Machine learning and molecular design of self-assembling π-conjugated oligopeptides

Bryce A. Thurston[a] and Andrew L. Ferguson[a,b,c]

[a]Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801, USA.; [b]Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, 1304 W Green Street, Urbana, IL 61801, USA. [c] Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, IL 61801, USA.

ARTICLE HISTORY
Compiled May 2, 2018

ABSTRACT
Self-assembling oligopeptides present a means to fabricate biocompatible supramolecular aggregates with engineered electronic and optical functionality. We conducted molecular dynamics simulations to probe the thermodynamics and morphologies of self-assembling synthetic oligopeptides with Asp-X_3-X_2-X_1-Π-X_1-X_2-X_3-Asp architectures. Dimerisation and trimerisation free energies were computed for a range of Asp-X_1-X_2-X_3 amino acid sequences within the symmetric tetrapeptide wings, and for perylenediimide (PDI) and naphthalenediimide (NDI) conjugated Π cores that mediate hydrophobic stacking and electron delocalisation along the backbone of the self-assembled nanostructure. Insertion of the larger PDI cores elevated oligomerisation free energies by a factor of 2-3 relative to NDI and also improved alignment of the oligopeptides within the stack. Training of a quantitative structure-property relationship (QSPR) model over the molecular simulation data revealed key physicochemical determinants of the observed oligomerisation free energies and produced a quantitative predictive model for the oligomerisation thermodynamics. Oligopeptides with moderate dimerisation and trimerisation free energies of ∼(-25) k_BT produced aggregates with the best in-register parallel stacking, and we used this criterion within our QSPR model to perform high-throughput virtual screening of oligopeptide chemical space and identify promising candidates for the spontaneous assembly of ordered nanoaggregates. We identified a small number of oligopeptide candidates for direct testing in large scale molecular simulations, and discover a novel chemistry DAVG-PDI-GVAD previously unstudied by experiment or simulation that produces well-aligned nanoaggregates expected to possess good optical and electronic functionality.

KEYWORDS
π-conjugated oligopeptides; molecular dynamics simulation; supramolecular peptides; self-assembly; quantitative structure-property relationships

1. Introduction

Self-assembling oligopeptides present an attractive vehicle for the synthesis of nanoaggregates with attractive structural properties and biological function [1–6]. Peptide assemblies can be controlled by a variety of properties such as primary structure, salt

CONTACT Andrew L. Ferguson. Email: alf@illinois.edu. Phone: (217) 300-2354. Fax: (217) 333-2736

concentration, and pH [7–13], allowing for many degrees of freedom with which to influence peptide assembly. Assembled aggregates have great potential applications in drug delivery, antimicrobials, vaccination, and regenerative medicine [1–3,7,8,14–17]. In addition to standard amino acids, synthetic oligopeptides can be functionalised with polymeric π-conjugated inserts [11,18–21], endowing the self-assembled supramolecular aggregates with optoelectronic and photophysical activity. Such materials have applications in a variety of systems including organic light emitting diodes (OLEDs), organic field-effect transistors (OFETs), and organic solar cells [22–31], and are attractive for their biocompatibility and water solubility [8,32–39].

The chemical sequence space accessible to synthetic oligopeptides is vast, and it is of value to understand the microscopic molecular forces and mechanisms governing assembly in order to provide rational principles to guide experimental peptide design towards candidates with good assembly behaviour. Such properties, however, can be difficult to access experimentally. As a result, computational studies provide an attractive method to elucidate these interactions [32,40–47]. While computer simulation can provide a great deal of insight, all-atom and even coarse-grained simulation at the length and time scales relevant to supramolecular assembly can make such calculations prohibitively expensive. This computational expense, coupled with the enormous palette of possible oligopeptide chemistries, makes it infeasible to directly evaluate the self-assembly behaviour of every possible chemistry. Quantitative structure-activity relationship/quantitative structure-property relationship (QSAR/QSPR) models present a means to develop inexpensive predictive models of molecular behaviour that can be used to perform high-throughput computational screening of chemical space to evaluate vastly more candidate peptides than would be possible by simulation and/or experimentation [48–51]. These models seek a predictive relationship between molecular behaviour that is expensive to evaluate and a set of physicochemical molecular descriptors that are inexpensive to compute or measure. It is typically implicitly assumed that similar molecules behave similarly and that the input descriptors are sufficient to predict the desired molecular behaviour [52]. The training of such models is a form of supervised learning, by which the relationship is extracted from a limited set of training data, validated against some test data, and then used to perform high-throughput "virtual screening". The use of QSPR models has a long history in chemometrics [53,54], and they have been applied to a diversity of peptidic systems in the context of structure, binding, drug loading, antimicrobial activity, and aggregation [52,55–61].

In this work, we conduct molecular dynamics simulations of a limited number of synthetic oligopeptide chemistries, and use these data to train QSPR models to predict oligomerisation thermodynamics from molecular physicochemical descriptors. These models are then used to inform the important determinants of assembly behaviour and perform high-throughput computational screening to identify peptide candidates with good predicted assembly behaviours. We focus our study on a class of synthetic oligopeptides with a peptide-Π-peptide symmetric triblock architecture of the form Asp-X_3-X_2-X_1-Π-X_1-X_2-X_3-Asp, where $\{X_1, X_2, X_3\}$ are amino acids from the set {Ala (A), Phe (F), Gly (G), Ile (I), Val (V)} and the Π insert is either a naphthalenediimide (NDI) or a perylenediimide (PDI) conjugated core (Figure 1). The peptide family represents a flexible archetype that may be readily synthesized by on-resin dimerization [62], and which has been previously shown in a number of prior computational [10,32,45,46,63] and experimental studies [9,32,35,46,62,64] to possess a variety of desirable properties. Specifically, these biocompatible and water soluble oligopeptides exist as dispersed small aggregates at neutral pH that are triggered to

assemble into micron-sized pseudo-1D fibrils upon acidification due to protonation of the terminal Asp residues. This eliminates the electrostatic repulsion between the Asp residues, and promotes assembly by hydrophobic, hydrogen bonding, and π-stacking interactions. Delocalisation of electrons between the π-conjugated cores provides the assemblies with functional electronic and photophysical properties, including electron transport and exciton migration, fluorescence, and gate voltage dependent current, which make such peptides viable materials to be used in biosensing, tracking molecular delivery to cells, energy transport and harvesting, imaging, field effect transistors, and other bioelectronic applications [32,35,38,38,39,45,62,65–67]. Previous studies have probed the role of N-to-C polarity, peptide concentration, pH, and particular peptide sequences and core chemistries upon assembly [10,32,45,46,62,68], but no work to date has sought to develop predictive physicochemical models of assembly to identify the important determinants promoting the formation of the ordered pseudo-1D assemblies required for good optoelectronic functionality and enable virtual screening of peptide sequence space. It is the principal motivation of this work to achieve these goals, and computationally test the model predictions by direct simulation of the assembly behaviour of identified candidates in large-scale molecular simulations.

[Figure 1 about here.]

We structure our work around two hypotheses. First, we propose that physicochemical molecular descriptors can be used to develop predictive models of oligomerisation free energies. A prerequisite to predicting large-scale many-body aggregation is a proper understanding of the mechanisms and thermodynamics of oligomerisation [47]. We demonstrate that QSPR models can ably predict oligopeptide dimerisation and trimerisation free energies for non-polar oligopeptides from small numbers of molecular descriptors, revealing the important physicochemical determinants of association and setting the stage for our second hypothesis. Second, we propose that the large-scale assembly behaviour can be predicted from the thermodynamics of oligomerisation. We ground this conjecture in the well-known principle for self-assembling systems in general [69,70], and experimental findings for this oligopeptide family in particular [32,64,67,68], that interactions between self-associating building blocks should be sufficiently strong to mediate assembly, but not so strong as to prevent mutual rearrangements into ordered structures as opposed to kinetically trapped states. Our results provide good support that our QSPR model can accurately identify non-polar oligopeptides possessing intermediate oligomerisation thermodynamics, and that these chemistries robustly assemble into aggregates with good in-register parallel stacking between neighbouring molecules. We deploy our model to perform high-throughput screening of oligopeptide chemical space, and identify a number of novel candidate sequences that form well-ordered parallel-stacked nanoaggregates in large-scale molecular simulation. The structure of the remainder of this manuscript is as follows. In Section 2, we describe our simulation methodology and QSPR model development. In Section 3, we report the results of our free energy and alignment simulations, the implementation of a QSPR model to predict simulation results, and a high-throughput screening of chemistries based on this model. Finally, in Section 4, we close with our conclusions and outlook for future work.

