Formalizing Coarse-Grained Representations of Anisotropic Interactions at Multimeric Protein Interfaces using Virtual Sites

Luc F. Christians[1], Ethan V. Halingstad[1], Emiel Kram[1], Evan M. Okolovitch[1], Alexander J. Pak[1,2,3,*]

[1] Department of Chemical and Biological Engineering, Colorado School of Mines, Golden, CO, 80401, USA

[2] Quantitative Biosciences and Engineering Program, Colorado School of Mines, Golden, CO, 80401, USA

[3] Materials Science Program, Colorado School of Mines, Golden, CO, 80401, USA

[*] Corresponding author: apak@mines.edu

Abstract

Molecular simulations of biomacromolecules that assemble into multimeric complexes remain a challenge due to computationally inaccessible length and time scales. Low-resolution and implicit-solvent coarse-grained modeling approaches using traditional nonbonded interactions (both pairwise and spherically isotropic) have been able to partially address this gap. However, these models may fail to capture the complex, anisotropic interactions present at macromolecular interfaces unless higher-order interaction potentials are incorporated at the expense of computational cost. In this work, we introduce an alternate and systematic approach to represent directional interactions at protein-protein interfaces using virtual sites restricted to pairwise interactions. We show that virtual site interaction parameters can be optimized within a relative entropy minimization framework using only information from known statistics between coarse-grained sites. We compare our virtual site models to traditional coarse-grained models using two case studies of multimeric protein assemblies and find that the virtual site models predict pairwise correlations with higher fidelity and more importantly, assembly behavior that is morphologically consistent with experiments. Our study underscores the importance of anisotropic interaction representations and paves the way for more accurate yet computationally efficient coarse-grained simulations of macromolecular assembly in future research.

Introduction

Molecular dynamics (MD) simulations have been widely used to investigate the relationships between molecular phenomena and macroscopic behavior, offering spatial or temporal resolutions that are difficult to probe experimentally. MD simulations at atomic resolution are most common, allowing high spatial (~Ångstroms) and temporal (~femtoseconds) resolution yet limited to length and time scales on the order of nanometers and microseconds.[1, 2] For the study of macromolecular systems involving dynamical behavior that requires longer (and possibly hierarchical) length and time scale dependence, such as commonly seen in biology and soft materials, coarse-grained (CG) modeling and simulation is an attractive alternative.[3-6]

One of the benefits of CG models is that of computational cost, as CG models represent sets of fine-grained (FG) particles (e.g., atoms) as pseudo-particles (i.e., CG sites), thereby reducing the complexity of modeled macromolecules and facilitating simulations of larger systems over longer times compared to those possible using atomistic models. The CG modeling process involves two steps: mapping and parameterization.[4-6] Mapping determines the correspondence between FG and CG particles while parameterization determines the effective interactions between CG particles. One strategy to determine CG mappings and parameters is to derive CG models that reproduce microscopic statistics from FG simulations, a strategy called the “bottom-up” approach. Constructing CG models in this way provides a direct means to first hypothesize then test the importance of molecular features (from CG mappings) and correlations (from CG interactions) on observed macroscopic behavior.

Within the field of bottom-up CG modeling, various systematic algorithms have been proposed for parameterization and, to a lesser extent, mapping. Mapping algorithms follow the convention of chemistry where complex molecules are broken down into clusters representing chemical moieties or functional groups. For biomolecules, mapping is often a linearly weighted average of local atoms, for example using the center-of-mass, which can be variationally optimized,[7] determined through graph theoretic methods,[8] or modified in the context of dynamic linear mappings.[9, 10] Parameterization algorithms have sought to derive CG models such that the sampled

distributions recapitulate the many-body configurational distribution of their FG counterparts mapped to the CG phase space.[11] Under this criterion, which is known as thermodynamic consistency, the ideal effective CG Hamiltonian is the CG-mapped many-body potential of mean force (PMF). Several methods have been proposed, including structure-matching methods[12-17] to iteratively capture correlation functions and variational methods[11, 18-24] that minimize least-squared differences in forces or the Kullback-Leiber (KL) divergence. However, using an arbitrarily complex basis set to describe the CG Hamiltonian is both challenging and impractical. Instead, prior studies have focused on simplified basis functions that recapitulate a reduced set of microscopic statistics. For instance, nonbonded interactions are traditionally represented using functions of pairwise distances, e.g., Lennard-Jones or Coulomb interactions, which have also been demonstrably successful in atomistic modeling.[25]

In recent years, CG models have moved toward increasingly lower resolutions to bridge increasingly larger length and time scales. Hence, while the pairwise approximation for the CG Hamiltonian may be acceptable for high-resolution CG models, i.e., those mapping around 4-to-1 heavy atoms per CG site, it is unlikely that this approximation will hold for low-resolution CG models. This problem is exacerbated by the fact that CG mappings often represent CG sites as spherically isotropic particles, as it is well-known that macromolecular interactions are typically anisotropic or highly specific.[26, 27] Recognizing this limitation, several solutions have been proposed. One solution is to include non-isotropic descriptors such as orientation vectors to delineate anisotropic interactions, e.g., via Gay-Berne potentials.[28-31] Another solution is to include higher-order interactions following the many-body expansion principle, such as analytical three-body potentials[32-36] or data-driven approximations of N-body potentials.[37-40] However, introducing non-isotropic or many-body interactions inevitably reduces the efficiency gains from CG modeling due to their increased complexity and computational cost.

An alternative approach to represent non-isotropic and/or higher-order CG interactions, while retaining the low computational cost of pairwise interactions, is to introduce “virtual” sites that may not explicitly represent sets of FG particles. Virtual sites can interact with real sites through pairwise interactions and aim to impart subtle

anisotropic projections of forces acting upon real sites. One prototypical example of this idea is the atomistic TIP4P water model.[41] Similar types of virtual sites have been used in the context of high-resolution CG models, notably for sterols and for aromatic hydrocarbons.[42, 43] Most recently, virtual sites have been implemented in low-resolution CG models to represent anisotropic interactions in lipids, biopolymers, and viral capsid proteins.[9, 44-53] While these studies demonstrate the viability of virtual sites as representations of directional interactions, a lack of systematic rules to derive their effective interactions limits their widespread adoption. To date, existing bottom-up methodologies have focused on dipole-dipole interactions through a center-of-symmetry framework or solvent-mediated interactions through representations of the hydration layer.[9, 43] Yet, a formal framework to define virtual site interactions in the context of biomacromolecular interfaces is still needed.

In this work, we present a systematic methodology to derive virtual site interactions with an emphasis on representations of anisotropic interactions for multimeric biomolecules. We then use our methodology to test the importance of anisotropic representations as compared to isotropic representations when modeling the self-assembly behavior of two different protein systems. The first protein is Q, which is an engineered α-helical protein that forms pentameric coiled-coils that further assemble into thermo-responsive nanofibers with upper critical solution temperature behavior.[54, 55] Coiled-coil proteins leverage both knob-in-hole interactions and amphiphilicity as driving forces for oligomerization.[56, 57] The second protein is the bacterial microcompartment (BMC) hexamer (H) shell protein BMC-H from *Haliangium ochraceum*[58, 59]; when expressed by itself, BMC-H assembles into hexameric nanosheets that roll into rosette-like shapes.[60, 61] For both systems, we report CG models using spherically-symmetric pairwise interactions and virtual site CG (VCG) models using our presented methodology. We then compare the structures, thermodynamics, and kinetics of protein assemblies using both models and argue that anisotropic representations of protein-protein interactions result in higher fidelity models that are essential for protein assembly studies.

Methods

Summary of computational methods. To derive each CG and VCG model, we first performed all-atom (AA) MD simulations to generate reference statistics. Next, we mapped the all-atom trajectories into VCG phase space and parameterized both CG and VCG models. Then, we compared the CG/VCG energetics to that of all-atom PMF calculations. Finally, we performed CGMD simulations of protein assembly using both CG and VCG models. The details of each step of the process are described below.

All-atom molecular dynamics simulations. We performed AAMD simulations of two axially aligned then stacked Q pentamers and two adjacent BMC-H hexamers. The atomic model for Q pentamers was approximated using CCBuilder2.0.[62] As the hierarchical structure of Q fibrils is unknown, we aligned Q pentamers based on complimentary charges at the two termini in an attempt to find a stable inter-coiled-coil interface. However, the two coiled-coils tended to separate, and we instead focused on intra-coiled-coil statistics. The atomic model for two BMC-H hexamers was isolated from a 3.5 Å resolution structure solved using X-ray crystallography[58, 61]. All structures were processed then simulated using GROMACS 2021[63] using parameters summarized in **Table S1**. First, a triclinic box was defined leaving a 2 nm buffer between the protein and domain edges; both systems utilized periodic boundary conditions. Next, each system was solvated in an aqueous solution with monovalent salt concentrations consistent with prior experiments (Q: 500 mM, BMC: 150 mM)[55, 61]. Proteins and ions were modeled using the CHARMM36m force field (February 2021)[64] and water was modeled using the TIP3P model[41]. Energy minimization was performed using steepest descent. Next, temperatures were equilibrated under the canonical (NVT) ensemble using the stochastic velocity rescaling thermostat[65] with parameters shown in **Table S1**. All heavy atoms were restrained using a harmonic force with spring constant 1000 kJ/mol/nm^2. Then, equilibration under the isobaric-isothermal (NPT) ensemble was performed using the stochastic velocity rescaling thermostat[65] and the Parrinello-Rahman barostat[66] using the parameters reported in **Table S1**. Finally, simulations were extended under the NVT ensemble using the stochastic velocity rescaling thermostat[65]

and the settings defined in **Table S1**. All simulations were run using a 2 fs timestep. Harmonic restraints to a single Cα in each monomer within one of the two supramolecular subunits present in each simulation (i.e., pentamer in Q and hexamer in BMC-H) were applied with spring constant 1000 kJ/mol/nm^2 to prevent drift. The final 300 ns and 415 ns of each Q and BMC-H trajectory, respectively, was used as reference data for CG modeling. Four independent replicates for each system were generated. Analysis of the root mean squared deviation of backbone α-carbons across the first 200 ns with respect to each atomic model shows that the CHARMM36m force-field maintains the atomic structure with minimal deviation and equilibrates within 50 ns (**Figure S1**).

