HHS Public Access

Author manuscript

J Phys Chem B. Author manuscript; available in PMC 2023 August 18.

Published in final edited form as:

J Phys Chem B. 2023 January 26; 127(3): 685–697. doi:10.1021/acs.jpcb.2c07672.

Network Hamiltonian Models for Unstructured Protein Aggregates, w/Application to γD-Crystallin

Elizabeth M. Diessner[†], **J. Alfredo Freites**[†], **Douglas J. Tobias**[†], **Carter T. Butts**[‡]

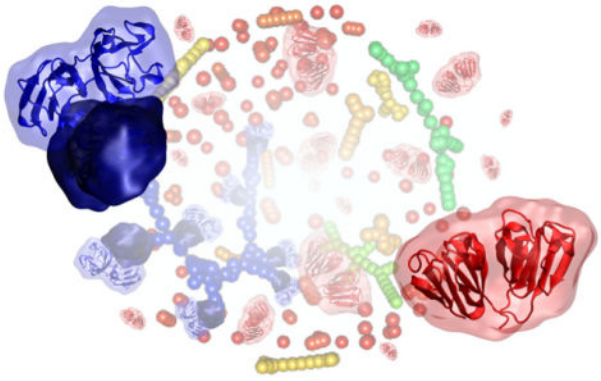
[†]Department of Chemistry, University of California, Irvine, CA 92697

[‡]Departments of Sociology, Statistics, Computer Science, and EECS,University of California, Irvine, CA 92697

Abstract

Network Hamiltonian models (NHMs) are a framework for topological coarse-graining of protein-protein interactions, in which each node corresponds to a protein, and edges are drawn between nodes representing proteins that are non-covalently bound. Here, this framework is applied to aggregates of γD-crystallin, a structural protein of the eye lens implicated in cataract disease. The NHMs in this study are generated from atomistic simulations of equilibrium distributions of wild-type and the cataract-causing variant W42R in solution, performed by Wong, E. K.; Prytkova, V.; Freites, J. A.; Butts, C. T.; Tobias, D. J. Molecular Mechanism of Aggregation of the Cataract-Related γD-Crystallin W42R Variant from Multiscale Atomistic Simulations. *Biochemistry* **2019**, 58 (35), 3691–3699. Network models are shown to successfully reproduce the aggregate size and structure observed in the atomistic simulation, and provide information about the transient protein-protein interactions therein. The system size is scaled from the original 375 monomers to a system of 10000 monomers, revealing a lowering of the upper tail of the aggregate size distribution of the W42R variant. Extrapolation to higher and lower concentrations is also performed. These results provide an example of the utility of NHMs for coarse-grained simulation of protein systems, as well as their ability to scale to large system sizes and high concentrations, reducing computational costs while retaining topological information about the system.

Graphical Abstract



buttsc@uci.edu, Phone: +1 (949) 824-8591.

Introduction

Protein aggregation is implicated in a wide range of diseases, including Alzheimer's, Parkinson's, type II diabetes, and cataract.[1,2] Aggregation can occur in a variety of biological environments, and in systems varying from intrinsically disordered proteins (IDPs) to proteins whose function depends on maintaining the stability of their native structure over the length of a human life-time (e.g., the structural crystallins of the human eye lens). The structures of the aggregates that result from this diverse set of proteins also vary, from the highly ordered amyloid fibrils associated with Alzheimers,[3] to the amorphous aggregates of crystallin that form cataracts.[4]

Molecular simulations of protein aggregation are important tools, along with experimental measurement, for probing the mechanics and interactions between proteins that lead to the formation of aggregates.[5,6] Monte Carlo (MC) simulations in particular have been used for studies of aggregation.[7,8] In regards to proteins, the convention is to simulate protein-protein interactions between rigid-body proteins with a single conformation.[9,10] To introduce some conformational flexibility, Wong et al.[11] studied the aggregation of γD-crystallin (γ-Dc) using the multiconformation Monte Carlo (mcMC) algorithm,[12,13] which employs a library of structures using conformations of the γ-Dc protein generated using single-protein and two-protein MD simulation trajectories. MC trial moves then are chosen among rigid-body translations, rotations, and conformation changes from the library of γ-Dc structures.

However, these simulations are still limited by the computational cost of modeling each conformation as part of an all-atom simulation. Coarse-graining these models in turn allows for simulation of longer time-scales, as well as increased complexity in terms of the number of molecules being observed in one simulation.[6] A wide range of coarse-graining approaches have been proposed for studying protein structure, dynamics, and interaction.[14]

Alternatively, models aimed at protein-protein interaction sometimes take a more radical approach. For instance, patchy sphere models represent an entire protein as a single sphere, with "patches" on the sphere surface that have unique interactions properties.[15] Patchy particles have been used for simulating self-assembly,[15,16] as well as protein phase behavior such as in the case of γ-Dc.[17–20]

While all of the above schemes work by modeling the physics of aggregate objects (chains, beads, etc.) within an explicit, Euclidean space, it is also possible to treat molecular systems *topologically*, representing systems in terms of patterns of interactions among subunits. For instance, Benson and Daggett[21] represent proteins as graphs whose nodes represent chemical moieties, and whose edges represent spatially defined contacts; this representation has been used for e.g. comparative analysis of conformational ensembles[22] or protein classes.[23] Further coarsening can be employed to represent entire residues with a single node, which has been used for e.g. identification of active sites,[24] studying transient structure in IDPs,[25] and analysis of protein dynamics.[26] While most applications of topological coarse-graining have been descriptive, it is also possible to directly model protein structure and/or interaction via its graph representation (see e.g.[25,27–29]). We employ this latter strategy in the context of modeling protein aggregation.

In prior work, topological coarse-graining has been used to model the formation of amyloid fibrils, by defining a free energy landscape (and a corresponding kinetic model) on the set of possible aggregate structures.[28,30] Aggregates in this approach are represented by *aggregation graphs*, where each node corresponds to a protein monomer, and edges join nodes whose respective proteins are non-covalently bound. Models of this type have been able to recapitulate the topology of experimentally determined fibril structures, while being efficient enough to simulate entire aggregation processes (from monomers to mature fibrils) in minutes on consumer hardware. This high degree of computational efficiency is obtained by implicitly integrating over spatial degrees of freedom, working only with binding and unbinding events; this allows both fibril topology and the structure of intermediate and transition states to be probed, for much larger systems and at longer timescales than would be accessible to conventional approaches. The specific approach employed for such models (here referred to as network Hamiltonian models (NHM)) borrows from a large body of computational and statistical theory on exponential family models of random graphs, originally developed to model social networks (see e.g.[31–33]).

While network Hamiltonian models have been used to model the structure of highly ordered aggregates, they have not to date been used to capture disordered aggregates of the type involved in cataract disease. Here, we consider a case involving *unstructured* aggregates, specifically transient aggregation states of γ-Dc as observed in atomistic simulations under physiologically relevant conditions by Wong et al.[11]. We show that a low-dimensional NHM can reproduce the topological structure of aggregates from both WT and W42R γ-Dc. We also show how these models can be used to produce equilibrium draws from much larger systems, facilitating the scaling-up of more detailed simulations to the bulk regime; as we show, this provides both confirmation in this case that many aspects of the small-scale model generalize to large systems, and insights into a specific system size effect in γ-Dc simulations with hundreds of monomers or fewer.

Interaction and Aggregation in γ-Dc

γ-Dc is a structural protein in the human eye lens that is composed of two double-Greek key domains.[34] γ-Dc is expressed in the fiber cells of the eye lens, along with other crystallins from the α, β and γ families, during embryonic development.[35] In order to ensure the transparency of the lens required for sight, other organelles such as the nucleus and rybosomes are removed from the fiber cells as the eye matures, leaving differential concentrations of the water soluble crystallins in each cell. The crystallins must maintain short-range interactions with each other to minimize light scattering while at high concentration (exceeding 400 g/L in humans), resulting in a dense liquid with transient local interactions among monomers.[36]

The high structural stability and weak interaction propensity among structural crystallins, along with the presence of α-crystallins to act as holdase chaperones for unfolded β and γ-crystallins prevent irreversible aggregation from occurring between WT γ-Dc for much of a human life-time.[37] However, as the number of α-crystallins available to chaperone β and γ-crystallins decreases with time, cataract are more likely to form. These cataract are the result of aggregation of (in this case) γ-Dc monomers, arising from e.g. damage from

attack by reactive oxygen species (e.g., hydroxyl radicals generated from UV exposure) or from random interactions occurring when hydrophobic surfaces are exposed due to natural fluctuations away from the native state of γ-Dc.[34]

In the case of the congenital cataract-causing γ-Dc variant W42R, the point-mutation of a buried tryptophan residue in the N-terminal domain (NTD) results in the protein possessing a locally stable conformation that exposes the hydrophobic surfaces of the NTD, making W42R more susceptible to NTD-NTD interactions with other monomers.[11,38] Otherwise, similar structures are found in both crystals and solution for both the WT and W42R variant.[39] We exploit this similarity between the WT and W42R variant structures in the process of coarse-graining - the functional difference between the two structures can be approximated in terms of their rates of aggregation-forming interactions with other monomers, which we recapitulate using network Hamiltonian models.

Network Hamiltonian Models and Aggregation Graphs

An *aggregation graph*, $G = (V, E)$, is a network whose vertices (V) represent protein monomers, and whose edges (E) are drawn between pairs of monomers that are non-covalently bound.[28] An aggregation graph can be seen as a form of *topological coarse-graining*,[40] which flexibly and succinctly represents the structure of connections among proteins while abstracting away other aspects of structure; aggregation graphs have been employed in prior work to model the structure and kinetics of amyloid fibrils,[28,30,41] and related topological representations have also been used to study structure and dynamics in both folded[22,42–44] and intrinsically disordered[25,40] protein systems.

While the aggregation graphs of amyloid fibrils are highly ordered, this is not true of all aggregates; indeed, here we are specifically interested in unstructured aggregates. Fig. 1 shows an aggregation graph derived from atomistic simulations of γ-Dc from Wong et al.[11], indicating the relationship between individual monomers and the resulting topology. While such aggregates are highly disordered, they nevertheless have numerous statistical regularities, which may be used both to gain insights into the aggregation process and model their formation.

