PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER



Cite this: *Phys. Chem. Chem. Phys.*, 2022, **24**, 1553

# A hybrid coarse-grained model for structure, solvation and assembly of lipid-like peptides†

Akash Banerjee, Chien Yu Lu and Meenakshi Dutt *

Received 14th September 2021,
Accepted 16th December 2021

DOI: 10.1039/d1cp04205j

rsc.li/pccp

Reconstituted photosynthetic proteins which are activated upon exposure to solar energy hold enormous potential for powering future solid state devices and solar cells. The functionality and integration of these proteins into such devices has been successfully enabled by lipid-like peptides. Yet, a fundamental understanding of the organization of these peptides with respect to the photosynthetic proteins and themselves remains unknown and is critical for guiding the design of such light-activated devices. This study investigates the relative organization of one such peptide sequence $V_6K_2$ (V: valine and K: lysine) within assemblies. Given the expansive spatiotemporal scales associated with this study, a hybrid coarse-grained (CG) model which captures the structure, conformation and aggregation of the peptide is adopted. The CG model uses a combination of iterative Boltzmann inversion and force matching to provide insight into the relative organization of $V_6K_2$ in assemblies. The CG model reproduces the structure of a $V_6K_2$ peptide sequence along with its all atom (AA) solvation structure. The relative organization of multiple peptides in an assembly, as captured by CG simulations, is in agreement with corresponding results from AA simulations. Also, a backmapping procedure reintroduces the AA details of the peptides within the aggregates captured by the CG model to demonstrate the relative organization of the peptides. Furthermore, a large number of peptides self-assemble into an elongated micelle in the CG simulation, which is consistent with experimental findings. The coarse-graining procedure is tested for transferability to longer peptide sequences, and hence can be extended to other amphiphilic peptide sequences.

## Introduction

The creation of novel solar cells which functionally integrate photosynthetic proteins into a synthetic matrix can harness solar energy to power a diverse array of nanoelectronics and solid-state devices.[1] Such cells will require the integration of fully functional reconstituted photosynthetic proteins from either plants, algae or bacteria. Reconstituting proteins such that they preserve all their functionality is nontrivial and has been achieved with lipid-like peptides.[1–6] Furthermore, small-sized assemblies of the lipid-like peptides and stabilized photosynthetic protein complexes can aggregate into three dimensional networks which can be used to power macroscale devices using solar energy. A rational design of such networks requires a fundamental understanding of how the peptides organize relative to the protein complexes and themselves. This study examines the relative organization of one such lipid-like peptide sequence, namely $V_6K_2$ (V: valine, K: lysine) in an aggregate encompassing this sequence.

Lipid-like peptides encompass 1–2 charged amino acids (head group) and 6–8 hydrophobic amino acids (tail group). The electrostatic interactions along with the hydrophobic effect drive the assembly of these peptides into regular structures, *e.g.*, nanotubes, fibers or vesicles.[7] Experimental studies have explored tuning the electrostatic and hydrophobic effects by synthesizing lipid-like peptide sequences of different head and tail lengths.[8,9] These modifications yield different nanostructures, each having unique material properties.[8–10] However, experimental

*Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. E-mail: meenakshi.dutt@rutgers.edu*

† Electronic supplementary information (ESI) available: Comparison of a IBI and FM potential, water–water RDF in a single peptide system, water–water RDFs in multiple peptide systems, water–water RDFs in 128 peptide systems, transferability of peptide–water potentials from $V_2K$ to $V_6K_2$, ions coordinated with lysine residues, stepwise increase in peptide chain length, schematic representations of CG bonds, comparison of angle distributions between the AA and CG representations of the $V_6K_2$ peptide, comparison of dihedral distributions between the AA and CG representations of the $V_6K_2$ peptide, intermolecular peptide–peptide and peptide–ion RDFs for the 25 $V_2K$ peptide system, comparison of peptide conformation prior to addition of 1–3–5 angle potentials, comparison of radius of gyration for the AA and CG representation of the $V_6K_2$ peptide, comparison of internal structural properties for the AA and CG $V_6K_2$ peptide, comparison of AA and CG time scales, gain in computational efficiency by coarse-graining, radius of gyration of peptides in multi-peptide systems, intermolecular RDFs of $V_6K_2$ peptides in multi-peptide simulations, contact maps, peptide–water interactions in a 8 peptide system across three resolutions, comparison of effective size of the micelle in AA, CG and backmapped-atomistic configurations. DOI link for simulation files: 10. 5281/zenodo.5715326. See DOI: 10.1039/d1cp04205j

approaches are constrained in their ability to resolve physical phenomena across diverse spatiotemporal scales, and are unable to resolve the relative organization of the peptides. A suitable computational approach which captures the large spatiotemporal scales associated with self-assembly and the chemical structure of the molecules can overcome this challenge.

All atom (AA) models provide a detailed representation of the chemical structure of peptides[11,12] and explicitly account for all intra- and intermolecular degrees of freedom (DOFs). Earlier computational studies have used AA models to investigate the aggregation of peptides. These studies are limited to either the early stages of aggregation,[13–15] or the examination of the stability of pre-assembled aggregates.[16,17] AA models are computationally very expensive when examining assembly of numerous, randomly dispersed peptides in an aqueous environment.

The formation of nanostructures occurring over large spatiotemporal scales requires models which simultaneously capture the effective chemistry of the peptides while having significantly fewer DOFs than the AA models. Coarse-graining approaches reduce selected DOFs in an AA model while preserving the effective chemistry of the molecule.[18] Groups of atoms in the AA model are mapped onto superatoms or pseudo atoms, also known as coarse-grained (CG) beads. An appropriate mapping scheme preserves the general structural description of the AA representation of the peptide in the CG model. The coarse-graining process reduces the number of coordinates and hence the DOFs that define the peptide, thereby significantly reducing the computational cost of using the CG model. In this study, the Molecular Dynamics (MD) technique is used to simulate the CG model. Similar to the CG MD approach, the Dissipative Particle Dynamics (DPD)[19–22] method is also suitable for investigating phenomena which span multiple spatiotemporal scales as it allows for a large time step and very soft potentials. However, unlike CG MD, DPD is unable to capture the structure of individual molecules in solution and in assemblies due to the soft potentials. The DPD potential allows for very large overlaps between the particles which introduces enormous challenges when back-mapping the DPD representation into the AA representation. Whereas CG MD has stiffer potentials, a smaller time step and no overlap between the MD beads. These features of CG MD allow it to capture the structure of the molecules and makes it easier to backmap the CG representation into the AA representation. Hence, CG MD is more suitable for this specific study which requires a CG potential which can capture the chemical structure of the molecules while simultaneously resolving their assembly.

The scientific questions determine the coarse-graining methodology. A top-down approach[18] is used to study peptide systems where thermodynamic properties are of particular interest, *e.g.*, peptide assembly. On the other hand, a bottom-up approach[23] is used to study structural details and local interactions. Both coarse-graining approaches neglect certain DOFs, and focus on investigating specific properties of interest.

Top-down approaches have been particularly successful in capturing the assembly of various peptides. Some approaches employ a scheme[24] that neglects the structural details of individual peptides. This technique drastically reduces the computational cost of simulating the assembly of a large number of peptides. An earlier study uses this approach to study the relationship between local peptide properties and the mechanism underlying peptide aggregation.[25] Nucleation and growth kinetics in the self-assembly process are also investigated with similar models.[26] Alternatively, an intermediate coarse-graining scheme[27] is applied to preserve some structural details during the self-assembly process. These models employ a higher number of beads per residue to preserve chemical specificity. One such example is the Martini model[28] which has been extensively used for simulating the assembly of peptides. Each CG bead is classified on the basis of the charge and polarity of the underlying AA fragment. The pairwise nonbonded potentials between the CG beads are parameterized to fit the free energy of partitioning between aqueous and hydrophobic solutions obtained from experiments. This model is successful for yielding self-assembled nanostructures for various classes of peptides.[29–38] The impact of factors like the initial concentration of peptides,[33,39] peptide sequence,[33,34] peptide chain length[32] and external conditions[38] on peptide self-assembly have been investigated with the Martini model. Elongated nanostructures are of particular interest to the current study. The Martini model is tested on amphiphilic peptides that self-assemble into elongated micelles.[40] In addition, the Martini model is used to study the aggregation of peptide amphiphiles into relatively larger, elongated nanostructures such as nanofibers.[30] Another example of a top-down approach is the PRIME model[41] which is used in conjunction with discontinuous molecular dynamics. Spontaneous formation of nanofibers has been resolved with the PRIME model.[42] Some of these models have a predetermined secondary structure,[28] whereas other models reproduce the AA secondary structure with special potentials.[43,44] However, the conformations of the peptides or their detailed local interactions within an aggregate cannot be resolved by a top-down approach.

Investigations of the local structure and forces require a bottom-up approach.[23] The approach is targeted towards reproducing specific properties in an AA system. Iterative Boltzmann Inversion (IBI)[45,46] is extensively used to preserve the chemical structure of peptides within aggregates.[47,48] Alternatively, Force Matching (FM)[49] minimizes the difference between the forces on the CG beads across AA and CG resolutions. FM is used to construct a generalized force field for any arbitrary peptide sequence.[50] Analogous approaches such as inverse Monte Carlo and relative entropy minimize the difference between a certain parameter across the AA and CG resolutions.[51,52] In addition, a reformulation of the relative entropy method preserves structural details of individual peptides within amorphous aggregates.[53] All these methods approximate the many-body potential of mean force (PMF) with effective pairwise CG potentials. Also, the AA distributions functions (for specific DOFs) are reproduced in the CG simulations. Although some correlations between DOFs cannot be reproduced,[54] there is significant agreement in the structural features corresponding to the underlying AA models. It is noted that there are more sophisticated approaches that better approximate the many-body PMF with local-density potentials[55,56] or 3-body potentials.[57,58]

 

Systems with multiple peptides in aqueous solution involve complex interactions which cannot be resolved by a single coarse-graining approach. This difficulty can be addressed by coarse-graining approaches[59] which combine multiple techniques to resolve different features or interactions in the system. For example, in an earlier study on peptide aggregation,[47] the bonded and nonbonded interaction potentials are extracted from AA trajectories. The bonded potentials and the peptide–water nonbonded potentials are resolved by applying IBI. The remaining nonbonded potentials are resolved by sampling the free energy of the underlying AA fragments. The resulting CG model efficiently reproduces the conformational flexibility of individual peptides, and the interactions between multiple peptides. Alternate methods such as the Adaptive Resolution Scheme (AdResS) enables AA and CG resolutions of a system in spatially localized regions of a simulation box.[60] The AA and CG regions are interfaced by a region with hybrid resolution. The AA, CG and hybrid resolutions are part of a single simulation, and the scheme allows exchange of the different resolutions of a molecule across the boundary of a region.[61,62] This approach is powerful and effective for scientific problems requiring a specific resolution in fixed spatial regions of the corresponding systems. Hence, a suitable hybrid CG technique is required which can simultaneously resolve on-demand the individual structure, conformation and aggregation behavior of peptides, independent of their spatial location within the simulation box.