All-atom umbrella sampling simulations. Umbrella sampling (US) simulations[67] were performed to compute the PMF associated with binding within a supramolecular subunit (i.e., intra-pentamer/hexamer) and between supramolecular subunits (i.e., inter-pentamer/hexamer). Each intra-subunit system was prepared by isolating two adjacent monomers within a supramolecular subunit while the inter-subunit systems used the same preparation as above. The first principal axis of each structure was aligned to a coordinate axis using VMD.[68] Each system was prepared using identical conditions to that of the AAMD simulations; exceptions and US simulation-specific parameters are reported in **Table S2**. After constant NVT followed by constant NPT equilibration, US windows were prepared by pulling and pushing monomers/subunits along the coordinate axis using center-of-mass distance as the metric; moving harmonic biases were applied under the constant NVT ensemble using GROMACS 2021[63] and PLUMED 2.7[69] with constants reported in **Table S2**. The final 25 and 30 ns of each Q and BMC-H window trajectory, respectively, was used for the final PMF calculation, which was performed using the weighted histogram analysis method (WHAM).[67, 70] For the Q system, a 2D PMF was initially computed with a final 1D PMF representing the minimum free energy path determined using the string method (**Figure S2**).[71] Histograms for each window are shown in **Figure S3**.

Integrating coarse-grained models with virtual sites. Following prior low-resolution CG models,[3, 72] we define the CG Hamiltonian in four parts:

$$U(R^N) \approx \sum_{i=1}^{N} \sum_{j=i+1}^{N} U_{bonded}(r_{ij}) + U_{coul}(r_{ij}) + U_{excl}(r_{ij}) + U_{attr}(r_{ij}) \quad (1)$$

which includes bonded (U_{bonded}), Coulombic (U_{coul}), excluded volume (U_{excl}), and attractive (U_{attr}) contributions based on pair distance r_{ij} between CG sites i and j across all N sites. The U_{bonded} term represents all intramolecular interactions while the latter three terms (U_{coul}, U_{excl}, and U_{attr}) represent intermolecular interactions due to electrostatics, sterics, and close contacts, respectively. Prior CG models[43, 73] have defined virtual sites as centers for binding interactions between two real CG sites that form a close contact (**Figure 1**), and we follow that same convention here. As such, virtual sites only contribute to the U_{attr} term. As shown in **Figure 1**, the virtual site v is bonded to CG site i and interacts with CG site j using a nonbonded potential that favors the overlap between the virtual site and CG site j. Here, we define the bonded interaction using a harmonic potential and the nonbonded interaction using a Gaussian potential such that U_{attr} can be approximated by two contributions:

$$U_{attr}(r_{ij}) = U_{VCG}(r_{iv}, r_{vj}) = K_{iv}(r_{iv} - r_{iv,0})^2 - A_{vj} \exp(-B_{vj}(r_{vj})^2) \quad (2)$$

where A_{vj} and B_{vj} are Gaussian parameters, K_{iv} and $r_{iv,0}$ are harmonic parameters, and r_{iv} (or r_{vj}) is the pair distance between CG sites i and v (or v and j).

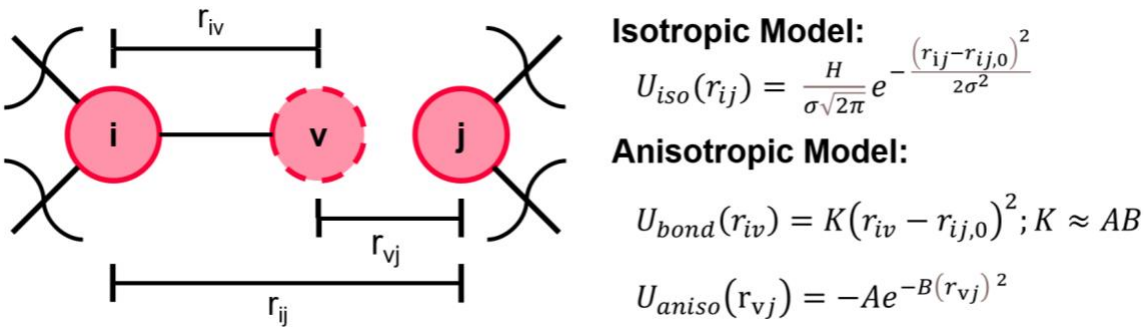


Figure 1. Coarse-grained modeling framework using virtual sites to represent anisotropic interactions. Schematic of a virtual CG site v (dashed red circle) serving as a directional binding interaction between real CG sites i and j (red circles) at the interface between two macromolecules. The virtual site is bonded to CG site j via a harmonic potential and interacts with CG site j via a Gaussian potential; the two potentials are coupled. In the equivalent isotropic model, CG sites i and j interact directly through a Gaussian potential.

To optimize the unknown interaction parameters, we use the relative entropy minimization (REM)[74] method, which aims to minimize the KL divergence:

$$S_{rel} = \int P_{AA}(R^N) \ln\left(\frac{P_{AA}(R^N)}{P_{CG}(R^N)}\right) dR^N \tag{3}$$

where S_{rel} is the relative entropy, P_{AA} is the configurational probability density in the atomistic ensemble, P_{CG} is the configurational probability density in the CG ensemble, and R^N is the CG configuration which can also be determined by mapping from atomic configurations (i.e., $M(r^n)$). Under the constant NVT ensemble, Eq. 3 can be reformulated as:

$$S_{rel} = \beta\langle U_{CG} - U_{AA}\rangle_{AA} - \beta(A_{CG} - A_{AA}) \tag{4}$$

where $U_{CG/AA}$ is the CG/AA internal energy and $A_{CG/AA}$ is the configurational part of the CG/AA Helmholtz free energy, which in itself is a function of $U_{CG/AA}$. One can numerically minimize S_{rel} with respect to the parameters λ that define U using gradient descent:

$$\lambda^{t+1} = \lambda^{t} - \chi\beta\left(\left\langle\frac{\partial U}{\partial\lambda}\right\rangle_{AA} - \left\langle\frac{\partial U}{\partial\lambda}\right\rangle_{CG}\right) + \eta(0,s) \tag{5}$$

where $\beta = (k_B T)^{-1}$, t is the iteration number, χ is the learning rate, and $\langle\,\rangle_{AA}$ and $\langle\,\rangle_{CG}$ are the averages over the respective AA and CG ensembles. Note that previous implementations used the Newton-Raphson method[74] but we chose to use an alternative method due to the nonconvex nature of our parameter space. Taking inspiration from perturbed gradient descent,[75] we add random noise sampled from a Gaussian distribution $\eta(0,s)$ to help escape from saddle points. As U is approximated with a pairwise basis according to Eq. 1, we can evaluate the ensemble-averaged quantities as follows:

$$\left\langle\left(\frac{\partial U}{\partial\lambda}\right)\right\rangle_{AA/CG} = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(\int\frac{\partial U(r_{ij})}{\partial\lambda}P_{AA/CG}(r_{ij})\,dr_{ij}\right) \tag{6}$$

Up to this point, we have described the REM method in a general sense. However, Eq. 6 is not readily applicable to the virtual site interactions defined in Eq. 2 as no explicit statistics for $P_{AA}(r_{iv})$ or $P_{AA}(r_{vj})$ exist. Instead, consider the fact that under an idealized binding state induced by the virtual site, we expect $P(r_{iv}) \approx P(r_{ij})$ and $P(r_{vj}) \approx P(r_{iv} - r_{iv,0})$. Then, it is possible to iteratively optimize the virtual site interaction using $P(r_{ij})$ statistics if the harmonic and Gaussian contributions to Eq. 2 are coupled to each other. The Gaussian interaction term in Eq. 2 can be expanded around $r_{vj} = 0$ as an infinite power series:

$$-A_{vj}\exp\left(-B_{vj}\left(r_{vj}\right)^{2}\right) \approx -A_{vj}\sum_{k=0}^{\infty}\frac{\left(-B_{vj}\left(r_{vj}\right)^{2}\right)^{k}}{k!} \tag{7}$$

Taking Eq. 7 to first-order ($k = 1$) and relating the Gaussian force to the harmonic force implied by Eq. 2 through the approximation $P(r_{vj}) \approx P(r_{iv} - r_{iv,0})$ shows that:

$$2K_{iv}(r_{iv} - r_{iv,0}) \approx 2A_{vj}B_{vj}(r_{iv} - r_{iv,0}) \quad (8)$$

Or more simply:

$$K_{iv} \approx A_{vj}B_{vj} \quad (9)$$

which shows that the spring constant of the virtual site bond can be related to the parameters of the virtual site binding interaction, which in turn are related to r_{ij} statistics through the approximation $r_{ij} \approx r_{iv}$. We note that we also tried to optimize K_{iv}, A_{vj}, and B_{vj} independently but our attempts always led to the divergence of parameters K_{iv} and A_{vj}, which we attribute to the fact that the equations for K_{iv} and A_{vj} emphasize the longer-range and shorter-range regions of the probability distribution, respectively. To summarize, we propose that the virtual site interaction in Eq. 2 can be coupled to real CG site statistics in the following manner:

1. Approximate $P(r_{vj})$ as $P(r_{ij} - r_{iv,0})$
2. Apply Eq. 5 to iteratively update A_{vj} and B_{vj}
3. Apply Eq. 9 to iteratively update K_{iv}

In comparison, the isotropic analog of Eq. 2 can be defined as:

$$U_{attr}(r_{ij}) = U_{CG}(r_{ij}) = \frac{H_{ij}}{\sigma_{ij}\sqrt{2\pi}} \exp\left(-\frac{(r_{ij}-r_{ij,0})^2}{2\sigma_{ij}^2}\right) \quad (10)$$

where H_{ij} and σ_{ij} are the Gaussian parameters and $r_{ij,0} = r_{iv,0}$. Hence, Eq. 5 can be directly applied to iteratively update H_{ij} and σ_{ij} without approximation.

The remaining terms in Eq. 2 can be defined as follows. We represent the intramolecular interactions through a harmonic bond network with each bond represented by:

$$U_{bonded}(r_{ij}) = K_{ij}(r_{ij} - r_{ij,0})^2 \quad (11)$$

where K_{ij} is the spring constant and $r_{ij,0}$ is the minimum energy distance. Electrostatic interactions were defined using the Yukawa potential:

$$U_{coul}(r_{ij}) = \frac{Cq_iq_j}{\epsilon_r r_{ij}} \exp(-\kappa r_{ij}) \quad (12)$$

where C is a unit conversion constant, q is the partial charge, ϵ_r is the dielectric constant, and κ is the inverse Debye screening length. Excluded volume repulsions were defined as:

$$U_{excl}(r_{ij}) = D_{ij}\left[1 + \cos\left(\frac{\pi r_{ij}}{r_{c,ij}}\right)\right] \text{ for } r_{ij} < r_c \quad (13)$$

where D_{ij} is the amplitude and $r_{c,ij}$ is the cutoff distance.

