Statistical Potentials from the Gaussian Scaling Behaviour of Chain Fragments Buried within Protein Globules

Stefano Zamuner[1], Flavio Seno[2], and Antonio Trovato[3,*]

[1]Institute of Physics, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
[2]Department of Physics and Astronomy, University of Padova, Via Marzolo 8, I-35131 Padova, Italy
[3]INFN, Sezione di Padova, Via Marzolo 8, I-35131 Padova, Italy
[*]antonio.trovato@unipd.it

ABSTRACT

Knowledge-based approaches use the statistics collected from protein data-bank structures to estimate effective interaction potentials between amino acid pairs. Empirical relations are typically employed that are based on the crucial choice of a reference state associated to the null interaction case. Despite their significant effectiveness, the physical interpretation of knowledge-based potentials has been repeatedly questioned, with no consensus on the choice of the reference state. Here we use the fact that the Flory theorem, originally derived for chains in a dense polymer melt, holds also for chain fragments within the core of globular proteins. After verifying that the ensuing Gaussian statistics, a hallmark of effectively non-interacting polymer chains, holds for a wide range of fragment lengths, we use it to define a 'bona fide' reference state. Notably, despite the latter does depend on fragment length, deviations from it do not. This allows to estimate an effective interaction potential which is not biased by the presence of correlations due to the connectivity of the protein chain. We show how different sequence-independent effective statistical potentials can be derived using this approach by coarse-graining the protein representation at varying levels. The possibility of defining sequence-dependent potentials is explored.

Introduction

Proteins are linear flexible hetero-polymers, made up of 20 different natural amino-acid species[1]. Most natural proteins in solution have roughly compact shapes, and thus are usually referred to as globular proteins. The fundamental fact about globular protein sequences is their ability to attain a compact native three-dimensional folded conformation in physiological conditions[2].

The biological functionality of proteins is intimately related to their native structures and to the dynamical properties encoded in them[3]. Quantitative theoretical modeling requires in principle a detailed description at atomic level, for example to take accurately into account the subtle yet dramatic effects that can be brought about by a single residue mutation.

On the other hand, processes such as protein self-assembly and aggregation, involve time scales and system sizes which are currently unattainable by atomistic models[4]. Several schemes were thus developed to coarse-grain the representation of protein structures, and of the physical interactions between the representing entities, at a low resolution level[5].

The surprising success of coarse-graining approaches in computational protein science is related to the presence of robust qualitative emergent properties in protein systems, amenable to prediction by low resolution models[6]. For example, the native topology both shapes equilibrium fluctuations and determines folding and unfolding pathways, allowing for successful predictions by structure-based coarse-grained models[7].

An even more remarkable example of successful coarse-graining is the use of statistical potentials, as both effective interaction potentials to be used in folding simulations[8–10], and scoring functions employed in different contexts such as protein structure and function prediction[11], "de novo" protein design[12], model quality assessment[13, 14], aggregation propensity prediction[15–17], protein-protein interactions[18–22], prediction of binding affinities and of stability changes upon mutations[23–26], and many others. Statistical "knowledge-based" potentials can be introduced at different coarse-graining levels, including atomic resolution (in this case, the coarse-graining is due to solvent molecules being integrated out). They renounce a physics-based description of the effective interactions between representing entities; interactions are instead parameterized using the statistics empirically collected from the Protein Data Bank[27] (PDB).

In paradigmatic examples[28, 29], "contact statistical potentials" evaluate the effective interaction between a pair of amino acid residues based on the observed frequency of contacts between that pair in PDB structures. This approach can be generalized to many different observables, such as solvent accessibility, backbone dihedrals, orientation-dependent or many body interactions[30–34]. The conversion of empirical frequencies into an energy function is normally done employing Boltzmann inversion, as originally suggested by Sippl for pairwise "distance dependent potentials", in analogy to the pairwise potentials of mean force[35].

A crucial element in the definition of statistical potentials via Boltzmann inversion is the choice of a "reference state". The probability distribution observed in the latter is used to normalize the statistics collected over the PDB structures for a given residue pair. The reference state should then be taken as an ensemble of protein-like structures where no specific direct interactions between amino acids are present. A simple choice is to consider the ensemble of all residue pairs from the PDB structures[36], but still many different recipes are possible to define the reference state[37–40]. Beside the uncertainty in the reference state definition, the very use of Boltzmann inversion for the statistics collected from different PDB structures was extensively debated[41,42], in particular with reference to chain connectivity. The Boltzmann inversion has been justified by using information theory arguments within a Bayesian approach[43]. In this context, statistical potentials are considered as statistical preferences that can be obtained "a posteriori" from empirical data, whereas the reference state contains the "a priori" information about the system.

In this work, we propose to use a reference state for deriving pairwise distance dependent potentials based on purely polymer physics considerations. Our strategy can be used at different levels of coarse-graining.

In particular, we use the fact that a properly filtered data set of protein "fragments" from PDB structures exhibit Gaussian statistics, the one expected for ideal chains in the absence of any interaction. This property had been already uncovered by Banavar et al.[44], who found that fragments buried within globular proteins obeys the same Flory theorem[45,46] derived for polymer melts, that is concentrated solutions of different chains. The same theorem has been shown to hold for fragments buried in the interior of single compact polymer chains, when selected with appropriate constraints[47]. The Flory theorem states that, for polymer chains from within a dense polymer melt, excluded volume repulsion is effectively canceled by solvent-mediated attractive forces between the monomers. Therefore the chains exhibit statistics which are characteristic of random walk behavior.

The first purpose of this work is to confirm the existence of a Flory regime for buried protein fragments when a much larger data set of proteins is considered. We then take advantage of this fact by using the ensuing Gaussian reference state in order to obtain an unbiased estimation of a distance dependent effective interaction potential between aminoacids[36,43,48]. The statistical potentials estimated with this strategy could be either sequence independent or sequence dependent.

In the first part of the paper, by analyzing a data set of 7500 non-redundant proteins, we confirm that the Flory theorem holds for compact native structures with a good degree of accuracy. This is achieved by showing that the properly rescaled distributions of the fragment end-to-end distances collapse to the same Maxwell distribution when fragment lengths larger than $m_{min} = 70$ and smaller than $N^{2/3}$, where N is the length of the protein chain, are considered. The upper cut-off is introduced in order to select buried fragments[47]. The lower one is instead necessary to achieve a uniform Kuhn length. Our results extend the findings of Ref.[44], showing that the Gaussian scaling holds for fragments in a larger range of sizes, provided a non uniform Kuhn length is considered.

As a consequence of the validity of the Flory theorem, we can assume that the excluded volume repulsion is effectively canceled by solvent-mediated attractive forces between the monomers. We therefore interpret systematic deviations from the expected Gaussian behavior which are visible in the short range regime as an effective intra-molecular interaction.

