



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Self-organized emergence of folded protein-like network structures from geometric constraints

Citation for published version:

Molkenthin, N, Mühle, S, Mey, ASJS & Timme, M 2020, 'Self-organized emergence of folded protein-like network structures from geometric constraints', *PLoS ONE*, vol. 15, no. 2, pp. e0229230. <https://doi.org/10.1371/journal.pone.0229230>

Digital Object Identifier (DOI):

[10.1371/journal.pone.0229230](https://doi.org/10.1371/journal.pone.0229230)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

PLoS ONE

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



RESEARCH ARTICLE

Self-organized emergence of folded protein-like network structures from geometric constraints

Nora Molkenhain^{1,2*}, Steffen Mühle³, Antonia S. J. S. Mey⁴, Marc Timme^{1,2}

1 Chair for Network Dynamics, Institute for Theoretical Physics and Center for Advancing Electronics Dresden (cfaed), Technical University of Dresden, Dresden, Germany, **2** Network Dynamics, Max Planck Institute for Dynamics and Self-Organization (MPIDS), Göttingen, Germany, **3** University of Göttingen, Third Institute of Physics – Biophysics, Göttingen, Germany, **4** EaStCHEM School of Chemistry, University of Edinburgh, Edinburgh, Scotland, United Kingdom

* molkenhain@nld.ds.mpg.de

OPEN ACCESS

Citation: Molkenhain N, Mühle S, Mey ASJS, Timme M (2020) Self-organized emergence of folded protein-like network structures from geometric constraints. PLoS ONE 15(2): e0229230. <https://doi.org/10.1371/journal.pone.0229230>

Editor: Yaakov Koby Levy, Weizmann Institute of Science, ISRAEL

Received: December 2, 2019

Accepted: January 31, 2020

Published: February 27, 2020

Copyright: © 2020 Molkenhain et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data is available at <https://github.com/ppxasjsm/Geometric-constraints-protein-folding>.

Funding: We acknowledge partial support by the Max Planck Society [NM, MT], the Deutsche Forschungsgemeinschaft (DFG) through SFB 755 (project A05) [SM], the Engineering and Physical Sciences Research Council (EPSRC) UK under grant no. EP/P022138/1 [AM], and the DFG through funding the Clusters of Excellence Center for Advancing Electronics Dresden (cfaed) and

Abstract

The intricate three-dimensional geometries of protein tertiary structures underlie protein function and emerge through a folding process from one-dimensional chains of amino acids. The exact spatial sequence and configuration of amino acids, the biochemical environment and the temporal sequence of distinct interactions yield a complex folding process that cannot yet be easily tracked for all proteins. To gain qualitative insights into the fundamental mechanisms behind the folding dynamics and generic features of the folded structure, we propose a simple model of structure formation that takes into account only fundamental geometric constraints and otherwise assumes randomly paired connections. We find that despite its simplicity, the model results in a network ensemble consistent with key overall features of the ensemble of Protein Residue Networks we obtained from more than 1000 biological protein geometries as available through the Protein Data Base. Specifically, the distribution of the number of interaction neighbors a unit (amino acid) has, the scaling of the structure's spatial extent with chain length, the eigenvalue spectrum and the scaling of the smallest relaxation time with chain length are all consistent between model and real proteins. These results indicate that geometric constraints alone may already account for a number of generic features of protein tertiary structures.

1. Introduction

Proteins consist of sequences of amino acids. The resulting *primary structure* of a protein, is expected to provide constraints for the folded three-dimensional (3D) structure of a globular protein, its *tertiary structure*. The problem of predicting the 3D structure of an amino acid sequence in an aqueous solution is known as the protein folding problem consisting of three sub-problems: First, to find the chemically active folded state; second, to uncover the pathway to get to that state; and third, to develop computational tools capable of accurately predicting the folded state [1–6]. Many different avenues have been taken to explore solutions towards

Physics of Life (PoL) [MT] The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

this problem, ranging from atomistic models using molecular dynamics approaches [7], to coarse grained models e.g [8], and to machine learning-based and heuristic physical models that disregard the atomistic details of the amino acid sequence [9, 10]. While much progress has been made improving molecular dynamics simulations using atomistic detail, the folding process of long chains is computationally highly expensive or even infeasible, and still requires access to purpose build massively parallel computers such as Anton [11], or distributed computing projects such as folding@home in order to generate quantitative data [12]. The other avenue often explored for structure models is tested in community-wide challenges such as the ‘Critical Assessment of Protein Structure Prediction’ (CASP) [13–15]. CASP is run every other year to see if a protein’s tertiary structure can be predicted based on its primary sequence of protein structures unresolved at the time of the challenge [16]. Predictions have improved drastically over previous CASP challenges [1], however, often rely on existing structural information in the protein data base (PDB) and homology modeling, comparing new proteins based on existing insights from known template proteins using computational models such as HHPred [17] or I-TASSER [18] or, more recently, machine-learning based predictions of the distance matrix [19]. These approaches support accurate prediction of 3D structures, yet by construction limit insights into fundamental physical mechanisms and constraints underlying the folding processes and final structures observed in the many and various proteins observed in nature.

In a complementary approach, a number of theories and structure analyses have been conducted into broader mechanisms of the folding process. Examples include tube models as in [20], where it was shown that secondary structures, such as helices and β -sheets, arise from explicit hydrogen bonds. Alternatively the behaviour of the chain can be expressed through a heuristic field equation of the backbone curvature, as in [21], where helices and sheets constitute the energy minimum. For this reason we here want to focus on the tertiary structure.

We propose an approach to further understand geometry and formation processes using a complex network framework. The 3D tertiary structures in our model arise from chain-like primary protein structures without comparing to specifically chosen protein structures available on the PDB, and without using complex molecular dynamics simulations. First, we analyze 1122 protein structures from the PDB, consider them as an ensemble of network structures representing protein tertiary structures, and quantify overall properties of this ensemble. In particular we (i) uncover the scaling of the diameter of proteins with their chain length, (ii) reveal the distribution of the number of other amino acids any given amino acid closely interacts with and (iii) find the distribution of second largest eigenvalues of their associated graph Laplacians, characterizing the most persistent time scales on which proteins are dynamically responding to perturbations. Second, we propose and analyze a simple stochastic process modeling the folding of chains of units. The minimal model takes into account geometric constraints only and does not consider any other protein property. The model process keeps connected units connected, forbids geometric overlap of units (volume exclusion) and connects randomly chosen units if geometrically permitted. Based only on such random monomer interactions and geometric constraints, akin to those in Lennard-Jones clusters and sticky hard spheres [22, 23], the 3D structures self-organizing through the simple model process are consistent with those of real protein ensembles in all of the above-mentioned features simultaneously.