This study examines the relative organization of the $V_6K_2$ peptide sequence in assemblies. This requires a bottom-up CG model which preserves the individual structure and conformation of the peptide while capturing its assembly. Here, AA trajectories are sampled to build CG tabulated potentials. A structure-based coarse-graining technique, *i.e.*, IBI is employed to derive the bonded potentials. This ensures that the individual structure of the peptides in the CG model is preserved. Additionally, IBI is used to derive the peptide–water nonbonded potentials. This ensures that the structure of the interface between the peptide molecules and water in the CG model is preserved. IBI works perfectly for bonded and peptide–water nonbonded potentials as a well-defined structural property of the AA reference system is available. This is achieved by sampling a single AA peptide in water (dilute solution). In the case of peptide–peptide nonbonded potentials, IBI generates potentials that are not transferable to other peptide concentrations. The potential is generated by sampling multiple AA peptides in water (concentrated solution). IBI generates multi-well peptide–peptide nonbonded potentials (see Fig. S1B, ESI†) that accurately captures the structure of the interactions in the reference system. However, since it is highly coupled to the underlying structure of the reference system, these potentials are expected to have lower transferability. On the other hand, a force-based method, *i.e.*, FM produces single welled potentials (see Fig. S1D, ESI†) with the same reference system. Due to its non-specific nature towards the structure of the interaction, it is surmised that FM potentials could be transferable to other systems within a limited peptide concentration range. Also, this method matches forces between the AA reference and CG models at different scales. In this manner, all forces that are inherent to the reference system that govern self-assembly of the peptides are reproduced in the CG model. FM has been extensively used to generate coarse-grained potentials for peptides.[50] In addition, it works well with long peptide sequences[63] like the one in this study. Thus, a hybrid scheme involving the IBI and FM methods aims at preserving the structure of individual peptides, the structure of the peptide–water interface, and the forces between the peptides that govern assembly.

In this study, a CG model for the $V_6K_2$ peptide sequence[6,64] is developed. The development of a CG model for such a peptide sequence poses particular challenges as the DOFs associated with the long hydrophobic tail group may be correlated. Hence, a stepwise procedure to build CG potentials of a short peptide ($V_2K$), and test its transferability on longer peptides ($V_4K$ and $V_6K_2$) is developed. The CG model resolves the structure of the $V_6K_2$ peptide sequence. The solvation structure of individual peptides is in good agreement with the corresponding results from the AA simulations. These peptides self-assemble into micelles. The relative organization of the peptides and the effective size of the micelle is in agreement with corresponding results from AA simulations. Furthermore, a large number of these peptides self-assemble into an elongated micelle which is consistent with experimental findings.[64,65]

# Methods

### Mapping scheme

All the peptides in the $V_nK_m$ ($n$: number of valine residues and $m$: number of lysine residues) series are coarse-grained using the same mapping scheme. These peptides encompass a hydrophobic tail and hydrophilic head group. The hydrophobic tail group includes valine residues. As shown in Fig. 1, $V_{Bx}$ (V: valine; B: backbone; $x$: index of valine residue) CG beads represent the valine backbone residues that form a linear chain. Each valine backbone bead is associated with a side chain bead, *i.e.*, $V_{Sx}$ (V: valine; S: side chain; $x$: index of valine residue). The hydrophilic head group consists of lysine residues. Since lysine has a larger excluded volume, two CG beads represent the side chain $K_{Sy1}$ and $K_{Sy2}$, and one CG bead represents the backbone bead $K_{By}$ (where K: lysine; B/S: backbone/side chain; $y$: index of lysine residue). On average, there are two to three heavy atoms per CG bead which is similar to another coarse-graining scheme for amino acids.[50] Capping residues are added on the C terminus ($NH_2$) and N terminus ($CH_3CO$).[6]

All the CG models include an explicit description of water and monovalent chloride ions. A one-site mapping scheme is used for both components. The development of the CG potential generates a set of effective pairwise potentials which account for the van der Waals and electrostatic interactions present in the atomistic simulations. Inclusion of separate electrostatic and van der Waals terms into the effective CG potentials will be pursued in a future study.

### Coarse-grained potential development

The Boltzmann inversion (BI) method has been extensively used for structure-based coarse-graining.[45,54,66] BI depends on

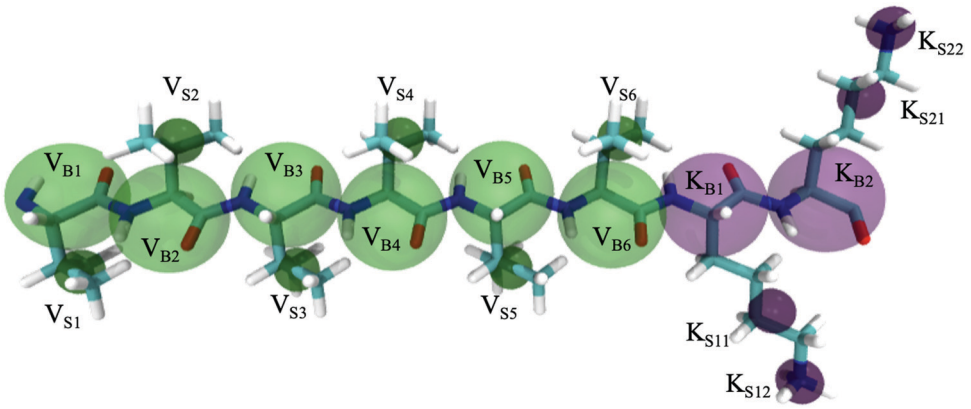


Fig. 1 The coarse-graining scheme for $V_6K_2$. Green and purple beads represent valine and lysine residues, respectively. The smaller green beads ($V_{Sx}$, V: valine; S: side chain; *x*: index of valine residue) represents the side chain of the valine residues. The larger green beads ($V_{Bx}$, V: valine; B: backbone; *x*: index of valine residue) represents the backbone of the valine residues. Similarly, the smaller and larger purple beads represent the lysine side chain ($K_{Sy1}$ and $K_{Sy2}$) and backbone ($K_{By}$) residues, respectively (K: lysine; B/S: backbone/side chain; *y*: index of lysine residue). The sizes of the backbone and side chain beads are kept different for visual clarity (bead sizes are not representative of their mass). The capping residues are not shown for clarity.

a structural property of the reference system, such as the RDF of a pairwise interaction. These distributions can be measured using particle trajectories from AA simulations. As per the mapping scheme, the CG coordinates are mapped onto the AA coordinates. Consequently, the distribution functions between the CG-mapped coordinates provide a reference for the equivalent interaction in the CG model. These CG-mapped distribution functions are representative of the sampling of the AA particle trajectories. Hence, extended AA particle trajectories sampled at frequent intervals will generate smooth reference distributions. The reference distributions are inverted *via* BI using eqn (1)–(4) to obtain tabulated potentials:[47,48,66]

$$U^{\mathrm{CG}}(d, T) = -k_{\mathrm{B}}T \ln\left(\frac{P^{\mathrm{CG}}(d, T)}{d^2}\right) \quad (1)$$

$$U^{\mathrm{CG}}(\theta, T) = -k_{\mathrm{B}}T \ln\left(\frac{P^{\mathrm{CG}}(\theta, T)}{\sin(\theta)}\right) \quad (2)$$

$$U^{\mathrm{CG}}(\phi, T) = -k_{\mathrm{B}}T \ln(P^{\mathrm{CG}}(\phi, T)) \quad (3)$$

$$U^{\mathrm{CG}}(r, T) = -k_{\mathrm{B}}T \ln(g(r, t)) \quad (4)$$

Eqn (1)–(4) represent the potentials obtained from BI for bonded and nonbonded interactions. $P^{\mathrm{CG}}$ is the normalized distribution function of a bonded interaction as a function of a degree of freedom (*i.e.*, $d$ is the bond distance, $\theta$ is an angle and $\phi$ is a dihedral angle). This idea can be extended to nonbonded interactions where the normalized distribution would be the radial distribution function ($g(r)$) of the pairwise interaction [eqn (4)]. All potentials are state dependent and hence associated with a specific temperature of $T$. $k_{\mathrm{B}}$ is the Boltzmann constant.

The potential obtained *via* BI works well for simple molecules like polymers.[67] However, the BI scheme cannot be directly applied for complex molecules such as peptides. This can be explained by the assumptions underlying the BI scheme: (i) the bonded and nonbonded potential energy terms should be treated separately. (ii) All bonded DOFs should be uncorrelated (eqn (5)). In the event that these assumptions do not hold true, which is typically the case for peptides,[47,48,54,67] the IBI scheme is applied to refine the potentials. Further corrective schemes like addition of special bonded potentials are used in this study. This is discussed in the Results and discussions.

$$P^{\mathrm{CG}}(d, \theta, \phi, T) = P^{\mathrm{CG}}(d, T) \times P^{\mathrm{CG}}(\theta, T) \times P^{\mathrm{CG}}(\phi, T) \quad (5)$$

The IBI scheme is highly efficient in correcting Boltzmann inverted bonded and nonbonded potentials.[66] The procedure is described by eqn (6).

$$V_{i+1}(\rho) = V_i(\rho) + k_{\mathrm{B}}T \ln\left[\frac{P_i(\rho, T)}{P^{\mathrm{ref}}(\rho, T)}\right] \quad (6)$$

$\rho$ is representative of any arbitrary DOF ($d$, $\theta$, $\phi$, $r$). First, a CG simulation is run using an initial estimate of the potential ($V_i(\rho)$). Next, the resulting CG bonded and nonbonded distributions are examined. If there is a difference between a CG distribution ($P_i(\rho, T)$) and the corresponding reference distribution (CG-mapped distributions that are representative of atomistic sampling, *i.e.*, $P^{\mathrm{ref}}(\rho, T)$), the underlying CG potential is corrected. The difference between the potential of mean force of the CG distribution and the reference distribution (namely, second term in eqn (6)) is the correction used to develop a new potential, *i.e.*, $V_{i+1}(\rho)$. The process is repeated until the CG distribution converges with the corresponding reference distribution within a tolerance value.

The FM[49] scheme is extensively used to build CG tabulated potentials with the help of forces from AA trajectories.[63] In FM, the AA forces on the CG-mapped coordinates are projected onto the CG model. The method employs a least square approach to minimize the difference between forces in the AA and CG resolutions.

$$\chi^2 = \sum_{m}^{M} \sum_{n}^{N} \left| F_{mn}^{\mathrm{ref}} - F_{mn}^{\mathrm{CG}} \right|^2 \quad (7)$$

 

$M$ is the number of frames (or, MD configurations) and $N$ is the total number of CG beads. $F_{mn}^{ref}$ refers to the forces in the reference simulation (AA), acting on the $n$th CG bead in the $m$th configuration. Similarly, $F_{mn}^{CG}$ refers to the corresponding CG force. The FM scheme converts eqn (7) into a set of linear equations.[68] These equations can be solved only when they are overdetermined. That is, the number of parameters should be less than the product of the number of beads ($N$) and configurations ($M$). This can be ensured by extensive sampling of the AA system so as to yield a large value for $M$.

