Topological decoding of biomolecular fold complexity

Scalvini, B.

CHAPTER 2:

TOPOLOGICAL PRINCIPLES OF PROTEIN FOLDING

What is the topology of a protein and what governs protein folding to a specific topology? This is a fundamental question in biology. The protein folding reaction is a critically important cellular process, which is failing in many prevalent diseases. Understanding protein folding is also key to the design of new proteins for applications. However, our ability to predict the folding of a protein chain is quite limited and much is still unknown about the topological principles of folding. Current predictors of folding kinetics, including the contact order and size, present a limited predictive power, suggesting that these models are fundamentally incomplete.

Here, we use a newly developed mathematical framework to define and extract the topology of a native protein conformation beyond knot theory, and investigate the relationship between native topology and folding kinetics in experimentally characterized proteins. We show that not only the folding rate, but also the mechanistic insight into folding mechanisms can be inferred from topological parameters.

We identify basic topological features that speed up or slow down the folding process. The approach enabled the decomposition of protein 3D conformation into topologically independent elementary folding units, called circuits. The number of circuits correlates significantly with the folding rate, offering not only an efficient kinetic predictor, but also a tool for a deeper understanding of theoretical folding models. This study contributes to recent work that reveals the critical relevance of topology to protein folding with a new, contact-based, mathematically rigorous perspective. We show that topology can predict folding kinetics when geometry-based predictors like contact order and size fail.

Publications associated to this chapter:

Barbara Scalvini, Vahid Sheikhhassani, Alireza Mashaghi, *Topological principles of protein folding*, PCCP, 23, 37, 21316-21328, September 21, 2021, https://doi.org/10.1039/D1CP03390E

1. INTRODUCTION

Over the last 20 years, it has been hypothesized that protein folding rates and mechanisms can be inferred from the native state topology [1]. The importance of local intra-chain contacts for small one-domain proteins emerged with the definition of Contact Order (CO), a “topological” parameter still widely used to date to predict protein folding rates [2]. This parameter was then coupled with size (length of the protein) with the introduction of absolute CO [3], to allow for better description of the folding kinetics of larger proteins. For such proteins, the folding pathway may be characterized by kinetic traps and escape from low free energy conformations [4][5]. In more recent years, other models have been suggested for folding rate prediction, based on total contact distance[6], a small selection of contact information [7], cumulative torsion angle[8] and other structural information [9]–[13]. Moreover, an evolution of the concept of contact order called partial contact order was envisioned in order to follow the progression of such topological descriptor from the unfolded to the folded state[14]. The partial contact order pCO takes into account the likelihood that a certain contact is formed, and the associated reduction of loop entropy [14].However, contact distance, contact order and protein length are not inherently topological properties, if topology is to be intended in the mathematical sense of the word. Topology is a mathematical concept characterizing the properties of objects which remain unaltered through continuous, invertible transformations such as stretching, shrinking and bending. A first step to introduce topology-based predictors for the quantification of entanglement was taken by Marco Baiesi et al. [15]–[17]. Drawing from knot theory, the concept of Gaussian entanglement was first applied to the intertwined backbones of domain-swapped protein dimers [17], and then to non-overlapping looping sub-chains of the same protein, where it proved to complement absolute CO in folding rate prediction on a set of 48 proteins [15]. However informative, these topologically inspired descriptors often concern a fairly limited portion of the available protein datasets, with about 15% of dimers displaying significant intertwining [17], and 32% of proteins from the CATH database showing non trivial Gaussian entanglement [16]. Topological concepts such as writhe and torsion were also applied to the protein backbone, yielding good results for folding rate prediction and revealing the role of handedness of proteins at both local and global organization levels [18].

Previous topological efforts to quantify the relation between the native state three-dimensional arrangement and folding kinetics relied on the concept of entanglement as defined by knot theory, and focused on the entanglement of the backbone. However, knots are rare in proteins, and knotted proteins generally yield very low folding rates [19]. Other topologically inspired descriptors

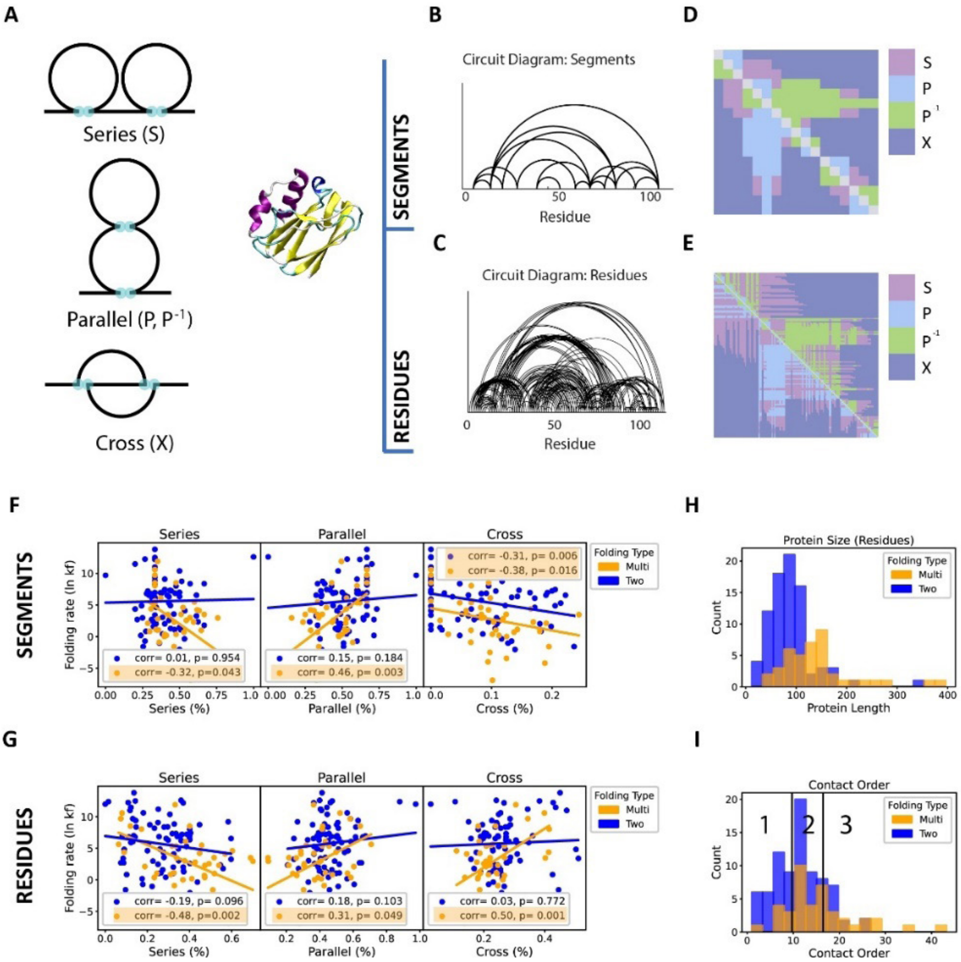


Figure 1. Segment and Residue-based CT parameters correlate with folding rate. A Three pairwise arrangements of CT: series, parallel and cross. The inner contact is in parallel relation (P) with the outer contact, while the outer contact is in inverse parallel relation (P-1) with the inner contact. ***B*** Circuit diagram for segment-based contacts. ***C*** Circuit diagram for residue-based contacts. ***D*** CT matrix for segment-based contacts. ***E*** CT matrix for residue-based contacts. Segment and residue-based contacts offer very different resolution into protein topological arrangement, for the same protein (pseudoazurin, PDB code: 1ADW). ***F*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for segment-based contacts. ***G*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts. ***H*** Size distribution (number of residues) for two-state and multi-state folders. ***I*** Contact Order distribution of the dataset. The dataset was divided into three sub-datasets (Lower, Average and Upper CO) by setting an upper (16.47) and lower (9.72) limit.

drawn from knot theory such as Gaussian entanglement also rely on the concept of backbone entanglement.

The effect of entanglement of proteins with no knots and no slipknots to folding rates has been studied in Panagiotou and Plaxco[18]and Baiesi, Orlandini et al. [15]–[17]. A mathematically rigorous topology concept, termed circuit topology (CT), has recently been proposed to describe the topology of unknots [20][21]. Circuit topology, in its first order definition, ignores possible backbone entanglement, and focuses only on the intra-chain contacts present in the native protein structure. Contacts are considered to be fixed. This allows circuit topology to provide a topological description of unknotted yet folded linear chains [20]–[23], a type of description which is complementary to that provided by Gaussian entanglement [15]–[17], writhe and torsion [18]. Moreover, contact-based topological descriptors represent a very natural framework for proteins, since contacts often have not only geometrical but also biological relevance. The circuit topology framework allows us to readily combine our descriptors with information such as the energy of a contact, for example. The vast majority of proteins present intra-chain contacts, making our analysis applicable virtually to all proteins. Once contacts in a structures have been identified, they are classified based on their pairwise topological arrangement (Figure 1A). According to CT, contacts can be in either one of three possible relations with each other: series (S), parallel (P) and cross (X) (Figure 1A). Series and parallel relations also include a subset of relations, called *concerted relations*, in which one of the two contact sites is shared between the two contacts. We call these relations concerted parallel (CP) and concerted series (CS). Here, we present a first order analysis, therefore CP and CS will be included in the main sets and counted respectively as parallel and series. We note that CT was already suggested to have an impact on the folding dynamics of model polymers [22][24], although its relevance to protein folding has not been evaluated.

Here, we show how the three fundamental topological relations S, P, and X display differential patterns of correlation with folding rate, providing insight into which types of topological arrangements facilitate folding and which hinder it. We define as *zipper effect* the mechanism with which a predominance of series arrangement slows folding, while parallel and cross arrangements (the so-called *entangled relations*) yield higher folding rates. It is important to note that here the word 'entangled' is used in a broad sense, since we are dealing with unknots. Parallel and cross are designated as entangled because the two loops forming the relation are not independent of each other. We show that both two-state and multi-state class proteins display statistical evidence of the zipper effect, if we only consider the topology of short range, attractive energy contacts. Lastly, we will show how proteins can be decomposed into *topological circuits* [25], that is, topologically independent units. The number of these circuits normalized by size correlates positively with the logarithm of folding rate ln(k_f), suggesting that the localization of contacts inside topological circuits might play a role in facilitating folding efficiency.

2. RESULTS

2.1. Topological parameters as kinetic predictors

Circuit topology utilizes contacts as basic elements for topological classification. However, a suitable definition of contacts is widely dependent on the purpose of the study. For folding rate prediction, contacts between residues have mostly been used for quantifying parameters such as CO [2]. Here, we will also consider contacts between residues. However, this is not the only choice; the flexibility of the CT framework allows us to consider other type of protein building blocks which can form contacts; one can define *segments* of proteins which correspond to secondary structure elements, and perform CT analysis on the contacts created by these coarse-grained structures. In Figure 1 we can see the CT diagram of segment-segment (Figure 1B) and residue-residue (Figure 1C) contacts, and their respective CT matrices, from which the frequencies of CT topological relations can readily be extracted (Figure 1D and 1E). Strikingly, these CT frequencies correlate with the logarithm of folding rate. The two choices of contact definition provide very different structural resolution, and we expectedly observe different degrees of correlation with folding rate. Contacts were retrieved from PDB structures, by defining a spatial cutoff for atom-atom distance (5.0 Å), and a threshold for the minimum number of atoms to be found in spatial proximity below the cut-off in order to consider the two residues/segments in contact (5 atoms for residues, 10 atoms for segments). Our main conclusions are robust with respect to the choice of parameters. For other cut - off choices, see Supplementary Information.

Next, we investigated whether the observed correlations depend on folding pathway complexity. Many proteins fold and unfold with one main fast event, by a simple two-state transition. These “two-state folders” [11][26][27] have gathered much of the attention of scientific inquiry in the past, and their folding rates correlate with relative CO[2]. On the other hand, proteins with more intricate multi-state transitions – “multi-state folders” – have shown a strong dependency of their folding kinetics on protein length (and not CO) [28]. Notably, CT parameters also provide differential patterns of correlations for two- and multi-state folders, at first sight (Figure 1F, 1G). For both segment and residue analyses we find statistically significant negative correlation between ln(kf) and series in multi-state folders (respectively $r = -0.32$, $p = 0.043$ and $r = -0.48$, $p = 0.002$). This