Coarse-grained mapping and parameterization. CG sites were mapped to Cα positions in the AA reference data with solvent integrated out. The final CG models were mapped using the essential dynamics coarse-graining method.[76] Virtual sites were defined based on protein-protein pair distance distributions. The proximity and height of peaks in the pair distributions were used to quantify the likelihood of a close contact; sharp peaks with heights greater than 0.004 and widths smaller than 1.0 Å were selected to limit the number of virtual sites (see **Table S3** and **Table S4** for complete list). These constraints were adjusted to ensure each interface was represented. The first frame of the CG-mapped atomistic trajectory was used to initialize the isotropic CG model. The VCG model was initialized by adding virtual sites to the position of CG site j (see **Figure 1**) for each close contact pair. The masses and charges for CG sites were computed using the sum of all residues used to define the CG site. Virtual sites were given no charge with masses set to their real site counterpart. Intramolecular bonds were parameterized using a heteroelastic network model (HENM)[77] with a cutoff of 15 Å

and 15.5 Å for the Q and BMC-H systems, respectively. For close contact pairs, $r_{ij,0}$ in Eq. 2 and Eq. 10 was set to the position of the peak in the pair distribution. For all intermolecular pairs, $r_{c,ij}$ for steric interactions in Eq. 13 was determined using the onset distance for non-zero density in the pair distribution up to a max $r_{c,ij}$ of 12 Å. The κ in Eq. 12 was determined using experimental monovalent salt concentrations and temperatures.

For Q, we investigated additional virtual sites (denoted as HP) that represent the steric interactions of hydrophobic side chains populating the interior of the coiled-coil pore. The positions of these HP sites were based on CG sites of A27, L34, L41, and L48, which are the hydrophobic residues in each monomer that contribute to the buried hydrophobic core. All HP sites were positioned 45% of the distance along the vector connecting each CG site to the center of mass of all five corresponding CG sites. Parameters for U_{bonded} and U_{excl} were defined via Boltzmann Inversion based on CG-mapped distributions of the HP sites; bonds between real and HP sites were defined for mean distances less than 10 Å.

To perform REM optimization, CG simulations were run at each iteration using LAMMPS[78] (2 Jun 2022) and Moltemplate[79] under constant NVT using the Langevin thermostat.[80] Details on the integration timestep, equilibration time, production time, and thermostat settings are provided in **Table S5**. As convexity is not guaranteed, multiple initial guesses for all parameters were tested. After each CG simulation (at each iteration), Eq. 5 was applied to determine the next iteration of parameters. For the CG model, parameters H_{ij} and σ_{ij} of Eq. 10 were iterated using REM. For the VCG model, parameters A_{vj} and B_{vj} of Eq. 2 were iterated using REM and K_{iv} was solved using Eq. 9. A learning rate schedule (see **Table S6-9**) was applied to gradually decrease the learning rate while the change in each parameter was capped (see **Table S6-9**) to prevent large changes; both strategies helped to smooth convergence. A total of 300 iterations for each trial set of parameters were performed. The mean squared error (MSE) was computed at each iteration as:

$$MSE = \sum_I \sum_r \left(P_{AA,I}(r) - P_{CG,I}(r) \right)^2 \quad (14)$$

where I is the index over all close contact pairs. Models yielding the lowest MSE were selected as the final CG/VCG models. All derivatives used in Eq. 6 are shown in **Table S10**. Final CG/VCG model parameters are shown in **Tables S11-13**.

For protein Q, we quantified three structural characteristics of the coiled-coil. First, we computed the root mean squared deviation (RMSD) compared to the atomic model using:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(|X_i - Y_i|)^2} \quad (15)$$

where X_i represents the CG-mapped positions of site i in the atomistic dataset, Y_i represents the positions in the CG dataset, and N represents the number of CG sites. To assess the symmetry, two metrics were used: a metric to quantify pentameric symmetry (ψ_5^{2D}) and the standard deviation of the distance from the centroid ($\sigma_{centroid}$) for each ring of residues (between residues 18 to 48) within the pentamer. ψ_5^{2D} was calculated using:

$$\psi_5^{2D} = \frac{1}{5}\sum_{i=0}^{4}\cos\left(5\,arccos\left(\frac{\vec{r_i}\cdot\overrightarrow{r_{i+1}}}{|\vec{r_i}||\overrightarrow{r_{i+1}}|}\right)\right) \quad (16)$$

where $\vec{r_i}$ represents the vector from the centroid to a CG site i within the ring of residues; $\psi_5^{2D} = 1$ indicates perfect five-fold symmetry and decreases toward zero with decreasing symmetry. $\sigma_{centroid}$was calculated using:

$$\sigma_{centroid} = \sqrt{\frac{1}{5NM}\sum_{k=0}^{M}\sum_{j=0}^{N}\sum_{i=0}^{4}(|r_{ic}| - \langle|r_{ic}|\rangle)^2} \quad (17)$$

where M represents the number of rings, N is the number of frames, $|r_{ic}|$ is the distance between the residue and its corresponding ring centroid, and $\langle|r_{ic}|\rangle$ is the mean distance; larger $\sigma_{centroid}$ corresponds to less uniformity with $\sigma_{centroid} = 0$ indicative of perfect uniformity.

Coarse-grained molecular dynamics simulations and analysis. Optimized CG and VCG models were analyzed by running assembly and pulling CGMD simulations using LAMMPS[78] (2 Jun 2022). Assembly simulations were set up with evenly dispersed monomers to achieve an initial monomer concentration of ~10-fold experimental concentrations to maintain monomer concentrations at or above experimental

conditions while monomers are depleted during assembly and to accelerate assembly kinetics (see **Table S14** for complete details). To simulate the hierarchical assembly of BMC-H with a fixed solution-state ratio of assembly-competent to assembly-incompetent monomer populations, BMC-H monomers were switched between active and inactive states based on a prior algorithm used in the study of viral protein assembly.[48, 51] In the inactive state, the U_{coul} and U_{attr} potentials were turned off. A constant solution-state monomers concentration of 4 mM (~2-fold experimental concentrations[60, 81]) was maintained to promote protein multimerization. In contrast, our Q simulations focused on the assembly of coiled-coil oligomers. Thus, instead of controlling concentration throughout the course of the simulation, Q simulations were initialized at 10 mM and allowed to equilibrate. Simulations were run under constant NVT using the Langevin thermostat[80] with details provided in **Table S14**. To determine the extent of assembly, supramolecular structures were modeled as graphs using NetworkX 3.1[82] where each monomer was treated as a node and close contacts (defined in **Table S15**) as edges. The size of each subgraph was used to represent the size of each assembled supramolecular structure. The shape of each assembled structure was visualized using VMD.[68]

To approximate the CG binding strengths within a supramolecular subunit and between subunits, pulling CGMD simulations were performed between two Q/BMC-H monomers and two BMC-H hexamers, respectively. Harmonic restraints were applied to a single CG site within one of the monomers or subunits present in each simulation. The harmonic restraint was shifted periodically, separating the two groups throughout the simulation (see **Table S16** for details). After each shift, simulations were run under constant NVT using the Langevin thermostat[80] for the total time shown in **Table S16**. Potential energies were saved every 5,000 steps and the final 30 ns of data was averaged to approximate the CG binding energies that were compared to AA PMFs.

Results and Discussion

Coarse-Grained Model of Protein Q. Protein Q is a pentameric coiled-coil protein capable of assembling into hydrogels composed of physically crosslinked fibrils.[54, 55]

The assembly is hierarchical with α-helical monomers of Q forming coiled-coil oligomers followed by fibrillization. In general, coiled-coils are stabilized by both their amphiphilic nature and knob-in-hole interactions between adjacent monomers.[57] Fibrillization, on the other hand, has been attributed to electrostatic interactions between complimentary charges at the N- and C-termini, a characteristic known as sticky ends.[83] In the case of Q, fibrils are 2.5-20 nm in diameter while gelation may take days to complete. [54, 55] Our interest in Q is motivated by the different types of interactions involved in Q assembly, making Q an attractive case study to compare CG models represented by anisotropic or isotropic interactions. As an atomic model for the inter-coiled-coil interface is unavailable at present, we focused our study in intra-coiled-coil interactions.

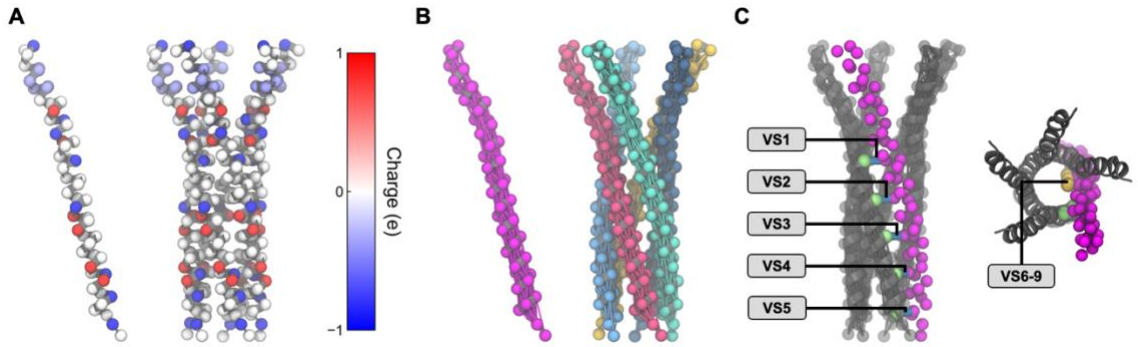


Figure 2. The coarse-grained (CG) model for Q. (A) The charge profile of the CG monomer (left) and pentamer (right) with CG sites colored by charge. (B) The bonds defined by the HENM using a 15 Å cutoff. (C) The VS pairs are represented by the blue bond between the virtual sites (green) and the CG sites (magenta). The N-terminal view shows the virtual sites (yellow) used to represent interactions in the inner hydrophobic region.

Our proposed VCG model for Q represents the knob-in-hole and amphiphilic interactions of Q using virtual sites. Monomers were modeled using a 1 CG site per residue resolution (**Figures 2A-B**). As shown in **Figure 2C**, virtual sites 1 to 5 (VS1-5) represent attractive, directional knob-in-hole interactions while virtual sites 6 to 9 (VS6-9) represent steric hindrance by hydrophobic side chains in the coiled-coil pore interior, which we modeled as purely repulsive. In total, three models were parameterized: an isotropic model with no virtual sites (CG model), a model with only VS1-5 (VCG model), and one with VS1-5 and VS6-9 (VCG+HP model). Both virtual site models (VCG and

VCG+HP) had VS1-5 trained using REM while the VS6-9 interactions were trained via Boltzmann Inversion independently from VS1-5; the presence or absence of VS6-9 interactions had negligible impact on VS1-5 training (**Figure S4**).