Therefore, we devoted the second part of this manuscript to exploit the feasibility of this idea, by deriving the effective potential for "all" protein fragments at different coarse-graining levels: CA-based, all heavy atoms, all atoms (including hydrogen atoms). The derived potentials consistently change in the three cases. In particular, a power law repulsive term is present at short length whereas the potential vanishes at 20 Å in all cases. Well defined minima with negative energy are present for the atomistic resolutions for distances compatible with the sum of Van der Waals radii or with hydrogen bond geometry.

If the analysis is repeated by classifying the protein segments according to the amino-acid types which occupy the first and the last position along the fragment, we can have a direct measure of the effective interactions between amino-acid types, as a function of the distance. This method could be of great interest for a wide range of applications in protein physics.

Results

In order to assess the hypothesis that the Flory theorem holds for fragments buried in the interior of globular proteins, we analyzed a large data-set of 7500 globular proteins. This proteins ensemble was obtained by refining the TOP8000 data-bank[49] after removal of the non globular structures as explained in Methods. In data-set pruning, each protein is represented as a polymer whose monomers are placed in the C_α atomic position of the N amino-acids. The logarithmic plot of the radius of gyration of these polymers versus their length N is shown in Figure S1 for the full TOP8000 data-bank. The proteins in the final pruned dataset (highlighted in Fig S1) have been selected in such a way that their radius of gyration scales as $N^{1/3}$, as expected for globular proteins.

Long enough buried protein fragments follow Gaussian statistics: the thermal exponent

To investigate the validity of the Flory theorem, we analyzed an ensemble of protein fragments of different lengths, extracted from the pruned database. For any given chain of length N we considered only fragments with length $m < N^{2/3}$ so that they belong to a part of the protein which is likely to be far from globule boundaries and thus buried within the globule interior[47]. Figure 1a shows in a logarithmic scale the behavior of the average end-to-end distance of such fragments as a function of their length m, when the end-to-end distance is evaluated using C_α backbone atoms.

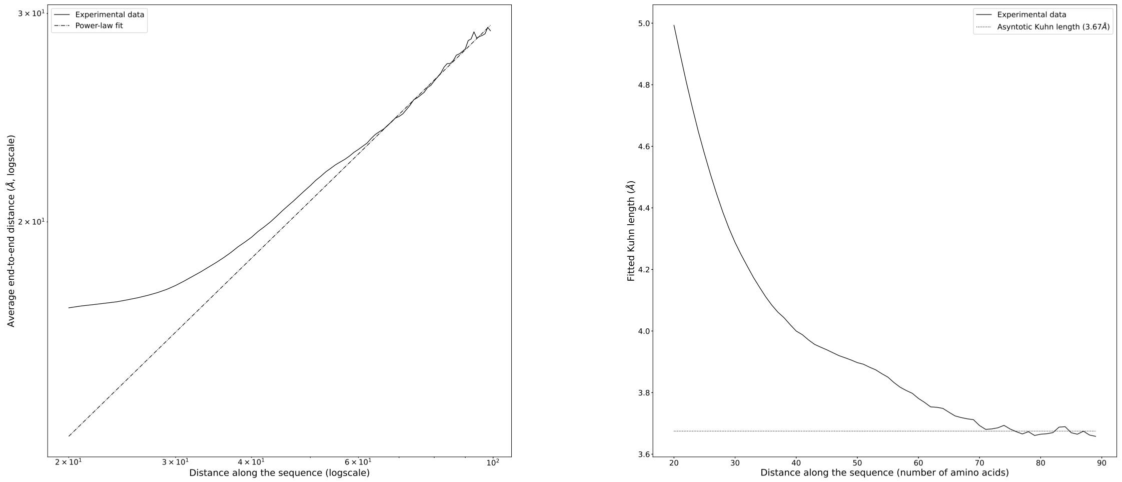

Figure 1. (**a**) Log-log plot of the average CA end-to-end distance R of protein fragments versus fragment length m. The plot was obtained by averaging over all fragments of length m from the data set selected as shown in Figure S1. For any given m, R was determined as the average over all fragments of that length in proteins whose overall lengths are larger than $m^{\frac{3}{2}}$, in order to consider only fragments likely to be buried in the globule interior[47]. The error bars are of the order of the size of the symbols. The Flory regime, e.g. $R \sim m^{\frac{1}{2}}$ is reached when $m > 70$. For $m \geq 90$ the statistical analysis deteriorates due to the fast decrease of available data with increasing m. (**b**) The Kuhn length b, obtained by maximizing the likelihood to Maxwellian distributions (3) of the empirical CA end-to-end distance data, plotted versus the length m of the protein fragments considered in the statistical analysis. The values of b decrease monotonically and reach a plateau in the region $70 < m < 90$. The plateau uniform value is estimated to be $b^* = 3.67$ Å. Only in this region all the rescaled empirical distributions collapse (see Figure 3), thereby showing the existence of the Flory regime. The number of fragments in the ensembles which are analyzed decreases with m as well. For $m \geq 90$ the ensemble population becomes too small to allow good estimations.

The Flory regime requires a scaling law $R \sim N^\nu$ with $\nu = \frac{1}{2}$. Our data for CA end-to-end distances shows that such a regime is valid only for the longer fragments, approximately when $m > 70$. This can be explained by the presence of secondary structures that introduce a strong bias in the scaling behavior for short chains. This behavior can be understood by looking at Figure S2 which shows the average tangent-tangent correlation function as a function of sequence separation along the chain. For short sequence separations the direction of the chain is highly correlated reflecting the existence of short, effectively one-dimensional, rigid motifs such as α-helices and β-strands. On the other hand, the sharp anti-correlation minimum at $m = 13$ reveals a bending propensity in the opposite direction. This finds its counterpart in the almost flat behaviour of the average end-to-end distance in Figure 1a for $m \gtrsim 20$. This picture is confirmed by noticing that the value at which the correlation function reaches again zero ($m \sim 25$) is almost twice as much as the value at the anti-correlation minimum, suggesting that for $m \sim 25$ protein chains are expected to loop back on themselves significantly more than for other values of sequence separation. In fact, the above observation is consistent with the peak described in[50] for the probability of loop formation. This analysis suggests that the Flory regime can not be observed for short fragments because of the effects induced by secondary structures.