2. Methods

2.1. Explicit solvent simulations

Molecular dynamics (MD) simulations of peptides in explicit solvent were conducted using GROMACS 4.6.7 [71,72] in order to compute free energy profiles for peptide collapse and dimerisation that we subsequently used to parametrise an implicit solvent model [45]. Peptide geometries were obtained using the GlycoBioChem PRODRG2 Server [73], and modelled using AMBER99SB [74,75]. Terminal Asp residues were fully protonated to order simulate a low pH environment. The NDI and PDI cores are non-standard groups within AMBER99SB force field. Bonded parameters for the cores were determined using the parmchk2 method from Antechamber [76]. Native AMBER99SB parameters were unavailable for three bond angle interaction types, that instead were adopted from the Generalized Amber Force Field (GAFF) [77]. In keeping with methodology used for the derivation of partial charges for the AMBER force field [78], we compute partial charges on the core atoms by means of the Restrained Electrostatic Potential (RESP) method [79] using the RESP ESP charge Derive Server (REDS) [80]. Cores were parametrised as fragments by adding N-methylamide groups to either side and enforcing charge neutrality. The server computes charges utilising a single configuration in two different orientations in each run [79,81], and employs Gaussian09 [82] at the Hartree Fock/6-31G(d) level of theory to obtain the partial charges. Peptides were placed in a rhombic dodecahedral box with periodic boundary conditions and solvated in TIP3P water [83]. The box size was sufficiently large to accommodate the umbrella sampling calculations detailed in Section 2.4. The system was subjected to steepest descent energy minimization until the maximum force on any given atom was less than a threshold of 1000 kJ/mol.nm. Atomic velocities were initialized from a Maxwell distribution at 298 K and the system equilibrated for 100 ps in an NVT ensemble at a temperature of 298 K using a stochastic velocity rescaling thermostat [84] with a time constant of 0.5 ps, and finally for 100 ps in an NPT ensemble using the same thermostat and a Parrinello-Rahman barostat [85,86] at a pressure of 1 atm with a time constant of 1 ps and a compressibility of 4.5 $\times 10^{-5}$ bar^{-1}. Production runs were conducted in the NPT ensemble using the same barostat and a Nosé-Hoover thermostat [87,88] with a time constant of 0.5 ps to maintain a temperature of 298 K. The equations of motion were integrated using the leap-frog algorithm with a 2 fs time step [89]. Electrostatic interactions were treated using Particle Mesh Ewald (PME) with a cutoff of 1.0 nm and a 0.12 nm Fourier grid spacing that were optimized during runtime [90]. Lennard-Jones interactions were shifted smoothly to zero at 1.0 nm. Bond lengths were fixed using the LINCS algorithm [91], and Lorentz-Berthelot combining rules were used to determine interaction parameters between unlike atoms [92]. Execution speeds of 3.3 ns/day were achieved on one core of an Intel i7-4820K processor.

2.2. Implicit solvent simulations

To reach the long length and time scales necessary to observe peptide self-assembly and realize computational efficiency gains to enable us to simulate more molecular chemistries, we parametrised an implicit solvent model similar to that we previously employed in our study of the self-assembly of Asp-Phe-Ala-Gly-OPV3-Gly-Ala-Phe-Asp peptides [45]. Peptides were modelled using the AMBER99SB force field as described above [74,75], but the water solvent is now represented implicitly using the

Generalized Born model to treat polar interactions between peptide and solvent, and the solvent accessible surface area approximation to treat nonpolar interactions [93]. Nonpolar interactions were treated using an analytical continuum electrostatic (ACE) approximation [94] with a value of 2.259 kJ/mol.nm^2 for the surface tension [95]. Born radii were calculated using the method of Onufriev, Bashford, and Case with a relative dielectric constant of 78.3 and the standard parameter set of $\alpha = 1$, $\beta = 0.8$, and $\gamma = 4.85$ [96]. Coulombic interactions were treated using a cutoff of 3.4 nm and a dielectric offset of 0.009 nm. Lennard-Jones interactions were shifted smoothly to zero at 3.4 nm. All simulations were conducted in the NVT ensemble at 298 K by integrating the Langevin equation with a friction constant of 0.5 ps^{-1} [95]. Following our previous work [45], we rescale the non-bonded interactions within the AMBER99SB force field to compensate for the use of an implicit solvent model that overestimates inter-residue interaction strengths [97–99]. As detailed in the Appendix, we compute the optimal rescaling factor of $\alpha = 0.75$ from a best fit of the potential of mean force profiles for single peptide collapse (Section 2.3) and peptide dimerisation (Section 2.4) to those computed in explicit solvent for a representative peptide chemistry Asp-Phe-Ala-Gly-NDI-Gly-Ala-Phe-Asp. Execution speeds of 33 ns/day were achieved on one core of an Intel i7-4820K processor, representing a ∼10 fold speedup relative to the explicit solvent model.

2.3. Potential of mean force for peptide collapse

The potential of mean force (PMF) profile in the head-to-tail extent of a single oligopeptide specifies the free energy of the molecule as a function of its linear extent, quantifying the relative favourability of extended and collapsed configurations [100–102]. We determine PMF profiles for isolated oligopeptides in both explicit and implicit solvent and use these profiles to parametrise the implicit solvent model (Section 2.2). The PMF profiles were calculated by performing umbrella sampling in the molecular head-to-tail distance ($h2t$) defined as the distance between the C_α atoms in the terminal aspartic acid residues [100]. Umbrella windows were placed at evenly spaced 0.1 nm intervals over the range $h2t$ = 0-4.0 nm. Initial configurations for each umbrella window were obtained from non-equilibrium pulling of an initially fully extended peptide to induce collapse. Harmonic biasing potentials with a force constant of 1000 kJ/mol.nm^2 were applied in each umbrella window, and simulations run for 20 ns discarding the first 1 ns for equilibration. The unbiased PMF profile was estimated from the biased umbrella sampling data by solving the WHAM equations [101] using the g_wham program within GROMACS 4.6.7 [71,72]. The PMFs resulting from each of the two independent umbrella sampling calculations are mutually aligned within the large-$h2t$ bond-stretching regions of the PMFs, and then averaged. Uncertainties in each individual PMF are estimated from 100 rounds of bootstrap resampling, and in the average by standard propagation of uncertainties.

2.4. Potential of mean force for peptide oligomerisation

We also computed the PMF profiles for the formation of oligopeptide dimers ($n = 2$) and trimers ($n = 3$) as a function of the centre of mass separation r_{COM} between a monomer and a preassembled $(n - 1)$ oligopeptide stack. Initial stacks of n peptides were prepared by first stretching a monomer to its maximum head-to-tail extent, replicating it n times, and constructing parallel stacks of the copies with inter-monomer

separations of 0.45 nm, corresponding approximately to the global free energy minimum of the dimerisation PMF for stacked oligopeptides [45]. The system was allowed to equilibrate for 20 ps with the positions of the core atoms restrained, and then for another 20 ps with core restraints removed. This procedure allowed for peptides to relax into a well-stacked configuration. We then simulated the system for 1.5 ns, and used the system geometry at 0.5 ns, 1.0 ns, and 1.5 ns as the starting point for a non-equilibrium pulling runs. Nonequilibrium pulling is applied to the centre of mass separation r_{COM} between a terminal monomer and the remaining $(n-1)$ monomer stack. The configurations over the course of each pull are used to initialize umbrella sampling runs at evenly spaced 0.1 nm intervals over the range r_{COM} = 0-2.5 nm in the case of dimer aggregation and over the range r_{COM} = 0-3.0 nm in the case of trimer aggregation, where the upper bound of the range is specified to be sufficiently large that the two groups are effectively non-interacting. Harmonic biasing potentials with a force constant of 1000 kJ/mol.nm^2 were applied in each umbrella window, and simulations run for 20 ns discarding the first 1 ns for equilibration. The unbiased PMF profile was estimated from the biased umbrella sampling data by solving the WHAM equations [101] using the g_wham program within GROMACS 4.6.7 [71,72]. The $\Delta F_{\mathrm{corr}} = -2k_BT\ \ln(r_{\mathrm{COM}})$ correction is applied to each PMF to remove the purely entropic effects attributable to restraining the two groups to a particular separation [95,103,104]. The PMFs resulting from each of the three independent umbrella sampling calculations are mutually aligned within the large-r_{COM} plateau regions of the PMFs where the two groups are non-interacting, and then averaged. Uncertainties in each individual PMF are estimated from 100 rounds of bootstrap resampling, and in the average by standard propagation of uncertainties. The dimerisation ΔF_2 and trimerisation ΔF_3 free energies are defined as the difference in free energy between the large-r_{COM} non-interacting plateau of the PMF and the global free energy minimum containing the associated configurations.

