What geometrically constrained models can tell us about real-world protein contact maps

IOP Publishing Phys. Biol. 20 (2023) 046004 https://doi.org/10.1088/1478-3975/acd543

Physical Biology

OPEN ACCESS

RECEIVED
29 December 2022

REVISED
2 May 2023

ACCEPTED FOR PUBLICATION
11 May 2023

PUBLISHED
26 May 2023

PAPER

What geometrically constrained models can tell us about real-world protein contact maps

J Jasmin Güven[1,*], **Nora Molkenthin**[2,*,**], **Steffen Mühle**[3] **and Antonia S J S Mey**[1,**]

[1] EaStCHEM School of Chemistry, University of Edinburgh, Edinburgh, United Kingdom
[2] Potsdam Institute for Climate Impact Research, Potsdam, Germany
[3] Max-Planck Institute for Dynamics and Self-Organization (MPIDS), Am Faßberg 17, 37077 Göttingen, Germany
[*] These authors contributed equally to this work.
[**] Authors to whom any correspondence should be addressed.

E-mail: nora.molkenthin@pik-potsdam.de and antonia.mey@ed.ac.uk

Keywords: biophysics, geometric constraints, protein folding

Supplementary material for this article is available online

Abstract

The mechanisms by which a protein's 3D structure can be determined based on its amino acid sequence have long been one of the key mysteries of biophysics. Often simplistic models, such as those derived from geometric constraints, capture bulk real-world 3D protein-protein properties well. One approach is using protein contact maps (PCMs) to better understand proteins' properties. In this study, we explore the emergent behaviour of contact maps for different geometrically constrained models and compare them to real-world protein systems. Specifically, we derive an analytical approximation for the distribution of amino acid distances, denoted as $P(s)$, using a mean-field approach based on a geometric constraint model. This approximation is then validated for amino acid distance distributions generated from a 2D and 3D version of the geometrically constrained random interaction model. For real protein data, we show how the analytical approximation can be used to fit amino acid distance distributions of protein chain lengths of $L \approx 100$, $L \approx 200$, and $L \approx 300$ generated from two different methods of evaluating a PCM, a simple cutoff based method and a shadow map based method. We present evidence that geometric constraints are sufficient to model the amino acid distance distributions of protein chains in bulk and amino acid sequences only play a secondary role, regardless of the definition of the PCM.

1. Introduction

Proteins, the molecular machines of every living organism, perform vital tasks required for life to persist. These range from transport (e.g. haemoglobin) [1], signal transduction (e.g. rhodopsin) [2], immune responses (e.g. antibodies), and hormonal regulation (e.g. insulin) [3, 4]. All natural proteins are made of 20 different amino acids which dictate the 3D conformations the proteins adopt in order to function [5]. One of the great challenges has been to understand how the primary structure, i.e. the amino acid sequence, can lead to the fully folded functional protein, known as the protein folding problem [6]. The last 50 years have seen many different routes to computationally predict biologically active—or native—protein structures without having to solve a crystal structure of the protein [7]. This folding problem can loosely be grouped into two aspects. Firstly, can we predict the dynamics and kinetics of folding, i.e. the pathway from a nascent protein to its folded native state? Secondly, can we simply predict the folded structure without necessarily requiring information on what path was followed?

When looking at the folding kinetics and pathways often the famous funnel picture comes to mind in which a protein goes down a free energy funnel to its folded state [8, 9]. Historically different routes have been taken to try and generate an ensemble of folding trajectories to understand the kinetics and folding paths, ranging from lattice models [10–15], coarse-grained models [16, 17] to atomistic

molecular dynamics simulations [18]. The challenge here is that often the dynamics are slow and kinetically frustrated by metastable states on the folding path and therefore cannot easily be used to predict the resulting native state of a protein for a large number of sequences.

For the challenge of predicting a native structure without the necessity of also understanding the folding pathway, non-simulation-based approaches have dominated. The critical assessment of structure prediction, first run in 1994, allowed for a plethora of different structure prediction approaches such as Rosetta [19, 20], as well as more physics-based simulation approaches discussed above to be compared systematically and as a result facilitate the development of new methods. AlphaFold 2, a machine learning model for structure prediction, has revolutionised this field by providing models for native structures that have experimental accuracy [21, 22]. This model was trained on structural data from the Research Collaboratory for Structural Bioinformatics Protein Data Bank [23] (RCSB PDB). A clear step change is that it does provide a rich resource of structures to understand emergent patterns in proteins on a more fundamental level.

Often using 1- or 2-dimensional features to study properties of folded proteins can already give good insights into protein structure. One example of such a feature is protein contact maps (PCMs) [24–26]. A PCM is a matrix representation of spatially close amino acids, given a certain cutoff distance as determined typically from a protein structure. PCMs have shown promise in understanding protein folding patterns around for example the co-evolution of native contacts [27], incorporating information from 3D-folds [28], as well as revealing allosteric communication pathways [29, 30].

In the past, it has been shown that the topology of amino acids has a significant contribution to the folding mechanism of a protein, and as such, the resulting native structure and its corresponding PCM [17]. Therefore, using topological models, such as geometrically constrained models to generate simplistic maximally compact structures that resemble a protein's native conformations, is a valid approach [31].

From a physicist's perspective, it is interesting to understand the emergent behaviour, in terms of folding or structure, in the ensemble of all protein structures. For example, early works showed that for a set of 11 proteins, the relative contact order (CO) is correlated with the folding rate of these proteins [32]. The CO takes an average of the amino acid distance (we will introduce the amino acid distance later as s) between all pairs of amino acids in contact in the native state and normalises this according to the chain length of the protein.

Here, we are interested in understanding what underlying features of PCMs can reveal on the scale of the proteome rather than a set of individual proteins. As such, different heuristic models for generating simplified PCMs have been introduced in the past by Atilgan *et al* [33] and Bartoli *et al* [34]. Atilgan *et al* proposed a random shuffle model to generate artificial PCMs with similar properties to real protein models. To this end, Bartoli *et al* built a model for PCMs in which they assume ad hoc that proteins have a distribution of amino acid distances $P(s)$ that follows $P(s) \approx s^{-1}$. The *amino acid distance* s is the separation of two connected amino acids along the backbone chain and gives rise to the simplest implementation of a connection probability in a protein. Bartoli *et al* justified their approximation only heuristically in the sense that resulting PCMs have similar properties to real-world PCMs.

