Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary Information

Topological principles of protein folding

Barbara Scalvini[1], Vahid Sheikhhassani[1], Alireza Mashaghi[1,*]

[1]Medical Systems Biophysics and Bioengineering, Leiden Academic Centre for Drug Research, Leiden University, Leiden, the Netherlands
*Correspondence: a.mashaghi.tabari@lacdr.leidenuniv.nl

1. Correlations between size and CT parameters

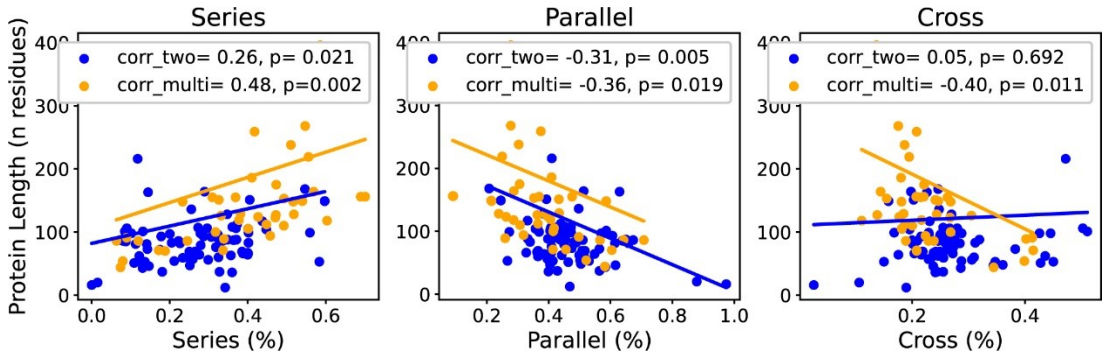


Figure S1. Relationship between CT parameters and protein size. All CT parameters are normalized by the number of contacts in a protein, making it possible to compare proteins with different contacts and sizes. However, a non-trivial relationship between size and CT parameters exists, because of the 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.