The Versatile Object-oriented Toolkit for Coarse-Graining Applications (VOTCA) package (version 1.5)[69] is used to determine the CG potentials. FM is used to determine the CG peptide–peptide nonbonded potentials. The peptide topology is modified to exclude bonded and intramolecular nonbonded interactions. This approach is termed as force matching with exclusions.[59] This modification works well when all the bonded and nonbonded potentials are used in conjunction to run CG simulations. The reference AA simulations generate 20 000 frames of force data. The force data from the AA simulation is reevaluated with the modified topology. The FM algorithm organizes the data into smaller blocks so as to have 100 frames in each block. Cubic splines are used to fit the force data across all blocks. The grid spacing between the minimum and maximum distances of a particular interaction can control the number of splines. An optimal number of splines would result in smooth CG potentials that preserve most of the atomistic details. A grid spacing of 0.03 nm is used which results in an average of 30 splines for each interaction.

The IBI method is used to determine the other nonbonded potentials, namely water–water and peptide–water potentials. Reference RDFs are generated from 100 000 frames of AA trajectory. Details of the IBI process for specific nonbonded interactions are discussed in the Results and discussions. Also, potentials involving ions are discussed in the Results and discussions.

For bonded potentials, 100 000 frames of configurations from the AA simulations are used. This results in smooth bonded distribution functions. BI is applied to obtain the CG bonded potentials and run the CG simulations. Any discrepancy in the comparison of the CG distributions with the reference distributions is addressed by iteratively correcting the CG bonded potential using IBI.

### All atom simulations

The AMBER99SB[70] force field is used to simulate all the $V_nK_m$ peptides. A previous study on $V_nK_m$ peptides employs the same force field. The AMBER force field accounts for effects such as hydrogen bonds between peptides and water. These interactions are critical to resolve chemical details such as secondary structure and peptide solvation. Hence, the AMBER force field generates good reference trajectories and forces for building the CG potentials. The AA modeling conditions are analogous to an earlier study.[47] The GROMACS Molecular Dynamics package (2016.1)[71–73] is used run initial energy minimization and equilibration simulations. The 2020.2 version of GROMACS is used to compute the AA trajectories and forces. The Lennard-Jones cut-off is 1.4 nm and the long range dispersion correction for energy and pressure is employed. The Particle Mesh Ewald (PME) algorithm is used[74] for the long range electrostatics. The system samples the *NPT* ensemble. The temperature is set at 300 K using the velocity-rescaling thermostat (with a stochastic term)[75,76] and a pressure of 1 bar using the Parrinello–Rahman barostat.[77] All bond distances are constrained by the LINCS algorithm.[78] A 2 fs time step (using the leap-frog integrator) is employed to run the AA simulations. An explicit description of water using the SPC/E[79] model is used. Details of runtimes and sampling rates are provided in the Results and discussions. The end-to-end distance of the peptide sequence of interest, *i.e.*, $V_6K_2$ is within the range determined by experiments.[6]

### Coarse-grained simulations

The final AA configuration file is converted to a CG representation using the csg_map tool in the VOTCA package.[69] The volume of the CG simulation box is the same as that corresponding to the equilibrated AA system. The 2018 version of GROMACS is used to run the CG simulations. At the time of development, tabulated potentials were not compatible with the 2020.2 version of GROMACS. The CG simulation is run at constant volume with a leap-frog stochastic dynamics integrator with a timestep of 2 fs.[80] The inverse friction coefficient is 1 ps. Bonded and nonbonded tabulated potentials are used for the CG simulations. Details of runtimes are provided in the Results and discussions.

### Backmapped-atomistic simulations

A code provided by a previous study[81] is used to project the AA coordinates onto the final CG configuration. Next, the backmapped-atomistic configuration is relaxed using the following steps:[81]

(1) 500 steps of energy minimization that excludes nonbonded interactions between peptides, water molecules and ions.

(2) 500 steps of energy minimization without any exclusions.

(3) A series of 4 short position restrained *NVT* simulation runs wherein the timestep is increased in the following order: 0.2, 0.5, 1 and 2 femtoseconds.

Finally, the backmapped-atomistic configuration is simulated in the same way as the original AA simulations without any position restraints. The root-mean-square deviation (RMSD) between the backmapped-atomistic and underlying CG configuration is measured at different intervals. It is noted that other sophisticated approaches[82] could also be employed for reintroducing atomistic details in a CG configuration.

## Results and discussion

A bottom-up coarse-graining approach is employed to preserve the structural properties of $V_6K_2$ while capturing its aggregation characteristics. AA simulation data is used to derive CG bonded and nonbonded potentials. These potentials work in conjunction to reproduce the overall structure of a single $V_6K_2$ peptide in aqueous solution. The AA solvation structure of the peptide is

reproduced in the CG model. In addition, internal DOFs like the linear extension of the hydrophobic tail group and the RDF of the backbone beads are in agreement with corresponding results from the AA simulations. The CG model is tested on systems that are not a part of the parameterization. Systems with multiple $V_6K_2$ peptides in aqueous solution are used for these tests. First, the structure of individual peptides in multi-peptide systems is assessed. Next, the relative organization of peptides in a micelle formed by 8 peptides is compared across the AA and CG resolutions. The effective size of the micelle formed in the AA and CG simulations are also compared. Furthermore, the morphology of an aggregate formed by a large number of peptides is validated against the supramolecular structure of self-assembled peptides reported by an experimental study.[64,65] The outcomes of the tests demonstrate that the CG model preserves the individual structure of the peptide while capturing its assembly. The steps followed to develop the CG model are summarized in Fig. 2.

### Water–water potentials

A system with 2100 water molecules in a simulation box (of dimensions 4 nm) is used to sample the water–water nonbonded interactions. The simulation is run in the *NPT* ensemble for 200 ns, and the trajectories from the final 50 ns are sampled for building a reference RDF. The RDF is inverted using BI to generate a CG potential. This potential serves as an initial estimate for the CG simulation. The volume of the CG simulation box is the same as the equilibrated AA simulation box. The simulation is run in the *NVT* ensemble using the leap-frog stochastic dynamics integrator[80] with a timestep of 2 fs. The resulting RDF for CG water–water interactions does not match with the reference RDF. Hence, IBI is applied to correct the CG potential. This correction requires 300 IBI steps, each running for 300 picoseconds. The resulting CG water–water RDF in a system with a single $V_6K_2$ peptide is in perfect agreement with the corresponding reference RDF (see Fig. S2, ESI†).

The water–water interactions do not incorporate any effects from other molecules (*i.e.*, peptides and ions). Hence, the CG water–water nonbonded potential is transferable across systems with varying peptide concentration. This is supported by the agreement between AA and CG RDFs of water–water interactions at different peptide concentrations (see Fig. S3, ESI†).

It is noted that a high concentration of peptides could disrupt the packing of water molecules. In this study, the formation of elongated micelles is reported with 128 CG peptides (discussed later). The water–water RDF of the corresponding AA simulation shows a slight disruption in the packing of water (see Fig. S4, ESI†). This is based on the observation that the water–water RDF does not decay to 1 at a distance of 1 nm (the water–water RDF of pure water decays to 1 at a distance below 1 nm).

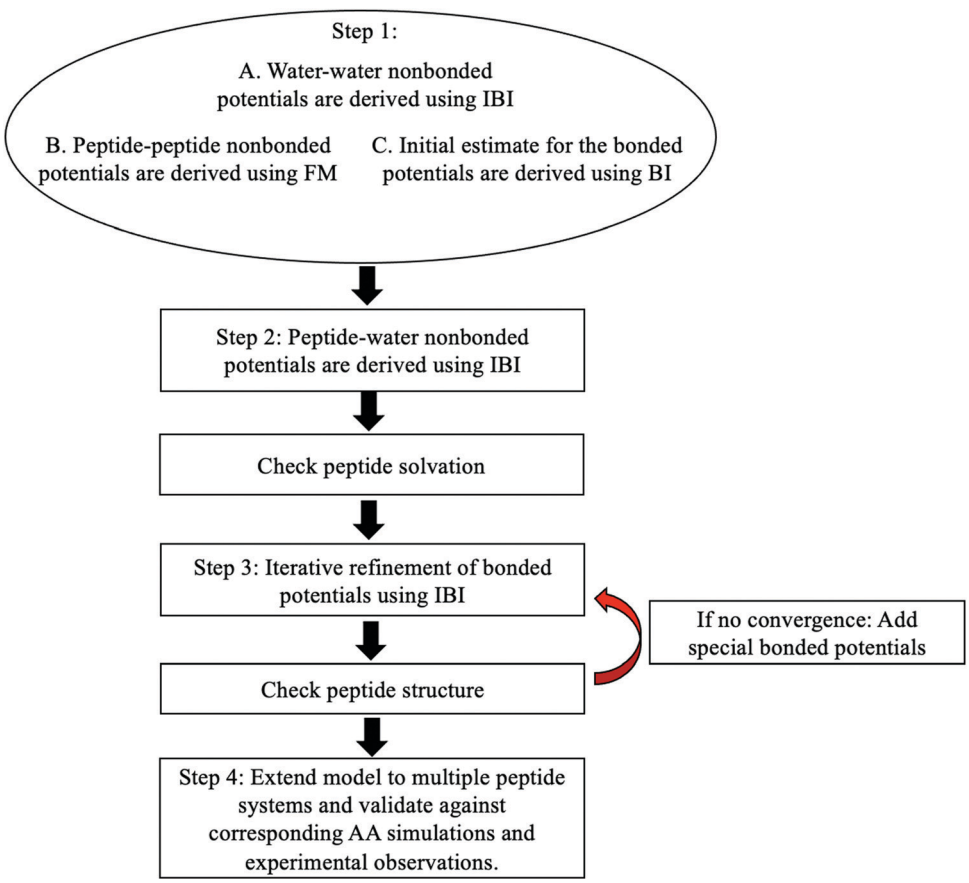


Fig. 2 The workflow for the systematic bottom-up coarse-graining of the $V_6K_2$ peptide sequence in aqueous solution.

 

The CG model is able to reproduce the water–water RDF of the AA 128 peptide model. This demonstrates that the CG water–water potential is transferable to systems where the packing of water is slightly disrupted. However, for higher concentrations, one can expect that these potentials may result in errors. One may need to make further refinements to the force field to account for the disruption of the packing of water molecules.

### Initial estimates for bonded potentials

The BI method is used to generate potentials for all bonded interactions. An AA system consisting of a single $V_2K$ peptide in aqueous solution serves as the reference system. The system is simulated for 500 ns. 100 000 coordinate frames of AA trajectory are processed (from the final 100 ns of AA simulation) to produce smooth CG bonded potentials. These potentials can be used as an initial estimate for bonded potentials in any $V_nK_m$ peptide sequence. These potentials can approximately model the internal structure of these peptides, and can be refined using IBI.