These results suggest that beyond the details of pairwise interaction of amino acids, from intermediate scales of a few amino acids to the full spatial extent of proteins, geometric constraints play an important role in structure formation and strongly impact the final protein tertiary structure. Our insights may put into perspective the influence of the specific details of

sequences of amino acids relative to simpler geometric constraints on structure forming processes of proteins.

II. Methods

A. Ensemble analysis of protein residue networks

With their modular polymer structure and their complex interaction patterns, proteins lend themselves naturally to a description as ensembles of complex networks. The mathematical object of a graph, simply termed network, represents a structure of nodes (units) and links, each describing an interaction between two units [24–26]. Networks and graphs have been used to describe the structure of a wide variety of systems, as different as social networks [27–29] and the global climate system [30, 31]. In this article, we analyze an ensemble of 1122 protein tertiary structures of chain lengths ranging from $N = 8$ to $N = 1500$ amino acids. Detailed structures have been experimentally determined to great accuracy and stored in the protein data bank (PDB) [32]. Part of the information stored in the PDB are the coordinates $x_i \in \mathbb{R}^3$ of the individual amino acid's central carbon atoms C_{α} where i indexes the amino acid's position along the chain.

Given such geometric data, the structures resulting from protein folding are commonly expressed as protein residue networks (PRN's) [33–36], in which the central carbon atom of each amino acid is taken to be a node and a link represents the interaction of two nodes if their spatial distance is small, i.e. less than a distance d_c apart.

Here, the distance between the amino acids indexed i and j is given by the Euclidean distance metric $d_{i,j} = \|x_i - x_j\|$. An adjacency matrix A_{ij} encodes the topology of a network, its entries are 1 if $d_{i,j} \leq d_c$, i.e. the units are considered connected, and 0 otherwise. The distance matrix resulting from PDB data thus defines the adjacency matrix as

$$A_{ij}^{\text{PDB}} = \begin{cases} 0, & \text{if } d_{i,j} > d_c \text{ or } i = j \\ 1, & \text{if } d_{i,j} \leq d_c. \end{cases} \quad (1)$$

The threshold of the PRN is commonly chosen between $d_c = 4 \text{ \AA}$ (approximate length of a peptide bond [35]) and $d_c = 8 \text{ \AA}$, reflecting an upper bound for a significant interaction to occur between two units [35]. Here, we created the PRNs of 1122 proteins selected from the PDB list in [37], covering a range of chain lengths N for comparison to simulations. Their geometric structures have been determined previously via NMR and x-ray studies. We choose a threshold value of $d_c = 6.5 \text{ \AA}$ to calibrate the average degree (the degree k_i of node i counts the number of nodes it is connected to) of nodes in the PRNs to the average degree found in the model simulations in the range of large $N \in [200, 400]$, Fig 1a. The average degree k grows with N and appears to saturate at a value determined by d_c . The ratio of this cutoff threshold and the unit size in the model, which we take half their mean distance, constitutes the only free structural parameter we employ in the current study.

The degree distribution of the resulting network ensemble, displayed in Fig 1b, is unimodal and covers effective degrees between $k = 2$ and $k = 11$. Interestingly, the degree distribution resulting from simulations of the model ensemble we are about to introduce below is statistically indistinguishable from those of the network of real PRNs (no additional fit parameter), Fig 1b. Equally, other quantifiers obtained from the simple, geometry-only model ensemble agree surprisingly well with those obtained from our data analysis of the experimentally obtained protein structures. For the network measures and manipulations NetworkX [38] was used.

