CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences

Kiersten M. Ruff,¹ Tyler S. Harmon,² and Rohit V. Pappu^{3,a)}

¹*Computational and Systems Biology Program and Center for Biological Systems Engineering, Washington University in St. Louis, St. Louis, Missouri 63130-4899, USA*

²*Department of Physics and Center for Biological Systems Engineering, Washington University in St. Louis, St. Louis, Missouri 63130-4899, USA*

³*Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, CB 1097, St. Louis, Missouri 63130-4899, USA*

(Received 16 August 2015; accepted 21 October 2015; published online 9 November 2015)

We report the development and deployment of a coarse-graining method that is well suited for computer simulations of aggregation and phase separation of protein sequences with block-copolymeric architectures. Our algorithm, named CAMELOT for Coarse-grained simulations Aided by MachinE Learning Optimization and Training, leverages information from converged all atom simulations that is used to determine a suitable resolution and parameterize the coarse-grained model. To parameterize a system-specific coarse-grained model, we use a combination of Boltzmann inversion, non-linear regression, and a Gaussian process Bayesian optimization approach. The accuracy of the coarse-grained model is demonstrated through direct comparisons to results from all atom simulations. We demonstrate the utility of our coarse-graining approach using the block-copolymeric sequence from the exon 1 encoded sequence of the huntingtin protein. This sequence comprises of 17 residues from the N-terminal end of huntingtin (N17) followed by a polyglutamine (polyQ) tract. Simulations based on the CAMELOT approach are used to show that the adsorption and unfolding of the wild type N17 and its sequence variants on the surface of polyQ tracts engender a patchy colloid like architecture that promotes the formation of linear aggregates. These results provide a plausible explanation for experimental observations, which show that N17 accelerates the formation of linear aggregates in block-copolymeric N17-polyQ sequences. The CAMELOT approach is versatile and is generalizable for simulating the aggregation and phase behavior of a range of block-copolymeric protein sequences. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4935066>]

I. INTRODUCTION

Protein sequences with block-copolymeric architectures can aggregate and drive reversible liquid-liquid demixing or sol-gel transitions that give rise to dense liquid or gel-like phases.^{1–17} A subset of these sequences also form amorphous or semi-crystalline solids such as amyloid fibers.^{1,18–30} The sequence blocks in these proteins are either well-folded domains or intrinsically disordered regions (IDRs). The latter refer to regions that fail to fold into unique three-dimensional structures as autonomous units.^{31–34} Accordingly, depending on their architecture, protein sequences that drive aggregation and phase separation may be classified as being intrinsically disordered or partially disordered block-copolymers (Figure 1). Molecular simulations can play a prominent role in uncovering the physical principles that govern the phase behavior of block-copolymeric proteins. Since aggregation and phase separation are driven by collective interactions among large numbers of polymeric molecules, the underlying physics of phase transitions and the demands of computational efficiency mandate the development and deployment of coarse-grained simulations.^{35–38} Here, we report the design and

implementation of a method for systematic coarse-graining that can be adapted to the two classes of block-copolymeric protein sequences depicted in Figure 1. The algorithm is named CAMELOT and it stands for Coarse-grained simulations Aided by MachinE Learning Optimization and Training. We demonstrate the utility of the CAMELOT method by deploying it in Langevin dynamics simulations of archetypal block-copolymeric sequences that are based on the exon 1 encoded intrinsically disordered region of the huntingtin protein.

There are two distinct paradigms for coarse-graining.³⁹ One is based on fixed resolution transferrable models with consensus parameters.^{40–43} The CAMELOT method follows a different paradigm whereby system-specific coarse-grained models are designed using information gleaned from prior finer-grained simulations. The general outline of the CAMELOT method is as follows: For the block-copolymeric sequence of interest, we first perform converged all atom simulations aided by enhanced sampling methods to obtain descriptions of conformational ensembles for individual molecules and oligomers. We use the information gleaned from the all atom simulations to prescribe the resolution, forcefield architecture, and parameters for the coarse-grained model. The questions we wish to answer and the length scales that we wish to access in our simulations will dictate the choice

^{a)}Electronic mail: pappu@wustl.edu

FIG. 1. Example (a) intrinsically disordered block-copolymeric and (b) partially disordered block-copolymeric sequences implicated in aggregation and phase separation. Sequences can be broken up into blocks consisting of ordered domains (circles) or intrinsically disordered regions with compositional biases (rounded rectangles). (a) Htt Exon 1 (UniProt ID: P42858) and Sup35p NM (UniProt ID: P05453) are archetypal intrinsically disordered block-copolymeric sequences. Htt Exon 1 forms insoluble inclusions that are the pathological hallmarks of Huntington's disease.⁷⁸ Sup35p NM forms insoluble amyloid fibrils and propagates as a prion.^{128,129} (b) NCK1 (UniProt ID: P16333) and Whi3p (UniProt ID: Q75E28) are archetypal partially disordered block-copolymeric sequences. NCK1 is involved in the formation of liquid-like micron-sized membraneless organelles important for signal integration.¹³⁰ Whi3p is involved in the formation of functional assemblies important for branching and cell cycle control in filamentous fungi.¹⁸

of resolution. The parameterization of the model of choice is based on a combination of Boltzmann inversion,^{44,45} regression analysis, and a Gaussian process Bayesian optimization approach.⁴⁶ Our design of CAMELOT is guided by the successes of previous methods that include the force-matching algorithm of Voth and coworkers,^{47–58} the Yvon-Born-Green formalism of Noid and coworkers,^{59–61} and the relative entropy method of Shell and coworkers.^{62–64}

The remainder of the text is organized as follows: Section II describes the CAMELOT algorithm. Section III describes the numerical methods we use for all atom and coarse-grained simulations. In Section IV, we apply the CAMELOT method in coarse-grained simulations of archetypal block-copolymeric sequences. Our results help to establish the accuracy of the CAMELOT method. Section V demonstrates our ability to answer specific questions regarding the impact of N-terminal flanking sequences as modulators of the aggregation of polyglutamine tracts. Section VI discusses the physical insights that emerge regarding the polyglutamine containing systems and possible generalizations of the CAMELOT approach to studies of aggregation and phase behavior in other block-copolymeric systems.

II. CAMELOT METHOD

A. Choosing the resolutions for coarse-grained simulations

In coarse-grained simulations we wish to achieve the optimal balance of *relevance*, *efficiency*, and *accuracy*. Relevance refers to the fact that aggregation is a multimolecular phenomenon and requires the incorporation of $O(10^3)$ molecules in each simulation. However, such simulations need to be efficient while also being accurate. This tripartite balance

among relevance, efficiency, and accuracy is not achievable using all atom representations of molecules and this in turn necessitates the use of coarse-grained models. However, the latter can become generic and inaccurate for the specific system in question if the model is not designed to achieve both efficiency and accuracy. Block-copolymeric systems are well suited for coarse-grained simulations because the degree of coupling within and between blocks is distinctly different. The block-copolymeric nature of IDRs emerges from compositional-to-conformational relationships of these sequences.^{34,65} These relationships are governed by specific compositional biases such as the fraction of charged residues and the presence or absence of stretches of polar amino acids. Depending on the compositional biases, IDRs can adopt either globular or random-coil-like conformations. In contrast to IDRs, the ordered domains are essentially semi-rigid globules. Accordingly, we deploy two categories of coarse-grained beads whereby the beads mimic either individual residues or groups of residues (domains) that are modeled as colloidal beads.

A residue bead represents a single amino acid residue that is centered on its center-of-mass. In the CAMELOT model, there are three distinct groups of residue beads. The charged amino acids (Arg, Asp, Glu, Lys) belong to *group-Ch*, the strongly interacting residues (Asn, His, Ile, Leu, Gln, Met, Phe, Trp, Tyr, Val) belong to *group-S*, and the weakly interacting amino acids (Ala, Cys, Gly, Pro, Ser, Thr) belong to *group-W*. We derived these groupings from the amino acid solubility data of Auton and Bolen.⁶⁶ These data quantify the free energy changes associated with the transfer of an amino acid from an aqueous solution to a crystalline state. The solubility data are relevant because they capture the competition that characterizes the transfer of proteins from their dispersed states in aqueous solutions to aggregates that are rich in protein-protein as opposed to protein-solvent interactions.

We use colloidal beads for groups of residues that correspond to ordered globular domains or IDRs that adopt globular conformations. We capture the effects of internal structures within colloidal beads using suitable inter-particle potentials that we glean from all atom simulations of pairwise associations of ordered domains and/or IDRs. The use of colloidal beads reduces the number of beads in the system enabling the inclusion of $O(10^3)$ to $O(10^4)$ molecules in coarse-grained simulations. This allows us to access length scales that are directly relevant to experimental studies of aggregation and phase separation of block-copolymeric sequences.

B. Effective energy functions for coarse-grained simulations

The effective energy functions for CAMELOT based coarse-grained simulations take the form

$$W_{\text{eff}} = W_b + W_\theta + W_\phi + W_{LJ} + W_{el} + W_C. \quad (1)$$

In Equation (1), the terms on the right hand side, respectively, correspond to bond length, bond angle, dihedral angle, Lennard-Jones, electrostatic, and colloidal bead interaction potentials. The choice of the resolution for coarse-grained simulations fixes the terms in Equation (1). If each residue is modeled as a single bead, then $W_C = 0$. Conversely, if every molecule is modeled as a single colloidal bead, then all terms except W_C become zero. All of the terms in Equation (1) are part of the effective energy function for hybrid resolutions that include residue specific beads and domain specific colloidal beads that correspond to groups of residues. The effective energy function is parameterized against data gathered from all atom simulations that are obtained at a specific temperature T . The specific parameters for W_{eff} are, therefore, valid for the particular temperature T at which the all atom simulations are performed. Each of the terms in Equation (1) is discussed in detail below.

1. Bonded interactions

Flexible bonds connect pairs of consecutive beads and there are preferred values for the lengths of these bonds and the angles that define the junctions at pairs of bonds. In terms of N_b and N_θ , the total number of bonds and bond angles in the system, the bond length and bond angle terms in Equation (1) are written as

$$W_b = \sum_{i=1}^{N_b} \frac{K_i(b_i - b_{0i})^2}{2}, \quad (2)$$

$$W_\theta = \sum_{i=1}^{N_\theta} \frac{L_i(\theta_i - \theta_{0i})^2}{2}.$$

In Equation (2), K_i and L_i are the force constants that quantify the penalties associated with deforming bonds and bending bond angles beyond the corresponding equilibrium values of b_{0i} and θ_{0i} . The energies associated with rotations around bonds that connect four consecutive beads are defined in terms of the dihedral angle potential, which is a Fourier series of the form

$$W_\phi = \sum_{i=1}^{N_\phi} \sum_{n=1}^3 V_{ni} [1 - \cos(n\phi_i - \phi_{ni})]. \quad (3)$$

In Equation (3), N_ϕ refers to the number of dihedral angles, n is the number of terms in the Fourier series, V_{ni} is the amplitude, and ϕ_{ni} is the phase.