During REM, CG statistics generated from CG simulations using the trial model in each iteration are compared to CG mapped statistics from the atomistic dataset to update the U_{attr} parameters until those that minimize the loss are found. The values of these parameters are shown in **Figure 3A** for the CG model and **Figure 3B** for the VCG model. **Figure 3A** shows that the explored parameter space for VS1 and VS5 in the CG model was more varied compared to VS2-4, which is likely a result of VS1 and VS5 being in more flexible regions of the protein. **Figure 3B** shows a similar trend for the VCG model although exploratory behavior was only evident in VS5. Both the VCG and CG models predicted weaker interactions in VS1 and VS5 compared to VS2-4. Comparison of the MSE loss depicted in **Figure 3C** shows that the VCG model was able to achieve a lower aggregate MSE loss ($2.51 \times 10^{-4} \pm 6.40 \times 10^{-7}$) compared to that of the CG model ($2.60 \times 10^{-4} \pm 4.18 \times 10^{-7}$), indicating that the VCG model achieved higher fidelity to the reference atomistic data. To further analyze differences in fidelity, we investigated differences in structural correlations. **Figure 3D** depicts the predicted structures of the coiled-coil with the optimized VCG and CG model parameters. It is evident that the VCG model is able to maintain the expected cylindrical shape and pentameric symmetry of the atomistic coiled-coil better than the CG model (i.e., the pentamer in the latter is oblongated). In particular, the VCG model had a lower RMSD of 0.123±0.014 nm compared to the CG model (0.132±0.020 nm) while a higher ψ_5^{2D} (VCG: 0.923±0.092 and CG: 0.908±0.121) and a lower $\sigma_{centroid}$ (VCG: 0.172 nm and CG: 0.316 nm) jointly indicate that the VCG model is more uniform and symmetric than the CG model. Nonetheless, only minute differences were observed in the VS pair distributions shown in **Figure 3E**. Both the VCG and CG models were able to capture the first peak in the reference AA distribution and qualitatively match the peaks at longer distances. The similarity in the VCG and CG pair distributions suggests that pairwise correlations alone are likely insufficient to distinguish higher-order correlations, such as the pentameric symmetry of the Q oligomer. However, as the VCG model is fit to pairwise correlations through projection by a directional virtual site interaction, it

appears that this higher-order correlation can be implicitly preserved. Finally, we note that simulations of the optimized CG model ran at a rate of ~6180 timesteps/s while the VCG model ~4619 timesteps/s, indicating that the VCG model is only 25% slower than the CG model.

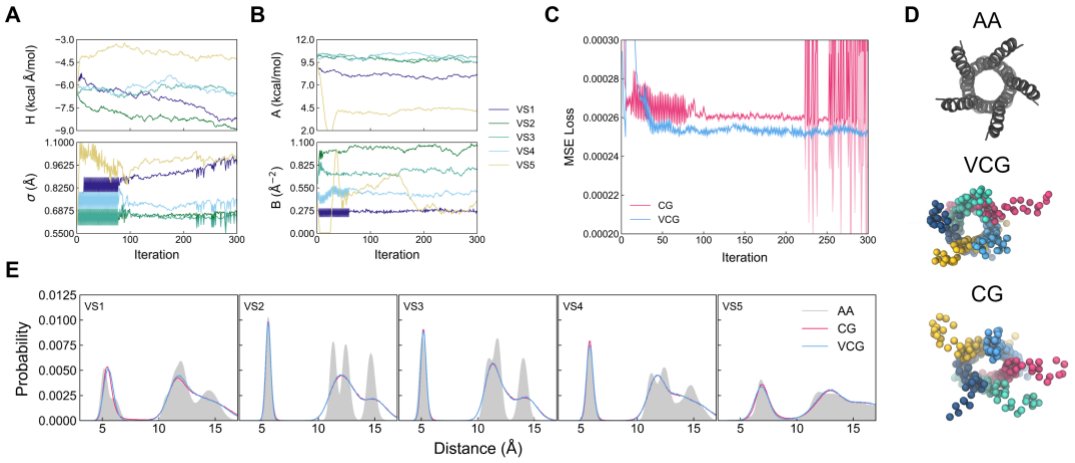


Figure 3. Coarse-grained (CG) model optimization for Q. Each U_{attr} parameter across iterations is shown for the (A) CG and (B) VCG models. (C) Comparison of the mean-squared error (MSE) of the CG and VCG models across iterations with standard deviations across four replicates shown as a shaded region. (D) Comparison of Q structures predicted by each optimized model based on the all-atom (AA) model of Q; each monomer in the CG snapshots is depicted by different colors. (E) Comparison of pair distance probability distributions between the real CG sites associated with each virtual site after model optimization. The data presented in (D) and (E) are from VCG iteration 243 and CG iteration 153.

Q Assembly into Coiled-Coil Oligomers. While the CG and VCG models of Q were able to maintain a pentameric coiled-coil structure and adhere to atomistic statistics, it would be informative to investigate the ability of such structures to form spontaneously. Here, we discuss the VCG+HP model (i.e., inclusion of VS6-9 as representations of hydrophobic core sterics) while results for the VCG model are shown in **Figure S5**. We first quantified the binding free energy between Q monomers using US simulations at atomic resolution and compared the resulting PMF to CG-derived energetics. As seen in **Figure 4A**, the monomer-monomer binding energy of the VCG+HP model (-26.33±0.06

kcal/mol) better recapitulates the atomistic binding free energy (-22.11 kcal/mol) compared to that of the CG model (-12.04±0.07 kcal/mol). The VCG+HP PMF also exhibits features of the AA PMF that are absent in the CG PMF, namely the presence of a metastable state at 10.0 Å and long-range attraction that begins as far as 19.5 Å. Overall, our free energy calculations show that the CG model predicts weaker and shorter-ranged attraction compared to the VCG+HP model, which further suggests that the kinetics of assembly and the stability of resulting oligomers may be slower and lower, respectively, for the CG model compared to the VCG+HP model. To test this hypothesis, we next performed CGMD assembly simulations.

Each CGMD assembly simulation started from configurations with evenly dispersed Q monomers at 10 mM concentration. **Figure 4B** shows the aggregate number of oligomerized Q over time. Interestingly, the CG model exhibited a faster initial rate of oligomerization compared to the VCG+HP model, although the maximum degree of oligomerization plateaued and was exceeded by the VCG+HP model. Despite the weaker binding energy shown in **Figure 4A**, the CG model showed faster assembly kinetics at lower time steps compared to the VCG+HP model, which we attribute to less stringent orientation requirements upon collision in the CG model case; at close separation distances, the VCG+HP model requires the two monomers be oriented according to the directionality of the virtual sites for successful binding. However, as larger multimers assemble at longer times, we speculate that the sterics imposed by the larger multimers implicitly enforce orientationally dependent association, which leads to slower assembly rates for both the VCG+HP and CG models. It is also possible that the additional mass of the virtual sites, which increases the mass of each Q monomer by ~16%, may contribute to slower assembly kinetics. However, when we tested the VCG+HP model with repartitioned masses such that the total mass is equivalent to the CG model, no noticeable differences in assembly rates or multimer assembly trends were observed (**Figure S6**).

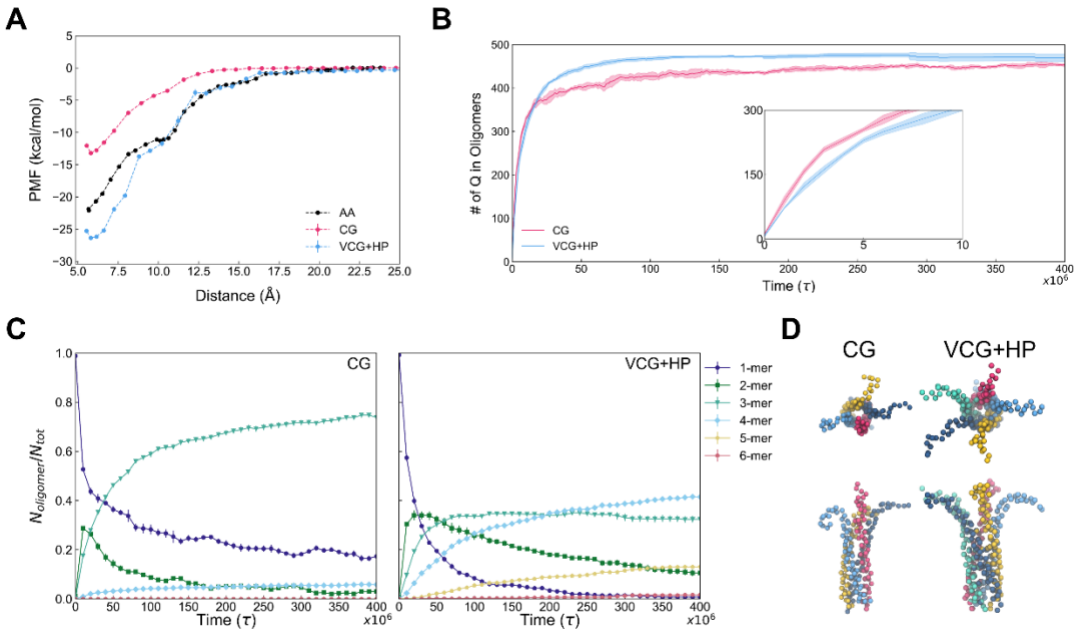


Figure 4. Q self-assembly into coiled-coil pentamers. (A) Comparison of the potential of mean force (PMF) as a function of distance between Q monomers for the CG, VCG+HP, and AA models. (B) The number of monomers in oligomers (of size 2 or more) over time for the CG and VCG+HP models; the inset shows the early stages of assembly. (C) The fraction of monomers classified by oligomer size to total monomers over time; classes ranging from monomers (circles) to hexamers (hexagons) for the CG (left) and VCG+HP (right) models are depicted. (D) Representative snapshots depicting the assembled coiled-coils predicted by the CG and VCG+HP assembly simulations.

While both models assembled into a variety of oligomers, not all assembled structures reflected the expected pentameric coiled-coil structure of Q. **Figure 4C** shows that the VCG+HP model was able to form pentamers with negligible higher-order oligomers formation (i.e., hexamers and above). Both tetramers and trimers were assembled as well, but we expect these to persist due to the depleting number of available monomers and dimers in the simulation domain. In contrast, the CG model tended to favor kinetically trapped “closed” trimer and tetramer structures that precluded recruitment of additional monomers or dimers to form pentamers. Analysis of pair distance distributions after assembly (**Figure S7**) reveals that both the CG and VCG+HP models capture the first peak of each distribution, although the VCG+HP is able to better replicate the atomistic distributions overall. The ability for the VCG+HP

model to form larger oligomers is driven by both the VS1-5 and VS6-9 interactions. The directional nature of both VS1-5 attraction and VS6-9 repulsion likely increased the preference for oligomers with pentameric symmetry; as shown in **Figure S5**, VS6-9 is not required for pentamer formation but instead accelerates their formation. Importantly, the VCG+HP model predicts the spontaneous formation of pentameric coiled-coils consistent with the AA model, as seen in **Figure 4D**. The CG model, on the other hand, was only able to predict tetrameric coiled-coils at most (see **Figure 4D**). It may be possible for the CG model to form pentamers beyond our simulated timescales. However, such an event is likely to be rare due to the high selectivity for kinetically trapped trimers and tetramers.