### Peptide–water potentials

To develop peptide–water nonbonded potentials, the peptide–peptide, water–water nonbonded and initial estimates of the bonded potentials need to be derived a priori. A single peptide in water is sampled for 1 microsecond. The simulation box has dimensions of 4 nm. Reference RDFs are constructed using 100 000 coordinate frames from the final 100 ns of AA trajectory. A 100 step IBI procedure generates CG potentials that reproduce the reference RDF. Each IBI step runs for 300 picoseconds. The transferability of these potentials is assessed by applying the peptide–water potentials developed for the CG $V_2K$ peptide on the CG $V_6K_2$ peptide. Fig. S5 (ESI†) shows that the solvation structure of the CG $V_6K_2$ peptide is primarily in agreement with the corresponding AA reference system. There are extremely minor differences between the AA and CG RDFs at some of the peak positions. Since the AA trajectory of the $V_6K_2$ peptide sequence was available at the time of force field development, it was used to build separate peptide–water potentials for the CG model of $V_6K_2$. This improves the agreement between the AA and CG peptide–water RDFs (see Fig. 3). The agreement between these RDFs demonstrates the correct solvation of the peptide in the CG representation.

### Potentials involving ions

An effective CG potential accounts for all underlying effects that are explicitly modelled in the AA representation of the system.

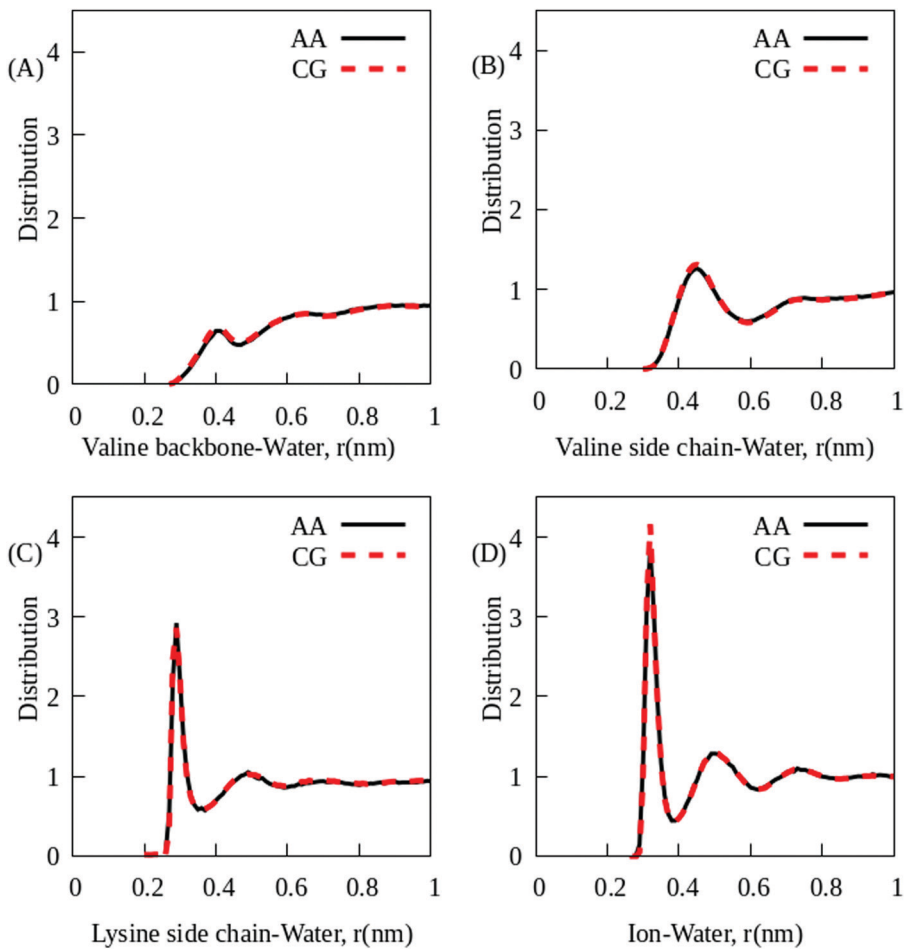


Fig. 3 Peptide–water and ion–water radial distribution functions. (A–C) shows the solvation of the peptide beads. (D) shows the same for the chloride beads. The black and red curves represent reference and CG distributions, respectively.

Hence, interactions involving ions are modelled with effective CG potentials that account for the electrostatic, van der Waals and other interactions present in the AA model. Three types of effective CG potentials are developed to account for the ions: ion–ion, peptide–ion and ion–water potentials. The objective of these potentials is to capture the electrostatic interactions between charged moieties, namely Chloride ions and lysine residues. The ion–ion and peptide–ion potentials are resolved using FM along with the peptide–peptide nonbonded potentials, thereby preserving the electrostatic forces between the moieties in the CG model. The procedure for building FM potentials is described later. Consequently, across all CG simulations, ions are well-coordinated with the lysine residues (Fig. S6, ESI†).

Finally, the ion–water potential is developed along with the peptide–water potentials. Fig. 3D shows agreement between the reference and CG RDFs. There is a minor discrepancy in the height of the first peak of the CG RDF. Further refinement of the potential is avoided as it impacts the accuracy of the peptide–water potentials (as ion–water and peptide–water potentials are resolved simultaneously using IBI). However, the ion–water potential ensures that the overall solvation structure of the CG ion beads is correct.

### Coarse-grained bonded potentials

The initial estimates for the potentials for all the bonded interactions are derived by BI. The potentials are further refined *via* IBI. This procedure is aimed at building a transferable CG model that could be applicable to longer peptides in the $V_nK_m$ series. The bonded potentials are constructed from an AA reference of a short peptide ($V_2K$), and the transferability of the potentials to longer peptide chains (for example: $V_6K_2$) is assessed. Since $V_6K_2$ is significantly longer than $V_2K$, the length of the peptide chain is increased in a stepwise manner ($V_2K \rightarrow V_4K \rightarrow V_6K_2$), as shown in Fig. S7 (ESI†). At every step, corrections are made to the set of potentials, and the updated set of potentials are transferred to the subsequent step. The solvent is explicitly represented in all AA and CG systems. The corrections in the CG potentials for the $V_4K$ and $V_6K_2$ sequences are almost identical. The CG model development for $V_6K_2$ is discussed in detail as it is the peptide sequence of interest.

Five types of bonds are defined for any peptide sequence in the $V_nK_m$ series: valine backbone ($V_{Bx}$–$V_{B(x+1)}$), valine side chain ($V_{Bx}$–$V_{Sx}$), valine–lysine connector ($V_{Bx}$–$K_{By}$), lysine long bond ($K_{By}$–$K_{Sy1}$) and lysine short bond ($K_{Sy1}$–$K_{Sy2}$). Schematic representations of these bonds are shown in Table S1 (ESI†).

The BI method is used to generate the potentials for the shorter bonds in the peptide, namely the valine side chain and lysine short bond (see the width of distributions in the lower panel of Fig. 4). These CG potentials are generated by processing the AA trajectory of a $V_2K$ peptide, and are transferable to the CG representation of a $V_6K_2$ peptide (see agreement between the AA and CG curves in Fig. 4D and E). On the other hand, BI is unable to generate suitable potentials for the longer bonds (see the width of distributions in the upper panel of Fig. 4). The resultant

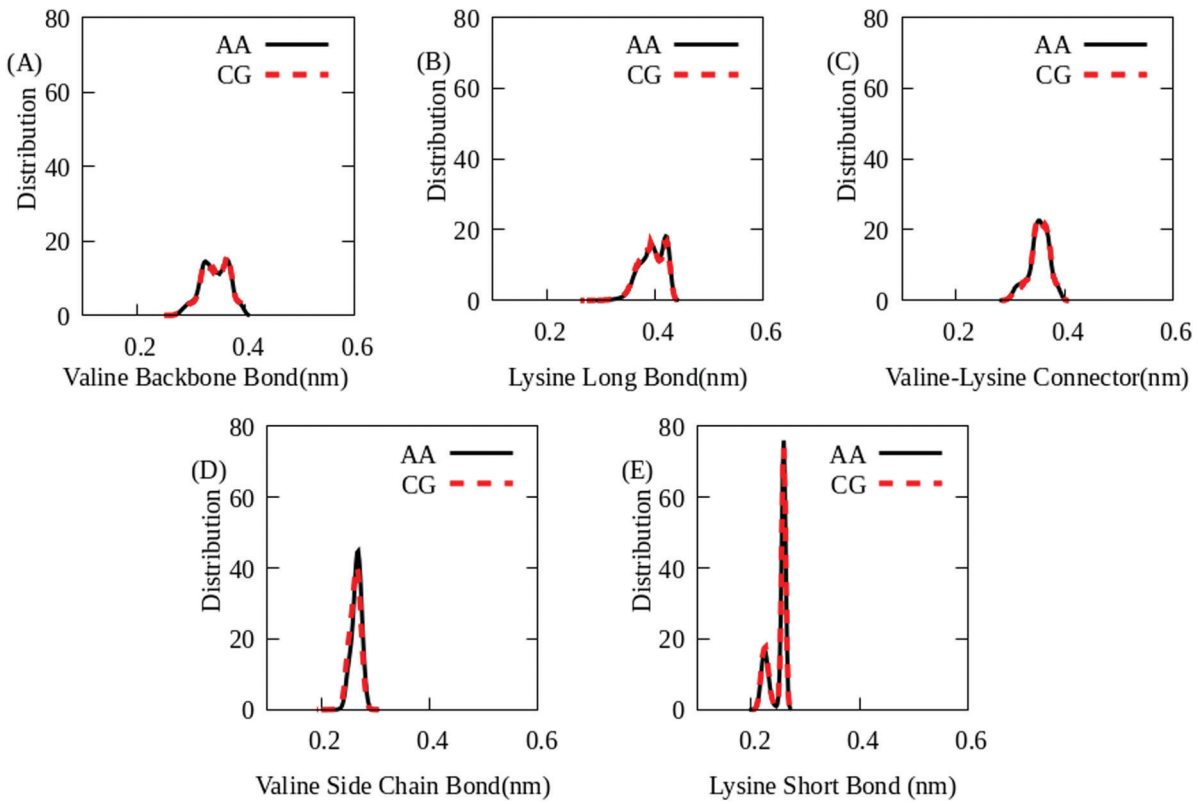


Fig. 4 Comparison of the reference (black) and CG (red) distributions for (A–C) long and (D and E) short bonds in the $V_6K_2$ peptide sequence.

 

CG distributions only sample one of the peaks in the underlying reference distribution. Hence, these potentials provide constrained conformations as they ignore significant AA details. Such errors arise whilst dealing with longer bonds, especially the ones that have a bimodal distribution. It is surmised that these types of bonds do not satisfy the assumptions of BI (see Methods). The CG potentials associated with these bonds are corrected by applying IBI. Reference distributions for these bonds in $V_6K_2$ are required to initiate the IBI process. The AA representation of a single $V_6K_2$ peptide in water is sampled for 1 microsecond. 100 000 coordinate frames from the final 100 ns of AA trajectory are processed to generate reference bond distributions. Three steps of IBI are applied to correct the bonded potentials. Each IBI step runs for 10 ns. Fig. 4 shows the resulting CG distributions to be in agreement with the reference distributions.

The BI method is unable to derive the correct CG angle potentials. Majority of the reference angles only reproduce one peak. Seven steps of IBI are applied to correct all the angle potentials (see Fig. S8, ESI†). Each IBI step runs for 20 ns.