2.5. Measurement of structural alignment

We ultimately seek to relate the ΔF_2 and ΔF_3 values predicted by our model to a measure of the quality of structural alignment within self-assembled aggregates formed by large numbers of peptides. Since we are interested in engineering peptides for optoelectronic functionality, our primary design objective is to establish good π-π stacking between neighbouring oligopeptides within the self-assembled stacks. Essentially, we are using good parallel stacking of the π-conjugated aromatic cores as a classical proxy for good quantum delocalisation of electrons over the backbone of the self-assembled stacks. We quantify the degree of structural alignment exhibited by a particular peptide chemistry by conducting 50 ns unbiased simulations of 64 peptides in a 50×50×50 nm^3 implicit solvent box, corresponding to a concentration of 0.85 mM. This concentration is both experimentally achievable and at which oligopeptide assembly has previously been observed [9]. The structure of aggregates is tracked as a function of time to monitor the formation of well-aligned parallel stacked clusters. The *association distance* between two oligopeptides is defined as [63,105],

$$R_{a,b}^{\mathrm{assoc}} = \min_{i\in a}\min_{j\in b}\ r_{ij}, \tag{1}$$

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

distance is defined as [63,105],

$$R_{a,b}^{\text{align}} = \max\left[\left(\max_{i\in(core\ a)}\ \min_{j\in(core\ b)}\ r_{ij}\right),\ \left(\max_{i'\in(core\ b)}\ \min_{j'\in(core\ a)}\ r_{i'j'}\right)\right]. \tag{2}$$

The first term in round brackets defines the minimum intermolecular distance from each atom i in the π-conjugated core of oligopeptide a to each atom j in the π-conjugated core of oligopeptide b, and selects the maximum of these. The second term in round brackets defines the reciprocal of this, computing the maximum minimum distance from any core atom i' in oligopeptide b to any core atom j' in oligopeptide a. In the square brackets we then take the maximum of the two terms. As has been previously observed, this distance is equivalent to the graph diameter [105,106]. This measure presents a relatively strict definition of molecular association, with small alignment distances only reported that if all atoms within the cores of the two molecules are in close proximity. Accordingly, it presents a means to identify whether the cores of a pair of oligopeptides are in a parallel stacked configuration with in-register alignment between all of the fused aromatic cores. Two oligopeptides are defined to be *aligned* if $R_{a,b}^{\text{align}} < 0.5$ nm. We specify the two cutoffs based on the observed mean separation of two peptides in a 20 ns unbiased run starting from a well-aligned, π-stacked configuration. Based on these definitions, we define the our alignment metric a for a particular snapshot of our molecular simulation as the ratio of the average size of aligned oligopeptide clusters to associated oligopeptide clusters,

$$a = \frac{\overline{n_a} - 1}{\overline{n_c} - 1}, \tag{3}$$

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

3. Results and Discussion

3.1. Dimerisation and trimerisation free energies computed by molecular simulation

Containing three independently mutable amino acid residues and NDI or PDI as potential π-conjugated cores, our Asp-X_3-X_2-X_1-Π-X_1-X_2-X_3-Asp peptide family comprises $20^3\times2 = 16{,}000$ members. Even with our implicit solvent model, exhaustive calculation of the dimerisation and trimerisation free energies from molecular simulation is computationally intractable. Accordingly, we instead perform these calculations over the restricted subset of oligopeptide chemistries DFAX-Π-XAFD, DFXG-Π-GXFD, and DXAG-Π-GAXD, where X ∈ {A, F, G, I, V} and Π ∈ {NDI, PDI}. This choice of 26 different chemistries was motivated by experimental work showing good assembly behaviours of peptides belonging to these and similar families [32,35,62,107], and the decision to avoid charged and/or polar residues that are expected to interfere with the triggerable low-pH association. We present in Table 1 the dimerisation ΔF_2 and trimerisation ΔF_3 free energies computed from the implicit solvent umbrella sampling

simulations described in Section 2.4.

[Table 1 about here.]

Analysis of the data reveals a significant difference between the free energies of interaction between PDI and NDI cores: the most strongly interacting NDI peptides (DFFG-NDI for dimers and DFAV-NDI for trimers) possess more shallow free energy wells that the most weakly interacting PDI peptides (DFAA-PDI for dimers and DFAV-PDI for trimers). Interestingly, we observe that stronger free energy changes for the formation of a dimer do not necessarily imply stronger free energy wells in the formation of a trimer, indicating the importance of going beyond purely pairwise interactions in characterizing multi-body assembly [47]. For example, DFAV-PDI has one of the largest ΔF_2 values but one of the lowest ΔF_3 values. While it is clear that the larger PDI π-conjugated cores tend to elevate oligomerisation free energies over that for NDI cores by a factor of 2-3, discerning more subtle trends based on peptide composition and sequence by inspection or intuition is challenging. In the following sections, we describe the development and interrogation of a QSPR model to assist in the discovery of the key determinants governing the thermodynamics of oligopeptide oligomerisation.

3.2. QSPR modelling of oligomerisation thermodynamics

We engage our hypothesis that oligomerisation thermodynamics can be predicted from physicochemical molecular descriptors by training a QSPR model to predict oligopeptide dimerisation and trimerisation free energies computed by umbrella sampling. Training is conducted over data for the 26 chemistries reported in Table 1. Although these particular oligopeptides represent no more than a small sampling of the 16,000 possible Asp-X_3-X_2-X_1-Π-X_1-X_2-X_3-Asp chemistries, we show that it was sufficient to produce QSPR models capable of quantitatively predicting oligomerisation free energies of a diversity of non-polar oligopeptides with diverse residue composition and sequence. Training of our QSPR model constitutes a form of supervised learning to regress a relationship of the form $\{\Delta F_2^i, \Delta F_3^i\} = f(\vec{d}_i)$, where ΔF_2 and ΔF_3 are the dimerisation and trimerisation free energies computed from molecular simulation, $\vec{d}$ is a vector of physicochemical molecular descriptors that can be inexpensively computed from the chemical sequence and/or three-dimensional structure of the peptide monomer, i indexes the particular peptide chemistry, and f is the functional mapping that is sought. Development of the QSPR model comprises four main steps: descriptor generation, descriptor cleaning, model construction, and model validation. An illustration of the QSPR training procedure is depicted in Figure 2.

[Figure 2 about here.]

3.2.1. Descriptor generation.

A molecular descriptor is a numerical quantity that can be computed directly from the molecular chemistry and/or structure [50,108]. The PaDEL software package [109] was used to compute a total of 1444 1D (dependent only on composition) and 2D (dependent on bond network) descriptors, and 431 3D (dependent on three dimensional structure) descriptors. These descriptors correspond to a number of physical and chemical attributes, examples of which include numbers of atoms of various types, autocorrelations between atoms separated by particular numbers of bonds weighted by

quantities such as electronegativity, and three dimensional weighted radial distribution functions. Peptide structures required by PaDEL were generated by steepest descent energy minimization from an initially extended configuration until the maximum force in the system does not exceed 1000 kJ/mol.nm. The resulting file is converted to an MDL MOL format using Babel [110] and aromatic bonds manually assigned. When descriptors involve atomic partial charges, PaDEL computes them using the Gasteiger-Marsili method [111]. The resultant descriptors produced over the 26 chemistries are Z-scored such that they are centred, standardised, and de-dimensionalised [112]. To increase the diversity of descriptors and potentially improve their interpretability, we apply the descriptor generation protocol to the entire oligopeptide, the π-conjugated core alone, and the variable X_1-X_2-X_3 amino acid triplet. In this manner, we generate a total of 5625 descriptors for each of the 26 chemistries.

3.2.2. Descriptor cleaning.

We clean the ensemble of 5625 descriptors to eliminate unstable, uninformative or redundant descriptors [112,113]. First, we eliminated 1230 descriptors with a sensitive dependence on the three dimensional structure of the oligopeptide. The oligopeptides are known to adopt a diversity of configurations both in isolation and within self-assembled aggregates, so we wished to discard descriptors that vary strongly with the peptide conformational state, and may be strongly influenced by the particulars of the methodology used to generate the initial peptide conformation. We compare descriptor values for all peptide chemistries in the training data set that are generated from our energy minimized peptide structure with those that are generated from the terminal configuration of a 20 ns simulation of an isolated peptide in implicit solvent, and select only those descriptors for which the root mean square deviation across all chemistries was less than a cutoff value of 0.15. Second, we eliminate 1686 descriptors that were found to be constant or nearly constant (defined as having a standard deviation less than 0.0001) over the 26 chemistries. Third, we removed 2462 highly correlated descriptors, identified from descriptor pairs possessing a Pearson correlation coefficient with magnitude in excess of $\rho = 0.90$. Highly correlated descriptors are removed in an iterative procedure. We first identify and retain the descriptor that is least correlated with all other descriptors, and then eliminate all descriptors with which it is highly correlated (i.e., $\rho \geq 0.90$). The next least correlated descriptor, of those that have not been selected or rejected, is then selected and those descriptors with which it is highly correlated are rejected, and so on. Together, these three cleaning protocols down-selected the number of descriptors from 5625 to 247.