In this paper, we provide a new analytical approximation that explains why the heuristic of s^{-1} is a good initial assumption for modelling PCMs, and offer additional correction terms. The analytical approximation is derived from a 2D geometrically constrained model which allows for a closed-form expression of $P(s)$. From this, we can recover the approximate heuristic of s^{-1}, but it also provides further correction terms. We validate that the approximation fits amino acid distance distributions obtained from simulations of the 2D geometrically constrained model [35] it is derived from. Furthermore, we show that this approximation also results in a valid fit for maximally compact structures and their resulting PCMs computed from simulation data of a 3D version of the constrained model [31], and even to real-world proteins. This means that the analytical approximation is a good starting point for artificially modelling PCMs in the future. In the same way simplistic folding dynamic models such as lattice models and geometric constrained models have been used to look at folding properties [35, 36], the same is true for generating ensembles of PCMs of native proteins.

The paper is structured as follows: in section 2, we provide a brief overview of the geometrically constrained models used in our study. In section 3, we describe our approach to constructing PCMs using simulations from the 2D and 3D versions of the geometrically constrained model, as well as from structural data in the RCSB PDB database. In section 4, we present our analytical approximation for modelling the amino acid distance distribution $P(s)$ in the 2D version of the geometrically constrained model. Next, in section 5, we report our findings from applying the 2D analytical approximation to PCMs generated from 2D and 3D simulations of the geometrically constrained model, as well as to the structural data from the RCSB PDB database. Overall, this section showcases the utility of our analytical approximation for accurately generating PCMs, and highlights the novel insights we gain from this approach.

2. 2D and 3D geometrically constrained models

Different approaches have been used in the past for geometrical models describing protein folding, some of which derived characteristics of the secondary structure from constraints on bond and torsion angles [37–39] or on the formation mechanisms of the tertiary structure [31, 35], and others are modelled by self-avoiding random walks on lattices [36, 40]. Here we will focus on models representing maximally folded structures in 2D, and 3D as introduced in [35] and [31], respectively. The 3D geometrical model and the analytical approximation of its 2D analogue both build on the idea that inherently geometrical objects, such as amino acids, imply that any resulting PCM network ensemble modelling them has to be spatially embedded.

In [35], we introduced a simplified, 2D version of this geometrically constrained model. It starts from a closed chain (i.e. a ring) of L unit discs and subsequently adds links, such that connected discs touch, yet no discs intersect. The advantage of this model is that it can be approximated by a purely topological simplification, which can be treated analytically. Figure 1 illustrates the construction of the 2D geometric model. The 3D version gives rise to the geometric constraints directly, rather than through an approximation by means of a topological constraint.

The topological constraints in the resulting network model are: (a) new links always form between two units that are part of the same face of the graph (region enclosed by a cycle in the network). This prevents the overlapping of discs. (b) No links form across the outer face. This prevents the enclosure of a unit by less than six other units (which is geometrically impossible) such that (c) the maximum degree of each unit is six, as six is the maximum number of unit discs one central unit disc can touch. (d) Once connected by a link, pairs of units do not disconnect. In [31], this geometrically constrained model was extended to 3D spheres using simulations to generate compact artificial protein-like polymer structures.

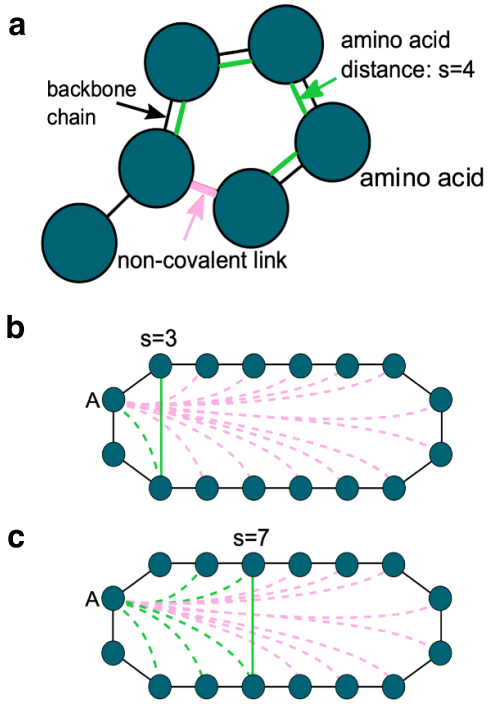

Figure 1. (a) The amino acid distance s (green) is defined as the number of amino acids (AAs) (here represented by a circle) along the protein backbone chain (black) between two connected AAs (pink). (b) Shows an example AA distance of length $s = 3$ (solid green line). When adding a new connection from node A, the previously created connection prohibits new ones (dashed pink lines). The green dotted link from A is still permitted and can be made. (c) Shows an example of an existing AS distance $s = 7$ (solid green), where there are now more permitted (dotted green) links and fewer prohibited (dotted pink) ones.

3. PCMs and shadow contact maps (SCMs)

For structural data of proteins, we are looking at two ways of generating PCMs. The first is a straightforward cutoff-based contact map, the second is based on SCMs as introduced by Noel, Whitford and Onuchic [28]. The complex interaction pattern between the amino acids in a protein can be naturally expressed as a network or graph, in which each amino acid is represented by a node and spatial proximity is encoded as a link. Whenever two central C_α atoms are closer together than a threshold d_c, they are connected, linked, or in 'close contact'. These connections can be determined from the 3D functional or folded structure of the protein and are encoded in a so-called (protein) contact map. This contact map is effectively an adjacency matrix $\mathbf{A}^{\text{struc}}$

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

The main idea behind SCMs is a shadowing radius r_S, which defines the size of the amino acids modelled as beads, in addition to the cutoff threshold d_c, which allows some non-physical contacts to be occluded. Namely, the shadow map excludes next-nearest neighbour contacts as well as contacts including an intervening atom. The SCM is constructed by first defining all contacts within the cutoff (similar to the PCM). Then, contacts are removed by considering if an atom falls within a 'shadow' produced by a second atom screening the light from the centre of a third atom [28]. Furthermore, contacts in the SCM are only recorded for distances of $|j - i| > 3$ for amino acids i and j.