2. Lennard-Jones interactions

The W_{LJ} term, which is calculated over non-bonded pairs of beads, is written as

$$W_{LJ} = \sum_{i=1}^{N_{nb}} \sum_{j<i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right], \quad r_{ij} < r_{\text{vdW}}^{(ij)}$$

$$= 0, \quad r_{ij} \geq r_{\text{vdW}}^{(ij)}. \quad (4)$$

In Equation (4), N_{nb} is the number of interacting beads for the Lennard-Jones potential, r_{ij} is the distance between beads i and j , σ_{ij} is the distance at which the inter-bead potential is zero, and ϵ_{ij} is the strength of the interaction. The set of free parameters for W_{LJ} is chosen to be $\mathbf{p}_{\text{vdW}} \equiv [\sigma_{ChCh}, \epsilon_{SS}, \epsilon_{SC_1}, \epsilon_{SC_2}, \dots, \epsilon_{SC_N}, \epsilon_{WW}, \epsilon_{WC_1}, \epsilon_{WC_2}, \dots, \epsilon_{WC_N}]$. Here, the subscript Ch refers to residue beads with a net charge, S refers to strongly interacting residue beads, W refers to weakly interacting residue beads, and C_N refers to colloidal beads of type N . All other parameters for the W_{LJ} term are prescribed either *a priori* or are determined by Lorentz-Berthelot mixing rules. The ϵ -values for the interactions between pairs of charged residue beads or between charged residue beads and colloidal beads are set to be 0.01 kcal/mol. The σ -value for beads corresponding to neutral residues is set to be equal to the average radii of gyration (R_g) values of these residues in the all atom simulations. The σ -values for interactions between any residue bead X and any colloidal bead C are set using the relation

$$\sigma_{XC} = \frac{R_g(C)}{(2)^{1/6}}. \quad (5)$$

In Equation (5), $R_g(C)$ is the average R_g of the group of residues that corresponds to the colloidal bead, C . For each interaction pair ij , the cutoff is set using the relation $r_{\text{vdW}}^{(ij)} = 2.5\sigma_{ij}$.

3. Electrostatic interactions

Residue beads that correspond to Lys, Arg, Glu, or Asp will have a net charge of ± 1 . The net charge of a colloidal bead is the sum over all charged amino acids within the colloid. The interactions, W_{el} , between beads with excess charge are written in terms of a Yukawa potential,

$$W_{el} = \sum_{i=1}^{n_c} \sum_{j<i} \frac{Cq_i q_j}{\epsilon r_{ij}} \exp\left(-\frac{r_{ij}}{l_D}\right) \quad \text{for } r_{ij} < r_c$$

$$= 0, \quad \text{for } r_{ij} \geq r_c. \quad (6)$$

Here, n_c denotes the number of beads that have an excess charge, q_i is the charge on bead i , $\epsilon = 78$ is the dielectric constant, $l_D = 10 \text{ \AA}$ is the Debye length for physiologically

relevant ionic strengths, and r_c is the distance past which the mean-field electrostatic interactions are zeroed out.

4. Interactions between colloidal beads

The coarse-grained model for a block-copolymeric sequence may have one or more colloidal beads. These beads are either identical to or different from one another. The number and types of colloidal beads are specified by the sequence encoded conformational properties of the block-copolymeric sequence of interest. The collective interactions among colloidal beads are modeled using the W_C term in Equation (1), where $W_C = \sum_{i=1}^{n_t} \sum_{k=1}^{g_i} \sum_{j=i+1}^{n_t} \sum_{l=1}^{g_j} w_C(\mathbf{r}_k^{(i)}, \mathbf{r}_l^{(j)})$. Here, n_t is the number of distinct types of colloidal beads, g_i is the number of colloidal beads of type i , and $\mathbf{r}_k^{(i)}$ is the position vector of bead k of type i . The functional form for w_c is not defined *a priori* but is determined using data gathered from simulations of pairs of colloidal beads represented in all atom detail. An example of the implementation of our approach for defining w_c is described in Subsection IV B for the sequence of an archetypal block-copolymer.

C. Parameterization of uncoupled W_{eff} terms: W_b , W_θ , W_ϕ , and w_C

We use a Boltzmann inversion procedure^{44,45} to extract parameters for uncoupled terms within W_{eff} . For a given observable, such as the bond length x , the observed probability distribution from all atom simulations is written as $\rho(x) \propto \exp[-W(x)/k_B T]$. Here, $\rho(x)$ is the probability density associated with the observable x , $W(x)$ is the effective potential in terms of the observable, k_B is the Boltzmann constant, and T is the simulation temperature.

Since W_b and W_θ are modeled using harmonic potentials, inversion of the Boltzmann relationship yields the analytical relationships for the parameters of the potentials in terms of the first and second moments, viz., $\langle b_i \rangle$, $\langle \theta_i \rangle$, $\langle b_i^2 \rangle$, and $\langle \theta_i^2 \rangle$ of the relevant probability distributions,

$$\begin{aligned} b_{0i} = \langle b_i \rangle, K_i &= \frac{k_B T}{\langle b_i^2 \rangle - \langle b_i \rangle^2}, \\ \theta_{0i} = \langle \theta_i \rangle, L_i &= \frac{k_B T}{\langle \theta_i^2 \rangle - \langle \theta_i \rangle^2}. \end{aligned} \quad (7)$$

Parameters for W_ϕ and w_C are obtained via non-linear regression analysis by fitting W_ϕ and w_C to the equation

$$W(x) = -k_B T \ln[P(x)]. \quad (8)$$

In Equation (8), $P(x)$ is either the probability distribution for each dihedral angle for W_ϕ or the pair correlation function between centers-of-mass of grouped residues that correspond to a given pair of colloidal beads for w_C .

D. Parameterization of W_{LJ}

1. Primary objective function

As described in Subsection II B 2, the set of free parameters for W_{LJ} is chosen to be $\mathbf{p}_{\text{vdW}} \equiv [\sigma_{ChCh}, \epsilon_{SS}, \epsilon_{SC_1}, \epsilon_{SC_2}, \dots,$

$\epsilon_{SC_N}, \epsilon_{WW}, \epsilon_{WC_1}, \epsilon_{WC_2}, \dots, \epsilon_{WC_N}]$. Given this set of free parameters, we seek values for $\mathbf{p}_{\text{vdW}}$ that minimize the objective function $\Omega(\mathbf{p}_{\text{vdW}})$, which is defined as

$$\Omega(\mathbf{p}_{\text{vdW}}) = 1 - \frac{1}{m} \sum_{i=1}^m \left(\frac{2 - \sum_{j=1}^{n_{\text{bin}}} |\rho_{ij}^{\text{AA}} - \rho_{ij}^{\text{CG}}(\mathbf{p}_{\text{vdW}})|}{2} \right). \quad (9)$$

Here, m is the number of distinct inter-bead distances in the coarse-grained model that correspond to pairs of beads that are separated by at least five bonds. Data for the histograms that quantify the probability densities for distances between pairs of interacting beads are recorded in simulations based on the all atom (AA) and coarse-grained (CG) models. For a pair of beads designated by a single index i , the densities within each of the bins j of the corresponding histograms are denoted as ρ_{ij}^{AA} and ρ_{ij}^{CG} . The optimal parameters for $\mathbf{p}_{\text{vdW}}$ should minimize $\Omega(\mathbf{p}_{\text{vdW}})$ and hence maximize the collective overlap among m pairs of histograms where $m = (n-5)(n-4)/2$. Here, n denotes the number of beads in the coarse-grained model.

2. Auxiliary objective function

The design of the primary objective function assumes that the overlap between probability densities of distances between all atom and coarse-grained models should be weighted equally for all pairs of beads. However, depending on the question of interest, the overlap of distributions for certain pairs of beads may be more important than others. In addition, it might be important to include observables that go beyond distance distributions. In these situations, an auxiliary objective function should be used in a post-processing step to refine some or all of the $\mathbf{p}_{\text{vdW}}$ parameters.

3. Gaussian process Bayesian optimization procedure for obtaining $\mathbf{p}_{\text{vdW}}$

Minimization of $\Omega(\mathbf{p}_{\text{vdW}})$ is non-trivial for two reasons: The objective function does not have an analytical form. Therefore, we cannot readily use gradient-based methods for optimization of $\Omega(\mathbf{p}_{\text{vdW}})$ to predict the optimal values of $\mathbf{p}_{\text{vdW}}$. Second, the objective function is expensive to evaluate because the calculation of $\Omega(\mathbf{p}_{\text{vdW}})$ requires that we perform Langevin dynamics (LD) simulations of individual molecules in the coarse-grained representation to calculate ρ_{ij}^{CG} . Bayesian optimization methods are well suited for such problems. These methods are efficient because they use prior knowledge that is generated during the optimization process to direct the sampling in high-dimensional spaces. Bayesian methods strike an optimal balance between *exploration* (sampling the parameter space, albeit with high uncertainty regarding the objective function) and *exploitation* (using prior knowledge and sampling where the objective function is likely to be minimized).

Our primary goal is to identify values of $\mathbf{p}_{\text{vdW}}$ that minimize the primary objective function Ω . In order to minimize Ω , we use a Bayesian optimization procedure. This requires the collection of observations $O_{1:t} = [(\mathbf{p}_{\text{vdW}, 1:t}, \Omega_{1:t})]$. Here,

$O_{1:t}$ denotes the set of consecutive observations $[O_1, \dots, O_t]$ from sampling the multidimensional space of parameters. The accumulation of observations is used to generate the *likelihood* function, $P(O_{1:t}|\Omega)$. The likelihood function is combined with the *prior* distribution $P(\Omega)$ to generate the *posterior* distribution, $P(\Omega|O_{1:t})$, for the unknown objective function Ω . The posterior probability for obtaining the desired low value of Ω given a collection of observations $O_{1:t}$ is calculated using Bayes theorem,

$$P(\Omega|O_{1:t}) \propto P(O_{1:t}|\Omega) P(\Omega). \quad (10)$$

In order to use Bayesian optimization, we need a model to estimate Ω . Here, we model Ω using a Gaussian process, which is a distribution of objective functions that is defined by its mean and covariance functions. A Gaussian process predicts the most likely values for objective functions and the uncertainties in the estimates for the most likely values. Each Gaussian process Bayesian optimization (GPBO) iteration involves three steps: (i) The algorithm utilizes the posterior distribution and the expected improvement acquisition function to determine the next set of $\mathbf{p}_{\text{vdW}}$ values. (ii) Given a choice for $\mathbf{p}_{\text{vdW}}$, we perform LD simulations for a single chain based on the coarse-grained model. (iii) Data from these simulations are used to calculate $\Omega(\mathbf{p}_{\text{vdW}})$ and the posterior distribution is updated based on the Gaussian process.⁶⁷

Depending on the complexity of the system, the GPBO procedure can be run in parallel with independent trials to ensure convergence to the same parameter space and/or over generations in which each subsequent generation utilizes information from the previous generation to shrink the parameter search space. An example of such a hierarchical approach is described for an example archetypal system in Subsection IV D 3. In general, a single generation, within which hundreds of GPBO iterations are conducted, appears to be reasonable for most systems containing fewer than seven parameters.

III. SIMULATION SETUP

A. All atom simulations

All atom simulations were performed using the CAM-PARI modeling package (<http://camapri.sourceforge.net>) utilizing the ABSINTH implicit solvation model^{68–71} and forcefield paradigm. This model is accurate—as measured against experimental data—and efficient for all atom simulations of conformational properties and intermolecular associations of intrinsically disordered proteins.⁷² In this work, the all atom simulations were based on the `abs3.2_opls.prm` parameter set. Additional details of the all atom simulations are provided in the supplementary material.⁷³

B. Coarse-grained Langevin dynamics simulations

We present the details of the LD simulations in Cartesian space that are used to guide the parameterization of the W_{LJ} terms and for simulations that use the optimized coarse-grained model. All LD simulations were performed using the LAMMPS simulation package (<http://lammps.sandia.gov>).

In these simulations, the Langevin equation shown in Equation (11) is integrated to propagate the positions and velocities of the residue specific and/or domain specific colloidal beads. The force on each of the beads labeled i is written as

$$\mathbf{F}_i = -\nabla W_{\text{eff},i} - \frac{m_i}{\gamma_i} \mathbf{v}_i + \mathbf{R}_i. \quad (11)$$

In Equation (11), $\mathbf{F}_i$ is the force exerted on bead i ; it is a sum of the negative gradient of the effective interactions specified by W_{eff} , the frictional force proportional to the velocity of the bead i , $\mathbf{v}_i$, and the random force $\mathbf{R}_i$ exerted by collisions of the beads with the bath. The random forces have a white noise spectrum in accord with the fluctuation dissipation theorem. The equation of motion is integrated using a velocity Verlet algorithm using an integration time step of 2 fs. The damping term is written as $\gamma_i = \alpha \gamma'_i$. Here, $\alpha = 2$ ps is a scaling factor and $\gamma'_i = \left(\frac{m_i}{6\pi\eta R_g^{(i)}} \right) \gamma_{\text{LYS}}^{-1}$, m_i is the mass of bead i , $\eta = 6.29 \times 10^{-4} \text{ kg m}^{-1} \text{ s}^{-1}$ is the viscosity of water at 315 K, $R_g^{(i)}$ is the average radius of gyration of bead i as calculated from data based on all atom simulations, and γ_{LYS}^{-1} is the inverse of the damping parameter for the lysine bead that is calculated using the Stokes-Einstein relationship. LD simulations of individual coarse-grained molecules were initiated by mapping equilibrated conformations drawn from the all atom simulations to the coarse-grained model.

