UNIVERSITY OF CALIFORNIA,
IRVINE

Using Network Models to Relate Local Interactions with Global Topology: Applications to Protein Interactions and Emergent Multi-Body Structures

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Elizabeth M. Diessner

Dissertation Committee:
Professor Carter T. Butts, Chair
Professor Ioan Andricioae
Professor Rachel W. Martin

2024

DEDICATION

To MeowMeow, and everyone else.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to acknowledge the publishers at Biochemistry and The Journal of Physical Chemistry B where materials in the following chapters have previously been published.

This research was supported by NSF awards 1633631 to E.M.D., and DMS-1361425 to C.T.B. and R.W.M., NASA Award 80NSSC20K0620 to R.W.M. and C.T.B., NIH award 1R01GM144964-01 to C.T.B., NIH grant 2R01EY021514 to R.W.M., the CIFAR Molecular Architecture of Life program, and the UC Irvine School of Physical Sciences Dean's Excellence Fund.

VITA

Elizabeth M. Diessner

EDUCATION

Doctor of Philosophy in Chemistry	**2024**
University of California, Irvine	*Irvine, CA*
Bachelor of Science in Analytical and Environmental Chemistry	**2019**
George Mason University	*Burke, VA*
Associates of Science in Science	**2016**
Northern Virginia Community College	*Annandale, VA*

RESEARCH EXPERIENCE

Graduate Research Assistant	**2020–2024**
University of California, Irvine	*Irvine, CA*
Undergraduate Research Assistant	**2018–2019**
George Mason University	*Burke, VA*

TEACHING EXPERIENCE

Teaching Assistant	**2019–2021**
University of California, Irvine	*Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Comparative Modeling and Analysis of Extremophilic D-Ala-D-Ala Carboxypeptidases **2023**
Biomolecules

Mutation Effects on Structure and Dynamics: Adaptive Evolution of the SARS-CoV-2 Main Protease **2023**
Biochemistry

Network Hamiltonian Models for Unstructured Protein Aggregates, with Application to γD-Crystallin **2023**
The Journal of Physical Chemistry B

Active Learn Module for Protein Structure Analysis Using Novel Enzymes **2022**
The Biophysicist

Neural Upscaling from Residue-Level Protein Structure Networks to Atomistic Structures **2021**
Biomolecules

A Cyclic Peptide Inhibitor of the SARS-CoV-2 Main Protease **2021**
European Journal of Medicinal Chemistry

Sequence Characterization and Molecular Modeling of Clinically Relevant Variants of the SARS-CoV-2 Main Protease **2020**
Biochemistry

ABSTRACT OF THE DISSERTATION

Using Network Models to Relate Local Interactions with Global Topology:
Applications to Protein Interactions and Emergent Multi-Body Structures

By

Elizabeth M. Diessner

Doctor of Philosophy in Chemistry

University of California, Irvine, 2024

Professor Carter T. Butts, Chair

Local interactions within and between proteins (or interacting objects in general) inherently determine the resulting global structure, whether that be a monomeric protein structure, a dimer or multimer, or a larger aggregate consisting of tens to thousands of proteins. For proteins, structure is canonically partitioned into four levels: primary, which describes the sequence of residues that make up the protein; secondary, the α-helices and β-sheets that result from hydrogen-bonding interactions between residues; tertiary, which describes (somewhat arbitrarily defined) domains of clustered secondary structures that are typically held together with salt-bridges; and finally, quaternary structures composed of multiple proteins interacting via hydrogen-bonding or other polar interactions. Variants are proteins with point mutations, or mutations occurring to a small number (typically one) of the amino acids in the primary structure. Point mutations can alter the higher-order structure and dynamics of the protein, and thus how it responds to its environment, making it susceptible to evolutionary forces that dampen or put emphasis on a given variant. Such changes in structure and dynamics can range from subtle deformations to changes in the way the protein folds, inhibiting function. Mutations that are favored by evolution provide information about how the protein's relationship with its environment affects its function and applies pressure to the adaptative evolution of the protein. The effects of mutations on protein structure,

function, and interactions are explored in chapters two and three of this text. To contrast, the fourth chapter takes a generalized approach by delving into the range of emergent multi-body structures that can arise from slight changes in environmental or structural parameters while remaining agnostic to any specific features of a single protein sequence.

Chapter 1

Introduction

1.1 Analysis of (M^{pro}) Variants

Chapter 2 describes analysis of the SARS-CoV-2 main protease (M^{pro}), in which the effects of point mutations (observed in 1253 variants collected from all reported clinical samples during the first year of the COVID-19 global pandemic) was studied. Protein structure networks (PSNs) were used to characterize changes in global cohesion - as well as cohesion of the domains - of the monomer and dimer structures of M^{pro}. Network analysis revealed significant trends towards less cohesive structures in all cases *except* the domain II of the dimer, which was observed to maintain similar levels of cohesion across all variants. This observation was made despite a lack of significant change in torsion angles of residue side-chains, as well as trends towards increasingly large and more hydrophobic residues. This combination of observations indicate that the protein's adaptation forgoes more stable internal interactions between residues within domains I and III, perhaps because they are not needed to maintain the structure necessary for functional dynamics in the thermodynamic environment of the human host. Domain II, however, saw the largest number of conserved residues (those that are not mutated in any variant in this sample), most of which are polar and aromatic, suggesting the internal residue interactions of domain II are vital to maintaining the structure and functional dynamics of the M^{pro}dimer.

In the case of M^{pro}, statistical analysis of mutations points out the obvious by stating which residues stay or go. However, supplementing that data with calculations using PSNs allows the story of M^{pro}'s evolution in human hosts to be rebuilt by aggregating the data provided by each variant regarding local interactions. Using the initial wild type (WT) variant as an anchor for the distribution of variants, these analyses allowed comparison of the movement of the total variant distribution relative to WT, which ultimately highlighted the importance of residue interactions occurring in domain II for function and dynamics of the dimer.

The following two chapters move beyond residue interactions *within* a protein that accommodate dimer formation, and instead focus on the interactions *between multiple proteins* that form large structured and unstructured aggregates.

1.2 Simulations of a γD-Crystallin Cataract Variant

γD-Crystallin (γ-Dc) is a structural protein of the eye lens that has evolved to maintain its structure for the span of a human lifetime without unfolding. However, unfolding events continue to occur naturally due to e.g., collisions or structural damage caused by exposure to sunlight, and as chaperone proteins such as the α-crystallins are depleted aggregation increases and cataracts form. The γ-Dc variant W42R is known to unfold more readily[172, 209], increasing the opportunity for longer-lasting interactions that can lead to aggregation and subsequent cataract formation. Chapter 3 reports on analyses of unstructured γ-Dc aggregates done using Network Hamiltonian Models (NHMs) that were generated from equilibrium distributions of simulated aggregation of both the WT and W42R variant, performed by Wong, et al.[209] NHMs provide a framework for inputting a set of network terms that describe local topological properties of the system, providing a model that can be used in a regression to return coefficients describing the relative influence of local topologies on the overall structure of the system. Using this model, the results of atomistic simulations of aggregation were recapitulated using minimal information about the patterns of contacts between individual proteins, where contacts are defined using a cutoff distance that was determined by Wong, et al.[209]

Due to the reduced amount of information needed to simulate the aggregates (only an adjacency matrix is needed) the system size can be scaled from the original 375 monomers of the atomistic simulations up to 10,000 monomers per simulation. The ability to scale the system by orders of magnitude not only allows simulation of systems that can be more easily corroborated by experimental data, but also allowed the effects of system size on the resulting structural and topological properties of the aggregates to be assessed. Such analyses give insight into hidden biases in simulation studies that arise from system size effects that were previously immeasurable when confined to using the more complex and costly all-atom simulations.

1.3 Phases of Structured Aggregation

The framework of NHMs allows for exploration of the effects of specific topological forces on the resulting multi-body aggregate structures. In Chapter 4, phases of fibrillar aggregates, a type of structured aggregate that is commonly observed from aggregation of intrinsically disordered peptides (IDPs), were mapped by varying the values of the coefficients on a minimal set of network terms that were included in the Network Hamiltonian. Without reference to a specific protein, these phases can be understood to be the result of slight changes to environmental parameters in which the aggregate was formed. Such parameters include those that affect local thermodynamics, such as pH, salinity, hydrophobicity, and pressure, as well as parameters describing differences in protein structures, such as sequence or specific electrostatic interactions between proteins. Environmental parameters are implicit in the model, only appearing in their effect on the energy of each interaction (as reflected in the network term coefficients), or as part of the unmodeled degrees of freedom that are represented by the reference measure.

This model revealed an intrinsic dependence of global topologies on the baseline edge coefficient value, a parameter that defines the energetic cost of adding or removing *any* edge, regardless of the local topology. In addition, the relationships among network terms that describe similar topological features was observed to directly influence the location of phase boundaries, and were defined using simple linear equations. System size effects were also analyzed in this study, and, in contrast with the results found in Chapter 3, show no dependence of structured aggregate formation on system size.

The protein-agnostic perspective in Chapter 4 highlights the importance of considering environmental effects when performing simulation studies of protein interactions, as slight changes to thermodynamic properties of the system can have measurable effects on the resulting structures that are being studied. This echos the findings in Chapter 3 based on the slight change of a single point mutation. Chapters 3 and 4 also bring to light the need to consider biases in simulations that arise from system size effects. Chapter 2 drives the point that single point-mutations can have far-reaching effects on protein interactions and functions, while utilizing the effects of evolutionary pressures on protein adaptation. In all, these studies detail the sensitive relationship between multi-body protein interactions and the environment in which they occur.

Chapter 2

Mutation Effects on Structure and Dynamics: Adaptive Evolution of the SARS-CoV-2 Main Protease

2.1 Abstract

The main protease of SARS-CoV-2 (M^{pro}) plays a critical role in viral replication; although it is relatively conserved, M^{pro} has nevertheless evolved over the course of the COVID-19 pandemic. Here, we examine phenotypic changes in clinically observed variants of M^{pro}, relative to the originally reported wild-type (WT) enzyme. Using atomistic molecular dynamics simulations, we examine effects of mutation on protein structure and dynamics. In addition to basic structural properties such as variation in surface area and torsion angles, we use protein structure networks (PSNs) and active site networks (ASNs) to evaluate functionally

relevant characters related to global cohesion and active site constraint. Substitution analysis shows a continuing trend toward more hydrophobic residues that is dependent on the location of the residue in primary, secondary, tertiary, and quaternary structure. Phylogenetic analysis provides additional evidence for the impact of selective pressure on mutation of M^{pro}. Overall, these analyses suggest evolutionary adaptation of M^{pro} toward more hydrophobicity and a less-constrained active site in response to the selective pressures of a novel host environment.

2.2 Introduction

The SARS-CoV-2 main protease (M^{pro}), also referred to as non-structural protein 5 (nsp5) or 3-chymotrypsin-like cysteine protease (3CL^{pro}), is a vital component of the coronavirus replication machinery [7]. During replication, the host ribosomes translate the SARS-CoV-2 non-structural proteins (nsps, i.e., enzymes) as a long polyprotein; this must then be cleaved into individual proteins to complete the expression and maturation process. In SARS-CoV and SARS-CoV-2, this cleavage function is performed by two proteases: the papain-like protease (PL^{pro}), and M^{pro} [230, 184]. The first three cleavage sites, corresponding to the release of nsp1-nsp3, are cleaved by PL^{pro}, with the remaining 11 cleavage sites handled by M^{pro}, including those needed to release M^{pro} itself [154, 215]. M^{pro} is thus necessary for maturation of the bulk of the proteins comprising the SARS-CoV-2 replicase [75]. M^{pro} also targets several proteins in the host cell, including key components of the cytokine and inflammatory responses [75, 127].

M^{pro} itself is a cysteine protease, in which hydrolysis is performed by a catalytic dyad composed of a neutral (protonated) cysteine (C145) and a histidine (H41); this mechanism is strongly conserved among coronaviruses [7, 8, 199]. M^{pro}'s active conformation is a homodimer [7], although limited activity of M^{pro} monomers has been reported [179]. Despite its

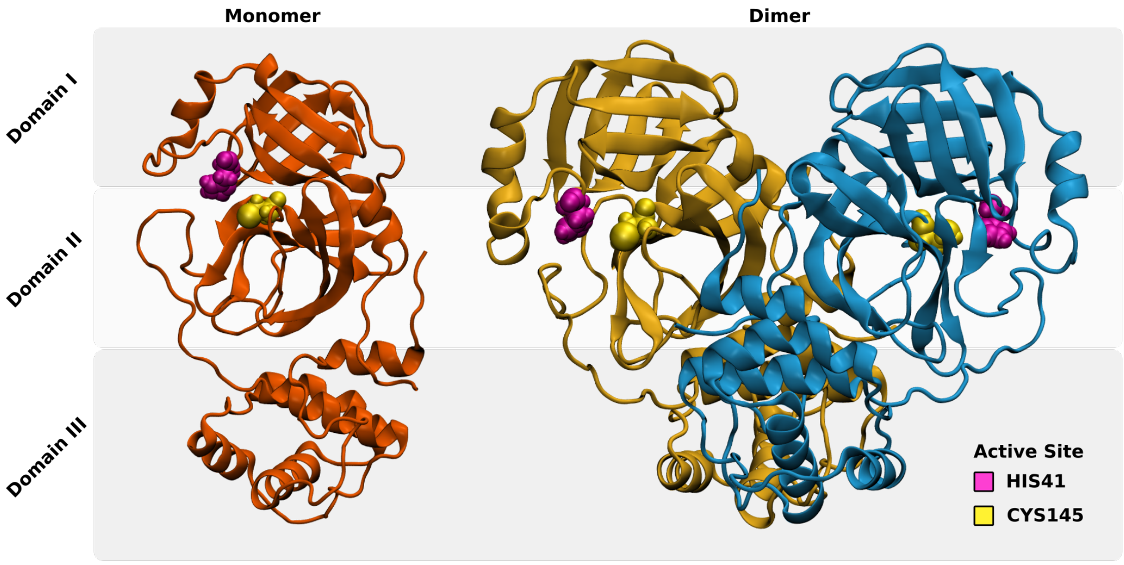


Figure 2.1: Monomer and dimer conformations of the wild-type SARS-CoV-2 main protease (M^{pro}), based on respective atomistic molecular dynamics simulations of the free monomer (left) and dimer (right); MD simulations were based on the 6Y2E PDB crystal structure of M^{pro} [225], as described in the Methods section. Note the three domains (highlighted, left); the active site straddles the cleft between domains I and II, and faces away from the dimerization interface.

greatly reduced activity, molecular modeling suggests that the monomer is likely to be stable under physiological conditions, with a conformation that is similar to its conformation in the active homodimer [53]. Monomer and dimer structures, labeled by domain, are shown in Figure 2.1.

SARS-CoV-2 is believed to have transferred to the human population from zoonotic origin [211, 9], and shares particular similarity with a number of bat coronaviruses [189]. While mutations to the infamous spike protein capture the attention of the public [36], other coronavirus proteins are also subject to evolutionary change, either due to neutral drift or as an adaptive response to environmental pressure. When adapting to a new host organism, selection pressure may be imposed by differences in the internal environment of host cells. For instance, bats experience a larger range of body temperatures compared with humans [181, 158], including periods of activity at very high temperature [122, 54]. Differences in host body temperatures impose different thermodynamic and kinetic constraints on the structure and activity of viral proteins within cells, which is a known factor limiting inter-species virus transmission [124, 123] as well as tissue tropism within a single host [197, 175].

As shown in studies of extremophilic organisms, the stability and catalytic efficiency of enzymes is dependent on their thermal environments [26, 125]. Proteins in organisms that regularly experience high temperatures require stronger and more extensive interactions among residues, such as disulfide bonds and salt bridges to maintain stability [107, 97, 74], whereas proteins in low-temperature regimes require greater internal flexibility to facilitate catalysis [188]. The large and abrupt fluctuations in body temperature of bats are representative of frequent thermodynamic changes that put different kinds of stress on proteins, which may require particular structural responses to maintain structure and activity [188].

Beyond structural effects, mutations may also affect dynamics. Changes to local structure near the active site are particularly relevant, since such changes can affect both protein-substrate interactions and catalysis. Stronger side-chain interactions within the active site, for instance, may increase constraint on the dynamics of the catalytic residues. At the same time, long-range effects of residue substitution are known [213, 13], suggesting that functionally relevant mutations may occur throughout the protein, as already observed for HIV protease [139] and SARS-CoV M^{pro} [15].

For SARS-CoV-2 M^{pro}, then, selection for successful replication in a novel host environment is likely to favor systematic changes in protein structure and dynamics, which in turn will favor specific patterns of substitution. Such patterns may or may not be evident from sequence alone, because many different mutations may lead to similar physical properties; however, if present, selection pressure should manifest as consistent differences between structural and dynamic properties of WT M^{pro} versus ecologically successful mutants. By contrast, functionally critical properties that must be conserved between human and prior hosts would be expected to remain similar for both WT and successful variants, and properties under neutral drift would be expected to show variation with no systematic change from WT. Examination of structure and dynamics across a large range of ecologically successful mutants compared to WT thus provides evidence regarding adaptation by M^{pro} to its new environment. Early studies have suggested that some M^{pro} variants do differ from WT in structure and dynamics [176, 134, 53], motivating a systematic comparative analysis.

In this study, we identify evidence of selective pressure on the evolutionary adaptation of M^{pro} by analyzing results from molecular dynamics simulations and network analysis of all 1253 clinically identified variants of M^{pro} that were reported to the GISAID database over the first year of the COVID-19 pandemic (i.e., before February 25, 2021). Focusing on clinically observed variants allows us to work with mutations that were both functional and ecologically successful, in that they could successfully infect human hosts "in the wild." To distinguish between effects arising directly from changes to the structure of the M^{pro} monomer and those emerging only in the dimeric state, we examine models of both the functional dimer and the free monomer in solution. Trends in physical properties of variants relative to WT are assessed using multiple techniques. Relative Solvent Accessibility (RSA) is used to calculate total surface area, providing preliminary information on the effect of mutation on global structure. Internal changes to structure are further investigated by analysis of Protein Structure Networks (PSNs) to observe changes in internal residue interaction rates. The effects of substitutions on local dynamics are observed by comparing variation in torsion angles - extracted from dynamic simulation trajectories - between and within variants. Active Site Networks (ASNs) of each variant are constructed to measure local constraint on the active site. Finally, we investigate trends in the physical properties of amino acid substitutions, and explore the ways the location of certain substitutions - or lack thereof - contribute to a response to selective pressure that may be guiding the adaptive evolution of M^{pro}.

The results of the following analyses provide a rich context for understanding the physical adaptation of M^{pro}, and suggest a number of targets for experimental investigation, which will be required to probe the impact of the observed mutations on catalytic activity and kinetic parameters. Compared with WT, variants are observed on average to have more solvent-accessible surface area (SASA), indicating either an increase in size of surface residues or a loosening of internal structure. In the monomeric state, M^{pro} is observed to have lower cohesion overall, contributing to the loosening of the structure, while the dimeric state conserves internal interactions in domain 2. Backbone torsion angles are generally similar between the monomeric and dimeric states, with mutations having the greatest impact on the backbone structure of residues in domain 2 of both states. The two active sites of the dimeric state trend towards less constraint on the catalytic residues, but the monomeric state shows no definite trend, despite the similar effects of mutation on the structure of the monomeric and dimeric states. The substitutions themselves generally trend towards more hydrophobic residues, with certain frequently occurring mutations near the active site showing a trend toward more hydrophilic residues. Frequently observed mutations, including some located near the active site, have occurred in several unique branches, indicating a possible benefit to M^{pro} function that is supported by selective pressure on the enzyme.

2.3 Methods

2.3.1 Sequence Preprocessing

Human-derived SARS-CoV-2 full genome sequences were retrieved from the GISAID EpiCoV database [103] on February 25, 2021 at 10:15 AM (PST). These were filtered for size and quality; those with <1 percent N content and lengths within +/-3 percent of the length of a designated WT sequence (RefSeq: NC_045512.2 [210]) (29,006 bp–30,800 bp inclusive) were retained for further processing. High-quality sequences were filtered for valid M^{pro} sequences, and then again for modellable M^{pro} sequences. For our purposes, "valid" sequences refer to those with no frameshifts, deletions, insertions, Ns, or non-standard IUPAC nucleotides (those other than A, C, U, G); "modellable" sequences are valid M^{pro} sequences with no non-synonymous mutations that result in either changes to the active site (H41 or C145) or premature stop codons, as the true functionality and/or translated structures of these variants are currently unknown. These M^{pro} sequences were located in and extracted from full genomes by using six 15-nucleotide keys, derived from the NC_045512.2 reference M^{pro} sequence (loc: 10,055–10,972). All sequence preprocessing was done using custom scripts in Python (v3.7.6) [196].

2.3.2 Alignments

All full genome alignments were performed using suggested MAFFT (v7.471) [101] protocols for SARS-CoV-2 (https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html). Full genomes were aligned to a WT reference (NC_045512.2), using the options "–auto" and "addfragments"; in order to retain site information for phylogenetic analysis, the "–keeplength" option was not used.

2.3.3 Clustering and Phylogenetic Tree

A phylogenetic tree was constructed for aligned full genomes that contained non-WT, modellable M^{pro} variants using FastTree (v2.1.11 SSE3) [149, 150] with OpenMP [55] (FastTreeMP); the "-fastest" option was used. This included 70,246 full genomes with non-synonymous M^{pro} mutations (considered "variants") and 34,909 full genomes with synonymous M^{pro} mutations (same protein sequence as WT). One WT full genome reference, (NC_045512.2) was also included. Visualizations were generated in R (v4.0.4) [156] using ggtree [214], ape [143], ggplot2 [204], treeio[201], tidyverse [205], ggtreeExtra [214], aplot [220], data.table [60], svglite [206].

2.3.4 Molecular Modelling of WT and Variant Structures

Monomer and dimer conformations of variant structures were predicted with MODELLER 9.23 [202] using the PDB structure 6Y2E [225] as the WT template. All structures underwent three rounds of annealing and MD refinement using "slow" optimization. The protonation states were corrected for the predicted cell environment using PROPKA 3.1 [140]. The corrected structures were minimized and equilibriated in explicit solvent. MD trajectories were then simulated from the corrected structures using NAMD [147] with a CHARMM36 [86] force field and TIP3P water at 310 K under periodic boundary conditions for a water box with a 10 Å margin in an NpT ensemble. Solvated models were energy-minimized for 10,000 iterations, then simulated once for 10 ps to make water box size adjustments (for PME calculations), and once more for a 10ns trajectory with sampled conformations saved every 20 ps. Temperature control was maintained via Langevin dynamics with a damping coefficient of 1/ps, and pressure control was performed via a Nose-Hoover Langevin [69] piston set at 1 atm. Visualizations and other static analyses are based on the final conformations from each trajectory, with full trajectories used for dynamic analyses. Visualizations were performed using VMD [87]. Solvent accessible surface area calculations were performed using the `dssp.pdb` function in the bio3d library in R [76].

2.3.5 Network Analysis

All frames from each respective simulated trajectory were individually translated into PSNs using scripts written using the statnet, Rpdb, and bio3d libraries in R [83, 29, 178, 76]. Vertices for each network follow the convention established by Benson and Daggett[19] - atoms are grouped into chemical moieties, each of which is represented by a node. Each residue is thus represented by a collection of nodes, and an edge (tie) is formed between two nodes when there exist respective atoms associated with each node that lie within a threshold distance of each other in the selected frame. The distance cutoff used here is 1.1 times the sum of the respective van der Waals radii of the two atoms. An ASN [62] was constructed for the active site of each variant structure by inducing a subgraph comprised of the nodes representing Cys 145, His 41, and all adjacent vertices from the respective PSN. PSNs and ASNs were calculated for all frames from each trajectory, all of which were used in the reported analyses.

Analyses of the PSNs used degree k-cores [169] to characterize the cohesion of each monomer and dimer chain, with the core number of each node (i.e., the highest k such that the node belongs to the kth core) being employed as a measure of local cohesion. Mean core numbers for vertices within each domain, and for the protein as a whole, were used to assess cohesion; all quantities were computed within each frame, with trajectory averages used as for structural comparison. Autocorrelation-corrected bootstrap standard errors were calculated to control for within-trajectory temporal autocorrelation in the trajectory means, and variant values were treated as significantly different from WT if they differed by more than two standard errors. Calculations were done using the sna library in R [30]. Analyses of ASNs included calculations of degree, triangle degree, core number, and connectivity, each averaged over the active site. Here, degree refers to the number of ties a particular node has - i.e. the total number of contacts. Triangle degree refers to the number of triangles containing a particular vertex, and core number for these analyses was assessed within-ASN (as opposed to core number within the broader PSN). Connectivity was measured using the log of the number of indirect paths between the two active site residues. Together, degree, triangle degree, core number, and connectivity give an indication of the freedom of movement within the active site. This gives an approximation of an active site state, which can be used to distinguish active site conformations which are more "open" or "closed." Quantitatively, we assess this via a *constraint score,* which is the score of each network on the first principal component of the combined and standardized degree, triangle degree, core number, and connectivity measures.

2.4 Results and Discussion

2.4.1 Variants Tend Toward Less Compact M^{pro} Structure

Surface area increases, but more so in the monomer than the dimer. Overall, the most common effect of mutations on the monomer conformation is to increase the surface area of the enzyme, as shown in Figure 2.2, with 46.3% of variants with increased surface area (p-value = 0.01 using an exact binomial test), 53.1% with no change (p-value = 0.03), and 0.6% with decreased surface area (p-value $<2.2\text{x}10^{-16}$). This could be a side effect of bulkier residues, or the result of a decrease in internal interactions. Alternatively, bulky and hydrophobic residue substitutions in the interior could cause the structure to expand outward to accommodate the larger side-chains.

The increase in surface area of the monomer is less pronounced in the dimeric conformation: although we do see a net tendency towards SASA increase (28.4% increased, 7.4% decreased, and 64.6% stay the same, p-values $<2.2\text{x}10^{-16}$ using an exact binomial test), fewer variants show significant differences, and the location of WT within the distribution is less skewed. This suggests that surface enlargement occurs disproportionately within the dimerization interface, resulting in a total surface area that is more conserved upon dimerization. That said, we still observe a significant bias towards higher-SASA dimers, which is consistent with selection favoring a somewhat looser, enlarged protein surface.

Global cohesion is lower in the majority of variants, except for domain 2 in the dimeric state. Looking at the impact of substitution on cohesion within free M^{pro} monomers, we see a consistent pattern of structural "loosening" relative to WT, with 78.5% of variants showing significantly lower levels of cohesion, versus 0.3% showing higher levels (p-value $<2\text{x}10^{-16}$ using an exact binomial test). PSNs measuring internal interaction rates between moieties show a decrease in internal cohesion in all domains of the monomer (Fig.

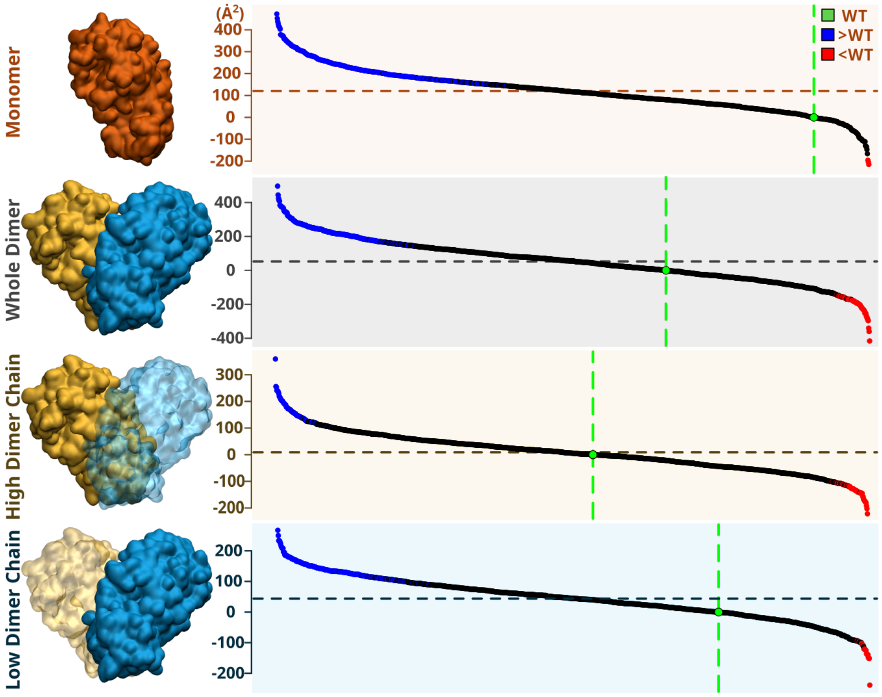


Figure 2.2: Total mean SASA distribution of the monomer, dimer, and each dimer chain, across variants. WT value is in green; trajectories significantly higher than WT are shown in blue, lower in red (black values do not differ significantly from WT). Substantially more variants show increased SASA versus WT than decreased SASA. This is particularly true for free monomers, suggesting that mutations act in part through modifications to interfacial surface that is buried in the dimer.

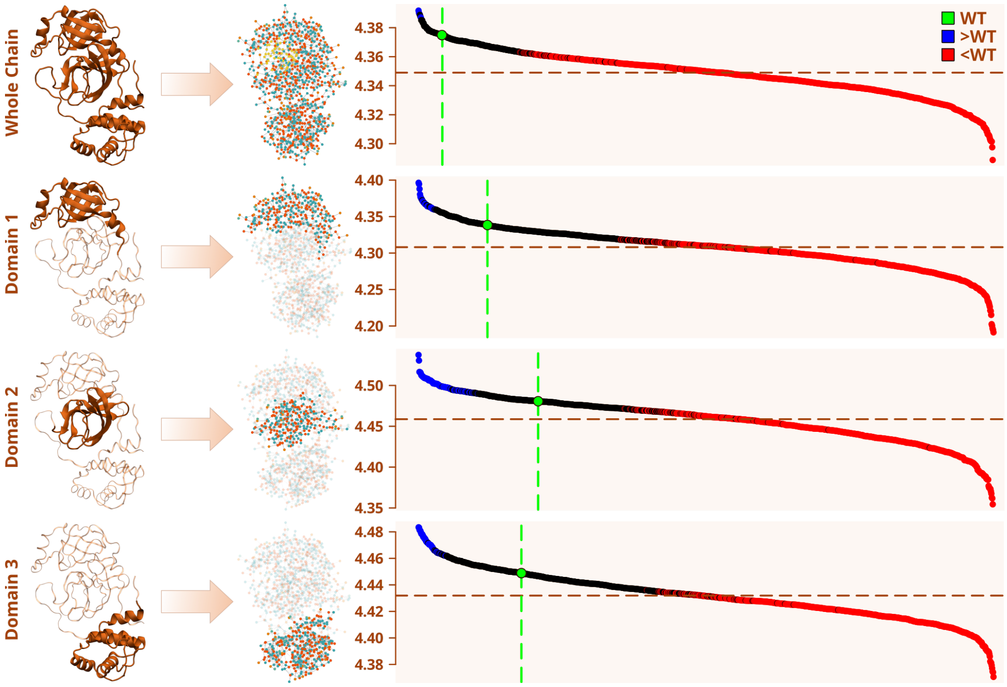


Figure 2.3: Mean cohesion of variants in the monomeric conformation, in decreasing order. WT is highlighted in green. Variants with mean cohesion scores significantly greater than WT are colored blue, and those significantly less than WT are colored red. A horizontal line through the distribution marks the grand mean. The majority of variants show less cohesion both for the monomer as a whole, and in each domain.