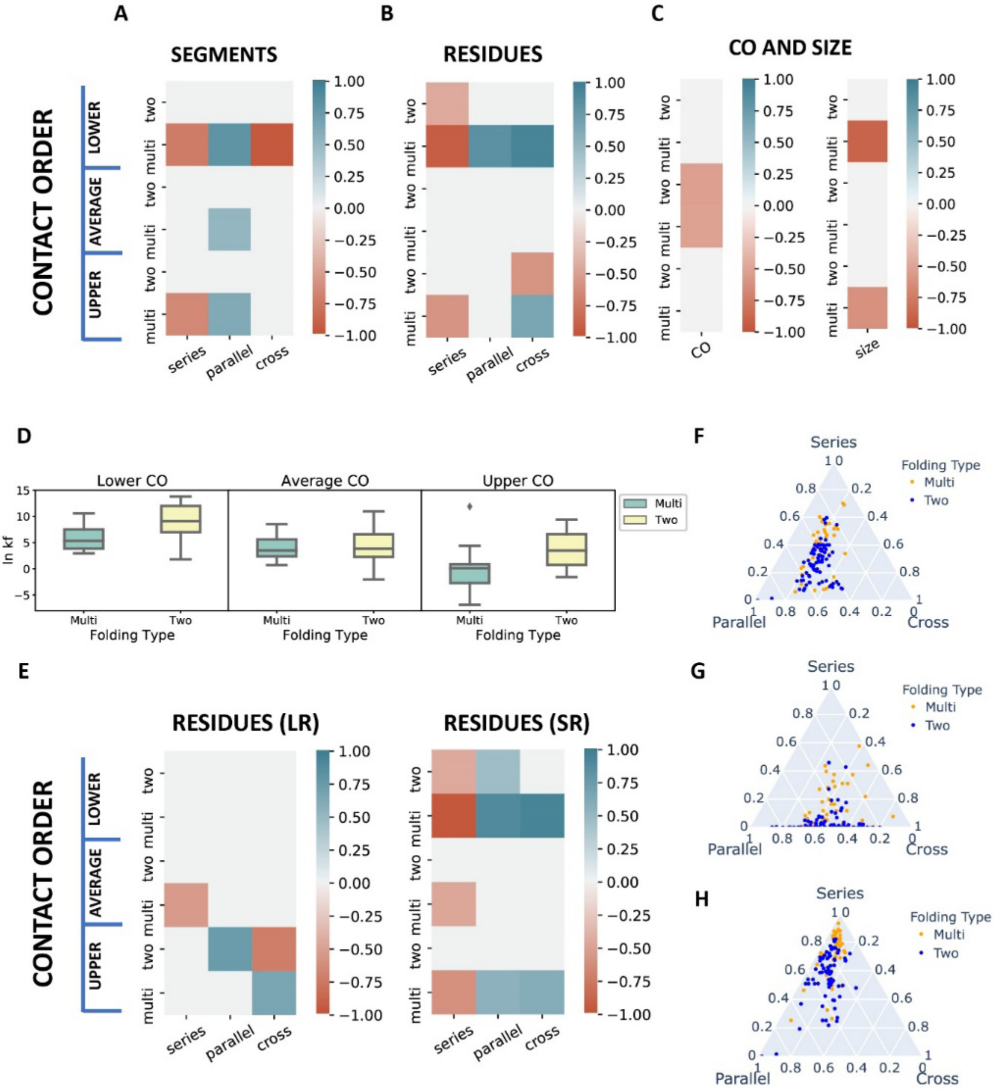


Figure 2. Classification based on Contact Order and length filtering highlight differential patterns of correlation. A Folding rate correlation map for segment-based CT, with CO classification. ***B*** Folding rate correlation map for residue-based CT, with CO classification. ***C*** Folding rate correlation map for Contact Order and Size, with CO classification. CT seems to be more informative than Contact Order for proteins with Lower and Upper Folding rate. ***D*** Boxplot of folding rate for different CO subsets. Slow folders populate the Upper CO sub-dataset, and display correlation between folding rate and long-range residue-based contacts. ***E*** Folding rate correlation map for residue-based CT, with CO classification. The two maps show only long-range contacts (on the left) and only short-range contacts (on the right). The threshold for range classification was set to 24 residues. ***F*** Triangular plot of the topological composition throughout the dataset, for residue-based CT. ***G*** Triangular plot of the topological composition throughout the dataset, for long-range residue-based CT. ***H*** Triangular plot of the topological composition throughout the dataset, for short-range residue-based CT.

is understandable as series relations favor delocalization along the chain, which seems to slow down folding of multi-state folders, but leaves two-state folders unaffected. On the other hand, two-state proteins display moderate negative correlation with cross relations (r = -0.31, p = 0.006), in their segment representation. These differences might be due to the different average size of the two-state and multi-state proteins (Figure 1H). Two-state proteins are generally smaller, therefore highly entangled topologies such as those favored by cross arrangement might be less likely to appear on the secondary structure level, for geometric and energetic constraints. The likelihood of finding such structures might increase for longer folding times. Therefore, it is not surprising to find a negative correlation between cross and folding rate in this instance. The folding rate in multi-state proteins is more affected by topology, showing evidence of statistically relevant zipper effect at both residue and secondary structure levels, having negative correlation with series and positive correlation with at least one of the two entangled relations – parallel for segments (r = 0.46, p = 0.003), both parallel and cross for residues (r = 0.31, p = 0.049 for parallel, r = 0.50, p = 0.001 for cross). Segment analysis yields correlation values for parallel (in multi-state folders) which are comparable to those obtained by Panagiotou and Plaxco for the writhe of the protein Primitive Path and the logarithm of folding rate [18]. Chain writhing is a mechanism (possibly the main one in proteins) which can indeed create parallel contact topologies; thus, in this case, contact topology might be seen as a proxy for backbone topology. However, no correlation is found for parallel topology in two-state proteins, indicating possibly that the protein is too short to produce substantial writhe. CT parameters are normalized by the number of contacts in the chain, making it possible to compare proteins with very different geometrical properties. However, due to the assembly principles of proteins and geometrical and steric constraints, a non-trivial relationship between size and CT parameters exists (Figure S1).

2.2. Disentangling the contributions of geometry and topology

We demonstrate that topology-based predictors complement CO and size, which are geometry-based predictors. In order to do so, we divided the dataset into three sub-datasets, based on their CO (Figure 1I): Upper, Average and Lower CO. CO values were retrieved from the ARCPro dataset [29] (cutoff value = 6 Å). Figure 2A and 2B show the correlation coefficients for three subsets, for two-state and multi-state proteins. Exact values can be seen in Tables S1 and S2. We also compare the CT correlation maps to those obtained by using CO and size on the same datasets (Figure 2C, Table S3). While CO is moderately accurate in predicting $\ln(k_f)$ for the Average CO dataset (r = -0.53, p = 4.5e^{-04} for two-state, r

= -0.51, p = 0.03 for multi-state), CT seems to obtain the best results for the two tails of the CO distribution, obtaining correlations as high as r = -0.93 (p = 0.002) for series and r = 0.94 (p = 0.001) for cross for multi-state proteins in the Lower CO range (Figure 2B). These results imply that CO and CT give in fact complementary information about folding kinetics. Also, CT is able to provide resolution for those proteins that have similar CO but present significant discrepancies in folding rate. Size also provides strong correlations for the Upper and Lower CO datasets (Figure 2C, Table S4), although only for multi-state proteins (r = -0.89, p = 0.01 for Lower CO and r = -0.61, p = 0.01), as expected. By combining the kinetic information on multi-state proteins provided by residue-based CT and size parameters, we see that not only the number of residues is impactful but also their topological arrangement, with contact delocalization favored by series relations being as efficient as protein length in promoting a slow folding process. Interestingly, CT on the segment level displays an opposite trend for cross relations for multi-state, Lower CO proteins, indicating that such a level of entanglement at the secondary structure level might actually be hindering folding for those members of the multi-state protein class which are smaller (Figure 1H) and have higher folding rates (Figure 2D). The fact that smaller multi-state proteins show similar correlations for the cross fraction to two-state folders might suggest a rather continuous transition with respect to size between multi-state and two-state classes, rather than two binary distinct folding styles [5][30].

2.3. Arrangement of short-range attractive contacts as a topological driver of folding

Here, we investigate how topology, interaction energy, and interaction range work together to regulate folding kinetics. The CO reflects the relative importance of local and non-local interactions in the molecule [2]. The conceptual background behind CO is that contacts between residues that are closer along the chain are less entropically costly, and therefore tend to happen early in the folding process. Therefore, simple proteins structures which are rich in local contacts tend to fold faster [1][31]. Broglia and Tiana [32] highlighted the role of local contacts by identifying a specific hierarchy, which involves the formation of early local elementary structures (LES), followed by the assembly of the LES into a post critical folding nucleus at a later stage. Moreover, evidence exists that natural selection favors folds with low contact order [33], and therefore structures rich in local contacts. Indeed, off-lattice models of protein folding showed that the suppression of local interactions prevents the structure from reaching the native conformation.[34] However, the respective role of local versus non local interactions is still a highly debated subject in literature. In silico studies of three model 36-mers on a cubic lattice suggested that non-local interactions are the primary determinant of protein folding [35].

We can ask ourselves if not only the relative number of local versus non-local contacts, but also their topological arrangement has an impact on folding kinetics. To address this question, we applied a 24 residue threshold in order to discriminate between short-range and long-range contacts prior to CT analysis (Figure 2E, Table S5 and S6). It is apparent to see that the topology of short-range contacts displays correlations which are higher in magnitude and also more widespread over the whole CO range, as opposed to long range contacts. Multi-state proteins in the Lower CO range still display the highest correlations between topological content and $\ln(k_f)$: r = -0.97, p = 1.9E-04 for series, r= 0.89, p = 0.007 for parallel, and r = 0.94, p = 0.002 for cross. The zipper effect appears to be confirmed in the results from the short-range correlation panel (Figure 2E): once local contacts are uncoupled from non-local contacts in CT analysis, negative correlations with folding rate are only seen for series relations, and positive correlations are observed with the *entangled relations*, cross and parallel. Short range contacts appear as the main topological folding drivers. This is compatible with the findings of Adesh Kumar and co-workers [36], who theorized that local contacts might be fundamental for the differentiation between the native-like conformations during folding, by Montecarlo simulation of three protein structures. However, correlations with long-range contacts also appear for the 'slow folding' Upper CO proteins (Figure 2D, 2E). Since non-local contacts along the chain are generally formed at a later stage during folding [32], they can only affect the folding process after longer characteristic times. This finding suggests that, for very fast folders, the impact of the topology of long-range contacts might be negligible.

Moreover, we find that short and long-range contacts are also qualitatively different with respect to the topological content. Figure 2F portrays in a triangular plot the percentages of series, cross and parallel for all residues. We can compare it to the topological content in long-range (Figure 2G) and short-range (Figure 2H) contacts; we see that non-local contacts are actually much richer in cross relations with respect to local contacts. This finding indicates that at the level of short-range contacts the high entanglement promoted by cross relations is unfavorable.

Contacts can also be discriminated by assigning energy-like quantities based on the statistical potential suggested by Paul Thomas and Ken Dill [37][38]. This procedure is a first order attempt to add bio-chemical information to contact topology: contacts can be filtered based on the sign of the potential matrix element associated to residue-residue interaction, resulting in 'repulsive energy contacts' (E>0) and 'attractive energy contacts' (E<0). The correlation map for energy filtering (Figure 3A, Table S7 and S8) clearly highlights how the topology of attracti-

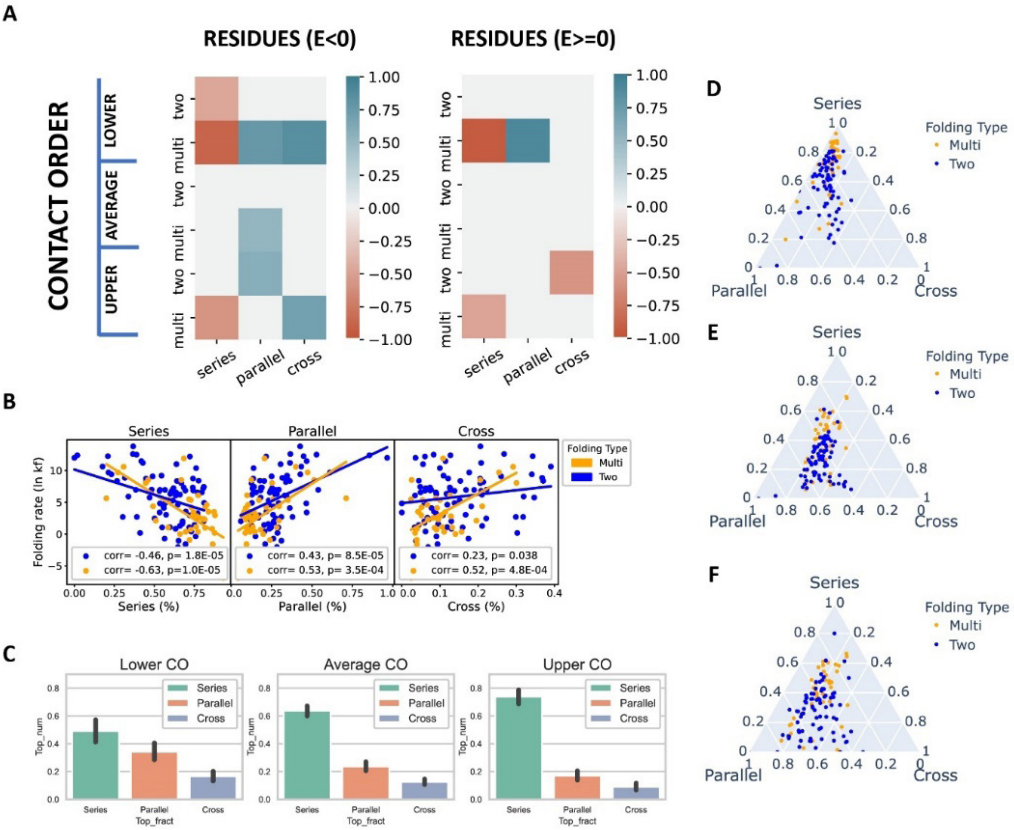


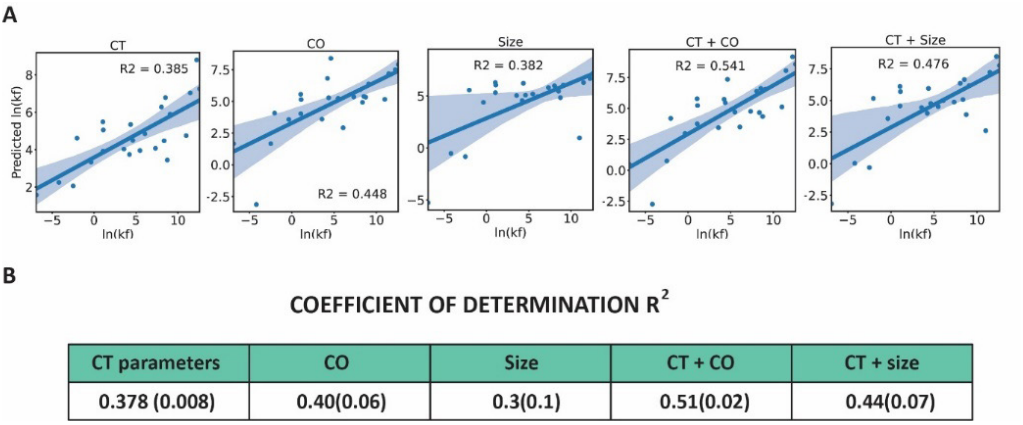
Figure 3. Classification based on Contact Order and energy filtering highlight the kinetic role of the topology of short-range attractive contacts. ***A*** Folding rate correlation maps for residue-based CT, with CO classification. The two maps show only negative energy contacts (on the left) and only positive energy contacts (on the right). ***B*** Scatterplot for residue-based CT fractions and folding rate: only short-range and negative energy contacts were included. With this type of filtering, both folding types display zipping effect, and all correlations are significant (p value ≤ 0.05). ***C*** Bar plot of topological fractions with respect to Contact Order classification, for negative energy/short-range residue-based CT. With increasing CO, we observe an increase in series fraction and a decrease in entangled fraction (parallel, cross). ***D*** Triangular plot of the topological composition throughout the dataset, for negative energy/short-range residue-based CT. ***E*** Triangular plot of the topological composition throughout the dataset, for negative energy residue-based CT. ***F*** Triangular plot of the topological composition throughout the dataset, for positive energy residue-based CT.

ve energy contacts plays the biggest role in folding kinetics. However, repulsive contacts can still correlate with slower folding processes, as in the case of Lower CO multi-state ($r = -0.95$, $p = 0.001$ for series), for Upper CO two-state proteins (cross, $r = -0.58$, $p = 0.05$) and Upper CO multi-state proteins (series, $r = -0.50$, $p = 0.048$).