Following Grazioli et al.[28], we may model the equilibrium behavior of G via a *network Hamiltonian* that operates on the topological degrees of freedom of the system (i.e., the patterns of bound interactions among protein monomers). Specifically, in equilibrium we model the probability of observing some specific graph microstate g as

$$\Pr\ (G = g \mid \phi, T) = \exp\ [-\mathscr{H}(g)/(k_B T)]h(g)/Z(\phi, T) \tag{1}$$

$$= \exp\ \left[-\left(\phi^T t(g) + k_B T t_e(g)\right)/(k_B T) - t_e(g)\log\ N - \log\ Z(\phi, T)\right], \tag{2}$$

where $\mathscr{H}$ is the graph or network Hamiltonian, expressed in terms of topological degrees of freedom t and energy parameters ϕ; N is the particle number; $h(g)$ is a reference measure accounting for the entropic contribution of unmodeled degrees of freedom; Z is the partition function; and T is the temperature. t_e, in particular, counts the edges of G. Here, we use the

contact-formation measure $h(g) = N^{-t_e(g)}$, and the bond vibration term $(k_B T t_e(g))$ suggested by Grazioli et al.,[28] which correct for (respectively) spatial limitations on edge formation and motional degrees of freedom that are coupled to the graph topology. Models based on Eq. 1 have been shown to be able to reproduce the structure of amyloid fibrils,[28,30] and can be extended to reproduce fibrillization kinetics. Here, we adapt these to the unstructured case.

Inference and model selection.—In practice, we do not know *a priori* which topological degrees of freedom will prove critical for our system of interest, nor do we know ϕ - rather, we observe random equilibrium draws from G, and seek to infer a Hamiltonian that reproduces the distribution of aggregation graphs. To this end, it is useful to observe that the model of Eq. 1 is equivalent to an exponential family random graph model (ERGM), a widely studied formalism for network modeling in the social and statistical sciences (see, e.g.,[32,33]). The ERGM parameterization of the model of Eq. 1 is given by

$$\Pr\left(G = g \mid \theta\right) = \exp\left[\theta^T t(g) + \log h(g) - \log Z(\theta)\right],$$

where t, h, and Z are as before, and θ is a real vector of model parameters. Model selection and inference for ERGMs are well-studied,[33] allowing us to infer θ and t (and hence $\mathscr{H}$) from the realized aggregation graphs. Specifically, we obtain ϕ from θ under the family of Eq. 2 via

$$\begin{aligned} &-\mathscr{H}(g)/(k_B T) + \log h(g) = \theta^T t(g) - \phi^T t(g)/(k_B T) - t_e(g) - t_c \log N = \theta^T t(g) \\ &\Rightarrow \phi_e = -k_B T(\theta_e + 1 + \log N),\ \phi_{s \neq e} = -k_B T \theta_{s \neq e}. \end{aligned} \tag{3}$$

Given a proposed set of model terms (i.e., choice of t), we perform parametric inference for θ using the pooled maximum likelihood (MLE) method of Yin and Butts[29], from which we can then infer ϕ using the relations of Eq. 3. As our goal here is to reproduce the distribution of aggregate sizes - corresponding to component sizes in the aggregation graph representation - we perform model selection by finding a term set that optimizes fit to the observed component distribution. Specifically, we first posit a set of candidate terms based on prior work and first principles, and then select models sequentially by minimizing distance between the simulated component size distribution under the model and the observed distribution (L2 norm of the log relative distribution). (See Methods for details.)

Model terms.—The terms in $\mathscr{H}$ reflect multi-body interactions, as reflected in the topological degrees of freedom of the aggregation graph. A large body of work exists on such terms in an ERGM context, including derivation from dependence constraints (i.e., Hammersley-Clifford[45]),[46,47] corrections for diminishing marginal effects,[48] and consequences for equilibrium behavior.[49–52] In the context of aggregation graphs, work on amyloid fibrils[28] has identified a number of terms that may be useful for capturing protein aggregation states per se; these include the null shared partner statistics (NSPs) and edgewise shared partner statistics (ESPs),[53] as well as cycle and star statistics. In the case

of γ-Dc, the highly skewed distribution of aggregate sizes also suggests terms specifically related to component sizes. These include monomer and dimer counts, as well as terms reflecting general tendencies that enhance or inhibit the formation of large aggregates. Specifically, we here introduce a term for this last effect based on non-central moments of the component size distribution. This term, which we refer to as *compsizesum*, has the form

$$t_C\left(g\right) = \sum_{i=1}^{N} S\left(g\right)_i i^{\gamma}, \tag{4}$$

where $S(g)_i$ is the count of components of size i within g, and γ is a fixed parameter governing the behavior of the statistic. We observe that $\gamma = 1$ simply returns the number of vertices, and is hence uninteresting; however, $\gamma = 2$ yields the sum of squared component sizes, and thus influences the variance of the component size distribution. Mechanistically, we also observe that the change in t_C associated with merging two components of sizes a and b is equal to $2ab$, and thus t_C directly reflects the impact of component size on the favorability of coalescence or dissolution: when the associated ϕ parameter is negative, this implies that contacts between larger aggregates are increasingly favored, while a positive ϕ indicates that such mergers become increasingly unfavorable as aggregate size increases.

For our analyses, we employ a subset of computationally scalable terms with relevance to the unstructured case; as we show, these terms are sufficient to produce models that can reproduce the observed distribution of γ-Dc aggregate sizes, along with other topological properties. The terms used are the following. The edge count (*edges*) parameterizes the base dissolution energy of a single edge[28] and is included in all models. The tendency to form extended versus “kinked” linear structures is influenced by open two-paths, as captured by null (i.e., unbonded) pairs bound to a single shared partner, or $NSP(1)$ s. Biases towards *monomers* and *dimers* are plausible, and captured by counts of the same (i.e., components of size 1 or 2, respectively). Closed triadic structures can be extremely stable, motivating consideration of counts of bound pairs (edges) with one ($ESP(1)$s) or two ($ESP(2)$s) shared partners. Higher-ordered edgewise shared partners must be handled carefully, as forces favoring excessively high shared partner counts easily lead to sharp transitions to extremely dense solid states that are not realistic for this system;[54,55] we thus employ the geometrically weighted edgewise shared partner (GWESP) statistic for higher-order triadic closure effects,[48,56] which constrains contributions of high-order ESPs to have geometrically declining marginal effects. The structures represented by these terms are represented schematically in Fig. 2.

Although all of these terms were considered in model evaluation, not all were ultimately selected for the final model. Our model selection procedure is described below.

Methods

Atomistic Simulation and Network Generation

Wong et al.[11] performed atomistic simulation of equilibrium distributions of WT and W42R γ-Dc using multi-conformation Monte Carlo (mcMC) methods;[12] here, we use the

network representation of aggregates generated from this study. mcMC simulations were performed for $N = 375$ proteins at 310K and 200g/K under periodic boundary conditions, using conformation libraries obtained from explicit solvent MD simulations under the CHARMM36 forcefield[57] in TIP3P water.[58] From these simulations, 14,000 and 16,000 frames were obtained for WT and W42R (respectively). Further details regarding the original simulation study can be found in Wong et al.[11].

Wong et al.[11] define aggregation graphs from the atomistic γ-Dc simulations as follows. Each vertex is associated with a single protein monomer, with one graph per frame; within a given network, two vertices are tied if they have respective domains whose centers of mass are within 31Å of each other. (This cutoff reflects the distance required for direct contact, as revealed by analysis of domain-domain radial distribution functions across simulation frames; see Wong et al.,[11] figure S3.) This resulted in 14,000 WT and 16,000 W42R aggregation graphs, which are employed for our present analysis. Network visualization and analysis was performed using the `statnet` library[59] for the R statistical computing system,[60] with the `network`[61] and `sna`[62] libraries used to compute descriptives and graphical layouts.

Component/Aggregate Size Distribution Estimation and Comparison

Component sizes for all networks were computed using the `sna` library. The component size distribution (the probability distribution for the size of a randomly chosen component) was estimated using a non-parametric Bayesian procedure, as follows. For an arbitrary graph of order N, the component size Z has support on $\mathcal{Z}_N = (1, \ldots, N)$. We model this as $Z \sim \text{Categorical}\ (\psi)$, where $\psi_i = \Pr\ (Z = i)$. We place a minimally informative Jeffreys prior on ψ, leading to $p(\psi) = \text{Dirichlet}\ (0.5)$, where the latter is the homogeneous N-dimensional Dirichlet distribution with concentration parameter 0.5. Given multiple observations of $Z, \boldsymbol{z} = (z_1, \ldots, Z_m)$, the corresponding posterior distribution is $p(\psi \mid \mathbf{Z} = \mathbf{z}) = \text{Dirichlet}\ (\mathrm{S} + 0.5)$, where $S_i = \sum_{j=1}^{m} I(z_j = i)$ is the observed count of components having size i. (This is an example of the well-known Dirichlet-multinomial model.[63]) Other posterior quantities are then easily calculated from the properties of the Dirichlet distribution; in particular, $\mathbf{E}\psi_i = (S_i + 0.5)/(m + N/2)$, and the posterior marginals of ψ_i are given by $\psi_i \sim \text{Beta}\ (S_i + 0.5, m - S_i + (N - 1)/2)$.

For model selection (as discussed below), we seek to compare the component size distributions arising from the network Hamiltonian model to the component size distributions obtained from atomistic simulations. Because we are particularly interested in tail events (i.e., the distribution of relatively rare, large aggregates), we use the L2 norm of the logged relative distribution[64] as our measure of discrepancy between distributions. I.e., given fixed distributions f, g over component sizes $\mathcal{Z}_N$, our discrepancy measure is

$$D\left(f, g\right) = \| \log\ f/g \| = \sum_{i=1}^{N} \left(\log\ f\left(i\right) - \log\ g\left(i\right) \right)^2, \tag{5}$$

where the informal notation f/g denotes the relative distribution over $\mathcal{X}_N$. In our case, we are interested in $D(f_{obs}, f_{sim})$, where f_{obs} is the observed or target component distribution and f_{sim} is the (simulated) distribution from our network model. However, neither distribution is known exactly. Thus, we instead minimize the posterior quantity $\mathbf{E}D(f_{obs}, f_{sim}) \mid \mathbf{z}_{obs}, \mathbf{z}_{sim}$, where $f_{obs} \sim \text{Dirichlet}\left(S^{obs} + 0.5\right)$ and $f_{sim} \sim \text{Dirichlet}\left(S^{sim} + 0.5\right)$ (with S^{obs} and S^{sim} the respective component count distributions from the atomistic and network Hamiltonian simulations, respectively). Although this has no closed form solution, we can calculate it straightforwardly by Monte Carlo quadrature,[65] exploiting the ease of taking draws from the Dirichlet distribution. (Note that our choice of prior ensures that $D(f_{obs}, f_{sim})$ has a finite expectation.) This approach allows us to automatically account for posterior uncertainty in component size distributions when making comparisons.