2.3), with a slightly reduced degree of loosening in domain 2. This suggests selection for increased flexibility at the level of individual proteins, possibly as a result of the more moderate thermal environment of the human host.

Is this monomer-level change retained upon dimerization? Fig. 2.4 shows that this pattern of reduced cohesion is largely preserved, with looser structures seen in entire dimerized chains, as well as internally within domains 1 and 3. Domain 2, however, shows a rather different pattern, with no clear evolutionary trend: indeed, a substantial fraction (33.1% in the high-cohesion chain and 13.6% in the low-cohesion chain, p-values $<2\text{x}10^{-16}$ using an exact binomial test) actually show enhanced cohesion versus WT. The presence of diversification (with some variants higher, others lower, and relatively few remaining similar) is compatible with the notion that domain 2 within the dimeric state is not being actively selected with respect to cohesion, and is subject to neutral drift. It is interesting to observe in this regard

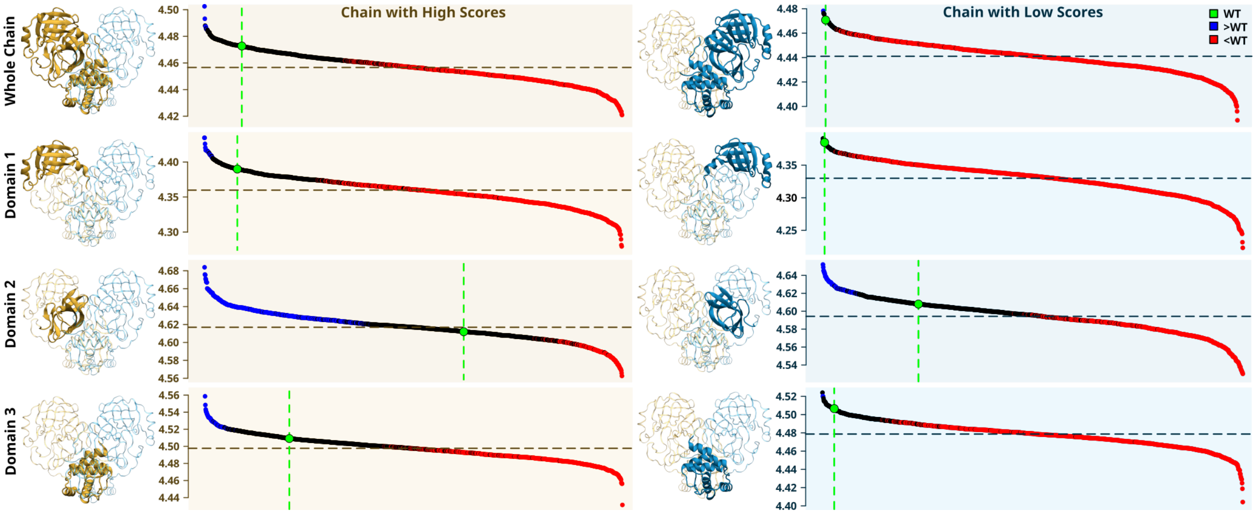


Figure 2.4: Mean cohesion values for of all variants in the dimeric state. The same plot style and color scheme are used as in Figure 2.3. To break homodimer symmetry, chains were labeled for analysis based on the observed mean cohesion score (left higher, right lower).

that we do see a cohesion-reducing trend for domain 2 in the monomer, and thus that the apparent direction of evolution is different for the components of active M^{pro} versus the active dimeric state itself; one plausible explanation is that the monomeric loosening within domain 2 arises as a side effect of overall selection for a less cohesive protein, but that interactions in the dimer interface do not preserve this property for that region in the dimeric state. Either way, we find no evidence that M^{pro} is being selected for a looser domain 2 structure in the dimer.

Local structural changes due to mutations show similar effects for free and dimerized monomers, despite cohesion differences. To assess local changes in backbone structure due to residue substitution, we compute the (angular) mean and variance for each backbone torsion angle in each trajectory for both free monomers and dimers. Using this, we compare the variance in angles within trajectory versus across trajectories, allowing us to determine the extent to which local structure differs across variants above and beyond natural variations due to protein dynamics. Figure 2.5 shows the log-ratio of the between-variant versus within-variant angular variance, plotted by residue. High log-ratio values (blue areas) show substantial sensitivity to mutations, while low log-ratio values (red areas) show little

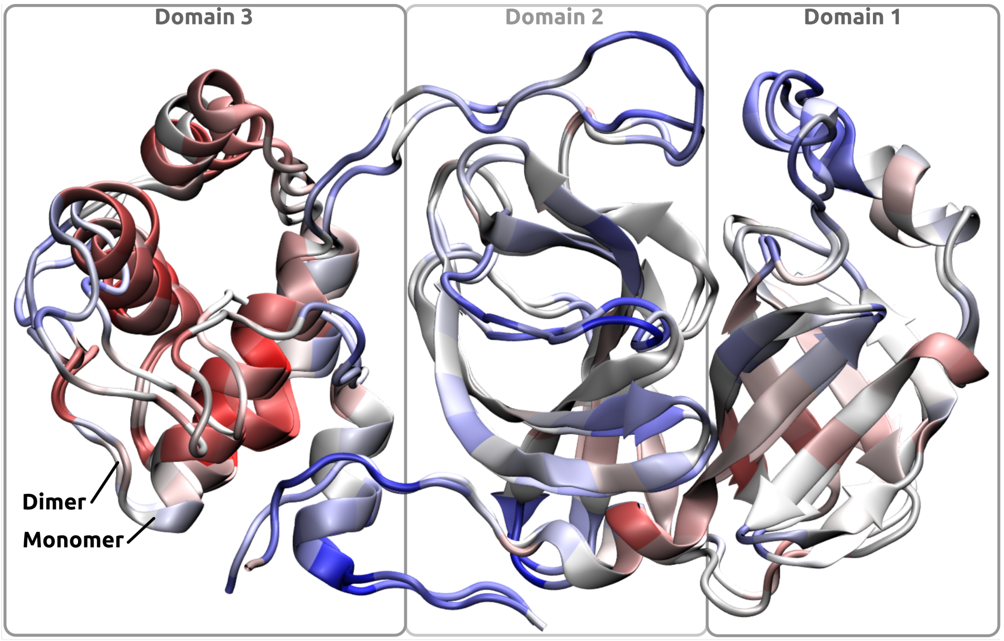


Figure 2.5: Comparison of monomer and dimer structures, with coloring corresponding to the log-ratio of between-chain variance and within-chain variance. Blue color shows higher between-chain variance, red shows higher within-chain variance. Free monomer and dimeric monomer structures are overlaid; both show very similar patterns of change in backbone torsion angles.

structural change relative to normal fluctuations due to protein dynamics. We see here that the bulk of the mutation effects are in or adjacent to domain 2, with domain 3 showing particularly low levels of sensitivity to observed substitutions. Taking these results in the context of the above findings regarding cohesion, we conclude that the cohesion changes seen in domains 1 and 3 are not due primarily to local deformation of the backbone in these regions of the protein, but more plausibly to a combination of side chain interactions and interactions with domain 2 residues (which do show greater change in torsion angle). Local deformation in domain 2 may thus be less important for the impact it has on domain 2 itself (which, as seen above, is inconsistent), versus its effect on the network of contacts in the neighboring domains (which both show consistent patterns of change).

Figure 2.5 also reveals that the pattern of backbone structure change in the free monomer is extremely similar to what is observed in the dimerized monomer, indicating that local structural changes are not strongly affected by dimerization. The immediate impact of mutation on local (backbone) structure thus depends only on interactions that are internal to the M^{pro} monomer itself, and are not related to interactions across the dimerization interface.

2.4.2 Mutations Increase Active Site Flexibility in the Active Dimer State

Mutations increase active site flexibility in the dimer, but not the free monomer. If mutations were selected to increase function of free monomers, the local structure around the active site of the monomer would be expected to show systematic change. This is not the case. As shown in Figure 2.6, constraint on the active site of the monomer does not trend in any direction; the presence of a large number of variants with either significantly higher (23.7%) or lower (11.2%) constraint levels suggests drift rather than conservation (*p*-values $<2.2\text{x}10^{-16}$ using an exact binomial test). By contrast, we see evidence of systematic selection for lower levels of active site constraint (looser structure) in the dimeric state. Not only are the grand means across variants lower for dimer active sites, but the majority (59.7% in higher scoring chain, 69.1% in lower scoring chain, *p*-values $<7.5\text{x}10^{-12}$, $<2.2\text{x}10^{-16}$, respectively using an exact binomial test) of variants have mean constraint scores that are significantly below WT. The presence of large differences in the dimer vs. monomer sites indicates that active site loosening is not driven by local structural changes to the monomer itself, but instead emerges from interaction between monomers in the dimer.

The decrease in constraint of the dimer active sites supports the hypothesis that the enzyme is increasing flexibility to adapt to the cellular environment of the human host. The difference between the changes in the properties of the dimer versus free monomer sites further sheds light on the dramatically higher activity of M^{pro} in the dimeric state: although earlier

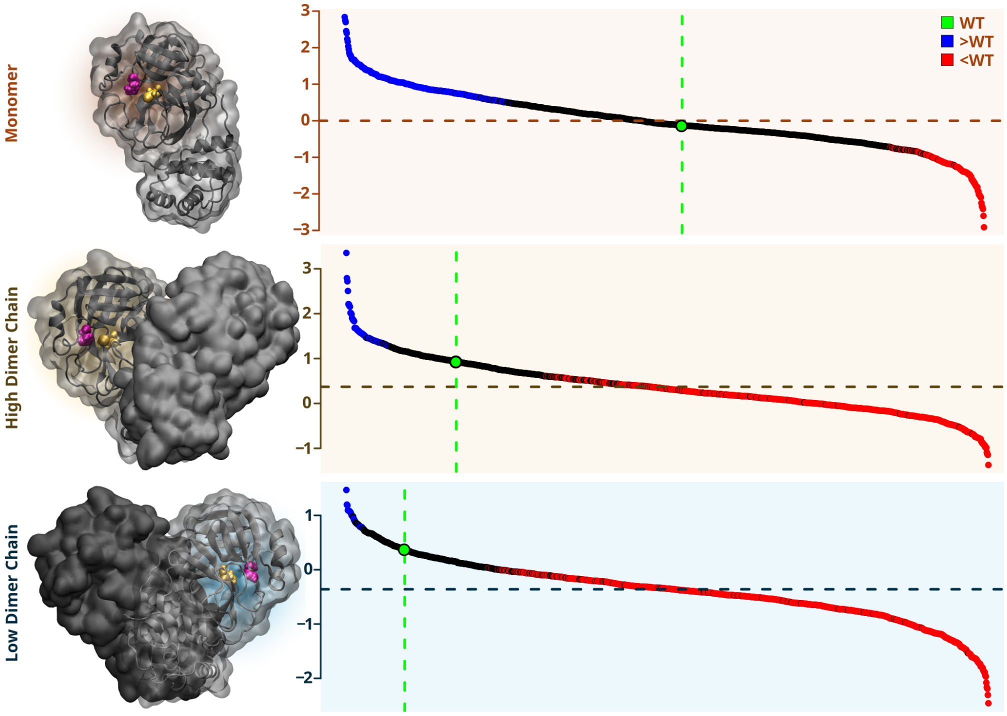


Figure 2.6: Mean ASN constraint scores by variant trajectory, for free monomeric and dimeric states; to break symmetry, dimeric active sites labeled based on mean constraint for analysis (middle high, bottom low). WT values indicated in green, grand mean indicted by horizontal line. Blue values are significantly more constrained than WT, red values are significantly less, black values not significant. Dimer active sites show reduced constraint for most variants, with no trend for the free monomer.

work [53] has shown that monomeric active site conformations do not differ markedly from dimeric ones, dimerization clearly shifts the equilibrium distribution of conformational states. Selection in this case appears to be operating on this shift, rather than on the underlying distribution, resulting in a pattern of changes that is selective for dimers while apparently neutral for free monomers.

2.4.3 Amino Acid Substitutions Favor Increased Size and Hydrophobicity

In general, substitutions increase hydrophobicity. Out of the 306 residues of the mature M^{pro} sequence, 269 have been substituted in at least one variant. To analyze the trends in properties of the substituted amino acids a substitution network was created by forming an adjacency matrix of substitutions. The rows and columns of the matrix were labeled with the 20 unique amino acids, and values in the matrix represented the frequency of each substitution occurring in the set of 1253 variants. This resulted in the network shown in Figure 2.7. Nodes represent unique amino acids, and edges represent the frequency of respective substitution. Substituted amino acids tend to be more hydrophobic and massive than their predecessors.

The large number of substitutions between certain residues, such as $L \rightarrow F$, $K \rightarrow R$, $G \rightarrow S$, and $A \rightarrow V$ indicate that these substitutions are highly favorable. These four substitutions are all examples of an exchange for a bulkier residue, and in the case of $G \rightarrow S$, a *more hydrophilic* residue. In the cases of $L \rightarrow F$ and $K \rightarrow R$, the substituted residues are able to form more complex intermolecular interactions, with a wider range of pi-stacking and cation-pi interactions available compared to the starting residues.

Frequent substitutions near the active site are either similar in hydrophobicity or more hydrophilic, while those in domain 2 are more hydrophobic. Figure 2.8

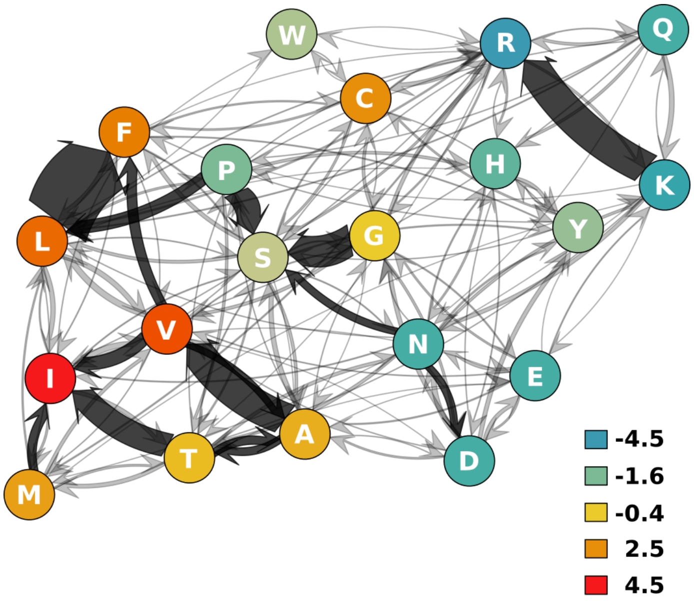


Figure 2.7: Substitution network showing the trends in residue substitutions. Nodes represent unique amino acids, with directed edges in the direction of the substitution. Edges are weighted by the number of substitutions observed, with darkened edges for substitutions which occurred more than 20 times. Nodes are colored by the corresponding hydrophobicity of the amino acid.

shows the frequencies of variants containing a substitution at a particular residue. The three most common substitutions among variants, L89F, K90R, and G15S, all occur in domain 1, and are all substitutions for bulkier residues. The decrease in cohesion of domain 1 could be caused by an increase in solvent interactions due, in part, to these three substitutions.

The first rug in Fig. 2.8 is the mean change in hydrophobicity, and shows a greater occurrence of hydrophobic substitutions in domain 2 than in either other domain. This also coincides with residues being more buried, as shown in the second rug by the darker blue coloring. An increase in the hydrophobicity of buried residues in domain 2 could be a response to a decreasing hydrophobic effect required to maintain the cohesion needed for certain dynamics resulting from internal interactions occurring between the dimer interface and the active site. While there are some structural changes upon dimerization in domain 2 due to substitutions, as seen in Fig. 2.5, those substitutions tend to be for more hydrophobic residues that are participating in the dimer interface. Increasing hydrophobicity at the dimer interface

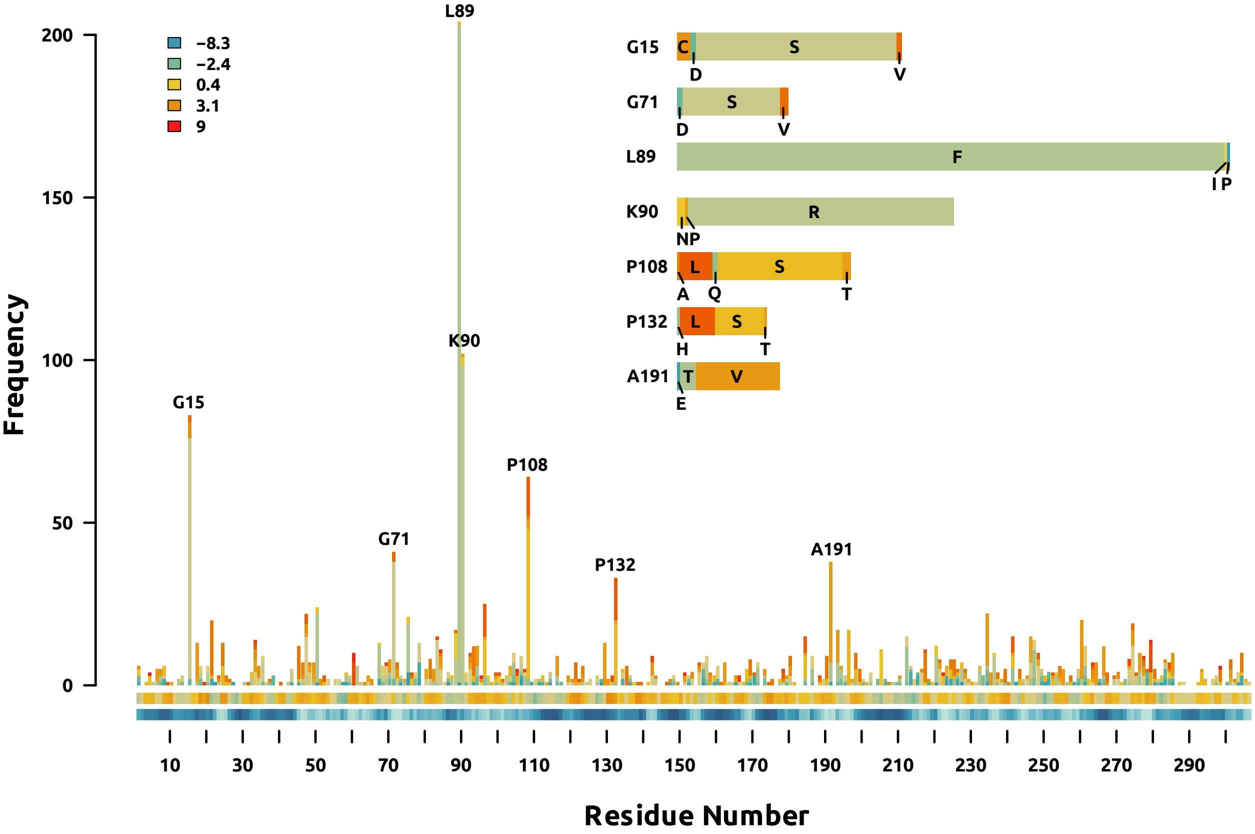


Figure 2.8: Frequency of substitutions along the main protease sequence. Colors indicate change in hydrophobicity resulting from the substitution, ranging from decreased hydrophobicity (blue) to increased hydrophobicity (red). A rectangular moving average of mean hydrophobicity change is shown below the bar plot using the same color scale. A rectangular moving average of the mean RSA of residues in the dimer conformation is shown at the bottom of the plot; darker values correspond to a more buried residue.

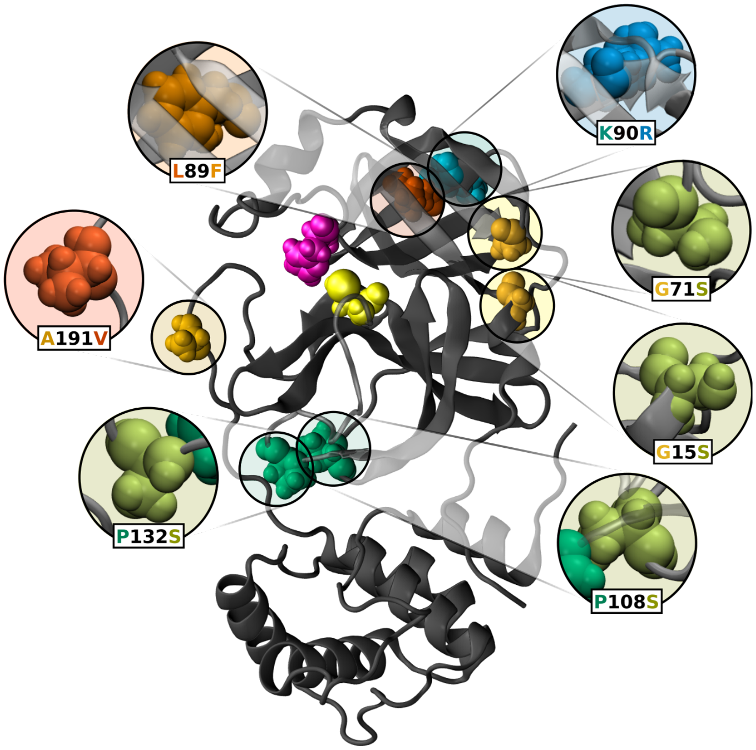


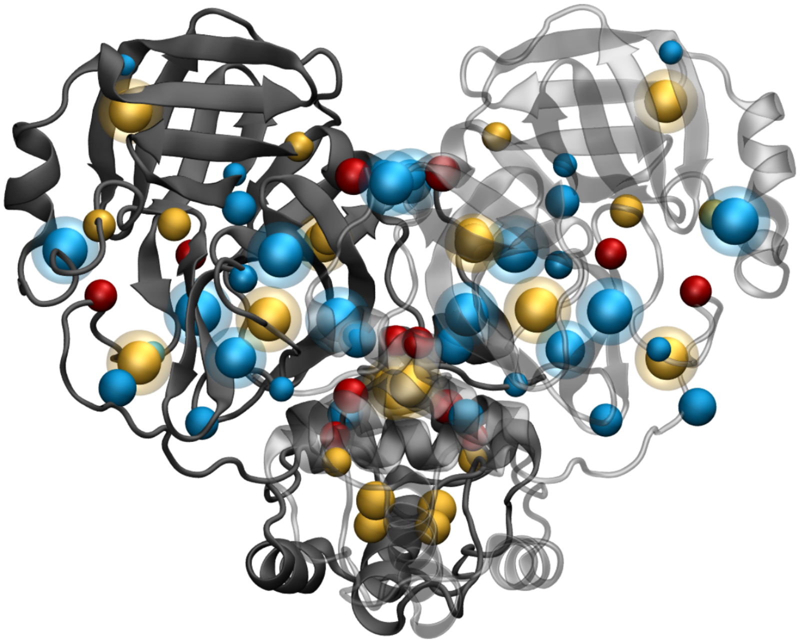
Figure 2.9: Locations of persistent substitutions in a single chain of the main protease structure are shown by the respective amino acid vdW representation colored according to hydrophobicity, as well as the catalytic C145 and H41.

would result in increased contact between the two chains due to the hydrophobic effect. Additionally, the location of more hydrophilic substitutions in regions where the chain is transitioning from the interior to the surface would cause a decrease in cohesion as those residues have stronger solvent interactions. Such regions are found in all three domains.

Persistent substitutions - those that occur most frequently - are shown in their location on one chain of the dimer in Figure 2.9. The most frequent substitution, L89F, is located between the folded β-sheets of domain 1. The substitution with a bulkier residue, phenylalanine, would push the β-sheets apart, reducing the cohesion of domain 1 and pulling the catalytic His41 back towards the β-fold. This change in structure of domain 1 would affect interactions between residues 43-50 and residues 186-190 on the unstructured loop between domains 2 and 3. There may also be some effect on the domain 1 residues near the N-terminus.

The substitution K90R would be expected behave similarly to L89F. However, this residue is facing out from the surface of the protein, and the substitution to arginine from lysine increases the number of potential hydrogen bonds that can be formed with the solvent in addition to increasing the bulk of the side chain. This may cause the domain 1 β-fold to be pulled open from the outside, instead of pushed from the inside. Variants with the K90R substitutions may see less effect on the interactions between residues 43-50 and 186-190, and more impact on the cohesion of domain 2 due to interactions between residues 97-105 in the unstructured loop between domains 1 and 2. The substitutions of G15S and G71S occur much closer to the dimer interface. Glycine and serine are both highly flexible residues, so the substitution at these locations may not have an appreciable effect on local structure. However, the polar nature of serine may cause it to respond to dynamics of other residues. For instance, a serine at residue 15 or 71 may interact with the polar hydroxyl group on Y154 of the opposite dimer chain, which would cause some correlation between the dynamics of domain 1 of one chain and domain 2 of the opposite chain.

P108 and P132 together form the ends of a loop that extends through domain 2 to interact at the dimer interface, forming a large part of the dimer interface. The location of the P108S and P132S substitutions may optimize their effect due to their connections with the dimer interface and proximity to the active site. Persistent mutations that are more hydrophilic are located away from the dimer interface, or else function to maintain the location of the interface by becoming less susceptible to the hydrophobic effect. These mutations are all located in domain 1, yet have limited impact on the active site except when in the dimer conformation. The more hydrophobic of the persistent mutations are located in domain 2, and have an influence on interaction at the dimer interface, while also having limited impact on the active site. The increasing hydrophobicity due to substitutions in domain 2 contributes to the conserved cohesion of the domain, as well as the increased influence of the dimer interface on local structure.



	Acidic	Aromatic	Nonpolar	Polar
Domain I	E14	Y54 F66	C16 P39 C44 F66	G2 G11 N28 G29 Y54 G79
Domain II	D176 D187	Y118 Y126 F140 F150 Y154 H172 Y182 F185	L115 F140 F150 F185	Y118 Y126 N133 S144 Y154 H172 Y182 G183 Q192
Domain III	D289 E290 D295	F291	A211 L268 L286 F291	N203 Q299

Figure 2.10: Conserved residues visualized in VMD using beads in their location on the dimer structure. Residues with a “halo” have an aromatic side-chain (Tyr, Phe, Hse). Blue are polar (Tyr, Asn, Gln, Ser, Hse, Gly), yellow are nonpolar (Phe, Cys, Ala, Leu, Pro). Acidic residues (Asp, Glu) are colored red.

Conserved residues are concentrated in domain 2, and tend to be polar. Substitutions in domain 2 for more hydrophobic residues may help to maintain the cohesion of the structure, as well as the dynamics resulting from interactions between the dimer interface and active site. Conserved residues may facilitate those dynamics to such an extent that any substitution that disrupts those interactions would inhibit function of the protein. This hypothesis is supported by the pattern of conserved residues in the dimer structure, shown in Figure 2.10.

Conserved residues in domain 2, located between the dimer interface and the active site, tend to be aromatic polar and nonpolar residues. Nonpolar residues that are conserved in domains 1 and 3 are by contrast non-aromatic. Acidic residues that are conserved are concentrated at the dimer interface near the N-termini, as well as on domain 2 at the active site. Polar residues other than Gly are concentrated around the active site, and at the dimer interface near the C-termini.

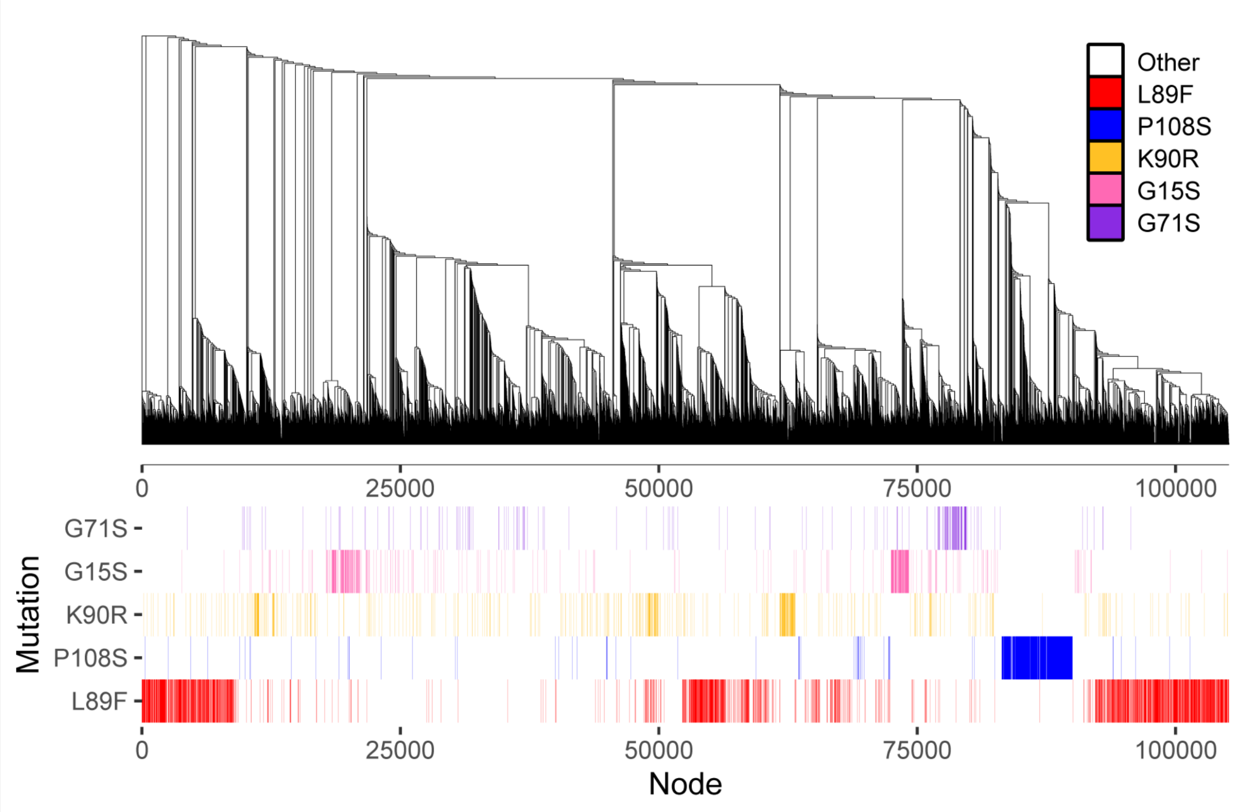


Figure 2.11: Phylogenetic tree (topology only) generated using all available full genomes from 1,253 M^{pro} variants as of February 25, 2021, including variants with multiple non-synonymous mutations, and one WT reference sequence [211]. The five most common mutations are indicated by colored lines: purple - G71S, pink - G15S, orange - K90R, blue - P108S, red - L89F.