3.2.3. Model construction.

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

we are operating. As such, we seek a relationship of the form,

$$\Delta F^i = c_0 + \sum_{j=1}^{N} c_j d_j^i + \epsilon_i, \tag{4}$$

where i indexes the particular oligopeptide chemistry, j indexes the descriptor, ΔF^i is either the dimerisation or trimerisation free energy computed for oligopeptide i, d_j^i is the j^{th} descriptor associated with chemistry i, $\{c_i\}_{i=0}^N$ are the regression coefficients to be determined, and ϵ_i is the residual error associated with chemistry i. In principle, regression may be conducted over all $N = 247$ descriptors retained after the cleaning procedure. Typically, however, it is valuable to perform some form of descriptor selection in order to generate more simple, interpretable, and generalizable models in which a large number of the c_i are constrained to be zero [112,113]. A number of means exist to perform wrapped and/or embedded descriptor selection, including genetic algorithms [114] and various flavours of regularization including ridge (L_2), LASSO (L_1), and elastic net (L_1 and L_2) [115–117]. In this work, we implement a combination of exhaustive and pseudo-greedy forward stepwise descriptor selection to regularize our models, and we avoid overfitting by dividing data into training and testing datasets and monitoring errors over the test set [118]. Specifically, we exhaustively compute all $\binom{247}{1} = 247$ univariate, $\binom{247}{2} = 30{,}381$ bivariate, and $\binom{247}{3} = 2{,}481{,}115$ trivariate MLR models by least squares fitting of Equation 4 over the 21 training chemistries. Exhaustive consideration of all $\binom{247}{4} = 151{,}348{,}015$ tetravariate models proved computationally expensive, so we instead greedily considered the 12,200 tetravariate models formed by adding one more descriptor to the top 50 trivariate models. Exhaustive computation of trivariate models proved to yield no significant benefit over trivariate models obtained using this greedy approach, so the final analysis was conducted using a greedy search approach for trivariate models as well and lending support for our use of this technique. Models were computed for 15 separate random divisions of data into training and testing sets in order to evaluate uncertainties in training and testing errors. In principle, we could have extended this stepwise selection procedure to models of arbitrary complexity, but in practice, we found tetravariate models or smaller were sufficient to predict dimerisation and trimerisation free energies with estimated uncertainties better that the accuracy with which these quantities were computed from our simulations. Furthermore, testing errors were observed to increase for both dimerisation and trimerisation free energies for higher than tetravariate models, indicating that the small size of our data set places us in a regime prone to overfitting. We avoid overfitting by controlling model complexity so as to minimise the testing error, but elect not to exacerbate the overfitting danger through the incorporation of nonlinear terms in our regression model. This also has the advantage of leading to more interpretable models formed from simple linear descriptor combinations.

3.2.4. Model validation.

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

[Figure 3 about here.]

In principle, we could have terminated our QSPR validation procedure here, and selected from the four top-ranked models with the smallest RMSE over the testing data as our terminal model. However, we seek to further improve our predictive accuracy by developing ensemble regressors that average over the top several models at each level of model complexity (i.e., the univariate, bivariate, trivariate, and tetravariate MLR models). It is well known that such ensemble models frequently exhibit better performance than any one of the constituent models alone [119–121]. We determine an appropriate number of top-ranked models over which to average at each level of model complexity by performing 15 rounds of shuffled cross-validation, in which we train the ensemble predictor on a randomly selected split of 80% of the training data and measure its prediction accuracy on the remaining 20%. We identify the optimal number of top-ranked models over which to average by identifying a knee in the curve of test RMSE against number of models participating in the average. This analysis identifies optimized ensemble predictors that average over the top one $N = 1$ order MLR models, top four $N = 2$ order models, top nine $N = 3$ order models, and top six $N = 4$ order models for ΔF_2, and the top three $N = 1$ order MLR models, top nine $N = 2$ order models, top four $N = 3$ order models, and top two $N = 4$ order models for ΔF_3. The performance of these ensemble models at each level of model complexity over the training and testing data is illustrated in Figure 4, from which we identify the $N = 2$ ensemble predictor to be optimal for prediction of ΔF_2, and the $N = 1$ ensemble predictor optimal for prediction of ΔF_3. The increase in testing error for models containing higher numbers of descriptors is a result of overfitting over our small dataset, and we avoid overfitting by selecting the model with the minimum testing error. The functional form of the best ΔF_2 ensemble model averaging over the four top-ranked $N = 2$ MLR models is,

$$\begin{aligned}\Delta F_2 =& \frac{1}{4}[(9.68 \times \text{MATS3c} + 4.58 \times \text{MATS1c\_wing} - 16.43) \\ &+ (9.85 \times \text{MATS3c} + 4.90 \times \text{ATSC4e\_wing} - 16.43) \\ &+ (-5.05 \times \text{SpMax6\_Bhs} - 10.73 \times \text{piPC3} - 16.43) \\ &+ (3.73 \times (\text{AVP\_5}) - 7.98 \times \text{piPC3} - 16.43)]\end{aligned} \tag{5}$$

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

$$\begin{aligned}\Delta F_3 =& \frac{1}{3}[(-7.80 \times \text{piPC3} - 19.30) \\ &+ (7.81 \times \text{maxHBint10} - 19.30) \\ &+ (7.83 \times \text{GATS2i} - 19.30)]\end{aligned} \tag{6}$$

An assessment of the statistical performance of these two ensemble models is presented in Table 2.

[Table 2 about here.]

The particular descriptors resolved in our terminal ΔF_2 QSPR model are the Moran autocorrelation [122] of lag 3 weighted by partial charges (MATS3c) and Moran autocorrelation of lag 1 of the amino acids on one side of the core minus the ASP residue on

the end weighted by partial charges (MATS1c_wing), and the centred Broto-Moreau autocorrelation [123] of lag 4 weighted by the Sanderson electronegativities [124] of the peptide wing (ATSC4e_wing), the 6^{th} largest eigenvalue of the modified Burden matrix [125] weighted by relative intrinsic state [126] (SpMax6_Bhs), conventional bond order ID number of order 3 [108] (piPC3), and the average valance path of order 5 [127] (AVP_5). The particular descriptors resolved in our terminal ΔF_3 QSPR model are the conventional bond order ID number of order 3 [108] (piPC3), the maximum electrotopological state [126] descriptor of strength for potential hydrogen bonds of path length 10 (maxHBint10), and the Geary autocorrelation [128] of lag 2 weighted by first ionization potential (GATS2i).

[Figure 4 about here.]

Importantly, the simple ensemble MLR QSPR models defined in Equations 5 and 6 provide quantitatively accurate predictions of the dimerisation and trimerisation free energies to within calculation accuracy of $\sim$4 k_BT using just a handful of easily calculable molecular properties. Indeed, computation of the eight molecular descriptors required by these two expressions for a particular oligopeptide chemistry requires only about 4 s of computation on one core of an Intel i7-4820K processor, amounting to approximately a 3 million-fold speedup over direct calculation of ΔF_2 and ΔF_3 by molecular simulation. Accordingly, these models can be used to perform high-throughput virtual screening for oligopeptide chemistries possessing desirable oligomerisation free energies.

3.3. Trained QSPR models provide molecular insight into determinants of oligomerisation thermodynamics

We now proceed to interrogate the particular descriptors and associated regression coefficients appearing in the optimal MLR ensemble predictors in Equations 5 and 6. In doing so we pick apart the physicochemical properties reflected in each descriptor, and develop insight into the key molecular features governing the oligomerisation thermodynamics.

MATS3c is the Moran autocorrelation [122] of lag 3 weighted by partial charges. Physically, this descriptor measures the correlation between atomic charges separated by three bonds. This descriptor appears twice with large positive regression coefficients in the expression for ΔF_2, indicating that large positive values of MATS3c favour large positive (i.e., unfavourable) values of ΔF_2. Bulkier aromatic residues and the larger PDI core tend to have lower values of MATS3c due to lower correlation between charges in atoms separated by three bonds in such residues compared to the higher correlation in the peptide backbone.

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

ATSC4e_wing is is the centred Broto-Moreau autocorrelation [123] of lag 4 weighted by Sanderson electronegativities [124] of the peptide wing. This descriptor also has

a positive correlation with ΔF_2. The Sanderson measure of electronegativity takes negative values for C and H atoms and positive values for N and O atoms. Due to spacing between atoms in the amino acid backbone, Ala and Gly residues tend to have lower values of this descriptor. Similar to MATS1c_wing, this descriptor appears with MATS3c and provides corrections to this term, in this case by predicting a larger free energy well for peptides containing more Ala and Gly residues.

piPC3 is the conventional bond order ID number of order 3 [108]. Physically, it measures the degree of branching in the bonded structure of the molecule. Aromatic bonds are weighted more heavily than single bonds so larger peptides, especially those containing large numbers of aromatic elements, with have larger values of this quantity. This descriptor appears twice in the expression for ΔF_2 and once in that for ΔF_3, in each case with large negative regression coefficients. PDI cores possess significantly larger values of piPC3 than NDI, with this descriptor reflecting the more favourable association free energies of the former relative to the latter. Less clear-cut correlations between residue size and lower free energies may play a role as well.

SpMax6_Bhs is the 6^{th} largest eigenvalue of the modified Burden matrix [125] weighted by relative intrinsic state [126]. It appears with piPC3 with a negative regression coefficient in the expression for ΔF_2. This descriptor does not have a simple physical interpretation, but is strongly negatively correlated $r = -0.922$ with the number of aromatic atoms. This descriptor appears to correct the piPC3 term by decreasing the free energy well for peptides containing a larger number of aromatic atoms, and increasing the free energy well for peptides containing fewer.