Model Selection and Parameter Estimation

Models were fit by maximum likelihood estimation (MLE), using the pooling method of Yin and Butts;[29] estimation was performed using the `ergm` package,[66] version 4.1.2, using the stochastic approximation method with respective base burn-in and thinning intervals of 5×10^4 and 2×10^4. For each candidate model, separate pooled MLEs were obtained for the respective collections of WT and W42R networks. Selection of the GWESP decay parameter was performed by grid search. Change statistics for the dimer count and summed component size terms were implemented via the `ergm.userterms` library.[67]

Models were chosen by forward selection, with the objective being minimization of the total expected L2 norm of the log relative distribution of the observed versus model-generated component distributions for WT and W42R. Specifically, for each fitted model we generate 5000 graph draws by Markov Chain Monte Carlo (MCMC) using the `ergm` library $\left(N^2\right.$ respective burn-in and thinning iterations for each trajectory, Tie-No-Tie sampler), obtaining the estimated posterior distribution of component sizes as described above. This was used to obtain $\mathbf{E}D(f_{obs}, f_{sim}) \mid \mathbf{z}_{obs}, \mathbf{z}_{sim}$ as described above for both WT and W42R, and the sum of the respective expected errors was taken as the figure of merit for the specified model. Terms were chosen to minimize this total error. Model search began with the base null model (edge-only); at each iteration, each currently non-incorporated term was added one at a time, and the addition providing the greatest total error reduction was kept for the next iteration. Model selection terminated when no term improved fit to the component size distribution. Table 1 shows the complete model selection trace, along with the errors at each step. In addition to the terms selected for the final model, terms for monomer count, dimer count, and ESP(2) counts were also evaluated; these were not found to improve fit to the component distributions, and were not selected. Parameter estimates (MLEs) and standard errors for the final models are shown in Table 2.

Extrapolative Simulation

Extrapolative simulation was performed by MCMC using the `ergm` library, using the default Tie-No-Tie sampler. Systematic pilot simulations using the final fitted models (not shown) indicated that, for graphs of order N, burn-in and thinning parameters of $250N$ provided

good convergence and mixing properties over a wide size range (with mixing improving with size). These settings were hence employed for all extrapolative simulations. Model parameters in ERGM (i.e., θ) space for the extrapolated models were obtained from the ϕ representation of Eq.1, with N adjustments as specified. Component size distributions and other metrics for the extrapolated network simulations were computed as described for the other simulations.

To extrapolate across concentration, it is necessary to add an additional adjustment to Eq. 2, to account for changes in the effective collision rate. Following Eq. 14 of Butts,[68] the first-order effect on the aggregation graph distribution of changing from baseline concentration C to extrapolated concentration C' is to shift the reference measure by a factor of $\left(\frac{C'}{C}\right)^{t_e(g)}$; this leads to the distribution

$$\Pr(G = g \mid \phi, T) = \exp\left[-\left(\phi^T t(g) + k_B T t_e(g)\right)/(k_B T) - t_e(g)\left(\log N - \log\frac{C'}{C}\right) - \log Z(\phi, T)\right].$$

Intuitively, multiplying the concentration by a factor α has the effect of shifting the edge parameter (in its θ representation) by $\log\ \alpha$, which is easily implemented. Thus, increasing the concentration will tend to increase the expected number of contacts per monomer, while decreasing concentration will reduce it. The net impact of concentration changes on the aggregation graph depends, however, on the full model. To examine the potential impact of concentration on aggregation in the γ-Dc models, we simulate 1000 graph draws for a large system ($N = 10000$) at concentrations of 100, 200, 300, and 400 g/L (with the original model having been calibrated based on mcMC simulations at 200 g/L).

Geometry Imputation

Although the aggregation graph is purely topological (i.e., it contains only information on bound interactions among monomers, and is not spatially explicit), we here perform an approximate geometry imputation to examine possible trends in aggregate shape driven by the underlying topology. Specifically, we map the topology of realized aggregates to a three-dimensional structure that is compatible with monomer size and bound interactions, and that conforms to a very simple but physically plausible model. Specifically, we proceed as follows. Given an aggregation graph, g, we first segment the aggregation graph into connected components (i.e., distinct aggregates) $g^{(1)}, \ldots, g^{(m)}$. (Component segmentation and other analyses performed using the `sna`[62] package.) For each component, $g^{(i)}$, three-dimensional coordinates are then assigned by a two-phase process. First, we employ a modified three-dimensional Kamada-Kawai[69] algorithm (KK) to obtain an initial layout, using the square root of the geodesic distance between vertices, scaled by twice the monomer radius, as the objective. The KK procedure attempts to find an assignment of coordinates to the vertex set that minimizes the sum of squared errors between the Euclidean distances among vertex coordinates and a target distance matrix; here, our choice of distance target approximates the expected distance under a random polymer model. Given the initial layout, we refine it to correct for overlapping vertices, ensure that bonded vertices are

in contact, and to prevent non-bonded vertices from being in contact. This is done via a simulated annealing procedure, minimizing a simple objective given by

$$\sum_{\{j,k\}} \left[E_{rep}(2r/d_{jk})^{12} + E_{bond} g_{jk}^{(i)} (2r - d_{jk})^2 \right],$$

where $E_{rep} = 1$ and $E_{bond} = 10$ are parameters governing repulsion and bonded interaction (respectively), r is the effective monomer radius, d_{jk} is the Euclidean distance between the coordinates of vertices j and k, $g_{jk}^{(i)} = 1$ if j is bound to k (else 0), and the sum is over all vertex pairs within the component. (Procedure implemented using `Rcpp`.[70]) The resulting coordinates reflect a plausible low-energy conformation for the aggregate, assuming that interactions among monomers are not angularly restricted beyond constraints induced by crowding and bound interactions. For an effective monomer radius, the geometric mean of their projected monomer lengths along their respective principle gyration axes were used; these were computed using the `bio3d` package,[71] based on PDB structures 1HK0[72] and 4GR7[39]. The resulting radii were 19.55Å for WT, and 20.05Å for W42R.

To probe possible relationships between geometry and size (in the sense of numbers of monomers per aggregate), we simulate 100 aggregation graph realizations from our estimated models for WT and W42R, extrapolating to a system with $N = 10^4$ monomers. Coordinates were obtained for each aggregate in each graph, using the above procedure. For each aggregate, the radius of gyration was computed (approximating each monomer by a sphere of its effective radius), and was scaled by the monomer radius of gyration to obtain the dimensionless statistic R_g/r_g (where r_g is the monomer radius of gyration). Using the above structures and libraries, the monomer r_g values were calculated to be 16.63 Å for WT and 16.72Å for W42R. We also examine geometry using an *elongation factor*, defined here as L_1/L_3, where L_i is the width of the aggregate when projected along its ith principal axis of gyration. Intuitively, an elongation factor of 1 indicates a spherical aggregate, with higher values indicating greater departures from sphericity. Likewise, R_g/r_g would be expected to scale as $N^{(1/3)}$ as N becomes large, for spherical aggregates.

Results

Topology of γ-Dc Aggregates

γ-Dc WT, W42R aggregates have skewed size distributions, with truncated upper tails.—Fig. 3 (top right) shows posterior means and 95% intervals for the aggregate size distributions; we observe monotone distributions in both cases, with sizes that scale as approximately $1/n^2$ for small aggregates. Size frequency in WT begins to drop off rapidly beyond approximately 10 monomers, with aggregates greater than 100 monomers being extremely rare. By contrast, W42R shows a much longer upper tail, with sizes becoming truncated only near the 200–250 range. Although this truncation point is still considerably smaller than the system size (375 monomers), it would be reasonable to suspect that it could be a finite-system artifact; as we show below, however, this does not appear to be the case.

Larger γ-Dc aggregates are dendritic, with locally kinked structure.—Fig. 3 (bottom) shows two representative topological γ-Dc aggregation graphs for WT and W42R (each selected by having the minimum discrepancy versus the overall component distribution), with vertices colored by component size. As can be seen, complex components found in either variant are relatively "loose," with extensive tree-like structures marked by continuous and occasionally branching paths, combined with local "kinks" resulting from triangulation. Although triangles are common relative to the sparsity of the graph, we see an absence of both large cliques and the highly regular linear structures seen in fibril formation. Qualitatively, WT and W42R appear to produce very similar types of aggregates (net of size); there are, however, statistical differences between them, as we show below.

Network Hamiltonian Modeling of γ-Dc Aggregates

Model parameters reveal topological drivers of aggregate structure.— Examination of reduction in prediction error for the component size distribution as a function of model terms (Table 1) shows that the key drivers of aggregate structure (in descending order of importance) are: the suppression of closed, chain-like structures (as evidenced by the positive NSP(1) energies (Table 2)); enhanced triadic closure (negative GWESP energies); and suppression of mergers between large aggregates (positive compsizesum energies). We also see an additional minor ESP(1) correction, which adjusts the closure pattern generated by GWESP but does not change the qualitative tendency towards local triangulation.

Quantitatively, we note that the base dissociation energy for a bond between two otherwise isolated monomers is low; although all such energies for coarse-grained models are necessarily approximate, we observe an effective net dissociation energy for such bonds of approximately 1 kcal/mol for WT, and 1.8 kcal/mol for W42R. To give some context for the nature of the interactions, this is roughly comparable to a weak hydrogen bond. While this may seem low, it is compatible with the observation that γ-Dc is overwhelmingly monomeric, and higher-order interactions are generally transient. As another point of comparison, Mills-Henry et al.[74] estimate the free energy of the γ-Dc domain interface - which would be expected to be a much stronger interaction than transient interactions between otherwise independent monomers - at approximately 4 kcal/mol. We observe that dissociation energies for W42R start off roughly 80% higher than WT, reflecting a greater net propensity for interaction.