Considering the results for length and energy-based contact filtering, it becomes clear that not all contact topologies are equally impactful when it comes to folding. Local, attractive energy contacts seem to be the topological drivers of the folding process. It is therefore natural to reconsider correlations for the whole dataset while considering short range, attractive energy contacts exclusively (Figure 3B). Interestingly, this type of filtering yields statistically significant correlations for both two-state and multi-state proteins, for all three topological relations. Even more notably, now both two-state and multi-state proteins show evidence of zipper effect, making the distinction between the two classes more quantitative than qualitative; correlations seem to be still more pronounced in the case of multi-state folders, but correlation trends are the same for the two classes. Figure 3C shows another evidence of zipper effect: with decreasing contact order (higher folding rate), there is a gradual increase in entangled relations. However, the triangular plot of the energy/length filtered dataset (Figure 3D) is a closer match to the short-range triangular plot (Figure 2H) rather than to the attractive energy plot (Figure 3E), indicating that the best predictor for the topological content is distance between contacts, and not energy. Moreover, the topological content for repulsive energy contacts (Figure 3F) does not look significantly different from the one for attractive energy contacts (Figure 3E).

2.4. Linear combination of CO and CT parameters as an improved folding rate predictor

The analysis outlined so far suggests complementarity between folding rate descriptors such as CT parameters and Contact Order. We see, for example, how the pre-filtering of data based on Contact Order is useful to uncover differential patterns of correlations for CT parameters (Figure 2 A, B, C, E, Figure 3 A). CO pre- filtering highlights also how proteins belonging to different CO ranges might be best described by CO, CT parameters or size, when it comes to folding rate prediction. It is then natural to ask whether these folding rate descriptors could be combined to produce more accurate folding rate predictions. In order to test this hypothesis, we envisioned a multilinear regression analysis of the dataset, using CT parameters, CT parameters combined with CO, and CT parameters combined with size as independent variables. Folding rate predictions yielded by using only CO and size are also reported for comparison. For the analysis, we use CT fractions derived from attractive energy short range contacts, since this unifies two- and multi-state folders for what concerns their correlation patterns with respect to CT (Figure 3B). All CT parameter values reported in this paper were previously normalized by the total sum of S, P and X relations in the protein. This normalization implies that, once we provide two CT fractions, the



COEFFICIENT OF DETERMINATION R^2

CT parameters	CO	Size	CT + CO	CT + size
0.378 (0.008)	0.40(0.06)	0.3(0.1)	0.51(0.02)	0.44(0.07)

Figure 4. The linear combination of CT parameters and CO allows for folding rate prediction with increased statistical significance. A Scatterplots of predicted folding rate (obtained with multilinear regression over CT fractions, CO, protein length and a combination of these parameters) and experimental Folding rate (ln kf), calculated over one of the 5 training/test sets combinations. ***B*** Average R^2 score for CT parameters (parallel, cross), CO, Size (protein length expressed in number of residues) and linear combination of CT parameters and CO, CT parameters and size. Numbers between parentheses indicate the standard deviation. The average was performed over 3 different choices of training/test subsets. Predictions obtained over test sets 1 and 5 were discarded by residual analysis, as their residual distribution did not satisfy the normality requirement.

third is automatically determined, as the sum of all three fractions needs to yield 1. This allows us to compare proteins with very different number of contacts. However, one of the three CT parameters is actually redundant, when it comes to multilinear regression analysis (MLR). Therefore, we decided to discard one and only use two CT parameters for folding rate prediction. Since the independent variables used for MLR should not be too highly correlated, we chose the two CT parameters which presented the lowest correlation coefficient when confronted with each other, parallel and cross (r = 0.23 , p = 0.011). The CT-based folding rate predictor is therefore defined as:

$$K_{CT} = c_P P + c_X X$$

where KCT is the predicted logarithm of folding rate, P and X are the parallel and cross fractions and cP and cX are coefficients which are calculated by the MLR model over the training set. Following the same reasoning, CT parameters can be combined with CO and size to obtain new predictors:

$$K_{CT+CO} = c_P P + c_X X + c_{CO} CO$$

$$K_{CT+L} = c_P P + c_X X + c_L L$$

where L is the size of the protein (number of residues), and c_L , c_{CO} coefficients calculated by the MLR model. In order to perform this analysis we relied on a freely available Python tool for machine learning and predictive data analysis, scikit-learn 0.24.2 [39]. Thanks to the scikit-learn cross-validator module, we divided the dataset into 5 consecutive folds (sub-sets). Iteratively, 4 of these 5 datasets were used as training set for the model, and the remaining one as test set for folding rate prediction. Folding rate predictions on one of these test sets can be seen in Figure 4A, for all predictors. Predictions for all test sets can be found in Figure S2. A useful parameter to quantify the goodness of our prediction (how well the MLR model is representative of our dataset) is the coefficient of determination R^2 [40]. The table in Figure 4B presents the average R^2 over the predictions from the 5 test sets: it is clear to see that both CO and size have higher predictive power when combined with CT than when they are used on their own, with K_{CT+CO} representing the best folding rate predictor. Folding rate predictions from the first and last test set were excluded from the comparison, as the residual (predicted folding rate – experimental folding rate) distribution from CO prediction did not satisfy the normality requirement (Table S9). However, R^2 values and adjusted R^2 values from all test sets can be seen in Table S10 and S11 respectively. The adjusted determination coefficient R^2_{adj} is a modified version of R^2 which takes into account the number of independent variables in the model. It discriminates whether the added variables provide an improvement to the prediction which is higher than what would be expected by the addition of random parameters. The R^2_{adj} coefficient confirms our general conclusions which identify K_{CT+CO} as the best predictor (Table S11). This result proves the complementarity in predictive power for CT parameters and CO for folding rate prediction, which we already hypothesized from the analysis presented in Figure 2B and 2C.

2.5. Circuits as elementary folding units

The circuit topology of a chain enables bottom-up analysis of fold architecture. We investigate whether higher order topological features are seen in proteins and whether they contribute to folding kinetics. It was previously suggested that protein folding might proceed in a step-wise manner from separately cooperative elementary units of about 20 residues, called foldons [41]. Analogously, we can look for the topological equivalent of folding sub-units, by exploiting a string notation of contacts, such as that utilized by generalized circuit topology [25]. The string notation allows for identification of *circuits* in the chain, formally defined as a segment of a string that consists only of pairs of letters. Circuits represent independent topological structures within a complex topology. Let us have a look at Figure 5A and 5B to clarify the notation. Letters are assigned to contacts in

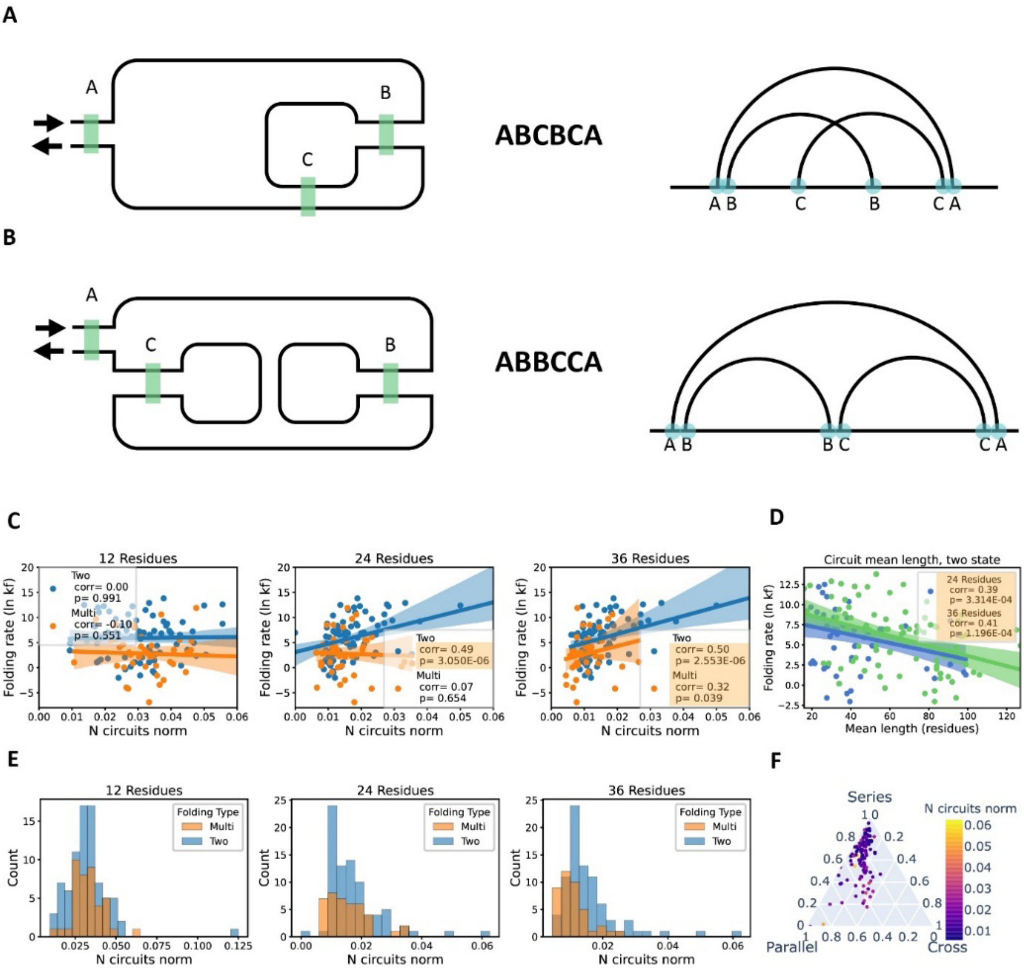


Figure 5. The number of topological circuits normalized by the size of the protein correlates positively with folding rate. A Example of circuit, with string and diagram representation. This circuit can be further decomposed, as BCBC is itself a circuit. ***B*** Example of circuit, with string and diagram representation. Also in this case, the circuit can be further decomposed. If we remove contact B, we would obtain circuit ACCA, leaving the topology of contact C and A unaffected. The same goes for contact C and circuit ABBA. Contact C and B together also form a circuit, BBCC. ***C*** Scatterplot for number of circuits normalized by size and folding rate. Legends display Spearman correlation coefficients. ***D*** Scatterplot for circuit mean length and folding rate, for 24 and 36 residue thresholds for long-range exclusion. No correlation was detected for the 12 residue threshold. ***E*** Histograms of the number of circuits, normalized by protein length, for all long-range exclusion thresholds. ***F*** Triangular plot of CT fractions for residue-based CT. The color code indicates the number of circuits normalized by size, calculated with 36 residues long-range exclusion threshold.

the order in which they appear along the chain. Each contact site will then be represented in the string by that letter; consequently, each letter will appear in the sequence twice, as each contact is formed by two contact sites (residues, in this case). Thus, if we take the diagram shown in Figure 5A, and we follow the chain from beginning to end, we first encounter contact A, then contact B, then C, B, C and finally A. Therefore, the string notation is **ABCBCA**. The choice of letter (or general symbol to identify the contact) is arbitrary: the notation is valid as long as the symbol used is unique to that contact in the string. Each segment of chain which consists of full pairs of letters represents a circuit. Therefore, the chain identified by **ABCBCA** is itself a circuit. **BCBC** is also a circuit, while **ABCB** is not. In Figure 5B, **ABBCCA** is a circuit, as also **BB** and **CC**. These are topologically independent units: the circuit **CC** could be removed, and the topology of **ABBA** would be unaffected: **ABBA** would still be a circuit. However, which circuit should we consider, when decomposing a longer chain? **ABBCCA** or the shorter **BB**, **CC**? This depends on the threshold we impose for the exclusion of long-range contacts. Three thresholds were tested on our dataset: 12, 24 and 36 residues. Imposing a threshold implies, for example, that contacts which are formed by residues that are more than 12 residues apart along the chain are erased, in order to reveal the self-contained topological sub structures of this length range. The retrieved number of circuits is related to the size of the protein in a non-trivial way [25]. The number of circuits in a protein can be considered as its topological size. Topological and geometrical size are clearly two closely related concepts. Here we show, however, that the information provided by these two parameters is not redundant. Correlations between protein size and number of circuits decrease as we increase the threshold for long range contact exclusion in circuit computation. Correlations go from being as high as r = 0.91 (p = 1.85e^{-47}) for t_{lr} = 12, to r = 0.64 (p = 2.98e^{-15}) for t_{lr} = 36. This consideration sets the lower boundary for our analysis, as for thresholds which are as low as 12 residues, the detected topological length size coincides with geometrical size. This conclusion becomes apparent when we normalize the number of circuits by protein length, and use the normalized number of circuits as folding rate predictor (Figure 5C). While we observe no correlation between normalized number of circuits and ln(k_f) for t_{lr} = 12, the correlation increases as we higher the threshold. The correlation is particularly pronounced for two-state folders, for which we observe significant correlations for both t_{lr} = 24 (r = 0.49, p = 3.05e^{-6}) and t_{lr} = 36 (r = 0.50, p = 2.55e^{-6}). Multi-state folders, on the other hand, only display correlation for t_{lr} =36, which is also weaker in magnitude as opposed to that of two-state proteins (r = 0.32, p = 0.039). This result is interesting especially if we consider how traditionally size as a folding rate predictor was particularly successful when applied to multi-state folders. Observing a significant, albeit weak correlation for multi-state folders for the normalized number of circuits indicates that topological and geometrical size are not always equivalent concepts. This consideration is particularly true when we consider two-state folders, where size generally provides only modest correlations. Indeed, for this dataset the correlation between protein length and ln(k_f) for two-state folders is r = - 0.28, p = 0.010 (Figure S3); this finding suggests that for two-state proteins the topological length might be a better descriptor for folding kinetics than geometrical size. In general, the correlations in Figure 5C suggest that, for proteins of comparable length, a subdivision in a higher number of topologically independent units might facilitate folding. Moreover, the size of the circuits also seems to matter for two-state folders, with proteins made up by smaller circuits folding faster (Figure 5D).

