \approx -5k_BT$). In simulations of hundreds of monomers over 70 ns we observe the spontaneous formation of up to undecameric aggregates under low-pH conditions. Modeling assembly as a continuous-time Markov process, we infer transition rates between different aggregate sizes and microsecond relaxation times for early-stage assembly. Our data suggests a hierarchical model of assembly in which peptides coalesce into small clusters over tens of nanoseconds followed by structural ripening and diffusion limited aggregation on longer time scales. This work provides new molecular-level understanding of early-stage assembly, and a means to study the impact of peptide sequence and aromatic core chemistry upon the thermodynamics, assembly kinetics, and morphology of the supramolecular aggregates.

Keywords: π-conjugated oligopeptides, molecular dynamics simulation, supramolecular peptides, self-assembly, Markov state model

PACS numbers: 87.14.ef, 87.15.A-, 87.10.Tf, 87.15.ap, 64.75.Yz

1. Introduction

The self-assembly of peptides is a promising methodology for the fabrication of novel macromolecular materials with desirable structural, functional, and biological properties [1, 2]. Such materials have potential applications in drug and protein delivery, material templating of inorganic structures, regenerative medicine, and antimicrobials [1–8]. Assembly can be triggered by environmental variables such as pH, temperature, and salt concentration [9–11]. The morphology of self-assembled peptidic aggregates can be tuned by the amino acid sequence, and endowed with useful electronic and optoelectronic properties by the incorporation of aromatic groups into the peptide backbone that establish inter-peptide π-conjugation upon assembly [10, 12, 13]. 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 [14–18]. 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 [19]. Conjugated peptides in particular offer a water soluble and biofunctional medium to fabricate self-assembled aggregates with tunable biological and electronic properties [10, 13, 17, 18, 20].

In this work, we perform 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 (Fig. 1). This prototypical peptide-Π-peptide triblock architecture presents a powerful and flexible architecture for self-assembling optoelectronic peptidic biomaterials [13, 21]. 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 [13, 22–24]. 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) [25, 26]. 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 [10]. 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 [13, 21, 22, 27, 28]. 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 [13, 21], but the elementary kinetic steps and molecular mechanisms of self-assembly remain unknown. This work establishes fundamental understanding of the early stages of DFAG-OPV3-GAFD self-assembly using molecular simulation.

[Figure 1 near here.]

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 work, 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 [29]. Building upon a body of prior experimental and simulation work [13, 22, 24], 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. Methods

2.1. *Explicit solvent simulations*

Molecular dynamics simulations were conducted using the GROMACS 4.6 simulation suite [30]. Initial DFAG-OPV3-GAFD peptide configurations were constructed with the assistance of the GlycoBioChem PRODRG2 Server [31]. 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 ≤ 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 [13]. Single peptides or pairs of peptides were placed in a $10\times10\times10$ 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 [32], and water with the simple point charge (SPC) model [33]. 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 2.3). 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} [34, 35]. 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 [36]. Bond lengths were fixed using the LINCS algorithm to improve efficiency [37]. 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 [38]. 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 [39]. 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. *Implicit solvent simulations*

Simulations were conducted using the GROMACS 4.6 simulation suite [30] in which solvent was modeled implicitly using the Generalized Born model with a relative dielectric constant of 78.3 [40–42]. 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 $\gamma = 4.85$ [43], 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 [44]. The Generalized Born model has been known to overestimate the stability of inter-residue interactions [45–47], but has also been shown to accurately treat solvent effects in simulating proteins [48–50] even when such proteins are stabilized by solvent effects [51, 52]. 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 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.

2.3. *Umbrella sampling*

We employed umbrella sampling to compute the the intramolecular and intermolecular peptide free energy landscapes along a preselected order parameter by applying artificial biasing potentials to enforce good sampling [53]. 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_α-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) [54] implemented in the g_wham module of GROMACS 4.6 [30] was applied to the umbrella sampling data to obtain the (relative) free energy of the system as a function of the umbrella coordinate. 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.

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 the long time and length scales required to observe peptide assembly. 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.

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 Fig. 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 near here.]

The PMF curves for peptides in the deprotonated (high-pH) state in Fig. 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 $\sim$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.

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 Fig. 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. [55], we find simulations conducted in implicit solvent greatly overestimate the free energy computed in explicit solvent, in this case by 40%.

[Figure 3 near here.]

The PMF curves for deprotonated peptides are presented in Fig. 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 [29], 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 Section 3.3 a systematic procedure to perform this reparameterization.

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 (Fig. 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 [56–59]. In a similar manner to Zhang *et al.* [60], 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 [55, 61, 62]. 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 [59, 63]. In the case of pair potentials, this reduces to matching of the pairwise PMFs [64]. More sophisticated procedures to adjust the implicit solvent force field are possible [61, 63], 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,

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

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 [42]. The dispersion interaction between atoms i and j is given by the Lennard-Jones function,

$$V_{LJ} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{2}$$

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 $\epsilon_{ij} = \sqrt{(\epsilon_i \epsilon_j)}$ [39]. 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,

$$\epsilon \to \alpha\epsilon \qquad q \to \sqrt{\alpha} q, \tag{3}$$

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

$$V^{nb} = V_C + V_{LJ} \to \alpha V^{nb}, \tag{4}$$

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,

$$\Delta\Delta F_{dimer}(\alpha) = |\Delta F^{\text{imp}}_{dimer}(\alpha) - \Delta F^{\text{exp}}_{dimer}|. \tag{5}$$

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 $\alpha = 0.6$-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 $\alpha = \{0.80, 0.82, 0.85, 0.87, 0.90, 1.00\}$ for protonated peptides and $\alpha = \{0.52, 0.60, 0.68, 0.70, 0.72, 0.76, 0.84, 0.92, 1.00\}$ for deprotonated peptides. We illustrate in Fig. 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 $\alpha^*_{\text{prot}} = 0.87$ accurately matches the explicit solvent dimerization free energy. For the deprotonated peptide, we determine an optimal value of $\alpha^*_{\text{deprot}} = 0.72$.

[Figure 4 near here.]

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 Fig. 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 Fig. 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 work, so the slightly poorer quality of the deprotonated reparameterization is not of undue concern. The PMFs for peptide collapse (Fig. 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.

[Figure 5 near here.]

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.

3.4.1. 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 Fig. 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 Fig. 5, the dimerization free energy is highly favorable at $\Delta F_{1+1} = -(14.5 \pm 1.2)\ k_BT$. The successive addition of two more monomers is also strongly spontaneous at $\Delta F_{2+1} = -(24.5 \pm 2.4)\ k_BT$ and $\Delta F_{3+1} = -(23.6 \pm 3.0)\ k_BT$, 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 $\Delta F_{2+2} = -(27.6 \pm 3.7)\ k_BT$. The free energy for tetramer formation 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 Fig. 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 [13, 22, 23].

[Figure 6 near here.]

In Fig. 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. Fig. 7 shows the monomeric addition free energy plateauing surprisingly quickly at the trimer, with $\Delta F_{2+1} \approx \Delta F_{3+1} \approx \Delta F_{4+1} \approx -25 k_BT$ 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 7 near here.]

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,

$$\begin{aligned}\Delta F_n =& \Delta U_n^{\text{intrapeptide}} + \Delta U_n^{\text{peptide-peptide}} + \Delta U_n^{\text{peptide-water}} \\ &+ \Delta U_n^{\text{water-water}} - T\Delta S_n + P\Delta V,\end{aligned} \tag{6}$$

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 computed by summing the monomeric addition free energies reported in Figs. 6 and 7. By performing an energetic analysis of our equilibrated peptide monomers, dimers, trimers, tetramers, and pentamers residing at the global free energy minimum of our 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 [65] we define the solvent-mediated contribution to the aggregation free energy as,

$$\begin{aligned}\Delta F_n^{\text{solvent}} &= \Delta U_n^{\text{peptide-water}} + \Delta U_n^{\text{water-water}} - T\Delta S_n^{\text{water}} \\ &\approx \Delta F_n - \Delta U_n^{\text{intrapeptide}} - \Delta U_n^{\text{peptide-peptide}}.\end{aligned} \tag{7}$$

We plot in Fig. 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 [65–67].

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,

$$\Delta U_n^{\text{peptide}-\text{peptide}} = \Delta U_n^{\text{wing}-\text{wing}} + \Delta U_n^{\text{core}-\text{core}} + \Delta U_n^{\text{wing}-\text{core}}, \tag{8}$$

where $\Delta U_n^{\text{wing}-\text{wing}}$ accounts for intermolecular dispersion and electrostatic interactions between the amino acid wings of all of the peptides in the aggregate of size n, $\Delta U_n^{\text{core}-\text{core}}$ is the interaction between the conjugated cores, and $\Delta U_n^{\text{wing}-\text{core}}$ accounts for the cross interactions between the wings and cores. We plot this decomposition in Fig. 8b. Interactions between the aromatic cores accounts 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. Fig. 6).

[Figure 8 near here.]

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 Fig. 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 our prior work [13, 21], 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 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 Fig. 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 [68, 69]. We detect hydrogen bonds using the g_hbond program within the GROMACS 4.6 simulation suite [30]. 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. Fig. 8).

[Figure 9 near here.]

3.4.2. 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 Fig. 10 the formation of dimers is slightly favorable with $\Delta F_{1+1} = -(4.0 \pm 0.7)\ k_BT$. Surprisingly, further monomeric addition continues to be favorable with $\Delta F_{2+1} = -(6.3 \pm 0.9)\ k_BT$, $\Delta F_{3+1} = -(6.0 \pm 0.8)\ k_BT$, and $\Delta F_{4+1} = -(4.6 \pm 1.2)\ k_BT$. 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 ordered 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 10 near here.]