While the qualitative behaviors of the WT and W42R energy functions are similar, we see further quantitative differences between the two. Extended conformations are less favorable for W42R than WT (as seen from the higher NSP(1) energy), though this must also be weighed against the higher baseline propensity of W42R to form contacts. Combining the ESP(1) and GWESP terms to examine the net energies associated with ESP(k) configurations, we find that ESP(1)s are overall much more favored in WT than W42R (−0.32 vs. 0.05 kcal/mol), and while this gap closes somewhat for ESP(2)s, it is still higher (−0.59 vs. −0.38 kcal/mol). This gap gradually narrows for higher order ESPs (−0.69 vs. −0.48 kcal/mol for ESP(3)s, and −0.74 vs. −0.56 kcal/mol for ESP(4)s), though it is still present. This suggests that, *prima facie*, triadic closure in WT is driven more by the

additional stability of triangulated structures, while the combination of enhanced interaction and instability/unfavorability of extended structures plays a larger role in W42R. Finally, while the compsizesum energy appears fairly small at first blush, we see that it is about an order of magnitude larger for WT and W42R. To put this term in perspective, it is helpful to consider the minimum component size such that a merger of two such components would produce a change in the compsizesum energy that exactly offsets the energy of a single baseline edge. For WT, this size is approximately 22 monomers, versus approximately 67 for W42R. Thus, self-inhibition is much weaker for the mutant than for wild type, plausibly playing a significant role in the ability of the latter to form larger components. Moreover, since the change in energy scales with the product of component sizes, we would expect to see growth in medium to large WT aggregates to be much more dependent upon incorporation of monomers of very small oligomers than W42R. This may provide more viable pathways to the formation of larger aggregates in the latter, with corresponding impact on aggregation kinetics.

Network Hamiltonian models recapitulate aggregate size and structure.—Fig. 4 shows predicted properties of aggregates from the network Hamiltonian models (based on MCMC simulation), versus the observed aggregation graphs. Despite the simplicity of the network models, we find that they do an excellent job of recapitulating both large-scale structure (component size distributions) and local structure (degree and ESP distributions) for both mutant and WT. In particular, both models recapitulate the $1/n^2$ small-aggregate scaling, and differences in tail weight. It should be noted that the ESP and degree statistics match well not only on means, but also on variances (as shown by 95% simulation intervals), demonstrating that they recapitulate variability in aggregate structure across realizations as well as overall tendencies.

Extrapolative Simulation of γ-Dc Aggregates

Larger systems at constant concentration yield similar aggregate sizes.—An obvious concern when simulating aggregation processes using atomistic methods is that we are restricted to relatively small system sizes; this both restricts the upper tail of the aggregation size distribution and creates artificial dependence in aggregate sizes. The latter arises from exhaustion: if, e.g., a system contains an aggregate of size M, then it must be the case that only $N - M$ monomers remain to form other aggregates. It is thus impossible to observe interactions among multiple aggregates of size $> N/2$, and every large aggregate is necessarily surrounded by much smaller aggregates (a condition that need not occur in bulk). While the truncation effect can only artificially reduce aggregate sizes, this last effect could either enhance or suppress the formation of larger aggregates (depending on the favorability of interactions between aggregates as a function of size).

In general, it is thus hard to know how system size effects will impact aggregate size, unless the maximum observed size is small compared to the number of monomers in the system. Here, however, the relative computational efficiency of the network Hamiltonian models allows us to simulate draws from much larger systems than are accessible via mcMC, permitting us to directly observe the impact of increasing system size on aggregation. In

particular, we here take draws from systems as large as 10^4 monomers, an increase of almost two orders of magnitude from our base case of $N = 375$.

Figure 5 shows the resulting posterior means and 95% intervals for aggregate size distributions, by variant and system size. Overall, we find that the size distributions seen in smaller systems remain similar as one approaches the bulk limit. We do not, in particular, see evidence of truncation effects (particularly for the W42R variant, where they might have been expected), suggesting that observed sizes are in fact due to the self-limiting properties of aggregate assembly and disassembly, and not to a lack of available monomers. Interestingly, we in fact see some sharpening and lowering of the upper tail of the size distribution as system size increases. This may result from mid-sized and smaller components competing with large components to recruit small components (since mergers become increasingly unfavorable with size), "starving" large components of monomers that they might otherwise recruit for further growth. Such competition is limited in the small-N case by the exhaustion mechanism described above, thus potentially allowing some components to grow slightly larger than would be possible in a bulk system. By being able to evaluate systems that are much larger than the largest components, we thus get a more realistic picture of bulk behavior.

Increasing concentration increases aggregate size.—Probing the high-concentration regime is another challenge for conventional Monte Carlo simulation methods, as close packing of proteins makes it difficult to propose moves without an extremely high clash (and hence rejection) rate. A potential asset of network Hamiltonian models is the ability to explore potential effects of concentration by simulating aggregation graphs from concentration-adjusted models, which do not suffer from this difficulty. For γ-Dc, Figure 6 shows posterior means and 95% intervals for aggregate size distributions, based on simulations with $N = 10^4$ and concentrations of 100, 200, 300, and 400 g/L (with 200 g/L being the concentration of the original system to which the models were fit). As expected, increasing concentration increases the mean aggregate size for both WT and W42R, although we do not observe a marked increase in the size of the very largest aggregates obtained for concentrations above 200 g/L. We do, however, see large aggregates occurring with higher frequency, particularly for W42R (where we see a marked flattening of the frequency distribution above ≈50 monomers at 400 g/L). We also see a larger mean shift for W42R versus WT, with the mean aggregate size at 400 g/L being 67% higher than the size at 200 g/L for WT (13.5 vs. 8.1) and 84% higher for W42R (33.8 vs. 18.3). Although reduced concentration lowers aggregate size, this is also more notable for WT than W42R (mean size 5.5 versus 10.1 , with a marked difference in the size of the largest aggregates). These results suggest that, beyond simply forming a small number of distinctively large aggregates, W42R at high concentration sustains larger populations of medium-to-large transient aggregates, which may place more monomers in locally crowded settings in which transient conformational changes (e.g., partial unfolding) potentially lead to irreversible aggregation.

Larger aggregates may be more compact, but slightly oblate.—Although our approach does not directly predict the three-dimensional structure of γ-Dc aggregates,

the aggregation graph may provide evidence regarding likely conformations. Using the procedure described above, we examine imputed geometric properties for all aggregates from samples of 100 draws from the WT and W42R models (respectively), with a system size of $N = 10^4$ monomers. Figure 7 shows the resulting relationships of scaled radius of gyration and elongation factors with aggregate size. While there is some deviation for small aggregates, medium to large aggregates (10 or more monomers) are predicted to have have nearly spherical R_g scaling; a linear fit of the log R_g/r_g ratios to log sizes for aggregates above this minimum lead to estimated scaling of $R_g/r_g \propto N^{0.333 \pm 0.002}$ for WT (with N here being the aggregate size), and $R_g/r_g \propto N^{0.313 \pm 0.001}$ for W42R. The elongation metric shows a slight deviation from spherical behavior, with large aggregates (100 or more monomers) tending towards an average of approximately 1.2 (i.e., the longest axis being 20% longer than the shortest). Although the R_g scaling coefficients are significantly different ($z = 15.52, p \ll 0.0001$), we would caution against drawing strong interpretations from such a small difference from a highly simplified geometric model. We would, however, suggest that the analysis shows that the topology of the aggregates does not constrain them to be far from spherical, nor does it constrain WT and W42R to produce aggregates that differ greatly in overall shape. Although tentative, the predicted trend in obliquity would seem to be a fruitful target for experimental examination.

Conclusion

Here we employed Exponential-family Random Graph Models to fit network Hamiltonian models to atomistic simulations of WT and W42R γ-Dc, allowing us to identify topological degrees of freedom that govern the formation of unstructured aggregates. The transient nature of the protein-protein interactions in the resulting models reflect the properties of the original mcMC simulation,[11] and are thus distinct from the highly durable intermolecular interactions seen in fibril formation. However, these transient interactions plausibly provide opportunities for damaged or partially unfolded γ-Dc to form longer-lived structures[75] (or, likewise to support more subtle surface interactions that have also been argued to promote aggregation[76]) and may hence provide insights into the process of cataract initiation. In keeping with this view, we see that the cataract-prone W42R mutant behaves in a manner much more conducive to structure formation, both in terms of the favorability of overall interaction and the tendency to form lower energy triadic structures. Given atomistic models or experimental data on durable aggregates, the same strategies followed here can also be used to model them.

Combining network analysis with mcMC simulations also offers the possibility of examining the relationships between conformational states and structural position within the aggregation graph. We did not pursue this avenue here, because preliminary examination of of the conformational states suggested that they did not show enough variation for such an analysis to be fruitful. However, in systems with greater variation in monomeric states, this approach would seem to be a useful direction. In particular, while our analysis implicitly marginalizes over monomeric states (their impacts on aggregation being indirectly reflected via the terms of the network Hamiltonian), it may in some cases be possible impute states

from simulated aggregation graphs, by training a model to predict the former from the latter using mcMC draws. This too would seem to be a useful direction for further work.

The scalability of network Hamiltonian models allows simulation of large systems, providing additional information on the impact of the system size on aggregation. For WT γ-Dc, the component size distribution did not change substantially from what was seen in the smaller, atomistic simulation, while the W42R variant system sees a decrease in observations of the largest aggregates (lighter upper tail) as system size increases. Our results suggest that this may arise from competition between mid-sized and large aggregates for monomers to incorporate, a phenomenon that is artificially suppressed in small simulations. Extrapolation to higher concentrations does show an increased population of large aggregates, particularly for W42R. Although we cannot directly determine geometry from these simulations, we can approximate it using simple spatial models. Applying that approach here suggests that we cannot immediately constrain the aggregates to being non-spheroidal in solution, though there is some evidence of obliquity. Better models for moving from topology to geometry for aggregation graphs (as has been explored at the atomistic scale for residue-level networks[40]) could further refine such predictions, and would be particularly valuable for providing better targets for e.g. light scattering experiments.

One interesting observation from the present models is the apparent self-limiting behavior of growing γ-Dc aggregates. This appears necessary to reproduce the results of the mcMC models, which even for W42R do not show aggregates that approach the limit of the system size ($N = 375$), and which manifests within the network Hamiltonian model by an inhibition for mergers between large aggregates. Such self-limiting behavior could be compatible with the formation of spherical structure, if more favorable attachment sites end up being buried as the aggregate grows, and one could conjecture that such a mechanism, if present, helps prevent pathological aggregation in the eye lens. However, we also reiterate that some modes of aggregation were not accessible to the mcMC model (e.g., those based on partial unfolding or refolding of monomers or disulfide bond formation[38,75]), and thus are not incorporated here; we therefore view this prediction as tentative. Formally, we observe that the essentially quadratic penalty for component mergers used in the models fitted here may be too sharp in some settings, and a softer function may be needed. Investigations with underlying models based on a wider range of systems would be fruitful in clarifying this issue.