2.4.4 Relationships between Variants

Clustering in phylogenetic tree shows independent occurrence of frequent mutations, supporting the selective pressure hypothesis. Frequent mutation is a form of adaptation in viruses [182], but while many rare variants exist in the population through luck, those that are observed in large numbers may be evidence of selective pressure [59]. Clustering patterns (Figure 2.11) have shown several large groups of recurring variants across disconnected lineages, supporting the hypothesis that this variation in sequence space may have also led to functional differences.

The most numerous mutations within the sample are, in increasing order: G71S, G15S, K90R, P108S, and L89F. These five were all present in a previous dataset from April, 2020 [53], though their prevalence in certain SARS-CoV-2 lineages were not necessarily as pronounced. Notably, G15S and K90R, which once dominated datasets over one year ago, have since been overtaken by L89F. Despite differences in raw counts, all five of these long-established mutations inhabit their own evolutionarily distinct clusters within the phylogenetic tree, often mimicking the large subtrees we saw in April, 2020 that were indicative of separate evolutionary events. Additionally, highly prolific mutations, including these five, have continued to remain viable in the population, co-occurring with secondary non-synonymous mutations that may impart their own structural or functional differences. For example, there are now 202 unique L89F variants (201 with at least one other amino acid mutation); in terms of mutational space, this means that nearly 1/6 of our unique variant dataset contains an L89F mutation. Although there is some overlap with other prominent mutations, much of that space is also taken up by G71S (36 variants), P108S (46 variants), G15S (74 variants), and K90R (97 variants).

Whole genome phylogenies are a useful tool in the study of viral evolution, but phylogenetic inferences should be made with the understanding that complex evolutionary dynamics are inherently difficult to capture. While neutral drift and selective mechanisms vie for control of genotypic diversity [126], factors like sequencing errors and sampling bias can disrupt attempts to accurately quantify their effects [71, 131]. The study of SARS-CoV-2 in particular is further complicated by large numbers of sequences with low sequence variation [131], making it difficult to draw meaningful conclusions from phylogenetic analyses alone. Because of regional variation in sequencing rates and pandemic policy, it is difficult to know if the rise of certain variants is truly due to fitness, as is often suspected. However, the trends observed here in total surface area, cohesion, torsion angle variance, and active site constraint speak to adaptations resulting from selective pressure, and reinforce evidence to that end observed in M^{pro}'s phylogeny.

2.5 Conclusion

Taken together, our analyses suggest that the SARS-CoV-2 main protease is evolving in response to selective pressure, possibly brought by the difference in cellular environments of bats and humans. The resulting adaptations are observed to affect the global structure and active site dynamics of the dimer conformation differently than the free monomer, despite having similar impacts on the local backbone structure of both states; in the case of active site constraint, the observed pattern of change vs. wild type is seen only in the dimeric state, and thus emerges from interactions between monomers. Adaptations tend to conserve interactions at the dimer interface and in domain 2, while allowing the rest of the protein, including the active site, to become more flexible in the dimeric state. The locations and properties of frequently occurring substitutions, as well as that of conserved residues, help elucidate the relationship between structure, dynamics, and function of M^{pro} as it is revealed by the process of selective adaptation.

As with any computational study, a major function of this work is to suggest targets for experimental investigation. Our findings suggest both general trends to be tested, and variants predicted to have extremal properties (relative to the ensemble); both tests of these hypothesized trends and examination of the relationship between the structural characteristics considered here and catalytic function would both shed light on M^{pro} evolution and help guide future computational studies. We also note that a number of other nsps (including the papain-like protease, PL^{pro} [135]) are also highly conserved within the beta-coronaviruses [40], suggesting mutation rates low enough to make computational studies like this one possible for such systems. Comparative analysis of changes seen across SARS-CoV-2 nsps in response to human host adaptation could provide deeper insights into ways in which evolutionary processes influence the molecular machines that carry out viral replication.

Chapter 3

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

3.1 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.

3.2 Introduction

Protein aggregation is implicated in a wide range of diseases, including Alzheimer's, Parkinson's, type II diabetes, and cataract[161, 47]. 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[137], to the amorphous aggregates of crystallin that form cataracts[95].

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[148, 133]. Monte Carlo (MC) simulations in particular have been used for studies of aggregation[44, 45]. In regards to proteins, the convention is to simulate protein-protein interactions between rigid-body proteins with a single conformation[116, 119]. To introduce some conformational flexibility, Wong, et al,[209] studied the aggregation of γD-crystallin (γ-Dc) using the multiconformation Monte Carlo (mcMC) algorithm[151, 121], 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[133]. A wide range of coarse-graining approaches have been proposed for studying protein structure, dynamics, and interaction[138].

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 [226]. Patchy particles have been used for simulating self-assembly [226, 207], as well as protein phase behavior such as in the case of γ-Dc[155, 102, 5, 115].

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 sub-units. For instance, Benson and Daggett[19] 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 [53] or protein classes [194]. Further coarsening can be employed to represent entire residues with a single node, which has been used for e.g. identification of active sites [6], studying transient structure in IDPs [78], and analysis of protein dynamics [176]. 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. [217, 78, 79, 218]). 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 [79, 222]. 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. [92, 120, 168]).

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[209]. 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.

3.2.1 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[170]. 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[24]. 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[57].

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[63]. 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[170].

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[172, 209]. Otherwise, similar structures are found in both crystals and solution for both the WT and W42R variant[96]. 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.

3.2.2 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 [79]. An aggregation graph can be seen as a form of *topological coarse-graining* [61], 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 [78, 222, 221], and related topological representations have also been used to study structure and dynamics in both folded [53, 34, 27, 163] and intrinsically disordered [78, 61] 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. 3.1 shows an aggregation graph derived from atomistic simulations of γ-Dc from Wong, et al[209], 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[78], 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

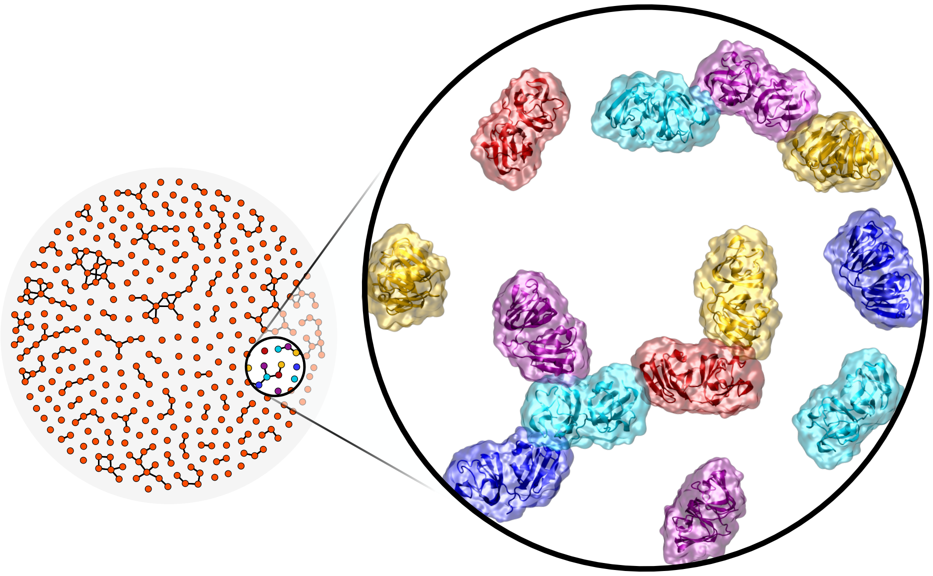


Figure 3.1: Example of an aggregation graph of the type studied here. Individual γ-Dc monomers are considered adjacent when they have respective domains whose centers of mass are within 31Å of each other in the atomistic model(see Methods). 2D graph representation shows underlying topology of the aggregate, without regard to spatial positions of the monomers.

model the probability of observing some specific graph microstate g as

$$\Pr(G = g|\phi, T) = \exp\left[-\mathcal{H}(g)/(k_B T)\right] h(g)/Z(\phi, T) \tag{3.1}$$

$$= \exp\left[-\left(\phi^T t(g) + k_B T t_e(g)\right)/(k_B T) \atop -t_e(g)\log N - \log Z(\phi, T)\right], \tag{3.2}$$

where $\mathcal{H}$ 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. [78], 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. 3.1 have been shown to be able to reproduce the structure of amyloid fibrils,[79, 222] 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. 3.1 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., [168, 120]). The ERGM parameterization of the model of Eq. 3.1 is given by

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

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 [168], allowing us to infer θ and t (and hence $\mathcal{H}$) from the realized aggregation graphs. Specifically, we obtain ϕ from θ under the family of Eq. 3.2 via

$$\begin{aligned} -\mathcal{H}(g)/(k_BT) + \log h(g) &= \theta^T t(g) \\ -\phi^T t(g)/(k_BT) - t_e(g) - t_e \log N &= \theta^T t(g) \\ \Rightarrow \quad \phi_e = -k_BT(\theta_e + 1 + \log N), \quad \phi_{s\neq e} &= -k_BT\theta_{s\neq e}. \end{aligned} \tag{3.3}$$

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[218], from which we can then infer ϕ using the relations of Eq. 3.3. 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 $\mathcal{H}$ 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 [21]) [70, 144], corrections for diminishing marginal effects [183], and consequences for equilibrium behavior [82, 31, 167, 33]. In the context of aggregation graphs, work on amyloid fibrils [79] 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) [90], 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

$$t_C(g) = \sum_{i=1}^{N} S(g)_i i^{\gamma}, \tag{3.4}$$

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 [79], 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 [187, 81]; we thus employ the geometrically weighted edgewise shared partner (GWESP) statistic for higher-order triadic closure effects [183, 88], which constrains contributions of high-order ESPs to have geometrically declining marginal effects. The structures represented by these terms are represented schematically in Fig. 3.2.

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.

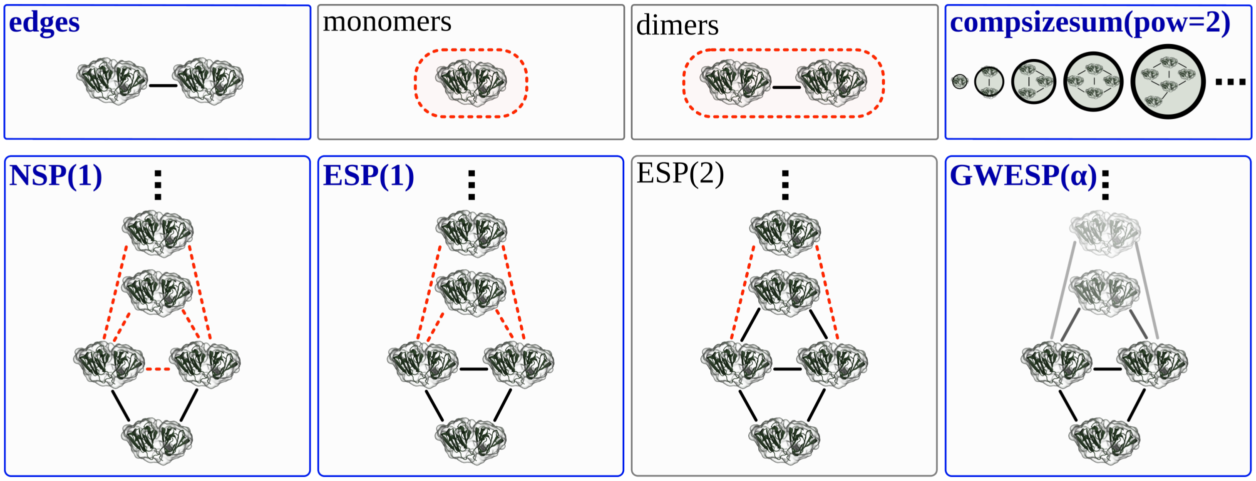


Figure 3.2: Schematic representation of candidate model terms for the γ-Dc network Hamiltonian. Black lines indicate edges that must be present in the specified configuration, while red dotted lines indicate edges that must not be present. Blue outline indicates terms selected in the final γ-Dc model. See text for details.

3.3 Methods

3.3.1 Atomistic Simulation and Network Generation

Wong et al. [209] performed atomistic simulation of equilibrium distributions of WT and W42R γ-Dc using multi-conformation Monte Carlo (mcMC) methods [151]; 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 [23] in TIP3P water [98]. 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[209].

Wong et al. [209] 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. [209], 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 [83] for the R statistical computing system [157], with the `network` [29] and `sna` [30] libraries used to compute descriptives and graphical layouts.

3.3.2 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, $\mathbf{z} = (z_1, \ldots, Z_m)$, the corresponding posterior distribution is $p(\psi|\mathbf{Z} = \mathbf{z}) = \text{Dirichlet}(\text{S} + 0.5)$, 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 [73].) 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 [84] as our measure of discrepancy between distributions. I.e., given fixed distributions f, g over component sizes $\mathcal{Z}_N$, our discrepancy measure is

$$D(f,g) = \| \log f/g \| = \sum_{i=1}^{N} \left(\log f(i) - \log g(i) \right)^2 , \tag{3.5}$$

where the informal notation f/g denotes the relative distribution over $\mathcal{Z}_N$. 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 $\mathbf{E}D(f_{obs}, f_{sim})|\mathbf{z}_{obs}, \mathbf{z}_{sim}$, where $f_{obs} \sim \text{Dirichlet}(S^{obs} + 0.5)$ and $f_{sim} \sim \text{Dirichlet}(S^{sim} + 0.5)$ (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 [99], 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.

3.3.3 Model Selection and Parameter Estimation

Models were fit by maximum likelihood estimation (MLE), using the pooling method of Yin and Butts [218]; estimation was performed using the `ergm` package [91], 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 [89].

		Model Terms			Error ($\Vert \log\left(f_{obs}/f_{sim}\right) \Vert$)			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 3.1: 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.

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 (N^2 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 $\mathbf{E}D(f_{obs}, f_{sim})|\mathbf{z}_{obs}, \mathbf{z}_{sim}$ 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 3.1 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 3.2.

3.3.4 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.3.1, 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. 3.2, to account for changes in the effective collision rate. Following Eq. 14 of Butts[32], 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 $(\frac{C'}{C})^{t_e(g)}$; this leads to the distribution

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

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).

3.3.5 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` [30] 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[100] 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

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

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`[64].) 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[76], based on PDB structures 1HK0[17] and 4GR7[96]. 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 16.63Å 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.

3.4 Results

3.4.1 Topology of γ-Dc Aggregates

γ-Dc WT, W42R aggregates have skewed size distributions, with truncated upper tails. Fig. 3.3 (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. 3.3 (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.

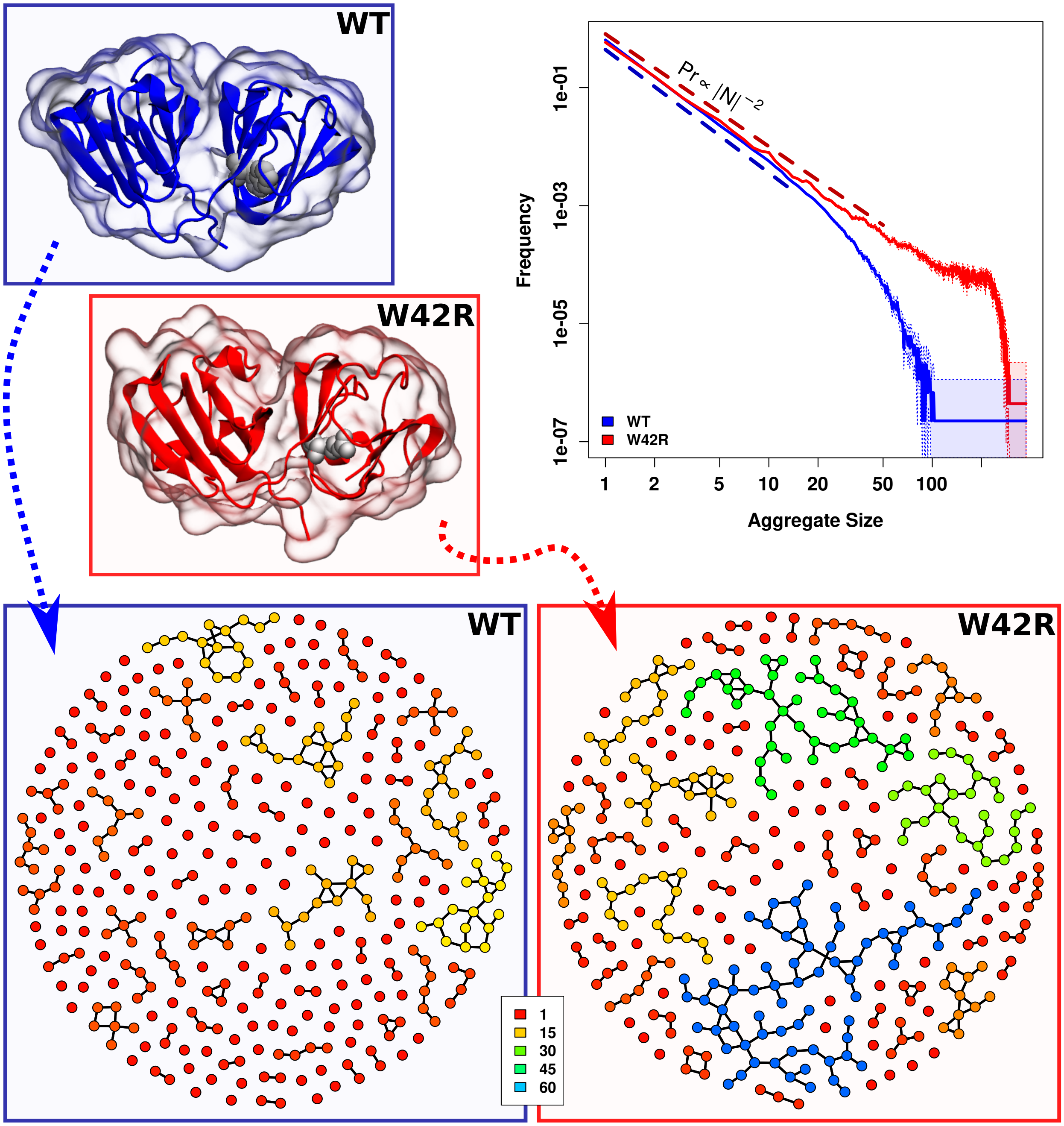


Figure 3.3: Aggregate sizes and topologies, from atomistic simulations by Wong et al. (2019). Top left: structures of WT (PDB 1HK0 [17]) and W42R (PDB 4GR7 [198]) monomers, with residue W42 highlighted. Trp to Arg substitution disrupts the N-terminal domain, increasing exposed hydrophobic surface area. 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).

3.4.2 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 3.1) 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 3.2)); 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.[129] 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

	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

Table 3.2: 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 ¡ 1×10^{-4}.

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. 3.4 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.

3.4.3 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.

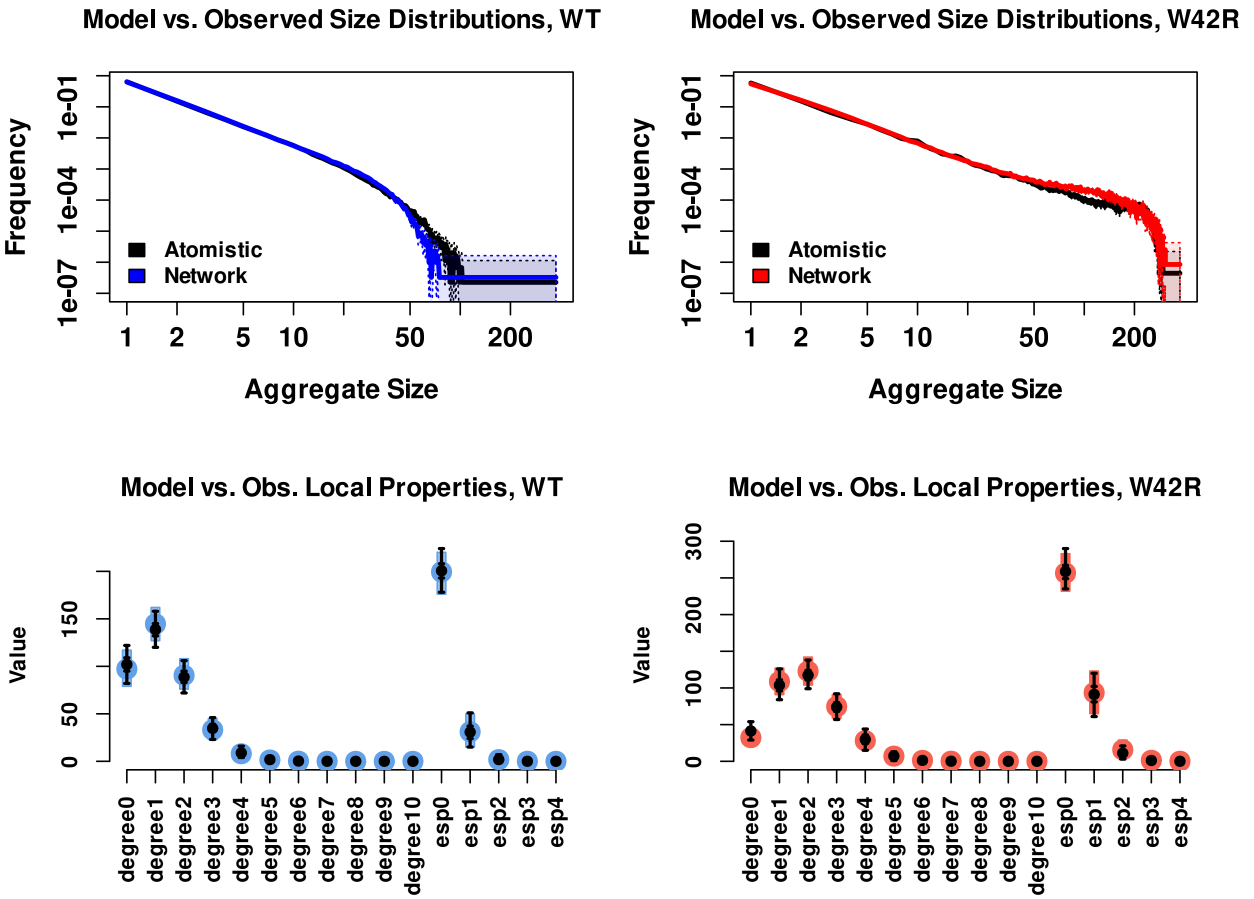


Figure 3.4: Model adequacy checks for the network Hamiltonian models. Top panels compare observed (black) to simulated (colored) aggregate size distributions (center line indicates posterior mean, shaded area 95% posterior intervals). Bottom panels compare observed (black) versus simulated (colored) 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.

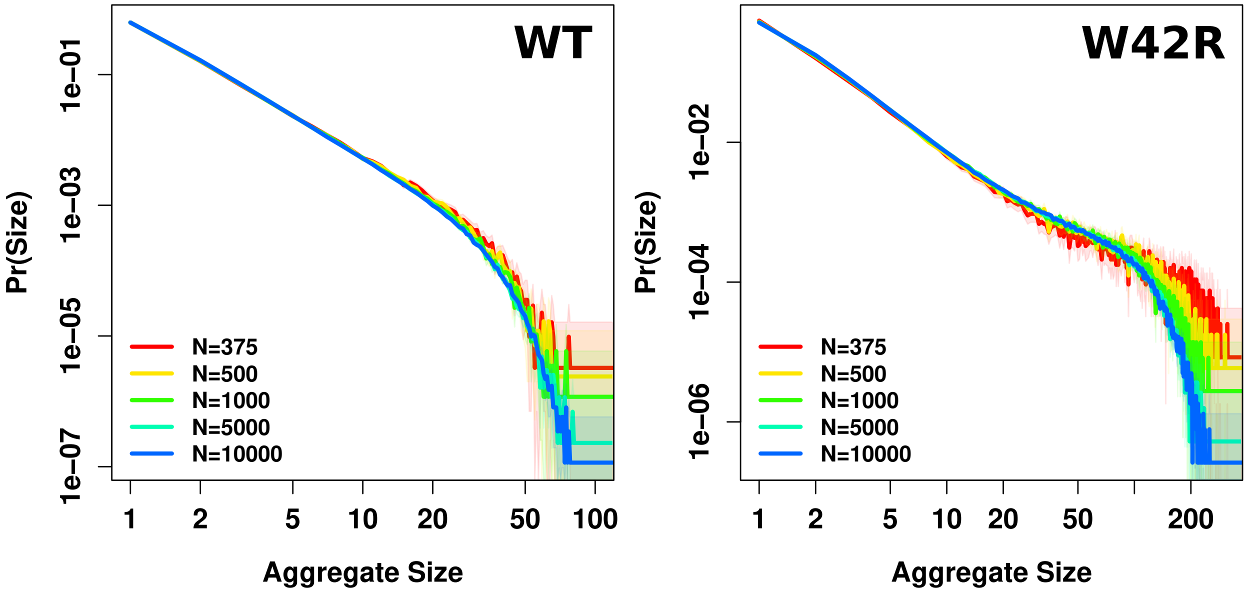


Figure 3.5: Predicted aggregate size distributions, by system size and variant. Center lines indicate posterior means; shaded areas indicate 95% posterior intervals. While distributions remain similar, maximum aggregate sizes decline more sharply when system sizes become large compared to the size of the largest aggregates.

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 3.5 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 unfavor-

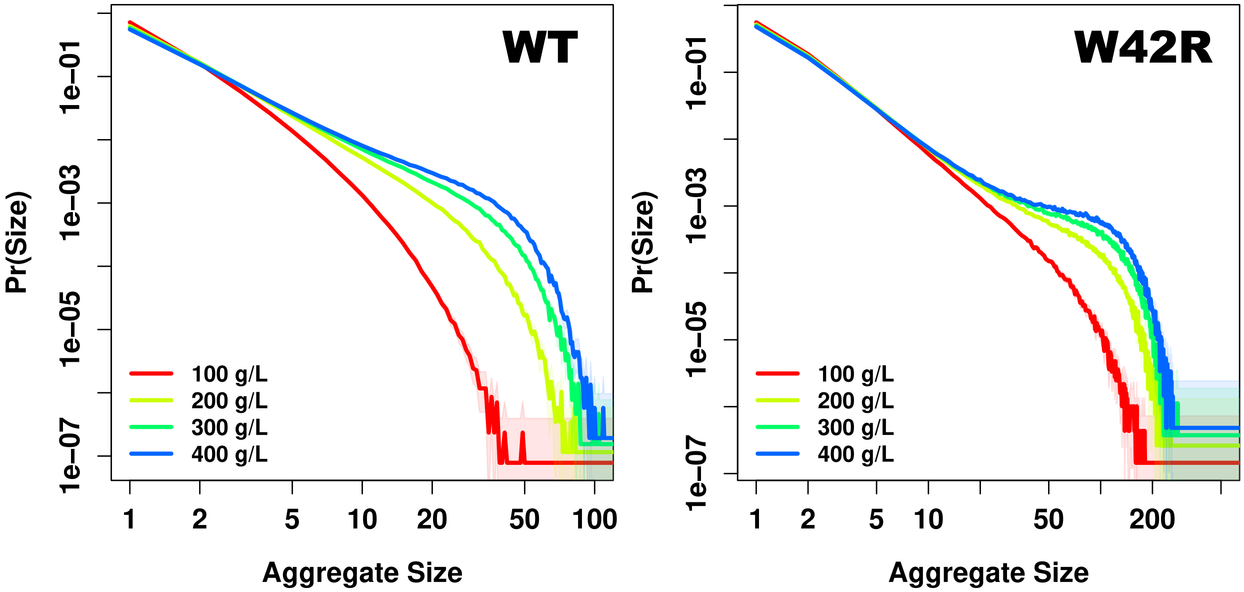


Figure 3.6: Predicted aggregate size distributions, by concentration and variant, at $N = 10^4$. Center lines indicate posterior means; shaded areas indicate 95% posterior intervals. Increased concentration favors growth of larger aggregates, particularly increasing the large-aggregate population in W42R.

able 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 3.6 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 3.7 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 $R_g/r_g \propto N^{0.333\pm0.002}$ for WT (with N here being the aggregate size), and $R_g/r_g \propto N^{0.313\pm0.001}$ for W42R. The elongation metric shows a slight deviation from spherical behavior, with large aggregates (100 or more monomers) tend-

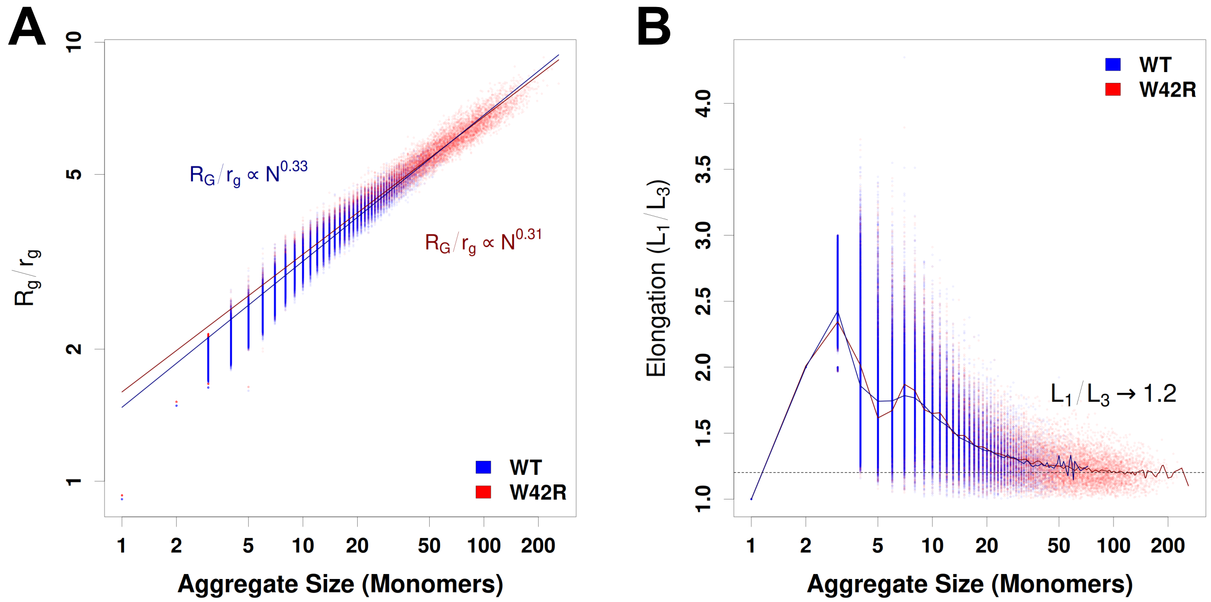


Figure 3.7: (A) Projected aggregate R_g over monomer radius of gyration (r_g) by aggregate size. For large aggregates (≥ 10 monomers), scaling is close to $N^{1/3}$, though slightly below for W42R. (B) Elongation factor (largest axis over shortest axis) by aggregate size; smoothing splines shown to indicate mean behavior. Larger aggregates approach a limiting elongation factor of approximately 1.2.

ing 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.