The dihedral potentials developed using AA references for $V_2K$ and $V_4K$ are partially transferable to the CG model of $V_6K_2$. To improve the agreement between the AA and CG dihedral distributions for $V_6K_2$, the AA reference trajectories for a single $V_6K_2$ peptide is sampled. The BI method processes 500 000 coordinate frames from the final 500 ns of AA trajectory. This procedure generates smooth dihedral potentials. Fig. S9 (ESI†) shows the CG dihedral potentials to sample the majority of the underlying AA distributions. Hence, the local structural details are preserved in the CG model. However, there are some minor inconsistencies: the CG potential for the $V_{Bx}$–$V_{B(x+1)}$–$V_{B(x+2)}$–$V_{B(x+3)}$ dihedral slightly over-samples the more likely conformation at −140 degree. On the other hand, the potential slightly under samples the less likely conformation at +140 degrees. The same behavior is observed for the $V_{Sx}$–$V_{Bx}$–$V_{B(x+1)}$–$V_{S(x+1)}$ dihedral. Further refinement of these potentials using IBI impacts other interactions in the CG model. Since majority of the AA details are preserved by the CG dihedrals, further refinement of the associated potentials is not performed.

### Peptide–peptide potentials

The main objective of these nonbonded potentials is to model the interpeptide interactions that govern the assembly of peptides, and the intramolecular nonbonded interactions that contribute to the structure of individual peptides. Both types of interactions are resolved by the same set of CG potentials. The FM method with exclusions[59] is employed to develop these potentials. This modification excludes all forces associated with bonded and intramolecular nonbonded interactions from the net forces in the system during the FM process. In this way, the FM process only matches the intermolecular forces rather than the forces associated with the bonded and intramolecular nonbonded interactions. A previous study[59] shows that the exclusion approach results in good agreement with the corresponding reference system. In the context of this work, it is surmised that intramolecular nonbonded interactions could significantly vary as the peptide chain length increases. Hence, the exclusion of intrapeptide forces could enhance the transferability of the peptide–peptide nonbonded potentials (namely, to longer peptide molecules in the $V_nK_m$ series). However, this process leads to loss in peptide backbone conformation as the CG intramolecular nonbonded potentials do not account for AA intramolecular nonbonded interactions. This is corrected by employing special bonded potentials which will be discussed in a later section. An AA representation of 25 $V_2K$ peptides solvated with 1600 water molecules in a simulation box (of dimensions 4 nm) serves as the reference system. This system has a high peptide concentration of 0.65 M so as to sample the reference forces that govern peptide assembly. Twenty-five monovalent chloride ions are added to maintain charge neutrality of the system. This system is simulated for 500 ns in the *NPT* ensemble. The final 100 ns of simulation trajectory is used to extract force data. The FM algorithm compiles the force data to generate CG potentials. The peptide–peptide nonbonded potentials will be best suited for systems with the same peptide concentration as the current system (*i.e.*, 0.65 M).

To access the accuracy of the peptide–peptide nonbonded potentials, the RDFs of the AA and CG representations of the 25 $V_2K$ peptide system are compared (see Fig. S10, ESI†). There is overall qualitative agreement between the AA and CG RDFs. The major peak positions of the RDFs are in agreement with each other. However, due to coarse-graining, the CG RDFs are unable to capture the ordering between the peptides. For example, Fig. S10A (ESI†) shows that the CG RDF averages over the atomistic details between 0.3 to 1 nm. This behavior could be a consequence of coarse-graining wherein the underlying effect of empty spaces in the AA fragments are smoothened out in the CG model. Also, since FM is not designed to match the structure of the underlying AA interactions, the CG model is unable to reproduce the local structure of these nonbonded interactions. A previous study on FM enhances the agreement with the underlying structure using 3-body potentials.[57,58] Currently, this approach is beyond the scope of this study and will be pursued in future. Further, Fig. S10D (ESI†) shows the distribution of ions around the peptides in the AA and CG systems. The position of the major peak of the CG RDF is in agreement with that of the AA RDF. However, the CG peak is higher than the AA peak, thereby indicating that the ion concentration near the peptide surface is higher in the CG simulation. This could be a consequence of implicitly accounting for electrostatics with effective CG potentials.

These nonbonded potentials contribute to the individual structure and assembly of the peptides. These features are examined in subsequent sections and serve as validation for the peptide–peptide nonbonded potentials.

### Peptide chain conformations

The equilibrium chain conformations of the peptide are governed by the CG model. The end-to-end distance is measured to characterize the chain conformation of the $V_6K_2$ peptide (see Fig. S11A, ESI†). This is the distance between the two capping residues at the ends of the peptide. In Fig. S11A (ESI†),

the peak at 0.25 nm in the CG distribution corresponds to a folded conformation of the CG peptide. This conformation could be attributed to the unrestrained flexibility of the CG peptide backbone and is due to the assumptions of the FM with exclusions method. The method derives the peptide–peptide nonbonded potentials by excluding bonded and intramolecular nonbonded interactions. Whereas this approach yields peptide–peptide nonbonded potentials which are transferable across the $V_nK_m$ series, the intramolecular nonbonded potentials are not derived explicitly. To obtain further clarity on the backbone flexibility, the degree of bending of the backbone in the AA and CG representation of the peptide is compared. This comparison requires the measurement of the 1–3–5 angles ($V_{Bx}$–$V_{B(x+2)}$–$V_{B(x+4)}$ and $V_{Bx}$–$V_{B(x+2)}$ – $K_{By}$) across the peptide backbone.

Fig. S11B (ESI†) shows that the region between 20–40 degrees is sampled by the CG simulation but not by the AA simulation. This could be a consequence of the unrestrained flexibility of the peptide backbone. Hence, 1–3–5 special angles are derived from BI to reduce the flexibility of the CG peptide backbone. These potentials are further refined by five steps of IBI. This procedure does not yield good quality potentials if the regular backbone angles (*e.g.*, $V_{Bx}$–$V_{B(x+1)}$–$V_{B(x+2)}$) are fixed. This observation could be attributed to the correlations between the conformations sampled by the 1–3–5 special angles and the regular backbone angles. Hence, these angles are simultaneously refined using IBI. Fig. 5A shows that these refinements result in perfect agreement between the end-to-end distances corresponding to the AA and CG representations of the peptide. This result indicates that all CG potentials work in conjunction to preserve the chain conformation of the peptide. However, less likely conformations between 40–80 degrees are not sampled by the CG model (see Fig. 5B). This minor inconsistency is ignored as the refinements are responsible for the good agreement in the chain conformation of the peptide.

Another measure of the chain conformation, namely the radius of gyration, corresponding to the AA and CG representations of the peptides are compared. This measurement quantifies the effective dimension of the peptide. Fig. S12 (ESI†) shows the radius of gyration of the CG representation of the peptide to be in agreement with the corresponding value for the AA representation of the peptide. This observation further demonstrates that the CG model preserves the chain conformation of the peptide. The solvation of the peptide remains unaffected by the refinement and addition of bonded potentials.

In addition to the chain conformation of the peptide, key internal structural properties such as the linear extension of the long hydrophobic tail group and the RDF of the backbone beads need to be validated. Fig. S13A (ESI†) compares the extension of the hydrophobic tail group using the AA and CG models. The less likely conformation yielding a distance of 0.8 nm is not sampled by the CG model. This could be attributed to minor inconsistencies in the CG model (*i.e.*, some of the CG dihedral potentials ignore the less likely conformations sampled in the AA simulation). However, given the chemical complexity of the amphiphilic peptide sequence (long hydrophobic and charged hydrophilic blocks), the internal structure of the CG peptide is well aligned with the reference data. The internal structure of the peptide backbone is further investigated using the RDF between the backbone beads. This measurement includes six valine and two lysine backbone beads. Fig. S13B (ESI†) shows the CG RDF to be in good agreement with the reference distribution. The RDF at distances greater than 1 nm (see inset of Fig. S13B, ESI†) is nearly in perfect agreement with the AA RDF. This result is significant as these interactions are not explicitly modeled by any specific CG potential. These interactions are cumulatively governed by bonded and nonbonded potentials. Hence, this result validates the use of a hybrid approach towards deriving bonded and nonbonded potentials.

### Time scales and computational efficiency

In CG simulations, the reduced DOFs and smooth interaction potentials can accelerate the dynamics of the molecules.[83,84] Earlier studies have compared the AA and CG timescales using various methods.[63,85] One approach is to compare the diffusion coefficients of a single peptide in the two representations. The self-diffusion coefficient is determined using AA and CG

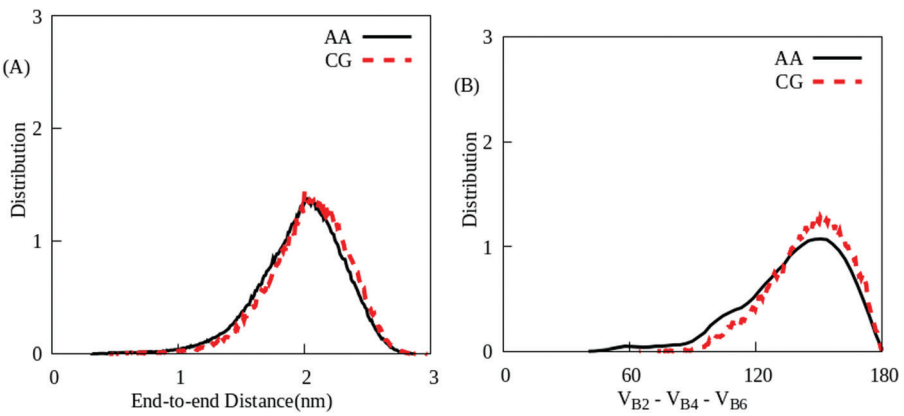


Fig. 5 The (A) end-to-end distance and a (B) selected 1–3–5 angle of a $V_6K_2$ peptide. The black and red curves represent the AA and CG distributions, respectively.

 

trajectories. The ratio of the coefficients in the CG ($D_{CG}$) representation to the corresponding value in the AA representation ($D_{AA}$) provides an estimate of the acceleration in the dynamics. Values of $D_{CG}$ and $D_{AA}$ are provided in Table S2 (ESI†). The speed-up factor, *i.e.*, $D_{CG}/D_{AA}$ is $\sim$1.5 for a single $V_6K_2$. This means that the effective CG simulation time is approximately 1.5 times that of the AA simulation time. However, in the case of multiple peptide systems, the speed-up factor could be a function of the peptide concentration. Hence, for systems with multiple peptides, the study will focus on the final equilibrated structures observed in the respective models. However, the performance (as reported by the MD software in nanoseconds per day) of the AA and CG models can be compared for multiple peptide systems. An 8 peptide system (discussed in the next section) is scaled across a selected number of CPUs on a supercomputer (see Fig. S14, ESI†). The system is solvated with approximately 2000 water molecules in a simulation box of dimension 4 nm. The CG model has a performance of 430 ns per day with 8 CPU cores. Whereas the AA model has a performance of 110 ns per day on the same computing resources. Hence, coarse-graining yields a gain in computational efficiency (ratio of the performance of the CG and AA models) by at least a factor of 4. Further, the CG model can resolve the self-assembly of 128 peptides into an elongated micelle within a short time span. This is not feasible with the corresponding AA model. It is noted that the gain in computational efficiency is not sufficiently high. This could be attributed to the bottom-up coarse-graining approach. Here, several bonded potentials consist of multiple energy wells of varying depths. These features are particularly observed in the dihedral and 1–3–5 angle potentials. The complex nature of these potentials captures the underlying chemistry of the AA fragments. Additionally, these potentials are used in tabulated form instead of fitting the energy values with an analytical function. Due to these reasons, a low CG simulation time step of 2 femtoseconds is employed, thereby resulting in a low gain in computational efficiency. It is noted that another bottom-up study reports a similar gain in computational efficiency.[47] The study accurately resolves the AA conformation of peptides in the CG model with the same CG simulation time step. In future, following a previous study on relatively longer peptides,[63] the mapping scheme of the $V_nK_m$ peptides will be modified to investigate its effect on the computational efficiency.