Network Hamiltonian models provide a flexible framework for describing interactions between proteins and the resulting structures, whether transient in nature as in the case of the present study, or the more durable structure of amyloid fibrils. Combined with experimental data or atomistic models, network Hamiltonian models can be used to extrapolate simulations of systems that are orders of magnitude larger than atomistic models, providing a convenient method for examining the underlying structure of large protein aggregates. Additionally, given the ability of network Hamiltonian models to determine distributions of aggregate sizes, these models may provide insight into the transient interactions which guide phenomena such as liquid-liquid phase separation and phase transitions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This research was supported by NASA award 80NSSC20K0620, and NIH award 1R01GM144964-01. The authors thank Rachel Martin for helpful comments.

References

(1). Santos J; Pujols J; Pallarès I; Iglesias V; Ventura S Computational Prediction of Protein Aggregation: Advances in Proteomics, Conformation-Specific Algorithms and Biotechnological Applications. Computational and Structural Biotechnology Journal 2020, 18, 1403–1413. [PubMed: 32637039]

(2). Chiti F; Dobson CM Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. Annual Review of Biochemistry 2017, 86, 27–68.

(3). Nilsberth C; Westlind-Danielsson A; Eckman CB; Condron MM; Axelman K; Forsell C; Stenh C; Luthman J; Teplow DB; Younkin SG; Näslund J; Lann-felt L The 'Arctic' APP Mutation (E693G) Causes Alzheimer's Disease by Enhanced Aβ Protofibril Formation. Nature Neuroscience 2001, 4, 887–893. [PubMed: 11528419]

(4). Jedziniak JA; Kinoshita JH; Yates EM; Hocker LO; Benedek GB On the Presence and Mechanism of Formation of Heavy Molecular Weight Aggregates in Human Normal and Cataractous Lenses. Experimental Eye Research 1973, 15, 185–192. [PubMed: 4692231]

(5). Prabakaran R; Rawat P; Thangakani AM; Kumar S; Gromiha MM Protein Aggregation: In Silico Algorithms and Applications. Biophysical Reviews 2021, 13, 71–89. [PubMed: 33747245]

(6). Morriss-Andrews A; Shea J-E Computational Studies of Protein Aggregation: Methods and Applications. Annual Review of Physical Chemistry 2015, 66, 643–666.

(7). Chen B; Siepmann JI A Novel Monte Carlo Algorithm for Simulating Strongly Associating Fluids: Applications to Water, Hydrogen Fluoride, and Acetic Acid. The Journal of Physical Chemistry B 2000, 104, 8725–8734.

(8). Chen B; Siepmann JI Improving the Efficiency of the Aggregation-Volume-Bias Monte Carlo Algorithm. The Journal of Physical Chemistry B 2001, 105, 11275–11282.

(9). Lomakin A; Asherie N; Benedek GB Monte Carlo Study of Phase Separation in Aqueous Protein Solutions. The Journal of Chemical Physics 1996, 104, 1646–1656.

(10). Lund M; Jönsson B A Mesoscopic Model for Protein-Protein Interactions in Solution. Biophysical Journal 2003, 85, 2940–2947. [PubMed: 14581196]

(11). Wong EK; Prytkova V; Freites JA; Butts CT; Tobias DJ Molecular Mechanism of Aggregation of the Cataract-Related γD-Crystallin W42R Variant from Multiscale Atomistic Simulations. Biochemistry 2019, 58, 3691–3699. [PubMed: 31393108]

(12). Prytkova V; Heyden M; Khago D; Freites JA; Butts CT; Martin RW; Tobias DJ Multi-Conformation Monte Carlo: A Method for Introducing Flexibility in Efficient Simulations of Many-Protein Systems. The Journal of Physical Chemistry B 2016, 120, 8115–8126. [PubMed: 27063730]

(13). Majumdar BB; Prytkova V; Wong EK; Freites JA; Tobias DJ; Heyden M Role of Conformational Flexibility in Monte Carlo Simulations of Many-Protein Systems. Journal of Chemical Theory and Computation 2019, 15, 1399–1408. [PubMed: 30633517]

(14). Noid WG Perspective: Coarse-grained Models for Biomolecular Systems. The Journal of Chemical Physics 2013, 139, 090901. [PubMed: 24028092]

(15). Zhang Z; Glotzer SC Self-Assembly of Patchy Particles. Nano Letters 2004, 4, 1407–1413. [PubMed: 29048902]

(16). Wilber AW; Doye JPK; Louis AA; Noya EG; Miller MA; Wong P Reversible Self-Assembly of Patchy Particles into Monodisperse Icosahedral Clusters. The Journal of Chemical Physics 2007, 127, 085106. [PubMed: 17764305]

(17). Quinn MK; Gnan N; James S; Ninarello A; Sciortino F; Zaccarelli E; McManus JJ How Fluorescent Labelling Alters the Solution Behaviour of Proteins. Physical Chemistry Chemical Physics 2015, 17, 31177–31187. [PubMed: 26542112]

(18). Khan AR; James S; Quinn MK; Altan I; Charbonneau P; McManus JJ Temperature-Dependent Interactions Explain Normal and Inverted Solubility in a γD-Crystallin Mutant. Biophysical Journal 2019, 117, 930–937. [PubMed: 31422822]

(19). Altan I; Khan AR; James S; Quinn MK; McManus JJ; Charbonneau P Using Schematic Models to Understand the Microscopic Basis for Inverted Solubility in γD-Crystallin. The Journal of Physical Chemistry B 2019, 123, 10061–10072. [PubMed: 31557434]

(20). Liu H; Kumar SK; Sciortino F Vapor-Liquid Coexistence of Patchy Models: Relevance to Protein Phase Behavior. The Journal of Chemical Physics 2007, 127, 084902. [PubMed: 17764289]

(21). Benson NC; Daggett V A Chemical Group Graph Representation for Efficient High-Throughput Analysis of Atomistic Protein Simulations. Journal of Bioinformatics and Computational Biology 2012, 10, 1250008.

(22). Cross TJ; Takahashi GR; Diessner EM; Crosby MG; Farahmand V; Zhuang S; Butts CT; Martin RW Sequence Characterization and Molecular Modeling of Clinically Relevant Variants of the SARS-CoV-2 Main Protease. Biochemistry 2020, 59, 3741–3756. [PubMed: 32931703]

(23). Unhelkar MH; Duong VT; Enendu KN; Kelly JE; Tahir S; Butts CT; Martin RW Structure Prediction and Network Analysis of Chitinases from the Cape Sundew, Drosera Capensis. Biochimica et Biophysica Acta (BBA) - General Subjects 2017, 1861, 636–643. [PubMed: 28040565]

(24). Amitai G; Shemesh A; Sitbon E; Shklar M; Netanely D; Venger I; Pietrokovski S Network Analysis of Protein Structures Identifies Functional Residues. Journal of Molecular Biology 2004, 344, 1135–1146. [PubMed: 15544817]

(25). Grazioli G; Martin RW; Butts CT Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods. Frontiers in Molecular Biosciences 2019, 6.

(26). Sheik Amamuddy O; Verkhivker GM; Tastan Bishop Ö Impact of Early Pandemic Stage Mutations on Molecular Dynamics of SARS-CoV-2 Mpro. Journal of Chemical Information and Modeling 2020, 60, 5080–5102. [PubMed: 32853525]

(27). Yaveroğlu ÖN; Fitzhugh SM; Kurant M; Markopoulou A; Butts CT; Pržulj N ergm.graphlets: A Package for ERG Modeling Based on Graphlet Statistics. Journal of Statistical Software 2015, 65, 1–29.

(28). Grazioli G; Yu Y; Unhelkar MH; Martin RW; Butts CT Network-Based Classification and Modeling of Amyloid Fibrils. The Journal of Physical Chemistry B 2019, 123, 5452–5462. [PubMed: 31095387]

(29). Yin F; Butts CT Highly Scalable Maximum Likelihood and Conjugate Bayesian Inference for ERGMs on Graph Sets with Equivalent Vertices. PLOS ONE 2022, 17, e0273039. [PubMed: 36018834]

(30). Yu Y; Grazioli G; Unhelkar MH; Martin RW; Butts CT Network Hamiltonian Models Reveal Pathways to Amyloid Fibril Formation. Scientific Reports 2020, 10, 15668. [PubMed: 32973286]

(31). Hunter DR; Krivitsky PN; Schweinberger M Computational Statistical Methods for Social Network Models. Journal of Computational and Graphical Statistics 2012, 21,856–882. [PubMed: 23828720]

(32). Lusher D, Koskinen J, Robins G, Eds. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications; Structural Analysis in the Social Sciences; Cambridge University Press: Cambridge, 2012.

(33). Schweinberger M; Krivitsky PN; Butts CT; Stewart JR Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios. Statistical Science 2020, 35, 627–662.

(34). Serebryany E; King JA The B γ-Crystallins: Native State Stability and Pathways to Aggregation. Progress in Biophysics and Molecular Biology 2014, 115, 32–41. [PubMed: 24835736]

(35). Bloemendal H; de Jong W; Jaenicke R; Lubsen NH; Slingsby C; Tardieu A Ageing and Vision: Structure, Stability and Function of Lens Crystallins. Progress in Biophysics and Molecular Biology 2004, 86, 407–485. [PubMed: 15302206]

(36). Delaye M; Tardieu A Short-Range Order of Crystallin Proteins Accounts for Eye Lens Transparency. Nature 1983, 302, 415–417. [PubMed: 6835373]

(37). Ecroyd H; Carver JA Crystallin Proteins and Amyloid Fibrils. Cellular and Molecular Life Sciences 2009, 66, 62–81. [PubMed: 18810322]

(38). Serebryany E; Woodard JC; Adkar BV; Shabab M; King JA; Shakhnovich EI An Internal Disulfide Locks a Misfolded Aggregation-prone Intermediate in Cataract-linked Mutants of Human γD-Crystallin*. Journal of Biological Chemistry 2016, 291, 19172–19183. [PubMed: 27417136]

(39). Ji F; Jung J; Koharudin LMI; Gronenborn AM The Human W42R γD-Crystallin Mutant Structure Provides a Link between Congenital and Age-related Cataracts. Journal of Biological Chemistry 2013, 288, 99–109. [PubMed: 23124202]

(40). Duong VT; Diessner EM; Grazioli G; Martin RW; Butts CT Neural Upscaling from Residue-Level Protein Structure Networks to Atomistic Structures. Biomolecules 2021, 11, 1788. [PubMed: 34944432]

(41). Yu Y; Grazioli G; Phillips NE; Butts CT Local Graph Stability in Exponential Family Random Graph Models. SIAM Journal on Applied Mathematics 2021, 81, 1389–1415.