Long enough buried protein fragments follow Gaussian statistics: Maxwellian distributions for end-to-end distances

In order to investigate in more detail the existence of a Flory regime we studied whether the end-to-end distance for fragments of length m follows the Maxwell-Boltzmann distribution described by

$$\mathscr{M}_m(R,b) = \frac{4\pi R^2}{(\frac{2}{3}\pi b^2 m)^{\frac{3}{2}}} \exp(-\frac{3R^2}{2b^2 m}) , \tag{1}$$

where the scale parameter b refers to the distance between consecutive monomers in an ideal Gaussian chain, namely the Kuhn length of the polymer. The distances between residues are computed in three different ways: as the distance between their alpha carbons, as the minimum distance between any atom of the two residues and as the minimum distance between any heavy atom of the residues. We will refer to these three different levels of coarse-graining as CA (carbon-alpha level), HH (hydrogen atoms level) and HV (heavy atoms level) respectively. For all three coarse-graining levels of description, we fitted the Kuhn length b of a Maxwell-Boltzmann distribution to maximize the likelihood that the empirical data have been drawn from it. This is done separately for any given m, so that the optimized Kuhn length $b\,(m)\,\sqrt{m}$ depends on fragment length m. In Figure 1b we plot the optimized b as a function of m for CA end-to-end distances. The values of b decrease towards a plateau beginning approximately at $m = 70$. For $m > 90$ the values of b change dramatically as a consequence of the poorer statistics (see Table 1). At the plateau we estimate $b(m) = b^* = 3.67$ Å for CA end-to-end distances. Similar results are obtained for HV and HH as well, as shown in Figure S3. The plateau values estimated for the Kuhn length are $b^* = 3.35$ Å for HV and $b^* = 3.27$ Å for HH.

Table 1. Number of buried ($m < N^{2/3}$) fragments in the pruned (see Figure S1) dataset as a function of fragment length.

fragment length	**number of fragments in the dataset**
20	1640293
30	1352751
40	991104
50	563178
60	276421
70	120448
90	24967

Figure 1b shows that the estimated Kuhn length is higher than in the plateau for lower values of m. This could explain the discrepancy with the higher value ($b^* = 3.75$ Å) obtained in[44], as a different range of values of m was used in that work.

Empirical probability distributions are inferred by raw data using Kernel Density Estimation (KDE), with kernel bandwidth estimated separately for each m by a maximum likelihood approach through a cross-validation procedure. The whole methodology is explained in detail in Methods.

In Figure 2, the empirical CA end-to-end distance probability distributions for four different fragment lengths ($m = 30,50,70,90$) are shown together with their best Maxwellian fit. Similar plots for HV and HH end-to-end distance distributions are shown in Figures S4 and S5. The competing effects of increasing m and of $b\,(m)$ decreasing with m are both visible in Figure 2.

It is interesting to observe that, as shown in Figure 2 and in Figures S4, S5, the Gaussian behavior of the end-to-end distances is mostly preserved even in a broader range of fragment lengths, $30 < m < 90$, but since $b(m)$ is not uniform for $m < 70$, we cannot talk about a Flory regime in that range. Nevertheless, the existence of a Gaussian reference state can be fruitfully exploited to derive an effective statistical potential in the full $30 < m < 90$ range.

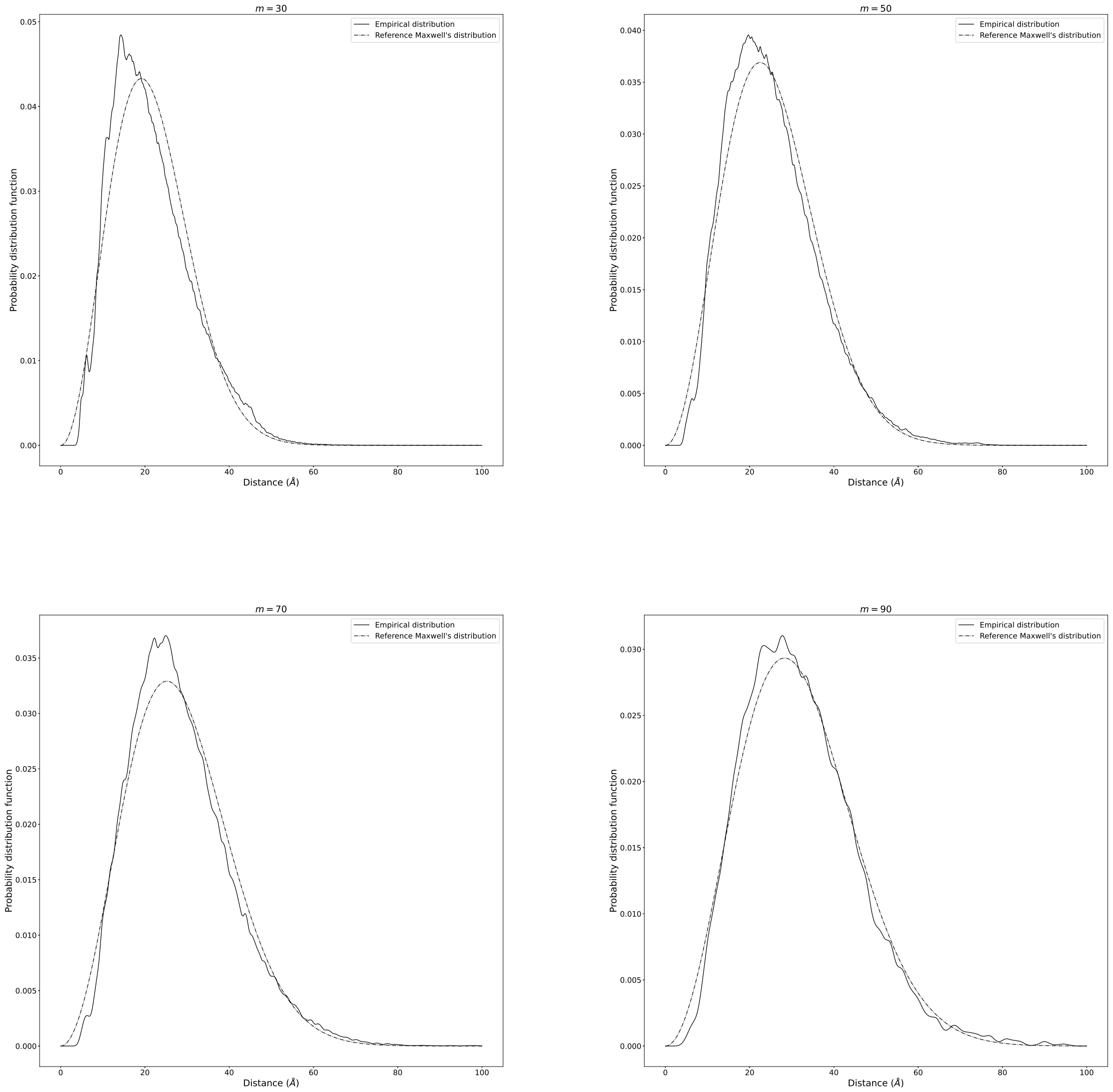

Figure 2. The CA end-to-end distance probability distributions for four different fragment lengths ($m = 30, 50, 70, 90$) are shown together with their best fits to the Maxwell distribution (1). The parameters b used in the plots are obtained maximizing the likelihood that the empirical data belong to the Maxwell distribution (1). The value of b decreases with m and reaches a plateau for $70 < m < 90$, corresponding to the Flory regime (see Figure 1b).