Overall, the VCG+HP model was able to assemble into experimentally consistent pentameric coiled-coils while the CG model could not. Our results demonstrate that the incorporation of virtual sites improves the fidelity of coiled-coil assembly not only by recapitulating the CG binding energy but also by introducing an orientation-dependent "entropic" barrier. In addition, while the presence of VS1-5 alone is sufficient for spontaneous pentamer assembly (see **Figure S5**), the rate and selectivity of pentamer formation is improved with the aid of VS6-9, demonstrating that balancing attractive (VS1-5) and repulsive (VS6-9) interactions through virtual sites is a viable CG modeling approach.

Coarse-Grained Model of BMC-H. BMCs are proteinaceous organelles that allow various bacterial species to thrive in diverse environments.[84] This role is done through the selective compartmentalization of enzymes and the controlled permeability of metabolites and toxic intermediates through the protein shell, which allows important biochemical processes to happen without interference or with high activity.[84] The protein shell of BMCs, which bear similarity to viral capsids, are composed of both hexameric (BMC-H or BMC-T) and pentameric (BMC-P) oligomers that tile into polyhedra.[85, 86] However, these polyhedral structures differ from viral capsids in size and composition, often irregular in shape and stoichiometry. Assembling different shell proteins into various forms, including icosahedral shells, tubules, and "swiss-roll" structures, suggests that the morphology and size of these constructs are highly tunable following

currently unknown mechanisms.[60, 61, 87] Here, we focus on BMC-H shell proteins from *Haliangium ochraceum*[58, 59] that spontaneously assemble into hexameric nanosheets that roll into a rosette shape.[60, 61] Our interest in BMC-H is motivated by the different hierarchies of oligomerization involved in BMC-H assembly, making BMC-H an attractive case study to compare CG models represented by anisotropic or isotropic interactions.

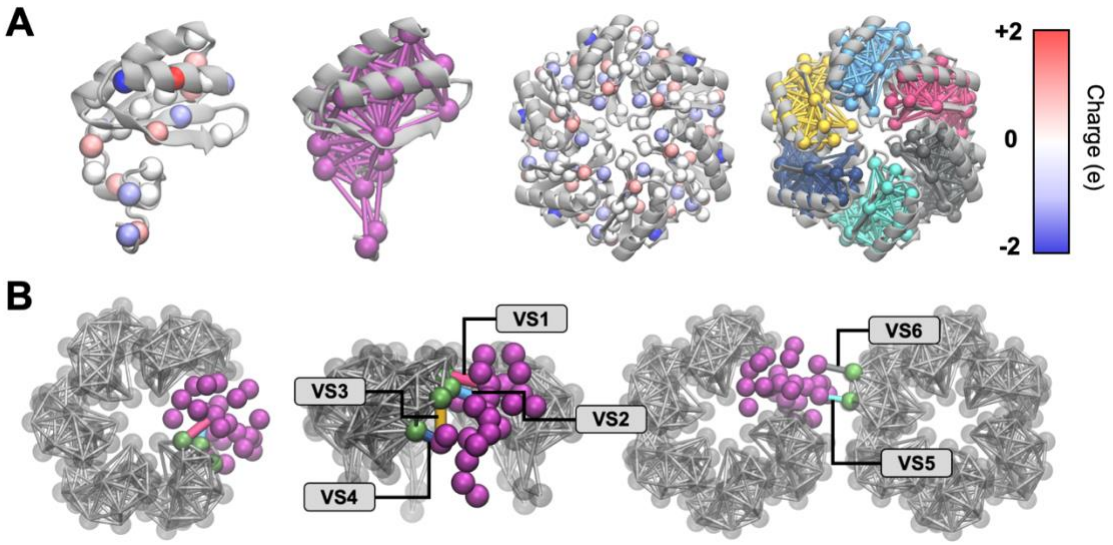


Figure 5. The coarse-grained (CG) model for BMC-H. (A) The charge profile (left-most) and bonds defined by the HENM (middle left) of the CG monomer and the equivalent for the hexamer (middle right and right-most). (B) The virtual site (VS) pairs are represented by the bond between the virtual sites (green) and the CG sites (magenta). The two left snapshots depict the intra-hexameric VS pairs while the two right snapshots depict the inter-hexameric VS pairs.

Our proposed VCG model for BMC-H consists of the 99 residue BMC-H monomer mapped to 26 CG sites with a net charge between ±2 and intra-protein fluctuations represented by an HENM bond network, as shown in **Figure 5A**. We modeled the intra-hexameric protein-protein contacts as VS1-4 and the inter-hexameric protein-protein contacts as VS5-6 (see **Figure 5B**). The same pairs of CG sites projected onto the virtual sites were used to create a CG model represented by

spherically isotropic interactions. In total, two models were fit: the VCG and CG models for BMC-H.

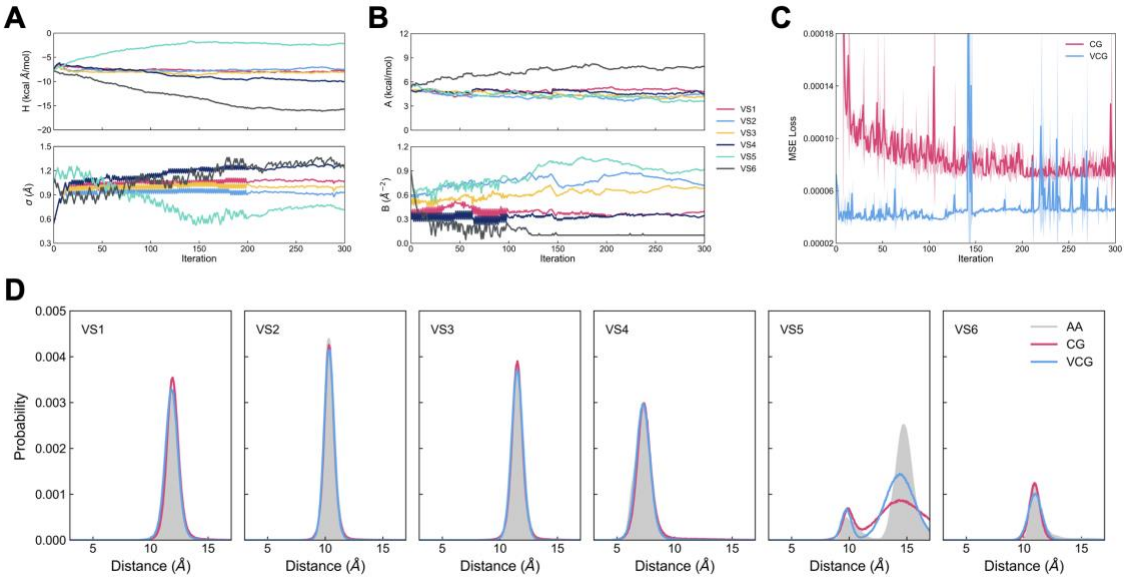


Figure 6. Coarse-grained (CG) model optimization for BMC-H. Each U_{attr} parameter across iterations is shown for the (A) CG and (B) VCG models. (C) Comparison of the mean-squared error (MSE) of the CG and VCG models across iterations with standard deviations across four replicates shown as a shaded region. (D) Comparison of pair distance probability distributions between the real CG sites associated with each virtual site after model optimization. The data presented are from VCG iteration 48 and CG iteration 205.

The values of the CG and VCG model parameters across REM iterations are shown in **Figure 6A** and **Figure 6B**, respectively. During both CG and VCG model optimization, we find that the intra-hexameric interaction parameters for VS1-3 find a minima and tend to remain stationary. However, VS5-6, and to a lesser extent, VS4, tend to explore a larger parameter space. As seen in **Figure 6C**, the VCG model initially explores a part of parameter space yielding an MSE loss as low as $3.69 \times 10^{-5} \pm 2.78 \times 10^{-7}$ but then enters a solution space with an MSE loss as low as $4.04 \times 10^{-5} \pm 1.75 \times 10^{-6}$, which highlights the complexity of this nonconvex optimization. Nonetheless, the VCG model is able to achieve a lower aggregate MSE loss compared to that of the CG model ($7.00 \times 10^{-5} \pm 3.99 \times 10^{-7}$), similar to our observations for the Q system. From the pair

distance distributions shown in **Figure 6D**, we find that both the CG and VCG models predict intra-hexameric distributions (i.e., VS1-4) that are consistent with the reference AA distributions with only minor differences observed between the CG and VCG models. More evident differences are observed in the inter-hexameric distributions (VS5-6) where both the CG and VCG models yield sharper distributions in the first peak for VS5 while the VCG model yields a broader distribution in the first peak for VS6. The mismatch between the AA and CG/VCG distributions for VS5-6 suggests that approximating the inter-hexameric interactions as Gaussians may be insufficient; in the future, other functional forms for pair interaction potentials, such as Lennard Jones or Morse, may be worthwhile to explore. However, no qualitative differences in BMC-H hexameric structure were observed in the case of both models compared to the AA model. While using a more complex basis may lead to quantitative improvement in accuracy with respect to pair distributions, the additional complexity and associated cost is not likely to be necessary. Finally, we note that simulations of the optimized CG model ran at a rate of ~7010 timesteps/s while that of the VCG model ran at ~6761 timesteps/s, indicating that the VCG model is only 4% slower than the CG model.

BMC-H Assembly into Hexameric Sheets. After optimization of both the CG and VCG models for BMC-H, we assessed the binding free energies at both the intra-hexameric and inter-hexameric interfaces using US simulations at atomic resolution and compared the resulting PMFs to CG-derived energetics. As seen in **Figure 7A**, the monomer-monomer binding energy of the VCG model (-14.18±0.08 kcal/mol) and of the CG model (-10.90±0.04 kcal/mol) both underestimate the atomistic binding free energy (-18.73±0.21 kcal/mol), although the VCG model is closer. However, the CG PMF exhibits a longer-range attraction at around 3.5 nm that is consistent with the AA PMF, unlike the VCG PMF where the attraction begins around 3.0 nm. As seen in **Figure 7B**, the hexamer-hexamer binding free energy of the VCG model (-21.83±0.65 kcal/mol) and of the CG model (-11.23±0.22 kcal/mol) both overestimate the atomistic binding free energy (-7.35±0.21 kcal/mol), although the CG model is closer. Overall, we find that both CG and VCG models tend to partition energetics equally between the intra- and inter-hexameric interfaces, while the AA PMFs suggest that the intra-hexameric

interface should have stronger binding affinity than that of the inter-hexameric interface. This discrepancy highlights a current limitation in the CG/VCG models, which could be addressed by redefining how protein-protein contacts are identified. More importantly, we find that the VCG model predicts stronger attraction compared to the CG model. However, as observed in the Q system, the stronger attraction may not be commensurate with faster assembly kinetics.