(42). Butts CT; Zhang X; Kelly JE; Roskamp KW; Unhelkar MH; Freites JA; Tahir S; Martin RW Sequence Comparison, Molecular Modeling, and Network Analysis Predict Structural Diversity in Cysteine Proteases from the Cape Sundew, Drosera Capensis. Computational and Structural Biotechnology Journal 2016, 14, 271–282. [PubMed: 27471585]

(43). Brinda KV; Vishveshwara S A Network Representation of Protein Structures: Implications for Protein Stability. Biophysical Journal 2005, 89, 4159–4170. [PubMed: 16150969]

(44). Sathyapriya R; Vishveshwara S Structure Networks of E. Coli Glutaminyl-tRNA Synthetase: Effects of Ligand Binding. Proteins: Structure, Function, and Bioinformatics 2007, 68, 541–550.

(45). Besag J Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society: Series B (Methodological) 1974, 36, 192–225.

(46). Frank O; Strauss D Markov Graphs. Journal of the American Statistical Association 1986, 81,832–842

(47). Pattison P; Robins G Neighborhood-Based Models for Social Networks. Sociological Methodology 2002, 32, 301–337.

(48). Snijders TAB; Pattison PE; Robins GL; Handcock MS New Specifications for Exponential Random Graph Models. Sociological Methodology 2006, 36, 99–153.

(49). Handcock MS Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers; National Academies Press: Washington, D.C., 2003; pp 229–240.

(50). Butts CT Bernoulli Graph Bounds for General Random Graphs. Sociological Methodology 2011 , 41, 299–345.

(51). Schweinberger M Instability, Sensitivity, and Degeneracy of Discrete Exponential Families. Journal of the American Statistical Association 2011, 106, 1361–1370. [PubMed: 22844170]

(52). Butts CT Phase Transitions in the Edge/Concurrent Vertex Model. The Journal of Mathematical Sociology 2021, 45, 135–147.

(53). Hunter DR; Handcock MS Inference in Curved Exponential Family Models for Networks. Journal of Computational and Graphical Statistics 2006, 15, 565–583.

(54). Strauss D On a General Class of Models for Interaction. SIAM Review 1986, 28, 513–527.

(55). Häggström O; Jonasson J Phase Transition in the Random Triangle Model. Journal of Applied Probability 1999, 36, 1101–1115.

(56). Hunter DR Curved Exponential Family Models for Social Networks. Social Networks 2007, 29, 216–230. [PubMed: 18311321]

(57). Best RB; Zhu X; Shim J; Lopes PEM; Mittal J; Feig M; MacKerell AD Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the

Backbone ϕ, ψ and Side-Chain $X1$ and $X2$ Dihedral Angles. Journal of Chemical Theory and Computation 2012, 8, 3257–3273. [PubMed: 23341755]

(58). Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML Comparison of Simple Potential Functions for Simulating Liquid Water. The Journal of Chemical Physics 1983, 79, 926–935.

(59). Handcock MS; Hunter DR; Butts CT; Goodreau SM; Morris M statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. Journal of Statistical Software 2008, 24, 1548–7660.

(60). R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2022.

(61). Butts CT network: A Package for Managing Relational Data in R. Journal of Statistical Software 2008, 24, 1–36. [PubMed: 18612375]

(62). Butts CT Social Network Analysis with sna. Journal of Statistical Software 2008, 24, 1–51 [PubMed: 18612375]

(63). Gelman A; Carlin JB; Stern HS; Dunson DB; Vehtari A; Rubin DB Bayesian Data Analysis, 3rd ed.; Chapman and Hall, 2013.

(64). Handcock MS; Morris M Relative Distribution Methods in the Social Sciences; Statistics for Social Science and Behavorial Sciences; Springer-Verlag: New York, 1999.

(65). Kalos MH; Whitlock PA Monte Carlo Methods. Vol. 1: Basics; Wiley-Interscience: USA, 1986.

(66). Hunter DR; Handcock MS; Butts CT; Goodreau SM; Morris M ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. Journal of Statistical Software 2008, 24, 1–29. [PubMed: 18612375]

(67). Hunter DR; Goodreau SM; Handcock MS ergm.userterms: A Template Package for Extending Statnet. Journal of Statistical Software 2013, 52, 1–25. [PubMed: 23761062]

(68). Butts CT A Dynamic Process Interpretation of the Sparse ERGM Reference Model. Journal of Mathematical Sociology 2019, 43, 40–57.

(69). Kamada T; Kawai S An Algorithm for Drawing General Undirected Graphs. Information Processing Letters 1989, 31, 7–15.

(70). Eddelbuettel D; Balamuta JJ Extending R with C++: A Brief Introduction to Rcpp. PeerJ Preprints 2017, 5, e3188v1.

(71). Grant BJ; Rodrigues AP; ElSawy KM; McCammon JA; Caves LS Bio3D: An R Package for the Comparative Analysis of Protein Structures. Bioinformatics 2006, 22, 2695–2696. [PubMed: 16940322]

(72). Basak A; Bateman O; Slingsby C; Pande A; Asherie N; Ogun O; Benedek GB; Pande J High-Resolution X-ray Crystal Structures of Human γD Crystallin (1.25Å) and the R58H Mutant (1.15Å) Associated with Aculeiform Cataract. Journal of Molecular Biology 2003, 328, 1137–1147. [PubMed: 12729747]

(73). Wang B; Yu C; Xi Y-B; Cai H-C; Wang J; Zhou S; Zhou S; Wu Y; Yan YB; Ma X; Xie L A Novel CRYGD Mutation (p.Trp43Arg) Causing Autosomal Dominant Congenital Cataract in a Chinese Family. Human Mutation 2011, 32, E1939–E1947. [PubMed: 21031598]

(74). Mills-Henry IA; Thol SL; Kosinski-Collins MS; Serebryany E; King JA Kinetic Stability of Long-Lived Human Lens γ-Crystallins and Their Isolated Double Greek Key Domains. Biophysical Journal 2019, 117, 269–280. [PubMed: 31266635]

(75). Serebryany E; King JA Wild-type Human γD-crystallin Promotes Aggregation of Its Oxidation-mimicking, Misfolding-prone W42Q Mutant. The Journal of Biological Chemistry 2015, 290, 11491–11503. [PubMed: 25787081]

(76). Boatz JC; Whitley MJ; Li M; Gronenborn AM; van der Wel PCA Cataract-associated P23T γD-crystallin Retains a Native-like Fold in Amorphous-looking Aggregates Formed at Physiological pH. Nature Communications 2017, 8, 15137.

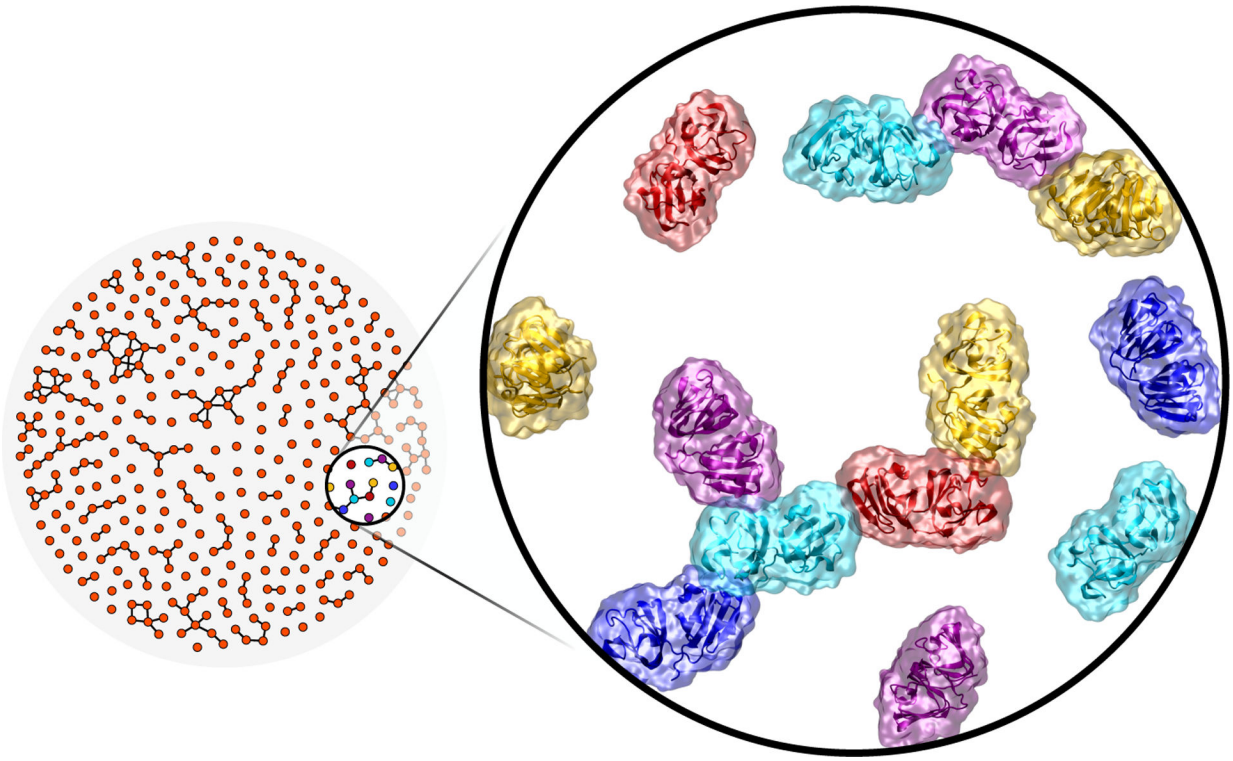


Figure 1:

Example of an aggregation graph of the type studied here. Individual γ-Dc monomers are considered adjacent when they have respective domains whose centers of mass are within 31Å of each other in the atomistic model(see Methods). 2D graph representation shows underlying topology of the aggregate, without regard to spatial positions of the monomers.