For multi-chain simulations at finite concentrations, there are at least $O(10^3)$ coarse-grained molecules in each simulation cell. These simulations were performed in the canonical ensemble at constant concentrations, which are controlled using periodic boundary conditions. We initiate the simulations by drawing an initial conformation at random from the equilibrated ensembles of all atom simulations and replicating them on three-dimensional lattices. This is followed by energy minimization using steepest descent and a subsequent equilibration based on 10^6 steps of LD simulations performed with time steps of 2 fs. Each final simulation involves 10^8 integration steps and for every combination of sequence and peptide concentration we performed multiple independent simulations.

IV. APPLICATION OF CAMELOT FOR SIMULATIONS OF BLOCK-COPOLYMERS WITH POLYGLUTAMINE TRACTS

Several proteins with polyglutamine tracts or glutamine-rich regions have been identified as drivers of aggregation and phase separation.^{74–76} The translation of genes with expanded CAG trinucleotide repeats leads to proteins with expanded polyglutamine (polyQ) tracts.⁷⁷ These form insoluble inclusions that are the pathological hallmarks of several neurodegenerative diseases including Huntington's disease.⁷⁸ Water is a poor solvent for homopolymeric polyQ tracts as well as polypeptide backbones.^{79,80} Accordingly, individual polyQ molecules form globular structures to minimize the chain-solvent interface. For polyQ molecules of a particular chain length, there exists a well-defined saturation concentration (c_s) that defines the boundary between soluble and insoluble phases.⁸¹ The insoluble phase is enriched in fibrillar aggregates

and the measured values of c_s decrease with increasing polyQ length. Additionally, for concentrations below c_s , there exists a second saturation concentration (c_c), which corresponds to the formation of spherical aggregates. Kinetics experiments initiated from fully disaggregated solutions that are supersaturated with respect to c_s indicate the early formation of spherical aggregates, 10-30 nm in size, that are precursors of fibrillar aggregates. Fibril formation is barrier-limited and appears to proceed via nucleated conformational conversion within liquid-like spheres⁸²⁻⁸⁴ in accord with the mechanism for crystallization that was proposed by ten Wolde and Frenkel.⁸⁵ In contrast, the formation of metastable liquid-like spheres does not involve discernible free energy barriers.⁸¹

Flanking sequence modules modulates the driving forces for and mechanisms of polyQ aggregation.^{76,81,86-93} This has been observed for the N-terminal stretch of huntingtin. For a given polyQ tract, the presence of N17, the 17-residue N-terminal flanking sequence module, leads to a lowering of c_s vis-à-vis the values measured for polyQ tracts. N17 narrows the gap between c_s and c_c and decreases the metastability of spherical aggregates. Accordingly, N17 helps accelerate the formation of fibrillar aggregates.^{94,95} For a given polyQ length, fibril formation proceeds without a discernible lag time if N17 is appended N-terminally to the polyQ tract. The curious effects of N17 are attributable to a *domain cross talk* between N17 and polyQ.^{96,97} According to this model, the N17 module adsorbs and unfolds on the surface of the polyQ domain.

This engenders a *patchy colloid*⁹⁸⁻¹⁰⁸ architecture for the N17-polyQ block-copolymer. In direct analogy with the physics of patchy colloids, the presence of an adsorbed N17 patch, with charged groups exposed on the surface, leads to a diminution of non-specific interactions of polyQ molecules.^{96,97} The patch on the colloidal particle breaks the spherical symmetry and imparts directional preferences to intermolecular encounters, thus promoting a distinct preference for linear aggregates. Here, we deploy our CAMELOT approach to test the applicability of the patchy colloid model for explaining the N17 enhanced formation of linear aggregates in block-copolymer sequences with polyQ tracts.

A. Determining the coarse-grained resolution from all atom simulations

1. Sequences of interest

For simplicity, the N-terminal and polyQ blocks are denoted as N- and Q-blocks, respectively (see Figure 2). We used sequence design to generate different types of N-block sequences. The different N-block sequences (see Figure 2) can be distinguished by the degree of adsorption between the N- and Q-blocks as shown in Figure 3. Here, we use data from all atom simulations to organize the different N-Q sequences along the ordinate of Figure 3 in ascending order of the degree of adsorption (d_A) between different N-blocks and a Q₄₀-block. The parameter d_A is calculated as

$$d_A = \left(\frac{V_I}{V_N + V_Q} \right),$$

$$V_I = \frac{\pi(R_{g,N} + R_{g,Q} - r_c)^2 (r_c^2 + 2r_c R_{g,Q} - 3R_{g,Q}^2 + 2r_c R_{g,N} + 6R_{g,N}R_{g,Q} - 3R_{g,N}^2)}{12r_c} \text{ if } R_{g,N} + R_{g,Q} > r_c, \quad (12)$$

$$V_I = 0 \text{ if } R_{g,N} + R_{g,Q} \leq r_c.$$

In Equation (12), V_I is the volume of the intersection between spherical envelopes corresponding to the N- and Q-blocks. The terms V_N and V_Q in the denominator were calculated using the conformation-specific R_g values for the N- and Q-blocks. In the definition of V_I , r_c is the distance between the centers-of-mass of the N- and Q-blocks whereas $R_{g,N}$ and $R_{g,Q}$ are the conformation-specific radii of gyration calculated over the atoms of the N- and Q-blocks, respectively. When $d_A = 0$, the spherical envelopes of the N- and Q-blocks do not intersect. The maximal degree of adsorption between a 17-residue N-terminal stretch and a globular polyQ domain is ~ 0.4 .

The values for d_A were calculated from ABSINTH-based all atom simulations of monomeric variants of different N-Q sequences. The abscissa in Figure 3 corresponds to the probabilities of forming dimers of N-blocks as autonomous units. These results were also extracted from ABSINTH-based all atom simulations with pairs of N-block molecules. The results summarized in Figure 3 make several points: The N-blocks show negligibly low likelihoods for self-interactions when compared to the high probability of self-associations between pairs of Q₄₀ molecules. By scrambling the sequence

of N17, we were able to design sequences that adsorb more strongly (N17_s) and more weakly (N17_w) to the polyQ domain when compared to the sequence of N17 that is drawn from the N-terminus of the huntingtin protein. Since a value of $d_A = 0$ cannot be achieved with the composition of N17, we designed a synthetic polyampholytic sequence, Glu-(Lys-Glu)₈ denoted as E(KE)₈, that helps us to achieve zero adsorption between the N- and Q-blocks. The four N-Q sequences allow us to titrate the effects of varying d_A on the sequence-encoded bias toward linear aggregates.

2. Choice of resolution

To test the hypothesis that the degree of adsorption between N- and Q-blocks modulates the bias towards linear aggregates, we need to understand the interplay between sequence-specific properties of flanking sequences and polyQ-mediated aggregation. Specifically, we seek a model that maintains sequence-specific properties of the flanking sequences while enabling simulations of $O(10^3)$ molecules in order to study the early stages of polyQ-mediated aggregation.

FIG. 2. N-Q sequences used for this study. N-block denotes different 17-residue N-terminal sequences and Q-block denotes a polyQ tract with 40 Gln residues. N17 denotes the wild type sequence (UniProt ID: P42858). Sequences of the N-block were designed in order to modulate the amino acid sequence (N17_w and N17_s) or to modulate the amino acid composition (E(KE)₈). Designs that modulate amino acid sequence maintain the wild type N17 composition but scramble the sequence. The choice of the polyampholytic sequence E(KE)₈ allowed for examination of N-block properties that could not be accessed using the N17 composition. In the sequences, hydrophobic residues are in black, polar residues are in green, and positively and negatively charged residues are shown in blue and red, respectively.

Experimental data and atomistic simulation results show that polyQ constructs adopt globular conformations in aqueous solutions.⁷⁹ Previous simulations have shown that $R_g = R_0 N^\nu$, for polyQ tracts. Here, R_g refers to the ensemble averaged radii of gyration, $\nu = 0.33$ for globules, and the pre-factor is $R_0 = 3.0$ Å. We analyzed the results from all atom simulations for $T = 315$ K to establish that the Q-block adopts globular conformations in the context of N-Q sequences. Further, in all of the N-Q constructs, the values of R_g calculated over the Q-blocks are concordant with a scaling exponent of $\nu = 0.33$.

Since the Q-block maintains its preference for globular conformations in all N-Q sequences examined here, we used the following architecture for the coarse-grained model: the residues of the N-block are modeled as residue beads whereas the residues of the Q-block are lumped together as a single colloidal bead (Figure 4). This choice allows for sequence-specific properties of the flanking sequences

to be maintained while enabling significant computational efficiency in simulations of $O(10^3)$ N-Q molecules.

In order to justify our choice of resolution, we deployed a network-based approach to provide an unbiased assessment of groups of residues that interact preferentially among themselves as opposed to interacting across groups. The idea is that residues that prefer to interact among themselves in the all atom simulations can be treated as a single entity upon coarse-graining. In network-based approaches, such groups are referred to as *communities*.^{109,110} The overall strategy, which is adapted from the work of Sethi *et al.*,¹¹¹ is as follows: each conformation drawn from the ensemble of conformations generated in an atomistic simulation is converted to a network of nodes and edges. The nodes are positions of C_α atoms and edges are drawn between pairs of nodes that are within 13 Å of each other. Communities based on these networks are determined using the Girvan-Newman algorithm combined with the network modularity score, Q . The final community structure for each conformation (network) is taken as the sub-network that yielded the highest Q -score.⁷³

Panels (a)–(d) in Figure 5 summarize the results obtained from analysis of the community structures for all N-Q constructs. Specifically, each panel plots the probability that two residues are observed to be in the same community. Hotter

FIG. 3. Degree of adsorption (d_A) to Q₄₀ versus the probability of forming N-block dimers for each N sequence. Results were extracted from at least 3 independent all atom simulations of monomers and monomer-dimer equilibria simulations, respectively. The dotted orange line quantifies the probability of Q₄₀ forming dimers. Dimers are defined by any two residues of differing molecules that are less than or equal to 3.5 Å apart. Insets show representative snapshots for each N-Q sequence. The black and orange translucent spheres correspond to the R_g 's of the N- and Q-blocks, respectively. The overlap between translucent spheres serves as a visualization of d_A .

FIG. 4. Architecture of the coarse-grained model compared to the all atom model for N17-Q₄₀. For the coarse-grained model, each residue in the N-block is modeled as a single bead, whereas all the residues of Q-block are modeled as a single colloidal bead. Here, Gln residues are in orange, hydrophobic residues are in black, positively and negatively charged residues are in blue and red, respectively, and other polar residues are in green.