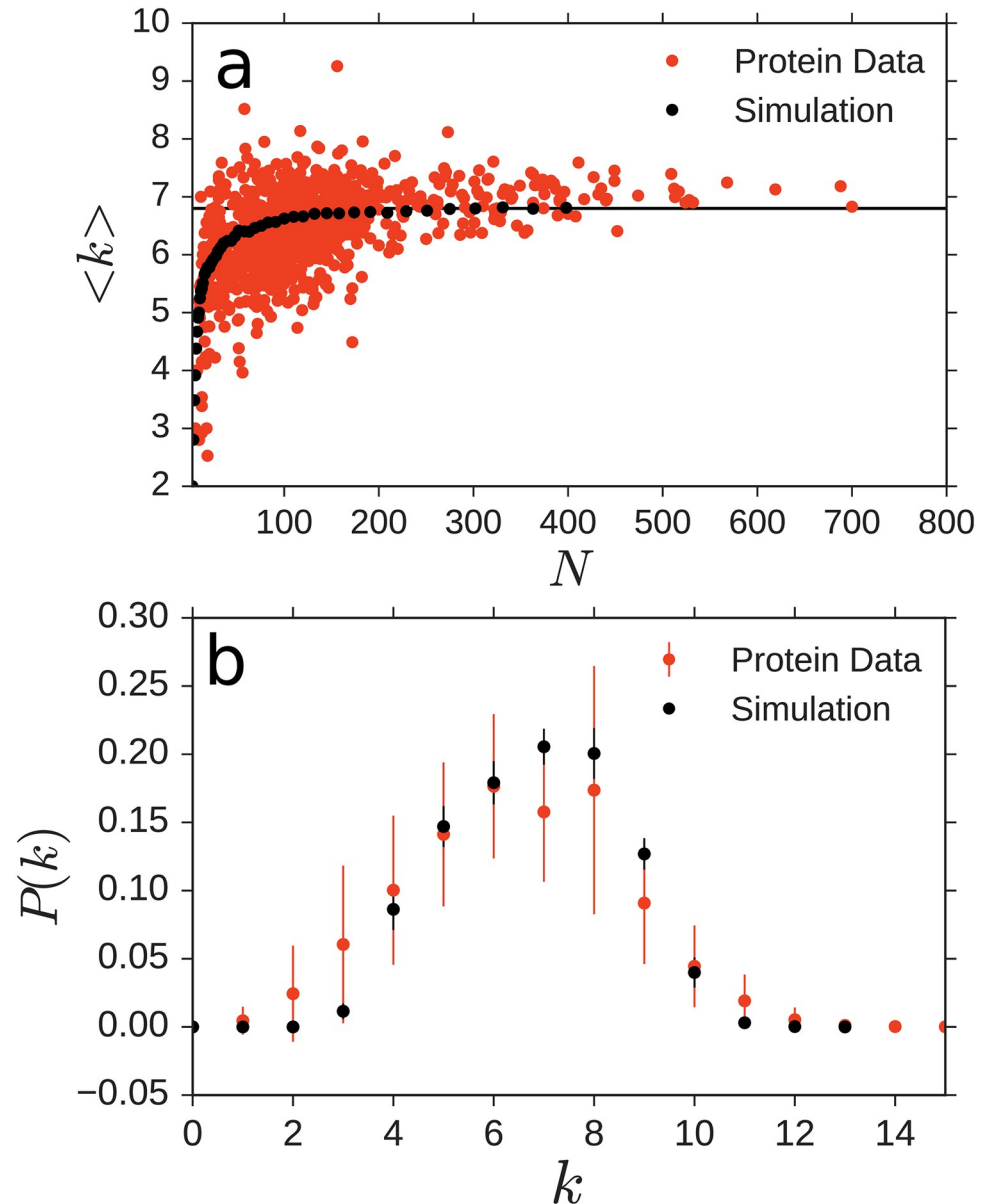


Fig 1. Degree distributions of simple model ensemble and real proteins are statistically indistinguishable. a) The average degree k of real protein ensemble (red dots) asymptotically saturates to $k \approx 6.8$ as the chain length N becomes large. The average degree of the nodes resulting from 30 model simulations for each chain length N , ranging from $N = 3$ to $N = 398$. b) The degree distribution $P(k)$ of the model simulation within the error margin is indistinguishable from that of real proteins (error bars indicate standard deviation of the distribution at each k).

<https://doi.org/10.1371/journal.pone.0229230.g001>

B. Simple model focused on geometric constraints

To better understand the impact of geometric constraints on the topology of protein tertiary structures, we introduce a random network formation model that takes into account geometric constraints and leaves out almost all other properties of real proteins, including heterogeneous sequences of amino acids, the amino acids' specific molecular properties, different forms of electrochemical interactions, conformational details of interactions between nearby amino-acids, and the influence of the fluid environment on protein folding. This formation

mechanism can be interpreted as the intersection of random graphs and self-avoiding random walks, which has vastly different properties from the two individual sets. We find that the simple, geometry-centered model already reproduces a range of overall topological properties of real protein residue networks well.

The model is built on the simple observation that proteins consist of a chain of close-to-identical units that interact in complex ways when folding, yet can not intersect, giving rise to geometric constraints. The individual units of the chain interact when they come into contact; typically there is an attraction that is the stronger the closer they are but repelling once they overlap. Depending on the specific amino acid, size, shape, and electromagnetic properties vary. In our model, however, all amino acids are represented as unit spheres and the interactions between each pair become very simple and identical across all pairs.

The model's initial state consists of a chain of N connected spheres, each of diameter and bond length of unity (later rescaled to match the mean distance between neighboring amino acids d_{mean}). A folding proceeds by sequentially picking random pairs of spheres (not connected with each other) and connecting them if possible, given the geometric constraints of volume exclusion. Here, volume exclusion also applies to co-moving other spheres connected either initially along the chain or through a previous step (see [S1 File](#)). The process repeats until all pairs are either connected or geometrically incapable of connecting. The adjacency matrix A^{sim} of the simulated chain keeps track of which spheres are linked to each other. Initially, it contains only zeros except for its secondary diagonal elements which equal 1 since neighboring spheres are connected via the backbone chain. The model is motivated by a two-dimensional model of network-based formation of aggregates where link constraints due to geometry in space have been approximately mapped to purely graph-theoretic constraints during network formation [39].

As described in the method section, the process of moving spheres towards each other is realized in a simple consistent way to satisfy all geometric constraints continuously in time. The forces and potentials employed, however, are *not* intended to reflect any physical forces or potentials created by amino acids. They plainly help to realize to attempt the joining of two randomly selected spheres.

Snapshots of the folding process are illustrated in [Fig 2](#), three examples of the final aggregates in [Fig 3](#). The aggregates are highly compact compared to the straight initial conditions. They are also much more compact than aggregates generated from self-avoiding random walks and close to, yet not quite maximally densely packed (see below), consistent with previous suggestions based on 2D aggregates [39].

All simulation details, including the code for reproducing the geometric constraint simulations, as well as the preparation and analysis of PDB files can be found in the following github repository: <https://github.com/ppxasjsm/Geometric-constraints-protein-folding>.

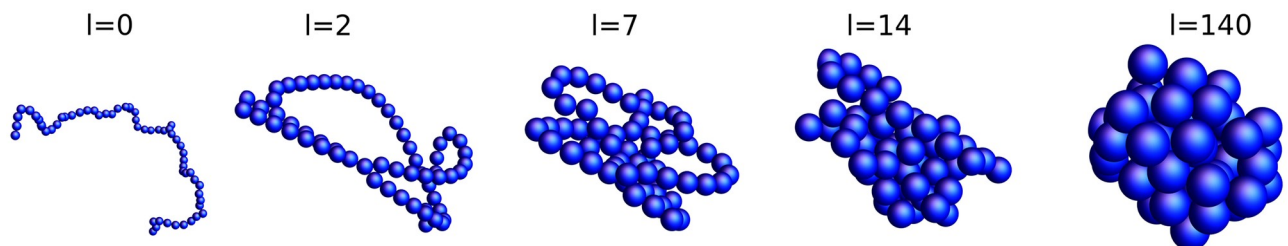


Fig 2. Model folding process at different times. Starting from an initial chain with $N = 60$, randomly picked units connect if geometrically possible. Shown here are examples after $l = 0, 2, 7, 14$ and 140 successful connection attempts.