In the region with uniform Kuhn length b ($70 < m < 90$) empirical distributions can be collapsed by rescaling the end-to-end distances by $\sqrt{m}$ and multiplying the probability distributions by the same quantity, as shown in Figure 3 for CA. This graph vividly shows the existence for globular proteins of a range of fragment lengths, in which their statistics is well described by Gaussian ideal chains with a uniform estimated Kuhn length close to the average distance between consecutive C_α's atoms. A data collapse of similar quality can be obtained for both HV and HH, as shown in Figure S6. Figure S6 also shows that the collapsed empirical distributions are more skewed with respect to the reference Maxwell distribution in the HV and HH cases.

Statistical potentials with a Gaussian reference state: sequence independent effective interaction

The Maxwell distribution (1) fits very well CA experimental data for large values of end-to-end distance, when the full cancellation of competing interactions, e.g. attractive and excluded volume, is effectively occurring. For short distances however, as we can see from Figure 3, there are important deviations from the ideal distribution. These are expected, since for an ideal chain excluded volume is absent even at short range, whereas for real protein chains it is anyway present. We then propose to use deviations from the ideal Gaussian behavior as a proxy of the effective short range interactions between protein residues.

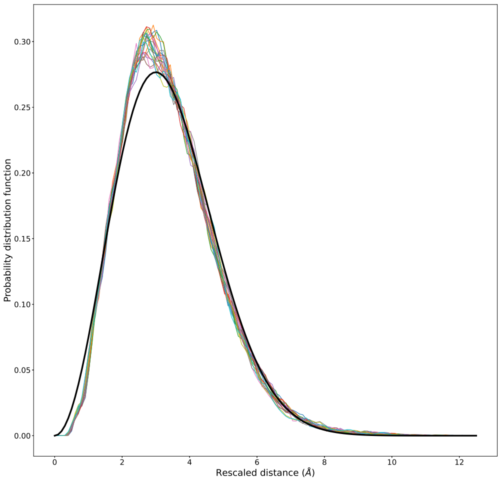

Figure 3. The rescaled empirical probability distributions as a function of the rescaled length $R/m^{1/2}$ for $70 < m < 90$. All curves collapse rather well together. The reference Maxwell distribution (1) evaluated for the plateau scale parameter $b^* = 3.67$ Å is shown for comparison. A significant deviation appears only for small distances and is due to the effect of excluded volume that at very short range can not disappear for real protein chains. It is worth to notice that, despite the presence of secondary structures, the value of b is very close to the average distance between consecutive C_α's atoms.

To test this hypothesis, we define through Boltzmann inversion a sequence independent statistical potential for any given fragment length m, as minus the logarithm of the ratio between the empirical probability density (already shown in Figure 2 and Figure 3 for different fragment lengths) and the reference Maxwell distribution:

$$V_m(R|b,w) = -\log\left(\frac{\mathscr{E}_m(R,w)}{\mathscr{M}_m(R,b)}\right) ; \tag{2}$$

where $\mathscr{E}_m(R,w)$ is the empirical end-to-end distance distribution for fragments of length m, obtained with KDE using a kernel bandwidth w (see Methods for details). Equation 2 highlights the dependence of such potential on the scale parameter b used for the reference state and on the kernel bandwidth w used to obtain the empirical distribution. It is worth noting that both parameters are obtained through a maximum likelihood approach.

We plot in Figure 4 the effective potentials $V_m(R)$ in the Flory regime $70 < m < 90$ in which $b(m) = b^*$ is uniform, for all coarse-graining levels used in this work. Remarkably, the curves for different fragment lengths m collapse nicely together, allowing to recover well defined effective potentials $V^*(R)$ that we define as the average of the potentials obtained from Equation (2) over all fragment lengths $70 < m < 90$ in the Flory regime (the resulting average potentials are shown in Figure S7 for all coarse-graining levels, along with the corresponding standard deviation). Such result pinpoints the existence of a robust underlying mechanism which is revealed by using the ratio between the empirical and the ideal reference distributions and which can allow for a consistent estimate of amino-acids interactions. Even more remarkably, we observe that while the empirical end-to-end distance distributions collapse upon rescaling (see Figure 3), deviations from the Maxwellian reference state collapse when rescaling back to the physical distance values (see Figure 4). For example, Figure S6 clearly shows, for the HH and HV cases, how the position of sharp small peak at short distances, that determines the main minimum of the statistical potential $V^*(R)$, drifts upon changing m when using rescaled distances.

The effective statistical potentials $V^*(R)$ differ significantly depending on the coarse-graining level, as in fact expected for physics-based interactions. The potentials obtained when considering all atoms, either with (HH) or without (HV) hydrogens share in fact similar features: a steep short range repulsive part and a series of well defined attractive minima with decreasing depth for increasing distance (see Table 2). At large distances the potential vanishes towards zero, although shallower minima can be still identified (see Table 2). However, in the more coarse-grained HV case, the first minimum is partially smoothed out and the barrier separating the first two minima becomes repulsive. In the even more coarse-grained CA case, the minima of the statistical potential get much more smoothed out and the potential becomes repulsive for all distances.

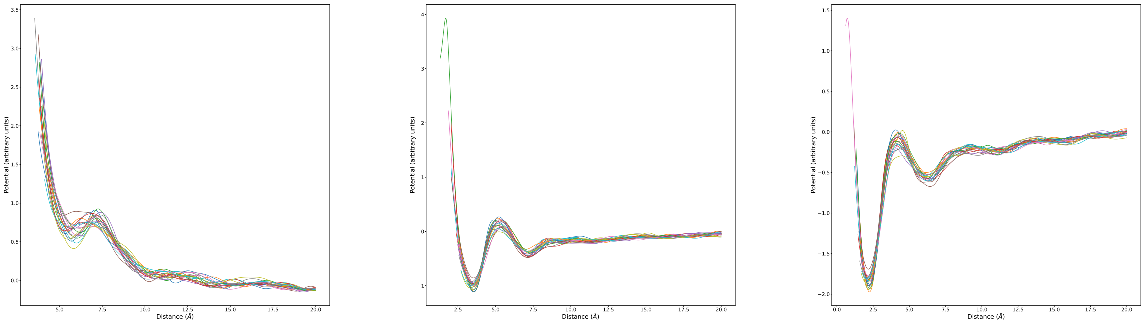

Figure 4. Effective potential $V_m(R)$ estimated for each $70 < m < 90$ in the Flory regime using Equation (2). Remarkably, the curves obtained with this procedure do not depend on the fragment length and can therefore be interpreted as an effective potential between the terminal fragment residues. In this case, where all fragments are considered regardless of the type of amino acids at their ends, the potential can be interpreted as a generic sequence independent interaction between all residues.