FIG. 5. Community analysis for each of the N-Q sequences. Conformations from all atom simulations were converted into networks wherein each C_{α} position was considered a node and a weighted edge $e_w = d_{ij}$ was drawn between two nodes if the distance between the C_{α} positions of residues i and j , d_{ij} , was less than or equal to 13 Å. Communities based on these networks were determined using the Girvan-Newman algorithm combined with the network modularity score, Q . The final community structure for each conformation (network) was taken as the sub-network that yielded the highest Q . Panels (a)-(d) show the probability that any two residues are in the same community for E(KE)₈-Q₄₀, N17_w-Q₄₀, N17-Q₄₀, and N17_s-Q₄₀, respectively. The bottom left corner of each plot corresponds to probabilities of intra-N communities (residue numbers 1-17), whereas the top right corner of each plot corresponds to the probabilities of intra-Q communities (residue numbers 18-57). For E(KE)₈-Q₄₀, N17_w-Q₄₀, and N17-Q₄₀, it is clearly shown that intra-block communities are more favorable (white or blue colors on the plots versus hotter colors for intra-N and intra-Q communities). Additionally, within the N- and Q-blocks there are no well-defined sub-communities.

colors imply that two residues have a higher probability of being part of the same community, whereas cooler colors imply that two residues have a low probability of being part of the same community. For E(KE)₈-Q₄₀, N17_w-Q₄₀, and N17-Q₄₀, the community analysis clearly shows that distinct communities involve residues within the N- and Q-blocks, respectively. Figure 6 provides further quantitative evidence in support of the choice of the coarse-grained model. Here, we quantify the probability that a residue from block X (N- or Q-) belongs to the same community as a residue from block Y (N- or Q-). Even though the coupling increases between the N- and Q-blocks for sequences such as N17_s-Q₄₀, residues prefer to be in communities with other residues that are part of the same block. Given the low probabilities for N- and Q-block residues to be part of the same community and the lack of evidence for sub-communities within the Q-block, the coarse-grained resolution in which each N-block residue is modeled as a residue bead and the Q-block is modeled as a colloidal bead is justified for the N-Q constructs.

FIG. 6. Average probabilities that a residue from block X is in the same community as a residue from block Y. Here, block X and block Y can refer to either N- or Q-blocks. The first two columns show the average intra-block probabilities for each N-Q sequence and the third column shows the inter-block probabilities. For all N-Q sequences, the intra-block probabilities are greater than the inter-block probabilities (hotter versus cooler colors). This implies that even as the coupling increases between residues across the N- and Q-blocks, residues still prefer to be in communities with other residues within the same block.

B. Effective energy function for coarse-grained simulations

The effective energy function for the coarse-grained model of N-Q constructs is given by Equation (1). In the current example, w_C denotes the colloidal potential between Q-blocks. We used data from 18 independent all atom simulation for pairs of Q₄₀ molecules to obtain the functional form for the w_C potential. These data yielded a pair correlation function $g(r)$ at the target temperature for the distance r between the centers-of-mass of the polyQ molecules with two minima that correspond to the two categories of interactions between homopolymeric globules, viz., docking and entanglement. The functional form of w_C is a sum of two terms, a Mie potential (w_E) to model entanglements and a Gaussian potential (w_D) to model the docking of globules.¹¹² Explicitly,

$$w_C = w_E + w_D,$$

$$w_E = \left(\frac{\gamma_r}{\gamma_r - \gamma_a} \right) \left(\frac{\gamma_r}{\gamma_a} \right)^{\left(\frac{\gamma_a}{\gamma_r - \gamma_a} \right)} \varepsilon_E \left[\left(\frac{\sigma_E}{r} \right)^{\gamma_r} - \left(\frac{\sigma_E}{r} \right)^{\gamma_a} \right], \quad (13)$$

$$w_D = -\varepsilon_D \exp \left[-\frac{(r - r_d)^2}{2\delta_d^2} \right].$$

In Equation (13), ε_E is the well depth of the entanglement potential, σ_E is the distance r between the colloidal beads for which the entanglement potential is zero, ε_D is the well depth of the docking potential, r_d is the inter-particle separation at which the docking potential is minimized, and δ_d controls the width of the well.

C. Parameterization of uncoupled W_{eff} terms: W_b , W_{θ} , W_{ϕ} , and w_C

The target potential of mean force $w_{C,\text{target}}(r)$ was obtained by Boltzmann inversion of $g(r)$ such that $w_{C,\text{target}}(r) = -k_B T \ln[g(r)]$. This potential of mean force, extracted from the all atom simulations, was used in a non-linear regression procedure to obtain the parameters for w_C in the coarse-grained model. The values of γ_{rep} and γ_{att} were set to be equal to 6 and 2,

FIG. 7. Interaction potential, w_C , for pairs of Q_{40} molecules. The potential of mean force for pairs of Q_{40} molecules was extracted from 18 independent all atom simulations at 315 K (orange curve) and was fit to Equation (13) using a non-linear regression procedure (black curve); w_C captures the two modes of interactions observed for pairs of Q_{40} molecules, namely, entanglement and docking. Representative snapshots of entanglement and docking states are shown as insets. In these snapshots, Q_{40} molecules are shown in atomic detail in orange and grey. Additionally, translucent spheres that correspond to the radius of gyration, R_g , of each molecule are drawn as a visual aid to distinguish between entanglement (large overlap between spheres) and docking (limited, if any, overlap between spheres) states.

respectively. The final parameters from the regression analysis are as follows:

$$\begin{aligned} \epsilon_E &= 3.72 \text{ kcal-mol}^{-1}, \\ \sigma_E &= 0.32R_g \text{ Å}, \\ \epsilon_D &= 3.92 \text{ kcal-mol}^{-1}, \\ r_d &= 1.78R_g \text{ Å}, \\ \delta_d &= 5.94 \text{ Å}. \end{aligned} \quad (14)$$

In Equation (14), R_g refers to the ensemble-averaged radius of gyration for monomeric forms of Q_{40} molecules that we calculate from all atom simulations. Figure 7 shows the potential of mean force extracted from all atom simulations of pairs of Q_{40} molecules at 315 K—the target potential $w_{C,\text{target}}(r)$ —and a comparison of the fit to this potential obtained via non-linear regression analysis that leads to the parameters summarized in Equation (14).

Parameters for the bonded and dihedral angle terms were also determined using the Boltzmann inversion procedure. Specifically, for each N-Q construct, we first calculated the positions of the centers-of-mass for residues within the N-block and over all of the residues within the Q-block. The simulation results at 315 K were used to extract the probability distributions for each of the bond lengths, bond angles, and dihedral angles that define the coarse-grained model. The all atom probability distributions were then used to generate the bond length, bond angle, and dihedral angle parameters as described in Subsection II C.

D. Parameterization of W_{LJ}

As described in Subsection II B 2, the set of free parameters for W_{LJ} is chosen to be $\mathbf{p}_{vdW} \equiv [\sigma_{ChCh}, \epsilon_{SS}, \epsilon_{SC_1}, \epsilon_{SC_2}, \dots,$

$\epsilon_{SC_N}, \epsilon_{WW}, \epsilon_{WC_1}, \epsilon_{WC_2}, \dots, \epsilon_{WC_N}]$, where Ch denotes a charged residue bead, S denotes a strongly interacting residue bead, W denotes a weakly interacting residue bead, and C_N denotes the colloidal bead of type N . Given that the N-Q constructs have only one colloidal bead, $\mathbf{p}_{vdW}$ reduces to five free parameters. Explicitly, $\mathbf{p}_{vdW} \equiv [\sigma_{ChCh}, \epsilon_{SS}, \epsilon_{SC}, \epsilon_{WW}, \epsilon_{WC}] \equiv [p_1, p_2, p_3, p_4, p_5]$. All other parameters were prescribed as described in Subsection II B 2 or determined through Lorentz-Berthelot mixing rules.

1. Primary objective function

For each of the N-Q sequences, we strive to obtain an optimal set of parameters for $\mathbf{p}_{vdW}$ such that the primary objective function, Ω , given by Equation (9) is minimized. The choice of parameters should maximize the overlaps between all distance pairs separated by at least five bonds between the all atom and coarse-grained models.

2. Auxiliary objective function

For the N-Q constructs, the Q-block is modeled as a colloidal bead in the coarse-grained model. Accordingly, for a sequence with n beads, only $(n-5)$ out of the m pairwise distances involve the Q-block. In machine learning approaches, additional consideration is given to specific features within the primary data through the inclusion of an auxiliary objective function. We added an auxiliary objective function Ω_{aux} in a post-processing step to ensure that we give extra weight to the interactions between the N- and Q-blocks in generating the optimal values for $\mathbf{p}_{vdW}$. This weighting scheme is important since our goal is to determine how the degree of adsorption between N- and Q-blocks modulates Q-block driven aggregation. The auxiliary objective function is defined as

$$\begin{aligned} \Omega_{\text{aux}}(\mathbf{p}_{vdW}) &= \left[\frac{\Omega(\mathbf{p}_{vdW}) + \Omega_{N-Q}(\mathbf{p}_{vdW})}{2} \right], \\ \Omega_{N-Q}(\mathbf{p}_{vdW}) &= \left[1 - \frac{1}{n-3} \sum_{i=1}^{n-3} \left(\frac{2 - \sum_{j=1}^{n_{\text{bin}}} |\rho_{ij}^{\text{AA}} - \rho_{ij}^{\text{CG}}(\mathbf{p}_{vdW})|}{2} \right) \right]. \end{aligned} \quad (15)$$

The term Ω_{N-Q} is evaluated over distances between the Q-block colloidal bead and beads of the N-block involving residues that are separated by three or more bonds from the colloidal bead.

3. GPBO procedure for parameterization of $\mathbf{p}_{vdW}$

For each N-Q sequence of interest, we use a multi-step procedure that involves several generations (see Figure 8). The search space for finding the optimal parameters is systematically narrowed from one generation to the next and the procedure terminates if the range does not narrow any further. For a given range in parameter space, we apply the GPBO procedure in an iterative procedure to refine the model using observations to identify the most likely values of $\mathbf{p}_{vdW}$ that

FIG. 8. Flow chart of the multi-step procedure used for the parameterization of W_{LJ} . This procedure combines 500 Gaussian process Bayesian optimization (GPBO) iterations based on the primary objective function, Ω , with a post-processing filtering of the optimal $\mathbf{p}_{vdw}$ sets from 20 independent trials based on the auxiliary objective function, Ω_{aux} . The post-processing filtering leads to a refinement of the parameter search space used for the next generation. The parameterization procedure is terminated if at least 8 generations have been completed and the overall change in the parameter range between consecutive generations is less than or equal to 0.05. The initial parameter range for each parameter p_i was $p_1 = [8.5 \text{ \AA} \text{ } 10.0 \text{ \AA}]$, $p_2 = [0.05 \text{ kcal/mol } 1.5 \text{ kcal/mol}]$, $p_3 = [0.3 \text{ kcal/mol } 2.0 \text{ kcal/mol}]$, $p_4 = [0.01 \text{ kcal/mol } 0.3 \text{ kcal/mol}]$, and $p_5 = [0.05 \text{ kcal/mol } 0.75 \text{ kcal/mol}]$.

will minimize Ω . We used the MATLAB implementation of Gardner and Weinberger to calculate the posterior probabilities and generate refined parameters.⁶⁷ Each trial involves 500 GPBO iterations and each generation involves 20 independent trials. Within each generation, we collect 20 separate estimates for $\mathbf{p}_{vdw}$ from each of the 20 trials. We then apply a post-processing step to generate x subsets of values for $\mathbf{p}_{vdw}$ by choosing the parameters that minimize the auxiliary objective function Ω_{aux} . Here, $x = 20$ if the generation number is less than three and $x = 10$ if the generation number is between three and six. Finally, $x = 5$ if the generation number is greater than six. The subsets are used to narrow the parameter range for the next generation of the overall optimization procedure. The parameterization procedure is terminated if the post-processing step does not lead to a further narrowing of the parameter range as defined by a numerical threshold (see Figure 8). The optimal final values for $\mathbf{p}_{vdw}$ are drawn from the final subset and correspond to the parameters that lead to the lowest value of Ω_{aux} . Figure 9 shows how the parameter space is explored in a single trial that involves 500 iterations of the GPBO procedure. Table I shows the final values for $\mathbf{p}_{vdw}$ that we obtain for each of the four N-Q sequences. Figure 10 summarizes the evolution of Ω and Ω_{aux} for N17-Q40 for each generation of the optimization procedure. These plots demonstrate the minimization of the two objective functions as the optimization progresses.