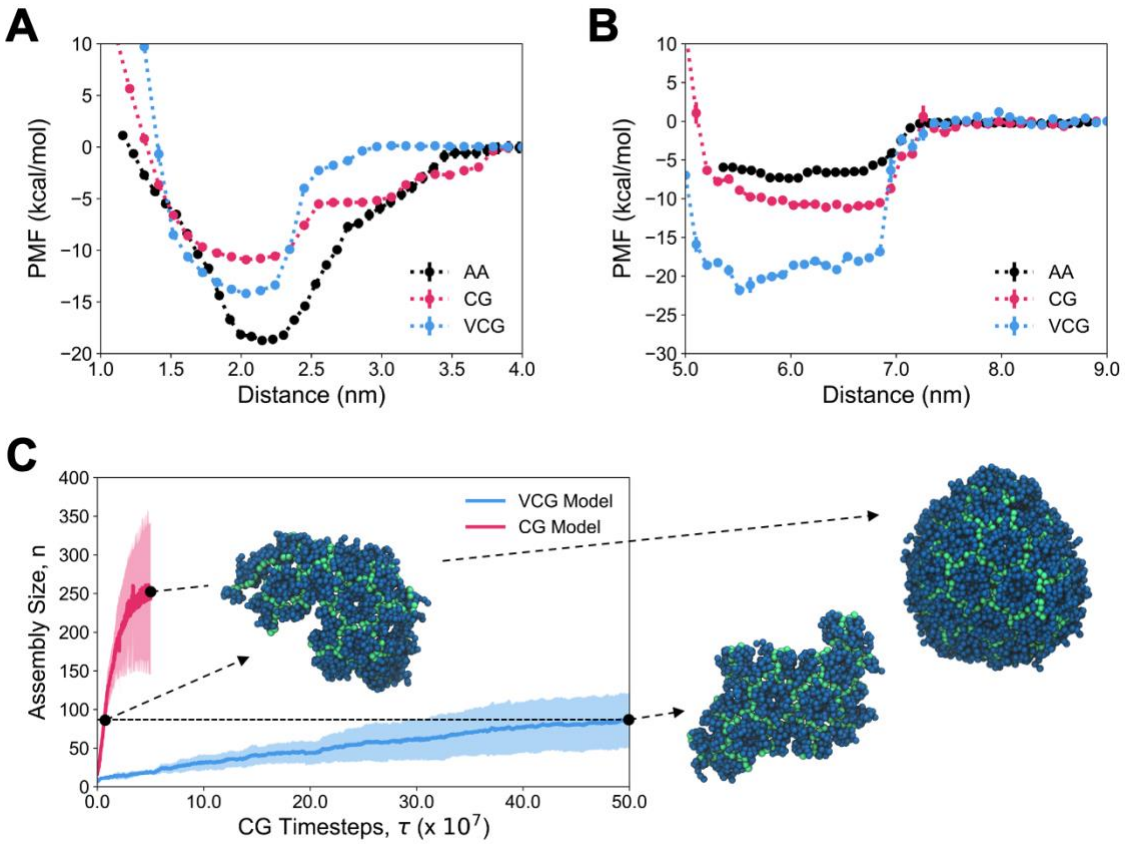


Figure 7. BMC-H self-assembly into hexameric sheets. (A) Comparison of the potential of mean force (PMF) as a function of distance between BMC-H monomers along the intra-hexameric interface for the CG, VCG, and AA models. (B) Comparison of the PMF as a function of distance between BMC-H hexamers along the inter-hexameric interface for the CG, VCG, and AA models. (C) The size of assembled BMC-H lattices over time for the CG and VCG models; the horizontal dashed line serves as a guide to the eye to show equivalent sizes in the CG and VCG assemblies. The arrows mark the time-points for each of the depicted snapshots. The snapshots show representative assembled lattices with inter-hexameric CG sites shown as green balls and the remaining CG sites shown as blue balls.

We performed CGMD assembly simulations of BMC-H at a fixed concentration of 4 mM from a reservoir of 10 mM protein. **Figure 7C** depicts the aggregate size of assembled BMC-H over time for both the CG and VCG models. It is evident that the CG model assembles at a rate that significantly exceeds that of the VCG model; the VCG model assembled into a lattice of size 86±36 BMC-H after $50x10^7$ time steps, while the CG model required $0.6x10^7$ time steps to achieve a comparable size. The morphologies predicted by the two models are also distinct. As seen in **Figure 7C** and **Figure S8**, comparison of the lattices predicted by the two models at a size around 85 reveals that the VCG model assembles with uniform hexamers while the CG model assembles with an assortment of hexamers, pentamers, and heptamers. The pentamers and heptamers observed in the CG model simulations are examples of kinetically trapped oligomers that could not anneal into hexamers before becoming enclosed. The CG model continued its rapid assembly until forming a ball-like structure composed of hexamers, pentamers, heptamers, and vacancies (see **Figure 7C**). Interestingly, the pair distance distributions for both the CG and VCG models after assembly (**Figure S9**) reveal that both models recapitulate atomistic statistics, which further suggests that pentameric, hexameric, and heptameric states are degenerate in the captured pairwise statistics. Prior experiments have shown that BMC-H expressed in *E. coli* spontaneously assemble into “swiss roll” or “rosette” structures that are likely curled hexameric sheets of BMC-H.[60, 61] Clearly, the ball-like structure predicted by the CG model is inconsistent with experimental morphologies. The lattice predicted by the VCG model, however, is consistent with hexameric sheets.

Similar to the Q system, we find that the VCG model is able to assemble into experimentally consistent morphologies while the CG model results in defective and kinetically trapped structures. Our results highlight the importance of anisotropic representations of protein-protein interactions, which we capture using virtual sites, that appear to reduce both assembly rates and defect production. We attribute the lower assembly rates using the VCG model as compared to the CG model to the same orientation-dependent entropic barrier noted in the Q study, which also compensates for the larger binding energies observed in the VCG model.

Conclusions

We present a systematic coarse-graining approach to model anisotropic interactions at protein-protein interfaces using virtual sites to facilitate low-resolution and implicit-solvent molecular dynamics simulations of multimeric biomacromolecules. As the virtual sites do not represent an explicit mapping from atomistic configurations, our methodological premise is that the virtual site interactions can be inferred from the statistics of the CG site pairs that the virtual sites are meant to represent. In this work, we show how virtual sites represented by a combination of harmonic bonds and Gaussian interactions can be coupled and related to CG site pair statistics, but we envision that our approach can be adapted for other functional forms. We also note that the VCG modeling framework is limited to pairwise interactions by design, such that the additional computational cost of the VCG model is negligible in comparison to CG models using pairwise interactions.

We show through two case studies of Q coiled-coil proteins and BMC-H shell proteins that the proposed VCG models outperform CG models that use equivalent spherically isotropic interactions in terms of both fidelity to the reference AA distributions and in terms of assembly into higher-order supramolecular structures. The observed discrepancies between reference AA PMFs and VCG PMFs, however, suggest that the positions and numbers of virtual sites defined in VCG models require further tuning. Interestingly, the VCG models consistently predicted larger binding energies compared to the CG models yet also resulted in slower assembly kinetics and avoidance of kinetic traps. We attribute this observation to the inherent anisotropy afforded by the virtual sites, which necessitates specific orientations at protein-protein interfaces to be explored, thereby introducing an entropic barrier for protein-protein association. In future work, it would be insightful to test different virtual site selection schemes and generate competing models using our methodology. For instance, one could adjust the probability distribution-based criteria that we used to select for candidate protein contacts. Alternatively, one could select protein contacts on the basis of protein-protein interaction types such as cation-π or salt-bridge interactions. By comparing VCG models that are

systematically constructed using different virtual site selection schemes, one can formally investigate the importance of specific protein-protein interactions on macroscopic properties of interest.

Given its systematic nature, our proposed methodology can be extended to other macromolecular systems, including prior work that leveraged virtual sites for intermolecular interactions.[44-52] Our methodology also bears similarity to the recent variational derivative REM (VD-REM) method reported by Sahrmann and coworkers.[53] The VD-REM method optimizes virtual site interaction parameters within a REM framework by approximating the conditional expectation of the potential energy derivative (with respect to virtual site parameters) using machine learning models, which, in turn, is related to the relative entropy derivative with respect to virtual site parameters (this derivative appears in the second term of Eq. 5). The current work sidesteps the need for a machine learning expectation estimator given the way our virtual sites are formulated. However, as the VD-REM approach is clearly applicable, it would be informative to investigate different model architectures for the estimator (e.g., gradient boost models or neural networks) to see if comparable model outcomes are predicted or if different higher-order correlations can be captured.

Finally, it is interesting to consider our VCG modeling framework in the context of prior low-resolution CG models that have successfully demonstrated the use of isotropic interactions.[72, 88, 89] While we did not systematically investigate the role of repulsive potentials in the present work, e.g., those introduced by the excluded volume interaction, we speculate that repulsive potentials in prior CG models may have had large enough radii to restrict the orientations available to interacting CG sites, which may have accomplished the same goal as our virtual site interactions. In our current work, the differences between the VCG and CG models are quite evident, which may be due to the smaller effective sizes assumed for each CG site.

In summary, we have demonstrated that virtual site representations are an effective means to increase the expressivity of conventional CG models using isotropic interactions and provide a starting point to extend our methodology to other macromolecular systems of interest.

Data Availability Statement

The data underlying this study are available in the published article, its Supporting Information, and openly availably at https://gitlab.com/pak-group/VCG_JPCB.

Supporting Information

Supporting information includes: RMSD analysis from the AAMD simulations (Figure S1), the two-dimensional PMF for Q binding (Figure S2), histograms for each AAMD umbrella sampling window (Figure S3), additional analysis on the CG/VCG/VCG+HP models for Q (Figures S4-S7), additional depictions and analysis on the CG/VCG models for BMC-H (Figures S8-S9), additional details on model optimization and molecular dynamics simulations (Tables S1-S16), and protein sequences for Q and BMC-H (Table S17).

Acknowledgements

This research was supported in part by the National Science Foundation under grant #2138620. This work used Bridges-2 at the Pittsburgh Supercomputing Center and Anvil at the Rosen Center for Advanced Computing through allocation BIO220015 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. L.F.C. acknowledges support from the Mines VPRTT Materials Science Graduate Fellowship Program.

References

(1) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129-1143. DOI: 10.1016/j.neuron.2018.08.011.

(2) Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular dynamics simulations of large macromolecular complexes. *Curr Opin Struc Biol* **2015**, *31*, 64-74. DOI: 10.1016/j.sbi.2015.03.007.

(3) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem Rev* **2016**, *116* (14), 7898-7936. DOI: 10.1021/acs.chemrev.6b00163.

(4) Pak, A. J.; Voth, G. A. Advances in coarse-grained modeling of macromolecular complexes. *Curr Opin Struc Biol* **2018**, *52*, 119-126. DOI: 10.1016/j.sbi.2018.11.005.