AVP_5 is the average valance path of order 5 [127]. It is positively correlated with ΔF_2 and appears with piPC3. AVP_5 can be thought of as a measure of molecular compactness: molecules with more paths of length 5 that contain heavier atoms with fewer valance electrons or atoms to which many hydrogen atoms are bound will have higher values for this descriptor. For the standard amino acids, this quantity will in general be larger when the ratio of hydrogen atoms to other atoms is larger. Accordingly, atoms containing Phe or Ile residues possess higher values, while NDI and PDI cores are not significantly differentiated from one another. This term appears to correct the piPC3 term that tends to overestimate the importance of residue size in determining ΔF_2.

maxHBint10 is the maximum electrotopological state [126] descriptor of strength for potential hydrogen bonds of path length 10. This descriptor is positively correlated with ΔF_3. Physically, the small size of the NDI core allows this path length to span across the core between two viable atoms, which is not possible for the PDI core. As a result, oligopeptides with NDI cores possess higher values for this descriptor than PDI peptides, resulting in less favourable trimerisation free energies. Furthermore, the presence of atoms with lower Kier-Hall electronegativity in residues adjacent to the NDI core or residues one position away from the PDI core leads to lower values for this descriptor and deeper free energy wells for trimerisation.

GATS2i is the Geary autocorrelation [128] of lag 2 weighted by first ionization potential. It has positive correlation with ΔF_3. This quantity is an inverse measure of autocorrelation (i.e., values > 1 indicate negative correlation and values < 1 indicate positive correlation) between the first ionization potential of atoms separated by two bonds. This quantity tends to reveal a weak negative correlation across all peptides, but the positive correlation between atoms in larger aromatic regions, such as PDI cores and Phe side chains, lowers the strength of this negative correlation and so the value of this descriptor. The second hydrogen in glycine residues also tends to decrease the strength of the negative correlation, so peptides containing more Gly will

have lower values for this descriptor.

In sum, the QSPR model has identified a small number of physicochemcial properties that are the principal determinants of dimerisation and trimerisation free energies. Synthesizing the above analyses provides the following three physical insights into the molecular mechanisms governing assembly. First, the larger PDI cores lead to deeper free energy wells for dimerisation and trimerisation. Each set of descriptors for dimerisation and each descriptor for trimerisation has some way of drawing a sharp distinction between PDI and NDI cores. Specifically, this model predicts a $\sim 15k_BT$ difference between PDI and NDI cores in both dimerisation and trimerisation free energies. Second, larger residues, especially Phe, are predicted to yield stronger free energies of aggregation, but that can easily be overestimated. Each pair of descriptors in the ΔF_2 model is characterized by the same general trend: one term distinguishes between PDI and NDI cores and overestimates the stability of larger residues, and the second term corrects for this. Specifically, this model predicts an average increase in well depth over the range of simulated chemistries of $\sim 2.5k_BT$ for replacing a given residue with Phe, with this effect larger for NDI than PDI cores. Third, larger residues with lower electronegativity nearer to the NDI core seem to play an important roll in stabilizing the NDI trimer, while such residues are not as important in PDI trimers.

3.4. Correlation of oligomerisation thermodynamics with self-assembled alignment quality.

Having developed a predictive QSPR model of dimerisation and trimerisation free energies, we now move to test our second hypothesis that oligomerisation thermodynamics can predict the large-scale self assembly behaviour. Following the precept that self-assembling building blocks should possess sufficiently strong interactions to stabilize self-assembled aggregates but not so strong as to impose kinetic trapping and prohibit mutual rearrangements and healing of defects to form ordered aggregates [69], it is our conjecture that peptide chemistries with intermediate ΔF_2 and ΔF_3 values should show the best assembly into well-ordered aggregates with in-register stacking of the π-conjugated cores. Experimental support for this assertion in the context of these oligopeptides comes from recent work showing significant differences in photophysical and conductive properties of assembled peptides resulting from variation of peptide amino acid sequence [32,35,107]. These results are hypothesized to be caused by kinetically trapped aggregates forming at early stages of assembly [32] and variations in local packing order [107]. While these results are rather qualitative and pertain to different π-conjugated cores than those studied here, they nevertheless support the hypothesis that the formation of kinetically trapped states, most likely resulting from overly attractive peptide interactions, have a negative impact on core alignment. To test this conjecture, we conduct large scale simulations of assembly during which we monitor the degree of in-register parallel stacking of the π-conjugated cores. The computational expense associated with these calculations precluded us from conducting these runs for all 26 chemistries, and so we judiciously selected DFAG-NDI, DFAG-PDI, DFAF-NDI, DFAF-PDI, DFAV-NDI, DFAV-PDI, DFAI-PDI, and DAAG-PDI as eight oligopeptides spanning a wide range of dimerisation and trimerisation free energies (cf. Table 1). We track alignment quality using the alignment metric a defined in Equation 3 as a measure of the probability that associated peptides will form well-aligned parallel stacks. The time evolution of this quantity over the 50 ns simulations are reported in Figure 5a, and the average a values over the equilibrated portion of

the runs reported alongside ΔF_2 and ΔF_3 in Table 1.

[Figure 5 about here.]

Our calculations reveal that oligopeptides with the smaller NDI core tend not to form well-aligned aggregates regardless of peptide wing chemistry. Peptides possessing a PDI core show two distinct groupings: DFAG-PDI and DAAG-PDI align most readily, while DFAV-PDI, DFAF-PDI, and DFAI-PDI do not align as well, although still better than those with an NDI core (Figure 5a). We quantify this relationship by fitting a bivariate Gaussian relating alignment quality and dimerisation and trimerisation free energies,

$$a = a_0 \exp\left(-\frac{(\Delta F_2 - \mu_{\Delta F_2})^2}{2\sigma^2_{\Delta F_2}} - \frac{(\Delta F_3 - \mu_{\Delta F_3})^2}{2\sigma^2_{\Delta F_3}}\right), \tag{7}$$

where a_0 = 0.548, $\mu_{\Delta F_2}$ = -22.2 k_BT, $\mu_{\Delta F_3}$ = -25.3 k_BT, and $\sigma_{\Delta F_2} = \sigma_{\Delta F_3}$ = 6.6 k_BT provide a good fit to the data (Figure 5b). This fit illuminates a "goldilocks" regime in which peptides possessing intermediate $\Delta F_2 \approx \Delta F_3 \approx$ -25 k_BT exhibit the best alignment, pointing towards an optimal trade off between sufficiently strong interaction strength to mediate assembly, but not so strong as to result in kinetic trapping in poorly ordered clusters.

3.5. High-throughput virtual screening

The good fit of the alignment metric to the dimerisation and trimerisation free energies – although based on a relatively small training set of only eight peptides – gives confidence that we can use our QSPR model to perform high-throughput screening of chemical space to identify peptides with ΔF_2 and ΔF_3 values predicted produce well-aligned stacks. Computing the eight molecular descriptors required by the QSPR model takes only 4 s per oligopeptide on a single Intel i7-4820K core, enabling traversal of orders of magnitude more chemistries than would be possible by molecular simulation. We search over the $17^3 \times 2$ = 9,826 chemistries in the Asp-X_3-X_2-X_1-Π-X_1-X_2-X_3-Asp peptide family, where Π ∈ {NDI, PDI} and {X_1, X_2, X_3} take on all possible natural amino acids with the exception of Lys, His, and Arg. These three residues are neglected since they are positively charged at low pH, and would therefore disrupt the pH-triggered assembly mechanism due to electrostatic repulsion. We present in Table 3 the predicted ΔF_2 and ΔF_3 values from our QSPR model (Equations 5 and 6) and alignment metric a from our bivariate Gaussian fit (Equation 7) for a selected fraction of the 9,826 chemistries.

[Table 3 about here.]

In order to test our model predictions, we select from our list four oligopeptide chemistries predicted to possess good alignment metrics, and also seven controls. We select DMPP-PDI and DAIA-PDI as the two highest ranked chemistries. We also select DAVG-PDI as the highest ranked chemistry possessing a Gly residue adjacent to the core, as experimental work has previously suggested this as an important factor in dictating good assembly [32,35,107]. We also select DWWW-NDI as the highest ranked NDI core chemistry. Finally we also select DWNN-PDI, DTCT-NDI, DWCG-PDI, DWYW-NDI, DSSW-PDI, DYGA-PDI, and DYGG-PDI as controls possessing a

range of predicted free energies and alignments, and constituent amino acids. Several of these oligopeptides contain polar amino acids, which were not contained in the training data set, and so allow us to assess the generalizability and transferability of our model. These 11 oligopeptides were then subjected to implicit solvent molecular simulation to evaluate their dimerisation and trimerisation free energies, and large-scale 40 ns simulations of 64 peptides at 0.85 mM to assess their alignment behaviours. We report the predicted and calculated values of ΔF_2, ΔF_3, and a in Table 4.

[Table 4 about here.]