In order to interpret the relatively small changes to the magnitudes of Ω and Ω_{aux} , we performed a sensitivity analysis

to assess how the objective functions change as we perturb individual parameters from the values that we obtain using the optimization procedure. Specifically, for a given parameter p_i ($i = 1, \dots, 5$), each panel in Figure 11 shows how the different objective functions, viz., Ω , Ω_{N-Q} , and Ω_{aux} change as a function of changes to individual p_i values. The objective functions are most sensitive to changes in parameters p_1 – p_3 whereas changes to p_4 and p_5 do not have a material impact on the objective functions. Deviation from $\min(\Omega)$ is most pronounced for p_1 and p_2 . This suggests that p_1 and p_2 are important for capturing N-N intramolecular interactions correctly. Such a result is reasonable given that p_1 defines the σ -value for charged beads and p_2 defines the ϵ -value between strongly interacting residue beads within the N-block. For p_3 , positive fraction changes lead to similar Ω values but lead to appreciable changes in Ω_{N-Q} . This result demonstrates the importance of including Ω_{aux} in a post-processing step to focus the optimization procedure to regions in the parameter space that minimize Ω as well as Ω_{aux} . The additional weight given to N-Q interactions highlights a central feature of machine learning approaches, whereby the learning of specific features that are relevant to the question of interest leads to systematic improvement of the objective function.

Prior to using the coarse-grained model for comparative assessments of the influence of different N-block sequences on polyQ aggregation, we tested the quality of final coarse-grained models for each sequence. Figure 12 summarizes results comparing the conformational properties obtained

FIG. 9. Visualization of the exploration of the 2-dimensional parameter space (p_2 , p_3) in a single trial for N17-Q₄₀. This single trial contained 500 GPBO iterations. Although the GPBO procedure explored the 5-dimensional parameter space (p_1 , p_2 , p_3 , p_4 , p_5), visualization of the GPBO procedure is only shown for the parameters p_2 and p_3 since these were the most sensitive parameters. (a) The binning in both dimensions was 0.2 kcal/mol. The color of each square indicates the average objective score, $\Omega(\mathbf{p}_{vdw})$, for iterations in which (p_2 , p_3) fall within this bin. Light blue colors indicate low average $\Omega(\mathbf{p}_{vdw})$ scores. The number within each square indicates the first iteration in which (p_2 , p_3) falls within this bin. The white squares indicate regions of the (p_2 , p_3) parameter space that were not sampled within the 500 GPBO iterations. (b) Each dot corresponds to the (p_2 , p_3) parameter set at a given iteration. As the iteration number increases, the colors move from cool (dark blue) to hot (dark red). Many regions of the (p_2 , p_3) parameter space are sampled early on. However, a majority of the iterations sample a similar (p_2 , p_3) parameter space. This parameter space corresponds to the low average objective score region of (a).

using the coarse-grained model to results obtained from the all atom model at 315 K. The comparisons are shown for N17-Q₄₀ and E(KE)₈-Q₄₀.

For each conformation drawn from the all atom and coarse-grained ensembles, we calculated the eigenvalues of the gyration tensor. These eigenvalues (λ_i , $i = 1, 2, 3$) were used to compute the radius of gyration (R_g) and the asphericity δ as shown below,¹¹³

$$R_g = \sqrt{\lambda_1 + \lambda_2 + \lambda_3}, \quad (16)$$

$$\delta = 1 - 3 \left(\frac{\lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3}{(\lambda_1 + \lambda_2 + \lambda_3)^2} \right).$$

Figure 12 shows the joint distributions of R_g and δ values derived from all atom simulations and the coarse-grained model for N17-Q₄₀ and E(KE)₈-Q₄₀. In each case, the two-dimensional space is tiled into a 3×3 grid. Figure 12 also shows the integrated probabilities within each cell of the 3×3 grid that are quantified for the two models. These results demonstrate that the global conformational properties of the coarse-grained model are congruent with those of the all atom model. Importantly, the quality of the comparative assessments is favorable and equivalent for N-Q sequences with very different types of N-block sequences. The computational efficiency of coarse-grained simulations increases by at least three orders of magnitude vis-à-vis the ABSINTH-based all atom simulations. The increased efficiency combined with the verifiable accuracy of the model justifies the deployment of the model in simulations of aggregation that involves ca. 10^3 N-Q molecules.

V. COARSE-GRAINED SIMULATIONS OF N-Q AGGREGATION

For each N-Q sequence, we performed eight independent LD simulations based on the coarse-grained model with residue beads for each N-block residue and a single colloidal bead for the Q-block residues. Each of these isothermal-isochoric simulations was performed in cubic boxes, 65.7 nm to a side, with periodic boundary conditions at 315 K. There were 512 N-Q molecules in the simulation cell and this corresponds to supersaturated solutions with a concentration of 3 mM. Each simulation was run for 200 ns. The total simulation time across eight independent simulations is 1.6 μ s. For a given N-Q sequence, a 200 ns simulation takes four weeks to complete on a single core with compute nodes comprising of 8 Intel E5430 Central Processing Units that share 4 GB of RAM. When comparing the simulation time scales to those for *in vitro* experiments, it is worth noting that the diffusivity of colloidal beads is $O(10^2)$ larger than in experiments. This combined with the reduction in the number of degrees of freedom implies that the simulation time scales access processes that are relevant for the formation of precursors of large-scale aggregates.

Aggregates are distinguishable by the number of molecules, the overall sizes, shapes, and packing densities. Our focus is on comparative assessments of the morphologies of aggregates that are the precursors of large individual aggregates that form over long time scales either due to barrier-limited processes or as a consequence of phenomena such as Ostwald ripening.^{114–116} Given the finite sizes of the simulation systems, the aggregates of interest are those that

TABLE I. Final $\mathbf{p}_{vdw}$ values for each N-Q sequence.

N-Q sequence	p_1 (Å)	p_2 (kcal/mol)	p_3 (kcal/mol)	p_4 (kcal/mol)	p_5 (kcal/mol)
N17-Q ₄₀	9.62	1.15	1.23	0.06	0.43
N17 _w -Q ₄₀	8.54	0.41	0.66	0.02	0.65
N17 _s -Q ₄₀	9.53	0.58	1.53	0.02	0.53
E(KE) ₈ -Q ₄₀	9.50	...	...	...	...

FIG. 10. Evolution of the average $\Omega(\mathbf{p}_{\text{vdW},o})$ (a) and average $\Omega_{\text{aux}}(\mathbf{p}_{\text{vdW},o})$ (b) over subsequent generations of the optimization procedure for N17-Q40. Here, $\mathbf{p}_{\text{vdW},o}$ corresponds to the optimal parameter set, i.e., the set of parameters that yielded $\min(\Omega)$ for each of the 20 independent trials in the generation. Here, $\Omega(\mathbf{p}_{\text{vdW},o})$ and $\Omega_{\text{aux}}(\mathbf{p}_{\text{vdW},o})$ generally improve upon subsequent generations. However, comparison of panels (a) and (b) shows that $\Omega(\mathbf{p}_{\text{vdW},o})$ and $\Omega_{\text{aux}}(\mathbf{p}_{\text{vdW},o})$ are not always correlated. Specifically, generation 7 has one of the lowest $\Omega(\mathbf{p}_{\text{vdW},o})$ values, but an intermediate $\Omega_{\text{aux}}(\mathbf{p}_{\text{vdW},o})$ value. Panels (c) and (d) plot the average $\Omega(\mathbf{p}_{\text{vdW},o})$ and average $\Omega_{\text{aux}}(\mathbf{p}_{\text{vdW},o})$ versus the overall change in the parameter range between consecutive generations. Colors move from cool to hot as the generations increase. Panels (c) and (d) show that at generation 12 the overall change in the parameter range is less than or equal to 0.05 and thus the optimization procedure was terminated.

incorporate at least an order of magnitude of the molecules in the central simulation cell. We analyzed the morphologies of aggregates that comprise of at least nine molecules using the last 120 ns of each simulation.⁷³ The robustness of this approach was established by applying similar analyses to all eight independent simulations for each construct.

In order to test the hypothesis that the degree of adsorption between N- and Q-blocks modulates the bias towards linear aggregates, we determined the preference for linear aggregates for each N-Q construct. We applied methods derived from network-based analysis described in Section IV A to distinguish between compact and linear aggregate types

FIG. 11. Parameter sensitivity analysis for N17-Q40. Given the final values for $\mathbf{p}_{\text{vdW}}$ obtained from the optimization procedure, individual parameters were systematically varied in increments of 0.05, the fractional change in order to determine the sensitivity of each individual parameter on Ω , $\Omega_{\text{N-Q}}$, and Ω_{aux} . Here, fraction change is defined as $\left(\frac{p_i^{\text{test}} - p_i^{\text{final}}}{p_i^{\text{final}}}\right)$, where p_i^{final} is the optimal value for parameter p_i and p_i^{test} is the parameter used for the given simulation. Four independent 200 ns monomer simulations were conducted and the average objective scores over the independent simulations are plotted. For each subplot, an individual parameter (p_1 , p_2 , p_3 , p_4 , or p_5) was varied from its final value, while all other parameters were held fixed at their final value. The objective scores are found to be least sensitive to parameters p_4 and p_5 .

FIG. 12. Comparison of conformational properties between all atom and coarse-grained models for N17-Q40 and E(KE)₈-Q40. A monomer's size and shape can be captured by its R_g and asphericity, δ , respectively. If $\delta = 0$, then the monomer adopts globular conformations, whereas if $\delta = 1$, then the monomer adopts rod-like conformations. R_g and δ values were determined from the eigenvalues of the gyration tensor over three and four independent monomer simulations for the all atom and coarse-grained models, respectively. For the all atom model, the gyration tensors were calculated over the centers-of-mass of each N-block residue and the center-of-mass of the Q-block. This choice allowed for a direct comparison to the gyration tensors calculated over the positions of the centers of beads in the coarse-grained model. Two-dimensional histograms of R_g versus δ are plotted for N17-Q40 all atom (1.a), N17-Q40 coarse-grained (1.b), E(KE)₈-Q40 all atom (2.a), and E(KE)₈-Q40 coarse-grained (2.b) models. Panels (1.c) and (2.c) quantify the integrated probabilities for the all atom and coarse-grained models within each cell of the 2-dimensional R_g versus δ plots for N17-Q40 and E(KE)₈-Q40, respectively. Insets of (1.c) and (2.c) show representative snapshots for the all atom and coarse-grained models. Here, Gln is colored orange, other polar residues green, hydrophobic residues black, and positively and negatively charged residues red and blue, respectively. In both the all atom and coarse-grained models, N17-Q40 prefers compact/globular conformations with N17 adsorbed on Q40. In contrast, in both the all atom and coarse-grained models, E(KE)₈-Q40 samples larger and more elongated conformations compared to N17-Q40.

based on the number of polyQ communities. The packing density of these communities determines the morphologies of aggregates. Pairs of polyQ colloidal beads are in the same community if the distance between the beads is less than or equal to 7 Å. The centroids of each polyQ-community within an aggregate are used to calculate the size of the aggregate defined as its radius of gyration and denoted as R_A . The values of R_A and the average size of an individual polyQ-community defined as the average radius of gyration of communities within the aggregate and denoted as R_Q are related to each other via the parameter χ , which is the packing density of polyQ communities. The quantities R_A , R_Q , χ , and the number of polyQ communities n_Q follow the relation

$$\begin{aligned} \chi R_A^3 &= n_Q R_Q^3 \\ \Rightarrow R_A^3 &= \left(\frac{n_Q R_Q^3}{\chi} \right). \end{aligned} \quad (17)$$

The size of an aggregate is governed by the size, number, and packing density of polyQ communities. Setting $\chi = 0.5$ yields a threshold value for R_A that allows us to distinguish between compact versus linear aggregates in terms of the average size and number of individual polyQ communities n_Q within an aggregate. The threshold value of R_A takes the

form

$$R_t = (2n_Q)^{1/3} R_Q. \quad (18)$$

Aggregates are compact if $R_A \leq R_t$ and they are linear if $R_A > R_t$. The threshold values of R_t are useful for distinguishing between the morphologies of aggregates with similar or identical numbers of polyQ communities.