The distribution of the normalized number of circuits for two and multi-state folders also contains crucial information, concerning the topological makeup of the two folder types (Figure 5E). While for t_{lr} = 12 and t_{lr} = 24 we do not observe any particular difference between the two distributions, for t_{lr} = 36 we actually observe a shift between the two, with two-state folders having a longer distribution tail towards high values of normalized number of circuits. For t_{lr}= 36 residues, the multi and two-state distributions for normalized number of circuits are statistically different, as quantified by the Mann-Whitney U test (p=5.05e^{-4}). There is still significant overlap between the two distributions for low values of normalized number of circuits, indicating, again, that the difference between two and multi-state folders is not binary. Nevertheless, the results suggest that topology might be informative not only of the speed but also of the quality of the folding process.

Concerning the topological content of the circuits, we do not observe a clear trend between the normalized number of circuits and topological fractions (Figure 5F). While short-range contacts contained inside one circuit tend to be in series with local contacts present in other circuits, we also find that a relatively high number of normalized circuits can be compatible with high percentages of entangled relations. This enrichments in cross and parallel fractions can be due to the fact that circuits favor tight knit local interaction and tend to bring protein strands closer together. Moreover, circuits can also create long-range entangled relationships with each other, which are generally ignored in the computation of circuits, if they happen for residues which are more distant along the chain than the threshold for long-range exclusion.

3. DISCUSSION

Thanks to the theoretical framework of CT, we were able to draw a correlation between topological properties and folding kinetics, disentangling the role of topology from that of geometry. A significant step in the direction of topological description of folding phenomena was undertaken by Nikolay V. Dokholyan et al, who demonstrated that average graph connectivity was a determinant of folding probability for pre-transition and post-transition states in the protein folding pathway [42]. Different approaches drawn from knot theory were also used to describe the entanglement, torsion and writhe of the protein backbone [15]–[18], devising topologically inspired descriptors which yielded fairly good correlations with the logarithm of the protein folding rate [15][18]. Here, we have taken a fundamentally new step forward, by showing how folding rate can be predicted by CT parameters. Circuit topology (as presented in this study) only focuses on contacts, therefore ignoring the entanglement of the backbone. Moreover, this method does not require cumbersome mathematical and computational operations to connect the ends of the chain, such as those applied by Sulkowska et al. [43]. CT not only considers the number of contacts in the protein, but also shows that there are differential patterns of correlations with respect to the topological arrangement of the contacts. Series, parallel and cross contacts are invariant with respect to shrinking, bending, stretching and other continuous transformations [20], thus present true topological features of protein folds. Our analysis reveals that CT and CO have complementary ranges of validity and can be coupled to predict with accuracy the folding rate of a protein within a wide range of sizes and folding complexity. Moreover, CT offers invaluable information about what type of topological arrangements favor or hinder folding, therefore adding a mechanistic insight to folding rate prediction. The evidence of zipper effect for short range, attractive energy contacts offers a generalized model for folding which resolves the qualitative discrepancy between two-state and multi-state proteins. This unified view is beneficial since often attribution to two-state or multi-state classes is somewhat arbitrary [30], and the folding state of a protein might also not be known *a priori*. Moreover, we found that zipper effect yields a particularly high correlation for multi-state folders, which were previously found to mainly correlate with protein length [28].

Although the current implementation of CT ignores backbone entanglement, one can consider a comparison between the correlation scores obtained by CT analysis and those extracted by other topologically inspired descriptors such as torsion, writhe, Gaussian linking number and its linear combinations with relative and absolute contact order [15][18]. It is natural to compare the results obtained in Figure 1F for segments to the analysis carried out by Panagiotou and plaxco, about torsion and writhe of the protein backbone. They obtained correlation scores as high as 0.48 and 0.45 for writhe and torsion respectively, with respect to the logarithm of the folding rate. We obtain comparable results when considering the parallel relation. However, we only obtain it for multi-state folders, while writhe and torsion correlation values were only provided for two-state folders[18]. A combination of the two approaches might provide a more complete description for protein folding kinetics at the secondary structure level. For what concerns Gaussian entanglement, correlations as high as -0.64 and -0.74 were obtained for two and multi-state proteins respectively [15], with correlations increasing when these scores were combined with RCO and ACO. However, these results were obtained on small datasets (26 two- and 22 multi-state proteins); it is important to take into account that this type of analysis is sensitive to the size and characteristics of the dataset [15]. In fact, CT provides comparable scores when applied to smaller subsections of the datasets, with sizes comparable to the ones in this study (Figure 2 and 3). Moreover, combining CT with traditionally used descriptors such as CO and protein length allows for an increase in the predictive power of both parameters (Figure 4). One might also consider the advantages of combining contact- and entanglement-based descriptors for protein folding prediction. Generalized CT [25][44] expands CT concepts to entangled subloops of a chain (the so-called *soft contacts*), therefore offering the opportunity for such complete description in future research.

The statistically significant correlations found between folding rate and the topology of short-range contacts, as well as the number of circuits, suggest that folding happens primarily at the circuit level. We might find parallels between the concept of topological circuits and the one of local elementary structure (LES) envisioned by the hierarchical model of protein folding [32][45]. Following this reasoning, one would envision a folding model in which folding occurs early on inside the circuits, and at a later stage the circuits are arranged with respect to each other, forming inter-circuit contacts. This type of folding model also matches the 'zipping and assembly' mechanism theorized by S. Bano Ozkan and co-workers [46]. This folding mechanism would be compatible with our observations that the topology of long-range contacts correlates with folding rate only for slow-folding proteins (Figure 2E). Circuits presumably represent the elementary topological units of folding. The correlations between normalized number of circuits and folding rate for two-state folders (and to a lesser extent, multi-state folders) indicate that, for proteins of comparable sizes, the ones that present multiple, small folding elementary units will fold faster. The high correlations obtained by two-state folders shed light on the nature of the different mechanisms experienced by two and multi-state proteins during folding. In particular, it would seem that topological length, as opposed to geometrical length, might play a role in folding rate prediction for two-state proteins.

These insights into the role of native topology offer not only new tools for theoretical understanding of protein kinetics, but also powerful principles for protein design. The framework of CT has already proved to be effective in the field of molecular engineering [47]. The zipper effect and circuit decomposition might provide an easily applicable topological prescription for obtaining proteins with the desired kinetic properties.

4. METHODS

All proteins, CO and kinetic information were retrieved from the ARCPro dataset[29]. Contact order retrieved from the ARCPro dataset was computed as the absolute contact order (ACO) based on a 6 Å cutoff for determining contacts in a multiple contact all-heavy atom method. Four proteins were excluded from analysis: 1FMK, 1M9S and 2BLM for incompleteness of structural information in the PDB files, and 1RA9 for incompleteness in kinetic information in the dataset. Therefore, the whole dataset for analysis comprised 122 proteins. The sub-datasets contained the following number of proteins: 36 proteins for Lower CO (multi-state: 7, two-state: 27), 58 proteins for Average CO (multi-state: 18, two-state: 40) and 28 for Upper CO (multi-state: 16, two-state: 12). The partitioning of the dataset into CO ranges was made by calculating mean x and standard deviation σ_x of the CO distribution and defining the following thresholds:

$$t_{Upper} = x + \frac{\sigma_x}{2}$$

$$t_{Lower} = x - \frac{\sigma_x}{2}.$$

Circuit Topology parameters were retrieved using custom-made Python code, which allows for energy, length filtering and circuit decomposition options. All PDBS are pre-processed automatically before analysis, in order to remove water molecules, hydrogen atoms and various binders. Only one chain (the first contained in the PDB) is selected.

Contacts between segments were calculated based on a distance cutoff of 5.0 Å and a cutoff in number of atoms equal to 10. See Supplementary information for distance cutoffs equal to 3.5, 4.0, 5.0, 5.5 and 6 Å (Figure S4). For the definition of segments, the secondary structure files of the proteins as produced by the free web service STRIDE[48] were used. Each secondary structural element as defined in the STRIDE file represents a segment, to which contacts formed by atoms included in the segment are assigned.

Contacts between residues were calculated based on a distance cutoff of 5.0 Å. Residues were deemed to be in contact when more than $n_a = 5$ atoms were found to be closer than the distance cutoff. We repeated the analysis for cutoffs equal to 3.5, 4.0, 4.5, 5.5 and 6.0 Å (Figure S5) and for $n_a = 1, 2, 3, 4, 5$ and 6 (Figure S6). The four closest neighbors of each residue were excluded from analysis. Each retrieved contact site in the protein structure was given an index. Indexes were given based on the order in which contact sites appeared along the protein chain, from left end to right end of the chain. In this way, each contact was characterized by the two indexes (i, j) of its constituent contact sites. In order to define the CT relation between two contacts, their contact indexes (i,j) and (r,s) were compared. CT relations were assigned based on the mathematical relations summarized below:

$$C_{i,j} S\, C_{r,s} \Leftrightarrow [i,j] \cap [r,s] = \varnothing$$

$$C_{i,j} P C_{r,s} \Leftrightarrow [i,j] \subset (r,s)$$

$$C_{i,j} X\, C_{r,s} \Leftrightarrow [i,j] \cap [r,s] \not\in \{[i,j],[r,s]\} \cup P(\{i,j,r,s\})$$

$$C_{i,j} CS\, C_{r,s} \Leftrightarrow (([i,j] \cap [r,s] = \{i\}) \vee ([i,j] \cap [r,s] = \{j\}))$$

$$C_{i,j} CP\, C_{r,s} \Leftrightarrow (([i,j] \subset [r,s]) \wedge (i = r \vee j = s))$$

P denotes the powerset i.e., all subsets of a set including the null set (ø). The topological relations introduced above are sufficient and necessary to describe the topology of any folded linear chain with binary contacts[20]. For simplicity, CP and CS relations were counted respectively as parallel and series in the analysis presented in this paper. One can readily adjust the set theoretic definition to reduce the relation set {P, S, X, CP, CS} to {P, S, X} and to make the parallel relation symmetric so that P = P^{-1}:

Series: $$C_{i,j} S\, C_{r,s} \Leftrightarrow [i,j] \cap [r,s] \subset \{i,j,r,s\}$$

Parallel: $$C_{i,j} P C_{r,s} \Leftrightarrow [i,j] \subset [r,s] \vee [r,s] \subset [i,j]$$

Cross: $$C_{i,j} X\, C_{r,s} \Leftrightarrow [i,j] \cap [r,s] \not\in \{[i,j],[r,s]\} \cup P(\{i,j,r,s\})$$

Correlation analysis for segments and residues subdivided in CO subgroups, for different distance cutoffs, can be seen in Figure S7 and Figure S8. Distance filtering (short range versus long range contacts) was carried out with a threshold for long range exclusion of 24 residues. The analysis was also repeated for thresholds equal to 12 and 36 residues (Figure S9). Energy filtering was carried out by exploiting the statistical potential matrix calculated by P. Thomas and K.Dill[37]. The Pearson correlation coefficient and two-tailed p value were calculated by custom-made data analysis Jupyter lab files. All correlation maps shown in the paper display correlations with p value ≤ 0.05.

Multilinear regression was performed by using an ordinary least squares Linear Regression from the linear_model module in scikit-learn 0.24.2. The subdivision into subsequent training and set tests was performed by model_selection module, with the KFold function. The five sets are formed respectively by: protein 1 to 25, protein 26 to 49, protein 50 to 73, protein 74 to 97 and 98 to 121. Indexes refer to those assigned to proteins in the ARCPro database.

Residuals from folding rate prediction were tested for normality with the Shapiro test. Distributions with P values < 0.05 were considered not normal. In order to evaluate the quality of the prediction, the determination coefficient R^2 was used, as calculated by the metrics.r2_score function:

$$R^2(y,\hat{y})=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

where $\hat{y}_i$ is the predicted value of the i-th data point, y_i is the corresponding true value, n the total number of samples and $\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$. The adjusted determination coefficient is defined as:

$$R^2_{adj}=1-\frac{(1-R^2)(n-1)}{n-p-1}$$

where n is the number of samples and p the number of predictors (independent variables).

Circuit decomposition and counting was performed by setting a threshold on the length of the circuits. Given the average length *l* of the circuits in a protein, and σl their standard deviation, the circuits with a length below a threshold $t_l = l + \frac{\sigma_l}{2}$ were discarded. This was done under the assumption that the folding speed of bigger circuits represents the bottleneck for folding rate, and therefore the smaller circuits are negligible. Results without application of threshold t_l are displayed in Figure S10.

5. REFERENCES

[1] D. Baker, "A surprising simplicity to protein folding," Nature, vol. 405, no. 6782, pp. 39–42, May 2000, doi: 10.1038/35011000.

[2] K. W. Plaxco, K. T. Simons, and D. Baker, "Contact order, transition state placement and the refolding rates of single domain proteins," J Mol Biol, vol. 277, no. 4, pp. 985–994, 1998, doi: 10.1006/jmbi.1998.1645.