(5) Jin, J. H. Y.; Pak, A. J.; Durumeric, A. E. P.; Loose, T. D.; Voth, G. A. Bottom-up Coarse-Graining: Principles and Perspectives. *J Chem Theory Comput* **2022**, *18* (10), 5759–5791. DOI: 10.1021/acs.jctc.2c00643.

(6) Noid, W. G. Perspective: Advances, Challenges, and Insight for Predictive Coarse-Grained Models. *J Phys Chem B* **2023**, *127* (19), 4174-4207. DOI: 10.1021/acs.jpcb.2c08731.

(7) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophysical Journal* **2008**, *95* (11), 5073-5083. DOI: 10.1529/biophysj.108.139626.

(8) Webb, M. A.; Delannoy, J. Y.; de Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *J Chem Theory Comput* **2019**, *15* (2), 1199-1208. DOI: 10.1021/acs.jctc.8b00920.

(9) Pak, A. J.; Dannenhoffer-Lafage, T.; Madsen, J. J.; Voth, G. A. Systematic Coarse-Grained Lipid Force Fields with Semiexplicit Solvation via Virtual Sites. *J Chem Theory Comput* **2019**, *15* (3), 2087-2100. DOI: 10.1021/acs.jctc.8b01033.

(10) Han, Y. N.; Dama, J. F.; Voth, G. A. Mesoscopic coarse-grained representations of fluids rigorously derived from atomistic models. *J Chem Phys* **2018**, *149* (4), 044104. DOI: 10.1063/1.5039738.

(11) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics* **2008**, *128* (24), 244114. DOI: 10.1063/1.2938860.

(12) Tschop, W.; Kremer, K.; Batoulis, J.; Burger, T.; Hahn, O. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polym* **1998**, *49* (2-3), 61-74. DOI: 10.1002/(Sici)1521-4044(199802)49:2/3<61::Aid-Apol61>3.0.Co;2-V.

(13) Reith, D.; Putz, M.; Muller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J Comput Chem* **2003**, *24* (13), 1624-1636. DOI: 10.1002/jcc.10307.

(14) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *J Chem Phys* **2014**, *140* (22), 224104 DOI: 10.1063/1.4880555.
(15) Moore, T. C.; Iacovella, C. R.; McCabe, C. Development of a Coarse-Grained Water Forcefield via Multistate Iterative Boltzmann Inversion. *Molec Model Simul* **2016**, 37-52. DOI: 10.1007/978-981-10-1128-3_3.
(16) Lyubartsev, A. P.; Laaksonen, A. Calculation of Effective Interaction Potentials from Radial-Distribution Functions - a Reverse Monte-Carlo Approach. *Phys Rev E* **1995**, *52* (4), 3730-3737. DOI: 10.1103/PhysRevE.52.3730.
(17) Lyubartsev, A. P.; Laaksonen, A. Osmotic and activity coefficients from effective potentials for hydrated ions. *Phys Rev E* **1997**, *55* (5), 5689-5696. DOI: 10.1103/PhysRevE.55.5689.
(18) Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109* (7), 2469-2473. DOI: 10.1021/jp044629q.
(19) Maragliano, L.; Vanden-Eijnden, E. Single-sweep methods for free energy calculations. *J Chem Phys* **2008**, *128* (18), 184110 DOI: 10.1063/1.2907241.
(20) Lu, L. Y.; Izvekov, S.; Das, A.; Andersen, H. C.; Voth, G. A. Efficient, Regularized, and Scalable Algorithms for Multiscale Coarse-Graining. *J Chem Theory Comput* **2010**, *6* (3), 954-965. DOI: 10.1021/ct900643r.
(21) Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics* **2008**, *129* (14), 144108. DOI: 10.1063/1.2992060.
(22) Chaimovich, A.; Shell, M. S. Relative entropy as a universal metric for multiscale errors. *Phys Rev E* **2010**, *81* (6), 060104. DOI: 10.1103/PhysRevE.81.060104.
(23) Chaimovich, A.; Shell, M. S. Coarse-graining errors and numerical optimization using a relative entropy framework. *J Chem Phys* **2011**, *134* (9), 094112. DOI: 10.1063/1.3557038.
(24) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Voth, G. A. Multiscale coarse-graining and structural correlations: Connections to liquid-state theory. *J Phys Chem B* **2007**, *111* (16), 4116-4127. DOI: 10.1021/jp068549t.
(25) Vanommeslaeghe, K.; Guvench, O.; MacKerell, A. D. Molecular Mechanics. *Curr Pharm Design* **2014**, *20* (20), 3281-3292. DOI: 10.2174/13816128113199990600.
(26) Roberts, C. J.; Blanco, M. A. Role of Anisotropic Interactions for Proteins and Patchy Nanoparticles. *J Phys Chem B* **2014**, *118* (44), 12599-12611. DOI: 10.1021/jp507886r.
(27) Whitelam, S.; Jack, R. L. The Statistical Mechanics of Dynamic Pathways to Self-Assembly. *Annu Rev Phys Chem* **2015**, *66*, 143-163. DOI: 10.1146/annurev-physchem-040214-121215.
(28) Gay, J. G.; Berne, B. J. Modification of the Overlap Potential to Mimic a Linear Site-Site Potential. *J Chem Phys* **1981**, *74* (6), 3316-3319. DOI: 10.1063/1.441483.
(29) Babadi, M.; Everaers, R.; Ejtehadi, M. R. Coarse-grained interaction potentials for anisotropic molecules. *J Chem Phys* **2006**, *124* (17), 174708. DOI: 10.1063/1.2179075.

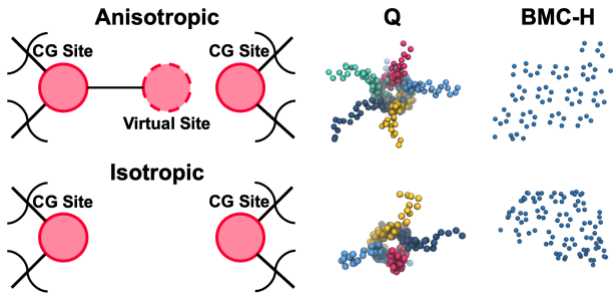
(30) Wu, J.; Zhen, X.; Shen, H. J.; Li, G. H.; Ren, P. Y. Gay-Berne and electrostatic multipole based coarse-grain potential in implicit solvent. *J Chem Phys* **2011**, *135* (15), 155104 DOI: 10.1063/1.3651626.
(31) Shen, H. J.; Li, Y.; Ren, P. Y.; Zhang, D. L.; Li, G. H. Anisotropic Coarse-Grained Model for Proteins Based On Gay-Berne and Electric Multipole Potentials. *J Chem Theory Comput* **2014**, *10* (2), 731-750. DOI: 10.1021/ct400974z.
(32) Stillinger, F. H.; Weber, T. A. Computer-Simulation of Local Order in Condensed Phases of Silicon. *Physical Review B* **1985**, *31* (8), 5262-5271. DOI: 10.1103/PhysRevB.31.5262.
(33) Lu, J. B.; Qiu, Y. Q.; Baron, R.; Molinero, V. Coarse-Graining of TIP4P/2005, TIP4P-Ew, SPC/E, and TIP3P to Monatomic Anisotropic Water Models Using Relative Entropy Minimization. *J Chem Theory Comput* **2014**, *10* (9), 4104-4120. DOI: 10.1021/ct500487h.
(34) Larini, L.; Lu, L. Y.; Voth, G. A. The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials. *J Chem Phys* **2010**, *132* (16), 164107. DOI: 10.1063/1.3394863.
(35) Das, A.; Andersen, H. C. The multiscale coarse-graining method. IX. A general method for construction of three body coarse-grained force fields. *J Chem Phys* **2012**, *136* (19), 194114. DOI: 10.1063/1.4705417.
(36) Scherer, C.; Andrienko, D. Understanding three-body contributions to coarse-grained force fields. *Physical Chemistry Chemical Physics* **2018**, *20* (34), 22387-22394. DOI: 10.1039/c8cp00746b.
(37) Zhang, L. F.; Han, J. Q.; Wang, H.; Car, R.; E, W. N. DeePCG: Constructing coarse-grained models via deep neural networks. *J Chem Phys* **2018**, *149* (3), 034101 DOI: 10.1063/1.5027645.
(38) Wang, J.; Olsson, S.; Wehmeyer, C.; Perez, A.; Charron, N. E.; de Fabritiis, G.; Noe, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *Acs Central Sci* **2019**, *5* (5), 755-767. DOI: 10.1021/acscentsci.8b00913.
(39) Husic, B. E.; Charron, N. E.; Lemm, D.; Wang, J.; Perez, A.; Majewski, M.; Kramer, A.; Chen, Y. Y.; Olsson, S.; de Fabritiis, G.; et al. Coarse graining molecular dynamics with graph neural networks. *J Chem Phys* **2020**, *153* (19), 194101. DOI: 10.1063/5.0026133.
(40) Ruza, J.; Wang, W. J.; Schwalbe-Koda, D.; Axelrod, S.; Harris, W. H.; Gomez-Bombarelli, R. Temperature-transferable coarse-graining of ionic liquids with dual graph convolutional neural networks. *J Chem Phys* **2020**, *153* (16), 164501. DOI: 10.1063/5.0022431.
(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79* (2), 926-935. DOI: 10.1063/1.445869.
(42) Melo, M. N.; Ingolfsson, H. I.; Marrink, S. J. Parameters for Martini sterols and hopanoids based on a virtual-site description. *J Chem Phys* **2015**, *143* (24), 243152 DOI: 10.1063/1.4937783.
(43) Jin, J.; Han, Y.; Voth, G. A. Coarse-graining involving virtual sites: Centers of symmetry coarse-graining. *J Chem Phys* **2019**, *150* (15), 154103. DOI: 10.1063/1.5067274.