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 [70]. 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 [71–73]. We performed five independent 20 ns simulations of a single isolated peptide in implicit solvent using the parameters given in Section 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 [74]. 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 2.1, 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 [75, 76], that has been shown to reliably approximate the true dynamical time scales of the atomic system [77]. 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 [78–80] resulting in dynamical time scales in good agreement with explicit solvent simulations.

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 [13, 21, 22]. Using the simulation parameters detailed in Section 2.2, 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 [22]. (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 unattainable by explicit solvent models, our implicit solvent model permits direct observation of peptide collision, dissociation, and structural relaxation events over the first 70 ns of assembly. Movies of one of our assembly trajectories at each concentration were rendered using VMD [81] and are provided in the Supplementary Information.

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 $r_{\mathrm{cut}} = 0.5$ 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. Fig. 5).

3.6.1. Low concentration

We present in Fig. 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 an 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 ∼10 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 1).

[Figure 11 near here.]

In Fig. 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 12 near here.]

3.6.2. High concentration

We present in Figs. 13 and 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, trimers, and tetramers Fig. 13, but relatively poorer agreement in the case of the higher order aggregates for which far fewer assembly or disassembly events occur Fig. 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. Fig. 15 and Table 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 13 near here.]

[Figure 14 near here.]

In Fig. 15 we show snapshots of a sequence of aggregation events leading to the formation of a 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 precise structural ordering. This finding suggests that larger aggregates may be 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 [13, 22–24].

[Figure 15 near here.]

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 [82–89]. 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 [84, 85, 90] – 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 [84].

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 ith element of this vector defines the probability with which a peptide monomer exists in an aggregate of size i [91]. 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 $r_{\mathrm{cut}} = 0.5$ 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, $q_{ii} = -\sum_{j=1}^{n} q_{ij}$ [82]. 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 [82]. 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 [82, 83, 90],

$$\mathbf{p}(t) = \mathbf{p}(0)e^{\mathbf{Q}t} = \mathbf{p}(0)\mathbf{T}(t), \tag{9}$$

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 [83, 84, 88]. We estimate the matrix exponential using the scaling squaring method of Lawson [92, 93],

$$\mathbf{T}(t) = e^{\mathbf{Q}t} \approx \left[r_m(\mathbf{Q}t/2^s)\right]^{2^s}, \tag{10}$$

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 [94] 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 [84].

We extract maximum likelihood estimator of the off-diagonal elements of $\mathbf{Q}$ from our simulation trajectories using the expression [90],

$$q_{ij} = \frac{N_{ij}}{R_i}, \tag{11}$$

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 Eqn. 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 $q_{ii} = -\sum_{j=1}^{n} q_{ij}$ [82]. We then employ Eqn. 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 Eqn. 9 and those measured directly from the simulation trajectory.

In Appendix A 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 B 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 [84].

3.7.1. Low concentration

In Table 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 Fig. 16, we compare the time evolution of the cluster size distribution predicted from the transition rate matrix using Eqn. 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, $1/|q_{ii}|$, to the characteristic time for the structural relaxation of aggregates of this size subsequent to their formation (c.f., Figs. 12 and 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 $\mathbf{T}$ as [84, 95],

$$t_i = -\frac{\tau}{\ln(|\lambda_i|)}, \tag{12}$$

where τ is the lag time used to construct the transition matrix and λ_i is a (possibly complex) eigenvalue of $\mathbf{T}$ 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 [96], and may therefore be related to experimental measurements of structural relaxation kinetics [89, 97]. 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 distributions with mean and standard deviation reported in Table 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 [88, 96], and suggests that simulations on the order of microseconds would be required to probe beyond early-stage assembly [96]. 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 1 near here.]

[Figure 16 near here.]

3.7.2. High concentration

In Table 2 and Fig. 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 $\tau = 100$ 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 or similar to 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 1 and 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} differs by only 0.4 μs^{-1} after error, and q_{13}, and q_{23} by $\sim$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 work [96].

[Table 2 near here.]

[Figure 17 near here.]

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. Aggregates exist as β-sheet-like stacks mediated by π-stacking of the aromatic cores and hydrogen bonding between the peptidic wings. We predict a favorable dimerization free energy of $\Delta F \approx -15 k_B T$, and compute a strongly favorable and approximately constant driving force for further monomer addition to the elongating stack of $\Delta F \approx -25 k_B T$. Interestingly, under high-pH conditions where the peptides are deprotonated and carry a net (-4) formal charge, we find dimerization to remain favorable with $\Delta F \approx -4 k_B T$ 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 $\Delta F \approx -5 k_B T$, 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 [10, 13, 21, 22].

The reparameterized implicit solvent model permitted us to access to 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 several tens of 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 and structural relaxation of of heavier aggregates proceeds on time scales of tens 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 [98–103]. 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 $(0.74 \pm 0.02) \times 10^{-5}$ 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. We are currently working to both secure computing time on large-scale cyberinfrastructure and to develop more highly coarse-grained models to perform these simulations.

In future work, we will employ 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 (i) the DXXX-OPV3-XXXD peptide family where X = {G,A,I,V,F} [13, 21, 22, 24, 28], and (ii) the DFAG-Π-GAFD family where Π corresponds to different conjugated cores of different characters and sizes, including oligophenylvinylenes, oligothiophenes, and rylene diimides [21, 23, 27] 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 [13, 20]. 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 [20]. 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.

References

[1] A. Dehsorkhi, V. Castelletto, and I. W. Hamley, J. Pept. Sci. **20**, 453 (2014).
[2] M. R. Caplan and D. A. Lauffenburger, Ind. Eng. Chem. Res. **41**, 403 (2001).
[3] V. F. Segers and R. T. Lee, Drug Discovery Today **12**, 561 (2007).
[4] K. Subramani and W. Ahmed, in *Micro and Nano Technologies*, edited by K. Subramani and W. Ahmed (William Andrew Publishing, Boston, 2012) pp. 209–224.
[5] H. Hosseinkhani, P.-D. Hong, and D.-S. Yu, Chem. Rev. **113**, 4837 (2013).
[6] L. C. Palmer, C. J. Newcomb, S. R. Kaltz, E. D. Spoerke, and S. I. Stupp, Chem. Rev. **108**, 4754 (2008).
[7] J. D. Hartgerink, E. Beniash, and S. I. Stupp, Science **294**, 1684 (2001).
[8] A. Aggeli, M. Bell, N. Boden, J. N. Keen, P. F. Knowles, T. C. B. McLeish, M. Pitkeathly, and S. E. Radford, Nature **386**, 259 (1997).
[9] M. A. Gauthier and H.-A. Klok, Chem. Commun. **23**, 2591 (2008).
[10] M. Zelzer and R. V. Ulijn, Chem. Soc. Rev. **39**, 3351 (2010).
[11] D. W. P. M. L owik, E. H. P. Leunissen, M. van den Heuvel, M. B. Hansen, and J. C. M. van Hest, Chem. Soc. Rev. **39**, 3394 (2010).
[12] S. Tsonchev, G. C. Schatz, and M. A. Ratner, J. Phys. Chem. B **108**, 8817 (2004).
[13] B. D. Wall, A. E. Zacca, A. M. Sanders, W. L. Wilson, A. L. Ferguson, and J. D. Tovar, Langmuir **30**, 5946 (2014).
[14] 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, Nature **347**, 539 (1990).
[15] L. Bian, E. Zhu, J. Tang, W. Tang, and F. Zhang, Prog. Polym. Sci. **37**, 1292 (2012).
[16] X. Guo, M. Baumgarten, and K. Müllen, Prog. Polym. Sci. **38**, 1832 (2013).
[17] S. H. Kim and J. R. Parquette, Nanoscale **4**, 6940 (2012).
[18] F. J. M. Hoeben, P. Jonkheijm, E. W. Meijer, and A. P. H. J. Schenning, Chem. Rev. **105**, 1491 (2005).
[19] White House Office of Science and Technology Policy, "Materials Genome Initiative for Global Competitiveness," (2011).
[20] H. A. M. Ardoña, K. Besar, M. Togninalli, H. E. Katz, and J. D. Tovar, J. Mater. Chem. C **3**, 6505 (2015).
[21] B. D. Wall, Y. Zhou, S. Mei, H. A. M. Ardoña, A. L. Ferguson, and J. D. Tovar, Langmuir **30**, 11375 (2014).
[22] A. B. Marciel, M. Tanyeri, B. D. Wall, J. D. Tovar, C. M. Schroeder, and W. L. Wilson, Adv. Mater. **25**, 6398 (2013).
[23] G. S. Vadehra, B. D. Wall, S. R. Diegelmann, and J. D. Tovar, Chem. Commun. **46**, 3947 (2010).
[24] B. D. Wall and J. D. Tovar, Pure Appl. Chem. **84**, 1039 (2012).
[25] A. M. Sanders, T. J. Dawidczyk, H. E. Katz, and J. D. Tovar, ACS Macro Letters **1**, 1326 (2012).
[26] H. A. M. Ardoña and J. D. Tovar, *Bioconjugate Chemistry*, Bioconjugate Chem. **in press (doi: 10.1021/acs.bioconjchem.5b00497)**, (2015).
[27] S. R. Diegelmann, J. M. Gorham, and J. D. Tovar, J. Am. Chem. Soc. **130**, 13840 (2008).
[28] B. D. Wall, S. R. Diegelmann, S. Zhang, T. J. Dawidczyk, W. L. Wilson, H. E. Katz, H.-Q. Mao, and J. D. Tovar, Adv. Mater. **23**, 5009 (2011).
[29] J. Mondal, X. Zhu, Q. Cui, and A. Yethiraj, J. Phys. Chem. C **114**, 13551 (2010).
[30] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[31] A. W. Schüttelkopf and D. M. F. van Aalten, Acta Crystallogr., Sect D: Biol. Crystallogr. **60**, 1355 (2004).
[32] A. D. MacKerell and N. K. Banavali, J. Comput. Chem. **21**, 105 (2000).
[33] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, *Intermolecular Forces* (Reidel, Dordrecht, 1981).
[34] W. F. Van Gunsteren and H. J. C. Berendsen, Mol. Simulat. **1**, 173 (1988).
[35] N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink, and H. J. C. Berendsen, J. Chem. Theory Comput. **8**, 3637 (2012).
[36] R. W. Hockney and J. W. Eastwood, *Computer Simulation using Particles* (CRC Press, 2010).
[37] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[38] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).