The positions and depths of the minima of the statistical potentials for different coarse-graining levels are reported in Table 2. Minima features are extracted using the $V^*(R)$ potential obtained in the Flory regime $70 < m < 90$. We observe that the deepest minimum in the HH case (2.21 Å) corresponds to twice the Van der Waals radius 1.1 Å for the hydrogen atom[51], whereas the deepest minimum in the HV case (3.54 Å) is within the distance range observed between donor nitrogen and acceptor oxygen atoms for hydrogen bonds occurring in proteins[52].

Table 2. Positions and values of the minima of the average effective statistical potential $V^*(R)$ for different coarse-graining levels.

	position (Å)			**value ($\kappa_B T$)**	
HH	**HV**	**CA**	**HH**	**HV**	**CA**
2.21	3.54	5.81	-1.91	-1.07	0.53
6.34	7.16	10.59	-0.61	-0.46	0.03
11.01	9.35	12.03	-0.26	-0.21	0.02
14.90	11.44	14.92	$-0,13$	-0.21	-0.08
15.69	15.90	19.55	-0.13	-0.12	-0.13
18.95	17.63		-0.06	-0.11	

In order to study the short distance repulsive behavior of the effective potential, more statistics is needed at very short distances. To this aim, we then consider the average $\overline{V}(R)$ of the statistical potentials defined by Equation (2), taken over the wider range of fragment lengths, $30 < m < 90$, for which the rescaled end-to-end distance distributions are close to Maxwellians (see Figure 2). The reference state is thus now defined with a non uniform Kuhn length $b(m)$. The quality of the collapse of the different $V_m(R)$, $30 < m < 90$, worsens, yet is still acceptable, as shown in Figure S8 for all coarse-graining levels.

The sequence independent effective potential is plotted in logarithmic scale in Figure 5 for the CA case, together with a linear regression fit for the short range region.

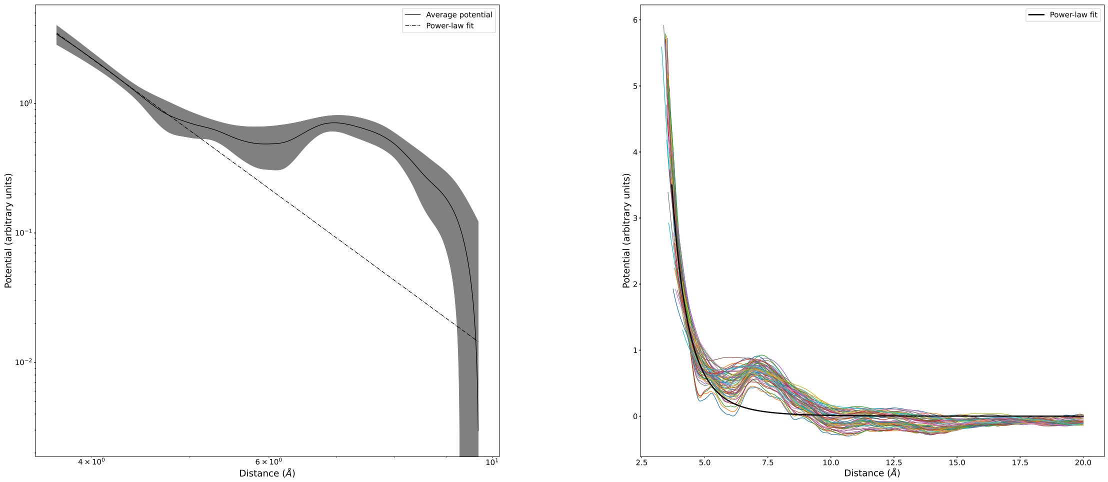

Figure 5. Potential $\overline{V}(R)$ for the CA case, obtained when averaging the statistical potentials (2) over all fragment lengths $30 < m < 90$, shown together with the power-law fit at short distances. The exponent estimate is -5.7 ± 0.3. (a) log-log scale; the standard deviation is also shown. (b) linear scale with data collapse of all $V_m(R)$ potentials for different values of sequence separation $30 < m < 90$.

The existence of a power law behavior seems clear. The resulting estimate for the exponent is -5.7 ± 0.3, which might be related to the presence of distinctive dipole-dipole interactions. Nevertheless, we caution that the above exponent estimate may depend on the limited range of short distances that can be probed with the available statistics. As a matter of fact, the use of a dipole-based description for peptide groups was already successfully proposed to perform coarse grained simulations of protein folding[53].

Statistical potentials with a Gaussian reference state: sequence dependent effective interactions

The analysis carried out in the previous subsection can be repeated by splitting the full data set according to the specific amino acid types found at the end of the considered protein fragments. The resulting statistical potential should be interpreted as an effective interaction between the terminal residues. The decreased statistics, unfortunately, pushes our approach to its very limit, even when considering the average potential $\overline{V}(R)$ over all sequence separation values $30 < m < 90$ and a reference state with variable b.

For completeness, we report in Figure 6 some examples of average statistical potentials $\overline{V}(R)$ derived in our approach for the CA case, involving two cysteine residues (CYS-CYS), two small non-polar residues (ALA-ALA), two charged residues (GLU-GLU) and two hydrohobic (LEU-LEU) residues.