3.5 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[209], 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[171] (or, likewise to support more subtle surface interactions that have also been argued to promote aggregation[25]) 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[61]) 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[171, 172]), 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.

Chapter 4

Production of Distinct Fibrillar, Oligomeric, and Other Aggregation States from Network Models of Multi-body Interaction

4.1 Abstract

Protein aggregation can produce a wide range of states, ranging from fibrillar structures and oligomers to unstructured and semi-structured gel phases. Recent work has shown that many of these states can be recapitulated by relatively simple, topological models specified in terms of multi-body interaction energies, providing a direct connection between aggregate intermolecular forces and aggregation products. Here, we examine a low-dimensional network Hamiltonian model (NHM) based on four types of multi-body interactions, previously shown to be sufficient to reproduce two common types of amyloid fibril structures. We

characterize the phase behavior of this NHM family, demonstrating the range of aggregation states possible with this set of interactions. As we show, fibrils arise from a balance between elongation-inducing and contact-inhibiting forces, existing in a regime bounded by gel-like and disaggregated phases; complex oligomers (including annular oligomers resembling those thought to be toxic species in Alzheimer's disease) also form distinct phases in this regime, controlled in part by closure-inducing forces. We show that phase structure is largely independent of system size, allowing generalization to macroscopic systems, and provide evidence of a rich structure of minor oligomeric phases that can arise from appropriate conditions.

4.2 Introduction

Protein aggregation is a fundamental biophysical process, implicated in both functional and disease processess. Among the most striking - and important - classes of protein aggregates are *fibrils*. Fibrils are highly structured peptide aggregates consisting of repeated patterns of interactions between protein monomers and oligomers. Fibrils are implicated in a wide variety of chronic diseases [185, 113, 208], while also having some functional roles [180]. Empirically, the repeating interactions that form fibril structures have been shown to be sensitive to a wide range of factors, including peptide sequence [20, 141, 42, 117, 152, 166, 173], oligomer shape [111, 80, 192, 16, 223], concentration [3, 228], electrostatic interactions [56, 111, 228, 174], the presence of lipids [108], changes in temperature or pressure [112, 162, 109, 191], and other environmental factors [66, 67, 229, 51, 28, 48, 50]. In many cases, the same protein or peptide is *polymorphic*, forming multiple distinct fibril structures with only slight changes to experimental parameters [229, 192, 174, 2], and even within the same sample [200, 228, 50, 193]. The multiple structures of polymorphic sequences have been shown to be related to variations in disease[193, 145, 118, 128, 185]. Understanding how the patterns of interactions between monomers differ could thus lead to a better understanding of the processes behind aggregation-related diseases.

The repetitive structure of fibrils is analogous to a 1-dimensional crystal structure (Fig. 4.1), with (for amyloid fibrils) a cross-β structure binding monomers along the fibril length [153, 164, 200, 52], and variations of a "steric zipper" forming a cohesive core [164, 65]. Many fibrils are formed by conserved dimeric building blocks [50], which can take a variety of conformations that then go on to influence the structures of oligomeric states to which they self-assemble [203, 192, 16]. As this implies, fibrils are only one of many possible protein aggregation states, others including complex oligomeric forms, gel-like phases, and sparse but largely unstructured aggregates. These phases are also of significant interest, with e.g. annular oligomers hypothesized to be toxic species in Alzheimer's and Parkinson's Diseases [38, 39, 41] and unstructured aggregates playing a central role in cataract [11, 14, 49]. In addition, combined phases of fibrillar and non-fibrillar aggregates have been observed in experimental settings [51] and may hold insight into the transition from disordered to ordered structure [72, 68]. Prediction of mixtures of aggregation states based on the variety of sensitive intermolecular interactions is thus a formidable challenge.

While intermolecular interactions are intrinsically governed by atomistic effects, these effects ultimately combine to produce the global structure of the aggregate. Taking a topological perspective, a minimal set of descriptors can be used to describe the basis for local interactions between monomers. Besides the interaction that exists between every pair of monomers, the local interaction that is observed in all aggregates of three or more monomers is the sequential chain of interacting monomers. Regardless of other environmental or structural parameters, one may take any aggregate and draw some line between interacting monomers that follows a series of monomers, with the minimum count of three monomers describing a locally elongated chain. In cases where the line does not include every monomer in the aggregate (without doubling back), monomers not included in the sequence may be considered branches. In cases where the monomer at the end of a series or branch of monomers is also interacting with the initial monomer of a series or branch, a loop or cycle is formed. These four local topologies, pair-wise interactions, series, branches, and cycles, can thus be described by topological network terms, such as those used in a Network Hamiltonian Model (NHM).

Network models of aggregation have recently emerged as a scalable solution for modeling protein aggregation [79, 58], characterizing the aggregation system in terms of its topology (i.e., bound interactions among protein monomers) and describing its behavior in terms of multi-body forces enhancing or inhibiting bound interactions found in local structure. NHMs, in particular, describe aggregating elements in terms of an energy function based on the topology of the system state (along with corrections for motional degrees of freedom), providing a statistical mechanical picture based interactions between peptides in a system that is not *a priori* assumed to form fibrils. This strength of NHMs, the ability to model larger systems of peptides that may form multiple aggregate phases within a single system, allows these models to produce results similar to those found by other coarse grain (CG) phenomenological models of monomer aggregation *as well as* provide estimates for the behavior of a larger system of *fibrils*, such as in systems explored by larger CG models[142, 219, 85, 51]. NHMs have been shown to be able to recapitulate the so-far observed classes of amyloid fibril topologies [79] as well as properties of cataract-associated aggregates [58]; kinetic extensions of NHMs have been shown to recapitulate experimentally observed stages in fibril formation [222], and finer-grained, residue-level NHMs have also been used to examine the behavior of intrinsically disordered proteins [78] and conformational variation across crystal structures of globular proteins [77]. Adapted from models originally developed to study social networks [92, 104], NHMs are also able to leverage a large body of computational and statistical work, greatly facilitating their use. Monte Carlo simulations from the equilibrium states of even fairly large NHMs (e.g., $10^2 - 10^4$ monomers) can often be easily performed on standard computing hardware, making them a very accessible tool for studying protein aggregation.

Predictions made by NHMs provide insight to the redistribution of intermolecular interactions that may result from differences in environmental conditions or peptide sequence. This can be used to inform a more targeted approach when using more fine-grained methods, such as molecular dynamics (MD) simulations or DFT calculations, to study specific inter- or intramolecular interactions that are suspected to play an integral role the formation of a given aggregate. MD simulations – arguably the theoretical workhorse of molecular biophysics – provide a method of exploring the atomistic interactions behind protein aggregation, especially in the case of proteins and peptides whose structure is difficult to elucidate due to low solubility or other factors [110]. MD is useful for exploring atomistic details, but can be difficult to scale to the large system sizes (hundreds or thousands of protein monomers) and time scales (hours to years) needed to capture the formation of fibrils and other complex aggregates [165, 227, 160, 79, 46]. Other coarse-grained (CG) models exist as alternatives to MD that simplify the aggregation system, trading atomistic detail for scalability (and, to the extent that the coarsened units are non-specific, generality). Molecular CG models may also appear to gloss over a large amount of useful information, but, as noted e.g. by [132], "In order to simulate the aggregation process itself, from monomers to large aggregates, one must be willing to sacrifice atomistic details and invoke coarse-grained models." There are many classes of CG models (see [132] for a more detailed review) at varying levels of coarseness. The most conservative are "systematic" CG models that map all-atom simulations to a CG representation [94, 159, 130, 43, 37], as well as relative entropy methods[177]. "Phenomenological" methods, such as the off-lattice[210, 35] and on-lattice[136, 114, 1] models, aim to probe the interaction space and determine the energetic parameters that can mimic the aggregation of different amino acid compositions by representing residues with as few as one bead per residue. The coarsest CG models typically represent the entire peptide as a single object, such as a rod[195], a stick[93], a tube[12], or a cuboid model [224], and take advantage of the simplicity of the model to scale simulations toward larger numbers of peptides aggregating into fibrils, using averaged interaction potentials to mimic intermolecular interactions. These molecular models are, however, biased by the shape chosen to represent fibril monomers, which has an effect on the interactions between monomers. This is not the case with NHMs, which focus exclusively on the interactions themselves, allowing for greater generalizability and application.

Here, we employ a NHM to examine the universe of aggregation states that can be created from a Hamiltonian based on small number of simple multi-body interactions, characterizing the resulting phase structure in terms of the balance between different topological forces that favor or disfavor aggregation. Prior work has shown the minimal four-term model (describing pair-wise interactions, series, branches, and cycles) to be sufficient to reproduce the frequently observed 1-ribbon and 2-ribbon fibril topologies [79, 222]. In this work, we also show that this model is capable of producing a wide range of other aggregation states, including oligomeric and gel-like phases. We find that the most important phase boundaries can be rationalized in terms of competing forces governing branching, elongation, and cyclization, providing a high-level view of the conditions that favor formation of fibrils versus other aggregates in equilibrium. We also show the presence of regimes containing large numbers of distinct oligomeric phases, suggesting fairly sensitive dependence on the balance of intermolecular forces under certain conditions. Because our model is directly specified in terms of the most basic intermolecular interactions, our findings should be broadly applicable to any fibril-forming system.

4.3 Methods

4.3.1 Modeling of Aggregation States

Following [79], we define protein aggregation states via *aggregation graphs.* An aggregation graph, $G = (V, E)$, is an undirected graph on a set of N protein monomers, V, whose edges

$\{i, j\} \in E$ represent pairs of monomers that are bound to one another. Examples are shown in Figure 4.1. The equilibrium state of the aggregation graph is specified by the NHM probability distribution

$$\Pr(G = g|\mathcal{H}, T) = \exp[-\mathcal{H}(g)/(k_B T) - \log Z(\mathcal{H}, T)]h(g) \tag{4.1}$$

where $g \in \mathbb{G}$ is a realized state of g (out of the ensemble of possible states, $\mathbb{G}$), $\mathcal{H}$ is the network Hamiltonian, $Z = \sum_{g' \in \mathcal{G}} \exp[-\mathcal{H}(g')/(k_B T)]h(g')$ is the partition function, T is the temperature, k_B is Boltzmann's constant, and h is a reference measure accounting for entropic effects of unmodeled degrees of freedom. This framework exploits the fact that motional degrees of freedom are largely time-scale separated from the (much slower) process of bond formation and dissolution, making it feasible to account for them indirectly by their average effects. While a kinetic extension of this model has been proposed [222], we here focus on equilibrium states. $\mathcal{H}$ is parameterized in terms of a set of topological degrees of freedom, t, that are real-valued functions of the aggregation graph; i.e., we take $\mathcal{H}(G) = \phi^\intercal t(G)$, with $t : \mathbb{G} \to \mathbb{R}^p$ and energy parameters $\phi \in \mathbb{R}^p$ indicating the potential per unit change in the corresponding elements of t. Defining the dimensionless parameter $\theta = -\phi/(k_B T)$ leads to the alternative parameterization $\Pr(G = g) \propto \exp[\theta^\intercal t(g)]h(g)$, a form which is known as an exponential family random graph model (ERGM) [168]; a considerable body of work exists on specification and simulation for models in ERGM form, which we leverage below.

As noted, $\mathcal{H}$ is specified by the choice of degrees of freedom (or *statistics*), t. Here we use a four-parameter family shown to produce 1-ribbon and 2-ribbon fibril structures [222], containing sufficient statistics (in ERGM terminology) t_e (the edge count), t_{nsp_1} (the count of NSP(1) configurations), t_{nsp_2} (the count of NSP(2) configurations), and t_{2s} (the count of 2-star configurations). These are defined as follows. t_e is the count of edges in the graph (i.e., the number of bound interactions among monomers). A *null shared k-partner* (NSP(k)) is a conformation involving two non-adjacent vertices i, j having exactly k partners in common.

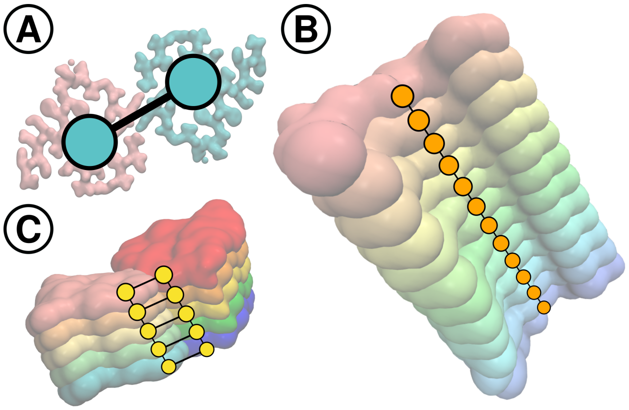


Figure 4.1: Examples of possible network representations of fibril structures found in the PDB. A) Dimer network structure of a single dimer block from PDB structure 7Q4M[216]. B) '1-ribbon' network structure super-imposed over PDB structure 2MXU[212]. C) '2-ribbon' super-imposed over the PDB structure 7Q4M.

t_{nsp_1} is thus the count of pairs of unbound monomers that are jointly bound to a third monomer, describing an elongated series of monomers, while t_{nsp_2} is the count of pairs of unbound monomers jointly bound to two other monomers, describing a closed-cycle effect. Finally, a 2-star is a configuration in which one node is bound to two others (inclusively - thus a node with four partners is the center of six 2-stars, see SI Fig. S10), and describes the number of branches on a given monomer. As suggested by [79, 222], ϕ_e can be thought of as the baseline energy for interaction between two otherwise unbound monomers, and ϕ_{2s} serves as a first-order approximation to the change in this energy associated with existing bound interactions (i.e., every existing tie to monomer i increments the energy of a new bound interaction with i by ϕ_{2s}). Jointly, these forces are sufficient to produce not only distinct fibril types, but many other aggregation states.

The graph terms corresponding to pair-wise interactions, elongation, branching, and cycle-formation represent net interaction energies resulting from combinations of intermolecular interactions, and thus vary depending on both peptide sequence and experimental conditions. However, these terms are agnostic of some spatial and energetic effects that are typically included for consistent model behavior when studying physical systems such as protein aggregates. In particular, maintaining consistent behavior at constant concentration requires accounting for the entropic contribution of spatial mixing (which is implicit in the topological model); following [79], we use the Krivitsky reference measure $h(G) = N^{-t_e(G)}$ for this purpose (see also [106, 32] for a more formal treatment). Likewise, maintaining consistent behavior across temperature requires accounting for energy within bond vibrations. As in prior work, we take such vibrational motion to be time-scale separated from formation or dissolution of edges, implying a total contribution to the graph Hamiltonian of $k_B T$ per edge (each edge having one potential and one kinetic degree of freedom, and oscillations being classical). The coefficient on the edge term, t_e, thus contains the force applied to each edge, ϕ_e, as well as the adjustment for temperature, $k_B T$. The remaining three terms also have coefficients representing the net forces (ϕ_{2s}, ϕ_{nsp_1}, ϕ_{nsp_2}) that influence the respective local topology (indicated by t_{2s}, t_{nsp_1}, and t_{nsp_2}, respectively). This results in a final graph Hamiltonian of the form

$$\mathcal{H}(g) = (\phi_e + k_B T)t_e(g) + \phi_{2s}t_{2s}(g) + \phi_{nsp_1}t_{nsp_1}(g) + \phi_{nsp_2}t_{nsp_2}(g). \tag{4.2}$$

The corresponding ERGM has parameter vector

$$\theta = (-\phi_e/(k_B T) - 1 - \log N, -\phi_{2s}/(k_B T), -\phi_{nsp_1}/(k_B T), -\phi_{nsp_2}/(k_B T)) \tag{4.3}$$

with statistics $t = (t_e, t_{2s}, t_{nsp_1}, t_{nsp_2})$ (where we have re-specified θ relative to the counting measure, as is common in practical use).

Given translation into ERGM form, equilibrium draws from the NHM can be obtained using standard Markov chain Monte Carlo (MCMC) techniques [22]; here, we use the default tie/random-dyad Metropolis-Hastings algorithm from the `ergm` package [91] within the `statnet` library [105] for simulation, with systems of size $N = 150$. (As noted below, we also perform targeted simulations with larger system sizes of $N = 750$ to verify stability of the model.)

4.3.2 Network Sampling

We define a parameter space over a range of values for parameters ϕ_{nsp_1}, ϕ_{nsp_2}, and ϕ_{2s}, with a constant value for ϕ_e. The range for each parameter value is determined based on physical grounds and on prior examination of regions that produce stable 1-ribbons and 2-ribbons [222, 221]. Parameter values follow the intuition for energy values that are measured in physical systems, where a more negative value indicates a lower energy and a more favorable local topology, and a more positive value indicates a disfavored topology. Thus, $\phi_e < 0$ is necessary for aggregation to be favorable, $\phi_{nsp_1} < 0$ encourages elongated structures, $\phi_{2s} \geq 0$ disfavors additional branching after an edge is added to a node, and $\phi_{nsp_2} < 0$ favors the formation of 4-cycles (cycles consisting of exactly four monomers). Conversely, positive parameter values for ϕ_{nsp_1} and ϕ_{nsp_2} result in less elongation (i.e. less aggregation beyond the formation of dimers) and less cyclization, respectively. We then employ a Halton sequence to sample 16,000 low-discrepancy points covering the parameter space, simulating a network for each combination of parameter values in the sample using the `statnet` packages in R[156, 91, 105]. Networks are simulated at three different ϕ_e values: -50, -66, and -81 kcal/mol (equivalent to θ_e values of 75, 100, and 125, respectively).

Following the initial, uniform survey of the parameter space, resulting phase boundaries within the space are refined by sampling additional networks in regions where network simulations are observed to produce fibrillar aggregates. For each initially sampled parameter value that is found to produce fibrils with the -66 kcal/mol ϕ_e value, 10 new points are sampled from the surrounding parameter space using a multivariate normal distribution (MVN) centered on the original point. This results in a total of 43,880 networks sampled from the parameter space, shown in Figure 4.2 as convex hulls encapsulating networks that produce given target structures.

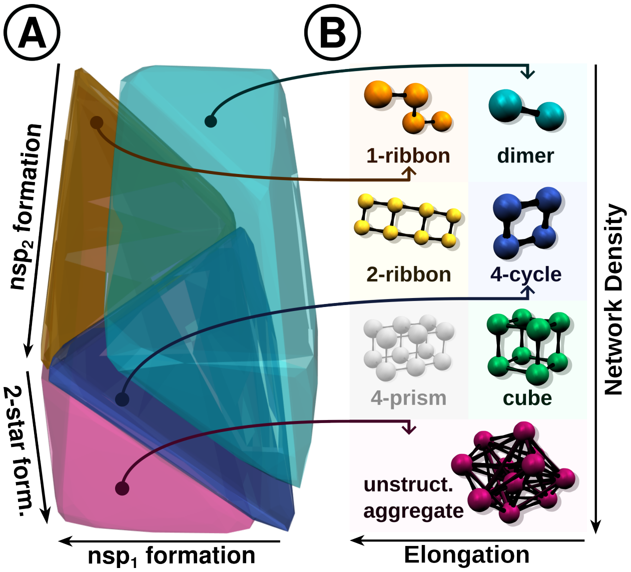


Figure 4.2: Networks were sampled from the space formed by ranges of parameter values on t_{2s}, t_{nsp_1}, and t_{nsp_2}. A) Convex hulls of networks with a majority of vertices forming 1-ribbons (orange), dimers (turquoise), 4-cycles (blue), and unstructured-aggregates (dark pink). The unfilled space between the dimer hull and the 1-ribbon and 4-cycle hulls contains networks with mixtures of dimers, 2-paths, and 4-cycles. B) Examples of the target structures, mapped according to naive descriptions of their network structures: elongation and tie density. All structures except for the 4-prism (grey) are observed in this study.

4.3.3 Yield Calculations

Each point in the parameter space is classified according to the average yield of target fibrils over five independent network simulations with $N = 150$ monomers. Vertices within each network are assigned a target fibril structure classification based on its local network structure. Graphlet orbits are used to identify the number of vertices that are locally in one of three target fibril structures: 1-ribbon, 2-ribbon, and 4-cycle oligomer [217]. Other target structures, such as "cubic" oligomers (cyclized 2-ribbons), dimers, and unstructured aggregates, are classified based on either the number of vertices that are included in a single structure, or using a heuristic classifier that uses component size and cycle information for a given structure within the network (see SI for details of classification). Examples of each structure, as well as the convex hulls of networks containing those structures in the parameter space, are shown in Fig. 4.2.

4.3.4 System Size Checks

To verify that the phase behavior is not heavily influenced by system size, we replicate simulations for a portion of our sampled space with a much larger number of protein monomers. Since our interest is in the inter-phase boundaries, we probe them by sampling points on a series of three parallel trajectories through the parameter space that intersect the previously determined boundaries (allowing us to determine whether the boundaries move as N increases). The trajectories are shown in Figure 4.6. Two sets of networks are simulated for each point on the three trajectories, one with N=150, and the second with N=750. This represents an increase of 5 times the system size of the original networks; our choice of reference measure implies that these systems are maintained at constant concentration (c.f. [58]).

We sample the trajectories using the equation $\phi_{2s} = (5/2)\phi_{nsp_1} + \phi_i$, where ϕ_i is the intercept and is equal to 287, 182, and 78 kcal/mol for the pure 1-ribbon phase, the 1-ribbon/2-ribbon interface, and the mixed 2-ribbon/oligomer phases, respectively. The value of ϕ_{nsp_2} is kept constant at -100 kcal/mol. Points are sampled equidistantly along this line, moving from the pure dimer phase towards the pure unstructured aggregate phase. We end sampling when at least five networks have been classified as unstructured aggregates. We then compute target structure yields for each of these networks for comparison of the boundary locations to determine any dependence on system size.

4.3.5 Phase Classification

We classify distinct phases of fibrillar and non-fibrillar aggregation states using points that produce networks with 100% yield of the corresponding aggregation product. The resulting phases are described as “pure” phases in following sections. Phases with more than one structure type are referred to as “mixed,” and are treated separately.

4.4 Results and Discussion

We begin by discussing the boundaries within the parameter space that separate phases of our target structures. Next, we determine dependence of the phase boundary location on the value of the ϕ_e parameter and discuss how phase boundaries impact our interpretation of intermolecular interactions. We then characterize the regions of the space containing mixed phases, and examine how mixtures respond to system size, N. Next, we describe how to calculate the change in energy that accompanies additional interactions, and discuss how changes in temperature would affect the system. Finally, we discuss the utility of NHMs and possible applications to empirical data.

4.4.1 Phase Boundaries

The points producing 100% yields of dimer, gel, or 1-ribbon structures produce the starkest boundaries. Examination of the phases plotted on the ϕ_{nsp_1}-ϕ_{2s} plane, shown in Figure 4.3A, suggests the parallel boundaries between the 1-ribbon phase and the unstructured-aggregate phase, the 1-ribbon phase and mixed oligomer phase, as well as the pure dimer phase and the mixed oligomer phase follow similar equations with unique intercept values, ϕ_0, corresponding to the specific boundary. Thus, the equation:

$$\phi_0 = \phi_{2s} + \phi_{nsp_1} \tag{4.4}$$

relates each boundary at its unique intercept to the difference in magnitude of the positive ϕ_{2s} and the negative ϕ_{nsp_1} values. Increasingly positive ϕ_{2s} inhibits additional interactions beyond the first interaction for any bound monomer. Increasingly negative ϕ_{nsp_1} favors elongation, which necessitates at least a second interaction for any bound monomer. The combination of these two values describes the net free energy required for bound monomers to accommodate additional interactions. Intuitively, when the magnitude of ϕ_{2s} is sufficiently greater than that of ϕ_{nsp_1} multiple interactions are ultimately disfavored, with the intercept (ϕ_0) determining the number of additional interactions any monomer can accommodate. The interpretation of exact intercept values is discussed below.

The ϕ_{nsp_1}-ϕ_{nsp_2} plane, shown in Figure 4.3B, shows the relationship between negative (favorable) values for both elongation and cyclization. Here, a boundary between the pure 1-ribbon phase and the 2-ribbon and 4-cycle mixed phases is observed with the equation:

$$\phi_0 = \phi_{nsp_2} - 2\phi_{nsp_1}. \tag{4.5}$$

The relationship between ϕ_{nsp_1} and ϕ_{nsp_2} is correlated, as t_{nsp_1} structures are required for cyclization (t_{nsp_2}) to occur. More precisely, the aggregation of four monomers into a cycle may be accomplished most simply by the combination of two dimers through two elongating interactions (ϕ_{nsp_1}). Correspondingly, the net free energy necessary to stabilize the 4-cycle structure is proportional to the energy of two elongating interactions by a proportionality constant, 2. When ϕ_{nsp_1} values are more negative elongation is favored over cyclization, and when ϕ_{nsp_2} is more negative (such that ϕ_{nsp_2} is sufficiently greater than $2\phi_{nsp_1}$) cyclization is favored over elongation. When ϕ_{nsp_2} is greater than $2\phi_{nsp_1}$ by a small amount, elongated structures containing 4-cycles are able to form. Further discussion of exact intercept values is found in following sections.

The plane given by (4.4) and (4.5), shown in Figure 4.3C, shows a boundary between the pure 1-ribbon phase and the majority 2-ribbon networks, and follows the equation:

$$(1/2)\phi_{nsp_2} - \phi_{nsp_1} = -(\phi_{2s} + \phi_{nsp_1}), \tag{4.6}$$

where the difference in the magnitudes of the intercepts from (4.4) and (4.5) determines

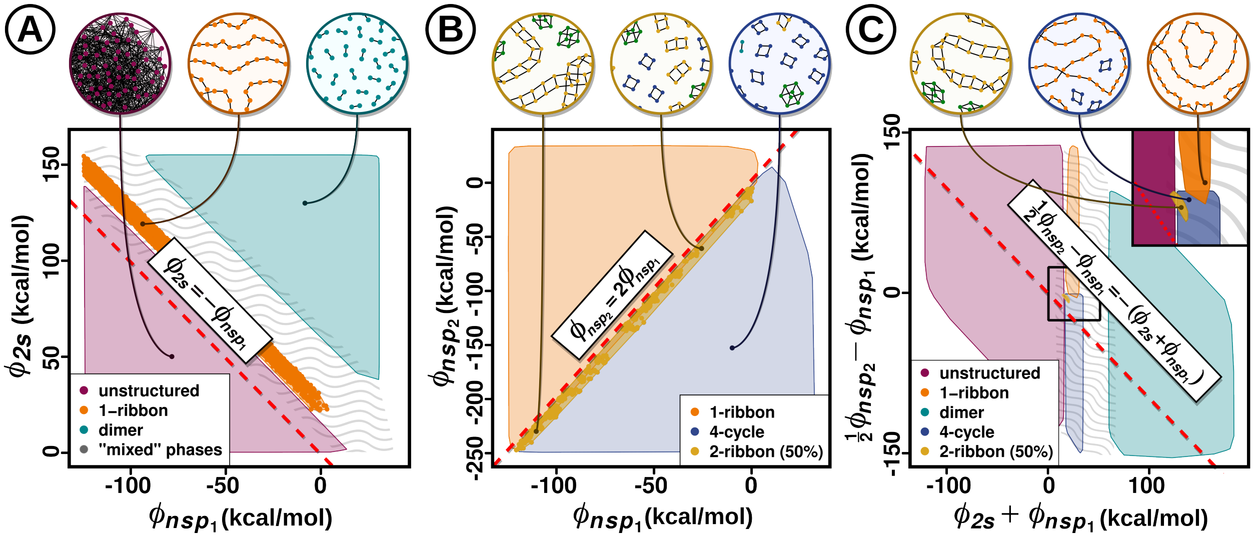


Figure 4.3: Plots show phase boundaries between regions of the parameter space producing networks with 100% of vertices in a type of target structure. Examples of networks from each phase are shown above. A) Convex hulls containing the pure dimer, pure 1-ribbon fibril, and pure unstructured-aggregate phases are plotted on the ϕ_{2s}-ϕ_{nsp_1} plane. Eq. 4.4 with ϕ_0 =0 is shown as a dashed red line. Phase boundaries are located at unique values of ϕ_0 that are shown to be multiples of ϕ_e^* (Fig. 4.4). Example networks (from left to right) are of the unstructured-aggregate phase, the 1-ribbon phase, and the dimer phase. B) The pure 1-ribbon fibril convex hull and the mixed oligomer convex hull are plotted on the ϕ_{nsp_1}-ϕ_{nsp_2} plane. Eq. 4.5 is shown as a dashed red line with ϕ_0 =0. Yellow points represent networks with a majority of vertices in 2-ribbon fibril structures. Examples show two mixed 2-ribbon networks and a mixed 4-cycle oligomer network containing 2-ribbon segments and cubic oligomers. C) Pure 1-ribbon, dimer, and unstructured-aggregate hulls, as well as the mixed oligomer hull, are plotted on the plane given by Eqs. 4.4 and 4.5. Eq. 4.6 is shown as a dashed red line with ϕ_0 =0. The window at top-right shows a close-up of the interface between phases of pure 1-ribbon, mixed oligomer, pure unstructured-aggregate, and networks with majority 2-ribbon fibrils (yellow hull). Example networks are from the majority 2-ribbon phase, the mixed oligomer phase in the region between the majority 2-ribbon phase and pure 1-ribbon phase, and the pure 1-ribbon phase. Examples were chosen to show the consistent elongating effect, and the increase in closure effects resulting in 4-cycles, 2-ribbons, and cubic oligomers at smaller ϕ_0 values.

the phase. This amounts to a comparison of the elongating effect that determines the number of interactions for any monomer and the cyclization effect that determines whether additional interactions to an elongated aggregate are between monomers of the same or different structures. Equation 4.6 can be further simplified to $(1/2)\phi_{nsp_2} = -\phi_{2s}$, making the location of the boundary a function of the competing effects of cyclization and branching. Note that for a 4-cycle to be formed on an elongated aggregate consisting of more than four bound monomers, at least one of the interacting peptides will need to be able to accommodate three total interactions (further discussion of the influence of pre-existing interactions on new interactions is found in the sections below). When the intercepts of (4.4) and (4.5) are of similar magnitude with a difference less than the net energy of an interaction between two monomers (ϕ_e), small changes in any one parameter may make the difference between forming a 1-ribbon or 2-ribbon fibril, or a 4-cycle oligomer. The relationship between intercept values and the ϕ_e parameter are discussed in the following section.