[39] M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids* (Oxford University Press, 1989).
[40] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, J. Am. Chem. Soc. **112**, 6127 (1990).
[41] R. Constanciel and R. Contreras, Theor. Chim. Acta **65**, 1 (1984).
[42] D. P. Fernández, Y. Mulev, A. R. H. Goodwin, and J. M. H. L. Sengers, J. Phys. Chem. Ref. Data **24**, 33 (1995).
[43] A. Onufriev, D. Bashford, and D. A. Case, Proteins: Struct., Funct., Bioinf. **55**, 383 (2004).
[44] M. Schaefer, C. Bartels, and M. Karplus, J. Mol. Biol. **284**, 835 (1998).
[45] R. G. Endres, *Molecular Simulation*, Molecular Simulation **31**, 773 (2005).
[46] Y. Pak, E. Kim, and S. Jang, The Journal of Chemical Physics **121**, 9184 (2004).
[47] D. R. Roe, A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling, *The Journal of Physical Chemistry B*, J. Phys. Chem. B **111**, 1846 (2007).
[48] T. Z. Lwin, R. Zhou, and R. Luo, The Journal of Chemical Physics **124**, 034902 (2006).
[49] F. Yaşar, P. Jiang, and U. H. Hansmann, Europhysics Letters **105**, 30008 (2014).
[50] G. Andreotti, I. Cabeza de Vaca, A. Poziello, M. C. Monti, V. Guallar, and M. V. Cubellis, Journal of Biological Chemistry **289**, 34900 (2014), http://www.jbc.org/content/289/50/34900.full.pdf+html .
[51] M. S. Shell, R. Ritterson, and K. A. Dill, *The Journal of Physical Chemistry B*, J. Phys. Chem. B **112**, 6878 (2008).
[52] H. R. Bureau, D. R. Merz, E. Hershkovits, S. Quirk, and R. Hernandez, PLoS ONE **10**, e0127034 (2015).
[53] G. Torrie and J. Valleau, J. Comput. Phys. **23**, 187 (1977).
[54] J. S. Hub, B. L. de Groot, and D. van der Spoel, J. Chem. Theory Comput. **6**, 3713 (2010).
[55] A. R. Brice and B. N. Dominy, Commun. Comput. Phys. **13**, 223 (2013).
[56] O. Allnér, L. Nilsson, and A. Villa, J. Chem. Theory Comput. **8**, 1493 (2012).
[57] L. Wang, B. E. Hingerty, A. Srinivasan, W. K. Olson, and S. Broyde, Biophys. J. **83**, 382 (2002).
[58] J. R. Spaeth, I. G. Kevrekidis, and A. Z. Panagiotopoulos, J. Chem. Phys. **134**, 164902 (2011).
[59] C. Maffeo, T. T. M. Ngo, T. Ha, and A. Aksimentiev, J. Chem. Theory Comput. **10**, 2891 (2014).
[60] L. Y. Zhang, E. Gallicchio, R. A. Friesner, and R. M. Levy, J. Comput. Chem. **22**, 591 (2001).
[61] A. Villa, C. Peter, and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. **11**, 2077 (2009).
[62] T. Sanghi and N. R. Aluru, J. Chem. Phys. **136**, 024102 (2012).
[63] W. G. Noid, J. Chem. Phys. **139**, 090901 (2013).
[64] J. C. Rasaiah, "Encyclopedia of chemical physics and physical chemistry: Fundamentals," (Taylor & Francis, 2001) Chap. A2.3 Statistical mechanics of strongly interacting systems: liquids and solids.
[65] D. Chandler, Nature **437**, 640 (2005).
[66] K. Lum, D. Chandler, and J. D. Weeks, *The Journal of Physical Chemistry B*, J. Phys. Chem. B **103**, 4570 (1999).
[67] M. V. Athawale, G. Goel, T. Ghosh, T. M. Truskett, and S. Garde, Proceedings of the National Academy of Sciences **104**, 733 (2007), http://www.pnas.org/content/104/3/733.full.pdf .
[68] R. Kumar, J. R. Schmidt, and J. L. Skinner, The Journal of Chemical Physics **126**, 204107 (2007).
[69] B. Hess, D. van der Spoel, and E. Lindahl, *GROMACS User Manual*, Royal Institute of Technology and Uppsala Univerity, 4th ed. (2013).
[70] N. Baker, D. Bashford, and D. Case, in *New Algorithms for Macromolecular Simulation*, Lecture Notes in Computational Science and Engineering, Vol. 49, edited by B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, and R. Skeel (Springer Berlin Heidelberg, 2006) pp. 263–295.
[71] R. D. Groot, J. Chem. Phys. **118**, 11265 (2003).
[72] R. Groot and K. Rabone, Biophys. J. **81**, 725 (2001).
[73] S. J. Marrink, A. H. de Vries, and A. E. Mark, J. Phys. Chem. B **108**, 750 (2004).
[74] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic Press, 2001).
[75] Y.-F. Hu, W.-J. Lv, S. Zhao, Y.-Z. Shang, H.-L. Wang, and H.-L. Liu, Chemical Engineering Science **134**, 813 (2015).
[76] O. Gereben and L. Pusztai, *The Journal of Physical Chemistry B*, J. Phys. Chem. B **119**,

3070 (2015).
[77] J. E. Basconi and M. R. Shirts, *Journal of Chemical Theory and Computation*, J. Chem. Theory Comput. **9**, 2887 (2013).
[78] J. Swails, D. M. York, and A. E. Roitberg, Journal of Chemical Theory and Computation **10**, 1341 (2013).
[79] O. Coskuner and O. Wise-Scira, ACS Chemical Neuroscience **4**, 1549 (2013).
[80] A. Atzori, A. E. Baker, M. Chiu, R. A. Bryce, and P. Bonnet, PLoS ONE **8**, e68723 (2013).
[81] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[82] S. Karlin, *A First Course in Stochastic Processes*, revised ed. (Academic Press, 2014).
[83] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[84] G. R. Bowman, in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, edited by G. R. Bowman, V. S. Pande, and F. Noe (Springer, 2014) Chap. 2, pp. 7–22.
[85] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).
[86] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[87] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. C. Ward, Y. Zhestkov, *et al.*, J. Phys. Chem. B **108**, 6582 (2004).
[88] J. Chodera, W. Swope, J. Pitera, and K. Dill, Multiscale Model. Simul. **5**, 1214 (2006).
[89] J. D. Chodera and F. Noé, Curr. Opin. Struct. Biol. **25**, 135 (2014).
[90] Y. Inamura, "Estimating Continuous Time Transition Matrices From Discretely Observed Data," Bank of Japan Working Paper Series **No.06 − E07** (2006).
[91] M. Rubinstein and R. H. Colby, *Polymer Physics* (OUP Oxford, 2003).
[92] J. Lawson, SIAM J. Numer. Anal. **4**, 372 (1967).
[93] N. Higham, SIAM. J. Matrix Anal. & Appl. **26**, 1179 (2005).
[94] A. Al-Mohy and N. Higham, SIAM. J. Matrix Anal. & Appl. **31**, 970 (2009).
[95] M. Kijima, "Markov processes for stochastic modeling," (CRC Press, 1997) Chap. 2, p. 64.
[96] M. Sarich, J.-H. Prinz, and C. Schütte, in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, edited by G. R. Bowman, V. S. Pande, and F. Noe (Springer, 2014) Chap. 3, pp. 23–44.
[97] G. R. Bowman, V. S. Pande, and F. Noe, in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, edited by G. R. Bowman, V. S. Pande, and F. Noe (Springer, 2014) Chap. 1, pp. 1–6.
[98] J. D. Schmit, K. Ghosh, and K. Dill, Biophys. J. **100**, 450 (2011).
[99] W. Yong, A. Lomakin, M. D. Kirkitadze, D. B. Teplow, S.-H. Chen, and G. B. Benedek, Proc. Natl. Acad. Sci. USA **99**, 150 (2002).
[100] J. E. Gillam and C. E. MacPhee, J. Phys. Condens. Matter **25**, 373101 (2013).
[101] J. M. Andrews and C. J. Roberts, J. Phys. Chem. B **111**, 7897 (2007).
[102] Y. Li and C. J. Roberts, J. Phys. Chem. B **113**, 7020 (2009).
[103] V. Wagoner, M. Cheon, I. Chang, and C. Hall, Proteins **79**, 2132 (2011).
[104] M. Campbell and S. Farrell, *Biochemistry*, 8th ed. (Cengage Learning, 2014).

FIGURES

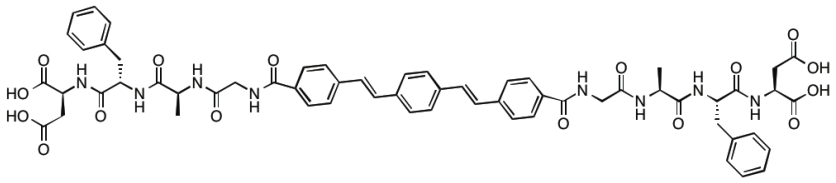

Figure 1. Chemical structure of the DFAG-OPV3-GAFD peptide studied in this work. The N-terminus to C-terminus directionality of each peptide wing proceeds away from the conjugated core, leading to an oligopeptide architecture possessing two C-termini. The pK_A of the carboxyl terminus and aspartic acid side chain residing at the C-termini are 2.09 and 3.86, respectively [104]. 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 [13]. 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.