3.1. Contact maps in a geometrically constrained protein model

Based on the simple observation that amino acids are objects in space that cannot overlap indefinitely, in [31] we introduced the 3D version of the

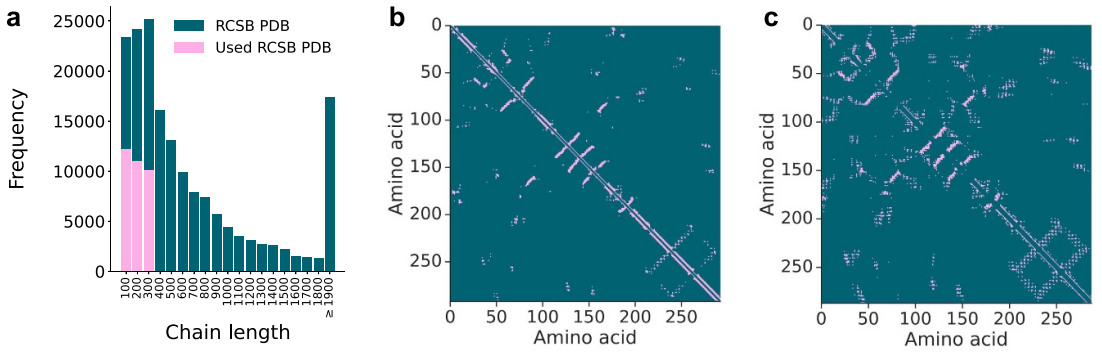

Figure 2. (a) RSCB PDB chain length frequencies. Pink bars show protein structures whose chain lengths match the resolution of the structure accounting for the whole sequence. Only pink RCSB PDB samples were used for the average contact map analysis. (b) Example of a protein contact map taken from RCSB PDB ID: 3UFG, with a cutoff of 8 Å. Pink squares indicate a contact present between amino acids. (c) Example of a shadow contact map constructed using an AlphaFold 2 predicted structure (UniProt ID: Q9PPK3) corresponding to the RCSB PDB structure shown in (b) with a distance cutoff of 8 Å and a shadow radius of 1 Å.

2D geometrically constrained model. Starting from a chain of identical spheres, each in contact only with the neighbours it is connected to, additional connections are made by randomly selecting two spheres and moving them towards each other until they touch. The new links formed that way cannot be broken in later steps and function as constraints on subsequent link formation steps. If the contact of the two selected spheres is geometrically impossible without breaking previously made connections or leading to overlaps of any spheres, the link is not made and is taken out of the pool of possible connections. This process is repeated until no more links can be formed without violating the geometric constraints and the final structure can be expressed as the adjacency matrix

$$A_{ij}^{\mathrm{sim}} = \begin{cases} 1, & \text{if } |i-j| = 1 \text{ or i and j connected} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

3.2. PCMs from structural protein data

The underlying formation mechanisms of individual protein folds are incredibly complex and depend on the surrounding solvent, as well as the specific amino acid sequence and their interactions. Here, we assess if the geometrically constrained protein model's contact map distributions capture real protein behaviour and make it a viable model for PCMs and provide a better explanation for the heuristic used by Bartoli *et al* [34]. We study the ensemble of PCMs from real protein structures and will look at 'averaged' contact maps over many different proteins that have different amino acid sequences, but their overall sequence or chain length is the same. We are specifically only interested in the contact maps and not the kinetics or folding paths of how the PCM was generated. We use the RCSB Protein Data Bank (RCSB PDB), a database of structurally resolved protein amino acid sequences through x-ray crystallography, nuclear magnetic resonance (NMR) or cryo-EM experiments. The RCSB PDB contains just over 200 000 protein structures to date [23]. The distribution of protein chain lengths in the RCSB PDB is illustrated in figure 2(a). These chain lengths vary from small fragments of less than 10 amino acids to large agglomerates of over 2000 amino acids. The most frequently occurring chain lengths, however, are between 100 and 300 amino acids as shown in figure 2(a). The pink subset of the bar plot shown in figure 2(a) is the set of structures used that fulfil the criterion on protein chain lengths of $L \approx 100$ ($85 \leqslant L \leqslant 115$) amino acids, $L \approx 200$ ($185 \leqslant L \leqslant 215$) amino acids and $L \approx 300$ ($285 \leqslant L \leqslant 315$) amino acids and represent all RCSB PDB structures used in our analysis of standard PCMs. The second database we use consists of synthetically generated structures using AlphaFold 2 with sequences taken from SwissProt [41] for both the standard PCM and SCM analysis. SwissProt is a manually curated database of protein sequences containing over 500 000 protein sequences, whose protein chain length distribution is shown in figure S2 in teal. The pink bars indicate the subset of AlphaFold 2 structures used for the second type of contact map analysis using the SCM. Only AlphaFold 2 per-residue confidence scores [21] of an average score of 90 or higher were used. Lastly, we removed structures that described the same protein but originated from different organisms, by selecting only unique protein names.

For the cutoff-based PCMs, RCSB PDB structures were downloaded in pdb format, read in with MDAnalysis version 2.0.0 [42] and only C_α were selected and used for the analysis. Proteins were categorised by length and placed into three groups. For

each of the sequences, all pairwise C_α distances were calculated, and two atoms were said to be in contact if their distance was less than 8 Å. See figure S7 of the supplementary material (SI) for different cutoffs of the contact maps. A cutoff of 8 Å is a typical value used in the literature [43]. The contacts (1 or 0) were then recorded in the adjacency matrix $\mathbf{A}^{\text{struc}}$. An example of a PCM is shown in figure 2(b).

We used AlphaFold 2 structures for the SCM approach as experimental PDB files consistently failed in the preprocessing step to define SCMs when using SMOG 2 [44] and SCM.jar [28]. As the structures in AlphaFold 2 are predicted from machine learning, no manual editing was required. Manual editing of the RCSB PDB (10 091 for $L \approx 300$, table S1 of the SI) structures which required loop modelling, or adding side chain information did not seem like a sensible choice. An example of an SCM is shown in figure 2(c).

We study the distribution of distances between connected amino acids in real and simulated proteins calculated from adjacency matrices through PCMs and SCMs. These distances were then histogrammed to calculate the mean amino acid distance distributions shown in section 5. The pseudo-algorithms (algorithm S1 and S2) describing the above two processes are shown in the SI. All code for the construction of the contact maps and their analysis is available on GitHub at [45]. The derived analytical approximation in section 4, was used for data fitting. All fits were carried out using SciPy 1.7.0 and with scripts for the fits available on GitHub at [45]. The secondary structure protein analysis was done using a local DSSP server [46], and the scripts for the processing of the DSSP output are available in the same GitHub repository [45].