2. Multilinear regression: CT parameters, Contact Order and Size.

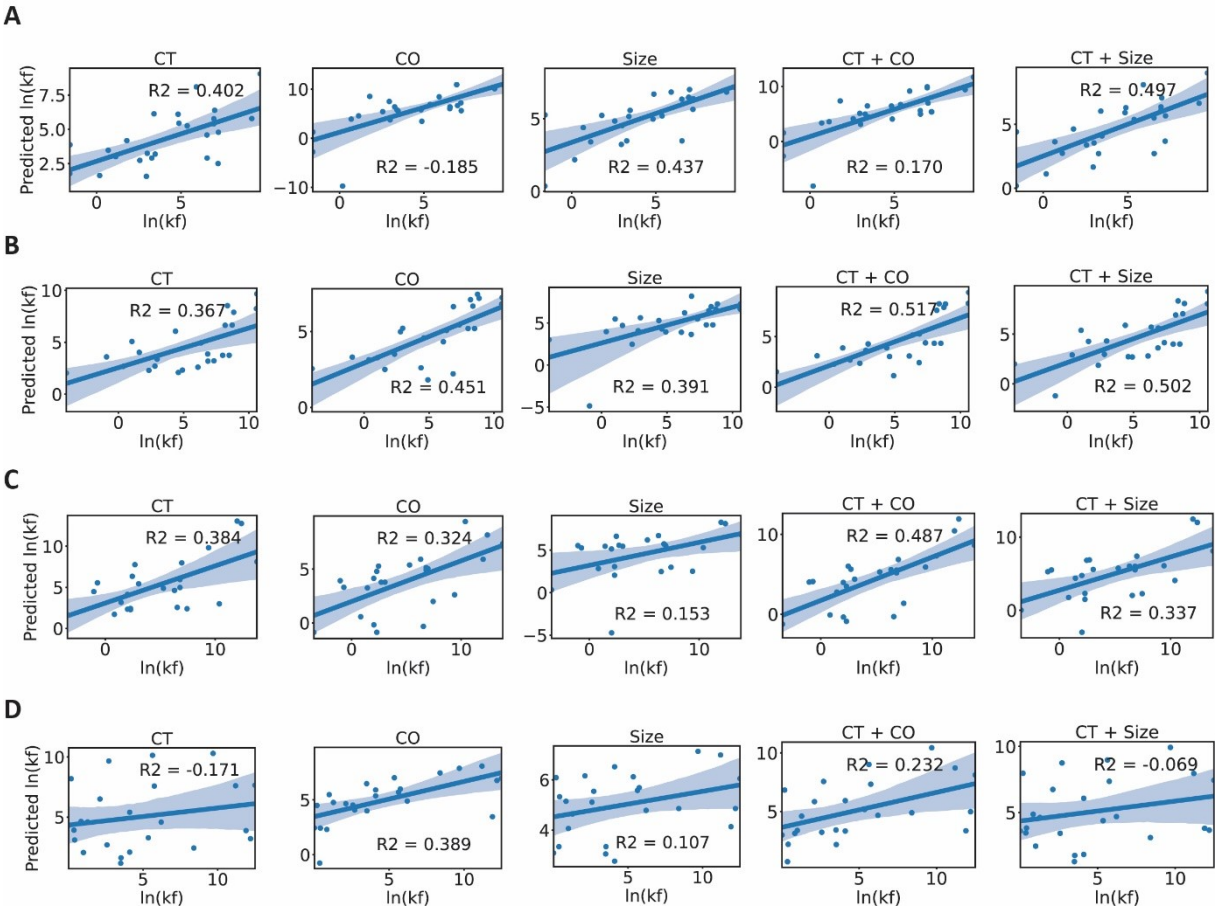


Figure S2. Circuit topology parameters in linear combination with traditional folding rate predictors such as CO and size allow for folding rate prediction with increased statistical significance. A Scatterplots of predicted folding rate (obtained with multilinear regression over CT fractions, CO, protein 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.

3. Correlation between size and folding rate.

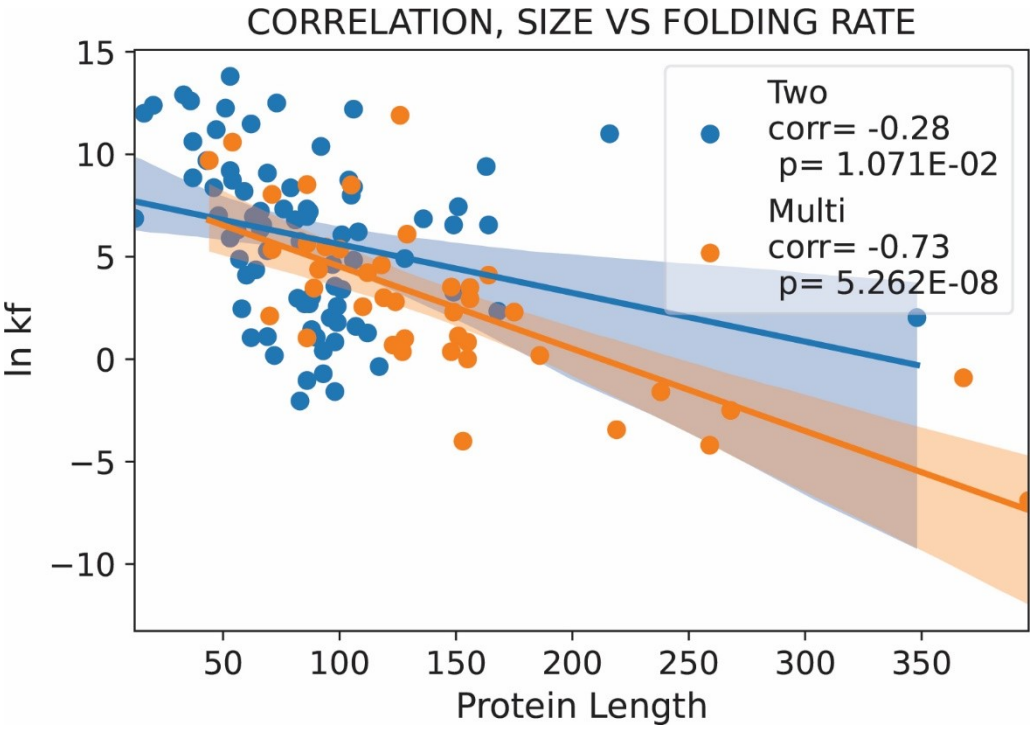


Figure S3. Protein size correlates with folding rate. Scatterplot of protein length versus folding rate (*ln kf*), for two- and multi-state folders.