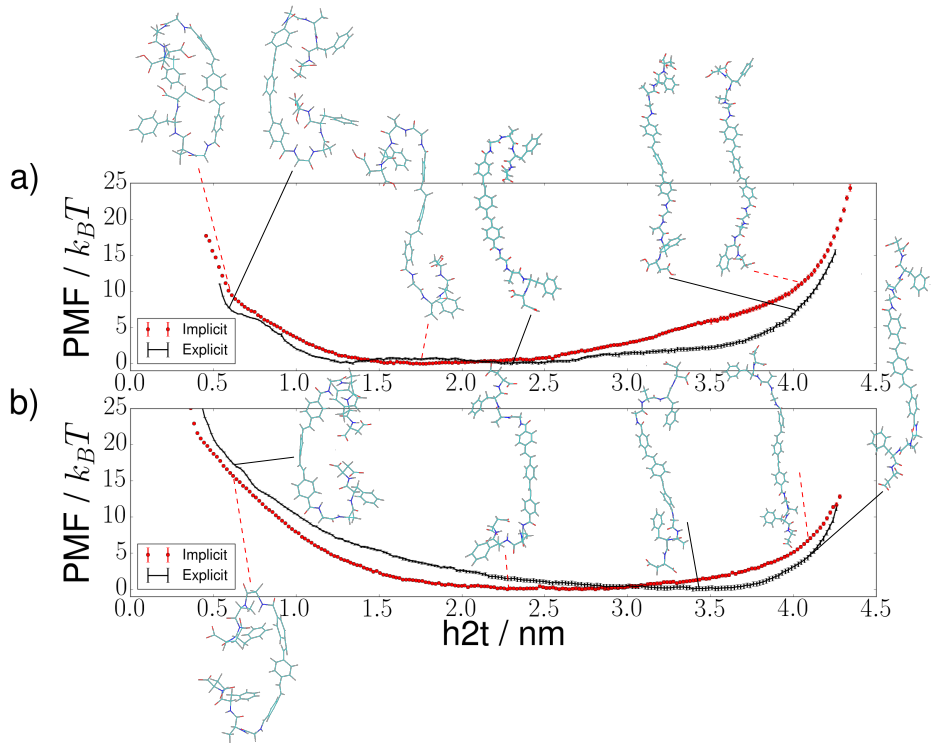

Figure 2. PMF curves for isolated DFAG-OPV3-GAFD peptides in explicit and implicit solvent in the (a) protonated (low-pH) and (b) deprotonated (high-pH) states parameterized by the intramolecular head-to-tail distance, $h2t$, between the C_α atoms of the terminal aspartic acid residues. Error bars in 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 [81].

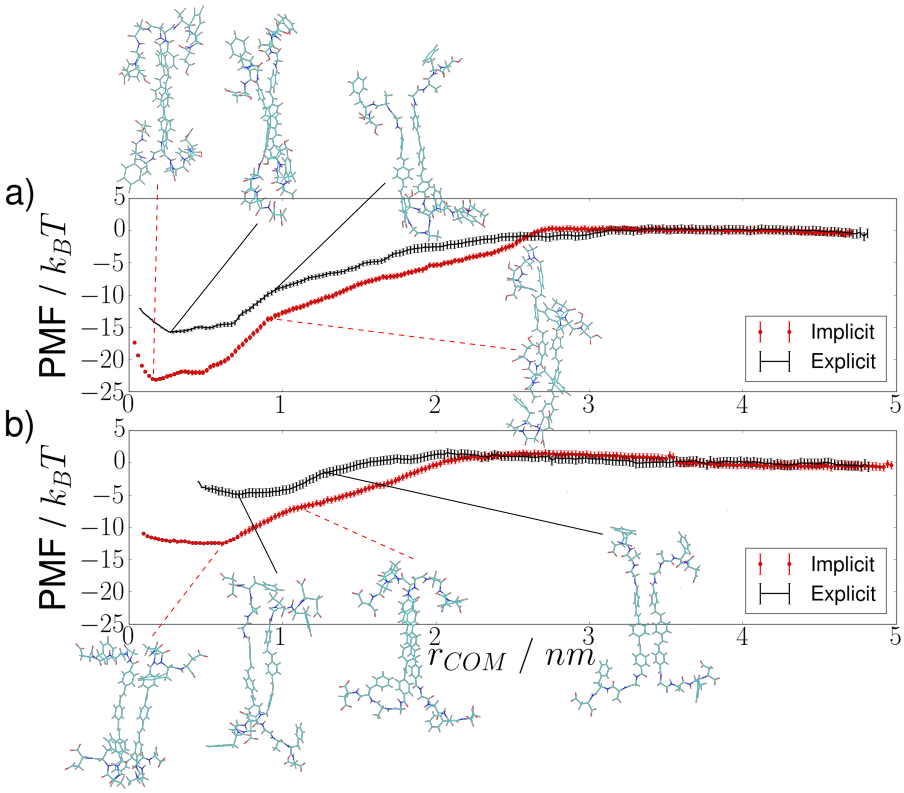

Figure 3. PMF curves for pairs of DFAG-OPV3-GAFD peptides in explicit and implicit solvent in the (a) protonated (low-pH) and (b) deprotonated (high-pH) states parameterized by peptide center of mass separation, r_{COM}. The dimerization free energy predicted by the explicit solvent model for the protonated (low-pH) peptides is strongly favorable at $\Delta F_{dimer}^{\mathrm{prot}} = -(15.7 \pm 0.7)\ k_BT$, whereas that for the deprotonated (high-pH) peptides is only weakly so at $\Delta F_{dimer}^{\mathrm{deprot}} = -(4.7 \pm 0.9)\ k_BT$.

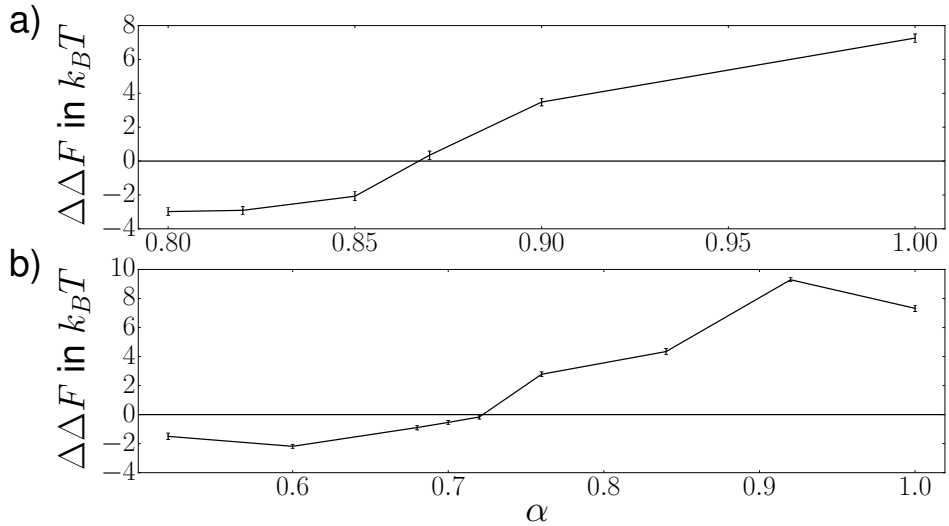

Figure 4. $\Delta\Delta F_{dimer}(\alpha) = |\Delta F^{\text{imp}}_{dimer}(\alpha) - \Delta F^{\text{exp}}_{dimer}|$ for (a) protonated peptides at values of the scaling parameter $\alpha = \{0.80, 0.82, 0.85, 0.87, 0.90, 1.00\}$, and (b) deprotonated peptides at $\alpha = \{0.52, 0.60, 0.68, 0.70, 0.72, 0.76, 0.84, 0.92, 1.00\}$.