4. Analytical approximation for the geometrically constrained model

To assess how we can best describe the distribution of the probability of amino acid distances $P(s)$ obtained from an average PCM for the geometrically constrained model, we consider the amino acid distance distribution for the simplified 2D model as proposed in [35]. We make the assumption that each distance is equally likely, from which we propose an analytical approximation for $P(s)$, derived from the geometric models, and schematically summarised in figure 1.

We introduce the auxiliary variable $F_k(s)$, defined to be the number of possible links with an amino acid distance of s if k links have been added before. Before any links are added, because we are looking for an average PCM over the whole proteome restricted to a particular chain length, we make the assumption that all amino acid distances are equally likely, so there are $F_0(s) = N$ possibilities of making a link of distance s each, with $2 \leqslant s < \frac{N}{2}$.

As we add more links, not only are links taken out of this pool, because they have already been realized, but each existing link can also geometrically prohibit other connections. Adding a link of length s_1 prohibits $2(s-1)$ links of length $s < s_1$ and $2(s_1-1)$ links of length $s \geqslant s_1$ (see figure 1(b)). To find the expected number of links still available in the second step, we average over all possible amino acid distances s_1 of the first link added.

The expected distribution of possible links after one step is thus given by the average over available link pools taken over all possible values of s_1:

$$\begin{aligned} F_1(s) &= \frac{1}{\frac{N}{2}-1}\sum_{s_1=2}^{N/2}\begin{cases} F_0(s)-2(s-1), \text{ for } s < s_1 \\ F_0(s)-2(s_1-1), \text{ for } s \geqslant s_1 \end{cases} \\ &\approx \frac{2}{N}\sum_{s_1=2}^{N/2}\begin{cases} N-2(s-1), \text{ for } s < s_1 \\ N-2(s_1-1), \text{ for } s \geqslant s_1. \end{cases} \end{aligned} \tag{3}$$

After $k-1$ links are made, the probability that the kth link, being randomly chosen from the pool of available links, has length s obeys

$$P^k_{\text{next}}(s) = \frac{F_k(s)}{C_k}, \tag{4}$$

where $C_k = \sum_{s=2}^{N/2} F_k(s)$. In subsequent steps, we use this probability to perform a weighted average over new links being added and get the same reduction (compare figure 1) but starting from the pool $F_{k-1}(s)$ of the step before rather than the initially available N links. The probability of a link of length s is thus the average overall added link probabilities from the first to the last added link. In 2D, this is repeated for $N-3$ steps until no more links can be added [35], giving

$$P(s) = \frac{1}{N-3}\sum_{k=0}^{N-4} P^k_{\text{next}}(s). \tag{5}$$

However, we must subtract fewer links to account for some of them having left the pool in earlier steps. To this end, we multiply the number of blocked links by

$$\frac{F_{k-1}(s)}{F_0(s)} = F_{k-1}(s)\frac{1}{N}. \tag{6}$$

This ignores any length-dependent bias in which links have been removed. Having considered some examples this assumption seems to hold well. This leads to the following recursion:

$$\begin{aligned} F_k(s) = \sum_{s_k=2}^{N/2} P^k_{\text{next}}(s_k)\Bigg(& F_{k-1}(s) - F_{k-1}(s) \\ & \times \frac{1}{N}\begin{cases} 2(s-1), \text{ for } s < s_k \\ 2(s_k-1)), \text{ for } s \geqslant s_k \end{cases}\Bigg), \end{aligned}$$

which generalizes (3) and simplifies to

$$F_k(s) = F_{k-1}(s)\left(1 - \frac{1}{N}\sum_{s_k=2}^{s} 2(s_k-1)P^k_{\text{next}}(s_k) - \frac{1}{N}\sum_{s_k=s+1}^{N/2} 2(s-1)P^k_{\text{next}}(s_k)\right)$$
$$= F_{k-1}(s)\left(1 - \frac{1}{N}\sum_{s_k=2}^{s} 2(s_k-1)P^k_{\text{next}}(s_k) - \frac{2(s-1)}{N}\left[\sum_{s_k=2}^{N/2} P^k_{\text{next}}(s_k) - \sum_{s_k=2}^{s} P^k_{\text{next}}(s_k)\right]\right)$$

and thus:

$$F_k(s) = F_{k-1}(s)\left(1 - \frac{2}{N}(s-1) + \frac{2}{N}s\sum_{s_k=2}^{s} P^k_{\text{next}}(s_k) - \frac{2}{N}\sum_{s_k=2}^{s} s_k P^k_{\text{next}}(s_k)\right). \quad (7)$$

Together with the initial condition $F_0(s) = N$, this expression constitutes an iteration rule from which $F_k(s)$ can be found for all k and s. Aiming to decouple those dynamics for different s, we make a heuristic guess $\tilde{P}^k_{\text{next}}(s_k)$ for $P^k_{\text{next}}(s_k)$. It consists of two separate approximations which are used in steps $k < a$ and steps $k \geqslant a$ respectively, where the threshold $1 \leqslant a \leqslant \frac{N}{2}$ is a free parameter. In the early stage, we use a uniform probability

$$\tilde{P}^{k\ll N/2}_{\text{next}}(s) \approx P^{k=0}_{\text{next}}(s) = \frac{2}{N}. \quad (8)$$

For later steps, the probability of a link to still being in the pool decreases with s. We thus approximate $P_{\text{next}}(s)$ with an ansatz for the $k-$average, namely, that it drops off as s^{-1}, leading to

$$\tilde{P}^{k\gg 1}_{\text{next}}(s) \approx \sum_{k=a}^{N-4} P^k_{\text{next}}(s) \approx \frac{s^{-1}}{H_{\frac{N}{2}} - 1}, \quad (9)$$

where $H_{\frac{N}{2}}$ is the harmonic number serving as a normalization factor. As we will see later, this approximation holds most precisely for intermediate values of k, with the $k-$average of $\tilde{P}^{k\gg 1}_{\text{next}}(s) \approx P(s)$, making it self-consistent. Errors for smaller and larger values of k are thought to approximately cancel out. Substituting (8) and (9) into (7) results in the following two recursions for the early and late evolution of the amino acid distance distribution respectively.