[3] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein, "Contact order revisited: Influence of protein size on the folding rate," Protein Science, vol. 12, no. 9, pp. 2057–2062, 2003, doi: 10.1110/ps.0302503.

[4] P. Sormanni et al., "Simultaneous quantification of protein order and disorder," Nature Chemical Biology, vol. 13, no. 4, pp. 339–342, Apr. 2017, doi: 10.1038/nchembio.2331.

[5] H. Kaya and H. S. Chan, "Solvation Effects and Driving Forces for Protein Thermodyna-

mic and Kinetic Cooperativity: How Adequate is Native-centric Topological Modeling?," Journal of Molecular Biology, vol. 326, no. 3, pp. 911–931, Feb. 2003, doi: 10.1016/S0022-2836(02)01434-1.

[6] H. Zhou and Y. Zhou, "Folding rate prediction using total contact distance," Biophysical Journal, vol. 82, no. 1, pp. 458–463, 2002, doi: 10.1016/S0006-3495(02)75410-6.

[7] L. Censoni and L. Martínez, "Prediction of kinetics of protein folding with non-redundant contact information," Bioinformatics, vol. 34, no. 23, pp. 4034–4038, 2018, doi: 10.1093/bioinformatics/bty478.

[8] Y. Li, Y. Zhang, and J. Lv, "An Effective Cumulative Torsion Angles Model for Prediction of Protein Folding Rates," Protein & Peptide Letters, vol. 27, no. 4, pp. 321–328, Mar. 2020, doi: 10.2174/0929866526666191014152207.

[9] D. N. Ivankov and A. V. Finkelstein, "Prediction of protein folding rates from the amino acid sequence-predicted secondary structure," Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 24, pp. 8942–8944, 2004, doi: 10.1073/pnas.0402659101.

[10] H. Gong, D. G. Isom, R. Srinivasan, and G. D. Rose, "Local secondary structure content predicts folding rates for simple, two-state proteins," Journal of Molecular Biology, vol. 327, no. 5, pp. 1149–1154, 2003, doi: 10.1016/S0022-2836(03)00211-0.

[11] M. M. Gromiha and S. Selvaraj, "Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction," Journal of Molecular Biology, vol. 310, no. 1, pp. 27–32, 2001, doi: 10.1006/jmbi.2001.4775.

[12] Z. Ouyang and J. Liang, "Predicting protein folding rates from geometric contact and amino acid sequence," Protein Science, vol. 17, no. 7, pp. 1256–1263, 2008, doi: 10.1110/ps.034660.108.

[13] M. M. Gromiha, "Multiple Contact Network Is a Key Determinant to Protein Folding Rates," Journal of Chemical Information and Modeling, vol. 49, no. 4, pp. 1130–1135, Apr. 2009, doi: 10.1021/ci800440x.

[14] L. L. Chavez, J. N. Onuchic, and C. Clementi, "Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates," Journal of the American Chemical Society, vol. 126, no. 27, pp. 8426–8432, 2004, doi: 10.1021/ja049510+.

[15] M. Baiesi, E. Orlandini, F. Seno, and A. Trovato, "Exploring the correlation between the folding rates of proteins and the entanglement of their native states," Journal of Physics A: Mathematical and Theoretical, vol. 50, no. 50, p. 504001, Dec. 2017, doi: 10.1088/1751-8121/aa97e7.

[16] M. Baiesi, E. Orlandini, F. Seno, and A. Trovato, "Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding," Sci Rep, vol. 9, no. 1, pp. 1–12, 2019, doi: 10.1038/s41598-019-44928-3.

[17] M. Baiesi, E. Orlandini, A. Trovato, and F. Seno, "Linking in domain-swapped protein dimers," Sci Rep, vol. 6, pp. 1–11, 2016, doi: 10.1038/srep33872.

[18] E. Panagiotou and K. W. Plaxco, "A topological study of protein folding kinetics," ArXiv, pp. 1–13, 2018, doi: 10.1090/conm/746/15010.

[19] E. Shakhnovich, “To knot or not to knot?,” Nature Materials, vol. 10, no. 2, pp. 84–86, Feb. 2011, doi: 10.1038/nmat2953.

[20] A. Mashaghi, R. J. Van Wijk, and S. J. Tans, “Circuit topology of proteins and nucleic acids,” Structure, vol. 22, no. 9, pp. 1227–1237, 2014, doi: 10.1016/j.str.2014.06.015.

[21] A. Mashaghi, “Circuit Topology of Folded Chains,” Not. Am. Math. Soc., vol. 68, pp. 420–423, 2021, doi: 10.1090/noti2241.

[22] M. Heidari, H. Schiessel, and A. Mashaghi, “Circuit Topology Analysis of Polymer Folding Reactions,” ACS Central Science, vol. 6, p. 839−847, May 2020, doi: 10.1021/acscentsci.0c00308.

[23] B. Scalvini et al., “Topology of Folded Molecular Chains: From Single Biomolecules to Engineered Origami,” Trends in Chemistry, vol. 2, no. 7, pp. 609–622, 2020, doi: 10.1016/j.trechm.2020.04.009.

[24] A. Mugler, S. J. Tans, and A. Mashaghi, “Circuit topology of self-interacting chains: Implications for folding and unfolding dynamics,” Physical Chemistry Chemical Physics, vol. 16, no. 41, pp. 22537–22544, 2014, doi: 10.1039/c4cp03402c.

[25] A. Golovnev and A. Mashaghi, “Generalized Circuit Topology of Folded Linear Chains,” iScience, vol. 23, no. 9, p. 101492, 2020, doi: 10.1016/j.isci.2020.101492.

[26] K. L. Maxwell et al., “Protein folding: Defining a ‘standard’ set of experimental conditions and a preliminary kinetic data set of two-state proteins,” Protein Science, vol. 14, no. 3, pp. 602–616, Mar. 2005, doi: 10.1110/ps.041205405.

[27] D. Barrick, “What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding?,” Physical Biology, vol. 6, no. 1, p. 015001, Feb. 2009, doi: 10.1088/1478-3975/6/1/015001.

[28] O. V. Galzitskaya, S. O. Garbuzynskiy, D. N. Ivankov, and A. V. Finkelstein, “Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics,” Proteins: Structure, Function, and Genetics, vol. 51, no. 2, pp. 162–166, May 2003, doi: 10.1002/prot.10343.

[29] A. S. Wagaman, A. Coburn, I. Brand-Thomas, B. Dash, and S. S. Jaswal, “A comprehensive database of verified experimental data on protein folding kinetics,” Protein Science, vol. 23, no. 12, pp. 1808–1812, 2014, doi: 10.1002/pro.2551.

[30] S. E. Jackson, “How do small single-domain proteins fold?,” Folding and Design, vol. 3, no. 4, pp. 81–91, 1998, doi: 10.1016/S1359-0278(98)00033-9.

[31] R. Doyle, K. Simons, H. Qian, and D. Baker, “Local Interactions and the Optimization of Protein Folding,” vol. 291, no. January, pp. 282–291, 1997.

[32] R. A. Broglia and G. Tiana, “Hierarchy of events in the folding of model proteins,” Journal of Chemical Physics, vol. 114, no. 16, pp. 7267–7273, 2001, doi: 10.1063/1.1361076.

[33] P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, and A. Laio, “Exploring the universe of protein structures beyond the protein data bank,” PLoS Computational Biology, vol. 6, no. 11, 2010, doi: 10.1371/journal.pcbi.1000957.

[34] J. C. Phys and A. Irba, “Local interactions and protein folding : A,” vol. 273, no. October 1996, 2019.

[35] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, “Impact of local and non-local interactions on thermodynamics and kinetics of protein folding,” Journal of Molecular Biology, vol. 252, no. 4, pp. 460–471, 1995, doi: 10.1006/jmbi.1995.0511.

[36] A. Kumar, A. Baruah, and P. Biswas, “Role of local and nonlocal interactions in folding and misfolding of globular proteins,” Journal of Chemical Physics, vol. 146, no. 6, 2017, doi: 10.1063/1.4975325.

[37] P. D. Thomastt and K. E. N. A. Dill, “An iterative method for extracting energy-like quantities from protein structures,” vol. 93, no. October, pp. 11628–11633, 1996.

[38] P. D. Thomas and K. a. Dill, “Structures : How Accurate Are They ? g g,” Journal of molecular biology, vol. 257, pp. 457–469, 1996.

[39] O. Kramer, “Scikit-Learn,” 2016, pp. 45–53. doi: 10.1007/978-3-319-33383-0_5.

[40] A. Di Bucchianico, “Coefficient of Determination (R 2),” in Encyclopedia of Statistics in Quality and Reliability, Chichester, UK: John Wiley & Sons, Ltd, 2008. doi: 10.1002/9780470061572.eqr173.

[41] S. W. Englander and L. Mayne, “The nature of protein folding pathways,” Proceedings of the National Academy of Sciences of the United States of America, vol. 111, no. 45, pp. 15873–15880, 2014, doi: 10.1073/pnas.1411798111.

[42] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, “Topological determinants of protein folding,” Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 13, pp. 8637–8641, 2002, doi: 10.1073/pnas.122076099.

[43] J. I. Sulkowska, E. J. Rawdon, K. C. Millett, J. N. Onuchic, and A. Stasiak, “Conservation of complex knotting and slipknotting patterns in proteins,” Proceedings of the National Academy of Sciences, vol. 109, no. 26, pp. E1715–E1723, Jun. 2012, doi: 10.1073/pnas.1205918109.

[44] L. Chains, “SS symmetry Coloring Invariant for Topological Circuits in Folded,” pp. 1–12, 2021.

[45] R. L. Baldwin and G. D. Rose, “Is protein folding hierarchic ? I . Local structure and peptide folding,” vol. 0004, no. 98, pp. 26–33, 1999.

[46] S. B. Ozkan, G. A. Wu, J. D. Chodera, and K. A. Dill, “Protein folding by zipping and assembly,” Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 29, pp. 11987–11992, 2007, doi: 10.1073/pnas.0703700104.

[47] V. Kočar et al., “Design principles for rapid folding of knotted DNA nanostructures,” Nat Commun, vol. 7, no. 1, p. 10803, Apr. 2016, doi: 10.1038/ncomms10803.

[48] M. Heinig and D. Frishman, “STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins,” Nucleic Acids Research, vol. 32, no. Web Server, pp. W500–W502, Jul. 2004, doi: 10.1093/nar/gkh429.

7. SUPPLEMENTARY

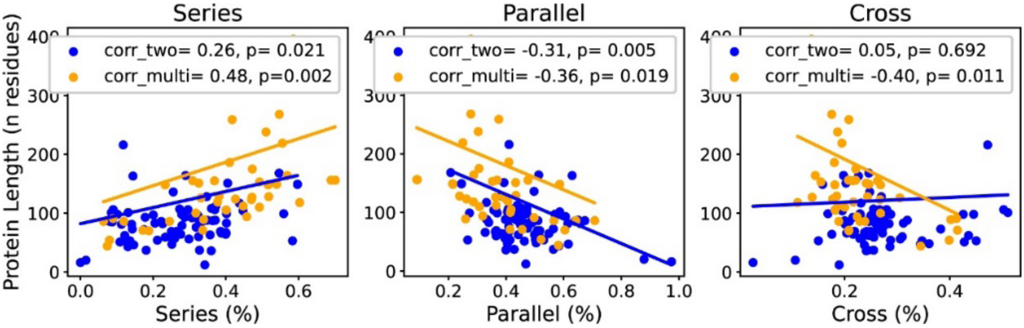


Figure S1. Relationship between CT parameters and protein size. All CT parameters are normalized by the number of contacts in a protein, making it possible to compare proteins with different contacts and sizes. However, a non-trivial relationship between size and CT parameters exists, because of the assembly principles of proteins and geometrical and steric constraints. Series topological content correlates positively with size, while proteins which are relatively richer in entangled fraction tend to be smaller.

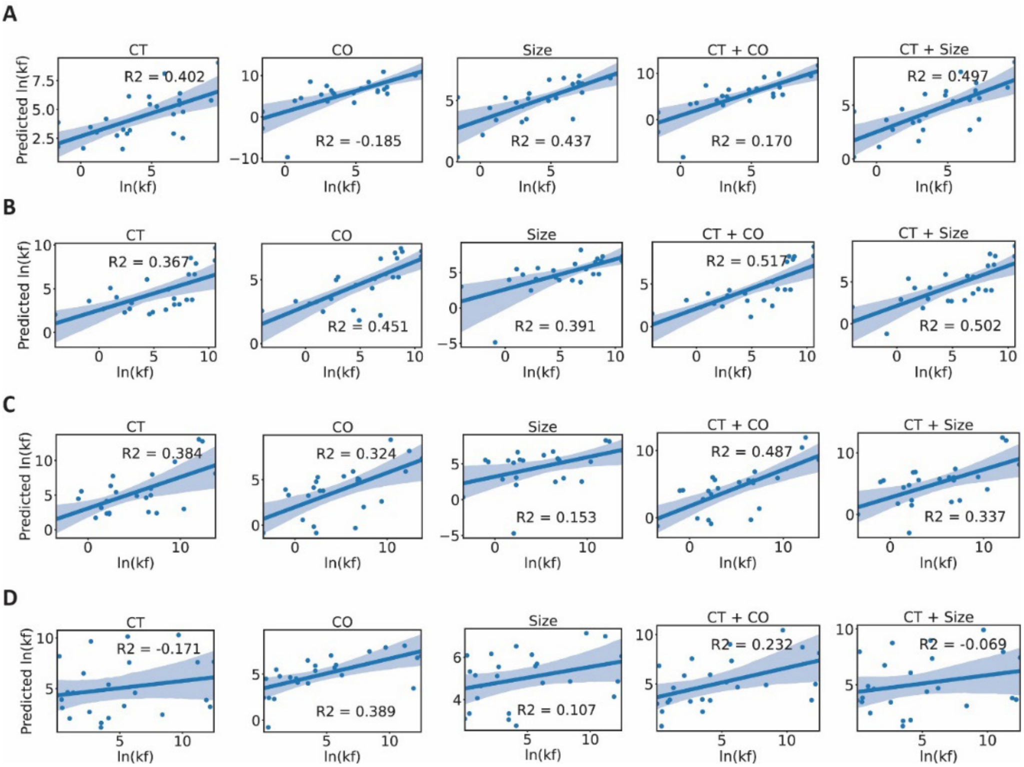


Figure S2. Circuit topology parameters in linear combination with traditional folding rate predictors such as CO and size allow for folding rate prediction with increased statistical significance. A Scatterplots of predicted folding rate (obtained with multilinear regression over CT fractions, CO, protein length and a combination of these parameters) and experimental Folding rate (ln kf), for the first training/test set combination. ***B*** Scatterplots of predicted folding rate (obtained with multilinear regression over CT fractions, CO, protein length and a combination of these parameters) and experimental Folding rate (ln kf), for the second training/test set combination. ***C*** Scatterplots of predicted folding rate (obtained with multilinear regression over CT fractions, CO, protein length and a combination of these parameters) and experimental Folding rate (ln kf), for the third training/test set combination. ***D*** Scatterplots of predicted folding rate (obtained with multilinear regression over CT fractions, CO, protein length and a combination of these parameters) and experimental Folding rate (ln kf), for the fifth training/test set combination.