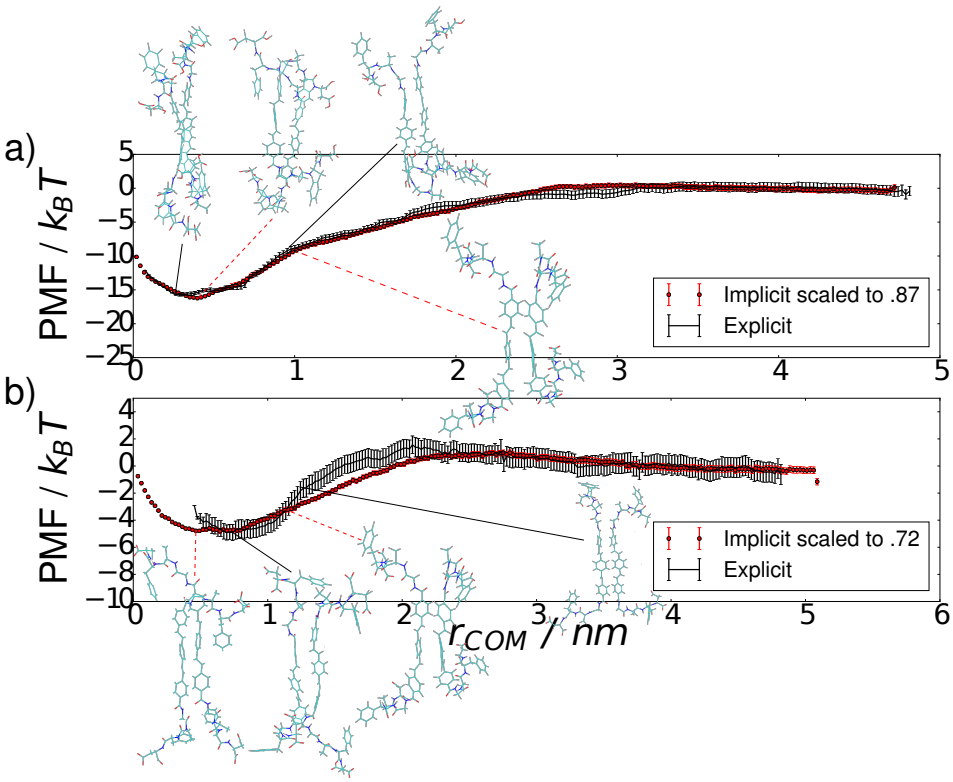

Figure 5. PMF curves for pairs of DFAG-OPV3-GAFD peptides in explicit and implicit solvent at the optimal value of (a) $\alpha^*_{\rm prot} = 0.87$ in the protonated (low-pH) and (b) $\alpha^*_{\rm deprot} = 0.72$ in the deprotonated (high-pH) states parameterized by peptide center of mass separation, $r_{\rm COM}$. The dimerization free energy predicted by the rescaled implicit solvent model for the protonated (low-pH) peptides is strongly favorable at $\Delta F^{\rm prot}_{dimer} = -(14.5 \pm 1.2)\ k_B T$, whereas that for the deprotonated (high-pH) peptides is only weakly so at $\Delta F^{\rm deprot}_{dimer} = -(4.0 \pm 0.7)\ k_B T$. Both values are in excellent agreement with the explicit solvent predictions of $\Delta F^{\rm prot}_{dimer} = -(15.7 \pm 0.7)\ k_B T$ and $\Delta F^{\rm deprot}_{dimer} = -(4.7 \pm 0.9)\ k_B T$.

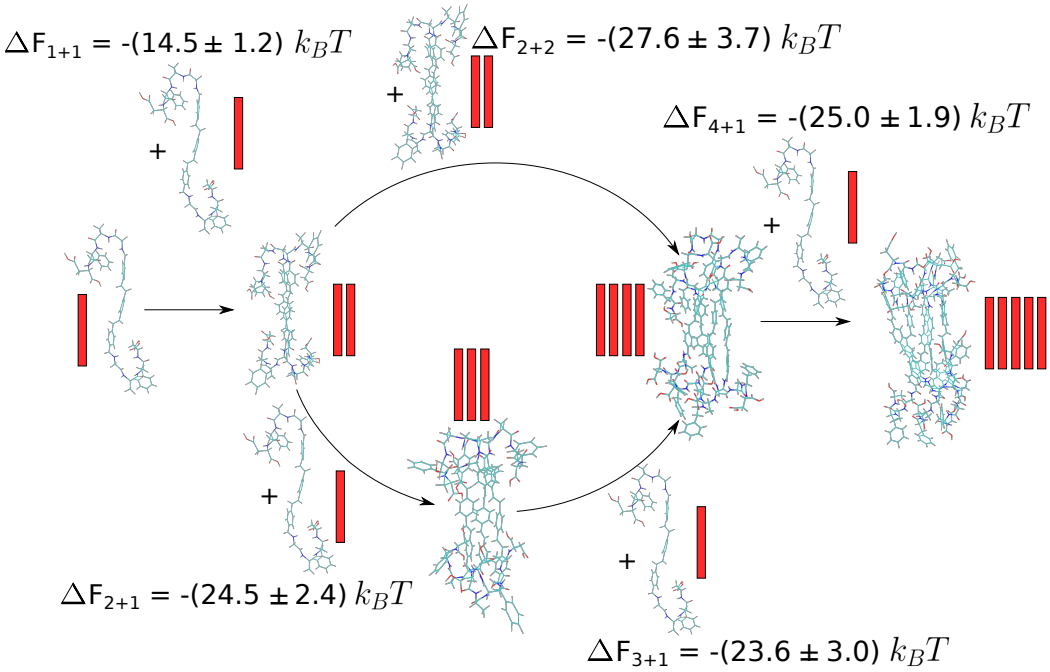

Figure 6. Elementary free energy changes for pentamer formation by monomeric addition (lower path) and dimer condensation plus monomeric addition (upper path) for protonated (low-pH) peptides computed by umbrella sampling using the reparameterized implicit solvent model. Uncertainties are computed by 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.

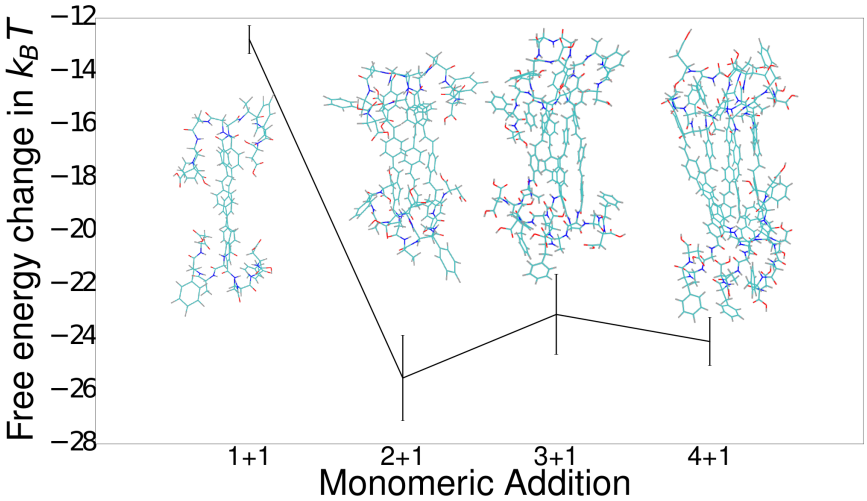

Figure 7. Free energy of monomeric addition to preassembled oligomers containing between one and four peptides under protonated (low-pH) conditions computed by umbrella sampling. Uncertainties are computed by standard propagation of errors over five independent runs accounting for the variance between 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 $\Delta F_{n+1} \approx -25 k_B T$.

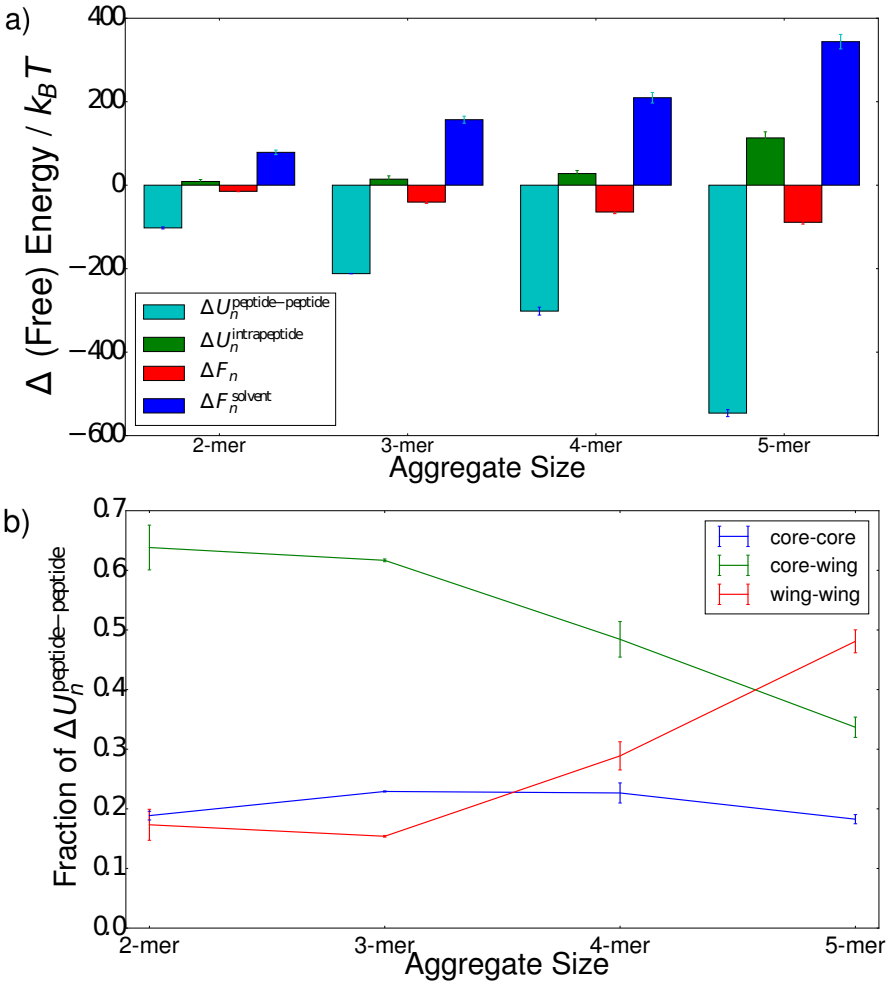

Figure 8. Decomposition of the energetic and entropic contributions to the free energy of association as a function of aggregate size for oligomers containing between two and five peptides. (a) The free energy of assembly of an aggregate of size n from non-interacting monomers ΔF_n and its decomposition into the change in the intramolecular, $\Delta U_n^{\text{intrapeptide}}$, and intermolecular, $\Delta U_n^{\text{peptide-peptide}}$, peptide energies, and the solvent-mediated contribution to aggregation, $\Delta F_n^{\text{solvent}} = \Delta U_n^{\text{peptide-water}} + \Delta U_n^{\text{water-water}} - T\Delta S_n^{\text{water}}$, which we estimate as $\Delta F_n^{\text{solvent}} \approx \Delta F_n - \Delta U_n^{\text{intrapeptide}} - \Delta U_n^{\text{peptide-peptide}}$, 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 $\Delta U_n^{\text{peptide-peptide}}$ 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.