Figure 13 summarizes the properties of the aggregates for different constructs by quantifying the joint probability histograms in the two-parameter space of n_Q and R_A . The results are drawn from simulations results for polyQ molecules with and without different N-block sequences. In each panel of Figure 13, the red curve is a plot of R_t . For a given value of n_Q , linear and compact aggregates lie above and below the curve for R_t . For each construct, we integrated the probabilities for forming linear aggregates (P_{LA}) defined as those that lie above the reference R_t curve. The corresponding probability for forming compact aggregates is $(1 - P_{LA})$ and the parameter of interest is the ratio $S = \left(\frac{P_{LA}}{1 - P_{LA}} \right)$. In Figure 14, we plot S obtained from the coarse-grained simulations against d_A , the degree of adsorption between N- and Q-blocks that we quantified from our analysis of the all atom simulations for the four constructs with 17-residue N-blocks. The Pearson product moment correlation coefficient, which measures the

FIG. 13. Two-dimensional histograms of the number of polyQ communities, n_Q , within an aggregate versus the radius of gyration of the aggregate, R_A . The final 120 ns of eight independent simulations was analyzed for each N-Q sequence. Only aggregates with greater than eight molecules were considered for these analyses. The trends are found to be the same if the cutoff for aggregate size was set to be six or ten (data not shown). In each plot, the red curve defines the cutoff between compact and linear aggregates, R_t , for a given n_Q based on Equation (18). As can be observed from panel (b), the choice of this threshold separates the two distinct density regions of aggregate sizes, R_A , for $n_Q = 3$ and 4. N17_W-Q₄₀, N17-Q₄₀, and N17_S-Q₄₀ form linear aggregates, whereas E(KE)₈-Q₄₀ and Q₄₀ form compact aggregates.

linear correlation between S and d_A , is 0.98. The implication is that the degree of adsorption between the N- and Q-blocks determines the sequence-encoded extent of interactions between N- and Q-blocks and this in turn determines the bias toward forming linear aggregates.

Figure 15 summarizes the frequencies of forming aggregates with different numbers of polyQ communities, n_Q , that correspond to different morphologies. Each panel includes

FIG. 14. The average degree of adsorption between N- and Q-blocks from all atom simulations of N-Q monomers (d_A) versus the relative probability of forming linear aggregates in the coarse-grained model (S). Here, r denotes the Pearson product moment correlation coefficient. The probability of forming linear aggregates is highly correlated with the degree of adsorption between N- and Q-blocks from all atom monomer simulations of N-Q sequences.

representative structures of aggregates formed by different constructs and these pictures provide visual support for the results summarized in Figure 14. The aggregate distributions for N17_W-Q₄₀, N17-Q₄₀, and N17_S-Q₄₀ are bimodal with preferences for small compact aggregates (red bars, $n_Q = 1$ or 2) and linear aggregates (green bars). The preference for forming linear aggregates increases from N17_W-Q₄₀ to N17-Q₄₀ to N17_S-Q₄₀ and this correlates with an increase in d_A . In contrast, Q₄₀ and E(KE)₈-Q₄₀ prefer larger compact aggregates and the distributions are essentially monodisperse.

The observations from the coarse-grained simulations yield predictive expectations regarding the outcomes of kinetics experiments. We expect a smaller barrier and smaller lag times associated with forming large linear aggregates if the sequence in question shows a pronounced bias toward forming linear aggregates. Our results indicate that sequences lacking the N-block favor compact aggregates as shown in panel (e) of Figures 13 and 15. This is consistent with pronounced lag times that have been observed for forming linear, fibrillar aggregates for polyQ molecules.⁸¹ Our results also show that the N17-Q₄₀ sequence has a clear bias toward forming linear aggregates. These findings are also consistent with the recent experimental data, which show that N17 engenders a sequence-encoded bias for the rapid formation of linear aggregates.

We tested the sensitivity of aggregate distributions to the choice of parameters ($\mathbf{p}_{\text{vdW}}$). For this, we selected the optimal set of $\mathbf{p}_{\text{vdW}}$ values from five randomly selected generations. Figure S6⁷³ shows the optimal $\mathbf{p}_{\text{vdW}}$ set for each of the five generations for N17-Q₄₀. The results of seven independent coarse-grained simulations containing 512 molecules

FIG. 15. Probability of forming aggregates with different numbers of polyQ communities, n_Q , and different morphologies (compact versus linear) for (a) N17_W-Q₄₀, (b) N17-Q₄₀, (c) N17_S-Q₄₀, (d) E(KE)₈-Q₄₀, and (e) Q₄₀. The black vertical line delineates between compact versus linear morphologies. The abscissa quantifies the number of polyQ communities, i.e., the number of distinct clusters of Q-beads within an aggregate using a cutoff of 7 Å. For example, if $n_Q = 3$, then the Q-beads cluster into 3 distinct communities (see bars labeled 3 in both the compact (red) and linear (green) regions for representative aggregates). Communities can consist of one to $O(10)$ molecules given the ability of the Q-beads to entangle with each other. Red bars represent the probabilities of compact aggregates and green bars represent the probabilities of linear aggregates with a given n_Q . Probabilities of forming distinct morphologies and n_Q values were taken from the last 120 ns of eight independent coarse-grained simulations containing 512 molecules for each of the N-Q sequences. Only aggregates with at least nine molecules were considered for this analysis. Insets for each panel show representative aggregates for the most probable aggregate types.

at 750 μM were analyzed for parameter sets drawn from each randomly chosen generation. The resultant aggregate distributions are shown in Figure S7.⁷³ Each distribution was compared to the distribution resulting from the final optimal parameter set for N17-Q₄₀. To quantify the correlation between aggregate distributions obtained using “suboptimal parameters” and the best set of parameters, we quantified the Pearson product moment correlation coefficients between the reference distribution and the distributions obtained using the supposedly suboptimal set of parameters. These values range between 0.97 and 0.99 indicating that the final choice of $\mathbf{p}_{\text{vdW}}$ values is relatively insensitive to the generation number. Therefore, a single generation of the GPBO procedure would have been sufficient for parameterizing the coarse-grained model for N-Q constructs.

VI. DISCUSSION

A. Physical significance of the findings from coarse-grained simulations of block-copolymeric molecules with polyQ tracts

Two distinct models have been put forth for the observed effects of N17 on polyQ aggregation. According to the *helix*

bundling model, N17 modules form oligomeric helical bundles that enable an increase in the local concentration of polyQ domains.⁹² A subsequent rate-limiting step corresponds to the conversion of polyQ domains to beta-sheet conformations that leads to the formation of linear aggregates, although the details of this process are not fully defined. The helix bundling model is appealing for its conceptual simplicity. However, it has two weaknesses: First, the bundling of N17 helices is observed at concentrations that are at least two orders of magnitude larger than the saturation concentration of typical polyQ domains.⁸¹ Hence, it becomes difficult to explain how the weak self-association of N17 modules contributes to increasing the effective concentration of polyQ modules when the measured driving forces suggest the opposite scenario. Second, none of the published kinetics data indicates the presence of discernible lag times for the aggregation of N17-polyQ molecules. This would be expected given the significant barrier that should accompany conversion from oligomers lacking beta-sheet structure to linear aggregates that are rich in beta-sheet structure. An alternative is the *domain cross talk* model⁹⁷ that is not focused on the internal conformational properties of individual blocks but focuses instead on generic properties that is reminiscent of the physics of patchy colloids. The model has four components: First, unfolding and adsorption

of the N-block on the Q-block is driven by interactions between uncharged residues within the N-block.⁹⁶ Second, adsorption of the N-block engenders a disordered patchy colloidal architecture¹⁰⁶ for N17-polyQ molecules with the charged residues in the N-block being exposed to solvent.⁹⁶ Third, the patch limits random intermolecular encounters between polyQ modules. And fourth, the conformation of the N-block patch is not rigid and this enables intermolecular interactions between the N-block of one N17-polyQ molecule and the Q-block of another N17-polyQ molecule.

The physics of patchy colloids is directly relevant for explaining the observed aggregation properties of N17-polyQ molecules. Isotropic attractions between uniformly sticky colloidal beads will lead to compact, hexagonally close-packed aggregates as opposed to linear aggregates. In patchy colloids, the interactions between surface adsorbed patches are repulsive. The morphologies of aggregates formed by patchy colloids are determined by a combination of the range of the repulsive interactions vis-à-vis the attractions, the number of patches per particle, and the relative sizes of repulsive patches versus those of colloidal beads.¹⁰⁵

Our findings for sequences with designed N-blocks support the transferability of the tenets of patchy colloids to explain the observations for N-Q sequences. The approach of using designed N-block sequences also results in a set of testable predictions. These predictions suggest that the polyampholytic N-block, E(KE)₈, will engender a strong preference for distinct architectures at the monomer level whereby the N-block forms an extended tail and the polyQ globule forms a compact head. The degree of adsorption between the N- and Q-blocks is zero for this sequence. Consequently, these sequences form a homogeneous distribution of compact as opposed to linear aggregates in simulations based on the coarse-grained model. Therefore, the aggregates formed by E(KE)₈-polyQ should be monodisperse species with solvent exposed polyampholytic tails resembling the architectures for reverse micelles. The conversion to large, linear fibrillar aggregates should be the slowest for such sequences. In contrast, the architectures of aggregates formed by N17_W-polyQ and N17_S-polyQ should show weaker and stronger intrinsic biases, respectively, toward linear aggregates when compared to the N17-polyQ sequence. We predict that these differences should be reflected in the measured driving forces for forming linear versus spherical aggregates and in the time scales for the conversion from spherical to linear aggregates. Importantly, these predictions require simulations that can access the relevant spatial scales and this is afforded by our methodology for coarse-graining that affords computational efficiency while preserving the ability to capture the effects of sequence-specific interactions.

B. Improvements and generalizations to the CAMELOT approach

It is imperative that we extend our current simulations to access at least $O(10^4)$ molecules to avoid finite size artifacts and make direct contact with the spatial scales that prevail in aggregation experiments. Such simulations combined with systematic concentration titrations will enable predictions of

phase diagrams for block-copolymeric sequences. In this work, we used a hybrid model that includes residue-specific beads for the N-block and a single colloidal bead for the residues of the Q-block. The interactions between Q-blocks were modeled using a colloidal potential that captures the two main features of the self-associations of homopolymeric globules, namely, entanglement and docking. The pair potential for colloidal beads can be generalized through additional considerations: It is conceivable that three-body effects will modulate the balance between docking and entanglement that are captured using pair potentials. We should be able to test for the presence of cooperative effects by parameterizing the W_C term against all atom simulations that include more than two polyQ molecules. It is conceivable that the barrier for converting to linear aggregates is lowered at larger length scales due to conformational fluctuations that enable the modulation of the range of attractive interactions between colloidal beads while also altering the balance between entanglements and docking. The CAMELOT approach incorporates the necessary versatility for modeling such emergent properties. This can be achieved by fixing the number of molecules and performing a series of independent simulations, each characterized by a quenched degree of adsorption and distinct functional forms for colloidal potentials to model differences in the range of attractive interactions and the energy scales for entanglements versus docking. The results of each set of independent simulations can be stitched together using appropriate methods to generate predictive phase diagrams to quantify how a specific block-copolymeric sequence encodes its aggregation landscape as a function of the degree of adsorption between N- and Q-blocks and features of the colloidal potentials.