$$F_{k\ll\frac{N}{2}}(s) \approx F_{k-1}(s)\left(1 - \frac{2}{N}(s-1) + \frac{4}{N^2}s(s-1) - \frac{2}{N^2}(s^2+s-2)\right)$$
$$= F_{k-1}(s)(1 - f_{\text{low}}(s)), \quad (10)$$

where

$$f_{\text{low}}(s) = -\frac{2}{N^2}s^2 + \left(\frac{2}{N} + \frac{6}{N^2}\right)s - \left(\frac{2}{N} + \frac{4}{N^2}\right),$$

and

$$F_{k\gg 1}(s) \approx F_{k-1}(s)\left(1 - \frac{2}{N}(s-1) + \frac{2}{N}s\sum_{s_k=2}^{s}\frac{1}{s_k(H_{\frac{N}{2}}-1)} - \frac{2}{N}\sum_{s_k=2}^{s} s_k\frac{1}{s_k(H_{\frac{N}{2}}-1)}\right)$$
$$= F_{k-1}(s)\left[1 - \left(\frac{2}{N} - \frac{2}{N}\frac{H_s - 1}{H_{\frac{N}{2}}-1} + \frac{2}{N(H_{\frac{N}{2}}-1)}\right)s + \frac{2}{N(H_{\frac{N}{2}}-1)} + \frac{2}{N}\right]$$
$$= F_{k-1}(s)(1 - f_{\text{high}}(s)), \quad (11)$$

where

$$f_{\text{high}}(s) = \frac{2}{N}\left(\frac{H_{\frac{N}{2}} - H_s + 1}{H_{\frac{N}{2}} - 1}s - \frac{H_{\frac{N}{2}}}{H_{\frac{N}{2}} - 1}\right).$$

We can now use the recursive expressions (10) and (11) to write down a closed expression for $F_k(s)$, using f_{low} for the first a steps and f_{high} for the rest:

$$F_k(s) = N\prod_{i=1}^{a}(1 - f_{\text{low}}(s))\prod_{i=a+1}^{k}\left(1 - f_{\text{high}}(s)\right)$$
$$= N\left(\frac{1 - f_{\text{low}}(s)}{1 - f_{\text{high}}(s)}\right)^a\left(1 - f_{\text{high}}(s)\right)^k. \quad (12)$$

This can now be used to find an approximation for the amino acid distance distribution $P(s)$. By inserting (12) into (5), using (4), we can state that the probability distribution of the realized amino acid distances of all added links is then given by the average over the available pools at each link addition step.

$$P(s) = \frac{1}{N-3}\sum_{k=0}^{N-4} P^k_{\text{next}}(s)$$
$$\approx \frac{N}{N-3}\left(\frac{1 - f_{\text{low}}(s)}{1 - f_{\text{high}}(s)}\right)^a \times \sum_{k=0}^{N-4}\frac{1}{\langle C_k\rangle_k}\left(1 - f_{\text{high}}(s)\right)^k \quad (13)$$

where $\langle C_k\rangle_k$ approximates the individual C_k's as the average of C_k over k. This approximation allows the use of the geometric series for solving the equation:

$$P(s) \approx \left(\frac{1 - f_{\text{low}}(s)}{1 - f_{\text{high}}(s)}\right)^a \frac{\Gamma}{f_{\text{high}}(s)}, \quad (14)$$

where $\Gamma = \frac{N}{(N-1)}\langle C_k\rangle_k$ collects all constant factors and is used as a free fitting parameter, as the details of the evolution of C_k, as well as the role of dimensionality (3D vs. 2D) are unknown. The resulting expression in (14) by visual inspection of figure 3 resembles a power law, thus justifying the earlier approximation of $P^k_{\text{next}}(s) \approx \tilde{P}^k_{\text{next}}(s)$ introduced in (9).

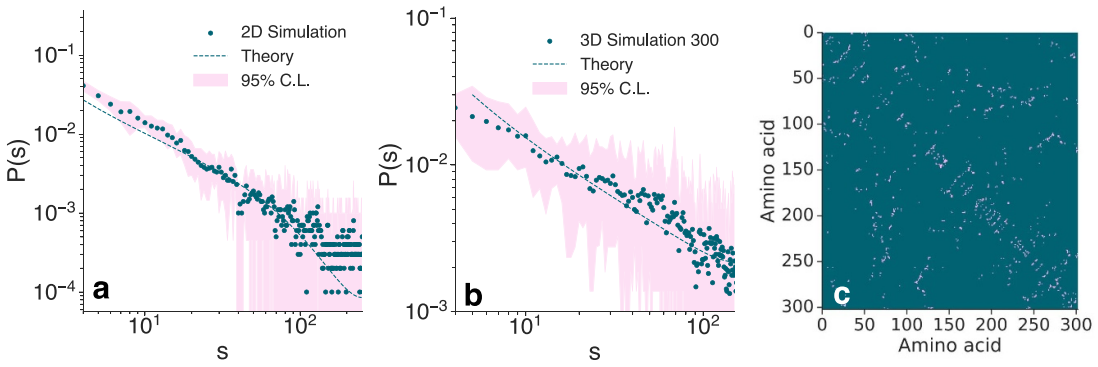

Figure 3. The analytical approximation fitted to amino acid distance distributions for (a) 2D and (b) 3D simulated data. The pink area shows the 95% confidence interval. (c) Example of a PCM taken from one of the 3D simulation runs. The pink squares indicate a contact between two nodes.

5. Results and discussion

Our analysis centres around understanding how the analytical approximation can be used to understand the PCMs generated from simulations in 2D and 3D from [31, 35], and protein structural data. We use RSCB PDB data for the standard PCMs and AlphaFold 2 data for SCMs. To this end, equation (14) is used as a fitting function in order to determine the two parameters Γ and a.

5.1. The analytical approximation fits PCMs generated from 2D and 3D simulation data

Figure 3 compares (14) with respect to (a) PCMs from 2D and (b) 3D simulation data introduced in [35] and [31], respectively. Figure 3(c) shows an example of an adjacency matrix $\mathbf{A}^{\text{sim}}$ as generated from a 3D simulation run. The 2D simulation data was generated from 10 repeats with chain lengths of 498 amino acids taken from [35]. The 3D simulation data consists of adjacency matrices computed from 30 simulation runs and was analysed as was explained in section 3. The amino acid distance distributions are plotted to show the mean frequencies of each amino acid distance up to $\frac{L}{2} \sim 250$ and $\frac{L}{2} \sim 150$ for 2D and 3D simulations, respectively. The shaded areas represent the 95% confidence intervals. The theoretical approximation from (14) was fitted to both 2D and 3D simulation data separately. The two parameters, a and Γ in (14), are given in table 1. See table S2 in the SI for the parameters for 3D simulation data in the $L \approx 100$ and $L \approx 200$ ranges. The fitted approximation captures the behaviour of the simulation data well for both 2D and 3D simulations.