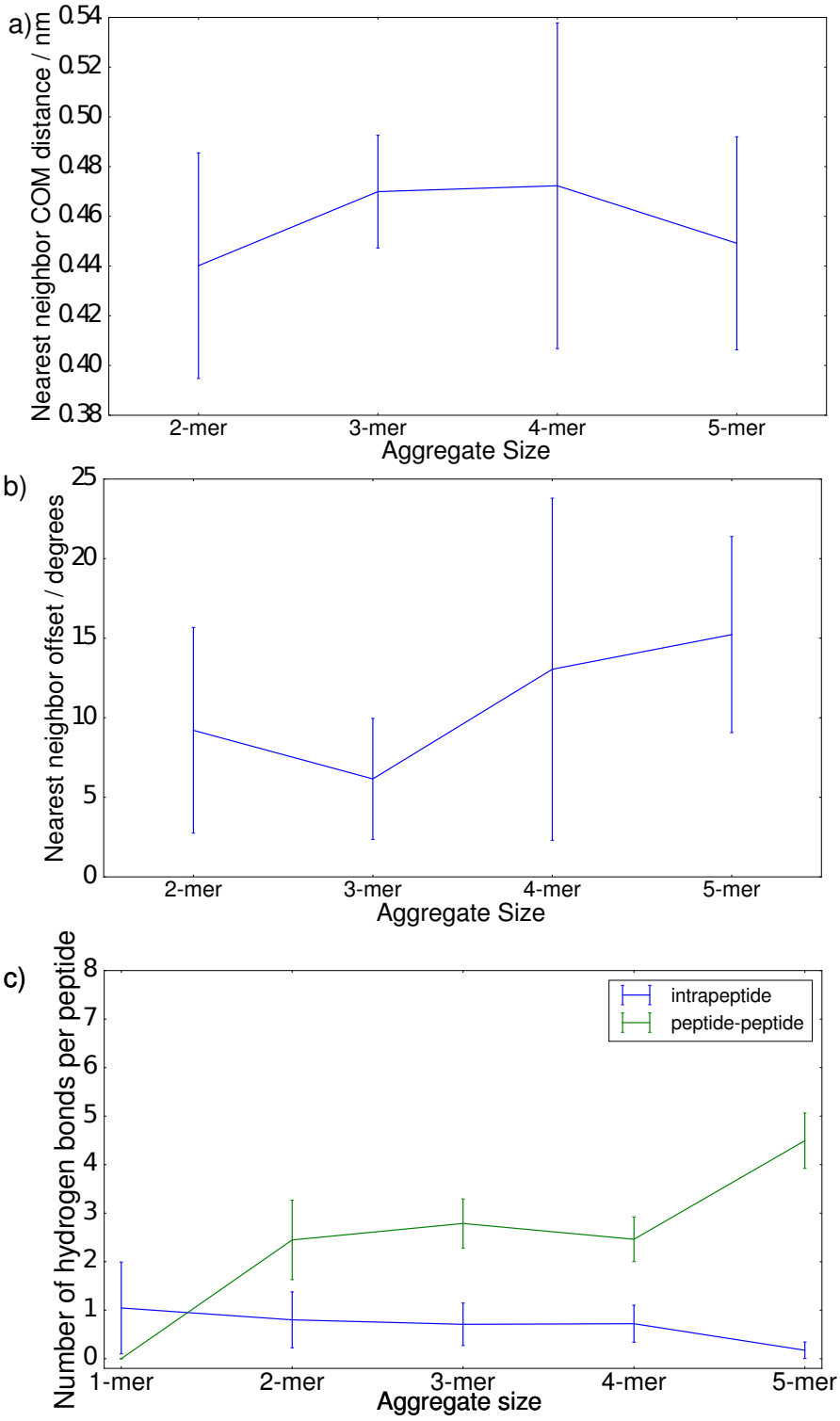

Figure 9. Structural characterization of oligomeric aggregates containing between two and five peptides. (a) Mean center of mass separation between nearest neighbor peptides in the aggregate. (b) Mean angular offset between nearest neighbor peptides computed by calculating the dot product between vectors 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.

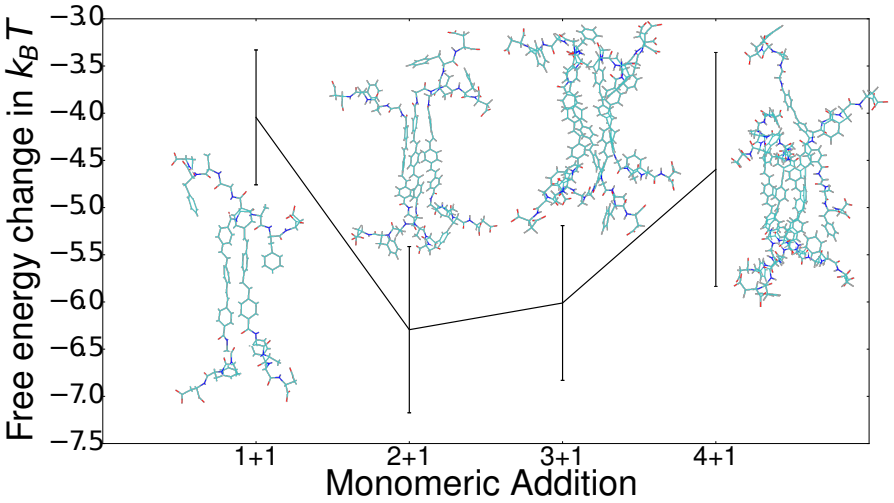

Figure 10. Free energy of monomeric addition to preassembled oligomers containing between one and four peptides under deprotonated (high-pH) conditions computed by umbrella sampling. Uncertainties are computed by standard propagation of errors over five independent runs accounting for the variance between 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.

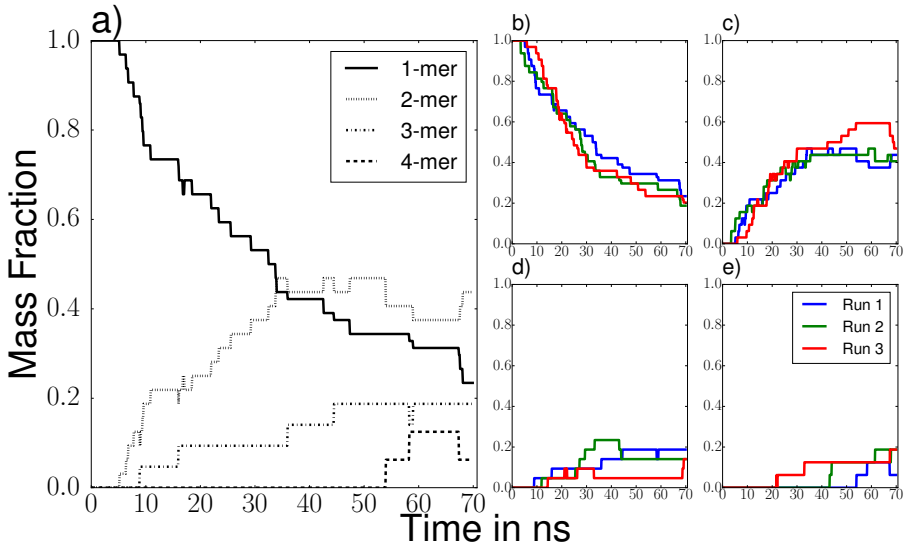

Figure 11. Mass fraction as a function of scaled simulation time for an unbiased simulation of 64 peptides at 0.85 mM for (a) one simulation, and (b-e) each of the three independent simulations tracking (b) 1-mers, (c) 2-mers, (d) 3-mers, and (e) 4-mers.

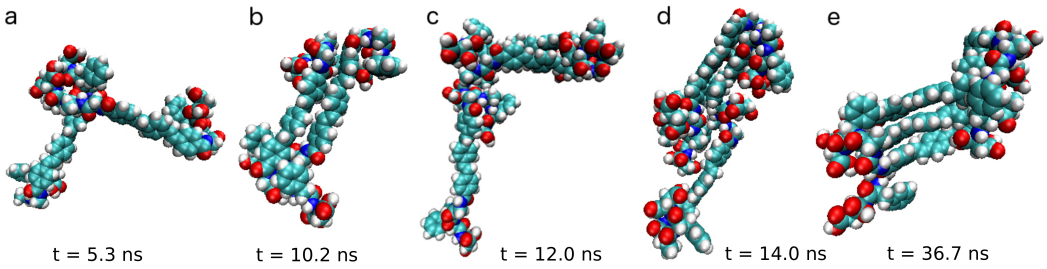

Figure 12. Representative snapshots of a trimerization event observed in Run 2 of the 0.85 mM concentration protonated (low-pH) peptide self-assembly simulations conducted over 70 ns using our reparameterized implicit solvent model. (a) Collision of two monomers at t = 5.3 ns induces rise to internal structural rearrangements until (b) the two conjugated cores are aligned to form the equilibrium dimer (cf. Fig. 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 at t = 36.7 ns (cf. Fig. 7).