The hybrid resolution used in this work has general applicability for other block-copolymeric systems involved in aggregation and phase separation (see Figure 1). This includes sequences with globular folded domains as established in Section II.⁷³ Our analysis of communities of residues gleaned from all atom simulation results for a sequence with a disordered bristle tethered to an ordered SH3 domain shows that a hybrid resolution with residue beads for the disordered bristle and a colloidal bead for the SH3 domain would be justifiable for efficient simulations of the phase behavior of these classes of molecules (see Figure 1). The colloidal bead could have specific residue beads adsorbed on its surface as patches or the interactions between SH3 domains could be modeled using directional potentials between colloidal beads. It is also possible to design hybrid resolutions with fewer residues per colloidal bead, thus interpolating between the purely bead per residue resolution on one end and the single colloidal bead for the entire sequence on the other end. Such interpolations would leverage the blob concept introduced by de Gennes,¹¹⁷ which has proven to be useful in understanding the phase behavior of synthetic polymers.¹¹⁸

The CAMELOT based coarse-grained simulations provide insights regarding the distributions of molecular aggregates. A particularly important question pertains to the conformational properties of individual molecules within aggregates and the barriers for interconversion between different conformational basins. These questions should be addressable using mixed resolution generalizations of the

CAMELOT approach that combine atomistic descriptions of an individual chain interacting with an explicitly modeled bath of coarse-grained molecules. The coupling between the all atom representation and the coarse-grained molecules in the bath can be modeled via a direct mapping between the all atom representation and the single bead per residue model. This would represent a generalized adaptation of the multiscale coarse-graining approach of Izvekov and Voth.¹¹⁹ Bottlenecks in conformational sampling for mixed resolution models can be overcome by adapting recent generalizations of the resolution exchange approach¹²⁰ that have been proposed by Chen and coworkers.^{121,122}

C. General connections between coarse-graining and machine learning

The CAMELOT approach is guided by the precepts that underlie force matching and other information-driven algorithms for systematic coarse-graining. Our use of machine learning aided methods is reminiscent of recent successes in reparameterizing molecular mechanics forcefields.¹²³ We envisage the prospect of using data from multiple resolutions aided by deep learning algorithms that use data from simulations based on multiple resolutions.^{124–126} The tenets of deep learning can be anchored in the framework of the renormalization group theory.¹²⁷ This should allow automated discoveries of the coupling between and correlated fluctuations of collective degrees of freedom that determine the sequence-specific phase diagrams for different block-copolymeric systems. As noted above, we would also like to be able to mix distinct resolutions with each other in order to extract finer structures, specifically conformational changes that are compatible with coarse-grained results. Such approaches will require a well-defined structure for the coarse-grained potentials across all resolutions and the functional form proposed for W_{eff} and the overall structure of the CAMELOT algorithm would appear to provide the necessary ingredients.

D. Final summary

We have developed a new method, named CAMELOT, that combines Boltzmann inversion, non-linear regression, and machine learning to develop a coarse-grained model whereby the appropriate resolution and optimal parameters are, respectively, inferred and learned from detailed all atom simulations. The CAMELOT algorithm is available as a general toolbox that can be plugged into generic molecular modeling suites such as the CAMPARI package (<http://campari.sourceforge.net>). This enables the automated transfer of information from the all atom simulations into a suitable coarse-grained model that is then run using the LAMMPS package (<http://lammps.sandia.gov>). In this work, we introduced the general framework for the CAMELOT approach and demonstrated its utility by deploying a hybrid resolution for coarse-grained simulations of molecules with polyQ tracts. The results obtained from the coarse-grained simulations are in accord with the hypothesis that adsorption of flanking sequences on the polyQ domain generates patchy colloid

architectures. The deployment of the proposed approach to study aggregation and phase separation in a range of block-copolymeric sequences should enable improved understanding of the sequence determinants of phase behavior such as the biases for liquids comprised of random coils versus solids based on linear, fibrillar aggregates.

ACKNOWLEDGMENTS

The National Institutes of Health supported this work through Grant No. 5R01NS056114. We acknowledge usage of computational resources provided by the Center for High Performance Computing at Washington University in St. Louis. We are grateful to Jacob Gardner and Kilian Weinberger for introducing us to Bayesian optimization methods and providing the initial code for implementing the GPBO part of this work. We thank Rahul Das, Alex Holehouse, and Jason Wagoner for helpful discussions during the early stages of this work. Gregory Bowman, Jianhan Chen, Gerhard Hummer, and Andreas Vitalis contributed several insightful critiques and suggestions that helped in revising the manuscript.

- ¹H. B. Schmidt and D. Goerlich, *Elife* **4**, e04251 (2015).
- ²L. H. Beun, I. M. Storm, M. W. T. Werten, F. A. de Wolf, M. A. C. Stuart, and R. de Vries, *Biomacromolecules* **15**, 3349 (2014).
- ³D. Ando, R. Zandi, Y. W. Kim, M. Colvin, M. Rexach, and A. Gopinathan, *Biophys. J.* **106**, 1997 (2014).
- ⁴L. Li, Z. Tong, X. Jia, and K. L. Kiick, *Soft Matter* **9**, 665 (2013).
- ⁵Z. Ge and S. Liu, *Chem. Soc. Rev.* **42**, 7289 (2013).
- ⁶R. L. DiMarco and S. C. Heilshorn, *Adv. Mater.* **24**, 3923 (2012).
- ⁷M. Humenik, M. Magdeburg, and T. Scheibel, *J. Struct. Biol.* **186**, 431 (2014).
- ⁸W. Huang, A. Rollett, and D. L. Kaplan, *Expert Opin. Drug Delivery* **12**, 779 (2015).
- ⁹O. Tokareva, M. Jacobsen, M. Buehler, J. Wong, and D. L. Kaplan, *Acta Biomater.* **10**, 1612 (2014).
- ¹⁰O. S. Rabotyagova, P. Cebe, and D. L. Kaplan, *Biomacromolecules* **12**, 269 (2011).
- ¹¹R. Valluzzi, S. Winkler, D. Wilson, and D. L. Kaplan, *Philos. Trans. R. Soc., B* **357**, 165 (2002).
- ¹²N. Srinivasan and S. Kumar, *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **4**, 204 (2012).
- ¹³S. Rauscher and R. Pomes, in *Fuzziness: Structural Disorder in Protein Complexes*, edited by M. Fuxreiter and P. Tompa (Springer Science & Business Media, 2012), p. 159.
- ¹⁴A. K. Buell, C. Galvagnion, R. Gaspar, E. Sparr, M. Vendruscolo, T. P. J. Knowles, S. Linse, and C. M. Dobson, *Proc. Natl. Acad. Sci. U. S. A.* **111**, 7671 (2014).
- ¹⁵R. V. Pappu, X. Wang, A. Vitalis, and S. L. Crick, *Arch. Biochem. Biophys.* **469**, 132 (2008).
- ¹⁶S. Treusch and S. Lindquist, *J. Cell Biol.* **197**, 369 (2012).
- ¹⁷J. A. Toretsky and P. E. Wright, *J. Cell Biol.* **206**, 579 (2014).
- ¹⁸C. Lee, P. Occhipinti, and A. S. Gladfelter, *J. Cell Biol.* **208**, 533 (2015).
- ¹⁹S. C. Weber and C. P. Brangwynne, *Curr. Biol.* **25**, 641 (2015).
- ²⁰S. Elbaum-Garfinkle, Y. Kim, K. Szczepaniak, C. C.-H. Chen, C. R. Eckmann, S. Myong, and C. P. Brangwynne, *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7189 (2015).
- ²¹L. Chiu Fan, C. P. Brangwynne, J. Gharakhani, A. A. Hyman, and F. Julicher, *Phys. Rev. Lett.* **111**, 088101 (2013).
- ²²C. P. Brangwynne, *J. Cell Biol.* **203**, 875 (2013).
- ²³V. N. Uversky, I. M. Kuznetsova, K. K. Turoverov, and B. Zaslavsky, *FEBS Lett.* **589**, 15 (2015).
- ²⁴A. Patel, H. O. Lee, L. Jawerth, S. Maharana, M. Jahnel, M. Y. Hein, S. Stoyanov, J. Mahamid, S. Saha, T. M. Franzmann, A. Pozniakovski, I. Poser, N. Maghelli, L. A. Royer, M. Weigert, E. W. Myers, S. Grill, D. Drechsel, A. A. Hyman, and S. Alberti, "A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation," *Cell* **162**(5), 1066 (2015).
- ²⁵N. Srinivasan, M. Bhagawati, B. Ananthanarayanan, and S. Kumar, *Nat. Commun.* **5**, 5145 (2014).