5.2. The analytical approximation can describe the amino acid distance distributions of structural data from many proteins

To understand how the ensemble of 'real-world' PCMs behave and how the standard PCM and SCM distributions compare we looked at data both from the RCSB PDB and AlphaFold 2. See the SI for a detailed validation using the Kolmogorov–Smirnov statistics of how the AlphaFold 2 amino acid distance distributions have the same underlying distribution as RCSB PDB data when using mean confidence scores of 90% or more (see figure S4). The number of structures used for the analysis in each chain length range is shown in table S1 of the SI for both RCSB and AlphaFold 2 data. All RCSB PBD structures were converted into PCMs, as described in section 3. From the PCMs we computed the amino acid distance distribution in the same way as was done for the 3D simulations. All AlphaFold 2 structures were processed and converted to the SCM-derived adjacency matrix.

Table 1. Fit parameters from fitting the analytical approximation to the 2D simulated data as well as the 3D simulated data in the $L \approx 300$ chain length range.

Simulation	a	$\Gamma\ (10^{-3})$
2D	6 ± 1	0.3 ± 0.1
3D 300	1 ± 1	7.0 ± 0.2

In the first instance, we investigated if the amino acid distance distribution from the structural data from the RCSB PDB can be modelled by the analytical approximation. In order to answer this, we look at the scaling of the distribution of RCSB PDB data in the first instance.

Figure 4(a) shows the distributions of the amino acid distances for RCSB PDB logarithmically scaled for the different protein chain lengths of $L \approx 100$ (light purple circles), $L \approx 200$ (medium purple squares), and $L \approx 300$ (palatinate triangles). For all three lengths, the distributions show an approximate power law decay in their tail, which is consistent with the simulated results.

Moreover, we fit the analytical approximation to the RCSB PDB amino acid distance distribution in figure 4(b) in the $L \approx 300$ range. The theoretical fit is represented by a solid black line. The parameters given by the analytical approximation are $a = 4 \pm 1$

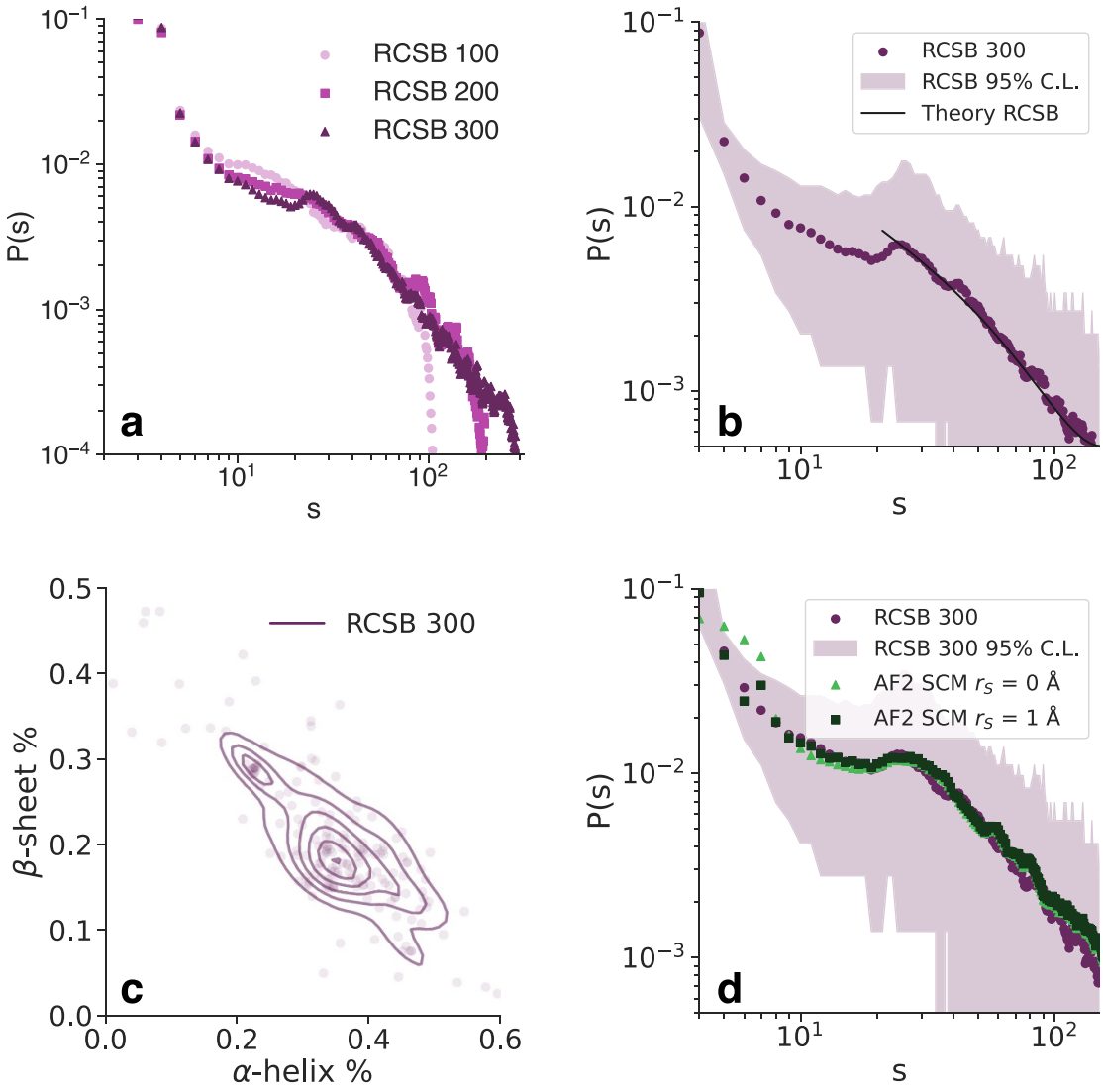

Figure 4. (a) Amino acid distance distributions for contact maps derived from the RCSB PDB database for protein chains in the ranges of $L \approx 100$, $L \approx 200$, and $L \approx 300$ amino acids. (b) Amino acid distance distributions of RCSB PDB for protein chains in the $L \approx 300$ range. The analytical approximation (solid line) is fitted to the tail of the data. The pink shaded area shows a 95% confidence level. (c) Secondary structure content distribution in the RCSB PDB structures in the $L \approx 300$ chain length range. The solid lines show the probability density function of the secondary structure content plotted with SciPy v.1.10.0. The scatter points show every 10th individual data point of secondary structure content in each structure in this length range. (d) Comparison between the amino acid distances from RCSB PDB and distances obtained from AlphaFold 2 using the SCM [28]. The light green triangles show the distribution of amino acid distances for a regular cutoff map with shadow radius of 0 Å and the dark squares show the distribution with a shadow radius of 1 Å.