<https://doi.org/10.1371/journal.pone.0229230.g002>

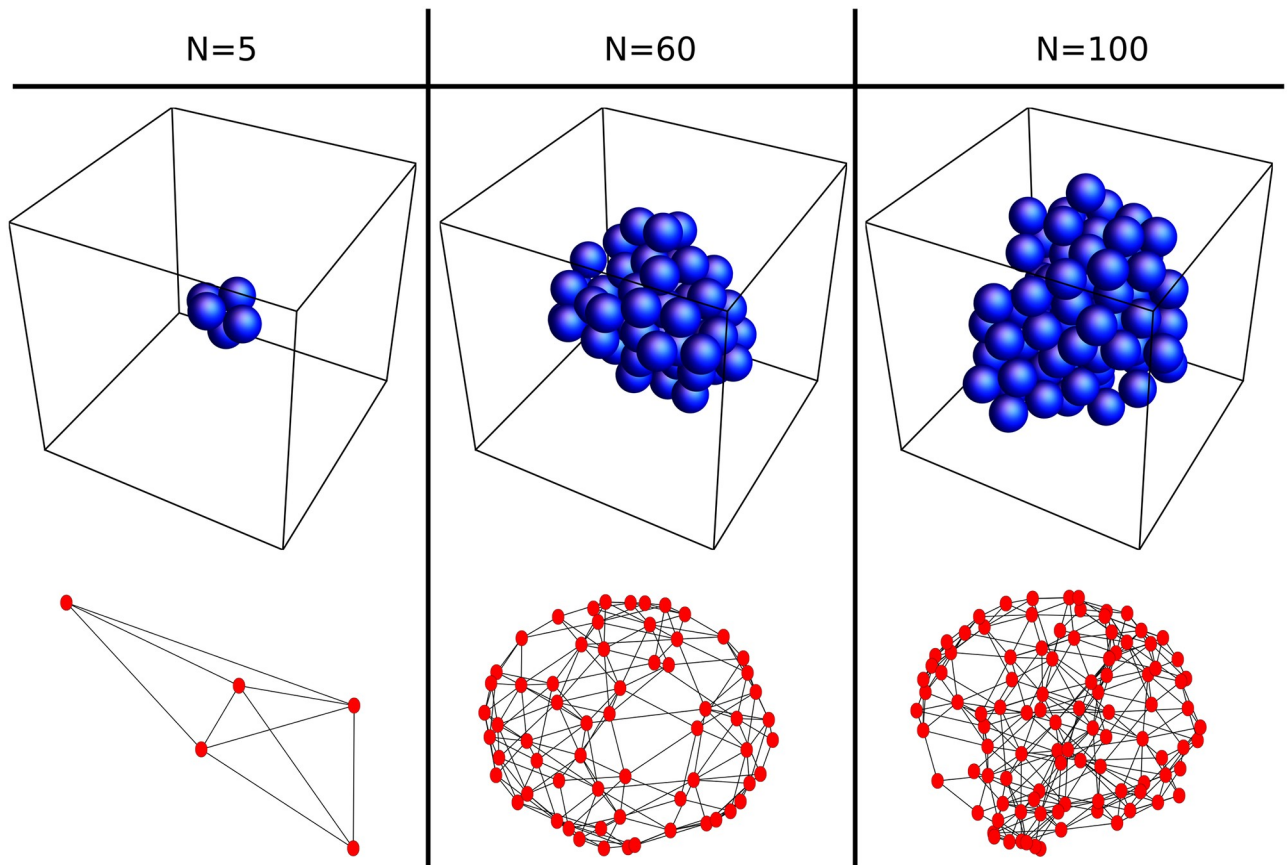


Fig 3. Final model aggregates. The final aggregates of the simulation for $N = \{5, 60, 100\}$ display the expected compactness. The corresponding networks are non-planar.

<https://doi.org/10.1371/journal.pone.0229230.g003>

III. Results

A. Spatial scaling of protein structures

The ensemble of protein tertiary structures exhibits an algebraic scaling law indicating that their radii of gyration R_g depend on their chain length N such that:

$$R_g \sim N^\nu, \tag{2}$$

as expected from a number of previous studies [21, 37, 39, 40]. As the overall geometry of a folded protein is often characterized by the locations of the central carbon atoms (C_α -atoms, one for each amino acid) of its backbone chain, its spatial extension is commonly measured by the radius of gyration

$$R_g = (N^{-1} \sum_i (x_i - \bar{x})^2)^{1/2}, \tag{3}$$

quantifying the average distance of units from the center of mass \bar{x} , where x_i is the location of unit $i \in \{1, \dots, N\}$. Our previous study [39] revealed that the scaling law indeed is algebraic and that the exponent ν is (slightly) larger than for space filling aggregates (where $\nu_{\text{SF}} = \frac{1}{3} = 0.3333 \dots$ in 3D) yet (far) smaller than for aggregates created through a self-avoiding random walk (where $\nu_{\text{RW}} = \frac{3}{5} = 0.6$ in 3D). That study found $\nu = 0.3916 \pm 0.0008$ for