### Multiple peptide systems

Prior to using the CG model for investigating self-assembly, the model must be examined for its ability to preserve the structure of a peptide in multi-peptide systems. Hence, the CG model is tested on systems encompassing 2, 4, 8 and 16 solvated peptide molecules. The corresponding AA systems are generated for comparison. The concentration of these systems ranges from 0.05 M to 0.29 M. With the exception of the 16 peptide system, all the other systems are simulated in a box of dimensions 4 nm. The 16 peptide system is placed in a simulation box of dimension 4.5 nm. Both AA and CG simulations start from an initial configuration where the peptides are randomly distributed, solvated in a simulation box and the system is equilibrated using the *NPT* ensemble for 5 ns. Next, the production AA and CG simulations are simulated for 500 ns and 200 ns, respectively. On some occasions, the CG simulations are run for longer durations to obtain better statistics. The final 100 000 coordinate frames are sampled for analysis.

Fig. 6 shows the end-to-end distances of peptides in the multi-peptide systems. There is good agreement between the AA and CG distributions for all except the 2 peptide system. The CG distributions for 2, 4 and 8 peptide systems are nearly the same, and is equivalent to that of the single peptide system. This result could be attributed to the bonded potentials that are fitted to the single peptide reference system. This could explain the disagreement in the 2 peptide system (see Fig. 6B). To improve the agreement between these distributions, the bonded potentials for this system can be further refined using a suitable reference (AA trajectory for the 2 peptide system). In this study, further refinement of the bonded potentials is not performed as there is overall agreement in the structure associated with AA and CG representations of the peptides. In addition, the CG model only captures the major peak of the 16 peptide system at $\sim$2.5 nm. The compact conformations between 1 nm and 2 nm are not sampled by the CG model. These conformations have a low probability of occurrence, and hence further refinements to the force field are not performed.

Fig. S15 (ESI†) compares the radius of gyration of the AA and CG representation of the peptides in multi-peptide systems. There is qualitative agreement between the AA and CG simulations for all except the 2 peptide system. Fig. S15 (ESI†) shows that the AA distributions becomes narrower as the peptide concentration is increased. This means that the conformation of these peptides becomes relatively restrained as the peptide concentration is increased. Except for the 2 peptide system, the CG model is able to capture this behavior. This is significant as the restraining effect was not a part of the parameterization. On some occasions, the individual AA peptides sample relatively compact conformations. This can be seen in the 8 and 16 peptide systems where the peaks of the AA distributions are at relatively smaller values ($\sim$0.82 nm) in comparison to the peaks of the CG distributions ($\sim$0.85 nm). Further, the discrepancy in the 2 peptide system could be attributed to frequent conformational changes (see Fig. S16A and B, ESI†). It is observed that both AA and CG simulations of 2 $V_6K_2$ peptides yield a peptide dimer (see inset in Fig. S16A and B, ESI†). The AA peptides organize antiparallel to each other in the dimer (Fig. S16A, ESI†). This organization between the peptides is preserved for the entire duration of the AA simulation. Whereas the CG peptides organize in both antiparallel and parallel orientations (Fig. S16B, ESI†). This means that the CG peptides frequently undergo conformational changes during the course of the simulation. This could explain the wider distribution of the radius of gyration values in the CG simulation (see Fig. S15B, ESI†). To summarize, the end-to-end distance and radius of gyration calculations show that the CG model is able to preserve the overall structure of $V_6K_2$ peptides even in the presence of interactions with neighboring peptides. The latter is not explicitly accounted for in the parameterization process, and hence demonstrates the robustness of the CG model.

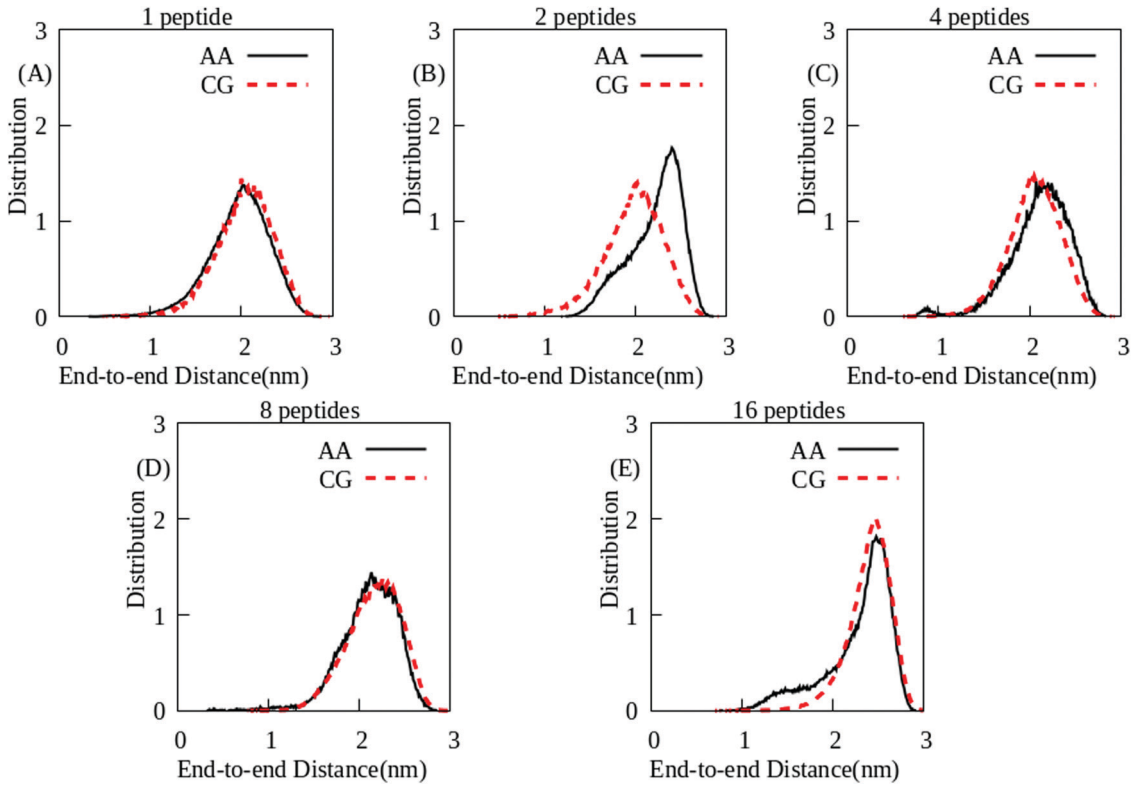


Fig. 6 The end-to-end distances of peptides in the (A) 1, (B) 2, (C) 4, (D) 8 and (E) 16 peptide systems. The black and red curves represent AA and CG distributions, respectively.

The assembly of the peptides is tested on the 8 peptide system with a peptide concentration of 0.21 M, which is within the same order of magnitude at which the peptide–peptide nonbonded potentials are developed. Both the AA and CG simulations yield a self-assembled micelle (see Fig. 7A and B).

The final CG configuration (Fig. 7B) is backmapped to AA coordinates to test the stability of the CG configuration in the AA representation. The backmapped-atomistic configuration is simulated for 1 ns without any restraints. The RMSD between the backmapped-atomistic and the initial CG (see Fig. 7B) configuration after 1 ns is 0.5 nm. An additional 10 ns of dynamics increases the RMSD to 0.8 nm (see the final backmapped-atomistic configuration in Fig. 7C). This deviation from the underlying CG configuration is very small given that the backmapped-atomistic simulation explicitly considers a higher number of DOFs.

To compare the relative organization of the peptides in a micelle, the interactions between the hydrophobic residues are investigated. Fig. S17 (ESI†) shows that the hydrophobic residues of the peptides are aligned with each other. Intermolecular RDFs between first (V1) and last (V6) valines in the $V_6K_2$ peptide sequence characterizes the relative orientation of these peptides (see Fig. S16E and F, ESI†). A higher count of either V1–V1 (black lines) or V6–V6 (red lines) interactions is indicative of parallel orientation between the peptides. Whereas a higher count of V1–V6 interactions (blue lines) is indicative of antiparallel orientation between the peptides. The antiparallel orientation is dominant in the AA 8 peptide system (Fig. S16E, ESI†). Whereas Fig. S16F (ESI†) shows that the antiparallel orientation is lost to a certain degree in the CG 8 peptide system.‡ In addition, the intermolecular RDFs show that the CG representation of the peptides prefer to pack tightly within the micelle. This observation is based on the starting points of the AA and CG RDFs. The starting point of the AA RDFs is ~0.5 nm and the starting point of the CG RDFs is ~0.25 nm. This discrepancy can be explained by the underlying effects modeled in the AA and CG simulations. The AA model employs the AMBER force field that accounts for interactions such as hydrogen bonding between peptides and water. These interactions result in higher number of peptide–water interactions. Fig. S18 (ESI†) reports higher solvation of AA peptides in comparison to the CG peptides. These interactions with water could hinder the AA peptides from packing tightly within the micelle. On the other hand, the CG model does not explicitly model interactions such as hydrogen bonding between peptides and water, thereby resulting in a tighter packing of the peptides.

The dimension of these micelles are compared by measuring their radius of gyration (see Fig. S19, ESI†). The AA simulation samples radius of gyration values between 1 and 3 nm. The CG simulation samples a slightly wider range which encompasses values between 1–3 nm. In addition, the probability of finding a compact conformation of the micelle

‡ Simulation files for the CG 1 and 8 peptide systems are available in a GitHub repository.[87]

 

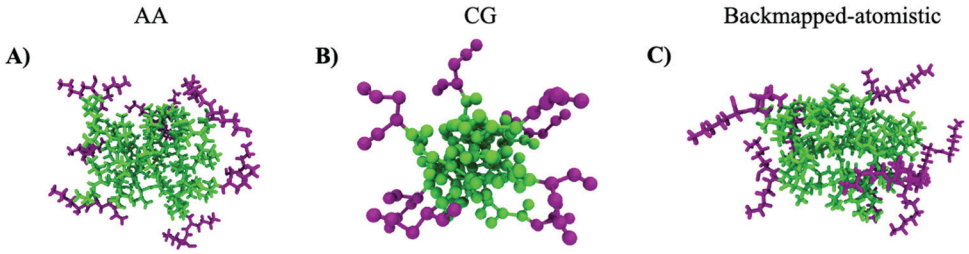


Fig. 7 A micelle consisting of eight peptides in (A) AA and (B) CG resolutions. (C) Equilibrated backmapped-atomistic conformation of the micelle. Color scheme – green: valine and purple: lysine.