Comparison of the predicted and calculated values of ΔF_2 and ΔF_3 show that our QSPR model accurately predicts the oligomerisation thermodynamics for both NDI and PDI oligopeptides containing non-polar amino acid residues (Table 4, upper). For all five such chemistries, the predicted and calculated quantities lie well within the estimated uncertainties. Importantly, this group of chemistries contains Met (M), Pro (P), Cys (C), and Trp (W) residues that were not part of the training ensemble, but the model is sufficiently transferable to give good predictive performance. Further, the model correctly predicts that the six Trp residues in the peptide wings lead to strong associations in the DWWW-NDI chemistry, despite the fact that very few of the NDI training examples had dimerisation and trimerisation free energies that were even half as large (Table 1). This indicates the model is able to accurately estimate the impact bulkier aromatic regions have on the free energies of aggregation. Considering now the six polar chemistries (Table 4, lower), we see poor agreement of the predicted and calculated free energies. These chemistries all contain one or more polar residues Ser (S), Asn (N), Thr (T), or Tyr (Y) containing OH or NH_2 polar moieties. The poor predictive performance of our QSPR model may be attributed to the fact that our training data contained only the non-polar residues Ala(A), Phe (F), Gly (G), Ile (I), and Val (V), and clearly demonstrates that our current model cannot be reliably extrapolated to strongly polar molecules.

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

Our results support our hypothesis that the ΔF_2 and ΔF_3 of non-polar peptide oligomers can be accurately predicted by our QSPR model, and these oligomerisation free energies used to identify non-polar oligopeptide chemistries – excluding those possessing high aromatic residue contents – likely to possess good alignment ($a \gtrsim 25\%$) within the self-assembled aggregates. In particular, we identify a chemistry DAVG-PDI previously unstudied by either simulation or experiment showing very high structural alignment propensity of $a = 0.64$. A representative snapshot of the equilibrium aggregates formed by this oligopeptide chemistry is presented in Figure 6.

[Figure 6 about here.]

4. Conclusions

We conducted molecular dynamics simulations and developed QSPR models to understand and engineer self-assembling π-conjugated Asp-X_3-X_2-X_1-II-X_1-X_2-X_3-Asp oligopeptides. These molecules exhibit pH-triggered assembly with in-register parallel π-π stacking between the conjugated aromatic cores leading to electronic delocalisation along the nanoaggregate backbones and the emergence of desirable optical and electronic properties. Our study was founded on two hypotheses: that physicochemical properties of the oligopeptides can be used to accurately predict dimerisation and trimerisation thermodynamics, and that chemistries possessing moderate oligomerisation free energies produce the best ordered nanoaggregates. To engage these hypotheses, we parametrised an implicit solvent molecular model against explicit solvent all-atom calculations, and used this efficient model to compute the dimerisation ΔF_2 and trimerisation ΔF_3 free energies for 26 oligopeptides generated from all Ala (A), Phe (F), Gly (G), Ile (I), and Val (V) point mutants – excluding the distal Asp (D) residues required for pH-triggered assembly – of DFAG-II-GAFD oligopeptides containing NDI and PDI cores. These results revealed the larger PDI cores to give rise to $\Delta F_2 \sim$ (-24) k_BT and $\Delta F_3 \sim$ (-27) k_BT compared to only $\Delta F_2 \sim$ (-9) k_BT and $\Delta F_3 \sim$ (-12) k_BT for NDI inserts. To parse more subtle trends based on the composition and sequence of the peptide wings, we parametrised a QSPR model based on eight molecular descriptors that was capable of quantitatively predicting the dimerisation and trimerisation of non-polar oligopeptides. The predictive performance for polar chemistries was poor, and attributable to the fact that the model was developed exclusively over non-polar training examples. The particular descriptors identified by the model are informative as to the underlying determinants of the oligomerisation thermodynamics. It predicts oligomerisation free energies to be $\sim$15 k_BT larger for PDI cores as compared with NDI, and bulkier residues, especially Phe, to increase free energies of association by $\sim$2.5 k_BT. Finally, we observe that amino acids having lower electronegativity near the peptide core may play an important role in stabilizing formation of the NDI trimer. In developing a qualitatively accurate QSPR model for non-polar oligopeptide dimerisation and trimerisation thermodynamics, we provide strong support for our first hypothesis. This result is weakened by the poor performance for polar chemistries, but we anticipate that expansion of the training set to encompass polar training examples can produce similarly accurate models for this class of molecules.

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

In future work, we aim to expand the training data to incorporate polar oligopeptide chemistries in order to build a more general and transferable QSPR model. Moreover, we would like to expand the training set to incorporate side chains of differing lengths

and a wider variety of Π cores, including oligophenylvinylenes, oligothiophenes, and other rylene diimides. We also propose to incorporate additional computational techniques, including deep learning techniques that obviate the need for descriptors [129–131], Markov state models parametrised by molecular simulation data to reach longer length and time scales [45,132], and time-dependent density functional theory (TD-DFT) to explicitly engage the electronic properties of the self-assembled aggregates. Finally, we will work with experimental collaborators to explicitly test the optimal designs identified under our computational screening protocol thereby guiding and accelerating experimental discovery efforts, and also incorporate the experimental results into our modelling paradigm to refine and improve our computational screens. These endeavours will continue to pave the way for design and realization of self-assembling oligopeptides as novel biocompatible supramolecular optoelectronic materials.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. DMR-1729011.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

[14] Yuqiao Sun, Wen Li, Xiaoli Wu, Na Zhang, Yongnu Zhang, Songying Ouyang, Xiyong Song, Xinyu Fang, Ramakrishna Seeram, Wei Xue, Liumin He, and Wutian Wu. Functional self-assembling peptide nanofiber hydrogels designed for nerve degeneration. *ACS Appl. Mater. Interfaces*, 8(3):2348–2359, January 2016.

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

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

[17] EY Lee, BM Fulan, GC Wong, and AL Ferguson. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48):13588–13593, 2016.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[36] Tejaswini S. Kale, Jeannette E. Marine, and John D. Tovar. Self-assembly and as-

sociated photophysics of dendron-appended peptide--peptide triblock macromolecules. *Macromolecules*, 50(14):5315–5322, July 2017.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[53] Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni,

Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. Qsar modeling: Where have you been? where are you going to? *J. Med. Chem.*, 57(12):4977–5010, June 2014.

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

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

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

[57] Xuan Xiao, Pu Wang, Wei-Zhong Lin, Jian-Hua Jia, and Kuo-Chen Chou. iamp-2l: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry*, 436(2):168–177, May 2013.

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

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

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

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

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

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

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

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

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

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

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

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

[70] Faifan Tantakitti, Job Boekhoven, Xin Wang, Roman V. Kazantsev, Tao Yu, Jiahe Li,

Ellen Zhuang, Roya Zandi, Julia H. Ortony, Christina J. Newcomb, Liam C. Palmer, Gajendra S. Shekhawat, Monica Olvera de la Cruz, George C. Schatz, and Samuel I. Stupp. Energy landscapes and functions of supramolecular systems. *Nature Materials*, 15:469–, January 2016.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[95] Berk Hess, David van der Spoel, and Erik Lindahl. *GROMACS User Manual*. Royal Institute of Technology and Uppsala Univerity, 4.6.3 edition, 2013.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[119] Manoj Bhasin and G.P.S. Raghava. Prediction of ctl epitopes using qm, svm and ann techniques. *Vaccine*, 22(23):3195–3204, 2004.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[134] A. R. Brice and B. N. Dominy. Examining electrostatic influences on base-flipping: A comparison of tip3p and gb solvent models. *Commun. Comput. Phys.*, 13(1):223–237, 2013.

[135] T. Sanghi and N. R. Aluru. Coarse-grained potential models for structural prediction of carbon dioxide (co2) in confined environments. *J. Chem. Phys.*, 136(2):–, 2012.

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

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

Appendix: Rescaling of implicit solvent non-bonded interactions

We demonstrated in our previous work that the GBSA implicit solvent model substantially overestimates the strength of non-bonded interactions between peptides, and that it is necessary to rescale these interactions in order to reliably reproduce the thermodynamics of peptide aggregation [45]. Following our previous protocol, we adopt the minimally invasive strategy of uniformly rescaling the non-bonded interactions within the peptide force field in implicit solvent to best reproduce the potential of mean force (PMF) profiles for single peptide collapse (Section 2.3) and peptide dimerisation (Section 2.4) computed in explicit solvent [45,133]. This rescaling protocol can be considered a form of PMF matching [134–136]. Despite its simplicity, we previously showed the approach to produce satisfactory performance in reproducing explicit solvent results [45]. We adopt DFAG-NDI-GAFD as a prototypical oligopeptide for which to perform the fitting procedure and ascertain the optimal value of the rescaling factor.

The non-bonded interactions comprise Coulombic and Lennard Jones interactions $V^{nb}(r) = V_C(r) + V_{LJ}(r)$ that are each pairwise decomposable functions of interatomic separation r. As detailed in Ref. [45], a uniform rescaling of these pairwise interactions $V^{nb} \rightarrow \alpha V^{nb}$ amounts to rescaling the Lennard-Jones interaction parameter by a factor of α and the partial charges by a factor of $\sqrt{\alpha}$. We define the optimal scaling factor as that which minimizes the error function,

$$\mathrm{RMSE}(\alpha) = \mathrm{RMSE}_1(\alpha) + \mathrm{RMSE}_2(\alpha), \tag{8}$$

where RMSE_1 and RMSE_2 are, respectively, the root mean squared error between the PMF for single peptide collapse and peptide dimerisation computed in explicit and implicit solvent. We computed the implicit solvent PMF curves at values of $\alpha = [0.50, 0.60, 0.70, 0.72, 0.75, 0.78, 0.80, 0.90, 1.00]$, and report in Figure 7 the values of RMSE, RMSE_1, and RMSE_2 as a function of α. From these results, we discern $\alpha = 0.75$ to be the optimal value of the scaling factor at which the PMF profiles for peptide collapse agree to within a root mean squared error of 0.7 k_BT and for peptide dimerisation within 1.3 k_BT.