- ²⁶L. Malinovska, S. Palm, K. Gibson, J. M. Verbavatz, and S. Alberti, "Dictyostelium discoideum has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation," *Proc. Natl. Acad. Sci.* **112**(20), E2620 (2015).
- ²⁷I. Kwon, S. Xiang, M. Kato, L. Wu, P. Theodoropoulos, T. Wang, J. Kim, J. Yun, Y. Xie, and S. L. McKnight, *Science* **345**, 1139 (2014).
- ²⁸J. Lai, C. H. Koh, M. Tjota, L. Pieuchot, V. Raman, K. B. Chandrababu, D. Yang, L. Wong, and G. Jedd, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 15781 (2012).
- ²⁹M. Kato, T. W. Han, S. Xie, K. Shi, X. Du, L. C. Wu, H. Mirzaei, E. J. Goldsmith, J. Longgood, J. Pei, N. V. Grishin, D. E. Frantz, J. W. Schneider, S. Chen, L. Li, M. R. Sawaya, D. Eisenberg, R. Tycko, and S. L. McKnight, *Cell* **149**, 753 (2012).
- ³⁰T. W. Han, M. Kato, S. Xie, L. C. Wu, H. Mirzaei, J. Pei, M. Chen, Y. Xie, J. Allen, G. Xiao, and S. L. McKnight, *Cell* **149**, 768 (2012).
- ³¹R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu, *Chem. Rev.* **114**, 6589 (2014).
- ³²V. N. Uversky, *Biotechnol. J.* **10**, 356 (2015).
- ³³N. K. Li, F. G. Quiroz, C. K. Hall, A. Chilkoti, and Y. G. Yingling, *Biomacromolecules* **15**, 3522 (2014).
- ³⁴R. K. Das, K. M. Ruff, and R. V. Pappu, *Curr. Opin. Struct. Biol.* **32**, 102 (2015).
- ³⁵K. Binder, M. Muller, P. Virnau, and L. G. MacDowell, in *Advanced Computer Simulation Approaches for Soft Matter Sciences I*, edited by C. Holm and K. Kremer (Springer, 2005), p. 1.
- ³⁶T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen, *Phys. Chem. Chem. Phys.* **11**, 1869 (2009).
- ³⁷E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodriguez-Ropero, and N. F. A. van der Vegt, *Soft Matter* **9**, 2108 (2013).
- ³⁸Y. Li, B. C. Abberton, M. Kroeger, and W. K. Liu, *Polymers* **5**, 751 (2013).
- ³⁹W. G. Noid, *J. Chem. Phys.* **139**, 090901 (2013).
- ⁴⁰S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, *J. Phys. Chem. B* **111**, 7812 (2007).
- ⁴¹P. Kar, S. M. Gopal, Y. M. Cheng, A. Predeus, and M. Feig, *J. Chem. Theory Comput.* **9**, 3769 (2013).
- ⁴²T. Frembgen-Kesner, C. T. Andrews, S. Li, N. Nguyen Anh, S. A. Shubert, A. Jain, O. J. Olayiwola, M. R. Weishaar, and A. H. Elcock, *J. Chem. Theory Comput.* **11**, 2341 (2015).
- ⁴³C. T. Andrews and A. H. Elcock, *J. Chem. Theory Comput.* **10**, 5178 (2014).
- ⁴⁴B. van Hoof, A. J. Markvoort, R. A. van Santen, and P. A. J. Hilbers, *Biophys. J.* **100**, 309a (2011).
- ⁴⁵T. C. Moore, C. R. Iacovella, and C. McCabe, *J. Chem. Phys.* **140**, 224104 (2014).
- ⁴⁶M. Seeger, *J. Int. Neural Syst.* **14**, 69 (2004).
- ⁴⁷A. Davtyan, J. F. Dama, G. A. Voth, and H. C. Andersen, *J. Chem. Phys.* **142**, 154104 (2015).
- ⁴⁸L. Lanyuan, J. F. Dama, and G. A. Voth, *J. Chem. Phys.* **139**, 121906 (2013).
- ⁴⁹J. F. Dama, A. V. Sinititsky, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth, *J. Chem. Theory Comput.* **9**, 2466 (2013).
- ⁵⁰R. D. Hills, Jr., L. Lu, and G. A. Voth, *PLoS Comput. Biol.* **6**, e1000827 (2010).
- ⁵¹L. Pu, S. Qiang, H. Daume III, and G. A. Voth, *J. Chem. Phys.* **129**, 214114 (2008).
- ⁵²S. Izvekov, J. M. J. Swanson, and G. A. Voth, *J. Phys. Chem. B* **112**, 4711 (2008).
- ⁵³W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, *J. Phys. Chem. B* **111**, 4116 (2007).
- ⁵⁴J. W. Chu and G. A. Voth, *Biophys. J.* **90**, 1572 (2006).
- ⁵⁵J. W. Chu, S. Izvekov, and G. A. Voth, *Mol. Simul.* **32**, 211 (2006).
- ⁵⁶Y. Wang and G. A. Voth, *J. Phys. Chem. B* **114**, 8735 (2010).
- ⁵⁷T. D. Hone, S. Izvekov, and G. A. Voth, *J. Chem. Phys.* **122**, 054105 (2005).
- ⁵⁸S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, *J. Chem. Phys.* **120**, 10896 (2004).
- ⁵⁹J. F. Rudzinski and W. G. Noid, *J. Phys. Chem. B* **118**, 8295 (2014).
- ⁶⁰J. W. Mullinax and W. G. Noid, *J. Phys. Chem. C* **114**, 5661 (2010).
- ⁶¹J. W. Mullinax and W. G. Noid, *Phys. Rev. Lett.* **103**, 198104 (2009).
- ⁶²A. Chaimovich and M. S. Shell, *J. Chem. Phys.* **134**, 094112 (2011).
- ⁶³A. Chaimovich and M. S. Shell, *Phys. Rev. E* **81**, 060104(R) (2010).
- ⁶⁴M. S. Shell, *J. Chem. Phys.* **129**, 144108 (2008).
- ⁶⁵R. K. Das and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13392 (2013).
- ⁶⁶M. Auton and D. W. Bolen, *Methods Enzymol.* **428**, 397 (2007).
- ⁶⁷J. Gardner, M. Kusner, Z. Xu, K. Weinberger, and J. Cunningham, in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (JMLR Workshop and Conference Proceedings, 2014), p. 937.
- ⁶⁸A. Mittal, R. K. Das, A. Vitalis, and R. V. Pappu, in *Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods*, edited by M. Fuxreiter (CRC Press, Boca Raton, FL, 2015).
- ⁶⁹A. Vitalis and R. V. Pappu, *J. Comput. Chem.* **30**, 673 (2009).
- ⁷⁰A. Radhakrishnan, A. Vitalis, A. H. Mao, A. T. Steffen, and R. V. Pappu, *J. Phys. Chem. B* **116**, 6862 (2012).
- ⁷¹A. Vitalis and R. V. Pappu, *Annu. Rep. Comput. Chem.* **5**, 49 (2009).
- ⁷²A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8183 (2010).
- ⁷³See supplementary material at <http://dx.doi.org/10.1063/1.4935066> for supplementary figures and additional analysis as well as description of methods.
- ⁷⁴C. C. Lee, R. H. Walters, and R. M. Murphy, *Biochemistry* **46**, 12810 (2007).
- ⁷⁵E. Landrum and R. Wetzel, *J. Biol. Chem.* **289**, 10254 (2014).
- ⁷⁶R. Wetzel, *J. Mol. Biol.* **421**, 466 (2012).
- ⁷⁷F. O. Walker, *Lancet* **369**, 218 (2007).
- ⁷⁸M. W. Becher, J. A. Kotzuk, A. H. Sharp, S. W. Davies, G. P. Bates, D. L. Price, and C. A. Ross, *Neurobiol. Dis.* **4**, 387 (1998).
- ⁷⁹S. L. Crick, M. Jayaraman, C. Frieden, R. Wetzel, and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16764 (2006).
- ⁸⁰A. S. Holehouse, K. Garai, N. Lyle, A. Vitalis, and R. V. Pappu, *J. Am. Chem. Soc.* **137**, 2984 (2015).
- ⁸¹S. L. Crick, K. M. Ruff, K. Garai, C. Frieden, and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20075 (2013).
- ⁸²T. R. Serio, A. G. Cashikar, A. S. Kowal, G. J. Sawicki, J. J. Mosleh, L. Serpell, M. F. Arnsdorf, and S. L. Lindquist, *Science* **289**, 1317 (2000).
- ⁸³R. H. Walters and R. M. Murphy, *J. Mol. Biol.* **412**, 505 (2011).
- ⁸⁴A. Vitalis and R. V. Pappu, *Biophys. Chem.* **159**, 14 (2011).
- ⁸⁵P. R. ten Wolde and D. Frenkel, *Science* **277**, 1975 (1997).
- ⁸⁶M. D. Tobelmann and R. M. Murphy, *Biophys. J.* **100**, 2773 (2011).
- ⁸⁷M. L. Duennwald, S. Jagadish, P. J. Muchowski, and S. Lindquist, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11045 (2006).
- ⁸⁸B. Sahoo, D. Singer, R. Kodali, T. Zuchner, and R. Wetzel, *Biochemistry* **53**, 3897 (2014).
- ⁸⁹R. Mishra, M. Jayaraman, B. P. Roland, E. Landrum, T. Fullam, R. Kodali, A. K. Thakur, I. Arduini, and R. Wetzel, *J. Mol. Biol.* **415**, 900 (2012).
- ⁹⁰R. Mishra, C. L. Hoop, R. Kodali, B. Sahoo, P. C. A. van der Wel, and R. Wetzel, *J. Mol. Biol.* **424**, 1 (2012).
- ⁹¹M. Jayaraman, R. Mishra, R. Kodali, A. K. Thakur, L. M. I. Koharudin, A. M. Gronenborn, and R. Wetzel, *Biochemistry* **51**, 2706 (2012).
- ⁹²M. Jayaraman, R. Kodali, B. Sahoo, A. K. Thakur, A. Mayasundari, R. Mishra, C. B. Peterson, and R. Wetzel, *J. Mol. Biol.* **415**, 881 (2012).
- ⁹³V. N. Sivanandam, M. Jayaraman, C. L. Hoop, R. Kodali, R. Wetzel, and P. C. A. van der Wel, *J. Am. Chem. Soc.* **133**, 4558 (2011).
- ⁹⁴S. Tam, C. Spiess, W. Auyeung, L. Joachimiak, B. Chen, M. A. Poirier, and J. Frydman, *Nat. Struct. Mol. Biol.* **16**, 1279 (2009).
- ⁹⁵A. K. Thakur, M. Jayaraman, R. Mishra, M. Thakur, V. M. Chellgren, I.-J. L. Byeon, D. H. Anjum, R. Kodali, T. P. Creamer, J. F. Conway, A. M. Gronenborn, and R. Wetzel, *Nat. Struct. Mol. Biol.* **16**, 380 (2009).
- ⁹⁶T. E. Williamson, A. Vitalis, S. L. Crick, and R. V. Pappu, *J. Mol. Biol.* **396**, 1295 (2010).
- ⁹⁷B. Kokona, Z. P. Rosenthal, and R. Fairman, *Biochemistry* **53**, 6738 (2014).
- ⁹⁸T. Vissers, F. Smallegange, G. Munao, Z. Preisler, and F. Sciortino, *J. Chem. Phys.* **140**, 144902 (2014).
- ⁹⁹Z. Preisler, T. Vissers, G. Munao, F. Smallegange, and F. Sciortino, *Soft Matter* **10**, 5121 (2014).
- ¹⁰⁰Z. Preisler, T. Vissers, F. Smallegange, G. Munao, and F. Sciortino, *J. Phys. Chem. B* **117**, 9540 (2013).
- ¹⁰¹F. Sciortino and E. Zaccarelli, *Curr. Opin. Solid State Mater. Sci.* **15**, 246 (2011).
- ¹⁰²B. Ruzicka, E. Zaccarelli, L. Zulian, R. Angelini, M. Sztucki, A. Moussaid, T. Narayanan, and F. Sciortino, *Nat. Mater.* **10**, 56 (2011).
- ¹⁰³F. Sciortino, A. Giacometti, and G. Pastore, *Phys. Chem. Chem. Phys.* **12**, 11869 (2010).
- ¹⁰⁴F. Sciortino, *Collect. Czech. Chem. Commun.* **75**, 349 (2010).
- ¹⁰⁵A. Giacometti, F. Lado, J. Largo, G. Pastore, and F. Sciortino, *J. Chem. Phys.* **132**, 174110 (2010).
- ¹⁰⁶E. Bianchi, P. Tarlaglia, E. Zaccarelli, and F. Sciortino, *J. Chem. Phys.* **128**, 144504 (2008).

- ¹⁰⁷G. Foffi and F. Sciortino, *J. Phys. Chem. B* **111**, 9702 (2007).
- ¹⁰⁸E. Bianchi, J. Largo, P. Tartaglia, E. Zaccarelli, and F. Sciortino, *Phys. Rev. Lett.* **97**, 168301 (2006).
- ¹⁰⁹M. E. J. Newman, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8577 (2006).
- ¹¹⁰E. A. Leicht and M. E. J. Newman, *Phys. Rev. Lett.* **100**, 118703 (2008).
- ¹¹¹A. Sethi, J. Tian, D. M. Vu, and S. Gnanakaran, *Biophys. J.* **103**, 748 (2012).
- ¹¹²K. M. Ruff, S. J. Khan, and R. V. Pappu, *Biophys. J.* **107**, 1226 (2014).
- ¹¹³M. O. Steinhauser, *J. Chem. Phys.* **122**, 094901 (2005).
- ¹¹⁴J. Zhang and M. Muthukumar, *J. Chem. Phys.* **130**, 035102 (2009).
- ¹¹⁵A. Baldan, *J. Mater. Sci.* **37**, 2171 (2002).
- ¹¹⁶A. Baldan, *J. Mater. Sci.* **37**, 2379 (2002).
- ¹¹⁷P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, London, 1979).
- ¹¹⁸T. Uematsu, C. Svanberg, and P. Jacobsson, *Macromolecules* **38**, 6227 (2005).
- ¹¹⁹S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).
- ¹²⁰E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, *Phys. Rev. Lett.* **96**, 028105 (2006).
- ¹²¹W. Zhang and J. Chen, *J. Chem. Theory Comput.* **10**, 918 (2014).
- ¹²²K. H. Lee and J. Chen, *J. Comput. Chem.* (published online 2015).
- ¹²³L.-P. Wang, T. J. Martinez, and V. S. Pande, *J. Phys. Chem. Lett.* **5**, 1885 (2014).
- ¹²⁴Y. Bengio, A. Courville, and P. Vincent, *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798 (2013).
- ¹²⁵R. Salakhutdinov and G. Hinton, *Neural Comput.* **24**, 1967 (2012).
- ¹²⁶H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, *J. Mach. Learn. Res.* **13**, 643 (2012).
- ¹²⁷P. B. Mehta and D. J. Schwab, e-print [arXiv:1410.3831](https://arxiv.org/abs/1410.3831) (2014).
- ¹²⁸S. Alberti, R. Halfmann, O. King, A. Kapila, and S. Lindquist, *Cell* **137**, 146 (2009).
- ¹²⁹R. Halfmann, S. Alberti, R. Krishnan, N. Lyle, C. W. O'Donnell, O. D. King, B. Berger, R. V. Pappu, and S. Lindquist, *Mol. Cell* **43**, 72 (2011).
- ¹³⁰P. Li, S. Banjade, H. C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q. X. Jiang, B. T. Nixon, and M. K. Rosen, *Nature* **483**, 336 (2012).