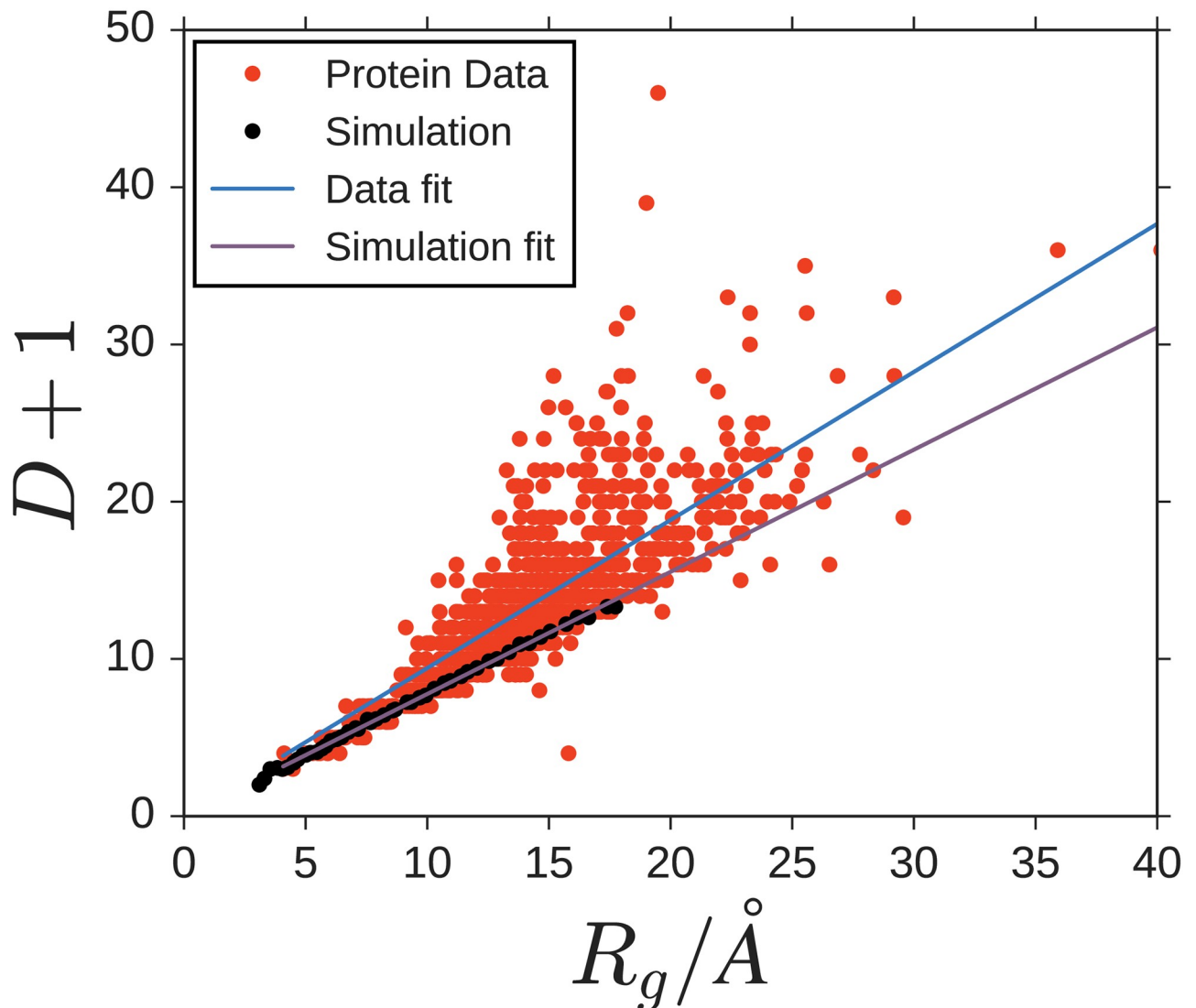


Fig 4. The network diameter D scales linearly with the radius of gyration R_g . This holds for both biological protein residue networks and simulated model networks. Scaling the model link length to the average link length of the PRN (see text for details), yields a scaling of the graph diameter of model networks within the experimentally observed range. The best fitting proportionality constant, however, differs, with $\frac{\partial D}{\partial R_g} = 0.942 \text{Å}^{-1}$ for experimental data and $\frac{\partial D}{\partial R_g} = 0.777 \text{Å}^{-1}$ for the model data.

<https://doi.org/10.1371/journal.pone.0229230.g004>

37162 proteins. For our smaller data set of 1122 proteins, we find $v_{exp} = 0.374 \pm 0.03$, see Fig 4 for illustration.

To compare the spatial extent of model aggregates, i.e. graph-theoretically defined networks of spheres, to biological proteins on the same footing, we first study how the network diameter D compares to the radius of gyration defined through Eq (3). The graph diameter is defined as the maximum number of links to be taken on the shortest link sequence (also referred to as shortest simple paths) between any pair of units in the PRN. We find that D is strongly linearly correlated with the spatial extent R_g of the PRN, Fig 4. Both the ensemble of biological proteins and the model ensembles studied exhibit a roughly proportional dependence of $D+1$ on R_g , with the slope obtained from the model data ($\frac{\partial}{\partial R_g} D = 0.777 \text{Å}^{-1}$) being lower and more

precisely determined than that obtained for the PRNs ($\frac{\partial}{\partial R_g} D = 0.942 \text{ \AA}^{-1}$). As proportionality factors do not affect the scaling, we thus also find

$$(D + 1) \sim N^{\nu}, \tag{4}$$

for both the PDB proteins and geometric-constraint model.

With the cutoff distance for the creation of networks chosen to be $d_c = 6.5 \text{ \AA}$ the resulting average link length in the biological proteins becomes $d_{mean} \approx 5.066 \text{ \AA}$, which in Fig 4 we substituted for the unit length of our model simulations. In the PRNs the network diameters are more dispersed. The lower bound of the experimental data fits well with the simulated structures, suggesting geometric constraints as a major driving mechanism influencing the spatial density during network formation.

Both ensembles show power-law scaling of the diameter. The exponent of $\nu_{sim} = 0.345 \pm 0.01$ of the simulation is very close to the value of $\nu_{exp} = 0.374 \pm 0.03$, measured in the PDB data. The plots are shown in Fig 5. Simulations for heterogeneous systems where the radii of individual units are drawn randomly from the uniform distribution on $[1 - a, 1 + a]$ for $a \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ increased the variance of the measurements for the radius of gyration, as expected. We did not observe any significant bias in the averages such that the scaling relations stay the same also for heterogeneous systems. The simulated results are found to align very well with the lower bound of folded protein diameters, suggesting that much of the discrepancy (constant factor shifting the measured results up in Fig 5) can be explained by the fact that the simulation only ceases to make new links when this is no longer geometrically possible. In real proteins on the other hand interactions range from Van-der-Waals interactions to hydrogen bonds and individual monomers vary in size and chemical properties and are subject to thermodynamic fluctuations. All this leads to larger gaps within the folded molecule and hence larger diameters of the PRN's.

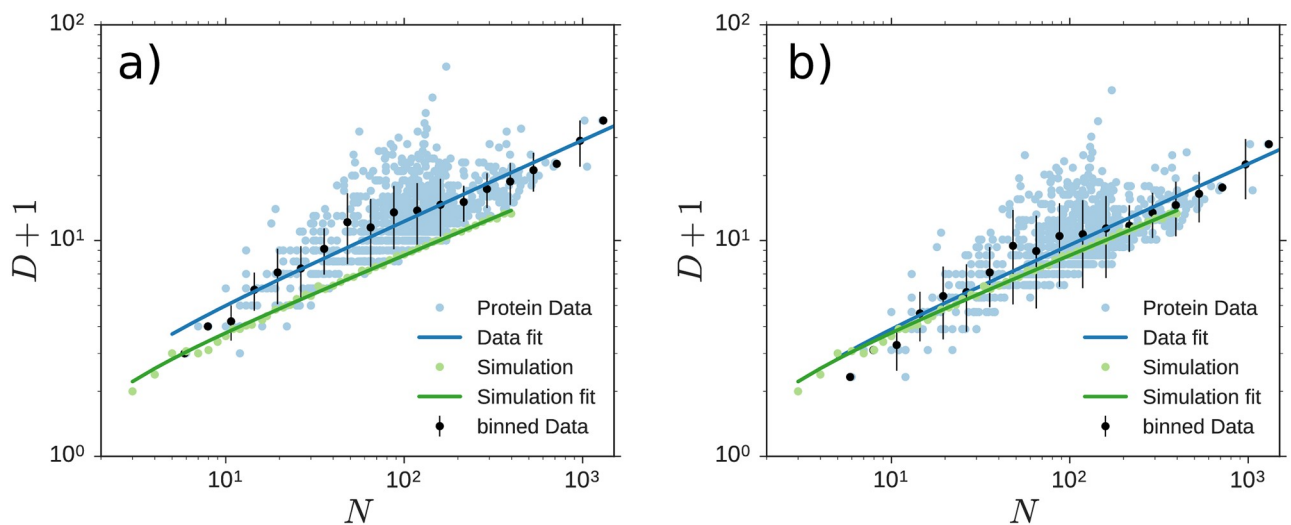


Fig 5. Diameter scaling with chain length. (a) The diameter D of simulated and measured PRN's scales according to Eq 4 with the chain length N . The model results coincide with the lower bound of measured results, which we attribute to the fact that we fold maximally. (b) Matching the proportional scaling relation between graph diameter D and radius of gyration (Fig 4) yields scaling relations between aggregate extent and chain length to be statistically indistinguishable between model and real proteins. For both panels, we simulated 30 random dynamic realizations each for 48 aggregate lengths N with logarithmically spaced between $N = 3$ to $N = 398$. The data displayed shows the network diameter averaged across realizations as a function of chain length.