4. **Correlations between folding rate and CT parameters, for segments-based CT: distance cutoff.**

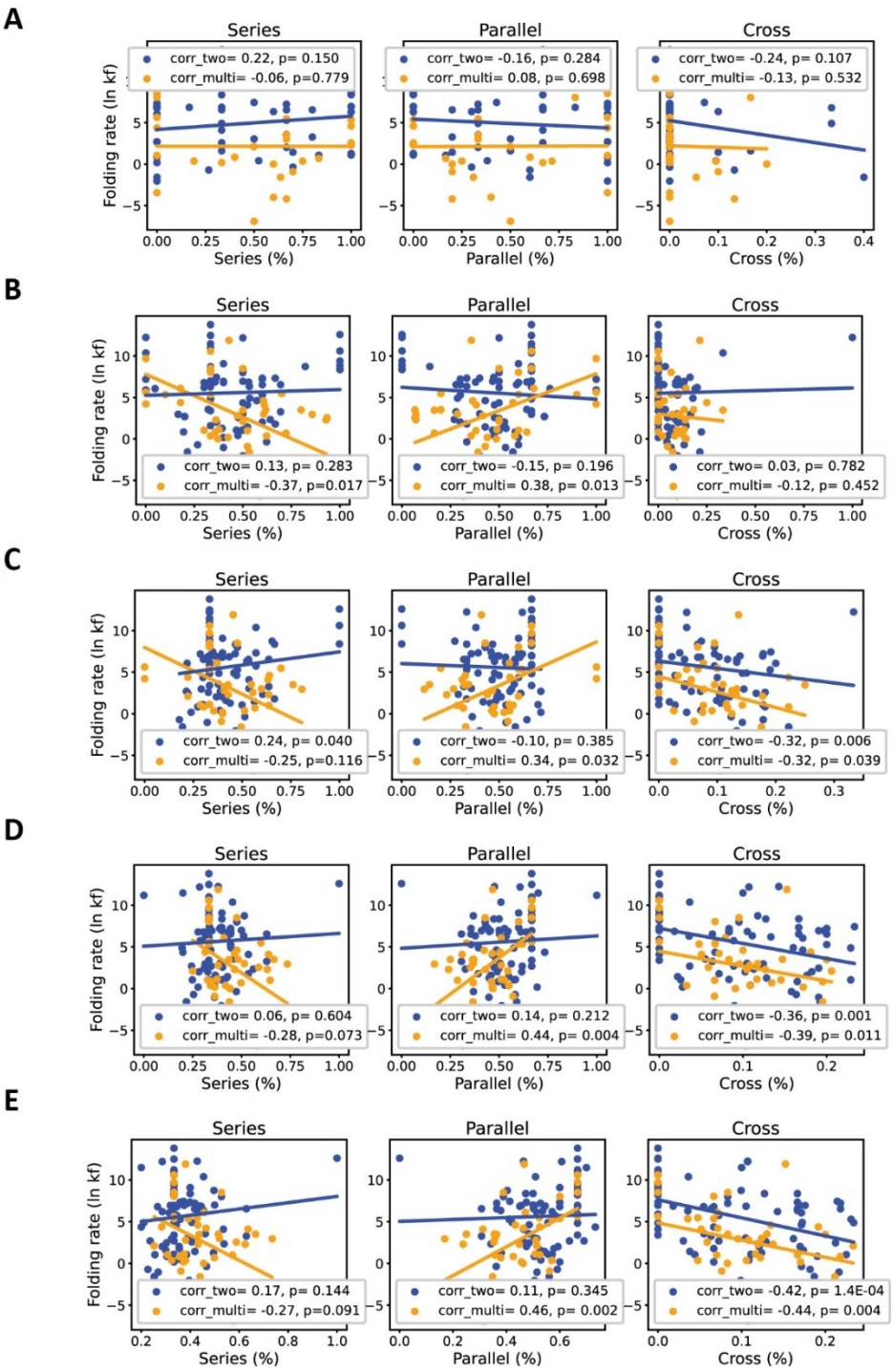


Figure S4. CT parameters for segment-based contacts correlate with folding rate, with distance cutoffs ranging from 4.0 to 6.0 Å. A Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (*ln kf*), for segment-based contacts, calculated with distance cutoff r = 3.5 Å. This 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.