(44) Perlmutter, J. D.; Qiao, C.; Hagan, M. F. Viral genome structures are optimal for capsid assembly. *Elife* **2013**, *2*, e00632. DOI: 10.7554/eLife.00632.
(45) Perlmutter, J. D.; Hagan, M. F. The Role of Packaging Sites in Efficient and Specific Virus Assembly. *J Mol Biol* **2015**, *427* (15), 2451-2467. DOI: 10.1016/j.jmb.2015.05.008.
(46) Ruiz-Herrero, T.; Hagan, M. F. Simulations Show that Virus Assembly and Budding Are Facilitated by Membrane Microdomains. *Biophysical Journal* **2015**, *108* (3), 585-595. DOI: 10.1016/j.bpj.2014.12.017.
(47) Wu, Z. J.; Beltran-Villegas, D. J.; Jayaraman, A. Development of a New Coarse-Grained Model to Simulate Assembly of Cellulose Chains Due to Hydrogen Bonding. *J Chem Theory Comput* **2020**, *16* (7), 4599-4614. DOI: 10.1021/acs.jctc.0c00225.
(48) Grime, J. M. A.; Dama, J. F.; Ganser-Pornillos, B. K.; Woodward, C. L.; Jensen, G. J.; Yeager, M.; Voth, G. A. Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. *Nat Commun* **2016**, *7*, 11568. DOI: 10.1038/ncomms11568.
(49) Pak, A. J.; Grime, J. M. A.; Sengupta, P.; Chen, A. K.; Durumeric, A. E. P.; Srivastava, A.; Yeager, M.; Briggs, J. A. G.; Lippincott-Schwartz, J.; Voth, G. A. Immature HIV-1 lattice assembly dynamics are regulated by scaffolding from nucleic acid and the plasma membrane. *P Natl Acad Sci USA* **2017**, *114* (47), E10056-E10065. DOI: 10.1073/pnas.1706600114.
(50) Pak, A. J.; Grime, J. M. A.; Yu, A.; Voth, G. A. Off-Pathway Assembly: A Broad-Spectrum Mechanism of Action for Drugs That Undermine Controlled HIV-1 Viral Capsid Formation. *Journal of the American Chemical Society* **2019**, *141* (26), 10214-10224. DOI: 10.1021/jacs.9b01413.
(51) Gupta, M.; Pak, A. J.; Voth, G. A. Critical mechanistic features of HIV-1 viral capsid assembly. *Science Advances* **2023**, *9*, eadd7434. DOI: 10.1126/sciadv.add7434.
(52) Yu, A.; Skorupka, K. A.; Pak, A. J.; Ganser-Pornillos, B. K.; Pornillos, O.; Voth, G. A. TRIM5 alpha self-assembly and compartmentalization of the HIV-1 viral capsid. *Nature Communications* **2020**, *11* (1), 1307 DOI: 10.1038/s41467-020-15106-1.
(53) Sahrmann, P. G.; Loose, T. D.; Durumeric, A. E. P.; Voth, G. A. Utilizing Machine Learning to Greatly Expand the Range and Accuracy of Bottom-Up Coarse-Grained Models through Virtual Particles. *J Chem Theory Comput* **2023**, *19* (14), 4402-4413. DOI: 10.1021/acs.jctc.2c01183.
(54) Hume, J.; Sun, J.; Jacquet, R.; Renfrew, P. D.; Martin, J. A.; Bonneau, R.; Gilchrist, M. L.; Montclare, J. K. Engineered Coiled-Coil Protein Microfibers. *Biomacromolecules* **2014**, *15* (10), 3503-3510. DOI: 10.1021/bm5004948.
(55) Hill, L. K.; Meleties, M.; Katyal, P.; Xie, X.; Delgado-Fukushima, E.; Jihad, T.; Liu, C. F.; O'Neill, S.; Tu, R. S.; Renfrew, P. D.; et al. Thermoresponsive Protein-Engineered Coiled-Coil Hydrogel for Sustained Small Molecule Release. *Biomacromolecules* **2019**, *20* (9), 3340-3351. DOI: 10.1021/acs.biomac.9b00107.
(56) Dawson, W. M.; Martin, F. J. O.; Rhys, G. G.; Shelley, K. L.; Brady, R. L.; Woolfson, D. N. Coiled coils 9-to-5: rational de novo design of alpha-helical barrels with tunable oligomeric states. *Chem Sci* **2021**, *12* (20), 6923-6928. DOI: 10.1039/d1sc00460c.

(57) Rhys, G. G.; Wood, C. W.; Lang, E. J. M.; Mulholland, A. J.; Brady, R. L.; Thomson, A. R.; Woolfson, D. N. Maintaining and breaking symmetry in homomeric coiled-coil assemblies. *Nature Communications* **2018**, *9*, 4132 DOI: 10.1038/s41467-018-06391-y.
(58) Sutter, M.; Greber, B.; Aussignargues, C.; Kerfeld, C. A. Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell. *Science* **2017**, *356*, 1293-1297. DOI: 10.1126/science.aan3289.
(59) Lassila, J. K.; Bernstein, S. L.; Kinney, J. N.; Axen, S. D.; Kerfeld, C. A. Assembly of Robust Bacterial Microcompartment Shells Using Building Blocks from an Organelle of Unknown Function. *J Mol Biol* **2014**, *426* (11), 2217-2228. DOI: 10.1016/j.jmb.2014.02.025.
(60) Hagen, A. R.; Plegaria, J. S.; Sloan, N.; Ferlez, B.; Aussignargues, C.; Burton, R.; Kerfeld, C. A. In Vitro Assembly of Diverse Bacterial Microcompartment Shell Architectures. *Nano Letters* **2018**, *18* (11), 7030-7037. DOI: 10.1021/acs.nanolett.8b02991.
(61) Sutter, M.; Faulkner, M.; Aussignargues, C.; Paasch, B. C.; Barrett, S.; Kerfeld, C. A.; Liu, L. N. Visualization of Bacterial Microcompartment Facet Assembly Using High-Speed Atomic Force Microscopy. *Nano Lett* **2016**, *16* (3), 1590-1595. DOI: 10.1021/acs.nanolett.5b04259.
(62) Wood, C. W.; Woolfson, D. N. CCBuilder 2.0: Powerful and accessible coiled-coil modeling. *Protein Sci* **2018**, *27* (1), 103-111. DOI: 10.1002/pro.3279.
(63) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19-25. DOI: 10.1016/j.softx.2015.06.001.
(64) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., Jr. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **2017**, *14* (1), 71-73. DOI: 10.1038/nmeth.4067.
(65) Bussi, G.; Zykova-Timan, T.; Parrinello, M. Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *J Chem Phys* **2009**, *130* (7), 074101. DOI: 10.1063/1.3073889.
(66) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52* (12), 7182-7190. DOI: 10.1063/1.328693.
(67) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Computer Physics Communications* **1977**, *23*, 187-199. DOI: 10.1016/0021-9991(77)90121-8.
(68) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33-38. DOI: 10.1016/0263-7855(96)00018-5.
(69) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014**, *185* (2), 604-613. DOI: 10.1016/j.cpc.2013.09.018.
(70) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J Comput Chem* **1992**, *13* (8), 1011-1021.

(71) E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J Chem Phys* **2007**, *126* (16), 164103. DOI: 10.1063/1.2720838.
(72) Pak, A. J.; Gupta, M.; Yeager, M.; Voth, G. A. Inositol Hexakisphosphate (IP6) Accelerates Immature HIV-1 Gag Protein Assembly toward Kinetically Trapped Morphologies. *J Am Chem Soc* **2022**, *144* (23), 10417-10428. DOI: 10.1021/jacs.2c02568.
(73) Pak, A. J.; Dannenhoffer-Lafage, T.; Madsen, J. J.; Voth, G. A. Systematic Coarse-Grained Lipid Force Fields with Semiexplicit Solvation via Virtual Sites. *J Chem Theory Comput* **2019**, *15* (3), 2087-2100. DOI: 10.1021/acs.jctc.8b01033 From NLM Medline.
(74) Chaimovich, A.; Shell, M. S. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. *Phys Chem Chem Phys* **2009**, *11* (12), 1901-1915. DOI: 10.1039/b818512c Medline.
(75) Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S. M.; Jordan, M. I. How to Escape Saddle Points Efficiently. *Pr Mach Learn Res* **2017**, *70*.
(76) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys J* **2008**, *95* (11), 5073-5083. DOI: 10.1529/biophysj.108.139626 From NLM Medline.
(77) Lyman, E.; Pfaendtner, J.; Voth, G. A. Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophys J* **2008**, *95* (9), 4183-4192. DOI: 10.1529/biophysj.108.139733.
(78) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **2022**, *271*. DOI: 10.1016/j.cpc.2021.108171.
(79) Jewett, A. I.; Zhuang, Z.; Shea, J.-E. Moltemplate a Coarse-Grained Model Assembly Tool. *Biophysical Journal* **2013**, *104* (2), 169a. DOI: 10.1016/j.bpj.2012.11.953.
(80) Schneider, T.; Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Physical Review B* **1978**, *17* (3), 1302-1322. DOI: 10.1103/PhysRevB.17.1302.
(81) Sutter, M.; McGuire, S.; Ferlez, B.; Kerfeld, C. A. Structural Characterization of a Synthetic Tandem-Domain Bacterial Microcompartment Shell Protein Capable of Forming Icosahedral Shell Assemblies. *Acs Synth Biol* **2019**, *8* (4), 668-674. DOI: 10.1021/acssynbio.9b00011.
(82) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the 7th Python in Science Conference* **2008**, 11 - 15.
(83) Burgess, N. C.; Sharp, T. H.; Thomas, F.; Wood, C. W.; Thomson, A. R.; Zaccai, N. R.; Brady, R. L.; Serpell, L. C.; Woolfson, D. N. Modular Design of Self-Assembling Peptide-Based Nanotubes. *Journal of the American Chemical Society* **2015**, *137* (33), 10554-10562. DOI: 10.1021/jacs.5b03973.

(84) Kerfeld, C. A.; Aussignargues, C.; Zarzycki, J.; Cai, F.; Sutter, M. Bacterial microcompartments. *Nat Rev Microbiol* **2018**, *16* (5), 277-290. DOI: 10.1038/nrmicro.2018.10.
(85) Ochoa, J. M.; Yeates, T. O. Recent structural insights into bacterial microcompartment shells. *Curr Opin Microbiol* **2021**, *62*, 51-60. DOI: 10.1016/j.mib.2021.04.007.
(86) Yeates, T. O.; Jorda, J.; Bobik, T. A. The Shells of BMC-Type Microcompartment Organelles in Bacteria. *J Mol Microb Biotech* **2013**, *23* (4-5), 290-299. DOI: 10.1159/000351347.
(87) Pang, A.; Frank, S.; Brown, I.; Warren, M. J.; Pickersgill, R. W. Structural Insights into Higher Order Assembly and Function of the Bacterial Microcompartment Protein PduA. *J Biol Chem* **2014**, *289* (32), 22377-22384. DOI: 10.1074/jbc.M114.569285.
(88) Yu, A.; Pak, A. J.; He, P.; Monje-Galvan, V.; Casalino, L.; Gaieb, Z.; Dommer, A. C.; Amaro, R. E.; Voth, G. A. A multiscale coarse-grained model of the SARS-CoV-2 virion. *Biophysical Journal* **2021**, *120* (6), 1097-1104. DOI: 10.1016/j.bpj.2020.10.048.
(89) Pak, A. J.; Yu, A.; Ke, Z. L.; Briggs, J. A. G.; Voth, G. A. Cooperative multivalent receptor binding promotes exposure of the SARS-CoV-2 fusion machinery core. *Nature Communications* **2022**, *13* (1), 1002 DOI: 10.1038/s41467-022-28654-5.


Anisotropic
Q
BMC-H
CG Site
CG Site
Virtual Site
Isotropic
CG Site
CG Site

TOC Figure