4.4.2 Dependence on ϕ_e

Noting that these networks were simulated with a fixed $\theta_e = 100$ (using the traditional ERGM form and terms), we observe the intercepts of (4.4) to be approximately located at multiples of θ_e: $-\theta_e/4$, $-\theta_e/3$, $-\theta_e/2$, and $-\theta_e$, for the unstructured-aggregate phase boundary, the pure 1-ribbon phase boundaries, and the pure dimer phase boundaries, respectively (see Figure 4.4). Because edge formation depends on vibrational and entropic contributions in addition to the edge potential (i.e., $\theta_e = -\phi_e/(k_B T) - (1 + \log N)$), phase boundaries do not fall as neatly on multiples of ϕ_e as they do on θ_e. As such, we introduce a parameter to describe the intercepts in relation to ϕ_e that includes the energetic effects of bond vibration and entropic effects of system size; the *net edge energy equivalent*, $\phi_e^* = -\phi_e - k_B T(1+\log N)$, formalizes the relationship of the phase boundaries with the edge potential.

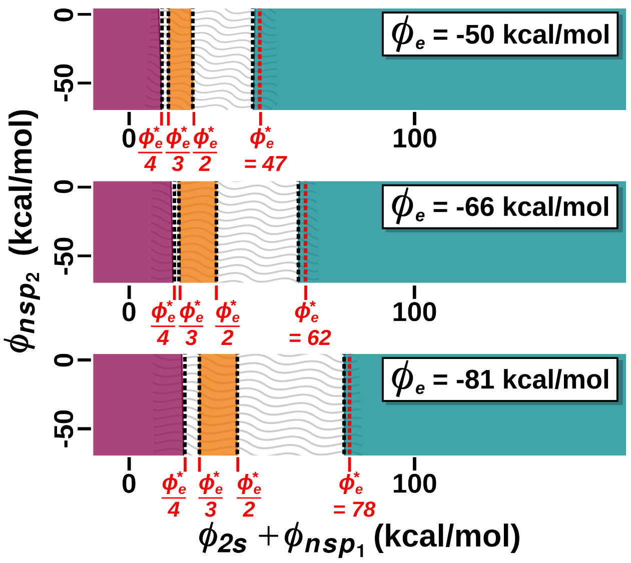


Figure 4.4: Segments of the ϕ_{2s} $+\phi_{nsp_1}$ boundaries (Eq. 4.4) plotted against ϕ_{nsp_2} show the effect of varying ϕ_e values on phase boundary locations. For each ϕ_e values, boundaries are observed to shift by a proportional amount toward or away from $\phi_0 = 0$. Boundary values relative to the net edge energy equivalent value, $\phi_e^* = -(\phi_e + k_BT(1 + \log N))$, are labelled in red.

Networks are sampled throughout the space at values of ϕ_e that are larger ($\phi_e =$ -81 kcal/mol, $\theta_e = 125$, $\phi_e^* = 78$ kcal/mol), and smaller ($\phi_e =$ -50 kcal/mol, $\theta_e = 75$, $\phi_e^* = 47$ kcal/mol) than the original analyses ($\phi_e =$ -66 kcal/mol, $\theta_e = 100$, $\phi_e^* = 62$ kcal/mol), while using the same ranges on ϕ_{nsp_1}, ϕ_{nsp_2}, and ϕ_{2s}. Changes to the magnitude of ϕ_e (and thus, ϕ_e^*) affects the location of phase boundaries in a predictable manner, as shown in Figure 4.4. For larger magnitudes of ϕ_e^*, we observed a shift of the pure phase boundaries toward higher values of ϕ_0 in Eq. 4.4 that remain equivalent to the relative values of ϕ_e^* described above. Similarly, smaller magnitudes of ϕ_e result in lower values of ϕ_0 that maintain the relative values of ϕ_e^*.

The linear relationship between phase boundary location and the relative value of ϕ_e^* highlights the advantage of conducting simulations and analyses using network potentials that are unbiased by local spatial effects such as the monomer shapes in molecular CG models: changes in state stability can be understood in terms of simple balances of multi-body forces, without having to specify the specific factors giving rise to those forces. In the case of proteins, the value of ϕ_e^* is related to the net free energy required for any two monomers (bound or free) to form a stable interaction, and may vary depending on physical parameters of the system being studied. For instance, a smaller ϕ_e value may be related to a peptide that has fewer degrees of freedom that can be employed for stabilizing an interaction, such as polar residues that are able to form salt bridges with other peptides, or surface area available for interaction[18, 110]. Smaller ϕ_e values result in smaller portions of the parameter space that produce networks with fibrils as there is less energy available for monomers to accommodate additional interactions beyond the formation of a dimer.

Specifically, 1-ribbon fibrils are formed for intercepts of Eq. 4.4 between $\phi_e^*/3$ and $\phi_e^*/2$. In this region, the third interaction for any monomer results in three local topologies of both t_{2s} and t_{nsp_1} (see SI Fig. S10B), resulting in $3\phi_{2s} - 3\phi_{nsp_1} > \phi_e^*$; i.e. the stabilizing effect of elongation (ϕ_{nsp_1}) is not enough to accommodate the strain placed on the existing interactions (ϕ_{2s}) by the addition of the third interaction, and the net free energy is less than that required (ϕ_e^*) for an edge to form. Similarly, networks formed between $\phi_e^*/2 < \phi_{2s} - \phi_{nsp_1} < \phi_e^*$ contain a mixture of dimers and proto-fibrils; while elongation may occur, the additional interactions are strenuous and prevent structures from forming that have more than three bound monomers. Networks sampled in the region $\phi_{2s} - \phi_{nsp_1} > \phi_e^*$ contain only dimers as bound monomers are incapable of accommodating additional interactions. Physical protein structures may experience this inhibitory effect due to electrostatic or other interactions that prevent a stable bond from forming [174]. Alternatively, elongation may be constrained by conformational limits of the peptide [146]. Given that dimers are a well known precursor to oligomers, protofibrils, and many other aggregation states [18, 203], the presence of a large region of the parameter space occupied by dimer phases is compatible with the view that this is an easily achieved state.

Unstructured aggregates form when $\phi_0 < \phi_e^*/4$, i.e. when the addition of a fourth edge to any vertex can be mediated by other vertices with fewer edges elsewhere in the gel-like network (note that a fourth edge on a vertex gives a count of six t_{2s} and t_{nsp_1}). For $\phi_0 < 0$, the magnitude of ϕ_{nsp_1} outweighs that of ϕ_{2s}, and edges may be added to any vertex such that the density of the graph is now only limited by the reference measure. This phase may include states in which peptides remain unfolded and disordered, resulting in transient interactions with other peptides that do not lead to stable patterns of interactions such as those needed to form β-sheets; alternately, it may also include states in which monomers remain globally well-folded, but sustain enough local disorder to sustain the interactions needed to condense into a gel.

4.4.3 Mixed Phases

The 2-ribbon, 4-cycle oligomer, and cubic oligomer structures are found to co-occur at varying proportions in the region of the parameter space between $\phi_e^*/4 < (\phi_{2s} + \phi_{nsp_1}) < \phi_e^*$, and surround the pure 1-ribbon phase. The majority of points sampled from the mixed phases produce networks largely composed of 4-cycle structures, with a much smaller quantity of 2-ribbon structures, and an even smaller quantity of cubic structures. 2-ribbon and cubic structures are located within the range given by $\phi_e^*/4 < (\phi_{2s} + \phi_{nsp_1}) < \phi_e^*/2$, with some dimer and 4-cycle structure found at values extending toward ϕ_e^*, as shown in Fig. 4.5A-C. The mixed phases appear to form in layers corresponding roughly to intercepts on (4.4), although the boundaries between them are not all parallel, and would require sampling at a finer scale than we have available here to determine precise locations and intersections.

The second phase boundary between the pure 1-ribbon phase and the mixed 2-ribbon and oligomer phases, shown in Fig. 4.5D-E, depends on the relative magnitudes of ϕ_{nsp_1} and ϕ_{nsp_2}, as given by Eq. 4.5. We note that, unlike 2-stars, NSP(k) terms are not "nested," such that one NSP(2) does not also contain two NSP(1) structures. At $\phi_0 = 0$ the closure-inducing force of ϕ_{nsp_2} is equivalent to twice the force of elongation by ϕ_{nsp_1}, forming the upper boundary of the majority 2-ribbon phase. For $\phi_0 > 0$, the cycle-closure effect of t_{nsp_2} is unfavorable

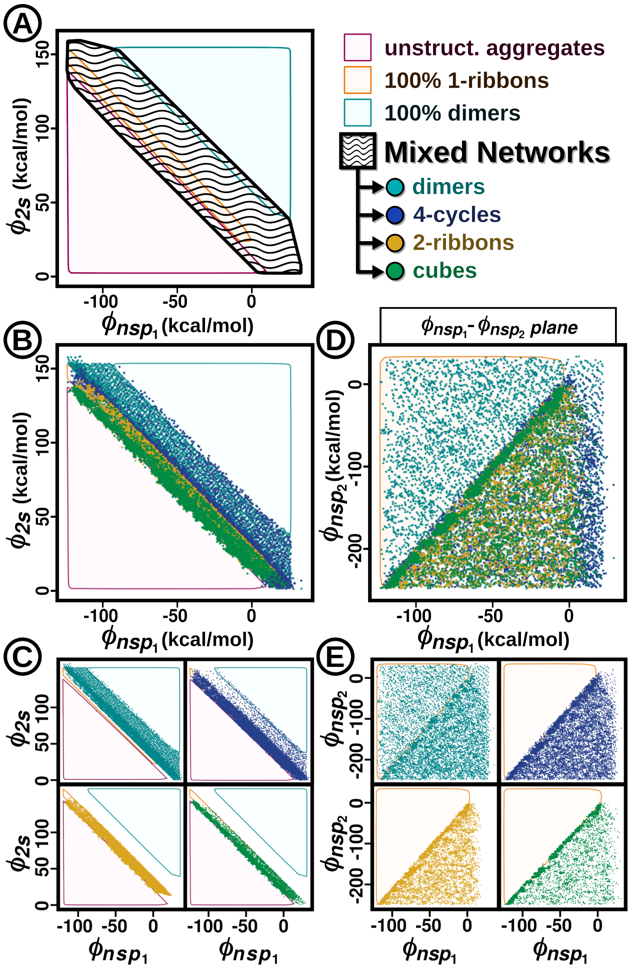


Figure 4.5: The region between pure dimer and pure unstructured-aggregate phases (shown in previous plots by wavey texture) contains mixtures of target structures. A) Copy of the plot in Fig 4.3A with the mixed network phases in bolded wave texture. Pure phases are in muted colors for visual reference. B) Mixed networks are plotted individually, colored by the target structures produced. All points contain multiple structure types. C) The same plot as in B, separated by structure type. (Clockwise: dimers (turquoise), 4-cycles (blue), cubes (green), and 2-ribbons (yellow). D) A copy of Fig. 4.3B, with mixed networks plotted as points. E) The same plot as in D, separated by structure type.

due to the necessary removal of two t_{nsp_1} for every t_{nsp_2}. Thus the elongation effect of t_{nsp_1} drives the structure formation. Large cycles may still form at $\phi_0 > 0$, however, it is more likely that a terminal vertex will form a tie with the terminal vertex of a different structure, of which there are multiple, than with the other terminal vertex of its own structure, of which there is only one. This is consistent with behavior of amyloid fibrils, which can be rigid and inflexible, and exist in an environment where other obstructions may make it difficult to form cycles from a single fibril. For $\phi_0 < 0$, we observe a mixture of 1-ribbons, 2-ribbons, 4-cycles, and "cubic" oligomers (annularized 2-ribbons), and as ϕ_0 becomes more negative the 4-cycle and cubic oligomer structures become dominant. Predominantly oligomeric networks form when $\phi_0 < -\phi_e^*/3$, or $\sum [2\phi_{nsp_1}, \phi_{nsp_2}]^T \cdot 3[t_{nsp_1}, t_{nsp_2}] < \phi_e^* t_e$, i.e. when formation of six t_{nsp_2} upon cyclization of the 2-ribbon into a cubic oligomer (see SI Fig S11) is more stabilizing than maintaining three t_{nsp_1} in the extended 2-ribbon structure . For peptide aggregates, this may occur as a result of environmental interactions such as protein concentration [10, 186, 229] or the presence of lipids [108, 4, 190].

The plane formed by Eqs. 4.4 and 4.5 (shown in Fig. 4.3C) reveals a third phase boundary between the pure 1-ribbon phase and the mixed 2-ribbon and oligomer phases described by (4.6), which is simplified to $(1/2)\phi_{nsp_2} = -\phi_{2s}$ as explained above. The 2-ribbon phase is found within $\phi_e^*/6 < \phi_0 < \phi_e^*/4$, with the 1-ribbon phase beginning at $\phi_0 > \phi_e^*/3$, and the mixed 4-cycle oligomer phase forming for all $\phi_0 < \phi_e^*/2$. This corresponds with our previous observations on the location of phase boundaries relative to the value of ϕ_e; while we have effectively normalized the elongation effects of $\phi_{nsp_1} \cdot t_{nsp_1}$ (the -ϕ_{nsp_1} on both sides of (4.6) can be separated from the Hamiltonian by dividing (4.1) by $\exp[\phi_{nsp_1} t_{nsp_1}/k_B T]$), the relationships observed in (4.4) and (4.5) are still present via ϕ_{2s} and ϕ_{nsp_2} and their respective relationships with ϕ_e^*.

The networks with 2-ribbon yields >50% are thus found only within the ranges $\phi_e^*/4 < (\phi_{2s} + \phi_{nsp_1}) < \phi_e^*/2$, $-\phi_e^*/3 < (\phi_{nsp_2} - 2\phi_{nsp_1}) < 0$, and $\phi_e^*/6 < ((1/2)\phi_{nsp_2} + \phi_{2s}) < \phi_e^*/4$.

This volume in the parameter space forms a thin sliver that follows the 1-ribbon boundary, and includes cube and 4-cycle structures. “Pure” 2-ribbon networks were not produced with the given network terms (always coexisting in equilibrium with structured oligomers), and may require additional terms in the Hamiltonian. Alternately, it may be the case that pure 2-ribbon equilibria do not arise in nature, as 2-ribbons seen from PDB structures of amyloid fibrils would have been obtained by removing non-fibrillar material prior to structure determination.

System Size Effects

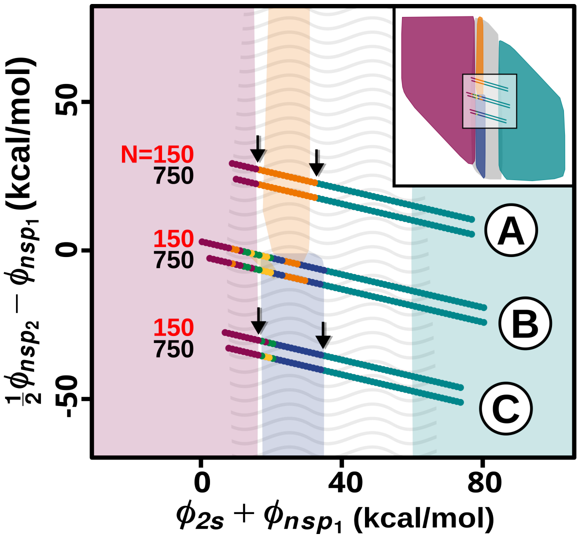


Figure 4.6: Sampling along $\phi_{2s} = \frac{5}{2}\phi_{nsp_1} + \phi_i$ shows effects of sample size (N) on phase boundary locations (with vertical adjustments to prevent overlap), with intercepts ϕ_i at 48, 1, and -47 kcal/mol on Eq. 4.5 for A, B, and C, respectively. The top trajectories for all three sets are networks with N=150, and the bottom are those with N=750. A) The pure 1-ribbon phase boundaries are the same for both system sizes, shown with arrows pointing to the boundary networks. B) The 1-ribbon/4-cycle oligomer boundary and the 1-ribbon/2-ribbon boundary are similar, with 1-ribbon and 2-ribbon networks appearing in the same locations for both system sizes. C) Mixed 4-cycle oligomer phase and unstructured-aggregate phase boundaries are the same for both system sizes, shown by arrows pointing to boundary networks. Points are colored according to the majority target structure (colors match those in Fig. 4.2)

As shown in Fig. 4.6, there is no significant movement of the phase boundaries after increasing the system to five times its original size. Moreover, the similarities between target structure yields indicate that the properties of the networks produced at these points in the parameter space are intrinsic properties of the model itself. The changes in system size do have some noticeable effects on the regions of the space between pure phases (i.e., regions of mixed phase). Notably, smaller networks have more fluctuation in relative proportions of target structures, with higher proportions of the less represented structures (see SI Fig. S9). We thus conclude that the pure phase boundaries are well-characterized for systems on the scale of $\geq 10^2$ monomers.

4.5 Discussion: Calculating the Net Energy of Interactions

The favorability of interaction depends on the number of interactions a free or bound monomer is already involved in, and the ability of the bound monomer to accommodate an additional interaction. As shown in Figure 4.7, the calculation of the total energy of a network system after an addition of an edge is impacted by the properties of the vertices between which the edge was formed. For the simple examples in Fig. 4.7, we focus on edge additions between a terminal vertex of a 1-ribbon and a vertex that is either part of a separate structure, or part of the same structure as the initiating vertex. This distinction is necessary, as edges that bind two separate structures are fundamentally incapable of forming cycles. On the other hand, edges that bind two vertices of the same structure inherently inhibit further elongation. Additionally, as noted previously, cycle-closure interactions within an elongated aggregate necessarily require at least one bound monomer to accommodate a third interaction (illustrated in the bottom row of Fig. 4.7).

This is relevant in discussions of fibrilization of peptides, as some of the most toxic species

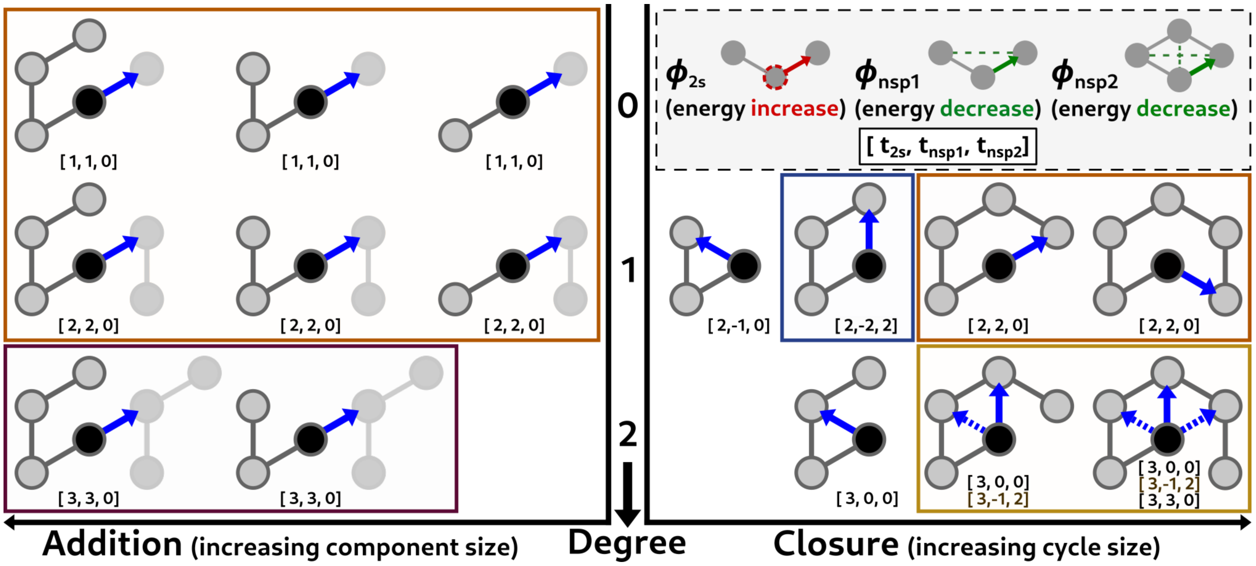


Figure 4.7: Change score calculations depend in part on the prior graph state. The degree of vertices prior to edge addition is shown in the figure as increasing by row, starting with degree = 0, with the added edge indicated by a blue arrow pointing towards the vertex whose degree is being indicated. When an edge is added between a terminal vertex of a 1-ribbon and a vertex with degree 0 or 1, the resulting structure is a 1-ribbon, with a change in graph energy equal to $\sum[\phi_{2s}, \phi_{nsp_1}, \phi_{nsp_2}]^T[t_{2s}, t_{nsp_1}, t_{nsp_2}]$ where the change score ($[t_{2s}, t_{nsp_1}, t_{nsp_2}]$) is either [1,1,0] or [2,2,0], as indicated below the example structures. When the added vertex increases the component size, the result is elongation, shown on the left half of the figure. When the added vertex is part of the same component, the result is cycle closure, shown on the right half of the figure. Some structures result in an increase in the total graph energy, such as the structures with degree 1 and 2 closure that form 3-cycles. As such, these structures are rarely observed in the networks studied here, and receive no classification. Otherwise, orange boxes indicate 1-ribbon structures, purple indicates unstructured-aggregates, blue indicates 4-cycle oligomers, and yellow indicates structures that may lead to 2-ribbon (or cubic oligomer) structures. These examples are far from exhaustive, but illustrate the dependence of fibril formation on the properties of surrounding monomers and aggregates.

are thought to be small oligomers, similar to the 4-cycle and "cubic" oligomers studied here. Formation of 4-cycles and cubic structures requires high energy bonds that can overcome the loss of entropy that is had by constraining the conformation of the peptides. As shown in the discussions for Eqs 4.5 and 4.6, cycles form only when elongating t_{nsp_1} topologies are less favorable than cyclizing t_{nsp_2} topologies, i.e. when interactions with separate structures or free monomers is unlikely or destructive and the formation of closed cycles allows aggregate structures to persist.

The value of ϕ_e^*, described as the *net edge energy equivalent* above, is equal to the magnitude of the base edge energy adjusted by $k_BT(1 + \log N)$. This adjustment accounts for the combined energetic and entropic effects of bond vibration and system size, returning ϕ_e^* to a value that, relative to other forces in the NHM, more closely resembles the network parameterization of θ_e. This is necessary for descriptions of phase boundaries using NHMs, as time-scale separated dynamics (vibration and diffusion) do impact edge formation, and must be accounted for.

4.6 Conclusion

NHMs are a scalable and insightful framework for studying protein aggregation, allowing both exploration of the space of possible multi-body interactions and rationalization of the resulting equilibria in terms of intermolecular forces. Examination of a simple, four-term model reveals a very rich structure, with multiple distinct fibrillar phases, a gel-like phase, and multiple oligomeric phases (including complex, annular oligomers). We find that the dominant drivers of the phase structure are respectively competition between hindrance of increasing contact numbers (on the one hand) and favorability of elongation of locally linear structures (on the other), and the competition between the latter elongation forces and the favorability of locally closed, loop-like structures. Elongation must dominate hindrance to obtain fibrils and complex oligomers, though if the overall favorability of bonding relative to hindrance becomes too large, the system collapses into a gel phase. 2-ribbon fibrils arise when elongation and formation of contacts is favored sufficiently for the mean contact number to grow beyond what a 1-ribbon can sustain, but below that of the gel phase; in addition, loop closure must be sufficiently favorable to lead to the characteristic “stacked 4-cycle” of the 2-ribbon, but not so strong as to cause the fibril to collapse into oligomers. This complex balance may explain why we do not see parameters with 100% 2-ribbon yield, though we do see cases with yields in excess of 50%.

Direct parameterization in terms of net intermolecular forces is a virtue for generality of insight, but has trade-offs with respect to atomistic detail: we cannot, on the basis of these models alone, specify which polypeptide sequences and experimental conditions reside within which part of the parameter space. However, specification of the intermolecular forces needed to give rise to specific aggregation states provides experimental targets for their measurement, and also motivates MD studies to predict forces of the type used here from atomistic models. The picture provided here is thus complementary to other experimental and theoretical techniques for probing the phenomenon of protein aggregation.

Chapter 5

Summary

5.1 Contributions

- My analysis of the SARS-CoV-2 Main Protease revealed trends in its evolution that describe how the viral reproduction process adapts to human hosts, thus allowing the virus to persist and proliferate despite society's preventative efforts as well as lineage truncation that naturally results from host deaths. Protein Structure Networks, typically used for studying structural properties of the protein itself, here provided hints about the effects of the host environment on the protein's adaptation; changes in structural cohesion give insight into the thermodynamic constraints placed on the protein by the environment. A closer look at the amino acid substitutions themselves gives clues about the residue interactions that are necessary for dimerization - and thus function of the protease - by revealing which residues persist despite hundreds of opportunities for mutation. The combination of these analyses tells an integral part of the story of how the reproductive process of SARS-CoV-2 preserves the structure and function of the Main Protease while adapting to new hosts.

- I used the framework of Network Hamiltonian Models to scale network simulations of WT γ-Dc and its W42R variant based on mcMC simulations by Wong, et al,[209] by two orders of magnitude, from 375 monomers to 10,000, which is unreachable by typical molecular dynamics simulations using today's computational resources. In the process, I was also able to examine the effects of system size on the simulation of aggregates, and discovered a change in the distribution of aggregate sizes in simulations of larger systems. While this may not necessarily be an artefact that would be observed in a physical system, it does highlight a detail that deserves attention when running computational simulations of disordered aggregates at various system sizes.

- I also used Network Hamiltonian Models to examine the emergent phases of ordered peptide aggregates by varying the coefficient on network terms in the Hamiltonian, revealing the sensitivity of aggregate structures to the balance of a minimal set of terms describing multi-body interactions. The ratios of elongation-inducing and contact-inhibiting forces are revealed to be directly related to the energy of each additional edge or contact; this relationship is thus shown to dominate the phase behavior of fibril formation, with secondary effects observed to result from the interaction between closure-inducing forces and forces that encourage aggregate growth. In contrast with the γ-Dc results, ordered aggregation patterns are observed to be consistent regardless of system size, which highlights a fundamental difference in the aggregation behavior of ordered versus disordered systems.

5.2 Limitations

- Data for study of the SARS-CoV-2 Main Protease was provided by clinical samples, which inherently includes limitations due to human reporting error, as well as limitations on the breadth of observable mutations that can occur to the protease. Additionally, the study done here was entirely computational, and would benefit from corroborating experimental studies. This is, however, unfeasible given the large number (1,253) variants. This large number also limits the computational power that can be devoted to simulation of each variant. As such, simulated dynamics trajectories were limited to 20ps, while simulations on the order of nano-seconds would have provided more information about the dynamics and structural relationships of the protease that govern dimerization and active site function.

- Simulations of γ-Dc aggregation in these analyses were inherently limited by the results of the original mcMC simulations[209], which were reliant upon conformations of γ-Dc that were sampled from single, and two monomer simulations. Additionally, the mcMC simulations are limited by system size, and a comparison of results at larger system sizes is unavailable. Finally, it is infeasible to corroborate the results of either simulation method with experimental results, as these are naturally very dense system that are difficult to reproduce, and experimental systems would necessarily involve much larger system sizes than can be computed.

- The simulation of phases involves computation of a large number of systems that are unlikely to have physical significance, and thus are difficult to study experimentally. Experimental system also are beholden to environmental effects that are not representable in the Network Hamiltonian Model framework, such as thermodynamic variables of volume, temperature, and pressure. Network Hamiltonian Models are also presently unable to capture interactions with surfaces, or other chemical species in the system that may affect the aggregate products. Finally, the work here describes fundamental phenomena that, while vital to understanding underlying forces that govern the process of fibril formation, can appear trivial when considering real systems, making the importance of this work difficult to communicate to experimental researchers.

5.3 Next Steps

- The process governing formation and dimerization of the SARS-CoV-2 Main Protease is still largely a mystery, however it is believed that the protease is initially anchored to the membrane of the endoplasmic reticulum by neighboring sections of the viral polypeptide as it is transcribed by host rybosomes. Computational simulations of the protease in hydrophobic media (such as a membrane) could be compared with identical simulations in water to observe differences in dynamics of the monomer and dimer structures, giving insight into how the protease is lysed from the viral polypeptide, and how dimerization might occur.

- The conformational states of WT gD-Crystallin and its W42R variant that are used in mcMC simulations by Wong, et al,[209] provide a potential connection between the topological patterns described by network terms and the chemical interactions that determine the strength of the physical interaction. Further analysis of the relationship between monomer conformations and their local graph environment could be performed using classic network analysis methods to more precisely define how the coarsened aggregate topology is impacted by the fine-grain details of chemistry.

- The fundamental nature of the phases described by network models of multi-body interactions leaves a wide array of possibilities for future directions. Additional network terms may be added to the Network Hamiltonian to explore how the added complexity affects aggregate topology given the known relationships shown in the present work. Alternatively, descriptions for other topologies, such as crystalline sheets or lattices, may be defined and searched for using the current model, perhaps with adjustments to the given coefficients. An experimental study may also be done using a chosen unique peptide sequence to determine how environmental parameters may be controlled to empirically reproduce a variety of fibrillar phases found in the current models. The boundaries are determined by the inputs.