5. Correlations between folding rate and CT parameters, for residue-based CT: distance cutoff

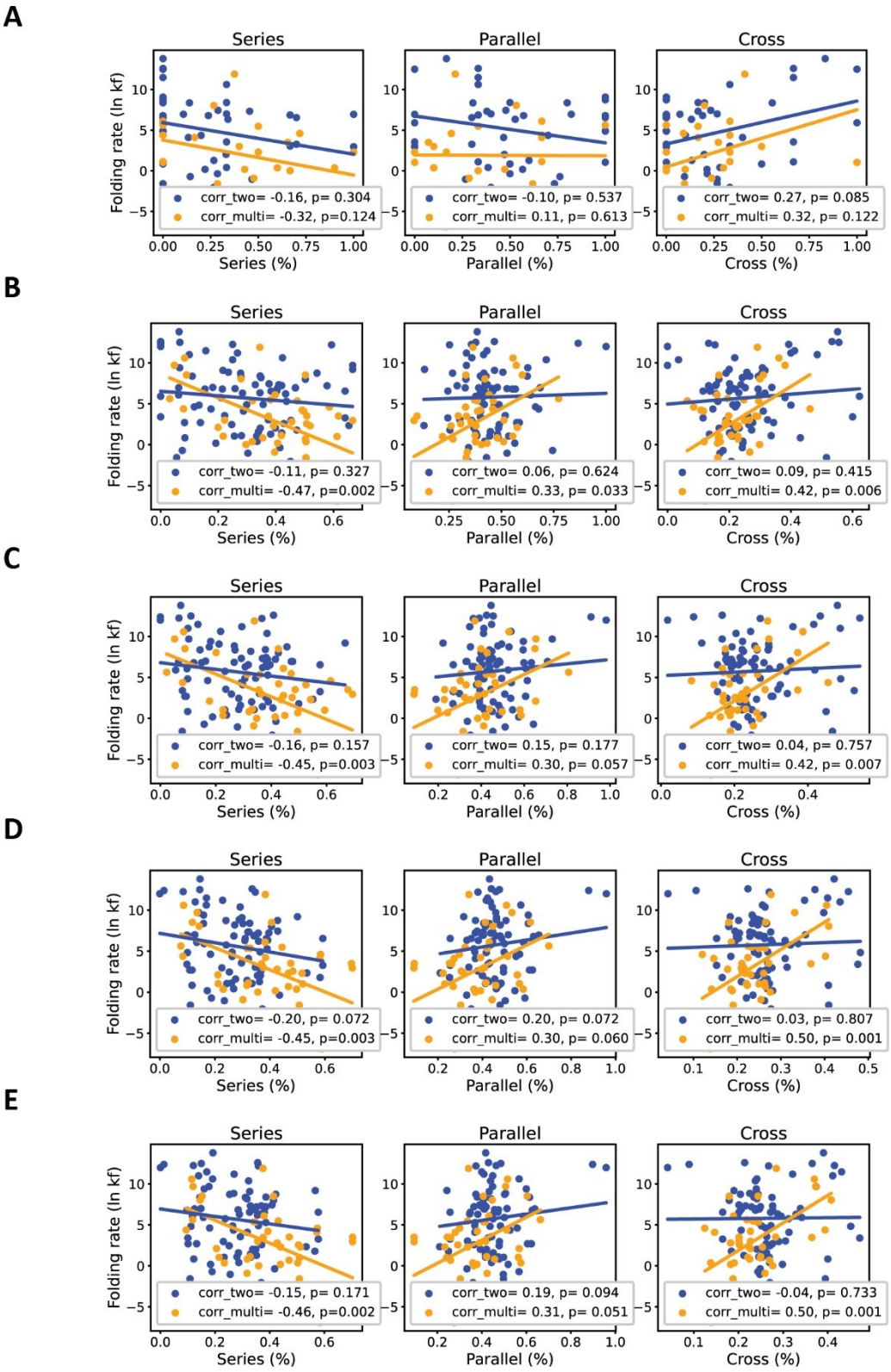
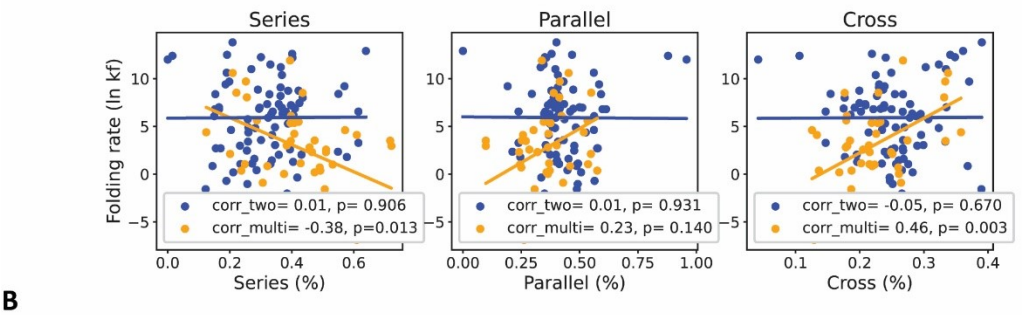