<https://doi.org/10.1371/journal.pone.0229230.g005>

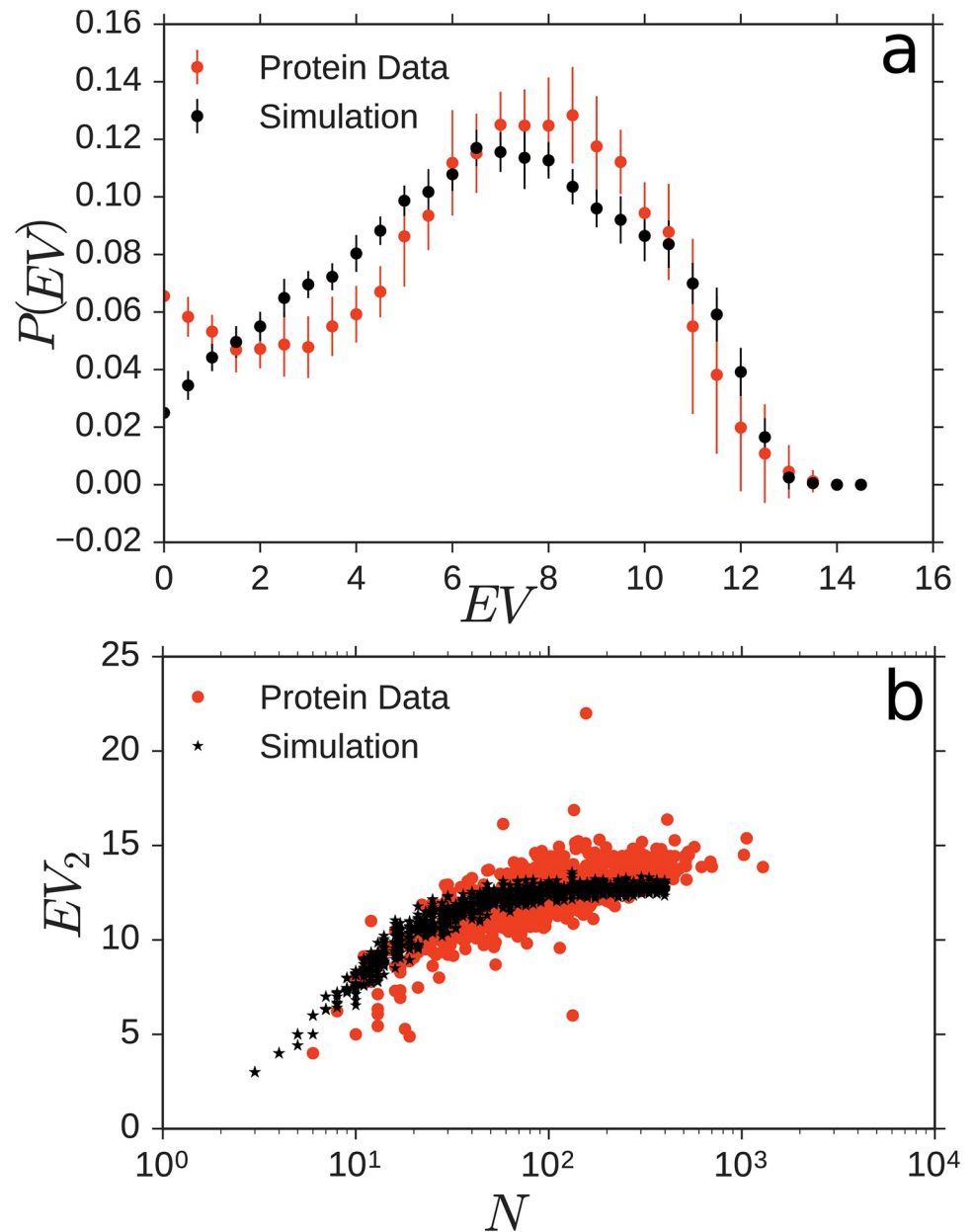


Fig 6. Model eigenvalue spectra of the graph Laplacian are similar to those of the PRN spectra. a) Histograms $P(EV)$ of eigenvalue spectra of PRN's with $N \approx 400$ and $r_c = 6.5 \text{ \AA}$ compared to model output at $N = 400$. b) Second largest eigenvalues EV_2 grow in similar ways for simulation and data. All eigenvalues $\lambda_1, \dots, \lambda_N$ for an $(N \times N)$ Laplacian matrix $L^s im = A^s im - diag_i(A_{ii})$ are computed using the routine provided by NetworkX [38]; the second largest eigenvalue $\lambda_2 = EV_2$ of those is plotted in panel b.

<https://doi.org/10.1371/journal.pone.0229230.g006>

B. Distribution and scaling of Laplacian eigenvalues

Lastly we explore the scaling of the second largest eigenvalue of the graph Laplacian with N in Fig 6 and find that it grows with N , approaching a saturation point of ≈ 15 for large N .

As two additional features roughly characterizing the dynamic properties of protein residue networks, we consider the distribution and scaling of Laplacian eigenvalues. The Laplacian of a network captures both its interaction topology and its relaxation and vibration properties

[41, 42]. If the PRN were made only of the central C_{α} atoms, the Laplacian would exactly quantify the networks vibrational and relaxational modes. As real PRNs are more complex, the Laplacian spectrum can be taken as a proxy for oscillatory and relaxation dynamics.

Because the eigenvalue spectra intrinsically scale with graph size (here: chain length), we have evaluated the spectra of simulated structures and PRNs of lengths of $N = 400 \pm 30$. Fig 6a shows the histogram of eigenvalues for the 18 PRNs (red) in that length range, accumulating all N eigenvalues for each of the 18 PRNs. For comparison, we computed 28 simulated structures (black), that fall in the same length range.

Both eigenvalue spectra exhibit a characteristic unimodal shape. The simulated structures have a more symmetric, slightly broader spectrum with a peak at $\lambda \approx 7$, while the PRN's have a slightly sharper peak at $\lambda \approx 8$ and higher probabilities for very small eigenvalues. Similarly, the second largest Laplacian eigenvalue exhibits the same qualitative scaling with chain length N for PRNs and geometric-constraint model. The second largest eigenvalue of a network's Laplacian quantifies the time scale of its slowest relaxing mode; as such, its scaling with chain length N indicates how intrinsic relaxation time scales change due to the aggregates becoming larger.