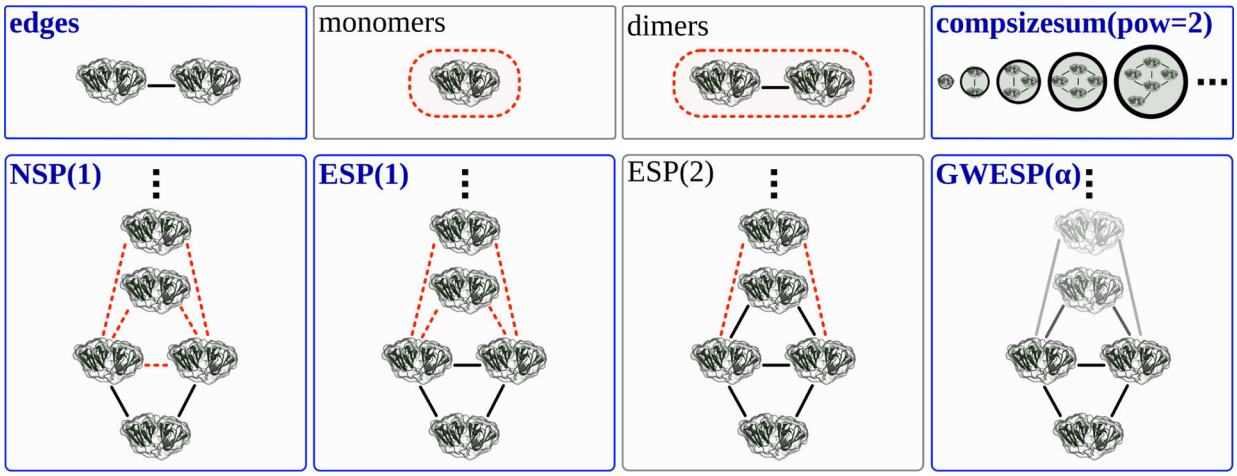


Figure 2:

Schematic representation of candidate model terms for the γ-Dc network Hamiltonian. Black lines indicate edges that must be present in the specified configuration, while red dotted lines indicate edges that must not be present. Blue outline indicates terms selected in the final γ-Dc model. See text for details.

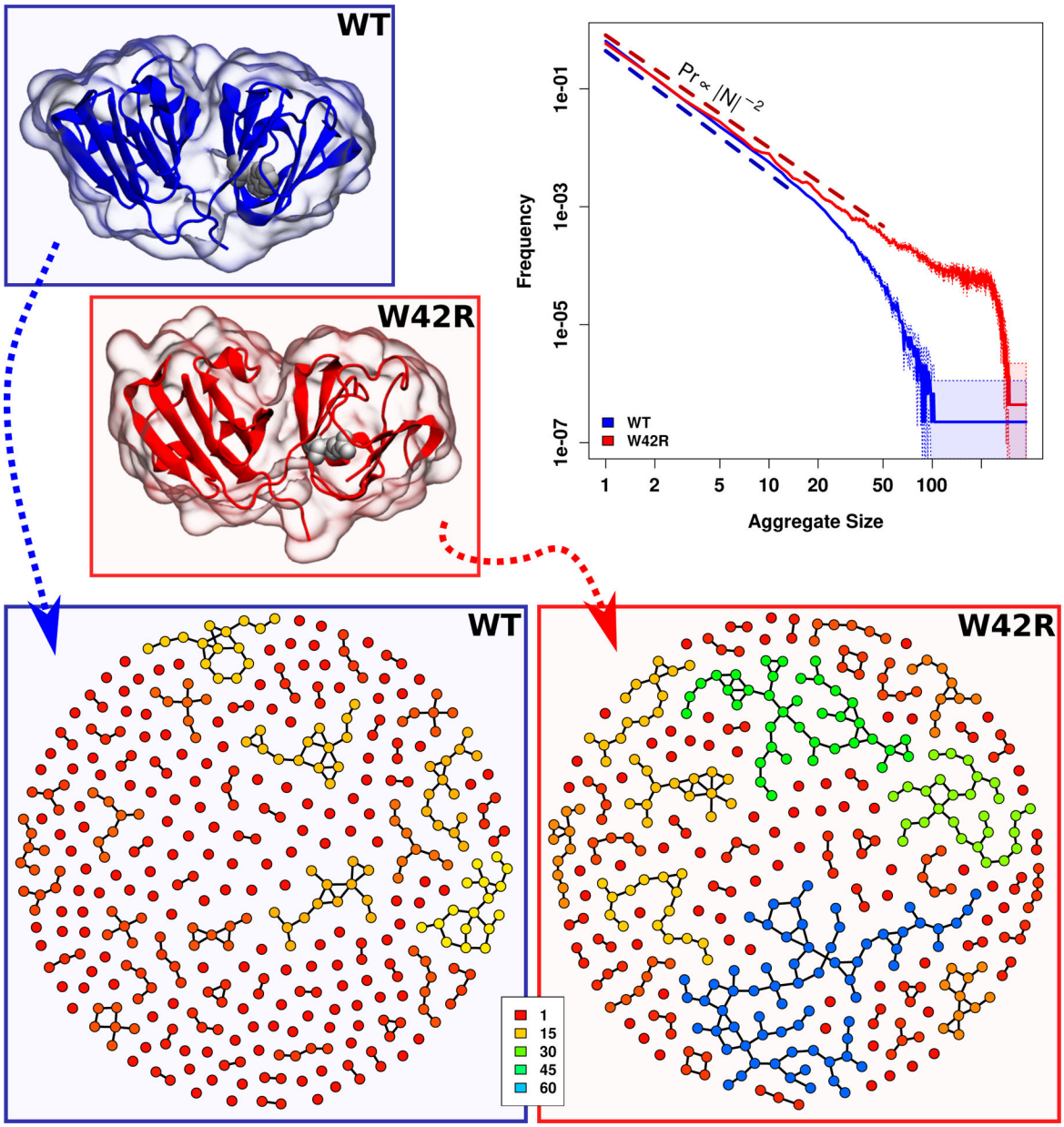


Figure 3:

Aggregate sizes and topologies, from atomistic simulations by Wong et al. (2019). Top left: structures of WT (PDB 1HK0[72]) and W42R (PDB 4GR7[73]) monomers, with residue W42 highlighted. Trp to Arg substitution disrupts the N-terminal domain, increasing exposed hydrophobic surface area. Top right: WT and W42R size distributions are similar for small aggregates, but W42R produces more large structures. Bottom: Representative examples of WT and W42R aggregation graphs illustrate typical differences in topology; vertex colors indicate component size, from red (free monomers) to blue (largest components).

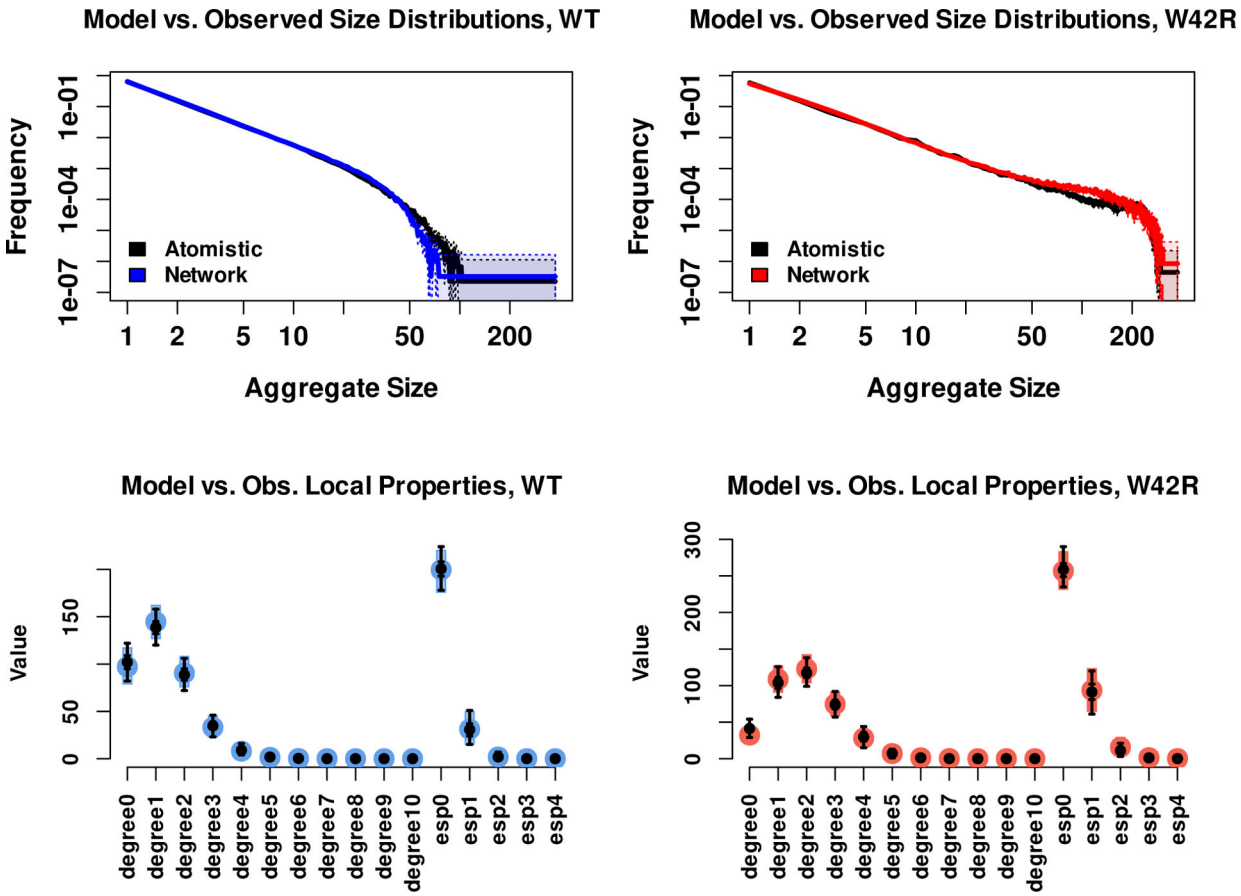


Figure 4:

Model adequacy checks for the network Hamiltonian models. Top panels compare observed (black) to simulated (colored) aggregate size distributions (center line indicates posterior mean, shaded area 95% posterior intervals). Bottom panels compare observed (black) versus simulated (colored) distributions of local structural properties, specifically degree and edgewise shared partner counts; dots indicate means, whiskers indicate 95% intervals. For both WT and W42R, the selected models successfully approximate the behavior of the atomistic simulations.