Bibliography

[1] S. Abeln, M. Vendruscolo, C. M. Dobson, and D. Frenkel. A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation. *PLOS ONE*, 9(1):e85185, Jan. 2014.

[2] M. Adachi, M. Noji, M. So, K. Sasahara, J. Kardos, H. Naiki, and Y. Goto. Aggregation-phase diagrams of $B2$-microglobulin reveal temperature and salt effects on competitive formation of amyloids versus amorphous aggregates. *Journal of Biological Chemistry*, 293(38):14775–14785, Sept. 2018.

[3] K. Afitska, A. Fucikova, V. V. Shvadchak, and D. A. Yushchenko. α-Synuclein aggregation at low concentrations. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1867(7):701–709, July 2019.

[4] H. Ahyayauch, M. Masserini, F. M. Goñi, and A. Alonso. The interaction of Aβ42 peptide in monomer, oligomer or fibril forms with sphingomyelin/cholesterol/ganglioside bilayers. *International Journal of Biological Macromolecules*, 168:611–619, Jan. 2021.

[5] I. Altan, A. R. Khan, S. James, M. K. Quinn, J. J. McManus, and P. Charbonneau. Using Schematic Models to Understand the Microscopic Basis for Inverted Solubility in γD-Crystallin. *J. Phys. Chem. B*, 123(47):10061–10072, Nov. 2019.

[6] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski. Network Analysis of Protein Structures Identifies Functional Residues. *Journal of Molecular Biology*, 344(4):1135–1146, Dec. 2004.

[7] K. Anand, G. Palm, J. Mesters, S. Siddell, J. Ziebuhr, and R. Hilgenfeld. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. *EMBO J.*, 21(13):3213–3224, July 2002.

[8] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, and R. Hilgenfeld. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. *Science*, 300(5626):1763–1767, June 2003.

[9] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry. The proximal origin of SARS-CoV-2. *Nat Med*, 26(4):450–452, Apr. 2020.

[10] T. Arakawa, D. Ejima, and T. Akuta. Protein aggregation under high concentration/density state during chromatographic and ultrafiltration processes. *International Journal of Biological Macromolecules*, 95:1153–1158, Feb. 2017.

[11] N. Asherie, J. Pande, A. Lomakin, O. Ogun, S. R. A. Hanson, J. B. Smith, and G. B. Benedek. Oligomerization and phase separation in globular protein solutions. *Biophysical Chemistry*, 75(3):213–227, Dec. 1998.

[12] S. Auer, F. Meersman, C. M. Dobson, and M. Vendruscolo. A Generic Mechanism of Emergence of Amyloid Protofilaments from Disordered Oligomeric Aggregates. *PLOS Computational Biology*, 4(11):e1000222, Nov. 2008.

[13] J. M. Axe and D. D. Boehr. Long-Range Interactions in the Alpha Subunit of Tryptophan Synthase Help to Coordinate Ligand Binding, Catalysis, and Substrate Channeling. *Journal of Molecular Biology*, 425(9):1527–1545, May 2013.

[14] K. J. Bari and S. Sharma. A Perspective on Biophysical Studies of Crystallin Aggregation and Implications for Cataract Formation. *J. Phys. Chem. B*, 124(49):11041–11054, Dec. 2020.

[15] J. Barrila, U. Bacha, and E. Freire. Long-Range Cooperative Interactions Modulate Dimerization in SARS 3CLpro. *Biochemistry*, 45(50):14908–14916, Dec. 2006.

[16] B. Barz, Q. Liao, and B. Strodel. Pathways of Amyloid-β Aggregation Depend on Oligomer Shape. *J. Am. Chem. Soc.*, 140(1):319–327, Jan. 2018.

[17] A. Basak, O. Bateman, C. Slingsby, A. Pande, N. Asherie, O. Ogun, G. B. Benedek, and J. Pande. 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*, 328(5):1137–1147, May 2003.

[18] G. Bellesia and J.-E. Shea. What Determines the Structure and Stability of KFFE Monomers, Dimers, and Protofibrils? *Biophysical Journal*, 96(3):875–886, Feb. 2009.

[19] N. C. Benson and V. Daggett. A Chemical Group Graph Representation for Efficient High-Throughput Analysis of Atomistic Protein Simulations. *J Bioinform Comput Biol*, 10(4):1250008, Aug. 2012.

[20] S. L. Bernstein, T. Wyttenbach, A. Baumketner, J.-E. Shea, G. Bitan, D. B. Teplow, and M. T. Bowers. Amyloid β-Protein: Monomer Structure and Early Aggregation States of Aβ42 and Its Pro19 Alloform. *J. Am. Chem. Soc.*, 127(7):2075–2084, Feb. 2005.

[21] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. R. Stat. Soc. Ser. B Methodol.*, 36(2):192–225, 1974.

[22] J. Besag. Markov chain Monte Carlo for statistical inference. Technical Report CSSS Working Paper 9, University of Washington, 2000.

[23] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell. 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. *J. Chem. Theory Comput.*, 8(9):3257–3273, Sept. 2012.

[24] H. Bloemendal, W. de Jong, R. Jaenicke, N. H. Lubsen, C. Slingsby, and A. Tardieu. Ageing and vision: Structure, stability and function of lens crystallins. *Progress in Biophysics and Molecular Biology*, 86(3):407–485, Nov. 2004.

[25] J. C. Boatz, M. J. Whitley, M. Li, A. M. Gronenborn, and P. C. A. van der Wel. Cataract-associated p23t γd-crystallin retains a native-like fold in amorphous-looking aggregates formed at physiological pH. *Nature Communications*, 8(1):15137, May 2017.

[26] M. Boob, Y. Wang, and M. Gruebele. Proteins: "Boil 'Em, Mash 'Em, Stick 'Em in a Stew". *J. Phys. Chem. B*, 123(40):8341–8350, Oct. 2019.

[27] K. V. Brinda, A. Surolia, and S. Vishveshwara. Insights into the quaternary association of proteins through structure graphs: A case study of lectins. *Biochem. J.*, 391(1):1–15, Oct. 2005.

[28] A. K. Buell, C. Galvagnion, R. Gaspar, E. Sparr, M. Vendruscolo, T. P. J. Knowles, S. Linse, and C. M. Dobson. Solution conditions determine the relative importance of nucleation and growth processes in α-synuclein aggregation. *Proc. Natl. Acad. Sci.*, 111(21):7671–7676, May 2014.

[29] C. T. Butts. **Network** : A Package for Managing Relational Data in *R*. *J. Stat. Soft.*, 24(2), 2008.

[30] C. T. Butts. Social Network Analysis with sna. *J. Stat. Softw.*, 24:1–51, May 2008.

[31] C. T. Butts. Bernoulli Graph Bounds for General Random Graphs. *Sociol. Methodol.*, 41(1):299–345, 2011.

[32] C. T. Butts. A dynamic process interpretation of the sparse ERGM reference model. *J. Math. Sociol.*, 43(1):40–57, Jan. 2019.

[33] C. T. Butts. Phase transitions in the edge/concurrent vertex model. *J. Math. Sociol.*, 45(3):135–147, July 2021.

[34] C. T. Butts, X. Zhang, J. E. Kelly, K. W. Roskamp, M. H. Unhelkar, J. A. Freites, S. Tahir, and R. W. Martin. Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Comput. Struct. Biotechnol. J.*, 14:271–282, 2016.

[35] A. Caflisch. Computational models for the prediction of polypeptide aggregation propensity. *Current Opinion in Chemical Biology*, 10(5):437–444, Oct. 2006.

[36] E. Callaway. Could new COVID variants undermine vaccines? Labs scramble to find out. *Nature*, 589(7841):177–178, Jan. 2021.

[37] S. P. Carmichael and M. S. Shell. A New Multiscale Algorithm and Its Application to Coarse-Grained Peptide Models for Self-Assembly. *J. Phys. Chem. B*, 116(29):8383–8393, July 2012.

[38] R. Carrotta, M. Manno, D. Bulone, V. Martorana, and P. L. S. Biagio. Protofibril Formation of Amyloid β-Protein at Low pH via a Non-cooperative Elongation Mechanism*. *Journal of Biological Chemistry*, 280(34):30001–30008, Aug. 2005.

[39] R. Cascella, S. W. Chen, A. Bigi, J. D. Camino, C. K. Xu, C. M. Dobson, F. Chiti, N. Cremades, and C. Cecchi. The release of toxic oligomers from α-synuclein fibrils induces dysfunction in neuronal cells. *Nat Commun*, 12(1):1814, Mar. 2021.

[40] C. Ceraolo and F. M. Giorgi. Genomic variance of the 2019-nCoV coronavirus. *Journal of Medical Virology*, 92(5):522–528, 2020.

[41] R. Chakraborty, S. Dey, P. Sil, S. S. Paul, D. Bhattacharyya, A. Bhunia, J. Sengupta, and K. Chattopadhyay. Conformational distortion in a fibril-forming oligomer arrests alpha-Synuclein fibrillation and minimizes its toxic effects. *Commun Biol*, 4(1):1–14, May 2021.

[42] T. Chakroun, V. Evsyukov, N.-P. Nykänen, M. Höllerhage, A. Schmidt, F. Kamp, V. C. Ruf, W. Wurst, T. W. Rösler, and G. U. Höglinger. Alpha-synuclein fragments trigger distinct aggregation pathways. *Cell Death Dis*, 11(2):1–16, Feb. 2020.

[43] Y. Chebaro, S. Pasquali, and P. Derreumaux. The Coarse-Grained OPEP Force Field for Non-Amyloid and Amyloid Proteins. *J. Phys. Chem. B*, 116(30):8741–8752, Aug. 2012.

[44] B. Chen and J. I. Siepmann. A Novel Monte Carlo Algorithm for Simulating Strongly Associating Fluids: Applications to Water, Hydrogen Fluoride, and Acetic Acid. *J. Phys. Chem. B*, 104(36):8725–8734, Sept. 2000.

[45] B. Chen and J. I. Siepmann. Improving the Efficiency of the Aggregation-Volume-Bias Monte Carlo Algorithm. *J. Phys. Chem. B*, 105(45):11275–11282, Nov. 2001.

[46] X. Chen, M. Chen, and P. G. Wolynes. Exploring the Interplay between Disordered and Ordered Oligomer Channels on the Aggregation Energy Landscapes of α-Synuclein. *J. Phys. Chem. B*, 126(28):5250–5261, July 2022.

[47] F. Chiti and C. M. Dobson. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *annurev-biochem*, 86:27–68, May 2017.

[48] I.-T. Chu, C. J. Stewart, S. L. Speer, and G. J. Pielak. A Difference between In Vitro and In-Cell Protein Dimer Formation. *Biochemistry*, 61(6):409–412, Mar. 2022.

[49] J. I. Clark. Self-assembly of protein aggregates in ageing disorders: The lens and cataract model. *Philos. Trans. R. Soc. B Biol. Sci.*, 368(1617):20120104, May 2013.

[50] W. Close, M. Neumann, A. Schmidt, M. Hora, K. Annamalai, M. Schmidt, B. Reif, V. Schmidt, N. Grigorieff, and M. Fändrich. Physical basis of amyloid fibril polymorphism. *Nat Commun*, 9(1):699, Feb. 2018.

[51] S. I. A. Cohen, S. Linse, L. M. Luheshi, E. Hellstrand, D. A. White, L. Rajah, D. E. Otzen, M. Vendruscolo, C. M. Dobson, and T. P. J. Knowles. Proliferation of amyloid-$B42$ aggregates occurs through a secondary nucleation mechanism. *Proc. Natl. Acad. Sci.*, 110(24):9758–9763, June 2013.

[52] M. T. Colvin, R. Silvers, Q. Z. Ni, T. V. Can, I. Sergeyev, M. Rosay, K. J. Donovan, B. Michael, J. Wall, S. Linse, and R. G. Griffin. Atomic Resolution Structure of Monomorphic Aβ42 Amyloid Fibrils. *J. Am. Chem. Soc.*, 138(30):9663–9674, Aug. 2016.

[53] T. J. Cross, G. R. Takahashi, E. M. Diessner, M. G. Crosby, V. Farahmand, S. Zhuang, C. T. Butts, and R. W. Martin. Sequence Characterization and Molecular Modeling of Clinically Relevant Variants of the SARS-CoV-2 Main Protease. *Biochemistry*, Sept. 2020.

[54] Z. J. Czenze, S. Naidoo, A. Kotze, and A. E. McKechnie. Bat thermoregulation in the heat: Limits to evaporative cooling capacity in three southern African bats. *Journal of Thermal Biology*, 89:102542, Apr. 2020.

[55] L. Dagum and R. Menon. OpenMP: An industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng.*, 5(1):46–55, Jan. 1998.

[56] A. De Simone, C. Kitchen, A. H. Kwan, M. Sunde, C. M. Dobson, and D. Frenkel. Intrinsic disorder modulates protein self-assembly and aggregation. *Proc. Natl. Acad. Sci.*, 109(18):6951–6956, May 2012.

[57] M. Delaye and A. Tardieu. Short-range order of crystallin proteins accounts for eye lens transparency. *Nature*, 302(5907):415–417, Mar. 1983.

[58] E. M. Diessner, J. A. Freites, D. J. Tobias, and C. T. Butts. Network Hamiltonian Models for Unstructured Protein Aggregates, with Application to γD-Crystallin. *J. Phys. Chem. B*, Jan. 2023.

[59] E. Domingo and C. Perales. Viral quasispecies. *PLOS Genetics*, 15(10):e1008271, Oct. 2019.

[60] M. Dowle, A. Srinivasan, J. Gorecki, M. Chirico, P. Stetsenko, T. Short, S. Lianoglou, E. Antonyan, M. Bonsch, H. Parsonage, S. Ritchie, K. Ren, X. Tan, R. Saporta, O. Seiskari, X. Dong, M. Lang, W. Iwasaki, S. Wenchel, K. Broman, T. Schmidt, D. Arenburg, E. Smith, F. Cocquemas, M. Gomez, P. Chataignon, N. Blaser, D. Selivanov, A. Riabushenko, C. Lee, D. Groves, D. Possenriede, F. Parages, D. Toth, M. Yaramaz-David, A. Perumal, J. Sams, M. Morgan, M. Quinn, R. Storey, M. Saraswat, M. Jacob, M. Schubmehl, D. Vaughan, T. Hocking, L. Silvestri, T. Barrett, J. Hester, A. Damico, S. Freundt, D. Simons, E. S. de Andrade, C. Miller, J. P.

Meldgaard, V. Tlapak, K. Ushey, D. Eddelbuettel, and B. Schwen. Data.table: Extension of 'data.frame', Sept. 2021.

[61] V. T. Duong, E. M. Diessner, G. Grazioli, R. W. Martin, and C. T. Butts. Neural Upscaling from Residue-Level Protein Structure Networks to Atomistic Structures. *Biomolecules*, 11(12):1788, Dec. 2021.

[62] V. T. Duong, M. H. Unhelkar, J. E. Kelly, S. H. Kim, C. T. Butts, and R. W. Martin. Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant Drosera capensis. *Integr. Biol.*, 10(12):768–779, Dec. 2018.

[63] H. Ecroyd and J. A. Carver. Crystallin proteins and amyloid fibrils. *Cell. Mol. Life Sci.*, 66(1):62–81, Jan. 2009.

[64] D. Eddelbuettel and R. Francois. Rcpp: Seamless R and C++ Integration. *J. Stat. Softw.*, 40:1–18, Apr. 2011.

[65] D. S. Eisenberg and M. R. Sawaya. Structural Studies of Amyloid Proteins at the Molecular Level. *Annu. Rev. Biochem.*, 86(1):69–95, 2017.

[66] J. R. Espinosa, J. A. Joseph, I. Sanchez-Burgos, A. Garaizar, D. Frenkel, and R. Collepardo-Guevara. Liquid network connectivity regulates the stability and composition of biomolecular condensates with many components. *Proc. Natl. Acad. Sci.*, 117(24):13238–13247, June 2020.

[67] Y. Fan, Y. Sun, W. Yu, Y. Tao, W. Xia, Y. Liu, Q. Zhao, Y. Tang, Y. Sun, F. Liu, Q. Cao, J. Wu, C. Liu, J. Wang, and D. Li. Conformational change of α-synuclein fibrils in cerebrospinal fluid from different clinical phases of Parkinson's disease. *Structure*, 31(1):78–87.e5, Jan. 2023.

[68] N. L. Fawzi, E.-H. Yap, Y. Okabe, K. L. Kohlstedt, S. P. Brown, and T. Head-Gordon. Contrasting Disease and Nondisease Protein Aggregation by Molecular Simulation. *Acc. Chem. Res.*, 41(8):1037–1047, Aug. 2008.

[69] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.*, 103(11):4613–4621, Sept. 1995.

[70] O. Frank and D. Strauss. Markov Graphs. *J. Am. Stat. Assoc.*, 81(395):832–842, Sept. 1986.

[71] S. D. W. Frost, O. G. Pybus, J. R. Gog, C. Viboud, S. Bonhoeffer, and T. Bedford. Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92, Mar. 2015.

[72] A. Garaizar, J. R. Espinosa, J. A. Joseph, G. Krainer, Y. Shen, T. P. Knowles, and R. Collepardo-Guevara. Aging can transform single-component protein condensates into multiphase architectures. *Proc. Natl. Acad. Sci.*, 119(26):e2119800119, June 2022.

[73] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis Third Edition (with Errors Fixed as of 15 February 2021).* Columbia University, feb 2021.

[74] R. González-Castro, M. A. Gómez-Lim, and F. Plisson. Cysteine-Rich Peptides: Hyperstable Scaffolds for Protein Engineering. *ChemBioChem*, 22(6):961–973, 2021.

[75] R. L. Graham, J. S. Sparks, L. D. Eckerle, A. C. Sims, and M. R. Denison. SARS coronavirus replicase proteins in pathogenesis. *Virus Research*, 133(1):88–100, Apr. 2008.

[76] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. D. Caves. Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696, Nov. 2006.

[77] G. Grazioli, C. T. Butts, and I. Andricioaei. Automated placement of interfaces in conformational kinetics calculations using machine learning. *J. Chem. Phys.*, 147(15):152727, Oct. 2017.

[78] G. Grazioli, R. W. Martin, and C. T. Butts. Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods. *Front. Mol. Biosci.*, 6, 2019.

[79] G. Grazioli, Y. Yu, M. H. Unhelkar, R. W. Martin, and C. T. Butts. Network-Based Classification and Modeling of Amyloid Fibrils. *J. Phys. Chem. B*, 123(26):5452–5462, July 2019.

[80] A. Guzzo, P. Delarue, A. Rojas, A. Nicolaï, G. G. Maisuradze, and P. Senet. Wild-Type α-Synuclein and Variants Occur in Different Disordered Dimers and Pre-Fibrillar Conformations in Early Stage of Aggregation. *Front. Mol. Biosci.*, 9, 2022.

[81] O. Häggström and J. Jonasson. Phase Transition in the Random Triangle Model. *J. Appl. Probab.*, 36(4):1101–1115, 1999.

[82] M. S. Handcock. Statistical Models for Social Networks: Inference and Degeneracy. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 229–240. National Academies Press, Washington, D.C., July 2003.

[83] M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris. Statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *J Stat Softw*, 24(1):1548–7660, 2008.

[84] M. S. Handcock and M. Morris. *Relative Distribution Methods in the Social Sciences.* Statistics for Social Science and Behavorial Sciences. Springer-Verlag, New York, 1999.

[85] T. Härd. Amyloid Fibrils: Formation, Polymorphism, and Inhibition. *J. Phys. Chem. Lett.*, 5(3):607–614, Feb. 2014.

[86] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat Methods*, 14(1):71–73, Jan. 2017.

[87] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8, Feb. 1996.

[88] D. R. Hunter. Curved exponential family models for social networks. *Social Networks*, 29(2):216–230, May 2007.

[89] D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Ergm.userterms: A Template Package for Extending statnet. *J. Stat. Softw.*, 52:1–25, Feb. 2013.

[90] D. R. Hunter and M. S. Handcock. Inference in Curved Exponential Family Models for Networks. *J. Comput. Graph. Stat.*, 15(3):565–583, Sept. 2006.

[91] D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *J. Stat. Softw.*, 24:1–29, May 2008.

[92] D. R. Hunter, P. N. Krivitsky, and M. Schweinberger. Computational Statistical Methods for Social Network Models. *J. Comput. Graph. Stat.*, 21(4):856–882, Oct. 2012.

[93] A. Irbäck, S. Æ. Jónsson, N. Linnemann, B. Linse, and S. Wallin. Aggregate Geometry in Amyloid Fibril Nucleation. *Phys. Rev. Lett.*, 110(5):058101, Jan. 2013.

[94] S. Izvekov and G. A. Voth. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B*, 109(7):2469–2473, Feb. 2005.

[95] J. A. Jedziniak, J. H. Kinoshita, E. M. Yates, L. O. Hocker, and G. B. Benedek. On the presence and mechanism of formation of heavy molecular weight aggregates in human normal and cataractous lenses. *Experimental Eye Research*, 15(2):185–192, Feb. 1973.

[96] F. Ji, J. Jung, L. M. I. Koharudin, and A. M. Gronenborn. The Human W42R γD-Crystallin Mutant Structure Provides a Link between Congenital and Age-related Cataracts. *Journal of Biological Chemistry*, 288(1):99–109, Jan. 2013.

[97] J. Jorda and T. O. Yeates. Widespread Disulfide Bonding in Proteins from Thermophilic Archaea. *Archaea*, 2011:e409156, Sept. 2011.

[98] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, July 1983.

[99] M. H. Kalos and P. A. Whitlock. *Monte Carlo Methods. Vol. 1: Basics.* Wiley-Interscience, USA, 1986.

[100] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[101] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780, Apr. 2013.

[102] A. R. Khan, S. James, M. K. Quinn, I. Altan, P. Charbonneau, and J. J. McManus. Temperature-Dependent Interactions Explain Normal and Inverted Solubility in a γD-Crystallin Mutant. *Biophysical Journal*, 117(5):930–937, Sept. 2019.

[103] S. Khare, C. Gurry, L. Freitas, M. B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R. T. Lee, W. Yeo, G. C. C. Team, and S. Maurer-Stroh. GISAID's Role in Pandemic Response. *CCDCW*, 3(49):1049–1051, Dec. 2021.

[104] P. N. Krivitsky. Exponential-Family Random Graph Models for Valued Networks. *Electron. J. Statist.*, 6(0):1100–1128, 2012.

[105] P. N. Krivitsky, M. S. Handcock, D. R. Hunter, C. T. Butts, C. Klumb, S. M. Goodreau, and M. Morris. Statnet: Tools for the Statistical Modeling of Network Data. Statnet Development Team, 2003/2022.

[106] P. N. Krivitsky, M. S. Handcock, and M. Morris. Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 8(4):319–339, July 2011.

[107] S. Kumar and R. Nussinov. How do thermophilic proteins deal with heat? *CMLS, Cell. Mol. Life Sci.*, 58(9):1216–1233, Aug. 2001.

[108] A. S. Kurochka, D. A. Yushchenko, P. Bouř, and V. V. Shvadchak. Influence of Lipid Membranes on α-Synuclein Aggregation. *ACS Chem. Neurosci.*, 12(5):825–830, Mar. 2021.

[109] Y. Kusumoto, A. Lomakin, D. B. Teplow, and G. B. Benedek. Temperature dependence of amyloid β-protein fibrillization. *Proc. Natl. Acad. Sci.*, 95(21):12277–12282, Oct. 1998.

[110] L. Larini and J.-E. Shea. Role of β-Hairpin Formation in Aggregation: The Self-Assembly of the Amyloid-β(25–35) Peptide. *Biophysical Journal*, 103(3):576–586, Aug. 2012.

[111] M. Lee, W.-M. Yau, J. M. Louis, and R. Tycko. Structures of brain-derived 42-residue amyloid-β fibril polymorphs with unusual molecular conformations and intermolecular interactions. *Proc. Natl. Acad. Sci.*, 120(11):e2218831120, Mar. 2023.

[112] S. W. Lee, H. Choi, G. Lee, Y. Choi, H. Lee, G. Kim, H. Lee, W. Lee, J. Park, and D. S. Yoon. Conformation Control of Amyloid Filaments by Repeated Thermal Perturbation. *ACS Macro Lett.*, 10(12):1549–1554, Dec. 2021.

[113] D. Li and C. Liu. Conformational strains of pathogenic amyloid proteins in neurodegenerative diseases. *Nat Rev Neurosci*, 23(9):523–534, Sept. 2022.

[114] M. S. Li, N. T. Co, G. Reddy, C.-K. Hu, J. E. Straub, and D. Thirumalai. Factors Governing Fibrillogenesis of Polypeptide Chains Revealed by Lattice Models. *Phys. Rev. Lett.*, 105(21):218101, Nov. 2010.

[115] H. Liu, S. K. Kumar, and F. Sciortino. Vapor-liquid coexistence of patchy models: Relevance to protein phase behavior. *J. Chem. Phys.*, 127(8):084902, Aug. 2007.

[116] A. Lomakin, N. Asherie, and G. B. Benedek. Monte Carlo study of phase separation in aqueous protein solutions. *J. Chem. Phys.*, 104(4):1646–1656, Jan. 1996.

[117] N. Louros, M. Ramakers, E. Michiels, K. Konstantoulea, C. Morelli, T. Garcia, N. Moonen, S. D'Haeyer, V. Goossens, D. R. Thal, D. Audenaert, F. Rousseau, and J. Schymkowitz. Mapping the sequence specificity of heterotypic amyloid interactions enables the identification of aggregation modifiers. *Nat Commun*, 13(1):1351, Mar. 2022.

[118] J.-X. Lu, W. Qiang, W.-M. Yau, C. D. Schwieters, S. C. Meredith, and R. Tycko. Molecular Structure of β-Amyloid Fibrils in Alzheimer's Disease Brain Tissue. *Cell*, 154(6):1257–1268, Sept. 2013.

[119] M. Lund and B. Jönsson. A Mesoscopic Model for Protein-Protein Interactions in Solution. *Biophysical Journal*, 85(5):2940–2947, Nov. 2003.

[120] D. Lusher, J. Koskinen, and G. Robins, editors. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge, 2012.

[121] B. B. Majumdar, V. Prytkova, E. K. Wong, J. A. Freites, D. J. Tobias, and M. Heyden. Role of Conformational Flexibility in Monte Carlo Simulations of Many-Protein Systems. *J. Chem. Theory Comput.*, 15(2):1399–1408, Feb. 2019.

[122] S. K. Maloney, G. N. Bronner, and R. Buffenstein. Thermoregulation in the Angolan Free-Tailed Bat Mops condylurus: A Small Mammal That Uses Hot Roosts. *Physiol. Biochem. Zool.*, 72(4):385–396, July 1999.

[123] B. Mänz, M. Schwemmle, and L. Brunotte. Adaptation of Avian Influenza A Virus Polymerase in Mammals To Overcome the Host Species Barrier. *J. Virol.*, 87(13):7200–7209, July 2013.

[124] P. Massin, S. van der Werf, and N. Naffakh. Residue 627 of PB2 Is a Determinant of Cold Sensitivity in RNA Replication of Avian Influenza Viruses. *J. Virol.*, 75(11):5398–5404, June 2001.

[125] G. M. Mathew, A. Madhavan, K. B. Arun, R. Sindhu, P. Binod, R. R. Singhania, R. K. Sukumaran, and A. Pandey. Thermophilic Chitinases: Structural, Functional and Engineering Attributes for Industrial Applications. *Appl Biochem Biotechnol*, 193(1):142–164, Jan. 2021.

[126] C. J. E. Metcalf, R. B. Birger, S. Funk, R. D. Kouyos, J. O. Lloyd-Smith, and V. A. A. Jansen. Five challenges in evolution and infectious diseases. *Epidemics*, 10:40–44, Mar. 2015.

[127] B. Meyer, J. Chiaravalli, S. Gellenoncourt, P. Brownridge, D. P. Bryne, L. A. Daly, A. Grauslys, M. Walter, F. Agou, L. A. Chakrabarti, C. S. Craik, C. E. Eyers, P. A. Eyers, Y. Gambin, A. R. Jones, E. Sierecki, E. Verdin, M. Vignuzzi, and E. Emmott. Characterising proteolysis during SARS-CoV-2 infection identifies viral cleavage sites and cellular targets with therapeutic potential. *Nat Commun*, 12(1):5553, Sept. 2021.

[128] M. Meyer-Luehmann, J. Coomaraswamy, T. Bolmont, S. Kaeser, C. Schaefer, E. Kilger, A. Neuenschwander, D. Abramowski, P. Frey, A. L. Jaton, J.-M. Vigouret, P. Paganetti, D. M. Walsh, P. M. Mathews, J. Ghiso, M. Staufenbiel, L. C. Walker, and M. Jucker. Exogenous Induction of Cerebral ssAmyloidogenesis Is Governedd byAgentt andHost. *Science*, 313(5794):1781–1784, Sept. 2006.

[129] I. A. Mills-Henry, S. L. Thol, M. S. Kosinski-Collins, E. Serebryany, and J. A. King. Kinetic stability of long-lived human lens γ-crystallins and their isolated double Greek key domains. *Biophys. J.*, 117:269–280, 2019.

[130] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.*, 4(5):819–834, May 2008.

[131] B. Morel, P. Barbera, L. Czech, B. Bettisworth, L. Hübner, S. Lutteropp, D. Serdari, E.-G. Kostaki, I. Mamais, A. M. Kozlov, P. Pavlidis, D. Paraskevis, and A. Stamatakis. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology and Evolution*, 38(5):1777–1791, May 2021.

[132] A. Morriss-Andrews and J.-E. Shea. Simulations of Protein Aggregation: Insights from Atomistic and Coarse-Grained Models. *J. Phys. Chem. Lett.*, 5(11):1899–1908, June 2014.

[133] A. Morriss-Andrews and J.-E. Shea. Computational Studies of Protein Aggregation: Methods and Applications. *Annu. Rev. Phys. Chem.*, 66:643–666, Feb. 2015.

[134] J. A. Mótyán, M. Mahdi, G. Hoffka, and J. Tőzsér. Potential Resistance of SARS-CoV-2 Main Protease (Mpro) against Protease Inhibitors: Lessons Learned from HIV-1 Protease. *Int. J. Mol. Sci.*, 23(7):3507, Jan. 2022.