The spectra and equally the scaling of the second largest eigenvalues are not indistinguishable between model and biological protein data yet overall exhibit similar properties. Whether or not spectra of model ensemble and PRN ensemble actually agree or disagree cannot be concluded without doubt from the data available, both because at (exactly) fixed chain length N there typically is no, one, or only very few proteins available in the real protein data set and because the model realizations at fixed N yield very similar spectra due to chain homogeneity. There is no unbiased way we know of to account for uncertainties in N and simultaneously inhomogeneities in the chain units such that a unambiguous conclusion can be drawn.

IV. Discussion

In this article we have proposed a simple model of spatial network formation taking into account geometric constraints only. Decoupling the constraints, that drive the folding process (geometry, sequence and solution) and focusing on the geometry allows us insights into the folding mechanisms behind the ensemble features. While this approach does not yield direct predictive power to find the native state of a specific sequence it may narrow down the landscape of possibilities.

We find that geometrically constrained random linking already leads to strong similarities of the resulting structures with protein residue networks in biology. Generalizing a 2D model of purely graph-theoretical network formation presented in [39] to 3D, the model is based upon random link additions with geometric constraints. As the topological shortcut is no longer possible, the geometric constraints are simulated directly. The simulation results were then compared to protein residue networks (PRN's), choosing the threshold such that the mean degrees of simulation results and PRN's matched. As a result, the degree distributions are within the error margins of each other.

The network diameter is linearly related to the radius of gyration in both simulation and data and matches when the simulation results are correctly scaled with the mean connection lengths. The network diameter scales with the chain length as a shifted power law with an exponent of $\nu_{sim} = 0.345 \pm 0.01$, which is in agreement with value of $\nu_{exp} = 0.374 \pm 0.03$, measured in PRN's. As in 2D, this is slightly less than space filling.

Furthermore, we have studied the Laplacian eigenvalue spectrum and the scaling of the second largest eigenvalue with system size, finding that the two systems are compatible. Using the findings from [41, 42] we can infer that the structure of vibrational modes and relaxation properties produced by the model are similar to those found in biological proteins.

These results can be taken as an indication that geometric constraints may be a mechanism behind the scaling behaviour of real protein structures, generating an ensemble also compatible on degree distribution and Laplacian spectrum. Further research, however, is necessary to determine how far the structural similarity reaches. For example by comparing further topological characteristics of PRN's vs. model simulations. If the analogy persists, the model could be extended to allow simple sequence features, such as hydrophobicity to attempt to get a simpler predictive model. This may give insights into the folding process, that are otherwise lost in simulation complexity.

Taken together, the above results indicate that coarse ensemble properties of protein tertiary structures are already induced by geometric constraints alone such that only finer scales of the folded structures of individual proteins may be controlled by the details of their amino acid sequences. Such simple models provide a new angle of analyzing protein structures at the coarse scale of ensembles and may help understand core mechanisms underlying the complex folding processes.

Supporting information

S1 File.
(PDF)

Author Contributions

Conceptualization: Nora Molkenhain, Marc Timme.

Data curation: Nora Molkenhain, Antonia S. J. S. Mey.

Funding acquisition: Marc Timme.

Investigation: Nora Molkenhain, Steffen Mühle, Antonia S. J. S. Mey.

Methodology: Nora Molkenhain, Steffen Mühle, Antonia S. J. S. Mey.

Resources: Marc Timme.

Software: Nora Molkenhain, Steffen Mühle, Antonia S. J. S. Mey.

Supervision: Marc Timme.

Visualization: Nora Molkenhain.

Writing – original draft: Nora Molkenhain, Steffen Mühle, Antonia S. J. S. Mey.

Writing – review & editing: Nora Molkenhain, Steffen Mühle, Antonia S. J. S. Mey, Marc Timme.

References

1. Dill Ken A. and MacCallum Justin L. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012. <https://doi.org/10.1126/science.1219021> PMID: 23180855
2. Scheraga Harold A., Khalilii Mey, and Liwo Adam. Protein-folding dynamics: overview of molecular simulation techniques. *Annual Review of Physical Chemistry*, 58:57–83, 2007. <https://doi.org/10.1146/annurev.physchem.58.032806.104614> PMID: 17034338
3. Mirny Leonid and Shakhnovich Eugene. Protein folding theory: from lattice to all-atom models. *Annual Review of Biophysics and Biomolecular Structure*, 30:361–396, 2001. <https://doi.org/10.1146/annurev.biophys.30.1.361> PMID: 11340064
4. Shakhnovich Eugene. Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry, and Biology Meet. *Chemical Reviews*, 106:1559–1588, 2006. <https://doi.org/10.1021/cr040425u> PMID: 16683745