(between 1–1.5 nm) is relatively higher in the CG model. These compact conformations are representative of the tight packing of the CG representation of the peptides in the micelle (see Fig. S16E and F, ESI†). As previously discussed, this could explain the under-solvation of the valine backbone and side chain residues in the CG representation (see Fig. S18, ESI†). This indicates some loss in transferability of the CG potentials. The degree of peptide solvation can be addressed by tuning the peptide–water potentials with a suitable reference system (*i.e.*, an AA system with 8 peptides). It is noted that reducing the strength of the interpeptide potentials does not correct the packing of the valine residues in the hydrophobic core of the micelle. However, further refinement of the peptide–water CG potentials is not deemed necessary as there is overall agreement on the size of the micelles. For the backmapped-atomistic system, the distribution of radius of gyration values is approximately at the center of the AA and CG distributions. The solvation of the backmapped-atomistic peptides is in agreement with corresponding results from the original AA simulation (see Fig. S18, ESI†), demonstrating the preservation of a solvation structure which is consistent with the AA description. These results demonstrate the stability of the aggregates sampled by the CG model.

It is noted that the individual conformations of the AA peptides are relatively more compact than CG peptides (see Fig. S15D, ESI†). However, the intermolecular RDFs (Fig. S16E and F, ESI†) shows that the AA peptides are not as tightly packed in the micelle as the CG peptides. Hence, even with individual AA peptides sampling compact conformations, the resultant micelle has a larger size.

To further understand the relative organization of the peptides, the interactions between all the hydrophobic residues are investigated. The probability of interactions between all possible valine residues within a cut-off distance (*i.e.*, 0.65 nm) are presented as a contact map[34] (see Fig. S17, ESI†). The backbone section of each valine residue is considered for this measurement. The CG-mapped coordinates of the AA model are used to compute the contact map for the AA systems. Hence, the interactions between the valine residues can be compared across the AA and CG resolutions. In the contact maps (see Fig. S17, ESI†), V1 represents the valine residue farthest away from the lysine residues, and V6 represents the valine residue closest to the lysine residues (see Fig. 1). The AA contact map shows a high probability of interactions between V2, V3 and V4 residues (see lower middle section of Fig. S17A, ESI†). This suggests that these valine residues from neighboring peptides are aligned which is a characteristic of a β-sheet-like organization. In addition, the high probability of interactions along an edge of the map (V4–V1, V5–V1 and V6–V1) indicate a preference towards antiparallel orientation between the peptide chains. The antiparallel orientation minimizes the repulsions between the lysine residues by increasing the distances between lysine residues. Furthermore, V2–V2 and V3–V3 interactions are preferred over V5–V5 and V6–V6. This indicates that the valine residues that are closer to the N terminus (*e.g.*: V2, V3) have a higher affinity towards each other as compared to the residues closer to the C terminus (*e.g.*: V5, V6). As reported in a previous study,[34] this result could be due to a cooperative hydrophobic effect (namely, V2, V3 have more hydrophobic neighbors as compared to V5, V6 that have Lysine neighbors on one side). It is surmised that the cooperative hydrophobic effect plays a role in stabilizing the peptide aggregate. Finally, it is noted that the AA contact map is highly asymmetric (for example, interactions like V1–V6 is not equivalent to V6–V1). This means that peptides are not perfectly aligned with each other in the micelle. It is surmised that the high magnitude of electrostatic repulsions between the lysine residues could prevent perfect alignment between the peptides.

Similarly, the CG contact map (Fig. S17B, ESI†) demonstrates the probability of interactions between the valine residues in the CG model. The plot shows that relative organization of the peptides is consistent with corresponding results using the AA description. A high probability of interactions between V2, V3 and V4 residues indicates β-sheet-like organization. In addition, the cooperative hydrophobic effect is reproduced in the CG model; namely, the V2–V2 and V3–V3 interactions are greater than the V5–V5 and V6–V6 interactions. However, due to the removal of selected AA features, a few discrepancies are observed; namely, the CG model shows a diminished preference for antiparallel orientation. Since lysine–lysine repulsions are not explicitly modeled by electrostatic potentials, the CG $V_6K_2$ peptides show a reduced tendency towards organizing in the antiparallel orientation. Furthermore, the distribution of hues is different in the AA and CG contact maps. The distribution of the hues in a contact map reflects the relative packing of hydrophobic residues in the micelle. The CG contact map reports a uniform distribution of hues: the packing of the hydrophobic residues gradually decreases from the center (V3 or V4) to the sides (V1 or V6) of

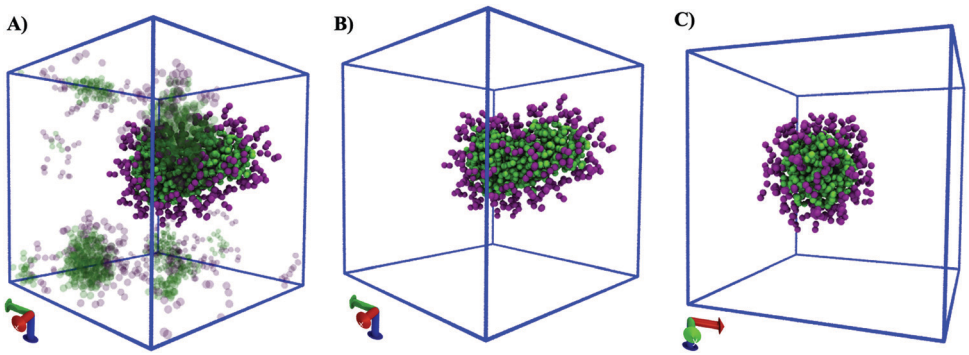


Fig. 8 Self-assembly of 128 CG $V_6K_2$ peptides. (A) Self-assembled nanostructures after $10^7$ timesteps of the CG MD simulation. The largest aggregate, *i.e.*, an elongated micelle is represented by regular beads, and the smaller aggregates are represented by translucent beads. (B) Side and (C) end views of the elongated micelle. The smaller aggregates are removed from (B) and (C) for clarity. Also, water and ion beads are removed from (A–C) for clarity. Color scheme – green: valine and purple: lysine.

the map. On the other hand, the AA contact map (Fig. S17A, ESI†) reports a relatively random distribution of hues. Thus, the hydrophobic residues in the AA simulation do not pack in a uniform manner as reported by the CG simulation. The discrepancy could be a consequence of coarse-graining wherein the empty spaces associated with the underlying AA fragments are smoothened out in the CG model. It is noted that the CG contact map is relatively more symmetric in comparison to the AA contact map. It is surmised that the peptides are more aligned with each other as the effect of electrostatic repulsions between lysine residues is reduced in the CG model. This could be a consequence of the using effective CG potentials.

Finally, Fig. S17C (ESI†) demonstrates the probability of interactions between the valine residues in the backmapped-atomistic model. It shows the reproduction of two structural properties; namely, the relative packing of hydrophobic residues, and preference towards the antiparallel orientation. Also, the contact map is asymmetric like the AA contact map (Fig. S17A, ESI†). All these features are reproduced as the CG model correctly samples the overall structure of the peptides.

In summary, the morphology of the final aggregate and relative organization of the peptides are in agreement with corresponding results from the AA simulations. Hence, all CG bonded and nonbonded potentials accurately model peptide aggregation. In addition, the CG model has successfully captured the relative organization of the peptides within an aggregate in a short span of time relative to the AA model.

Fig. S16 (ESI†) shows the intermolecular RDFs for 2, 4, 8 and 16 peptide systems. In the AA simulations, it is observed that peptides prefer to organize in a parallel orientation with increasing peptide concentration. For smaller peptide concentrations (namely the 2 peptide system), the AA simulation only samples the antiparallel orientation (Fig. S16A, ESI†). The corresponding CG simulation (Fig. S16B, ESI†) samples both antiparallel and parallel orientations (the antiparallel orientation is relatively preferred). On the other hand, at higher peptide concentrations (namely the 16 peptide system), the AA simulation primarily samples the parallel orientation (Fig. S16G, ESI†). The corresponding CG simulation primarily samples the same orientation (Fig. S16H, ESI†). This shows qualitative agreement between the intermolecular packing of peptides in the AA and CG simulations.

The ability of the CG model to capture the self-assembly of $V_6K_2$ peptides into large nanostructures reported by experiments is assessed. Earlier experiments have reported $V_6K_2$ peptides to assemble into nanorods for concentrations ranging from $\sim$0.05 to 0.5 mM.[64,65] In this computational study, a much higher peptide concentration of 0.21 M is chosen. This enables faster self-assembly of the peptides at a lower computational cost.[53,80] The system of interest encompasses 128 $V_6K_2$ peptides in the CG representation in a simulation box of dimension 10 nm. The peptides are solvated with approximately 26 000 water molecules, and 256 chloride ions are added to maintain charge neutrality of the system. It is observed that the majority of the peptides assemble into an elongated micelle (see Fig. 8) which is a precursor to nanorods and thereby consistent with experimental observations.[64,65] Another study using the Martini model has reported the formation of a large nanorod using longer simulations.[86] To permit large scale assembly within a bottom-up CG framework, appropriate references (an AA model of 128 $V_6K_2$ peptides) need to be integrated into the parametrization process. This approach is out of scope of the present study and will be pursued in a future work.

## Conclusions

In this study, a hybrid CG model is used to resolve the structure of a $V_6K_2$ peptide and capture its assembly in aqueous solution. The underlying AA solvation structure of the peptide is perfectly reproduced in the CG model. The CG model is developed with AA trajectories and forces as references, and hence accounts for complex DOFs that govern the structure and solvation of peptides. In addition, due to the reduced representation of the system, longer spatiotemporal scales can be easily accessed. The CG model provides a gain in computational efficiency by a factor of 4. The structure of individual $V_6K_2$ peptides are tested in multiple peptide systems. The potentials preserve the structure of the individual peptides in the presence of interactions with neighboring peptides. The aggregation of a few

 

peptides is successfully validated against corresponding results from AA simulations. In addition, the self-assembled nanostructure formed by a large number of these CG peptides is in agreement with prior experimental observations.[64,65]

This study outlines a method to resolve the structure of peptides within aggregates. This is critical towards understanding structure–function relationships of tunable peptide sequences. Hence, this method can be extended to other molecules that assemble into nanostructures. Future improvements to the CG model can be made with additional AA simulations to enhance the agreement with multiple peptide systems. In addition, the use of complex potentials to better approximate the many-body PMF could improve the transferability of the potentials across various systems.[55–58] This could resolve the structure of peptides within much larger nanostructures. Furthermore, the impact of local interactions on the properties of the nanostructure spanning large spatiotemporal scales, such as curvature and bending rigidity, can be investigated.

## Conflicts of interest

There are no conflicts of interest to declare.

## Acknowledgements

M. D. would like to acknowledge NSF CAREER DMR-1654325 for financial support. Portions of the research presented used computational resources supported by NSF XSEDE (allocation DMR-140125).