Figure S5. CT parameters for residue-based contacts correlate with folding rate, with distance cutoffs ranging from 4.0 to 6.0 Å. A Scatterplot of topological fractions (Series, Parallel and Cross) versus

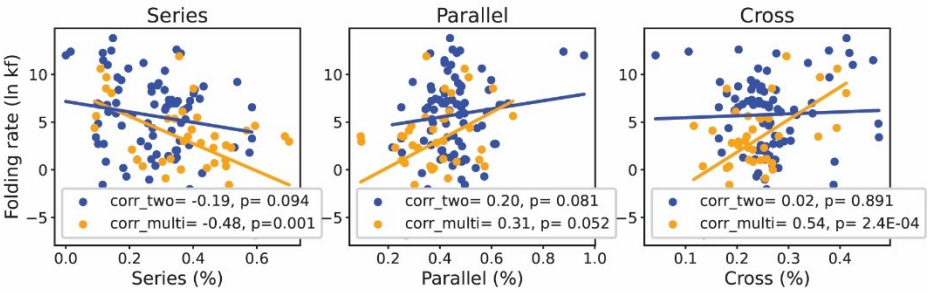
Folding rate (*ln kf*), for residue-based contacts, calculated with distance cutoff r = 3.5 Å. This cutoff represents the lower limit of our analysis, as 55 proteins out of 122 result devoid of contacts with this contact definition. There are no significant correlations between folding rate and CT parameters 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.

6. **Correlations between folding rate and CT parameters, for residue-based CT: number of atoms.**