and $\Gamma = (7.4 \pm 0.8) \cdot 10^{-4}$. The shaded area shows a 95% confidence level (C.L.). Similar fits for $L \approx 100$ and $L \approx 200$ results as well as amino acid distance distributions obtained from AlphaFold 2 structures are presented in figures S5(a), (b) and S3 of the SI, respectively.

We observe that for intermediate values of s, i.e. $30 \lesssim s \lesssim \frac{L}{2}$, the distributions are described well by the analytical approximation. However, the curves deviate for amino acid distances in the $5 \lesssim s \lesssim 30$ range (see figure 4(b)). Comparing this to the simulation distributions and their theoretical fits (figure 3), we see an under-representation of amino acid distances calculated from PCMs from RCSB PDB data.

To understand this under-representation better, we looked at the secondary structure content of the PCMs. The reasoning behind this is that the range in which the under-representation is observed is at a specific distance range, where contacts arise from the proteins' secondary structure elements. In figure 4(c), we show the secondary structure content distributions for proteins in the $L \approx 300$ chain length range from RCSB PDB. We see that proteins mostly contain secondary structures in the combined α-helix—β-sheet region. For protein chain length regions of $L \approx 100$ and $L \approx 200$ proteins with predominantly α and β-sheet regions exist, are shown in figures S5(c) and (d) of the SI, respectively.

Therefore we next looked at different thresholds used that define a contact between two amino acids. We changed the threshold value d_c from the original 8 Å to 10, 15 and 21 Å, see figure S4 in the SI. For 10 Å, the shape of the distribution was similar to that of in figure 4(b), but for 15 Å the under-representation starts to disappear, whereas for 21 Å it disappears almost completely. It is thus likely that with a threshold of 8 Å some inter-secondary structure contacts are being excluded, e.g. for large β-sheet content. Tuning this threshold value allows for the modulation of the under-representation of amino acid distances in the $5 \lesssim s \lesssim 30$ region. Compared to simulations where this is not observed due to the neglect of the existence of secondary structure.

Next we wanted to understand if a different definition of the PCM, namely using the SCMs will give a different overall behaviour. Figure 4(d) shows a comparison of amino acid distance distributions obtained from the PCMs and from the SCMs [28]. The SCM distributions were obtained from 6231 structures obtained from the AlphaFold 2 database in the $L \approx 300$ chain length range.

Figure 4(d) shows the RSCB PDB distribution in palatinate circles and the AlphaFold 2 SCM distribution with a shadow radius of $r_S = 1$ Å in dark green squares. In addition, we plot a 'standard' contact map with the AlphaFold 2 structures obtained from the SCM by setting the shadow radius to $r_S = 0$ Å. At the shorter distances of $s \approx 7$, the SCMs seem to produce more contacts than the PCM, but otherwise, the distributions overlap almost entirely. It is likely that the occlusion of contacts mostly only affects the shorter distances, as Noel, Whitford and Onuchich also suggest [28], and does not affect the intermediate distances, on which our analytical analysis is focused.

6. Conclusion

The tertiary structures of folded proteins have long been one of the most important mysteries of biophysics. While for individual structures, detailed molecular dynamics or statistical models are essential in approaching these questions, general statements for the ensemble of folded proteins can be made with much simpler models.

Here, we introduced and analysed the amino acid distance s and its probability distribution $P(s)$ for both real-world and geometrically constrained model simulations. For the equivalent model in 2D, we have used a mean-field approach to derive an analytical approximation for the amino acid distance distribution $P(s)$, and we show its agreement with the simulated geometrically constrained folding model and measured distributions from real-world proteins. Therefore we demonstrate that the geometrically constrained model's amino acid distance distribution $P(s)$ can match real-world data well. In addition, the derived analytical approximation can serve as a good basis to model and generate protein-like adjacency matrices as was done in [34]. It also highlights that the proposed heuristic of $P(s) \approx s^{-1}$ is a good starting point to describe amino acid distance distributions. However, here we managed to derive a more broad approximation for which a power law can be seen as a special case (14).

Gaining a better understanding of the ensemble of folded protein structures can help guide the way to a better understanding of the constraints within which structures may occur. Together with an understanding of secondary structure principles, such as that in [38, 39], this can help to narrow down the complex energy landscapes and find paths through them more effectively in the future.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/meyresearch/amino_acid_distance_distributions. All data generated or analysed during this study are included in this published article. All raw data and their analysis can be found on GitHub at [45], including information required for their reproduction.

Acknowledgment

We thank Marc Timme and Matteo Degiacomi for fruitful discussions.

Conflict of interests

The authors declare no competing interests.

ORCID iDs

J Jasmin Güven https://orcid.org/0000-0003-1555-0075
Nora Molkenthin https://orcid.org/0000-0002-7190-0714
Antonia S J S Mey https://orcid.org/0000-0001-7512-5252

References

[1] Ahmed M H, Ghatge M S and Safo M K 2020 Hemoglobin: structure, function and allostery *Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and Other Body Fluid Proteins, Subcellular Biochemistry* ed U Hoeger and J R Harris (Cham: Springer International Publishing) pp 345–82
[2] Nagata T and Inoue K 2021 Rhodopsins at a glance *J. Cell Sci.* **134** jcs258989
[3] Dill K A, Ozkan S B, Shell M S and Weikl T R 2008 The protein folding problem *Annu. Rev. Biophys.* **37** 289–316
[4] Dill K A and MacCallum J L 2012 The protein-folding problem, 50 years on *Science* **338** 1042–6
[5] Scheraga H A, Khalili M and Liwo A 2007 Protein-folding dynamics: overview of molecular simulation techniques *Annu. Rev. Phys. Chem.* **58** 57–83