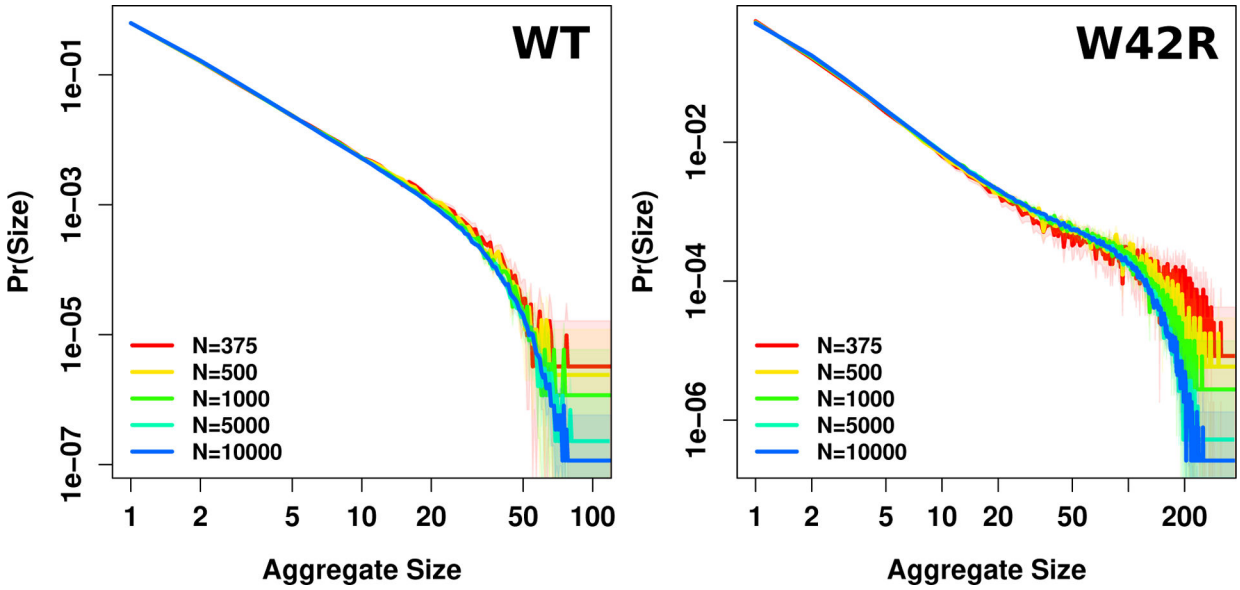


Figure 5:

Predicted aggregate size distributions, by system size and variant. Center lines indicate posterior means; shaded areas indicate 95% posterior intervals. While distributions remain similar, maximum aggregate sizes decline more sharply when system sizes become large compared to the size of the largest aggregates.

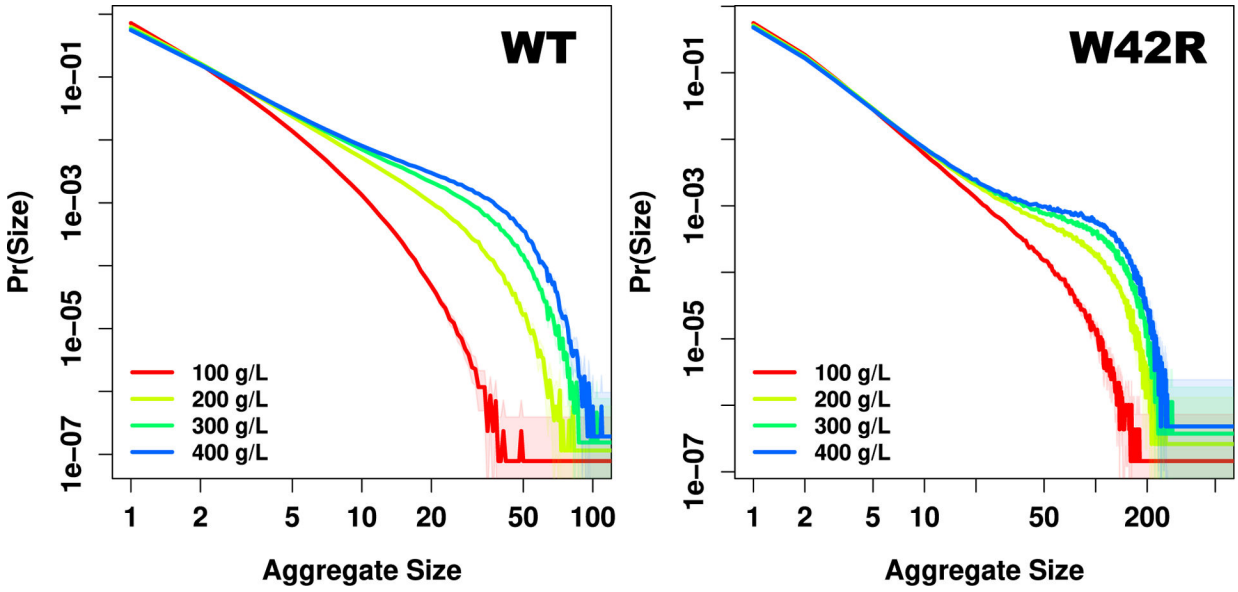


Figure 6:

Predicted aggregate size distributions, by concentration and variant, at $N = 10^4$. Center lines indicate posterior means; shaded areas indicate 95% posterior intervals. Increased concentration favors growth of larger aggregates, particularly increasing the large-aggregate population in W42R.

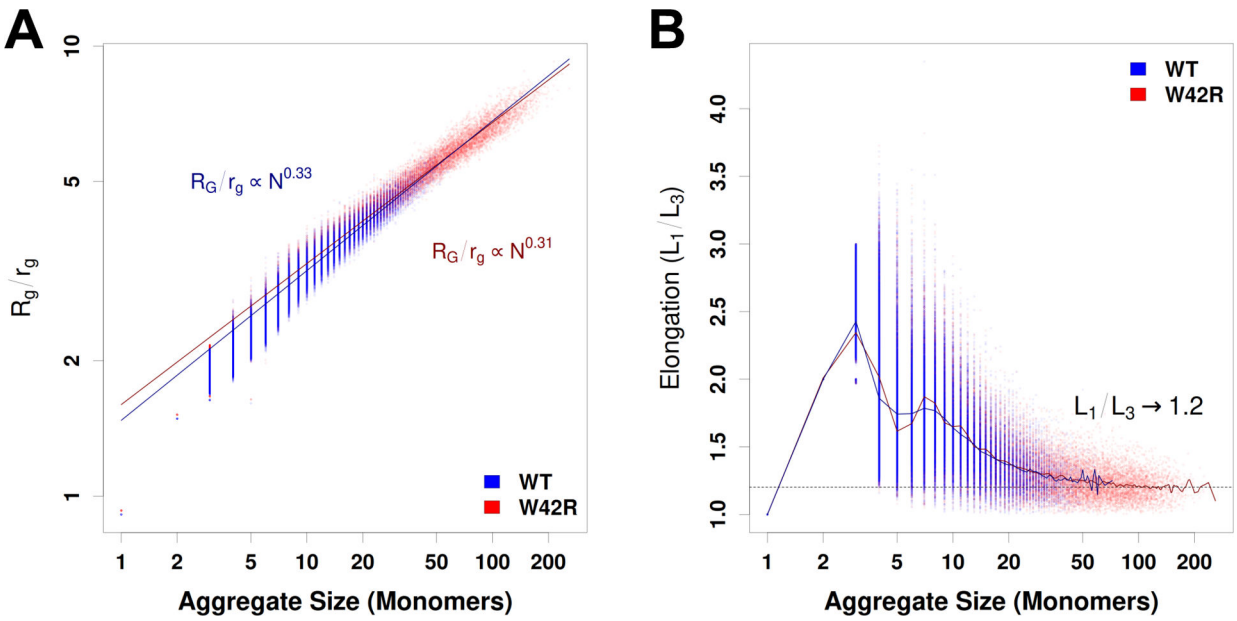


Figure 7:

(A) Projected aggregate R_g over monomer radius of gyration (r_g) by aggregate size. For large aggregates (≥ 10 monomers), scaling is close to $N^{1/3}$, though slightly below for W42R.

(B) Elongation factor (largest axis over shortest axis) by aggregate size; smoothing splines shown to indicate mean behavior. Larger aggregates approach a limiting elongation factor of approximately 1.2.

Table 1:

Selected models for γ-Dc aggregates, by selection stage. Columns 1–5 indicate included terms; terms selected by steepest descent, and no other terms were found to improve fit. Error for observed (f_{obs}) versus model-predicted (f_{sim}) aggregate size distributions given for WT, W42R, and combined cases. Relative gain shows fraction of total error reduction versus the baseline (edge-only) model.

Model Terms					Error($\|\|\log(f_{obs}/f_{sim})\|\|$)			**Rel. Gain**
edges	**NSP(1)**	GWESP(decay=α)	**compsizesum(power=2)**	**ESP(1)**	**WT**	**W42R**	**Total**	
TRUE	FALSE	FALSE	FALSE	FALSE	0.67	1.17	1.84	–
TRUE	TRUE	FALSE	FALSE	FALSE	0.37	0.68	1.05	43%
TRUE	TRUE	TRUE	FALSE	FALSE	0.29	0.26	0.55	27%
TRUE	TRUE	TRUE	TRUE	FALSE	0.24	0.17	0.41	8%
TRUE	TRUE	TRUE	TRUE	TRUE	0.24	0.15	0.38	2%

Table 2:

Estimated model coefficients for γ-Dc aggregate models; θ specifies ERGM form at simulated temperature and N, ϕ indicates equivalent Hamiltonian representation. All coefficients significant at $p < 1 \times 10^{-4}$; apparent zero standard errors indicate SE$< 1 \times 10^{-4}$.

	WT			W42R		
Term	$\hat{\theta}$	**Std. Err.**	$\hat{\phi}$ **(kcal/mol)**	$\hat{\theta}$	**Std. Err.**	$\hat{\phi}$ **(kcal/mol)**
edges	−5.2546	0.0061	−1.0302	−3.9911	0.0066	−1.8085
NSP(1)	−0.2163	0.0034	0.1332	−0.4036	0.0025	0.2486
GWESP(α)	1.2855	0.0132	−0.7919	1.1983	0.0090	−0.7382
α	0.5			0.3		
compsizesum(power=2)	−0.0016	0.0001	0.0010	−0.0003	0.0000	0.0002
ESP(1)	−0.1165	0.0144	0.0718	−0.2197	0.0098	0.1353