## Notes and references

1 R. Das, P. J. Kiley, M. Segal, J. Norville, A. A. Yu, L. Wang, S. A. Trammell, L. E. Reddick, R. Kumar, F. Stellacci, N. Lebedev, J. Schnur, B. D. Bruce, S. Zhang and M. Baldo, *Nano Lett.*, 2004, **4**, 1079–1083.
2 B. Ge, F. Yang, D. Yu, S. Liu and H. Xu, *PLoS One*, 2010, **5**, e10233.
3 X. Zhao, Y. Nagai, P. J. Reeves, P. Kiley, H. G. Khorana and S. Zhang, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 17707–17712.
4 J. I. Yeh, S. Du, A. Tortajada, J. Paulo and S. Zhang, *Biochemistry*, 2005, **44**, 16912–16919.
5 P. Kiley, X. Zhao, M. Vaughn, M. A. Baldo, B. D. Bruce and S. Zhang, *PLoS Biol.*, 2005, **3**(7), e230.
6 K. Matsumoto, M. Vaughn, B. D. Bruce, S. Koutsopoulos and S. Zhang, *J. Phys. Chem. B*, 2009, **113**, 75–83.
7 Q. Meng, Y. Kou, X. Ma, Y. Liang, L. Guo, C. Ni and K. Liu, *Langmuir*, 2012, **28**, 5017–5022.
8 S. Santoso, W. Hwang, H. Hartman and S. Zhang, *Nano Lett.*, 2002, **2**, 687–691.
9 G. von Maltzahn, S. Vauthey, S. Santoso and S. Zhang, *Langmuir*, 2003, **19**, 4332–4337.
10 S. Vauthey, S. Santoso, H. Gong, N. Watson and S. Zhang, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 5355–5360.
11 S. A. Hollingsworth and R. O. Dror, *Neuron*, 2018, **99**, 1129–1143.
12 R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu and D. E. Shaw, *Annu. Rev. Biophys.*, 2012, **41**, 429–452.
13 J. E. Straub and D. Thirumalai, *Annu. Rev. Phys. Chem.*, 2011, **62**, 437–463.
14 L. E. Buchanan, E. B. Dunkelberger, H. Q. Tran, P.-N. Cheng, C.-C. Chiu, P. Cao, D. P. Raleigh, J. J. de Pablo, J. S. Nowick and M. T. Zanni, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 19285–19290.
15 C. A. Miller, S. H. Gellman, N. L. Abbott and J. J. de Pablo, *Biophys. J.*, 2009, **96**, 4349–4362.
16 W. M. Berhanu, E. J. Alred, N. A. Bernhardt and U. H. E. Hansmann, *Phys. Procedia*, 2015, **68**, 61–68.
17 O.-S. Lee, S. I. Stupp and G. C. Schatz, *J. Am. Chem. Soc.*, 2011, **133**, 3677–3683.
18 W. G. Noid, *J. Chem. Phys.*, 2013, **139**, 090901.
19 R. D. Groot and P. B. Warren, *J. Chem. Phys.*, 1997, **107**, 4423.
20 M. Dutt, O. Kuksenok, S. R. Little and A. C. Balazs, *Nanoscale*, 2011, **3**, 240–250.
21 M. Dutt, O. Kuksenok, M. J. Nayhouse, S. R. Little and A. C. Balazs, *ACS Nano*, 2011, **5**(6), 4769–4782.
22 F. Aydin, G. Uppaladadium and M. Dutt, *Colloids Surf., B*, 2015, **128**, 268–275.
23 E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero and N. F. A. van der Vegt, *Soft Matter*, 2013, **9**, 2108–2119.
24 C. Wu and J.-E. Shea, *Curr. Opin. Struct. Biol.*, 2011, **21**, 209–220.
25 R. Pellarin, E. Guarnera and A. Caflisch, *J. Mol. Biol.*, 2007, **374**, 917–924.
26 J. Zhang and M. Muthukumar, *J. Chem. Phys.*, 2009, **130**, 035102.
27 H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole and S. J. Marrink, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 225–248.
28 L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S.-J. Marrink, *J. Chem. Theory Comput.*, 2008, **4**, 819–834.
29 P. W. J. M. Frederix, R. v. Ulijn, N. T. Hunt and T. Tuttle, *J. Phys. Chem. Lett.*, 2011, **2**, 2380–2384.
30 O.-S. Lee, V. Cho and G. C. Schatz, *Nano Lett.*, 2012, **12**, 4907–4913.
31 M. Mazza, R. Notman, J. Anwar, A. Rodger, M. Hicks, G. Parkinson, D. McCarthy, T. Daviter, J. Moger, N. Garrett, T. Mead, M. Briggs, A. G. Schätzlein and I. F. Uchegbu, *ACS Nano*, 2013, **7**, 1016–1026.
32 N. Thota, Y. Ma and J. Jiang, *RSC Adv.*, 2014, **4**, 60741–60748.
33 S. Mushnoori, K. Schmidt, V. Nanda and M. Dutt, *Org. Biomol. Chem.*, 2018, **16**, 2499–2507.
34 Y. Sun, Z. Qian, C. Guo and G. Wei, *Biomacromolecules*, 2015, **16**, 2940–2949.
35 C. Guo, Y. Luo, R. Zhou and G. Wei, *Nanoscale*, 2014, **6**, 2800–2811.
36 P. W. J. M. Frederix, G. G. Scott, Y. M. Abul-Haija, D. Kalafatovic, C. G. Pappas, N. Javid, N. T. Hunt, R. V. Ulijn and T. Tuttle, *Nat. Chem.*, 2015, **7**, 30–37.

37 G. G. Scott, P. J. McKnight, T. Tuttle and R. V. Ulijn, *Adv. Mater.*, 2016, **28**, 1381–1386.
38 J. Kwon, M. Lee and S. Na, *J. Comput. Chem.*, 2016, **37**, 1839–1846.
39 C. Guo, Y. Luo, R. Zhou and G. Wei, *ACS Nano*, 2012, **6**, 3907–3918.
40 N. Thota, Z. Luo, Z. Hu and J. Jiang, *J. Phys. Chem. B*, 2013, **117**, 9690–9698.
41 H. D. Nguyen and C. K. Hall, *J. Am. Chem. Soc.*, 2006, **128**, 1890–1901.
42 H. D. Nguyen and C. K. Hall, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 16180–16185.
43 R. A. Mansbach and A. L. Ferguson, *Org. Biomol. Chem.*, 2017, **15**, 5484–5502.
44 R. A. Mansbach and A. L. Ferguson, *J. Phys. Chem. B*, 2017, **121**, 1684–1706.
45 F. Müller-Plathe, *ChemPhysChem*, 2002, **3**, 754–769.
46 D. Reith, M. Pütz and F. Müller-Plathe, *J. Comput. Chem.*, 2003, **24**, 1624–1636.
47 A. Villa, N. F. A. van der Vegt and C. Peter, *Phys. Chem. Chem. Phys.*, 2009, **11**, 2068.
48 A. Villa, C. Peter and N. F. A. van der Vegt, *Phys. Chem. Chem. Phys.*, 2009, **11**, 2077.
49 S. Izvekov and G. A. Voth, *J. Phys. Chem. B*, 2005, **109**, 2469–2473.
50 R. D. Hills, L. Lu and G. A. Voth, *PLoS Comput. Biol.*, 2010, **6**, e1000827.
51 A. P. Lyubartsev and A. Laaksonen, *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.*, 1995, **52**, 3730–3737.
52 M. S. Shell, *J. Chem. Phys.*, 2008, **129**, 144108.
53 S. P. Carmichael and M. S. Shell, *J. Phys. Chem. B*, 2012, **116**, 8383–8393.
54 O. Bezkorovaynaya, A. Lukyanov, K. Kremer and C. Peter, *J. Comput. Chem.*, 2012, **33**, 937–949.
55 T. Sanyal and M. S. Shell, *J. Chem. Phys.*, 2016, **145**, 034109.
56 T. Sanyal and M. S. Shell, *J. Phys. Chem. B*, 2018, **122**(21), 5678–5693.
57 A. Das and H. C. Andersen, *J. Chem. Phys.*, 2012, **36**, 194114.
58 L. Larini, L. Lu and G. A. Voth, *J. Chem. Phys.*, 2010, **132**, 164107.
59 V. Rühle and C. Junghans, *Macromol. Theory Simul.*, 2011, **20**, 472–477.
60 M. Praprotnik, L. Delle Site and K. Kremer, *J. Chem. Phys.*, 2005, **123**, 224106.
61 M. Praprotnik, L. Delle Site and K. Kremer, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2006, **73**, 066701.
62 M. Praprotnik, L. Delle Site and K. Kremer, *J. Chem. Phys.*, 2007, **126**, 134902.
63 J. Zhou, I. F. Thorpe, S. Izvekov and G. A. Voth, *Biophys. J.*, 2007, **92**, 4289–4303.
64 M. K. Baumann, M. Textor and E. Reimhult, *Langmuir*, 2008, **24**, 7645–7647.
65 I. W. Hamley, *Soft Matter*, 2011, **7**, 4122.
66 D. Reith, M. Pütz and F. Müller-Plathe, *J. Comput. Chem.*, 2003, **24**, 1624–1636.
67 C. Peter and K. Kremer, *Soft Matter*, 2009, **5**, 4357.
68 S. Izvekov, M. Parrinello, C. J. Burnham and G. A. Voth, *J. Chem. Phys.*, 2004, **120**, 10896–10913.
69 V. Rühle, C. Junghans, A. Lukyanov, K. Kremer and D. Andrienko, *J. Chem. Theory Comput.*, 2009, **5**, 3211–3223.
70 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins: Struct., Funct., Bioinf.*, 2006, **65**, 712–725.
71 H. J. C. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43–56.
72 E. Lindahl, B. Hess and D. van der Spoel, *J. Mol. Model.*, 2001, **7**, 306–317.
73 D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.
74 T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
75 G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.
76 G. Bussi, T. Zykova-Timan and M. Parrinello, *J. Chem. Phys.*, 2009, **130**, 074101.
77 M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.
78 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.
79 H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269–6271.
80 W. F. van Gunsteren and H. J. C. Berendsen, *Mol. Simul.*, 1988, **1**, 173–185.
81 T. A. Wassenaar, K. Pluhackova, R. A. Böckmann, S. J. Marrink and D. P. Tieleman, *J. Chem. Theory Comput.*, 2014, **10**, 676–690.
82 J. Peng, C. Yuan, R. Ma and Z. Zhang, *J. Chem. Theory Comput.*, 2019, **15**(5), 3344–3353.
83 S. J. Marrink and D. P. Tieleman, *Chem. Soc. Rev.*, 2013, **42**, 6801–6822.
84 D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt and K. Kremer, *Phys. Chem. Chem. Phys.*, 2011, **13**, 10412–10420.
85 M. Zhao, J. Sampath, S. Alamdari, G. Shen, C.-L. Chen, C. J. Mundy, J. Pfaendtner and A. L. Ferguson, *J. Phys. Chem. B*, 2020, **124**, 7745–7764.
86 C. Y. Lu, Master's thesis, Rutgers, The State University of New Jersey, 2020.
87 Hybrid CG Models Peptides, (accessed November 2021), DOI: 10.5281/zenodo.5715326.


 