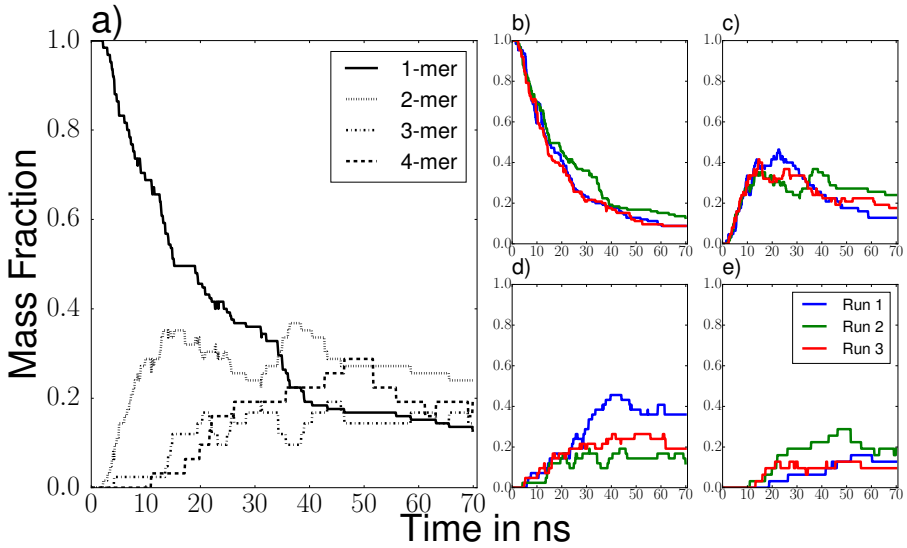

Figure 13. Mass fraction as a function of scaled simulation time for an unbiased simulation of 125 peptides at 1.66 mM for (a) one simulation, and (b-e) each of the three independent simulations tracking (b) 1-mers, (c) 2-mers, (d) 3-mers, and (e) 4-mers.

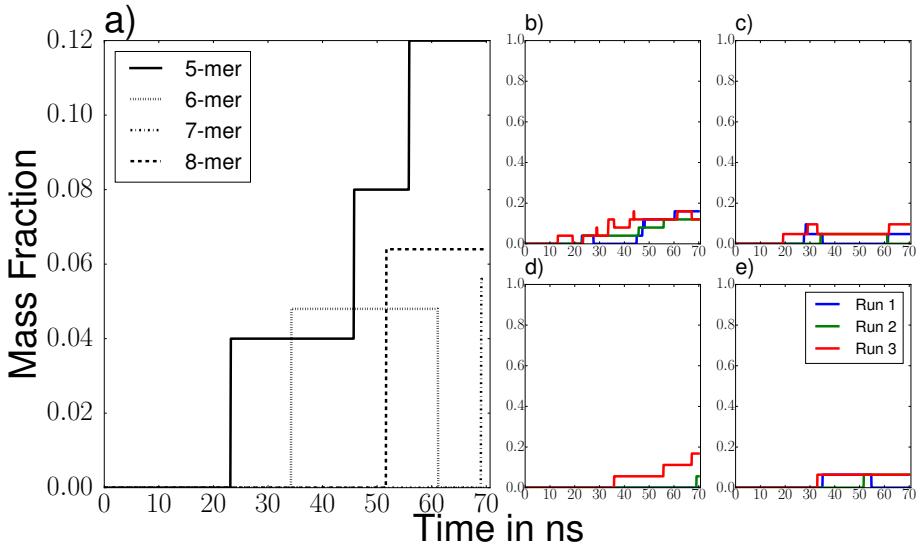

Figure 14. Mass fraction as a function of scaled simulation time for an unbiased simulation of 125 peptides at 1.66 mM for (a) one simulation, and (b-e) each of the three independent simulations tracking (b) 5-mers, (c) 6-mers, (d) 7-mers, and (e) 8-mers.

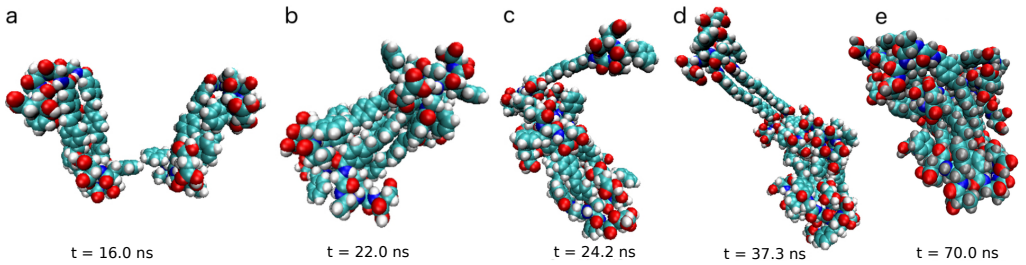

Figure 15. Representative snapshots of a heptamerization event observed in Run 3 of the 1.66 mM concentration protonated (low-pH) peptide self-assembly simulations conducted over 70 ns using our reparameterized implicit solvent model. (a) Collision of two pre-assembled dimers at t = 16 ns produces (b) a tetramer β-sheet after 6 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.

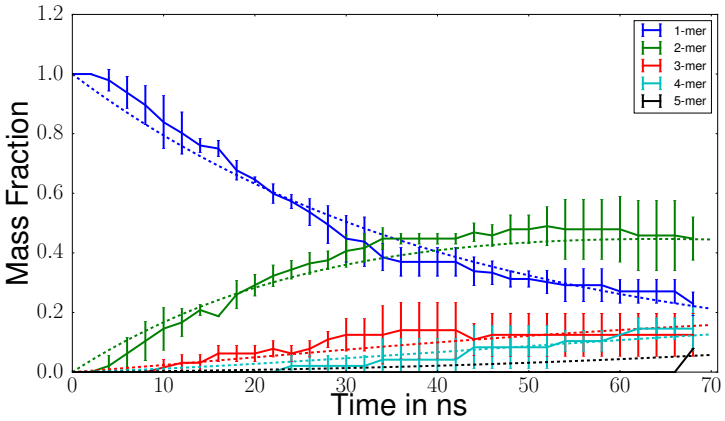

Figure 16. Comparison of the cluster size distribution in the 0.85 mM system predicted by the Markov model with a lag time of $\tau = 100$ ps (dashed lines) to that directly observed in the simulations (sold lines with every 20th point plotted) for aggregates of size n = 1-5. The simulation data is plotted 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.

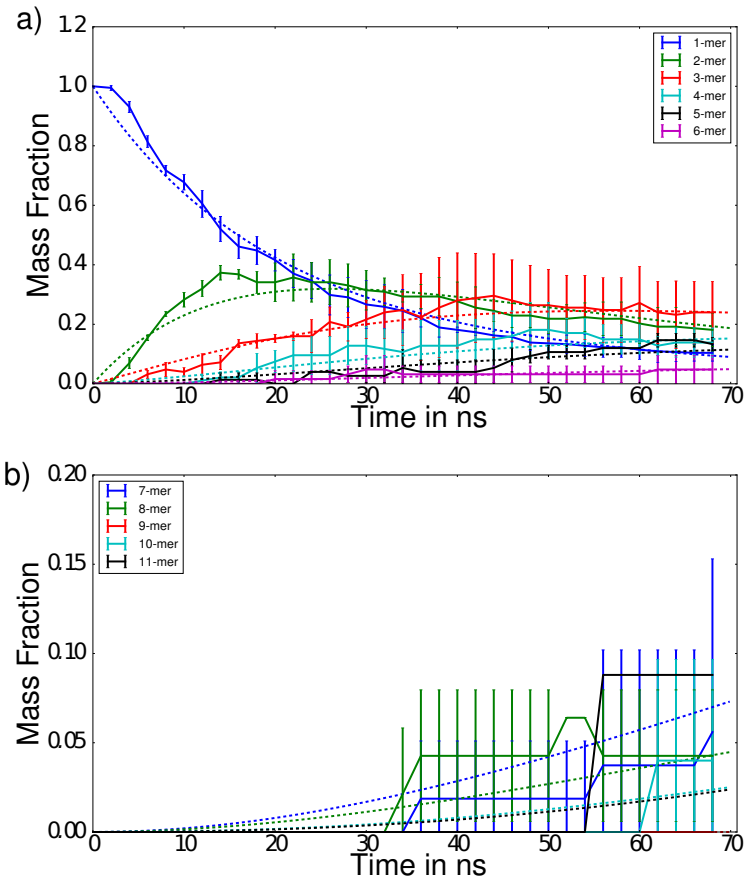

Figure 17. Comparison of the cluster size distribution in the 1.66 mM system predicted by the Markov model with a lag time of $\tau = 100$ ps (dashed lines) to that directly observed in the simulations (sold lines with every 20th point plotted) for aggregates of size (a) n = 1-6 and (b) n = 7-11. The 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.

TABLES

Table 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 $\tau = 100$ 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**
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

Table 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
From 1-mer	−41.3 ± 5.5	25.0 ± 4.4	11.7 ± 1.7	3.6 ± 0.9	2.1 ± 0.7
From 2-mer	25.0 ± 10.1	−56.3 ± 11.2	21.6 ± 2.9	4.6 ± 1.4	3.8 ± 1.4
From 3-mer	6.5 ± 4.5	6.8 ± 3.0	−42.9 ± 7.6	14.7 ± 3.5	8.4 ± 2.7
From 4-mer	1.8 ± 1.1	0.7 ± 0.7	5.5 ± 3.3	−28.0 ± 7.0	14.3 ± 5.1
From 5-mer	5.2 ± 4.8	13.3 ± 6.8	19.9 ± 10.2	20.9 ± 19.3	−107.0 ± 34.2
From 6-mer	18.9 ± 18.2	0.0 ± 0.0	8.3 ± 8.4	0.0 ± 0.0	94.3 ± 91.1
From 7-mer	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
From 9-mer	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
From 11-mer	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

...	To 6-mer	To 7-mer	To 8-mer	To 9-mer	To 10-mer	To 11-mer
	0.8 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	0.0 ± 0.0	0.7 ± 0.5	0.5 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	4.2 ± 2.2	1.2 ± 0.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.0 ± 1.1
	0.0 ± 0.0	2.6 ± 1.8	1.4 ± 1.4	0.0 ± 0.0	1.7 ± 1.7	0.0 ± 0.0
	42.5 ± 24.3	5.3 ± 3.8	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	−167.2 ± 102.3	0.0 ± 0.0	6.7 ± 6.8	0.0 ± 0.0	39.1 ± 40.3	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	−75.8 ± 83.3	0.0 ± 0.0	0.0 ± 0.0	75.8 ± 83.0
	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 ± 0.0	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

MOVIES

Movie 1. Movie of the 70 ns implicit solvent simulation of peptide self-assembly at a peptide concentration of 0.85 mM.