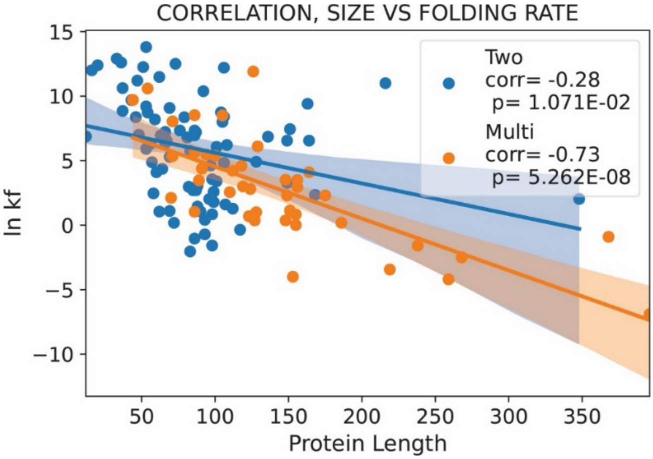


Figure S3. Protein size correlates with folding rate. Scatterplot of protein length versus folding rate (ln kf), for two- and multi-state folders.

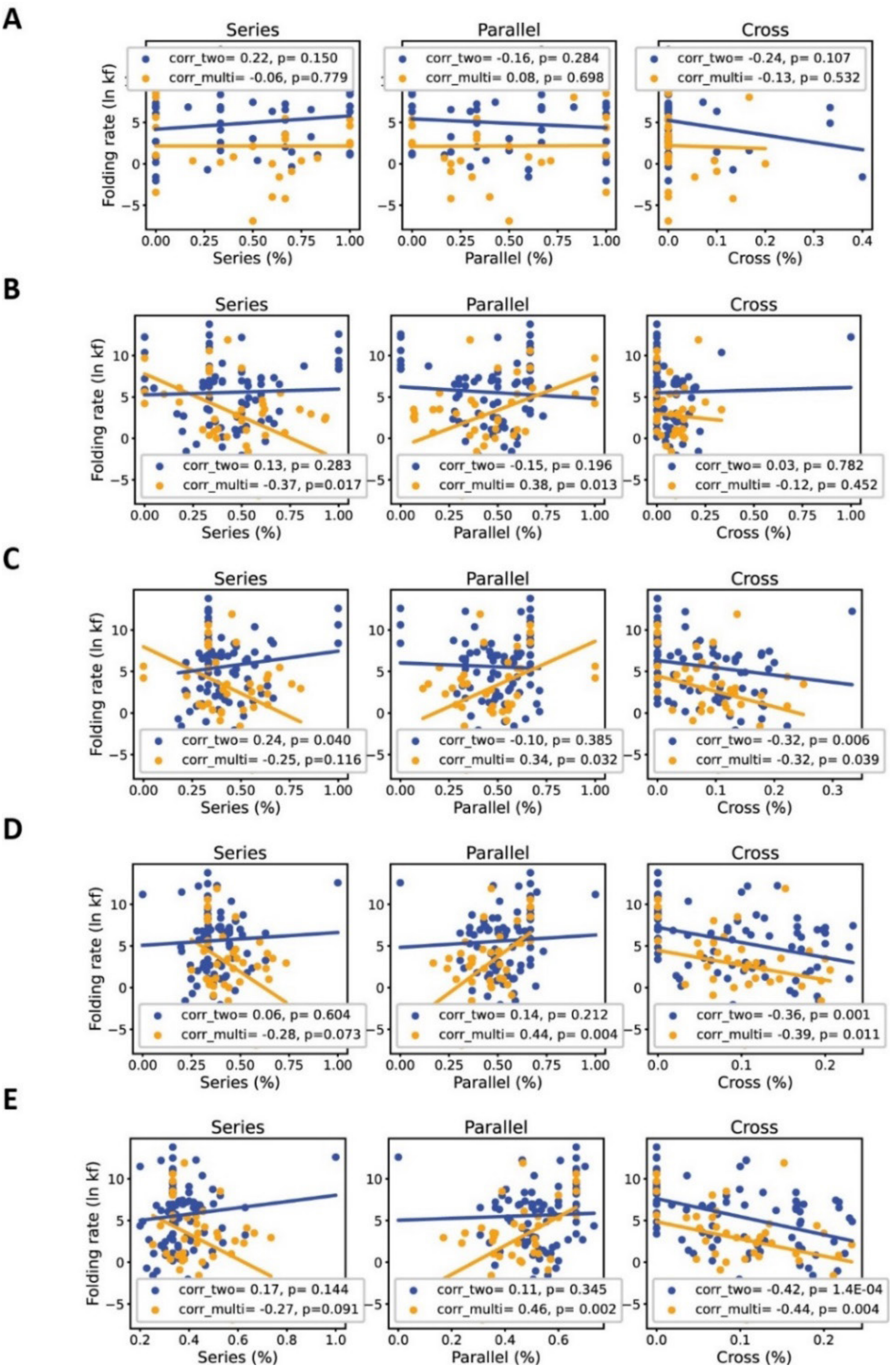

A
Series
corr_two= 0.22, p= 0.150
corr_multi= -0.06, p=0.779
Parallel
corr_two= -0.16, p= 0.284
corr_multi= 0.08, p= 0.698
Cross
corr_two= -0.24, p= 0.107
corr_multi= -0.13, p= 0.532
Folding rate (ln kf)
Series (%)
Parallel (%)
Cross (%)
B
Series
corr_two= 0.13, p= 0.283
corr_multi= -0.37, p=0.017
Parallel
corr_two= -0.15, p= 0.196
corr_multi= 0.38, p= 0.013
Cross
corr_two= 0.03, p= 0.782
corr_multi= -0.12, p= 0.452
Folding rate (ln kf)
Series (%)
Parallel (%)
Cross (%)
C
Series
corr_two= 0.24, p= 0.040
corr_multi= -0.25, p=0.116
Parallel
corr_two= -0.10, p= 0.385
corr_multi= 0.34, p= 0.032
Cross
corr_two= -0.32, p= 0.006
corr_multi= -0.32, p= 0.039
Folding rate (ln kf)
Series (%)
Parallel (%)
Cross (%)
D
Series
corr_two= 0.06, p= 0.604
corr_multi= -0.28, p=0.073
Parallel
corr_two= 0.14, p= 0.212
corr_multi= 0.44, p= 0.004
Cross
corr_two= -0.36, p= 0.001
corr_multi= -0.39, p= 0.011
Folding rate (ln kf)
Series (%)
Parallel (%)
Cross (%)
E
Series
corr_two= 0.17, p= 0.144
corr_multi= -0.27, p=0.091
Parallel
corr_two= 0.11, p= 0.345
corr_multi= 0.46, p= 0.002
Cross
corr_two= -0.42, p= 1.4E-04
corr_multi= -0.44, p= 0.004
Folding rate (ln kf)
Series (%)
Parallel (%)
Cross (%)

Figure S4. CT parameters for segment-based contacts correlate with folding rate, with distance cutoffs ranging from 4.0 to 6.0 Å. A Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for segment-based contacts, calculated with distance cutoff r = 3.5 Å. This cutoff represents the lower limit of our analysis, as 50 proteins out of 122 result devoid of contacts with this contact definition. There are no significant correlations between folding rate and CT parameters with this threshold. ***B*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for segment-based contacts, calculated with distance cutoff r = 4.0 Å. ***C*** Scatterplotof topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for segment-based contacts, calculated with distance cutoff r = 4.5 Å. ***D*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for segment-based contacts, calculated with distance cutoff r = 5.5 Å. ***E*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for segment-based contacts, calculated with distance cutoff r = 6.0 Å.

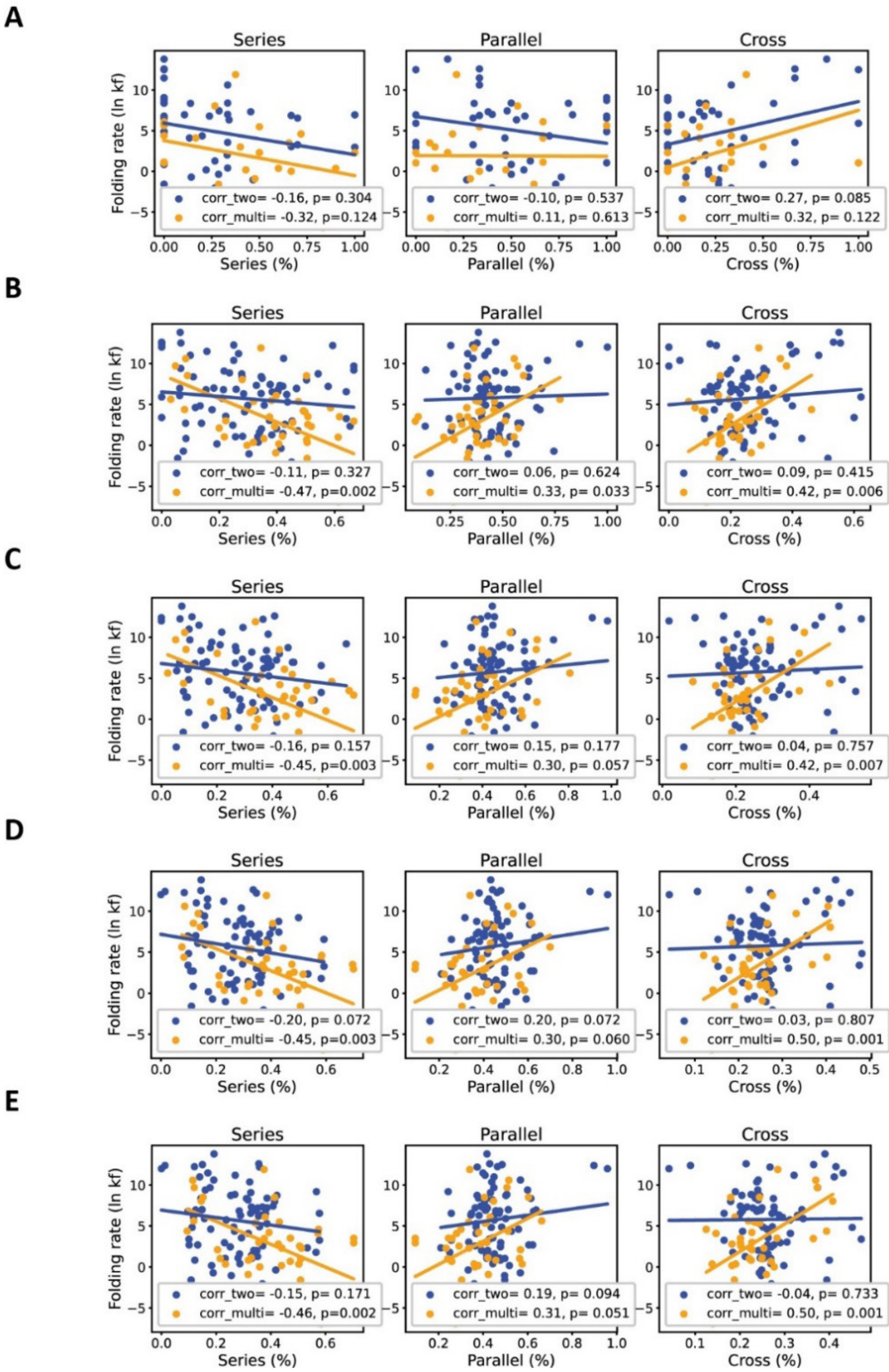
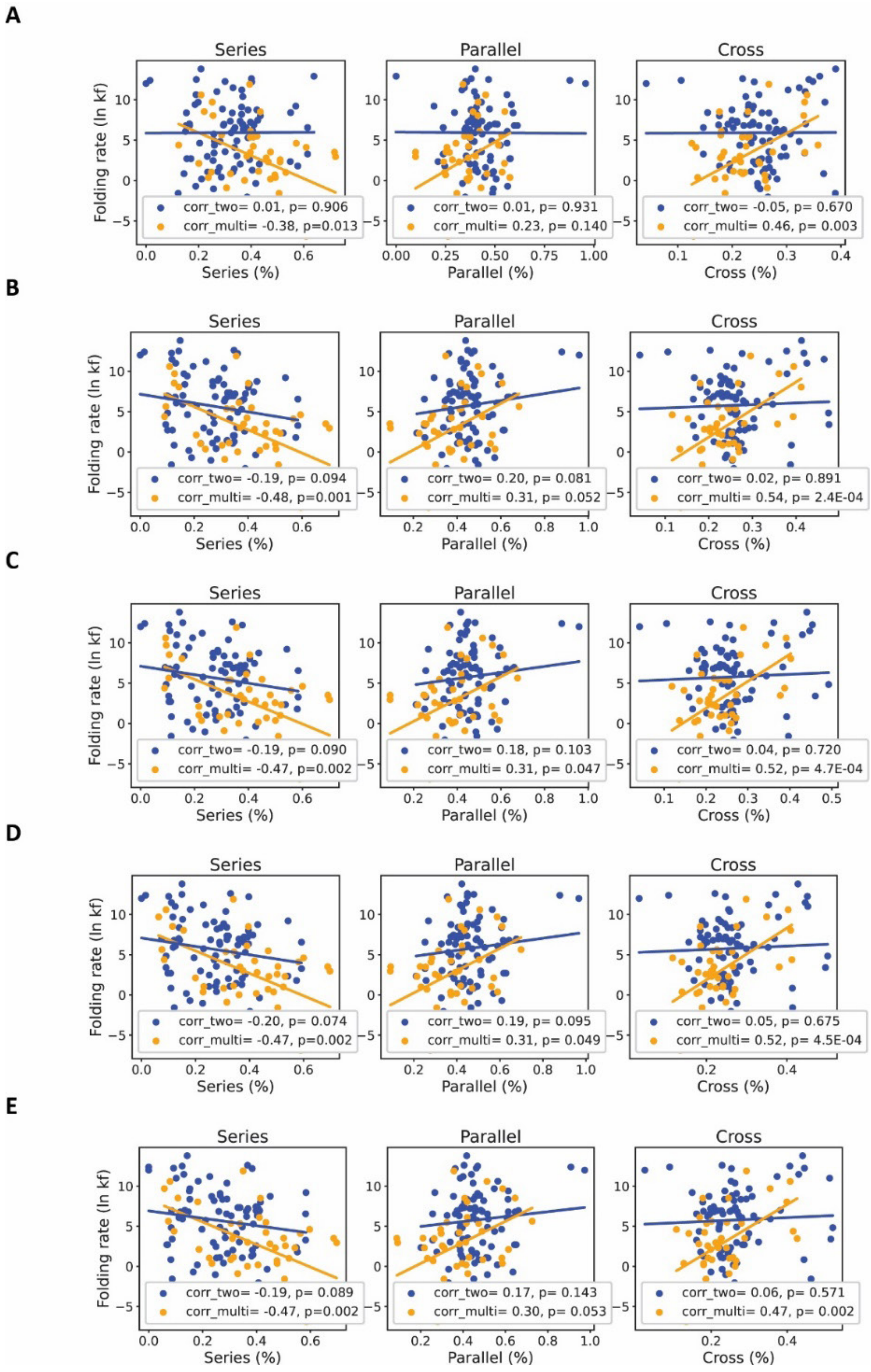