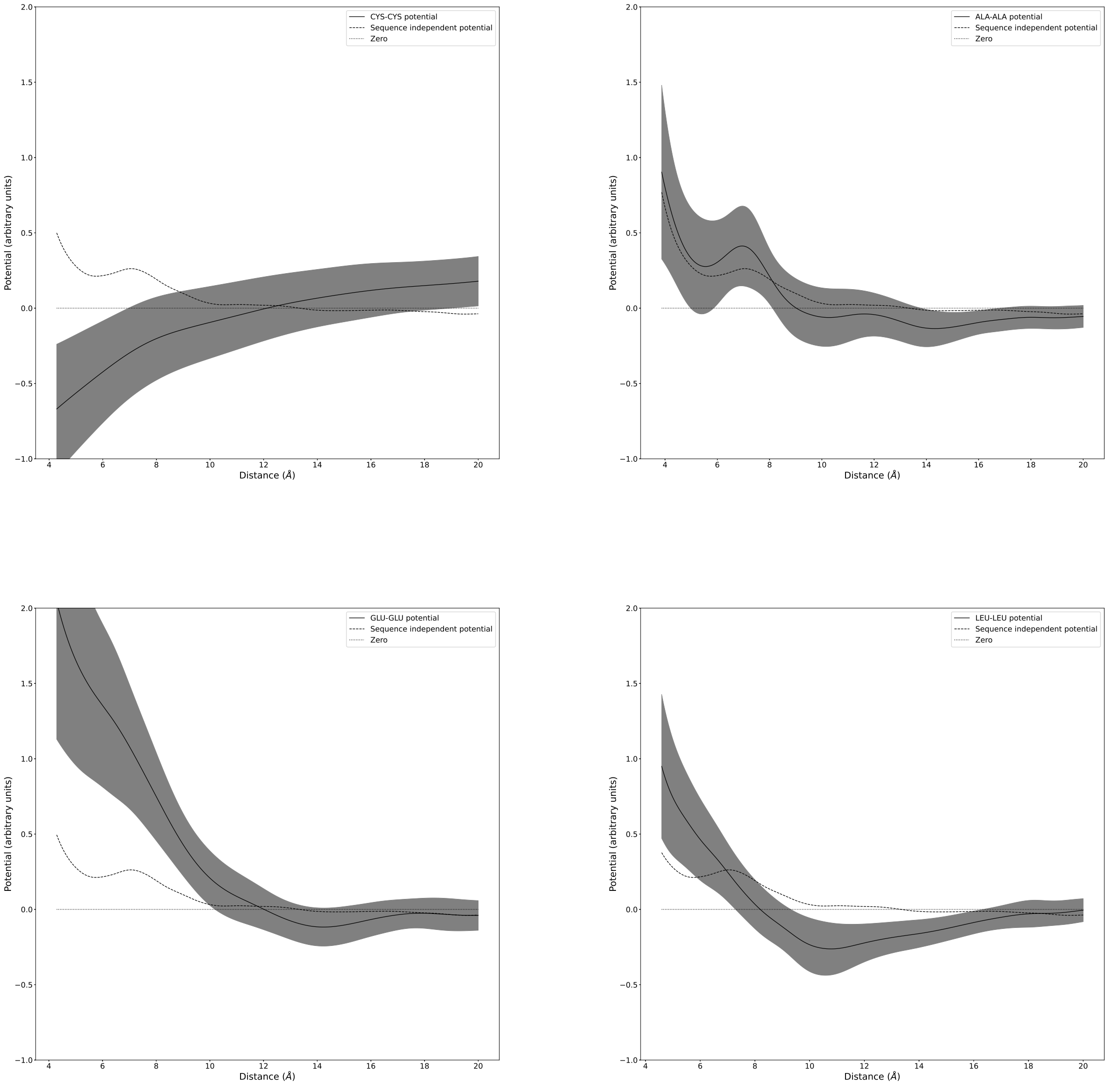

Figure 6. Examples of sequence-dependent potentials $\overline{V}(R)$ for the CA case, obtained when averaging the statistical potentials (2) over all fragment lengths $30 < m < 90$, with its standard deviation (gray areas). The sequence independent potential (dashed line) is shown as a reference. (a) Cysteine-Cysteine (b) Small non polar residues ALA-ALA. (c) Two negatively charged residue GLU-GLU. (d) Two hydrohobic residues (LEU-LEU)

.

It appears clear that the sequence-dependent potential can differ significantly from the average sequence-independent one. It is also interesting to notice how the obtained potentials reflect the physical-chemical properties of the amino acids. We indeed see that we obtain a strongly negative (attractive) interaction between two cysteines and a strongly repulsive one between two equally charged amino acids (GLU-GLU). The interactions between two small and non polar residues matches very closely the average behavior of the sequence-independent potential, whereas two hydrophobic residues, despite strongly repulsive at small distances, show an attractive interaction at longer distances. Similar plots for the HV and HH cases are shown in Figure S9 and Figure S10, respectively.

Discussion

In this paper we have shown that the statistical properties of long enough fragments buried in the interior of globular proteins are indistinguishable from those of Gaussian ideal chains in a polymer melt. The data set that we use[49] is based on experimentally derived protein native structures[27]. Figure 1a in fact shows that for sequence separations $70 < m < 90$ the average fragment end-to-end distance, computed between C_α atoms, scales like $m^{1/2}$, as expected for an ideal chain. Buried fragments are selected as previously proposed. At the same time, Figure 1b shows that the scale parameter $b(m)$, that maximizes the likelihood to the Maxwell distribution expected for ideal end-to-end distances, consistently plateaus to a uniform value $b^* = 3.67$ Å in the same sequence separation range.

On the other hand, Figure 2 shows that, even outside the range $70 < m < 90$ range where the Flory theorem seems to hold, empirical distributions are well approximated by Maxwell distributions in the whole sequence separation range $30 < m < 90$, with a scale parameter $b(m)$ non uniform for $m < 70$. The presence of secondary structure elements is seen to play a role (see Figure S2) only for even shorter sequence separations, $m < 20$, and is then fully compatible with the observation of ideal Gaussian statistics for longer fragments. The intriguing observation that medium-sized buried protein fragments follow ideal chain statistics with varying Kuhn length certainly deserves further investigation.

Figure 3 shows the remarkable data collapse of empirical end-to-end distance distributions in the Flory regime $70 < m < 90$, obtained upon rescaling $C_\alpha - C_\alpha$ distances by $m^{1/2}$. Notably, as shown in Figures S3, S4, S5, S6, we find similar results when computing fragment end-to-end distances with a more fine-grained representation of the protein chain, for either all atom (including hydrogen atoms, HH) or for all heavy atoms (excluding hydrogen atoms, HV).

The observed ideal chain behaviour is due to the compensation between excluded volume effects and amino acids interactions, as predicted by the Flory theorem in a polymer melt. Our results extends previous findings based on a smaller data set of protein structures. Moreover, we clearly show how the region in which the theorem applies should be determined. Another important observation emerges from our study: although the Gaussian statistics is valid for a large range of end-to-end distances, at short scales there are deviations due to the fact that the excluded volume effect cannot obviously fully disappear for real protein fragments. We exploit these deviations to extract effective interaction potentials between amino-acids at fragment ends by comparing the empirical probability distribution with the ideal one taken as a reference.

Following this approach, a different statistical potential can be computed separately for any given value of sequence separation for which the ideal statistics is a good approximation of the empirical distribution. The main result of this work, shown in Figure 4, is the collapse of the different statistical potentials in the Flory regime $70 < m < 90$. Most remarkably, while the reference states for different sequence separations collapse when rescaling distances, deviations from the reference states collapse when rescaling back to physical distances, a strong hint that the statistical potentials that we compute in this work do indeed capture physics-based effective interactions. Even more remarkably, different potentials are obtained for the different coarse-graining levels used in this work, again as expected for physics-based effective interaction potentials.

For all coarse-graining levels, the statistical potential vanishes at large distances. Well defined local minima can be observed, listed in Table 2, with the deepest ones corresponding for the atomistic resolutions to steric (sum of Van der Waals radii) or hydrogen bonding interactions. The potential mimima get smoothed when considering a coarser representation, as expected for a proper coarse-graining when the finer degrees of freedom are averaged out. Within the C_α representation, the statistical potential is basically always repulsive; this can be rationalized by observing that the ideal Gaussian reference state already takes into account the average hydrophobic attraction needed for stabilizing a protein globule.