Movie 2. Movie of the 70 ns implicit solvent simulation of peptide self-assembly at a peptide concentration of 1.66 mM.

Appendix A. Time dependence of transition rate matrix

By modeling the system as a time homogeneous continuous time Markov chain, we implicitly assume that the elements of the transition rate matrix $\mathbf{Q}$ 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 Fig. A1, we present a time resolved plot of the estimates of the transition rate matrix elements for the 0.85 mM concentration and in Fig. A2 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 rate 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. Figs. 11, 13 and 14) supports the assertion that the dynamical evolution is adequately modeled as a time homogeneous Markov process.

[Figure A1 near here.]

[Figure A2 near here.]

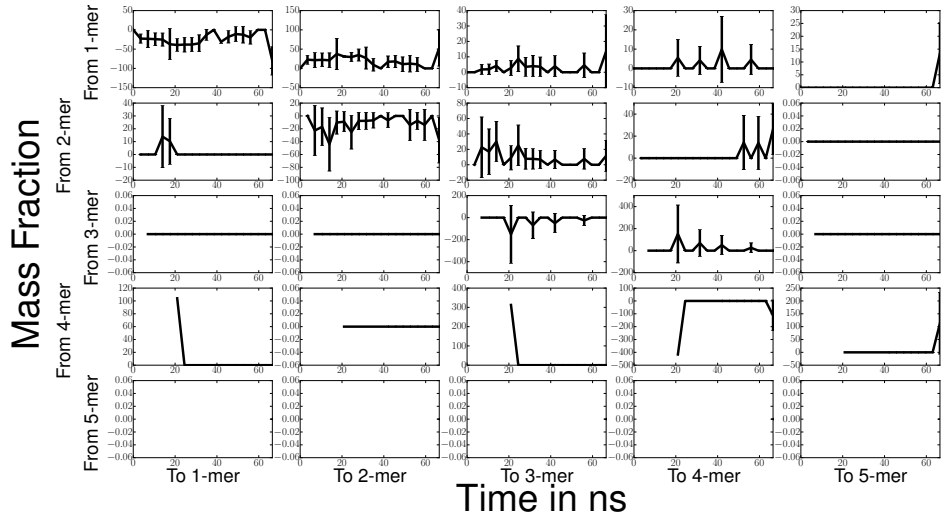

Figure A1. Time dependence of the transition rate matrix elements extracted from three 70 ns simulations of the aggregation of 64 protonated (low-pH) peptides at a concentration of 0.85 mM using our reparameterized implicit solvent model. Off-diagonal elements q_{ij} are reported over 3.5 ns time blocks 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.

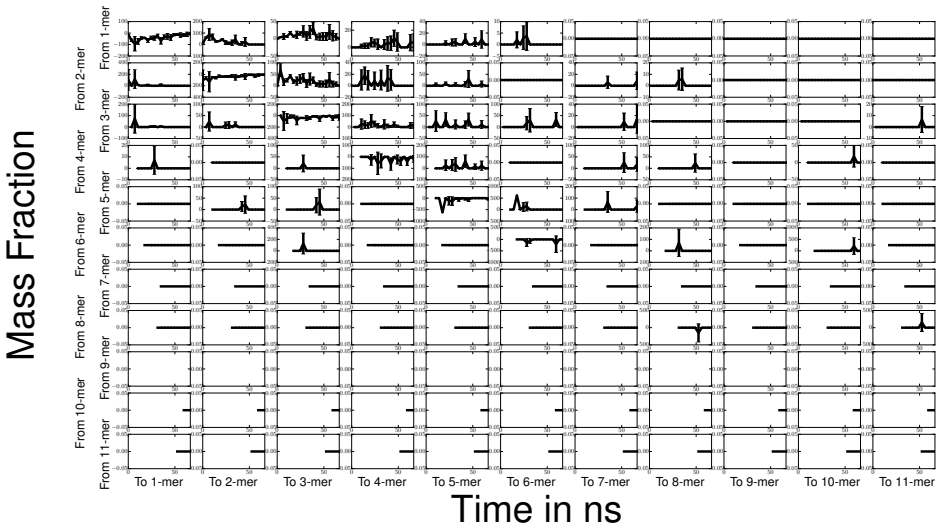

Figure A2. Time dependence of the transition rate matrix elements extracted from three 70 ns simulations of the aggregation of 125 protonated (low-pH) peptides at a concentration of 1.66 mM using our reparameterized implicit solvent model. Off-diagonal elements q_{ij} are reported over 3.5 ns time blocks 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.

Appendix B. Discrete time Markov chain (DTMC)

Our molecular simulation trajectories are continuous in time, making it possible to estimate the transition rate matrix $\mathbf{Q}$ 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 forming the matrix exponential $e^{\mathbf{Q}t}$ and applying Eqn. 9 [82]. 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. Eqn. 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 [84, 88],

$$\mathbf{p}(n\tau) = \mathbf{p}(0)\,[\mathbf{T}(\tau)]^n\,. \tag{B1}$$

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 [84, 88, 90],

$$t_{ij}(\tau) = \frac{N_{ij}(\tau)}{\sum_{m=1}^{M} N_{im}(\tau)}, \tag{B2}$$

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 [84]. The Chapman-Kolmogorov applied to a time homogeneous DTMC possessing the memoryless property states that [84],

$$\mathbf{T}(n\tau) = [\mathbf{T}(\tau)]^n\,, \tag{B3}$$

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 [84].

In Fig. B1, 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 Eqn. 9 and employing a lag time of 100 ps, and two DTMCs using Eqn. B1, and employing lag times of 100 ps and 400 ps. In Fig. B2 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 B1 near here.]

[Figure B2 near here.]

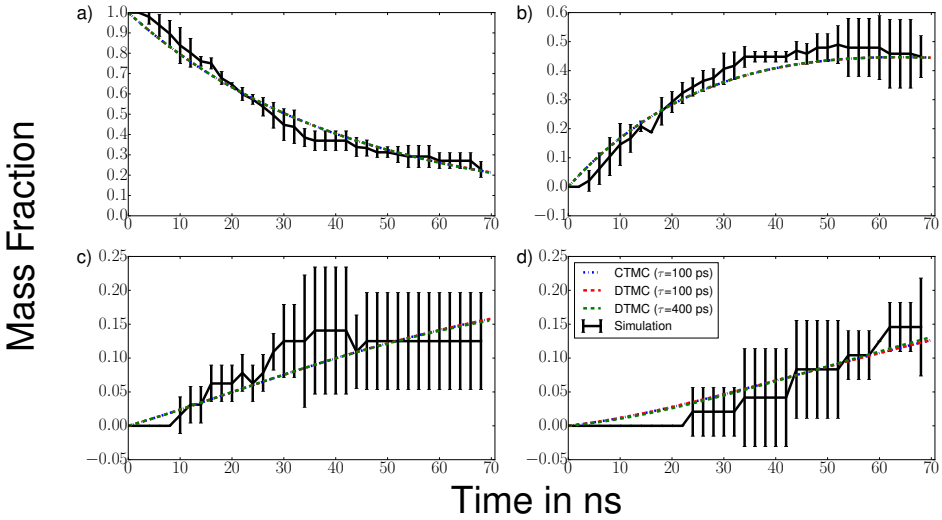

Figure B1. Comparison of the cluster size distribution in the 0.85 mM system predicted by the (i) CTMC with $\tau = 100$ ps (blue dot-dash lines), (ii) DTMC with a lag time of $\tau = 100$ ps (red dashed lines), (iii) DTMC with a lag time of $4\tau = 400$ ps (green dashed lines) to that directly observed 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.

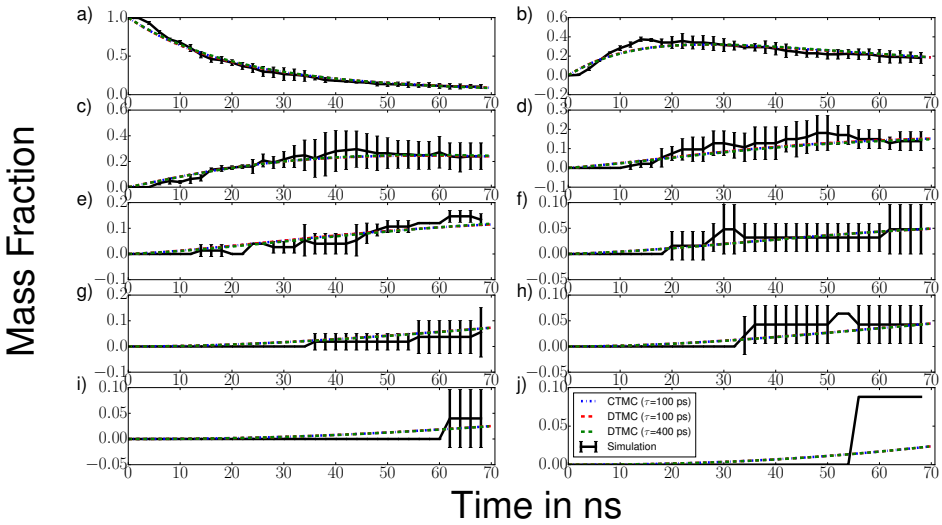

Figure B2. Comparison of the cluster size distribution in the 1.66 mM system predicted by the (i) CTMC with $\tau = 100$ ps (blue dot-dash lines), (ii) DTMC with a lag time of $\tau = 100$ ps (red dashed lines), (iii) DTMC with a lag time of $4\tau = 400$ ps (green dashed lines) to that directly observed 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.