5. Saunders Marissa G and Voth Gregory A. Coarse-graining methods for computational biology. *Annual Review of Biophysics*, 42:73–93, 2013. <https://doi.org/10.1146/annurev-biophys-083012-130348> PMID: 23451897
6. Mey Antonia SJS, Geissler Phillip L, and Garrahan Juan P. Rare-event trajectory ensemble analysis reveals metastable dynamical phases in lattice proteins. *Physical Review E*, 89(3):032109, 2014. <https://doi.org/10.1103/PhysRevE.89.032109>
7. Snow Christopher D, Nguyen Houbi, Pande Vijay S, and Gruebele Martin. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, 420(6911):102–106, 2002. <https://doi.org/10.1038/nature01160> PMID: 12422224
8. Clementi Cecilia. Coarse-grained models of protein folding: toy models or predictive tools? *Current Opinion in Structural Biology*, 18(1):10–15, 2008. <https://doi.org/10.1016/j.sbi.2007.10.005> PMID: 18160277
9. Cheng Jianlin and Baldi Pierre. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, 22(12):1456–1463, 2006. <https://doi.org/10.1093/bioinformatics/btl102> PMID: 16547073
10. Dill Ken A., Bromberg Sarina, Yue Kaizhi, Chan Hue Sun, Ftebig Klaus M., Yee David P., et al. Principles of protein folding—a perspective from simple exact models. *Protein Science*, 4(4):561–602, 1995. <https://doi.org/10.1002/pro.5560040401> PMID: 7613459
11. Shaw David E, Deneroff Martin M, Dror Ron O, Kuskin Jeffrey S, Larson Richard H, Salmon John K, Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008. <https://doi.org/10.1145/1364782.1364802>
12. Shirts Michael and Pande Vijay S. Screen savers of the world unite! *Science*, 290(5498):1903–1904, 2000. <https://doi.org/10.1126/science.290.5498.1903> PMID: 17742054
13. Moult John, Fidelis Krzysztof, Kryshtafovych Andriy, Rost Burkhard, Hubbard Tim, and Tramontano Anna. Critical assessment of methods of protein structure prediction—round vii. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):3–9, 2007. <https://doi.org/10.1002/prot.21767>
14. Moult John, Fidelis Krzysztof, Kryshtafovych Andriy, Schwede Torsten, and Tramontano Anna. Critical assessment of methods of protein structure prediction (CASP)—round xii. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15, 2018. <https://doi.org/10.1002/prot.25415>
15. Monastyrskyy Bohdan, Daniel D'Andrea, Fidelis Krzysztof, Tramontano Anna, and Kryshtafovych Andriy. New encouraging developments in contact prediction: Assessment of the CASP 11 results. *Proteins: Structure, Function, and Bioinformatics*, 84:131–144, 2016. <https://doi.org/10.1002/prot.24943>
16. Ogorzalek Tadeusz L., Hura Greg L., Belsom Adam, Burnett Kathryn H., Kryshtafovych Andriy, Tainer John A. et al. Small angle x-ray scattering prediction and cross-linking for data assisted protein structure prediction in CASP 12 with prospects for improved accuracy. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):202–214, 2018. <https://doi.org/10.1002/prot.25452>
17. Bradley Philip, S. Misura Kira M., and David Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005. <https://doi.org/10.1126/science.1113801> PMID: 16166519
18. Yang Jianyi, Yan Renxiang, Roy Amrsh, Xu Dong, Poisson Jonathan, and Zhang Yang. The I-TASSER suite: protein structure and function prediction. *Nature Methods*, 12(1):7, 2015. <https://doi.org/10.1038/nmeth.3213> PMID: 25549265
19. Xu Jinbo. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, 116(34):16856–16865, 2019. <https://doi.org/10.1073/pnas.1821309116>
20. Hubner Isaac A and Shakhnovich Eugene I. Geometric and physical considerations for realistic protein models. *Physical Review E*, 72(2):022901, 2005. <https://doi.org/10.1103/PhysRevE.72.022901>
21. Danielsson Ulf H., Lundgren Martin, and Niemi Antti J. Gauge field theory of chirally folded homopolymers with applications to folded proteins. *Physical Review E*, 82(2):1–5, 2010. <https://doi.org/10.1103/PhysRevE.82.021910>
22. Trombach Lukas, Hoy Robert S, Wales David J, and Peter Schwerdtfeger. From sticky-hard-sphere to lennard-jones-type clusters. *Physical Review E*, 97(4):043309, 2018. <https://doi.org/10.1103/PhysRevE.97.043309> PMID: 29758765
23. Uppenbrink Julia and Wales David J. Packing schemes for lennard-jones clusters of 13 to 150 atoms: minima, transition states and rearrangement mechanisms. *Journal of the Chemical Society, Faraday Transactions*, 87(2):215–222, 1991. <https://doi.org/10.1039/ft9918700215>
24. Albert R. and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002. <https://doi.org/10.1103/RevModPhys.74.47>

25. Boccaletti Stefano, Latora Vito, Moreno Yamir, Chavez Martin, and Hwang D-U. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006. <https://doi.org/10.1016/j.physrep.2005.10.009>
26. Newman Mark EJ. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. <https://doi.org/10.1137/S003614450342480>
27. Watts Duncan J and Strogatz Steven H. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 1998. <https://doi.org/10.1038/30918> PMID: 9623998
28. Girvan Michelle and Newman Mark EJ. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. <https://doi.org/10.1073/pnas.122653799>
29. Molkenhthn Nora, Schröder Malte, and Timme Marc. Adhesion-induced discontinuous transitions and classifying social networks. *Physical review letters*, 121(13):138301, 2018. <https://doi.org/10.1103/PhysRevLett.121.138301> PMID: 30312048
30. Donges Jonathan F, Zou Yong, Marwan Norbert, and Jürgen Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, 2009. <https://doi.org/10.1140/epjst/e2009-01098-2>
31. Molkenhthn Nora, Rehfeld Kira, Marwan Norbert, and Jürgen Kurths. Networks from flows—from dynamics to topology. *Scientific reports*, 4, 2014. <https://doi.org/10.1038/srep04119> PMID: 24535026
32. Berman Helen, Henrick Kim, and Nakamura Haruki. Announcing the worldwide protein data bank. *Nature Structural Biology*, 10(12), 2003. <https://doi.org/10.1038/nsb1203-980> PMID: 14634627
33. Dokholyan N. V., Li L, Ding F, and Shakhnovich E. Topological determinants of protein folding. *Proceedings of the National Academy of Sciences*, 99(13):8637–8641, 2002. <https://doi.org/10.1073/pnas.122076099>
34. Vendruscolo M., Dokholyan N. V., Paci E., and Karplus M. Small-world view of the amino acids that play a key role in protein folding. *Physical Review E*, 65:1–4, 2002. <https://doi.org/10.1103/PhysRevE.65.061910>
35. Di Paola L., De Ruvo M., Paci P., Santoni D., and Giuliani a. Protein contact networks: An emerging paradigm in chemistry. *Chemical Reviews*, 113:1598–1613, 2013. <https://doi.org/10.1021/cr3002356> PMID: 23186336
36. Estrada Ernesto. *The structure of complex networks—Theory and Applications*. Oxford University Press, 2011.
37. Hong Liu and Lei Jinzhi. Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity. *Journal of Polymer Science, Part B: Polymer Physics*, 47(2):207–214, 2009. <https://doi.org/10.1002/polb.21634>
38. Hagberg Aric A., Schult Daniel A., and Swart Pieter J. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11–15, 2008.
39. Molkenhthn Nora and Timme Marc. Scaling laws in spatial network formation. *Physical Review Letters*, 117(16):168301, 2016. <https://doi.org/10.1103/PhysRevLett.117.168301> PMID: 27792385
40. Molkenhthn N., Hu S., and Niemi A.J. Discrete nonlinear Schrödinger equation and polygonal solitons with applications to collapsed proteins. *Physical Review Letters*, 106(7):078102, 2011. <https://doi.org/10.1103/PhysRevLett.106.078102> PMID: 21405545
41. Estrada Ernesto and Hatano Naomichi. Resistance distance, information centrality, node vulnerability and vibrations in complex networks. In *Network Science*, pages 13–29. Springer, 2010. https://doi.org/10.1007/978-1-84996-396-1_2
42. Ren Jie and Li Baowen. Thermodynamic stability of small-world oscillator networks: A case study of proteins. *Physical Review E*, 79(5):051922, 2009. <https://doi.org/10.1103/PhysRevE.79.051922>