The short-range behaviour is not easy to investigate, since small values of end-to-end distance are scarcely sampled, and the use of shorter, more numerous, fragments is then required. Within the C_α representation, Figure 5 shows that a power law repulsion is found, with an exponent estimate consistent with -6. This could be related to the presence of peculiar dipole-dipole interactions.

Finally, we show in Figure 6 how the same approach can be used to derive sequence-dependent statistical potentials. Unfortunately the statistics available for buried protein fragments with a given pair of amino acid types at their end is barely enough to provide significant signals. Nonetheless, we observe trends consistent with what is expected from the physical-chemical features of the probed types of residue pairs.

An ideal chain reference state was already used to define statistical potentials, in order to take into account chain connectivity in a minimal way[39]. However, our work shows that the use of an ideal chain reference state is well justified only for buried protein fragments, being in fact rooted into the non trivial polymer physics properties of protein globules. We believe this finding may be a breakthrough, at least conceptually.

Further work is certainly needed to investigate in more detail the role of the constraint used to select buried protein fragments. In particular, relaxing that constraint could be a way to gather more statistics and obtain more reliable sequence-dependent potentials. The latter could then be used in several contexts, from protein structure prediction to drug design. Finally, we mention that it would be interesting to compare the results obtained here for buried fragments in protein globules, to the properties of fragments buried in the interior of polymer conformations sampled in the compact phase below the θ-point. In particular, it is interesting to speculate whether the Gaussian behaviour with non uniform Kuhn length found here for intermediate size fragments is peculiar to proteins or not.

Finally, we observe that the statistical properties uncovered in this work were derived analyzing ensembles built with different protein chains. Nonetheless, we may predict that the very same properties, reminiscent of ideal chain behaviour, could be observed for single protein chains in native conditions, for the specific case of Intrinsically Disordered Proteins (IDPs) that can form collapsed, globular ensembles while simultaneously exhibiting significant conformational heterogeneity[54]. This prediction could be in principle tested by means of single molecule FRET experiments in which fluorescent labels can be placed across different chain fragments, thereby providing a direct measurement of end-to-end fragment distances[55]. Similar experiments were in fact already carried on for IDPs that form extended heterogeneous ensembles[56].

Methods

Dataset preparation and analysis

Our database of reference is Top8000[49], which contains a set of 7957 high-resolution protein structures. The dataset has been filtered by excluding those structures of length N that do not exhibit a globular shape, i.e. whose end-to-end distance $R_e(N)$ does not scale as $N^{\frac{1}{3}}$. In order to achieve this we fit experimental data with the relation $R_e(N) = a * N^{\frac{1}{3}}$ and discard all the structures that fall more than three standard deviations apart from the fitted curve. 164 structures have been discarded in this phase.

The tangent vector to each residue has been computed as the difference between the coordinates of the subsequent and the previous residues along the chain. We reported the average cosine of the angle between the tangent vectors of pairs of residues as a function of their separation along the chain. As the tangent-tangent correlation goes to zero when $m \sim 30$, we decided to exclude from our analysis chain fragments shorter than 30 amino acids.

We therefore split all the protein chains in fragments of length $30 \leq m \leq N^{\frac{2}{3}}$ and grouped them by length. All other fragments have been discarded.

Reference and empirical distributions

We measured the end-to-end distance of all fragments of given length m.

We fitted the rescaled data at fixed m with a Maxwell distribution

$$\mathscr{M}_m(R,b) = \frac{4\pi R^2}{(\frac{2}{3}\pi b^2\, m)^{\frac{3}{2}}} \exp(-\frac{3R^2}{2b^2\, m})\,, \tag{3}$$

with a single free parameter b (the scale, a.k.a. Kuhn's length) by using the *scipy.stats* python package and a maximum likelihood fit.

The empirical distribution $\mathscr{E}_m(R,w)$ has been obtained by employing a Kernel Density Estimation (KDE) with a gaussian kernel:

$$\mathscr{E}_m(R,w) = \frac{1}{M}\sum_{\sigma} \frac{1}{\sqrt{2\pi w}} \exp\left(\frac{(R - r_\sigma)^2}{2\, w^2}\right). \tag{4}$$

The sum is extended over all M values r_σ in the dataset of end-to-end distances of fragments of length m. We used cross validation in order to estabilish the optimal kernel bandwidth w for fragment lenghts $m \in \{42, 48, 60, 64, 66, 72, 78, 84, 92\}$. We divided every set of end-to-end distances of fixed fragment length in five groups: an empirical distribution was computed using the data of four groups. The width of the gaussian kernel was therefore adjusted in order to maximize the likelihood that the data from the fifth group was obtained from the same empirical distribution.

In order to estimate the optimal bandwidth for all other datasets, we assumed the relation $w = a * n^s$ between the bandwidth w and the number of points n in the dataset. We fitted the parameters a and s by minimizing the RMSD with the cross-validated bandwidths (see Figure S11).

Potential

For every m, we estimated the potential as a function of the distance R by using the formula

$$V_m(R|b,w) = -\log\left(\frac{\mathscr{M}_m(R,b)}{\mathscr{E}_m(R,w)}\right). \tag{5}$$

As the potential does not depend on m, we finally computed $V(R|b,w)$ as the average over all possible fragment length of $V_m(R|b,w)$.

We fitted the short range repulsive part of the potential by minimizing the root mean square deviation between the logarithm of $V(R|b,w)$ and a linear function.

Residue-dependent analysis

We repeated the previous analysis while filtering the fragments depending on the amino acids they presented at their ends.