[135] M. Moustaqil, E. Ollivier, H.-P. Chiu, S. Van Tol, P. Rudolffi-Soto, C. Stevens, A. Bhumkar, D. J. B. Hunter, A. N. Freiberg, D. Jacques, B. Lee, E. Sierecki, and Y. Gambin. SARS-CoV-2 proteases PLpro and 3CLpro cleave IRF3 and critical modulators of inflammatory pathways (NLRP12 and TAB1): Implications for disease presentation across species. *Emerg. Microbes Infect.*, 10(1):178–195, Jan. 2021.

[136] R. Ni, S. Abeln, M. Schor, M. A. Cohen Stuart, and P. G. Bolhuis. Interplay between Folding and Assembly of Fibril-Forming Polypeptides. *Phys. Rev. Lett.*, 111(5):058101, July 2013.

[137] C. Nilsberth, A. Westlind-Danielsson, C. B. Eckman, M. M. Condron, K. Axelman, C. Forsell, C. Stenh, J. Luthman, D. B. Teplow, S. G. Younkin, J. Näslund, and L. Lannfelt. The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation. *Nat Neurosci*, 4(9):887–893, Sept. 2001.

[138] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.*, 139(9):090901, Sept. 2013.

[139] H. Ohtaka, A. Schön, and E. Freire. Multidrug Resistance to HIV-1 Protease Inhibition Requires Cooperative Coupling between Distal Mutations. *Biochemistry*, 42(46):13659–13666, Nov. 2003.

[140] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.*, 7(2):525–537, Feb. 2011.

[141] T. T. O'Malley, W. M. I. Witbold, S. Linse, and D. M. Walsh. The Aggregation Paths and Products of Aβ42 Dimers Are Distinct from Those of the Aβ42 Monomer. *Biochemistry*, 55(44):6150–6161, Nov. 2016.

[142] R. Paparcone, S. W. Cranford, and M. J. Buehler. Self-folding and aggregation of amyloid nanofibrils. *Nanoscale*, 3(4):1748–1755, Apr. 2011.

[143] E. Paradis and K. Schliep. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, Feb. 2019.

[144] P. Pattison and G. Robins. Neighborhood-Based Models for Social Networks. *Sociol. Methodol.*, 32(1):301–337, Aug. 2002.

[145] A. T. Petkova, R. D. Leapman, Z. Guo, W.-M. Yau, M. P. Mattson, and R. Tycko. Self-Propagating, Molecular-Level Polymorphism in Alzheimer's ß-Amyloid Fibrils. *Science*, 307(5707):262–265, Jan. 2005.

[146] T. M. Phan and J. D. Schmit. Conformational entropy limits the transition from nucleation to elongation in amyloid aggregation. *Biophysical Journal*, 121(15):2931–2939, Aug. 2022.

[147] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26(16):1781–1802, 2005.

[148] R. Prabakaran, P. Rawat, A. M. Thangakani, S. Kumar, and M. M. Gromiha. Protein aggregation: In silico algorithms and applications. *Biophys Rev*, 13(1):71–89, Feb. 2021.

[149] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26(7):1641–1650, July 2009.

[150] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, 5(3):e9490, Mar. 2010.

[151] V. Prytkova, M. Heyden, D. Khago, J. A. Freites, C. T. Butts, R. W. Martin, and D. J. Tobias. Multi-Conformation Monte Carlo: A Method for Introducing Flexibility in Efficient Simulations of Many-Protein Systems. *J. Phys. Chem. B*, 120(33):8115–8126, Aug. 2016.

[152] K. Przygońska, M. Pacewicz, W. Sadowska, J. Poznański, W. Bal, and M. Dadlez. His6, His13, and His14 residues in Aβ 1–40 peptide significantly and specifically affect oligomeric equilibria. *Sci Rep*, 9(1):9449, July 2019.

[153] W. Qiang, W.-M. Yau, Y. Luo, M. P. Mattson, and R. Tycko. Antiparallel β-sheet architecture in Iowa-mutant β-amyloid fibrils. *Proc. Natl. Acad. Sci.*, 109(12):4443–4448, Mar. 2012.

[154] Y. Qiu and K. Xu. Functional studies of the coronavirus nonstructural proteins. *STEMedicine*, 1(2):e39–e39, Mar. 2020.

[155] M. K. Quinn, N. Gnan, S. James, A. Ninarello, F. Sciortino, E. Zaccarelli, and J. J. McManus. How fluorescent labelling alters the solution behaviour of proteins. *Phys. Chem. Chem. Phys.*, 17(46):31177–31187, Nov. 2015.

[156] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[157] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022.

[158] M. J. Ramos Pereira, T. Stefanski Chaves, P. E. Bobrowiec, and G. Selbach Hofmann. How aerial insectivore bats of different sizes respond to nightly temperature shifts. *Int J Biometeorol*, 66(3):601–612, Mar. 2022.

[159] D. Reith, M. Pütz, and F. Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*, 24(13):1624–1636, 2003.

[160] A. Rojas, N. Maisuradze, K. Kachlishvili, H. A. Scheraga, and G. G. Maisuradze. Elucidating Important Sites and the Mechanism for Amyloid Fibril Formation by Coarse-Grained Molecular Dynamics. *ACS Chem. Neurosci.*, 8(1):201–209, Jan. 2017.

[161] J. Santos, J. Pujols, I. Pallarès, V. Iglesias, and S. Ventura. Computational prediction of protein aggregation: Advances in proteomics, conformation-specific algorithms and biotechnological applications. *Computational and Structural Biotechnology Journal*, 18:1403–1413, Jan. 2020.

[162] K. Sasahara, H. Yagi, H. Naiki, and Y. Goto. Heat-Triggered Conversion of Protofibrils into Mature Amyloid Fibrils of $B2$-Microglobulin. *Biochemistry*, 46(11):3286–3293, Mar. 2007.

[163] R. Sathyapriya and S. Vishveshwara. Structure networks of E. coli glutaminyl-tRNA synthetase: Effects of ligand binding. *Proteins Struct. Funct. Bioinforma.*, 68(2):541–550, 2007.

[164] M. R. Sawaya, S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. W. Wiltzius, H. T. McFarlane, A. Ø. Madsen, C. Riekel, and D. Eisenberg. Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature*, 447(7143):453–457, May 2007.

[165] E. Scalone, L. Broggini, C. Visentin, D. Erba, F. Bačić Toplek, K. Peqini, S. Pellegrino, S. Ricagno, C. Paissoni, and C. Camilloni. Multi-eGO: An in silico lens to look into protein aggregation kinetics at atomic resolution. *Proc. Natl. Acad. Sci.*, 119(26):e2203181119, June 2022.

[166] B. Schwarze, A. Korn, C. Höfling, U. Zeitschel, M. Krueger, S. Roßner, and D. Huster. Peptide backbone modifications of amyloid β (1–40) impact fibrillation behavior and neuronal toxicity. *Sci Rep*, 11(1):23767, Dec. 2021.

[167] M. Schweinberger. Instability, Sensitivity, and Degeneracy of Discrete Exponential Families. *J. Am. Stat. Assoc.*, 106(496):1361–1370, Dec. 2011.

[168] M. Schweinberger, P. N. Krivitsky, C. T. Butts, and J. R. Stewart. Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios. *Stat. Sci.*, 35(4):627–662, Nov. 2020.

[169] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.

[170] E. Serebryany and J. A. King. The $B\gamma$-crystallins: Native state stability and pathways to aggregation. *Progress in Biophysics and Molecular Biology*, 115(1):32–41, July 2014.

[171] E. Serebryany and J. A. King. Wild-type human γd-crystallin promotes aggregation of its oxidation-mimicking, misfolding-prone W42Q mutant. *The Journal of Biological Chemistry*, 290(18):11491–11503, 2015.

[172] E. Serebryany, J. C. Woodard, B. V. Adkar, M. Shabab, J. A. King, and E. I. Shakhnovich. An Internal Disulfide Locks a Misfolded Aggregation-prone Intermediate in Cataract-linked Mutants of Human γD-Crystallin*. *Journal of Biological Chemistry*, 291(36):19172–19183, Sept. 2016.

[173] M. Seuma, B. Lehner, and B. Bolognesi. An atlas of amyloid aggregation: The impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. *Nat Commun*, 13(1):7084, Nov. 2022.

[174] S. L. Shammas, T. P. J. Knowles, A. J. Baldwin, C. E. MacPhee, M. E. Welland, C. M. Dobson, and G. L. Devlin. Perturbation of the Stability of Amyloid Fibrils through Alteration of Electrostatic Interactions. *Biophysical Journal*, 100(11):2783–2791, June 2011.

[175] P. D. Shaw Stewart and J. L. Bach. Temperature dependent viral tropism: Understanding viral seasonality and pathogenicity as applied to the avoidance and treatment of endemic viral respiratory illnesses. *Rev. Med. Virol.*, 32(1):e2241, 2022.

[176] O. Sheik Amamuddy, G. M. Verkhivker, and Ö. Tastan Bishop. Impact of Early Pandemic Stage Mutations on Molecular Dynamics of SARS-CoV-2 Mpro. *J. Chem. Inf. Model.*, 60(10):5080–5102, Oct. 2020.

[177] M. S. Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics*, 129(14):144108, Oct. 2008.

[178] Y. Shen, F. Gao, M. Wang, and A. Li. RPdb: A database of experimentally verified cellular reprogramming records. *Bioinformatics*, 31(19):3237–3239, Oct. 2015.

[179] J. Shi, Z. Wei, and J. Song. Dissection Study on the Severe Acute Respiratory Syndrome 3C-like Protease Reveals the Critical Role of the Extra Domain in Dimerization of the Enzyme: DEFINING THE EXTRA DOMAIN AS A NEW TARGET FOR DESIGN OF HIGHLY SPECIFIC PROTEASE INHIBITORS *. *Journal of Biological Chemistry*, 279(23):24765–24773, June 2004.

[180] A. B. Siemer. What makes functional amyloids work? *Crit. Rev. Biochem. Mol. Biol.*, 57(4):399–411, July 2022.

[181] K. H. Skåra, C. Bech, M. A. Fjelldal, J. van der Kooij, R. Sørås, and C. Stawski. Energetics of whiskered bats in comparison to other bats of the family Vespertilionidae. *Biology Open*, 10(8):bio058640, July 2021.

[182] E. C. Smith and M. R. Denison. Implications of altered replication fidelity on the evolution and pathogenesis of coronaviruses. *Current Opinion in Virology*, 2(5):519–524, Oct. 2012.

[183] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New Specifications for Exponential Random Graph Models. *Sociol. Methodol.*, 36(1):99–153, Aug. 2006.

[184] Z. Song, Y. Xu, L. Bao, L. Zhang, P. Yu, Y. Qu, H. Zhu, W. Zhao, Y. Han, and C. Qin. From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses*, 11(1):59, Jan. 2019.

[185] C. Soto and S. Pritzkow. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nat Neurosci*, 21(10):1332–1340, Oct. 2018.

[186] C. J. Stewart, G. I. Olgenblum, A. Propst, D. Harries, and G. J. Pielak. Resolving the enthalpy of protein stabilization by macromolecular crowding. *Protein Sci.*, 32(3):e4573, 2023.

[187] D. Strauss. On a General Class of Models for Interaction. *SIAM Rev.*, 28(4):513–527, Dec. 1986.

[188] A. Suka, H. Oki, Y. Kato, K. Kawahara, T. Ohkubo, T. Maruno, Y. Kobayashi, S. Fujii, S. Wakai, L. Lisdiana, and Y. Sambongi. Stability of cytochromes c′ from psychrophilic and piezophilic Shewanella species: Implications for complex multiple adaptation to low temperature and high hydrostatic pressure. *Extremophiles*, 23(2):239–248, Mar. 2019.

[189] S. Temmam, K. Vongphayloth, E. Baquero, S. Munier, M. Bonomi, B. Regnault, B. Douangboubpha, Y. Karami, D. Chrétien, D. Sanamxay, V. Xayaphet, P. Paphaphanh, V. Lacoste, S. Somlor, K. Lakeomany, N. Phommavanh, P. Pérot, O. Dehan, F. Amara, F. Donati, T. Bigot, M. Nilges, F. A. Rey, S. van der Werf, P. T. Brey, and M. Eloit. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature*, 604(7905):330–336, Apr. 2022.

[190] C. Tempra, F. Scollo, M. Pannuzzo, F. Lolicato, and C. La Rosa. A unifying framework for amyloid-mediated membrane damage: The lipid-chaperone hypothesis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1870(4):140767, Apr. 2022.

[191] J. Torrent, D. Martin, S. Noinville, Y. Yin, M. Doumic, M. Moudjou, V. Béringue, and H. Rezaei. Pressure Reveals Unique Conformational Features in Prion Protein Fibril Diversity. *Sci Rep*, 9(1):2802, Feb. 2019.

[192] I. F. Tsigelny, Y. Sharikov, V. L. Kouznetsova, J. P. Greenberg, W. Wrasidlo, T. Gonzalez, P. Desplats, S. E. Michael, M. Trejo-Morales, C. R. Overk, and E. Masliah. Structural Diversity of Alzheimer's Disease Amyloid-β Dimers and Their Role in Oligomerization and Fibril Formation. *J. Alzheimers Dis.*, 39(3):583–600, Jan. 2014.

[193] R. Tycko. Physical and structural basis for polymorphism in amyloid fibrils. *Protein Sci.*, 23(11):1528–1539, 2014.

[194] M. H. Unhelkar, V. T. Duong, K. N. Enendu, J. E. Kelly, S. Tahir, C. T. Butts, and R. W. Martin. Structure prediction and network analysis of chitinases from the Cape sundew, Drosera capensis. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1861(3):636–643, Mar. 2017.

[195] R. Vácha and D. Frenkel. Relation between Molecular Shape and the Morphology of Self-Assembling Aggregates: A Simulation Study. *Biophysical Journal*, 101(6):1432–1439, Sept. 2011.

[196] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[197] P. V'kovski, M. Gultom, J. N. Kelly, S. Steiner, J. Russeil, B. Mangeat, E. Cora, J. Pezoldt, M. Holwerda, A. Kratzel, L. Laloli, M. Wider, J. Portmann, T. Tran, N. Ebert, H. Stalder, R. Hartmann, V. Gardeux, D. Alpern, B. Deplancke, V. Thiel, and R. Dijkman. Disparate temperature-dependent virus–host dynamics for SARS-CoV-2 and SARS-CoV in the human respiratory epithelium. *PLOS Biology*, 19(3):e3001158, Mar. 2021.

[198] B. Wang, C. Yu, Y.-B. Xi, H.-C. Cai, J. Wang, S. Zhou, S. Zhou, Y. Wu, Y.-B. Yan, X. Ma, and L. Xie. A novel CRYGD mutation (p.Trp43Arg) causing autosomal dominant congenital cataract in a Chinese family. *Hum. Mutat.*, 32(1):E1939–E1947, 2011.

[199] F. Wang, C. Chen, W. Tan, K. Yang, and H. Yang. Structure of Main Protease from Human Coronavirus NL63: Insights for Wide Spectrum Anti-Coronavirus Drug Design. *Sci Rep*, 6(1):22677, Mar. 2016.

[200] H. Wang, L. Duo, F. Hsu, C. Xue, Y. K. Lee, and Z. Guo. Polymorphic Aβ42 fibrils adopt similar secondary structure but differ in cross-strand side chain stacking interactions within the same β-sheet. *Sci Rep*, 10(1):5720, Mar. 2020.

[201] L.-G. Wang, T. T.-Y. Lam, S. Xu, Z. Dai, L. Zhou, T. Feng, P. Guo, C. W. Dunn, B. R. Jones, T. Bradley, H. Zhu, Y. Guan, Y. Jiang, and G. Yu. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution*, 37(2):599–603, Feb. 2020.

[202] B. Webb and A. Sali. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.*, 54(1):5.6.1–5.6.37, 2016.

[203] G. Wei, A. I. Jewett, and J.-E. Shea. Structural diversity of dimers of the Alzheimer amyloid- β (25–35) peptide and polymorphism of the resulting fibrils. *Phys. Chem. Chem. Phys.*, 12(14):3622–3629, 2010.

[204] H. Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[205] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the Tidyverse. *J. Open Source Softw.*, 4(43):1686, Nov. 2019.

[206] H. Wickham, L. Henry, T. L. Pedersen, T. J. Luciani, M. Decorde, and V. Lise. Svglite: An 'SVG' Graphics Device, 2022.

[207] A. W. Wilber, J. P. K. Doye, A. A. Louis, E. G. Noya, M. A. Miller, and P. Wong. Reversible self-assembly of patchy particles into monodisperse icosahedral clusters. *J. Chem. Phys.*, 127(8):085106, Aug. 2007.

[208] B. Winner, R. Jappelli, S. K. Maji, P. A. Desplats, L. Boyer, S. Aigner, C. Hetzer, T. Loher, M. Vilar, S. Campioni, C. Tzitzilonis, A. Soragni, S. Jessberger, H. Mira, A. Consiglio, E. Pham, E. Masliah, F. H. Gage, and R. Riek. In vivo demonstration that α-synuclein oligomers are toxic. *Proc. Natl. Acad. Sci.*, 108(10):4194–4199, Mar. 2011.

[209] E. K. Wong, V. Prytkova, J. A. Freites, C. T. Butts, and D. J. Tobias. Molecular Mechanism of Aggregation of the Cataract-Related γD-Crystallin W42R Variant from Multiscale Atomistic Simulations. *Biochemistry*, 58(35):3691–3699, Sept. 2019.

[210] C. Wu and J.-E. Shea. Coarse-grained models for protein aggregation. *Current Opinion in Structural Biology*, 21(2):209–220, Apr. 2011.

[211] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, Mar. 2020.

[212] Y. Xiao, B. Ma, D. McElheny, S. Parthasarathy, F. Long, M. Hoshi, R. Nussinov, and Y. Ishii. Aβ(1–42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nat Struct Mol Biol*, 22(6):499–505, June 2015.

[213] W. Xie, L. A. Nangle, W. Zhang, P. Schimmel, and X.-L. Yang. Long-range structural effects of a Charcot–Marie–Tooth disease-causing mutation in human glycyl-tRNA synthetase. *Proc. Natl. Acad. Sci.*, 104(24):9976–9981, June 2007.

[214] S. Xu, Z. Dai, P. Guo, X. Fu, S. Liu, L. Zhou, W. Tang, T. Feng, M. Chen, L. Zhan, T. Wu, E. Hu, Y. Jiang, X. Bo, and G. Yu. ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. *Molecular Biology and Evolution*, 38(9):4039–4042, Sept. 2021.

[215] S. Yan and G. Wu. Potential 3-chymotrypsin-like cysteine protease cleavage sites in the coronavirus polyproteins pp1a and pp1ab and their possible relevance to COVID-19 vaccine and drug development. *FASEB J.*, 35(5):e21573, 2021.

[216] Y. Yang, D. Arseni, W. Zhang, M. Huang, S. Lövestam, M. Schweighauser, A. Kotecha, A. G. Murzin, S. Y. Peak-Chew, J. Macdonald, I. Lavenir, H. J. Garringer, E. Gelpi, K. L. Newell, G. G. Kovacs, R. Vidal, B. Ghetti, B. Ryskeldi-Falcon, S. H. W. Scheres, and M. Goedert. Cryo-EM structures of amyloid-β 42 filaments from human brains. *Science*, 375(6577):167–172, Jan. 2022.

[217] Ö. N. Yaveroğlu, S. M. Fitzhugh, M. Kurant, A. Markopoulou, C. T. Butts, and N. Pržulj. Ergm.graphlets: A Package for ERG Modeling Based on Graphlet Statistics. *J. Stat. Softw.*, 65(1):1–29, June 2015.

[218] F. Yin and C. T. Butts. Highly Scalable Maximum Likelihood and Conjugate Bayesian Inference for ERGMs on Graph Sets with Equivalent Vertices. *PLOS One*, Aug. 2022.

[219] G. Yoon, J. Kwak, J. I. Kim, S. Na, and K. Eom. Mechanical Characterization of Amyloid Fibrils Using Coarse-Grained Normal Mode Analysis. *Adv. Funct. Mater.*, 21(18):3454–3463, 2011.

[220] G. Yu. Aplot: Decorate a 'ggplot' with Associated Information, Apr. 2022.

[221] Y. Yu, G. Grazioli, N. E. Phillips, and C. T. Butts. Local Graph Stability in Exponential Family Random Graph Models. *SIAM J. Appl. Math.*, pages 1389–1415, Jan. 2021.

[222] Y. Yu, G. Grazioli, M. H. Unhelkar, R. W. Martin, and C. T. Butts. Network Hamiltonian models reveal pathways to amyloid fibril formation. *Sci. Rep.*, 10(1):15668, Sept. 2020.

[223] A. A. H. Zanjani, N. P. Reynolds, A. Zhang, T. Schilling, R. Mezzenga, and J. T. Berryman. Kinetic Control of Parallel versus Antiparallel Amyloid Aggregation via Shape of the Growing Aggregate. *Sci Rep*, 9(1):15987, Nov. 2019.

[224] J. Zhang and M. Muthukumar. Simulations of nucleation and elongation of amyloid fibrils. *The Journal of Chemical Physics*, 130(3):035102, Jan. 2009.

[225] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, and R. Hilgenfeld. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, 368(6489):409–412, Apr. 2020.

[226] Z. Zhang and S. C. Glotzer. Self-Assembly of Patchy Particles. *Nano Lett.*, 4(8):1407–1413, Aug. 2004.

[227] W. Zheng, M.-Y. Tsai, M. Chen, and P. G. Wolynes. Exploring the aggregation free energy landscape of the amyloid-β protein (1–40). *Proc. Natl. Acad. Sci.*, 113(42):11835–11840, Oct. 2016.

[228] M. Ziaunys, A. Sakalauskas, K. Mikalauskaite, and V. Smirnovas. Polymorphism of Alpha-Synuclein Amyloid Fibrils Depends on Ionic Strength and Protein Concentration. *Int. J. Mol. Sci.*, 22(22):12382, Jan. 2021.

[229] M. Ziaunys, T. Sneideris, and V. Smirnovas. Formation of distinct prion protein amyloid fibrils under identical experimental conditions. *Sci Rep*, 10(1):4572, Mar. 2020.

[230] J. Ziebuhr. The coronavirus replicase. *Current Topics in Microbiology and Immunology*, 287:57–94, 2005.

Appendix A

Supporting Information for Chapter 2

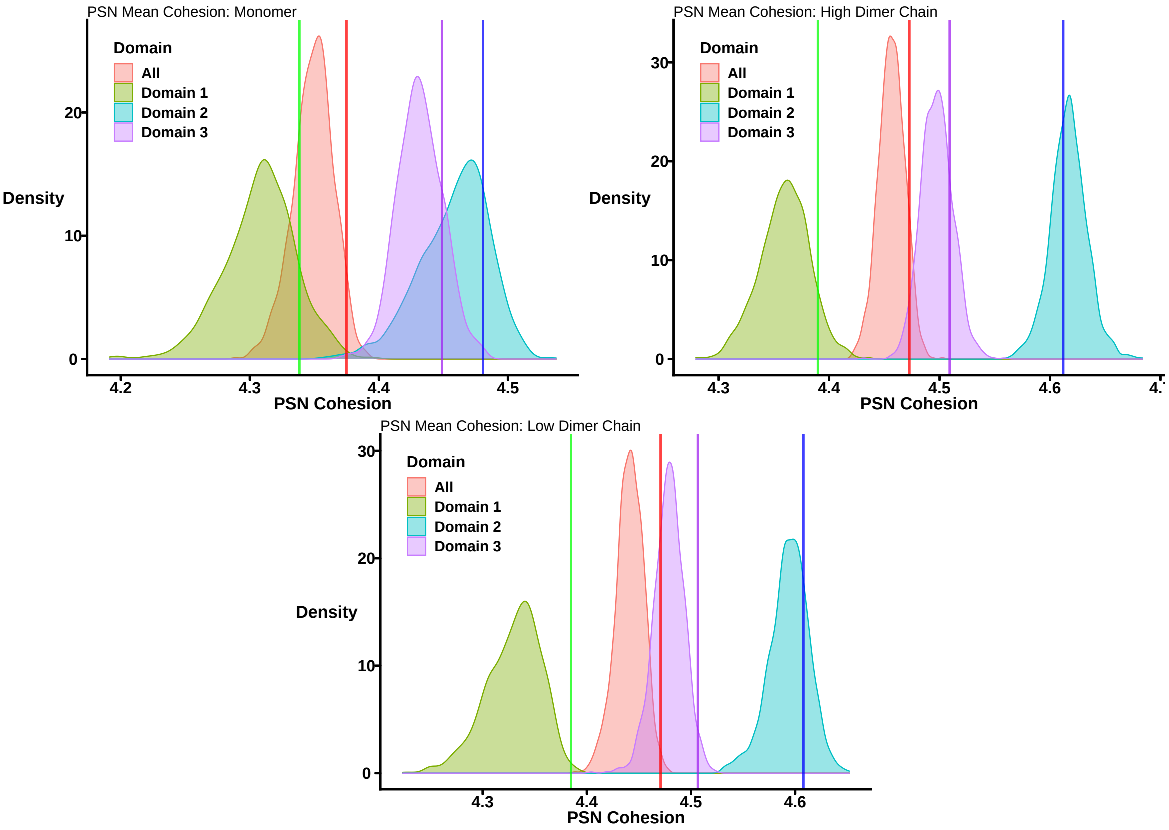


Figure A.1: Kernel density estimates of cohesion scores within each chain, pooled across variants; vertical lines indicate grand means.

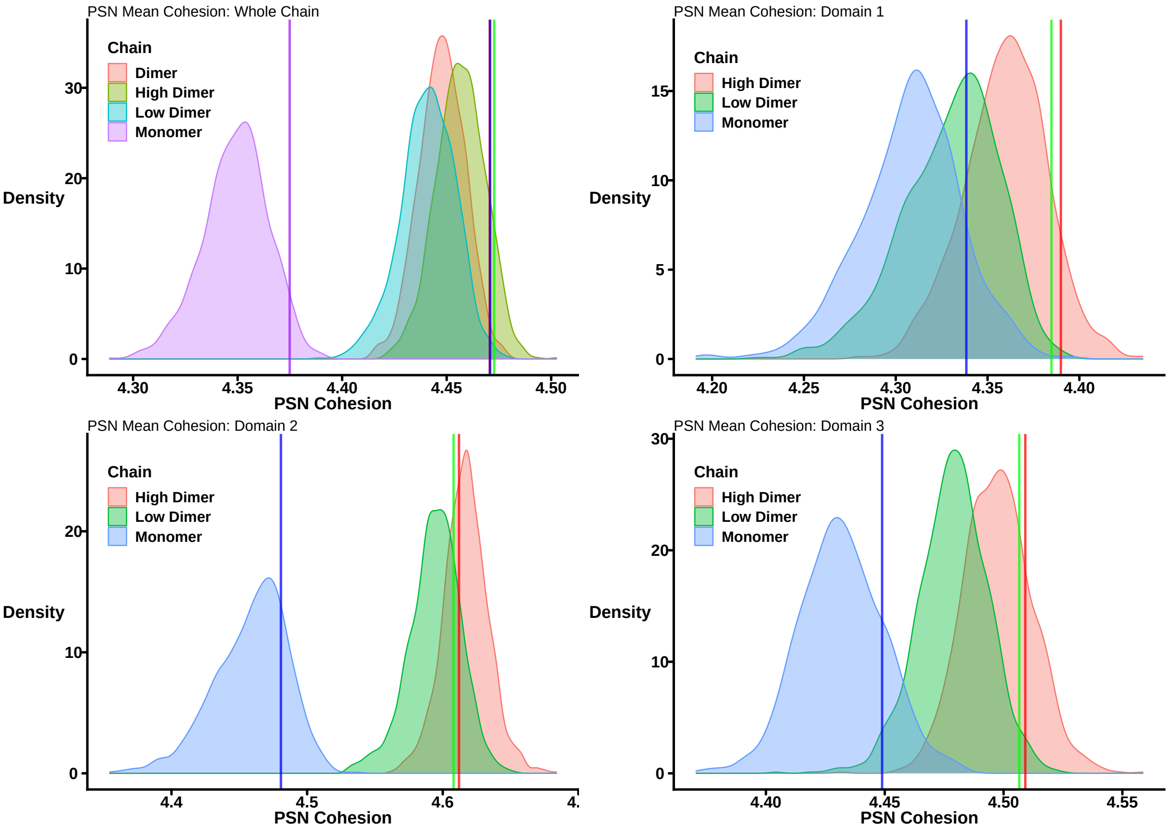


Figure A.2: Kernel density estimates of cohesion scores over the whole chain, and within each domain, pooled across variants; vertical lines indicate grand means.