A
B
C
D
E
Series
Parallel
Cross
Folding rate (ln kf)
Series (%)
Parallel (%)
Cross (%)
corr_two= -0.16, p= 0.304
corr_multi= -0.32, p=0.124
corr_two= -0.10, p= 0.537
corr_multi= 0.11, p= 0.613
corr_two= 0.27, p= 0.085
corr_multi= 0.32, p= 0.122
corr_two= -0.11, p= 0.327
corr_multi= -0.47, p=0.002
corr_two= 0.06, p= 0.624
corr_multi= 0.33, p= 0.033
corr_two= 0.09, p= 0.415
corr_multi= 0.42, p= 0.006
corr_two= -0.16, p= 0.157
corr_multi= -0.45, p=0.003
corr_two= 0.15, p= 0.177
corr_multi= 0.30, p= 0.057
corr_two= 0.04, p= 0.757
corr_multi= 0.42, p= 0.007
corr_two= -0.20, p= 0.072
corr_multi= -0.45, p=0.003
corr_two= 0.20, p= 0.072
corr_multi= 0.30, p= 0.060
corr_two= 0.03, p= 0.807
corr_multi= 0.50, p= 0.001
corr_two= -0.15, p= 0.171
corr_multi= -0.46, p=0.002
corr_two= 0.19, p= 0.094
corr_multi= 0.31, p= 0.051
corr_two= -0.04, p= 0.733
corr_multi= 0.50, p= 0.001

Figure S5. CT parameters for residue-based contacts correlate with folding rate, with distance cutoffs ranging from 4.0 to 6.0 Å. A Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with distance cutoff r = 3.5 Å. This cutoff represents the lower limit of our analysis, as 55 proteins out of 122 result devoid of contacts with this contact definition. There are no significant correlations between folding rate and CT parameters with this threshold. ***B*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with distance cutoff r = 4.0 Å. ***C*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with distance cutoff r = 4.5 Å. ***D*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with distance cutoff r = 5.5 Å. ***E*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with distance cutoff r = 6.0 Å.



A
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= 0.01, p= 0.906
corr_multi= -0.38, p=0.013
corr_two= 0.01, p= 0.931
corr_multi= 0.23, p= 0.140
corr_two= -0.05, p= 0.670
corr_multi= 0.46, p= 0.003
Series (%)
Parallel (%)
Cross (%)
B
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.19, p= 0.094
corr_multi= -0.48, p=0.001
corr_two= 0.20, p= 0.081
corr_multi= 0.31, p= 0.052
corr_two= 0.02, p= 0.891
corr_multi= 0.54, p= 2.4E-04
Series (%)
Parallel (%)
Cross (%)
C
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.19, p= 0.090
corr_multi= -0.47, p=0.002
corr_two= 0.18, p= 0.103
corr_multi= 0.31, p= 0.047
corr_two= 0.04, p= 0.720
corr_multi= 0.52, p= 4.7E-04
Series (%)
Parallel (%)
Cross (%)
D
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.20, p= 0.074
corr_multi= -0.47, p=0.002
corr_two= 0.19, p= 0.095
corr_multi= 0.31, p= 0.049
corr_two= 0.05, p= 0.675
corr_multi= 0.52, p= 4.5E-04
Series (%)
Parallel (%)
Cross (%)
E
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.19, p= 0.089
corr_multi= -0.47, p=0.002
corr_two= 0.17, p= 0.143
corr_multi= 0.30, p= 0.053
corr_two= 0.06, p= 0.571
corr_multi= 0.47, p= 0.002
Series (%)
Parallel (%)
Cross (%)

Figure S6. CT parameters for residue-based contacts correlate with folding rate, with r = 5.0 Å and n_a thresholds ranging from 1 to 6. ***A*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with n_a = 1. ***B*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with calculated with n_a = 2. ***C*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with calculated with n_a = 3. ***D*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with n_a = 4. ***E*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for residue-based contacts, calculated with n_a = 6.

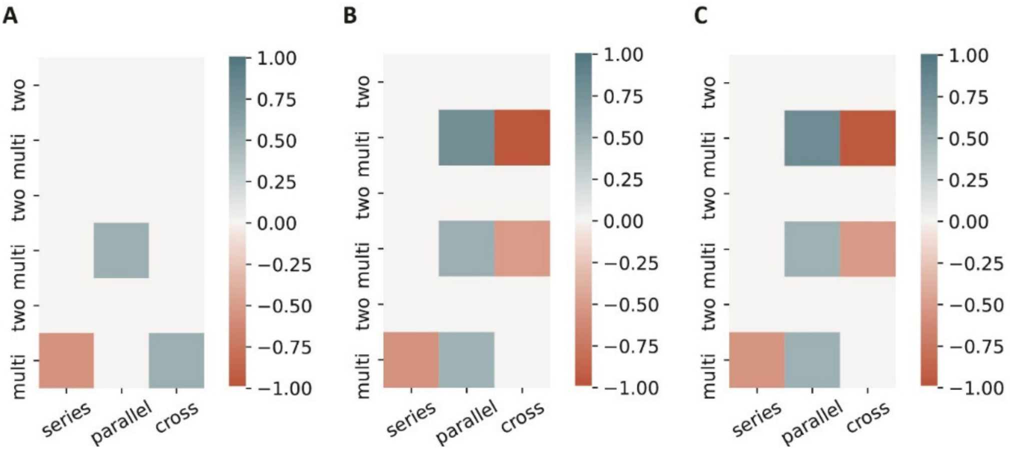


Figure S7. Segment-based CT parameters display differential patterns of correlation with folding rate, which can be highlighted by CO classification. A Folding rate correlation map for segment-based CT, with CO classification, calculated for distance cutoff r=4.0 Å. ***B*** Folding rate correlation map for segment-based CT, with CO classification, calculated for distance cutoff r=5.5 Å. ***C*** Folding rate correlation map for segment-based CT, with CO classification, calculated for distance cutoff r=6.0 Å. Analysis for distance cutoff r=4.5 Å yielded an empty correlation map.

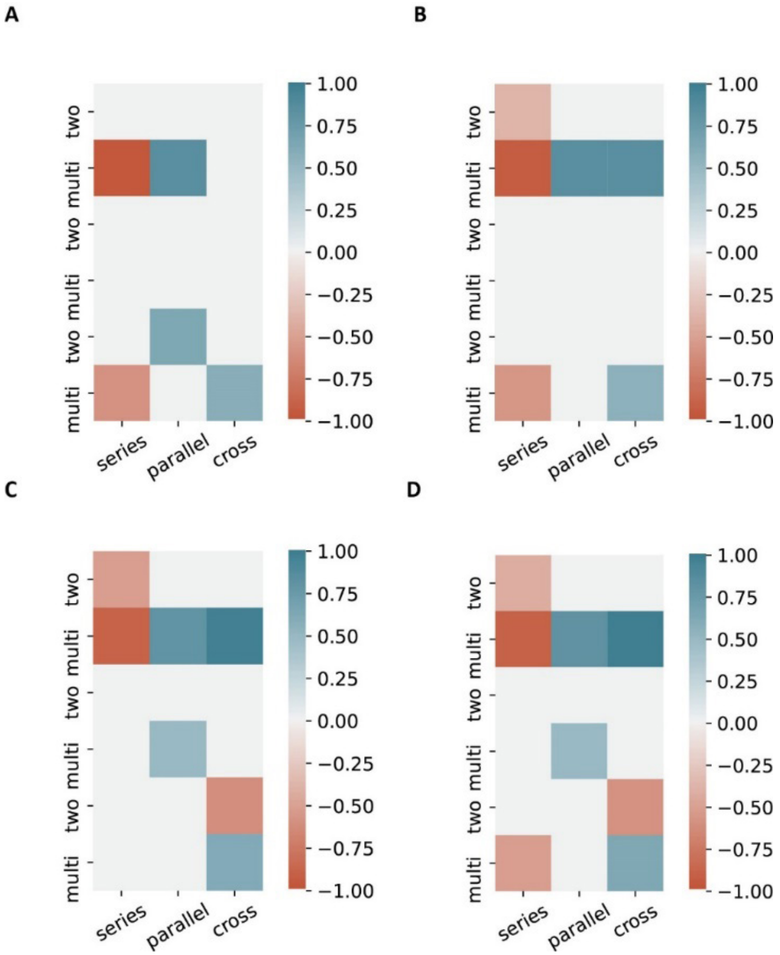


Figure S8. Residue-based CT parameters display differential patterns of correlation with folding rate, which can be highlighted by CO classification. A Folding rate correlation map for residue-based CT, with CO classification, calculated for distance cutoff r=4.0 Å. ***B*** Folding rate correlation map for residue-based CT, with CO classification, calculated for distance cutoff r=4.5 Å. ***C*** Folding rate correlation map for residue-based CT, with CO classification, calculated for distance cutoff r=5.5 Å. ***D*** Folding rate correlation map for residue-based CT, with CO classification, calculated for distance cutoff r=6.0 Å.