[Figure 7 about here.]

Table 1. Free energies of dimerisation ΔF_2 and trimerisation ΔF_3 computed from implicit solvent umbrella sampling calculations for the 26 chemistries in the families DFAX-Π-XAFD, DFXG-Π-GXFD, and DXAG-Π-GAXD, where X ∈ {A, F, G, I, V} and Π ∈ {NDI, PDI}. Values are computed as the mean over the three independent runs, and uncertainties are estimated by propagation of uncertainties and bootstrap resampling. Eight of the 26 peptides were selected for large-scale simulations to assess the alignment quality of the self-assembled supramolecular assemblies (Section 3.4). We report for these eight chemistries the alignment metric a (Equation 3) averaged over the equilibrated portion of simulations of the assembly of 64 oligopeptides at a concentration of 0.85 mM as a measure of the probability oligopeptides will assemble into well-aligned stacks with in-register parallel stacking between the π-conjugated cores. Uncertainties were estimated by five-fold block averaging the equilibrated portion of the trajectory.

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

Table 2. Statistical measures of our computed QSPR model for ΔF_2 and ΔF_3 for both training data and testing data. RMSE is the root mean square error of the model measured in k_BT, R^2 is the Pearson correlation coefficient, q^2 is the correlation coefficient over the leave one out cross validation of the training data, MAE is the mean average error of the model measured in k_BT, and R^2_{adj} is the adjusted correlation coefficient [137]. Error values are comparable to errors obtained in simulation. High values of the correlation coefficient and similar values of the adjusted correlation coefficient indicate the data are fit well by the model without overfitting.

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

Table 3. Dimerisation ΔF_2 and trimerisation ΔF_3 free energies predicted by Equations 5 and 6 and alignment metric a predicted by Equation 7 for a selected number of the 9,826 chemistries in the Asp-X_3-X_2-X_1-Π-X_1-X_2-X_3-Asp peptide family, where $\Pi \in$ {NDI, PDI} and {X_1, X_2, X_3} take on all possible natural amino acids with the exception of Lys, His, and Arg. Chemistries are ordered by the magnitude of the predicted alignment metric. Uncertainties in ΔF_2 and ΔF_3 are the mean error in prediction of the testing data over 15 rounds of shuffled cross validation. Uncertainties in a are estimated by applying Equation 7 to 10^5 $\{\Delta F_2, \Delta F_3\}$ pairs generated by sampling from a Gaussian distribution with the specified mean and standard deviation and taking the standard deviation of the result. DMMP-PDI and DAIA-PDI are selected for further simulation as the chemistries with the highest predicted alignment, DAVG-PDI is selected as the chemistry having a Gly residue nearest the core with the highest predicted alignment, and DWWW-NDI is selected as the NDI core with the highest predicted alignment. Finally, DWCG-PDI, DWYW-NDI, DSSW-PDI, DYGA-PDI, DYGG-PDI, DTCT-NDI, and DWNN-PDI are all selected as controls having a wide variety of predicted free energies, alignments, and constituent amino acids.

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

Table 4. Predicted and calculated dimerisation free energy ΔF_2, trimerisation free energy ΔF_3, and alignment metric a for the 11 oligopeptide chemistries selected from our high-throughput virtual screening. Chemistries above the horizontal line possess non-polar amino acid residues for which QSPR model predictions of the oligomerisation thermodynamics and alignment quality are generally very good. The polar oligopeptide chemistries – defined as those containing a difference in partial charge between any two bonded atoms greater than 1.0 e – reside below the line, for which the model predictions are relatively poor. Chemistries are ordered by the magnitude of the predicted alignment metric.

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

List of Figures

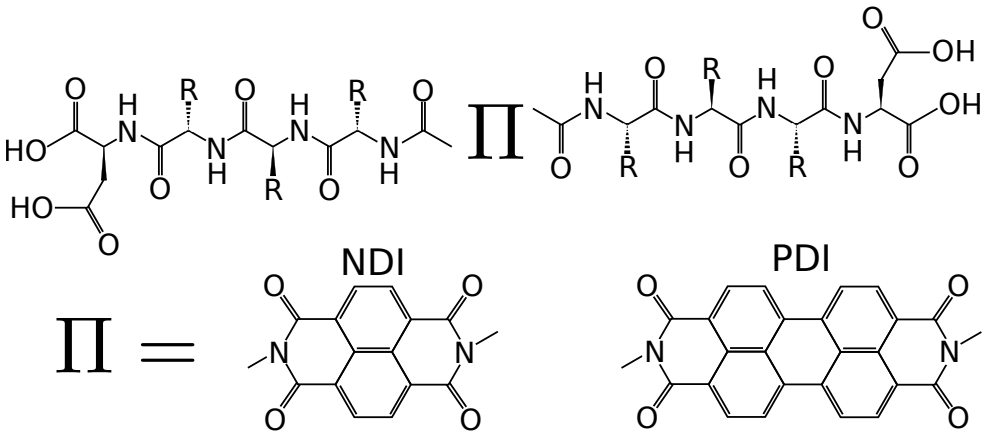

Figure 1. Chemical structure of the Asp-X_3-X_2-X_1-Π-X_1-X_2-X_3-Asp peptide family. {X_1, X_2, X_3} may be tuned to any one of the 20 natural amino acids, and the Π insert is a conjugated aromatic core, which in this work we restrict to be either a naphthalenediimide (NDI) or a perylenediimide (PDI) conjugated core. The N-to-C directionality of each peptide points away from the core, such that the oligopeptide sequence is antisymmetric and possesses two C-termini. The terminal aspartic acid residues and carboxyl termini are deprotonated at pH $\gtrsim 5$ such that the peptides carry a (-4) formal charge, and large-scale assembly is prohibited by electrostatic repulsion; at pH $\lesssim 1$, the termini protonate and assembly proceeds by π-π stacking, hydrogen bonding, and hydrophobic interactions [32].

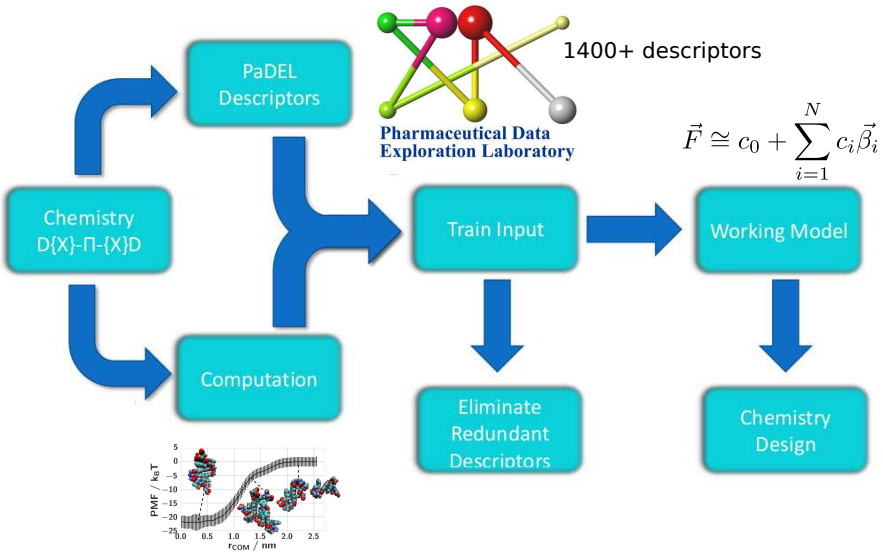

Figure 2. Schematic illustration of the QSPR model development protocol.