References

1. Creighton, T. E. *Proteins: structures and molecular properties* (Macmillan, 1993).
2. Anfinsen, C. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
3. Bahar, I., Lezon, T., Yang, L.-W. & Eyal, E. Global dynamics of proteins: Bridging between structure and function. *Annu. Rev. Biophys.* **39**, 23–42 (2010).
4. Levine, Z. & Shea, J.-E. Simulations of disordered proteins and systems with conformational heterogeneity. *Curr. Opin. Struct. Biol.* **43**, 95–103 (2017).
5. Noid, W. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **139** (2013).
6. Baker, D. A surprising simplicity to protein folding. *Nature* **405**, 39–42 (2000).
7. Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **18**, 10–15 (2008).
8. Yang, J., Chen, W., Skolnick, J. & Shakhnovich, E. All-atom ab initio folding of a diverse set of proteins. *Structure* **15**, 53–63 (2007).
9. Majek, P. & Elber, R. A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins. *Proteins: Struct. Funct. Bioinforma.* **76**, 822–836 (2009).
10. Xu, J., Huang, L. & Shakhnovich, E. The ensemble folding kinetics of the fbp28 ww domain revealed by an all-atom monte carlo simulation in a knowledge-based potential. *Proteins: Struct. Funct. Bioinforma.* **79**, 1704–1714 (2011).
11. Yang, J. *et al.* The i-tasser suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2014).
12. Leman, J. *et al.* Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nat. Methods* **17**, 665–680 (2020).
13. Cossio, P., Granata, D., Laio, A., Seno, F. & Trovato, A. A simple and efficient statistical potential for scoring ensembles of protein structures. *Sci. Reports* **2** (2012).
14. Studer, G. *et al.* Qmeandisco-distance constraints applied on model quality estimation. *Bioinforma. (Oxford, England)* **36**, 1765–1771 (2020).
15. Trovato, A., Chiti, F., Maritan, A. & Seno, F. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput. Biol.* **2**, 1608–1618 (2006).
16. Walsh, I., Seno, F., Tosatto, S. & Trovato, A. Pasta 2.0: An improved server for protein aggregation prediction. *Nucleic Acids Res.* **42**, W301–W307 (2014).
17. Orlando, G., Silva, A., MacEdo-Ribeiro, S., Raimondi, D. & Vranken, W. Accurate prediction of protein beta-aggregation with generalized statistical potentials. *Bioinformatics* **36**, 2076–2081 (2020).
18. De Vries, S. *et al.* Haddock versus haddock: New features and performance of haddock2.0 on the capri targets. *Proteins: Struct. Funct. Genet.* **69**, 726–733 (2007).
19. Sarti, E., Granata, D., Seno, F., Trovato, A. & Laio, A. Native fold and docking pose discrimination by the same residue-based scoring function. *Proteins: Struct. Funct. Bioinforma.* **83**, 621–630 (2015).
20. Sarti, E., Gladich, I., Zamuner, S., Correia, B. & Laio, A. Protein–protein structure prediction by scoring molecular dynamics trajectories of putative poses. *Proteins: Struct. Funct. Bioinforma.* **84**, 1312–1320 (2016).
21. Yu, J. *et al.* Interevdock: A docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Res.* **44**, W542–W549 (2016).
22. Battisti, A., Zamuner, S., Sarti, E. & Laio, A. Toward a unified scoring function for native state discrimination and drug-binding pocket recognition. *Phys. Chem. Chem. Phys.* **20**, 17148–17155 (2018).
23. Guerois, R., Nielsen, J. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
24. Vangone, A. & Bonvin, A. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife* **4** (2015).

25. Xiong, P., Zhang, C., Zheng, W. & Zhang, Y. Bindprofx: Assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *J. Mol. Biol.* **429**, 426–434 (2017).

26. Skrbic, T. *et al.* Vibrational entropy estimation can improve binding affinity prediction for non-obligatory protein complexes. *Proteins: Struct. Funct. Bioinforma.* **86**, 393–404 (2018).

27. Berman, H. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

28. Tanaka, S. & Scheraga, H. Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9**, 945–950 (1976).

29. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

30. Jones, D., Taylort, W. & Thornton, J. A new approach to protein fold recognition. *Nature* **358**, 86–89 (1992).

31. Buchete, N.-V., Straub, J. & Thirumalai, D. Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.* **14**, 225–232 (2004).

32. Dehouck, Y., Gilis, D. & Rooman, M. A new generation of statistical potentials for proteins. *Biophys. J.* **90**, 4010–4017 (2006).

33. Bhattacharyay, A., Trovato, A. & Seno, F. Simple solvation potential for coarse-grained models of proteins. *Proteins: Struct. Funct. Genet.* **67**, 285–292 (2007).

34. Sarti, E. *et al.* Bachscore. a tool for evaluating efficiently and reliably the quality of large sets of protein structures. *Comput. Phys. Commun.* **184**, 2860–2865 (2013).

35. Sippl, M. Calculation of conformational ensembles from potentials of mena force. an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883 (1990).

36. Samudrala, R. & Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. molecular biology* **275**, 895–916 (1998).

37. Betancourt, M. & Thirumalai, D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **8**, 361–369 (1999).

38. Skolnick, J., Kolinski, A. & Ortiz, A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins: Struct. Funct. Genet.* **38**, 3–16 (2000).

39. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS ONE* **5** (2010).

40. Hamelryck, T. *et al.* Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS ONE* **5** (2010).

41. Thomas, P. & Dill, K. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **257**, 457–469 (1996).

42. Ben-Naim, A. Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.* **107**, 3698–3706 (1997).

43. Sippl, M. J. Knowledge-based potentials for proteins. *Curr. opinion structural biology* **5**, 229–235 (1995).

44. Banavar, J. R., Hoang, T. X. & Maritan, A. Proteins and polymers. *The J. chemical physics* **122**, 234910 (2005).

45. Flory, P. J. *Principles of polymer chemistry* (Cornell University Press, 1953).

46. Orland, H. Flory theory revisited. *J. de Physique I* **4**, 101–114 (1994).

47. Lua, R., Borovinskiy, A. L. & Grosberg, A. Y. Fractal and statistical properties of large compact polymers: a computational study. *Polymer* **45**, 717–731 (2004).

48. Tobi, D. & Elber, R. Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins: Struct. Funct. Bioinforma.* **41**, 40–46 (2000).

49. Richardson, J. S. & Richardson, D. C. Studying and polishing the pdb's macromolecules. *Biopolymers* **99**, 170–182 (2013).

50. Berezovsky, I. N., Grosberg, A. Y. & Trifonov, E. N. Closed loops of nearly standard size: common basic element of protein structure. *Febs Lett.* **466**, 283–286 (2000).

51. Rowland, R. & Taylor, R. Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der waals radii. *J. Phys. Chem.* **100**, 7384–7391 (1996).

52. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

53. Liwo, A. *et al.* United-residue force field for off-lattice protein-structure simulations: Iii. origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *J. computational chemistry* **19**, 259–276 (1998).

54. Holehouse, A. & Pappu, R. Collapse transitions of proteins and the interplay among backbone, sidechain, and solvent interactions. *Annu. Rev. Biophys.* **47**, 19–39 (2018).

55. Soranno, A. Physical basis of the disorder-order transition. *Arch. Biochem. Biophys.* **685** (2020).

56. Zheng, W. *et al.* Inferring properties of disordered chains from fret transfer efficiencies. *J. Chem. Phys.* **148** (2018).

Author contributions statement

Conceptualization, F.S., A.T, and S.Z.; methodology, F.S., A.T, and S.Z.; software, S.Z.; validation, F.S., A.T, and S.Z.; formal analysis, S.Z.; investigation, F.S., A.T, and S.Z.; data curation, S.Z.; writing–original draft preparation, F.S., A.T, and S.Z.; writing–review and editing, F.S., A.T, and S.Z.; supervision, A.T. All authors have read and agreed to the published version of the manuscript.

Additional information

This research received no external funding. The authors declare no conflict of interest.