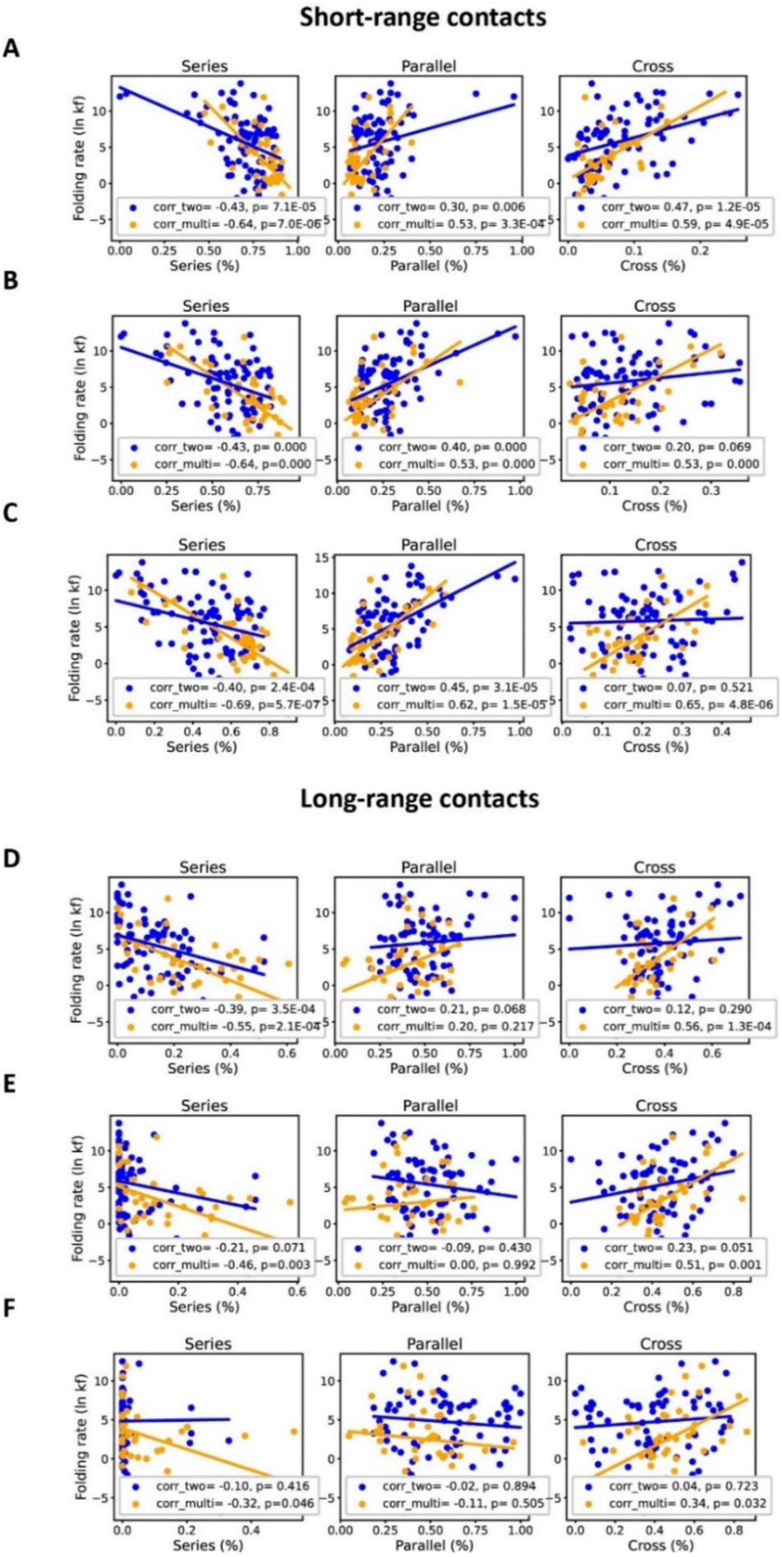

Short-range contacts
A
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.43, p= 7.1E-05
corr_multi= -0.64, p=7.0E-06
corr_two= 0.30, p= 0.006
corr_multi= 0.53, p= 3.3E-04
corr_two= 0.47, p= 1.2E-05
corr_multi= 0.59, p= 4.9E-05
Series (%)
Parallel (%)
Cross (%)
B
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.43, p= 0.000
corr_multi= -0.64, p=0.000
corr_two= 0.40, p= 0.000
corr_multi= 0.53, p= 0.000
corr_two= 0.20, p= 0.069
corr_multi= 0.53, p= 0.000
Series (%)
Parallel (%)
Cross (%)
C
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.40, p= 2.4E-04
corr_multi= -0.69, p=5.7E-07
corr_two= 0.45, p= 3.1E-05
corr_multi= 0.62, p= 1.5E-05
corr_two= 0.07, p= 0.521
corr_multi= 0.65, p= 4.8E-06
Series (%)
Parallel (%)
Cross (%)
Long-range contacts
D
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.39, p= 3.5E-04
corr_multi= -0.55, p=2.1E-04
corr_two= 0.21, p= 0.068
corr_multi= 0.20, p= 0.217
corr_two= 0.12, p= 0.290
corr_multi= 0.56, p= 1.3E-04
Series (%)
Parallel (%)
Cross (%)
E
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.21, p= 0.071
corr_multi= -0.46, p=0.003
corr_two= -0.09, p= 0.430
corr_multi= 0.00, p= 0.992
corr_two= 0.23, p= 0.051
corr_multi= 0.51, p= 0.001
Series (%)
Parallel (%)
Cross (%)
F
Series
Parallel
Cross
Folding rate (ln kf)
corr_two= -0.10, p= 0.416
corr_multi= -0.32, p=0.046
corr_two= -0.02, p= 0.894
corr_multi= -0.11, p= 0.505
corr_two= 0.04, p= 0.723
corr_multi= 0.34, p= 0.032
Series (%)
Parallel (%)
Cross (%)

Figure S9. The topology of local and non-local contacts impacts folding rate in different measures, with short-range contacts displaying overall higher correlations. A Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for short-range residue-based contacts, with a threshold of 12 residues. ***B*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for short-range residue-based contacts, with a threshold of 24 residues. ***C*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for short-range residue-based contacts, with a threshold of 36 residues. ***D*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for long-range residue-based contacts, with a threshold of 12 residues. ***E*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for long-range residue-based contacts, with a threshold of 24 residues. ***F*** Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (ln kf), for long-range residue-based contacts, with a threshold of 36 residues.

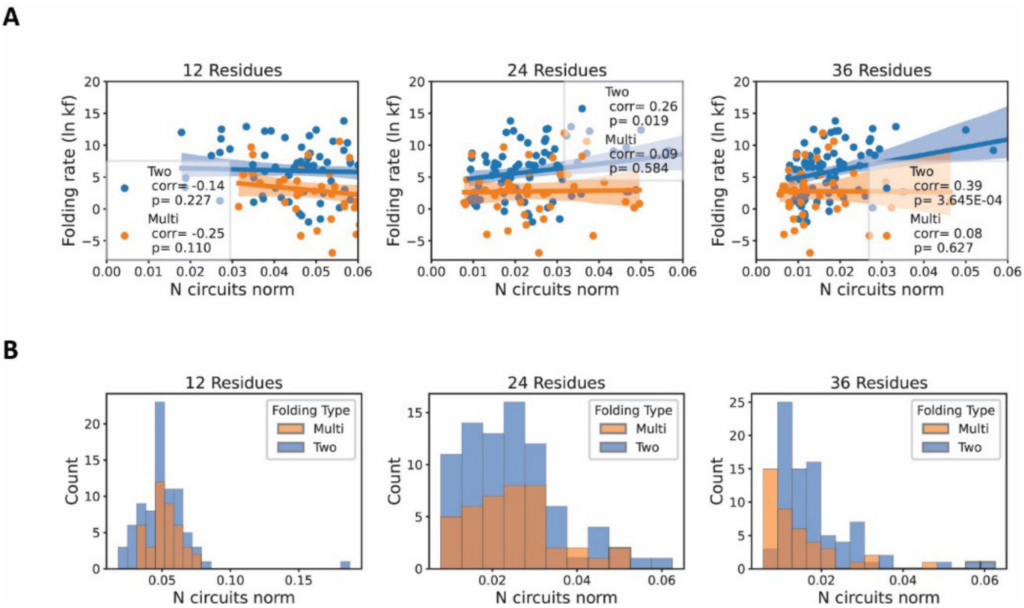


Figure S10. Folding rate correlates positively with the number of topological circuits composing the protein, normalized by size. A Scatterplot of number of circuits normalized by protein length versus folding rate (ln kf). Circuits we calculated with a threshold for long-range exclusion equal to 12, 24 and 36 residues. No additional threshold tl was applied (all circuits were computed regardless of their size). ***B*** Histogram of the number of circuits normalized by protein length for two and multi-state folders, for long-range exclusion equal to 12, 24 and 36 residues. No additional threshold t_l was applied.

SEGMENTS

1.a) LOWER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.10	0.631	0.15	0.497	-0.11	0.617
Multi	-0.75	0.050	0.82	0.025	-0.96	0.001

1.b) AVERAGE CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.09	0.600	-0.06	0.693	-0.02	0.879
Multi	-0.31	0.204	0.51	0.029	-0.45	0.058

1.c) HIGHER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.08	0.803	0.24	0.444	-0.40	0.198
Multi	-0.66	0.006	0.61	0.012	0.16	0.543

Table S1. Correlation coefficients for segment-based CT parameters, subdivided by CO classification. All correlation coefficients were calculated for distance cutoff r=5.0 Å and threshold n_a = 10.

RESIDUES

2.a) LOWER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.45	0.016	0.27	0.157	0.24	0.218
Multi	-0.93	0.002	0.83	0.021	0.94	0.001

2.b) AVERAGE CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.02	0.892	-0.08	0.607	0.08	0.615
Multi	-0.43	0.075	0.43	0.072	0.06	0.802

2.c) HIGHER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.01	0.973	0.53	0.075	-0.59	0.045
Multi	-0.58	0.019	0.33	0.206	0.65	0.006

Table S2. Correlation coefficients for residue-based CT parameters, subdivided by CO classification. All correlation coefficients were calculated for distance cutoff r=5.0 Å and threshold $n_a = 5$.

CONTACT ORDER

	LowerCO(r)	pvalue	AveCO(r)	pvalue	HigherCO(r)	pvalue
Two	-0.037	0.85	-0.529	0.00045	0.044	0.891
Multi	-0.605	0.15	-0.51	0.031	-0.273	0.306

Table S3. Correlation coefficients for contact order and folding rate, subdivided by CO classification. Contact order values refer to Absolute Contact Order (ACO), calculated for a distance cutoff r = 6 Å.

PROTEIN LENGTH

	LowerCO(r)	pvalue	AveCO(r)	pvalue	HigherCO(r)	pvalue
Two	-0.343	0.068	0.225	0.163	0.157	0.626
Multi	-0.889	0.007	-0.459	0.055	-0.607	0.013

Table S4. Correlation coefficients for protein length and folding rate, subdivided by CO classification. Protein length values are expressed in number of residues.

RESIDUES (LR)

2.a) LOWER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.16	0.445	-0.19	0.367	0.17	0.434
Multi	-0.69	0.087	0.39	0.393	0.62	0.135

2.b) AVERAGE CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.11	0.504	-0.16	0.332	0.28	0.075
Multi	-0.55	0.019	0.30	0.231	0.10	0.703

2.c) HIGHER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.13	0.683	0.74	0.006	-0.73	0.007
Multi	-0.48	0.060	0.12	0.649	0.65	0.006

Table S5. Correlation coefficients for long range residue-based CT parameters, subdivided by CO classification. All correlation coefficients were calculated for distance cutoff r=5.0 Å, n_a = 5 and a threshold of 24 residues for range exclusion.

RESIDUES (SR)

2.a) LOWER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.46	0.014	0.46	0.013	0.09	0.650
Multi	-0.97	1.9E-04	0.89	0.007	0.94	0.002

2.b) AVERAGE CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.03	0.846	-0.05	0.779	0.01	0.946
Multi	-0.47	0.048	0.41	0.094	0.18	0.471

2.c) HIGHER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.30	0.346	0.28	0.377	0.19	0.546
Multi	-0.61	0.012	0.55	0.028	0.59	0.016

Table S6. Correlation coefficients for short range residue-based CT parameters, subdivided by CO classification. All correlation coefficients were calculated for distance cutoff r=5.0 Å, n_a = 5 and a threshold of 24 residues for range exclusion.

RESIDUES (E<0)

2.a) LOWER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.47	0.012	0.27	0.160	0.29	0.133
Multi	-0.89	0.007	0.77	0.045	0.85	0.016

2.b) AVERAGE CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.04	0.829	-0.14	0.376	0.14	0.393
Multi	-0.46	0.054	0.51	0.030	0.07	0.785

2.c) HIGHER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.02	0.948	0.58	0.050	-0.48	0.112
Multi	-0.60	0.013	0.37	0.161	0.69	0.003

Table S7. Correlation coefficients for attractive energy residue-based CT parameters, subdivided by CO classification. All correlation coefficients were calculated for distance cutoff r=5.0 Å and threshold $n_a = 5$.

RESIDUES (E>0)

2.a) LOWER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.18	0.381	-0.07	0.744	0.28	0.161
Multi	-0.95	0.001	0.91	0.004	0.64	0.122

2.b) AVERAGE CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	-0.01	0.958	0.03	0.875	-0.03	0.876
Multi	-0.36	0.143	0.29	0.240	0.11	0.652

2.c) HIGHER CO

	series (r)	pvalue	parallel (r)	pvalue	cross (r)	pvalue
Two	0.10	0.761	0.32	0.313	-0.58	0.050
Multi	-0.50	0.048	0.26	0.329	0.49	0.056

Table S8. Correlation coefficients for repulsive energy residue-based CT parameters, subdivided by CO classification. All correlation coefficients were calculated for distance cutoff r=5.0 Å and threshold n_a = 5.

COEFFICIENT OF DETERMINATION (R^2)

Validation set	CT parameters	CO	Size	CT + CO	CT + size
1	0.402	-0.185	0.437	0.170	0.497
2	0.367	0.451	0.391	0.517	0.502
3	0.384	0.324	0.153	0.487	0.337
4	0.385	0.448	0.382	0.541	0.476
5	-0.171	0.389	0.107	0.232	-0.069

Table S9. R^2 coefficients for folding rate prediction, using multilinear regression over CT parameters, CO and size. The dataset was divided into 5 subsets. Of these, 4 were used as training set, while the remaining one was used as test set. This process was repeated iteratively so that each subset was used as test set once. The adjusted determination coefficient is higher when we combine CT parameters (parallel and cross) with traditional folding rate predictors such as CO and protein length. Validation sets 1 and 5 were then excluded from the computation of the average presented in Figure 4, since the residuals retrieved from these sets were not normally distributed (Figure S11).

ADJUSTED COEFFICIENT OF DETERMINATION (R^2_{adj})

Validation set	CT parameters	CO	Size	CT + CO	CT + size
1	0.348	-0.237	0.413	0.051	0.425
2	0.307	0.426	0.363	0.444	0.427
3	0.326	0.293	0.115	0.410	0.238
4	0.327	0.423	0.354	0.472	0.397
5	-0.282	0.361	0.067	0.116	-0.229

Table S10. Adjusted R^2 coefficients for folding rate prediction, using multilinear regression over CT parameters, CO and size. The dataset was divided into 5 subsets. Of these, 4 were used as training set, while the remaining one was used as test set. This process was repeated iteratively so that each subset was used as test set once. The adjusted determination coefficient is higher when we combine CT parameters (parallel and cross) with traditional folding rate predictors such as CO and protein length.

RESIDUAL ANALYSIS
Shapiro test, p values

Validation set	CT parameters	CO	Size	CT + CO	CT + size
1	0.997	0.002	0.382	0.006	0.821
2	0.201	0.187	0.094	0.479	0.243
3	0.417	0.308	0.291	0.927	0.356
4	0.275	0.178	0.934	0.386	0.831
5	0.710	0.029	0.103	0.333	0.511

Table S11. Residual analysis reveals residuals from folding rate prediction in the first and fifth validation sets are not normally distributed, when CO is used as independent variable in the linear regression. In order to verify normality, the Shapiro test was applied to the residuals distribution (Predicted (ln kf) - (ln kf)) for each validation set. P values which are lower than 0.05 indicate the distribution does not satisfy the hypothesis of normality.