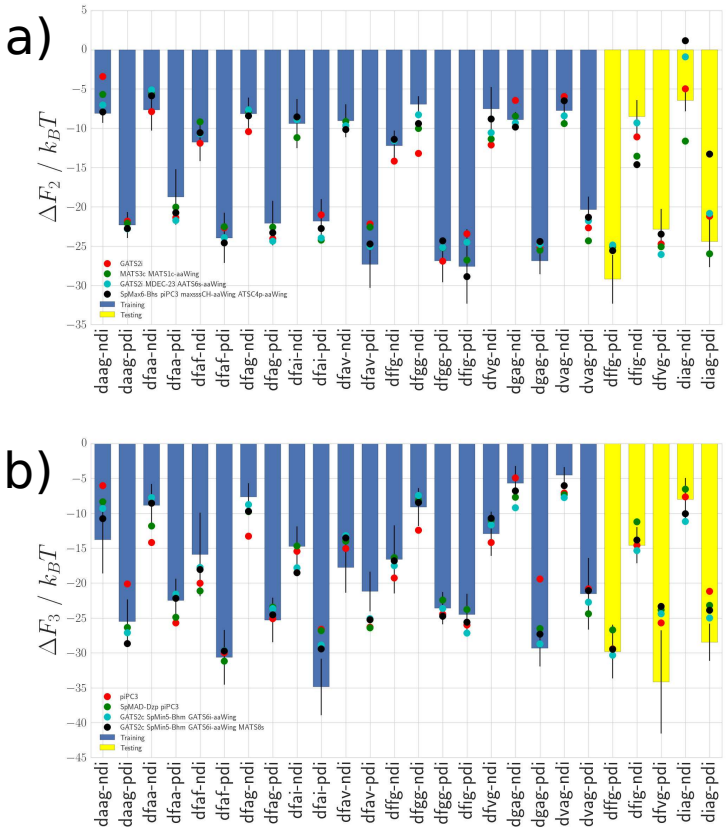

Figure 3. Performance of top-ranked MLR models comprising $N = 1$ (red dots), 2 (green dots), 3 (cyan dots), and 4 (black dots) molecular descriptors in predicting (a) dimerisation free energy and (b) trimerisation free energy computed in simulation. Free energies are reported in units of k_BT, where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

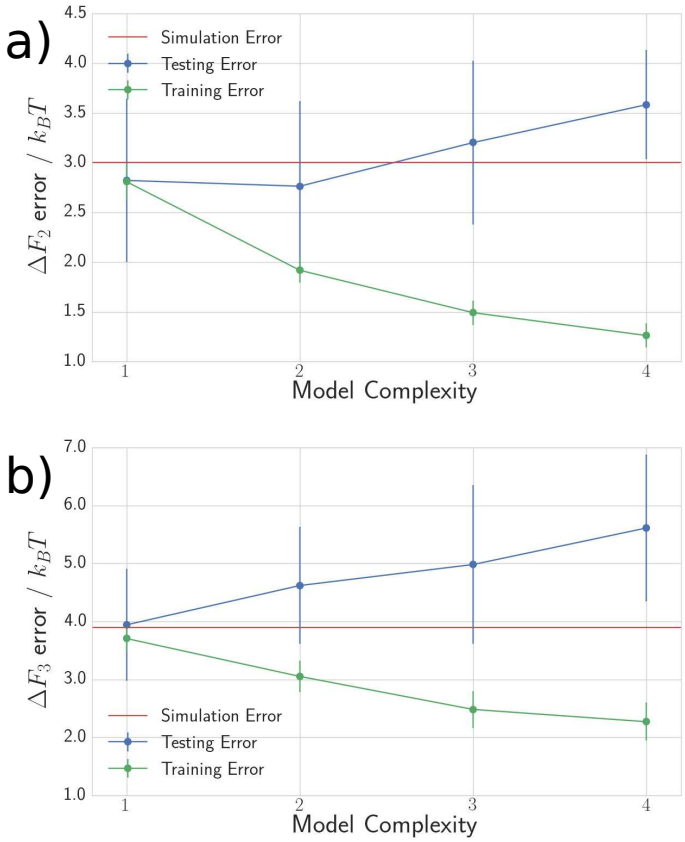

Figure 4. Performance of the optimal ensemble models at each level of model complexity over the training (green) and testing (blue) data in predicting the (a) dimerisation free energy and (b) trimerisation free energy computed in simulation. Free energies are reported in units of k_BT, where k_B is Boltzmann's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

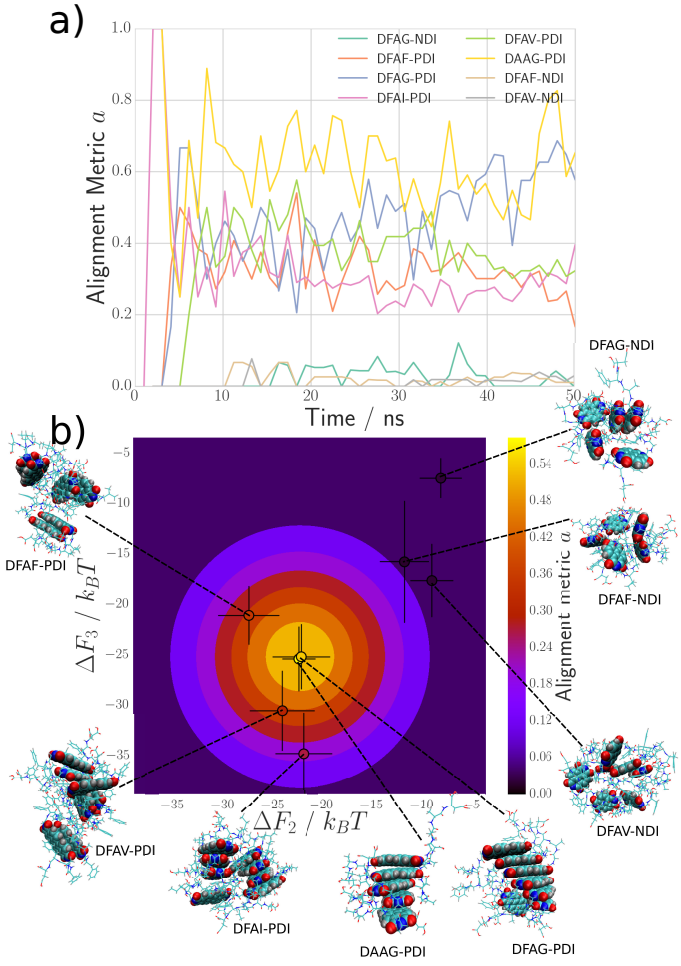

Figure 5. Alignment assessment of oligopeptide aggregates. (a) Time evolution of the alignment metric a (Equation 3) over the course of 50 ns runs of the self-assembly of 64 oligopeptides at a 0.85 mM initialized from randomly oriented monomers deposited over a grid. Values of a averaged over the equilibrated portion of the trajectory are reported in Table 1. (b) Scatter plot of the dimerisation ΔF_2 and trimerisation ΔF_3 free energies with points coloured by the computed alignment metric a. Characteristic snapshots of the oligopeptide aggregates extracted from our molecular simulations show that DFAG-PDI and DAAG-PDI tend to form well-aligned stacks, DFAV-PDI, DFAF-PDI, and DFAI-PDI show a weaker propensity for good alignment, and DFAV-NDI, DFAF-NDI, and DFAG-NDI do not associate into well-formed stacks. The contour plot represents a best fit bivariate Gaussian with $\mu_{\Delta F_2}$ = -22.2 k_BT, $\mu_{\Delta F_3}$ = -25.3 k_BT, and $\sigma_{\Delta F_2} = \sigma_{\Delta F_3}$ = 6.6 k_BT, where k_B is Boltzmann's constant ant T = 298 K. The data and fit support the assertion that intermediate ΔF_2 and ΔF_3 values result in the optimal oligopeptide alignment.

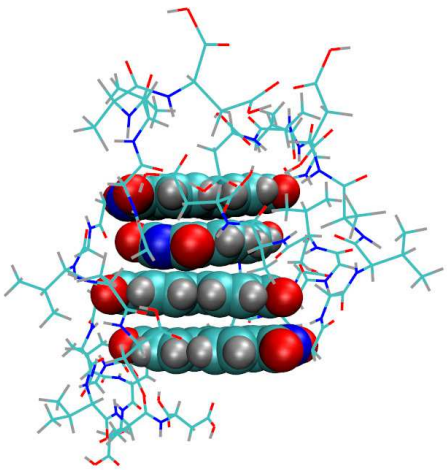

Figure 6. Representative snapshot of a self-assembled aggregate formed by the DAVG-PDI oligopeptide chemistry in 40 ns implicit solvent molecular dynamics simulations of 64 peptides at 0.85 mM.

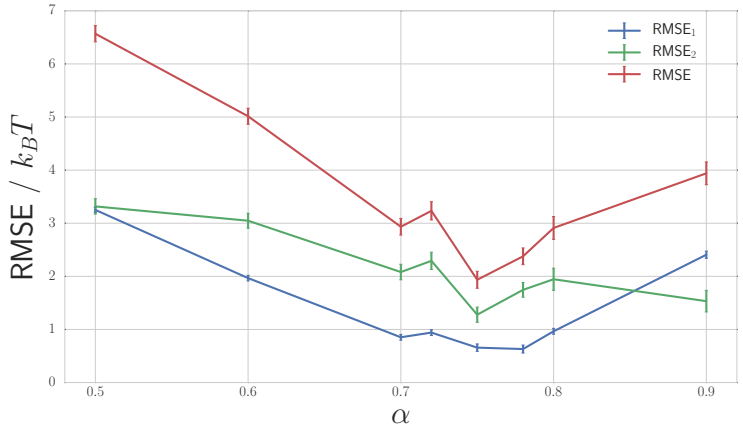

Figure 7. Root mean squared error between the PMF profiles for single peptide collapse and peptide dimerisation in implicit and explicit solvent as a function of the scaling factor for the implicit solvent non-bonded interactions. The agreement for single peptide collapse RMSE_1 and peptide dimerisation RMSE_2 both attain their optima at a scaling factor of $\alpha = 0.75$. Uncertainties are estimated by 100 bootstrap resamples of the simulation data used to compute the PMF profiles.