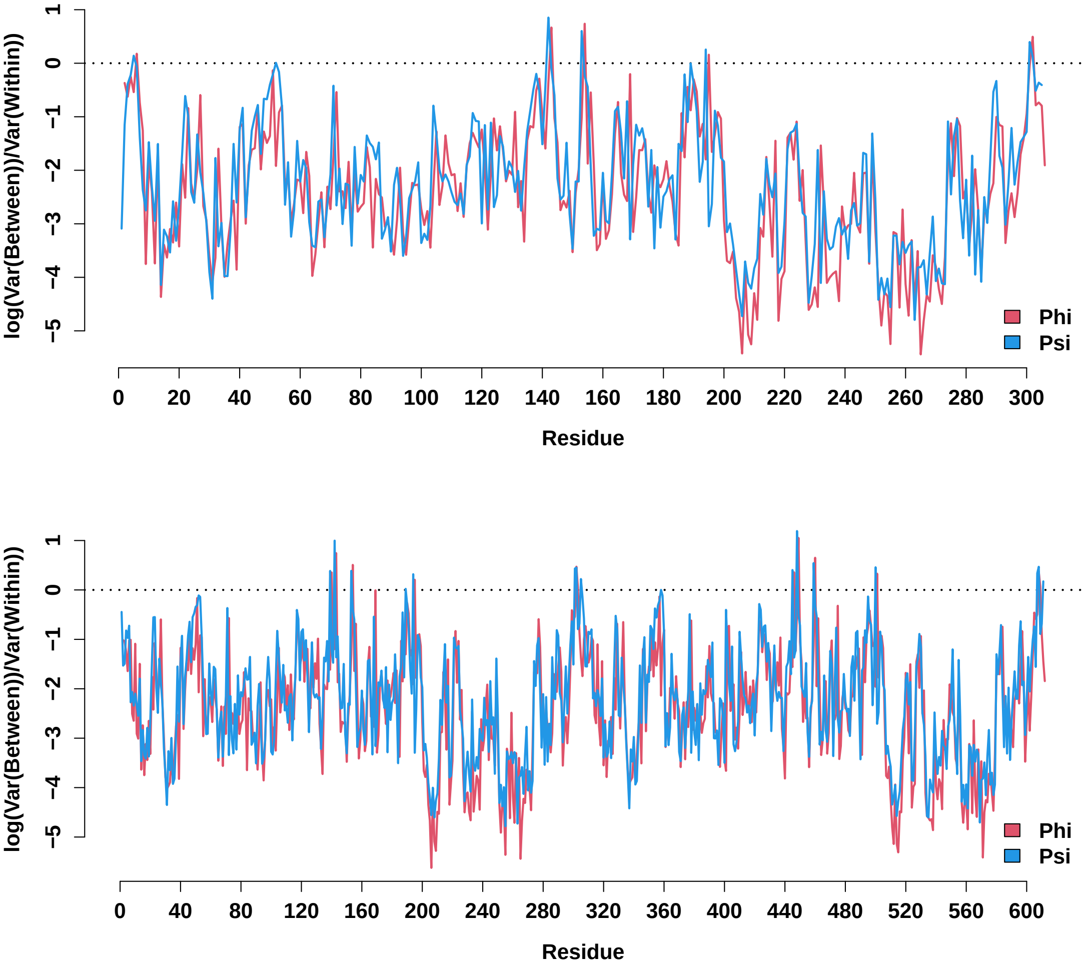


Figure A.3: Log ratio of torsion between variants over the torsion within variants, for each angle phi and psi; free monomer values shown in top panel, dimer values shown in bottom panel. Values were calculated by finding the angles between atoms in the DCD trajectory frames, in radians, then taking the angular mean and angular variance over each trajectory. Variance within-chain was then estimated by taking the mean of the trajectory variances for each variant sequence, and variance between-chains was estimated by taking the angular variance of the trajectory means for each variant sequence. Higher values indicate greater between-variant differences in mean angle, net of within-trajectory (dynamic) variation.

	Mean Difference	Std.Err	t value	Pr(>t)
Polar	0.04	0.01	3.19	0.0014
HydropathyKD	0.59	0.06	10.46	0.0000
Charge	0.02	0.01	2.16	0.0309
Aromatic	0.18	0.01	17.49	0.0000
Mass	11.12	0.66	16.76	0.0000
Volume	8.80	0.64	13.85	0.0000
Bulk	-0.01	0.00	-4.17	0.0000

Table A.1: Mean differences in amino acid side chain physical properties, for substituted residues. Substitution favors larger, more massive, and more hydrophobic residues.

	Fraction of conserved (#AA/37)	Fraction of all (#AA/306)	Fraction of AA type conserved (#AA conserved/#AA total)
Tyr	0.135	0.036	0.455
Phe	0.135	0.056	0.294
His	0.054	0.023	0.286
Cys	0.081	0.039	0.25
Asp	0.108	0.056	0.235
Glu	0.054	0.029	0.222
Gly	0.135	0.085	0.192
Asn	0.081	0.069	0.143
Gln	0.054	0.046	0.143
Leu	0.081	0.095	0.103
Pro	0.027	0.042	0.077
Ser	0.027	0.052	0.063
Ala	0.027	0.056	0.059

Table A.2: Distribution of conserved residues by amino acid type. Non-mutated residues: 2 11 14 16 28 29 39 41 44 54 66 79 115 118 126 133 140 144 145 150 154 172 176 182 183 185 187 192 203 211 268 286 289 290 291 295 299

Appendix B

Supporting Information for Chapter 3

B.1 Reproducibility Details

B.1.1 Atomistic Simulation

Input aggregate networks were generated by Wong, et al,[209] from atomistic simulations of equilibrium distributions of WT and W42R γ-Dc using the multi-conformation Monte Carlo (mcMC) methods described in Prytkova, et al[151].

MD simulations: Conformation libraries for mcMC steps were obtained from explicit solvent MD simulations under the CHARMM36 forcefield [23] in TIP3P water [98], as detailed in Wong, et al[209].

Parameters of mcMC simulations:

- No. of γ-Dc monomers (for both WT and W42R): $N = 375$ proteins

- Temperature: 310K
- Pressure: 200g/K
- periodic boundary conditions

Output of mcMC simulations:

- 14,000 frames WT equilibrium distributions
- 16,000 frames W42R equilibrium distributions

Further details needed for reproducing the original atomistic and mcMC simulation study can be found in Wong, et al[209].

B.1.2 Network Generation

Aggregation graph definitions by Wong, et al[209]:

- **vertex:** single protein monomer
- **edge:** occurs if two monomers 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. [209], figure S3.)

Network visualization and analysis software:

- R statistical computing system, version 4.2.0 [157]

- **Libraries:** `statnet` [83], `network` [29], `sna` [30], `ergm` [91], `ergm.userterms` [89], `parallel` [], `Rcpp`[]

B.1.3 Component/Aggregate Size Distribution Estimation

Component sizes: (computed with the `sna` library for all networks)

Distributions were estimated using a non-parametric Bayesian procedure:

Algorithm 1 Component size distribution estimation. Comparison with observed (atomistic) simulations are shown in Figure 4 of the main text.

```
procedure COMPONENTPOSTERIOR
    $nets \leftarrow \gamma$-Dc aggregation networks
    $jeffreys.prior \leftarrow 0.5$
    $n \leftarrow network.size(nets)$
    $comp.dist.obs \leftarrow \text{table}(\text{component.dist}(nets, \text{connected} = \text{“}weak\text{”})[[\text{“}csize\text{”}]], n)$
    $count \leftarrow \text{rowSums}(comp.dist.obs)$
    $posterior.mean = (count + jeffreys.prior)/\text{sum}(count + jeffreys.prior)$
    $posterior.param = count + jeffreys.prior$
```

L2 norm of the logged relative distribution [84] (our measure of discrepancy between distributions):

Algorithm 2 L2 norm of log relative size distribution for comparison of atomistic and network simulations.

```
procedure DISTRIBUTIONDISTANCES
    $obs.data \leftarrow \gamma$-Dc aggregation networks
    $sim.data \leftarrow$ componentPosterior distributions
    $comp.dist.obs \leftarrow \text{component.dist}(obs.data, \text{connected} = \text{“}weak\text{”})$
    $\text{sum}((\log((comp.dist.obs[[\text{“}cdist\text{”}]] + 0.5)/\text{sum}(comp.dist.obs[[\text{“}cdist\text{”}]] + 0.5)) - \log(sim.data[[\text{“}posterior.mean\text{”}]]))^2)$
```

B.1.4 Model Selection and Parameter Estimation

Models were fit by maximum likelihood estimation (MLE), using the pooling method of Yin and Butts [218]; estimation was performed using the `ergm` package [91], 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. The packages `sna` (version 2.6) and `ergm.components` (version 0.1) are also required.

Algorithm 3 Pooled ERGM MLE model fit. Resulting parameter coefficients are used in Algorithm 4 to assess aggregate size distributions of the estimated model.

```
procedure FITERGM
    $nets \leftarrow \gamma$-Dc aggregation networks
    $control \leftarrow$ control.ergm(main method = ”Stochastic Approximation”, MCMC.burnin = $5 * 10^4$, MCMC.interval = $2 * 10^4$, loglik = control.logLik.ergm(MCMC.burnin = 375, MCMC.interval = 375))
    if $gwesp(\alpha)$ then
        expand.grid($n$) for grid search on $\alpha$
    $f \leftarrow nets \sim edges + isolates + dimers + compsizesum + nsp(1) + gwesp(\alpha) + esp(1) + esp(2)$
    $fit \leftarrow$ ergmMSFit(formula = $f$, control = $control$)
```

Refer to Yin and Butts[218] for detailed description of the pooled MLE method being implemented in the function `ergmMSFit`. 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 [89].

B.1.5 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

Algorithm 4 Component size distribution simulation. Coefficients for the best fit for each model are given in Table 1 of the main text.

```
**procedure** componentSimulation
    $terms \leftarrow$ ERGM parameters
    $coef \leftarrow$ coef($fit$)
    $basenet \leftarrow$ observed$\gamma$-Dc aggregation nets
    $burnin \leftarrow 375^2$
    $thin \leftarrow 375^2$
    $control \leftarrow$ control.simulate.formula($burnin$, $thin$)
    $g \leftarrow$ simulate($basenet \sim terms$, $coef$, $control$)
    $draws \leftarrow 5000$
    $comp.dist.sim \leftarrow$ componentSimulation($terms$, $coef$, $basenet$, $draws$, $burnin$, $thin$)
    $i \leftarrow 2$
    $ncp \leftarrow 5$
    **while** $i =< ncp$ **do**
        $g \leftarrow$ simulate($g$[[length($g$)]] $\sim terms$, $coef$, $control$)
        $sim \leftarrow g$
        $i \leftarrow i + 1$
    $sim \leftarrow$ componentPosterior($sim$, $jeffreys.prior$)
    **return** $sim$
```

Algorithm 5 Component size distribution comparison by L2 norm. Results are displayed in Figure 4 of main text.

```
**procedure** componentCompare
    $opar \leftarrow comp.dist.obs$[["$posterior.param$"]]
    $spar \leftarrow comp.dist.sim$[["$posterior.param$"]]
    $metric \leftarrow$ "logL2"
    $tolerance \leftarrow 1 * 10^{-5}$
    $tot.est \leftarrow 0$
    **for** $i$ to length($opar$) **do**
        $reps = 64$
        $dv \leftarrow$ MonteCarloQuadratureFunction($opar$[[$i$]], $spar$[[$i$]], $reps$, $metric$)
        $mean.est \leftarrow dv[1]/reps$
        $mean.square.est \leftarrow dv[2]/reps$
        $sd.est \leftarrow ((mean.square.est - mean.est^2)/reps)^{1/2}$
        **while** ($reps < 1 * 10^6$) && (abs($sd.est/mean.est$) $>$ $tolerance$) **do**
            $dv \leftarrow$ MonteCarloQuadratureFunction($opar$[[$i$]], $spar$[[$i$]], $reps$)
            $mean.est \leftarrow (mean.est + dv[1]/reps)/2$
            $mean.square.est \leftarrow (mean.square.est + dv[2]/reps)/2$
            $sd.est \leftarrow ((mean.square.est - mean.est^2)/reps)^{1/2}$
        $tot.est \leftarrow tot.est + mean.est$
    **return** $tot.est$
```

Algorithm 6 Extrapolative simulations for large system sizes. Results are displayed in Figure 5 of main text.

```
procedure EXTRAPOLATIVESIM
    $mod \leftarrow$ ERGM base model
    $draws \leftarrow 1000$
    $n \leftarrow 375$
    $temp \leftarrow 310$
    $conc \leftarrow 200$
    $target \leftarrow \text{list(“compdist”, “graphs”, “stats”)}$
    $co \leftarrow \text{coef}(mod)$
    $co[-1] \leftarrow co[-1] * 310/temp$
    $co[1] \leftarrow 310/temp * (co[1] + 1 + \log(375)) - 1 - \log(n) + \log(conc/200)$
    $kB \leftarrow 1.987204259 * 10^{-3}$
    $phi \leftarrow (-kB * 310) * co$
    $phi[1] \leftarrow (-kB * 310) * (co[1] + 1 + log(375))$
    $net \leftarrow \text{network(rgraph}(n, \text{tp} = (mod[[\text{“target.stats”}]][1] * 2/375)/(n-1), \text{mode} = \text{“graph”}), \text{directed} = \texttt{FALSE})$
    $sim \leftarrow \text{simulate(“net} \sim mod\text{”}, co, draws, control)$
    return $sim$
```

and mixing properties over a wide size range (with mixing improving with size). Component size distributions and other metrics for the extrapolated network simulations were computed as described for the other simulations.

Appendix C

Supporting Information for Chapter 4

C.1 Fibril Classification

Classification of 1-ribbons, 2-ribbons, and 4-cycle oligomers was performed using the `ergm.graphlets` package along with rule-based methods in `R` [217]. The graphlet orbits determine the number of vertices adjacent to a given vertex belonging to specific local automorphism orbits, allowing that focal vertex to be classified by the structure of its surrounding vertices. For the structures encountered in the simulations for this paper, the following classification rules were employed. For each simulation frame, each vertex was classified into aggregation type as follows:

1-ribbons: the focal vertex ((is adjacent to one other vertex AND belongs to an open 2-path) OR (is adjacent to two other vertices AND is the center vertex of an open 2-path)) AND (belongs to a component containing at least 4 serially adjacent vertices satisfying the former criteria).

2-ribbons: the focal vertex ((is adjacent to two other vertices AND is the center of an open

2-path AND belongs to a chordless 4-cycle) OR (is adjacent to three other vertices AND is the center of one open 3-star AND belongs to two chordless 4-cycles)) AND (belongs to a component containing at least six vertices satisfying the former criterion)

4-cycles: the focal vertex belongs to a component of order 4 AND the focal vertex belongs to a chordless 4-cycle.

dimers: the focal vertex belongs to a component of order 2.

cubic oligomers: the focal vertex belongs to a component of order 8 in which (all vertices are automorphically equivalent AND all vertices have degree 3 AND the component contains no odd-length cycles AND all vertices belong to three 4-cycles AND all vertices belong to twelve 6-cycles).

unstructured aggregates ("gel"): the focal vertex belongs to a component containing at least 50% of the vertices in the network such that at least 90% of vertices within the component fail to meet any of the above classification criteria.

Yield for a given type is calculated from the fraction of vertices classified into the corresponding category. As described in the main text, much of the parameter space reliably produces phases that are dominated by a particular species; some regions (see e.g. Figure 5) produce mixed phases in which multiple aggregation states coexist. (See Figure S8 for additional examples.) We note that these classes do not exhaust the set of all aggregation states known to arise from NHMs, but were found to effectively summarize the set of states observed for the specific class of interactions studied here.

Table C.1: Edge-Dependent Phase Boundary Intercepts on Equation 3 ($\phi_{2s} + \phi_{nsp_1} = \phi_0$)

Phase Boundary	**Intercept**	**Edge Values**			**Rel. Val.**
ERGM notation	$\theta_e =$	75	100	125	
NHM notation	ϕ_e (kcal/mol) =	-50	-66	-81	
net edge energy equiv.	ϕ_e^* (kcal/mol) =	47	62	78	
Unstructured Aggregate	$\theta_0 =$	19	26	32	
	$\phi_0 =$	12	16	20	
	$=$	$\mathbf{0.25}\boldsymbol{\phi_e^*}$	$\mathbf{0.26}\boldsymbol{\phi_e^*}$	$\mathbf{0.26}\boldsymbol{\phi_e^*}$	$\approx \boldsymbol{\phi_e^*}/\mathbf{4}$
Pure 1-ribbon (lower)	$\theta_0 =$	22	28	40	
	$\phi_0 =$	14	17	25	
	$=$	$\mathbf{0.30}\boldsymbol{\phi_e^*}$	$\mathbf{0.28}\boldsymbol{\phi_e^*}$	$\mathbf{0.32}\boldsymbol{\phi_e^*}$	$\approx \boldsymbol{\phi_e^*}/\mathbf{3}$
Pure 1-ribbon (upper)	$\theta_0 =$	36	50	62	
	$\phi_0 =$	23	31	38	
	$=$	$\mathbf{0.49}\boldsymbol{\phi_e^*}$	$\mathbf{0.50}\boldsymbol{\phi_e^*}$	$\mathbf{0.49}\boldsymbol{\phi_e^*}$	$\approx \boldsymbol{\phi_e^*}/\mathbf{2}$
Pure Dimer	$\theta_0 =$	70	96	122	
	$\phi_0 =$	44	60	76	
	$=$	$\mathbf{0.94}\boldsymbol{\phi_e^*}$	$\mathbf{0.96}\boldsymbol{\phi_e^*}$	$\mathbf{0.98}\boldsymbol{\phi_e^*}$	$\approx \boldsymbol{\phi_e^*}$

Conversion between ERGM notation (θ) and NHM notation (ϕ) for an arbitrary term s can be performed as follows:

$$\phi_s = \begin{cases} -k_B T(\theta_s + 1 + \log N) & s = e \\ -k_B T \theta_s & s \neq e \end{cases}$$

$$\theta_s = \begin{cases} -\phi_s/(k_B T) - 1 - \log N & s = e \\ -\phi_s/(k_B T) & s \neq e \end{cases}$$

where T is the system temperature, N is the number of monomers, k_B is Boltzmann's constant, and e refers to the edge term. The additional offsets to the edge parameters reflect the effect of bond vibrations and entropic corrections for motional degrees of freedom, per [78]. (Note that this specification is in terms of the counting measure, with entropic corrections expressed as an offset to the edge parameter rather than as a separate $h(g)$

function. The two expressions are equivalent, with the form given here appropriate for use with e.g. `statnet` software [91].)

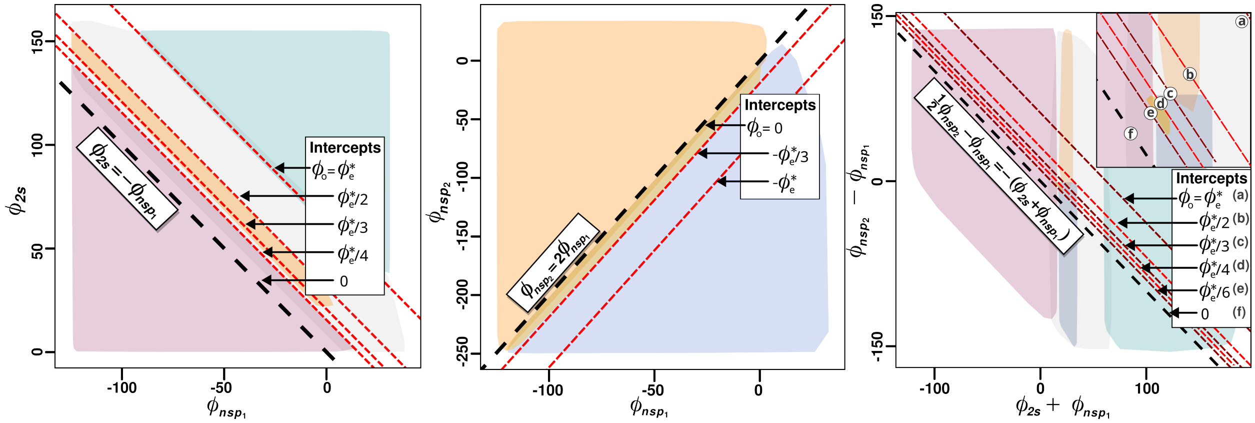


Figure C.1: Reproduction of Figure 2, showing boundaries of each phase marked by a dotted red line, with the corresponding unique intercept of each boundary shown in the legend and indicated by an arrow to the respective boundary. Intercepts should be interpreted as being added to the right-hand side of the respective equation indicated on each plot. I.e. plot (A) shows locations of boundaries with unique intercepts (denoted ϕ_0) on the equation $\phi_{2s} = -\phi_{nsp_1} + \phi_0$, plot (B) shows boundaries with unique intercepts on $\phi_{nsp_2} = 2\phi_{nsp_1} + \phi_0$, and plot (C) shows boundaries with unique intercepts on $\phi_{nsp_2}/2 - \phi_{nsp_1} = -\phi_{2s} - \phi_{nsp_1} + \phi_0$. As illustrated by these plots, the value of ϕ_0 for a given boundary is found to be equivalent to a simple fraction of ϕ_e^*, the *net edge equivalent energy*, which is an adjustment to the base ϕ_e value that is independent of spatial constraints and bond vibration energies, since those energy contributions have no effect on changes to structural calculations resulting from individual edge addition. As such, each ϕ_0 relates the location of the boundary in the given parameter space to the change in topology of the allowable structures that may be formed by the addition of an edge.

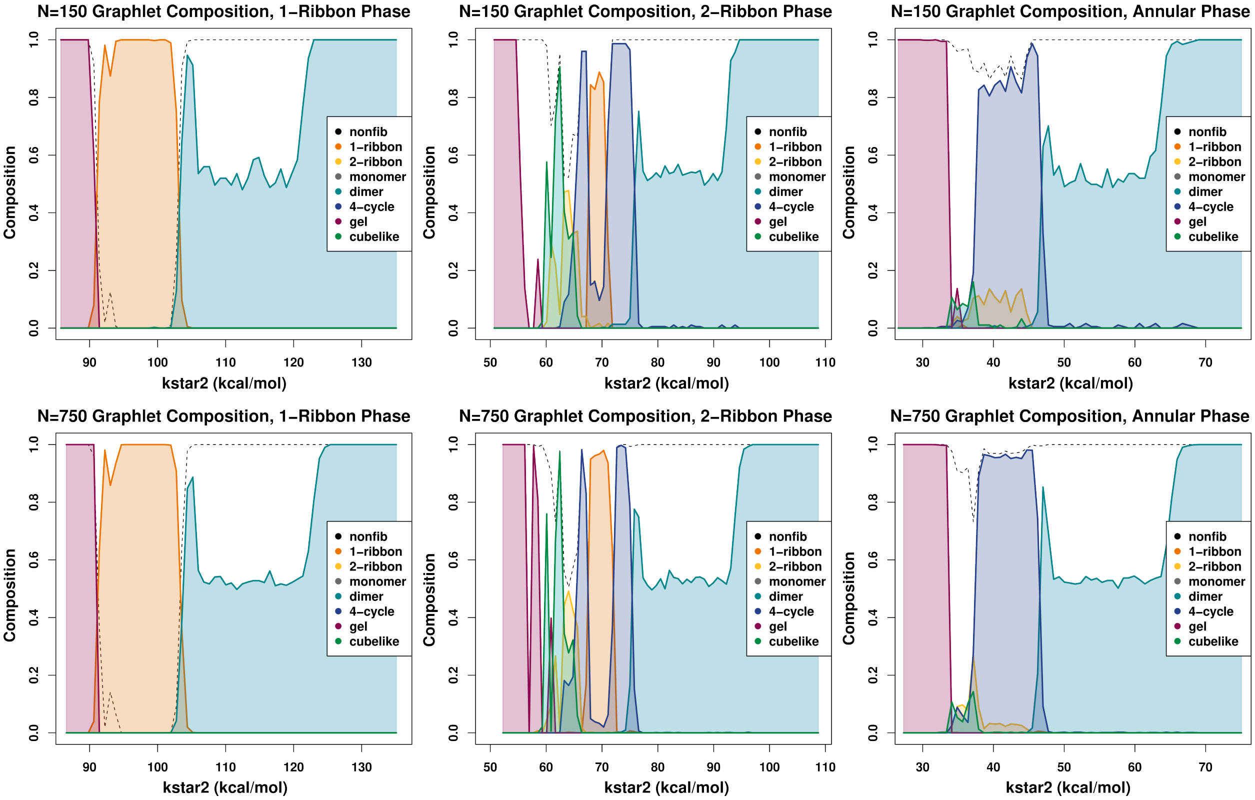


Figure C.2: Plots A-F show the composition of networks simulated from the sample trajectories depicted in Figure 4. In the first row, plots A-C correspond to networks simulated with 150 vertices, with plot A showing the trajectory going through the pure 1-ribbon phase boundaries, plot B showing the mixed 2-ribbon phase boundaries, and plot C showing the mixed 4-cycle and cubic oligomer phase boundaries. The second row, plots D-E, show the compositions of networks simulated with 750 vertices following the same sampling trajectories as the plots of the top row. Phase boundaries are observed to remain at fixed values of the parameter space, despite a five-fold increase in the system size of the simulated networks. Additionally, plots B and E highlight the stability of compositions of mixed phase in the parameter space, indicating that the phases of fibril and oligomer formation are intrinsically related to descriptions of topological interactions, and are independent of system size.

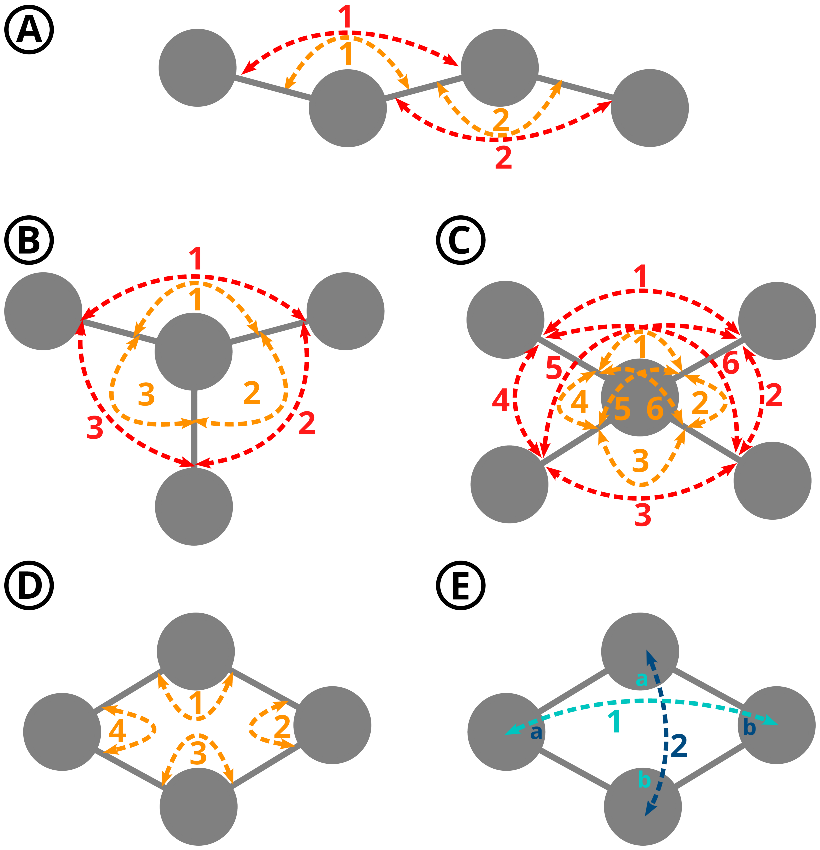


Figure C.3: Panels A-D show counts of t_{nsp_1}, t_{2s}, and t_{nsp_2} network topologies with red, yellow, and blue/teal dashed lines, respectively. Panel A depicts a minimal 1-ribbon fibril structure with 2 t_{2s} (yellow) and 2 t_{nsp_1} (red). Panel B shows the count of 3 t_{2s} and 3 t_{nsp_1} that result from adding a third edge to any vertex that already has exactly two edges, such as occurs at the $\phi_{2s} = -\phi_{nsp_1} + \phi_e^*/3$ boundary. Panel C shows the count of 6 t_{2s} and 6 t_{nsp_1} that results from the addition of a fourth edge to a vertex, such as at the $\phi_{nsp_2}/2 - \phi_{nsp_1} = -\phi_{2s} - \phi_{nsp_1} + \phi_e^*/6$ boundary where 2-ribbon fibrils share a boundary with the unstructured-aggregate (gel) phase. This highlights the geometrically increasing effect of adding an edge to a single vertex. Panel D shows the count of t_{2s} found in a 4-cycle oligomer, while panel E shows the count of 2 t_{nsp_2} in the same 4-cycle oligomer, with the null edge indicated by the blue and teal dashed lines, and the letters ’a’ and ’b’ indicating the two mutual partners being shared by each respective null edge. t_{nsp_1} do not occur in 4-cycles due to the non-nesting nature of the ”*shared partner*” network terms. Comparison of panels A, and E illustrates the relationship between t_{nsp_1} and t_{nsp_2} that determines the $\phi_{nsp_2} = 2\phi_{nsp_1} + \phi_e^*/4$ boundary between the 1-ribbon fibril phase and the 4-cycle oligomer phase. The addition of the fourth edge that forms the 4-cycle removes the 2 t_{nsp_1} found in the minimal 1-ribbon (panel A) by adding 2 t_{nsp_2}, meaning the decrease in energy got by having 2 t_{nsp_2} is large enough to compensate for losing 2 t_{nsp_1}, and exceeds that value by an amount equivalent to adding one t_e, thus losing the 1-ribbon structure while forming the 4-cycle.

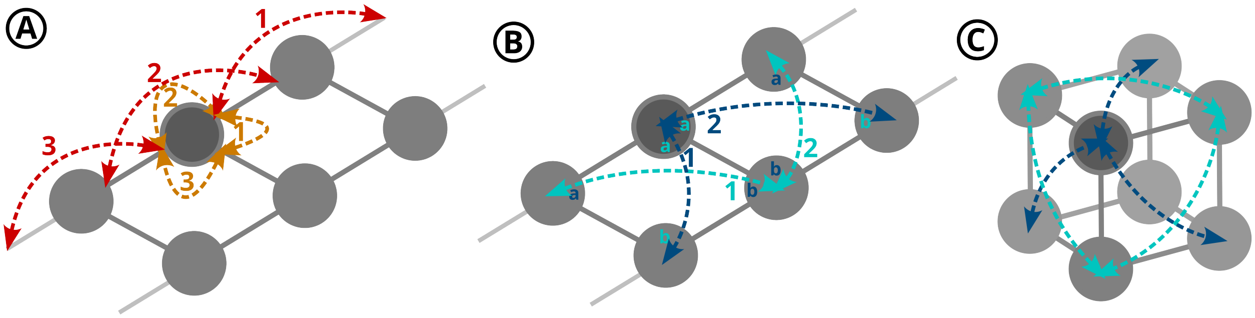


Figure C.4: Individual vertices experience sets of forces acting on them that differ according to structure. Panels A and B show the forces acting on a vertex within a 2-ribbon fibril, indicated with a dark gray shading. Panel A shows the ϕ_{2s} forces in yellow, and the ϕ_{nsp_1} forces in red. While the counts of t_{2s} and t_{nsp_1} are equivalent, the sets of edge and vertices involved in either structure type are different. Panel B shows the ϕ_{nsp_2} forces in blue and teal, with blue representing the nsp(2) structures in which the focal vertex is part of the null dyad, and teal representing the null dyads to which the focal vertex acts as a shared partner. The focal vertex is thus involved in a total of four nsp(2), three nsp(1), and three 2-star structures. Panel C shows a focal vertex that is part of a cubic oligomer, with ϕ_{nsp_2} forces illustrated in blue and teal as in panel B. In the case of the cube, the focal vertex is involved in six total nsp(2) structures and three 2-star structures, but cannot be described as being part of any nsp(1) structures. Thus, cubes occur at small negative values of ϕ_{nsp_1} and large negative values of ϕ_{nsp_2}, as described by the boundary $\phi_{nsp_2} = 2\phi_{nsp_1} - \phi_e^*/3$.