[6] Nassar R, Dignon G L, Razban R M and Dill K A 2021 The protein folding problem: the role of theory *J. Mol. Biol.* **433** 167126
[7] Marks D S, Hopf T A and Sander C 2012 Protein structure prediction from sequence variation *Nat. Biotechnol.* **30** 1072–80
[8] Creighton T E 1990 Protein folding *Biochem. J.* **270** 1–16
[9] Dobson. C M 2003 Protein folding and misfolding *Nature* **426** 884–90
[10] Šali A, Eugene S and Martin K 1994 Kinetics of protein folding: a lattice model study of the requirements for folding to the native state *J. Mol. Biol.* **235** 1614–36
[11] Yue K, Fiebig K M, Thomas P D, Chan H S, Shakhnovich E I and Dill K A 1995 A test of lattice protein folding algorithms *Proc. Natl Acad. Sci.* **92** 325–9
[12] Dill K A 1985 Theory for the folding and stability of globular proteins *Biochemistry* **24** 1501–9
[13] Socci N D and Onuchic Je N 1994 Folding kinetics of protein like heteropolymers *J. Chem. Phys.* **101** 1519
[14] Go N 1983 Theoretical studies of protein folding *Annu. Rev. Biophys. Bio.* **12** 183–210
[15] Nobuhiro G 1983 Protein folding as a stochastic process *J. Stat. Phys.* **30** 413–23
[16] Clementi C 2008 Coarse-grained models of protein folding: toy models or predictive tools? *Cur. Opt. Struc. Biol.* **18** 10–15
[17] Clementi C, Nymeyer H and Onuchic J N 2000 Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins *J. Mol. Biol.* **298** 937–53
[18] Lindorff-Larsen K, Piana S, Dror R O and Shaw D E 2011 How fast-folding proteins fold *Science* **334** 517–20
[19] Rohl C A, Strauss C E M, Misura K M S and Baker D 2004 Protein structure prediction using rosetta *Numerical Computer Methods, Part D* (*Methods in Enzymology* vol 383) (New York: Academic) pp 66–93
[20] Baker D and Sali A 2001 Protein structure prediction and structural genomics *Science* **294** 93–96
[21] John J *et al* 2021 Highly accurate protein structure prediction with AlphaFold *Nature* **596** 583–9
[22] Kryshtafovych A, Schwede T, Topf M, Fidelis K and Moult J 2021 Critical assessment of methods of protein structure prediction (CASP)—Round XIV *Proteins* **89** 1607–17
[23] Berman H M, Westbrook J, Feng Z, Gary Gilliland T N B, Weissig H, Shindyalov I N and Bourne P E 2000 The protein data bank *Nucleic Acids Res.* **28** 235–42
[24] Vendruscolo M, Dokholyan N V, Paci E and Karplus M 2002 Small-world view of the amino acids that play a key role in protein folding *Phys. Rev.* E **65** 1–4
[25] Di Paola L, De Ruvo M, Paci P, Santoni D and Giuliani A 2013 Protein contact networks: an emerging paradigm in chemistry *Chem. Rev.* **113** 1598–613
[26] Estrada E 2011 *The Structure of Complex Networks - Theory and Applications* (Oxford: Oxford University Press)
[27] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T and Weigt M 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families *Proc. Natl Acad. Sci.* **108** E1293–301
[28] Noel J K, Whitford P C and Onuchic Je N 2012 The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function *J. Phys. Chem.* B **116** 8692–702
[29] Menichetti G, Fariselli P and Remondini D 2016 Network measures for protein folding state discrimination *Sci. Rep. UKF* **6** 30367
[30] Dokholyan N V, Lewyn Li, Ding F and Shakhnovich E I 2002 Topological determinants of protein folding *Proc. Natl Acad. Sci.* **99** 8637–41
[31] Molkenthin N, Mühle S, Mey A S J S and Timme M 2020 Self-organized emergence of folded protein-like network structures from geometric constraints *PLoS One* **15** e0229230
[32] Plaxco K W, Simons K T and Baker D 1998 Contact order, transition state placement and the refolding rates of single domain proteins *J. Mol. Biol.* **277** 985–94
[33] Atilgan A R, Akan P and Baysal C 2004 Small-world communication of residues and significance for protein dynamics *Biophys. J.* **86** 85–91
[34] Bartoli L, Fariselli P and Casadio R 2008 The effect of backbone on the small-world properties of protein contact maps *Phys. Biol.* **4** L1–L5
[35] Molkenthin N and Timme M 2016 Scaling laws in spatial network formation *Phys. Rev. Lett.* **117** 168301
[36] Mey A S J S, Geissler P L and Juan P G 2014 Rare-event trajectory ensemble analysis reveals metastable dynamical phases in lattice proteins *Phys. Rev.* E **89** 032109
[37] Bhattacharjee S M, Giacometti A and Maritan A 2013 Flory theory for polymers *J. Phys.: Condens. Matter* **25** 503101
[38] Danielsson U H, Lundgren M and Niemi A J 2010 Gauge field theory of chirally folded homopolymers with applications to folded proteins *Phys. Rev.* E **82** 1–5
[39] Molkenthin N, Hu S and Niemi A J 2011 Discrete nonlinear Schrödinger equation and polygonal solitons with applications to collapsed proteins *Phys. Rev. Lett.* **106** 078102
[40] Hills R D and Brooks C L 2009 Insights from coarse-grained Gō models for protein folding and dynamics *Int. J. Mol. Sci.* **10** 889–905
[41] Bairoch A 2000 The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000 *Nucleic Acids Res.* **28** 45–48
[42] Gowers R J *et al* 2016 MDAnalysis: a python package for the rapid analysis of molecular dynamics simulations *Proc. 15th Python in Science Conf.* pp 98–105
[43] Duarte J M, Sathyapriya R, Stehr H, Filippis I and Lappe M 2010 Optimal contact definition for reconstruction of contact maps *BMC Bioinform.* **11** 283
[44] Noel J K, Levi M, Raghunathan M, Lammert H, Hayes R L, Onuchic Je N and Whitford P C 2016 SMOG 2: a versatile software package for generating structure-based models *PLOS Comp. Biol.* **12** e1004794
[45] Güven J J, Molkenthin N, Mühle S, Antonia S and Mey S J S 2022 Amino acid distance distributions (available at: https://github.com/meyresearch/amino_acid_distance_distributions)
[46] Kabsch W and Sander C 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features *Biopolymers* **22** 2577–637