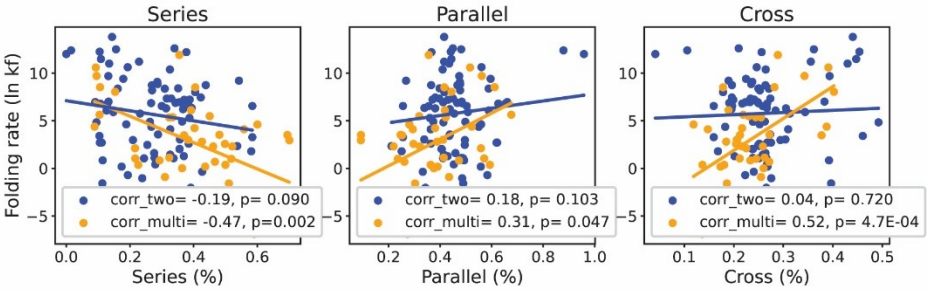
A



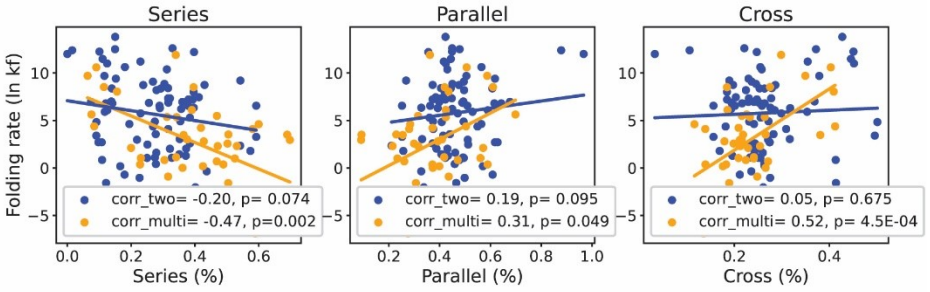
B



C



D



E

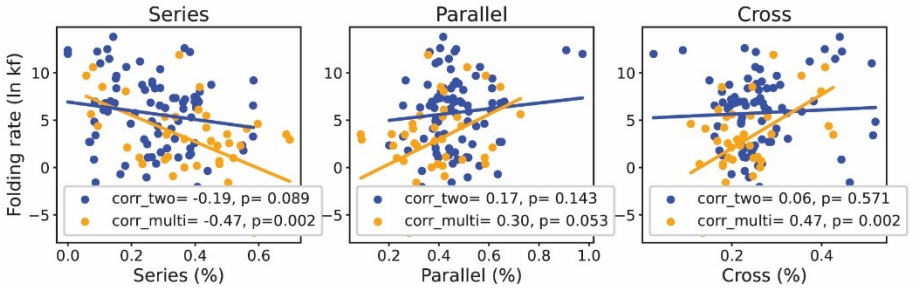


Figure S6. CT parameters for residue-based contacts correlate with folding rate, with r = 5.0 Å and n_a thresholds ranging from 1 to 6. A Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (*ln kf*), for residue-based contacts, calculated with n_a = 1. **B**

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.

7. **Correlations between folding rate and CT parameters (segments), with CO classification: distance cutoff.**

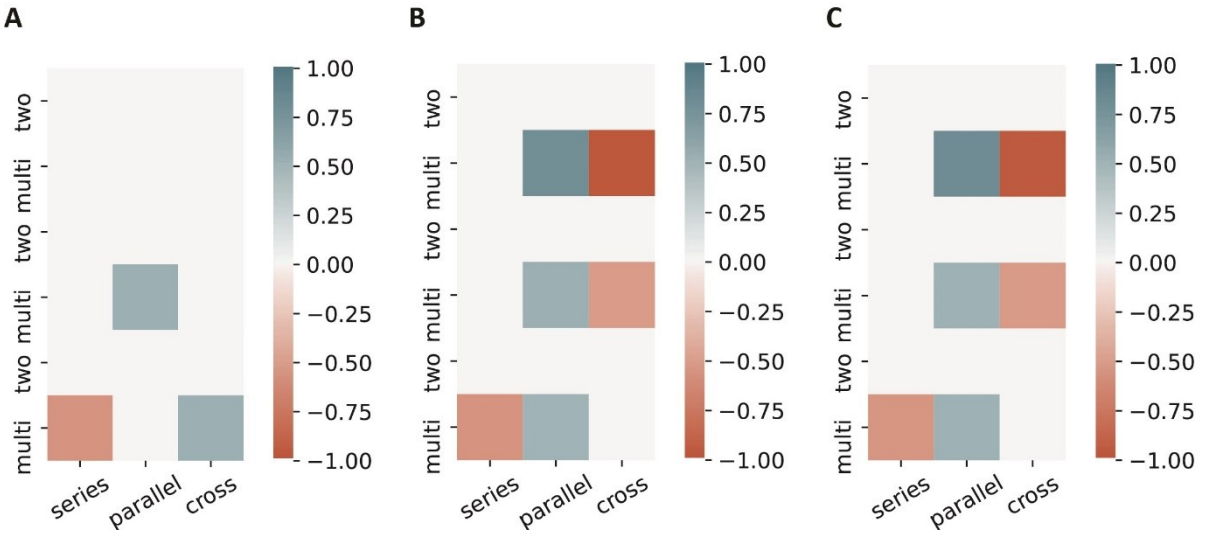


Figure S7. Segment-based CT parameters display differential patterns of correlation with folding rate, which can be highlighted by CO classification. A Folding rate correlation map for segment-based CT, with CO classification, calculated for distance cutoff r=4.0 Å. **B** Folding rate correlation map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

8. **Correlations between folding rate and CT parameters (residues), with CO classification: distance cutoff.**

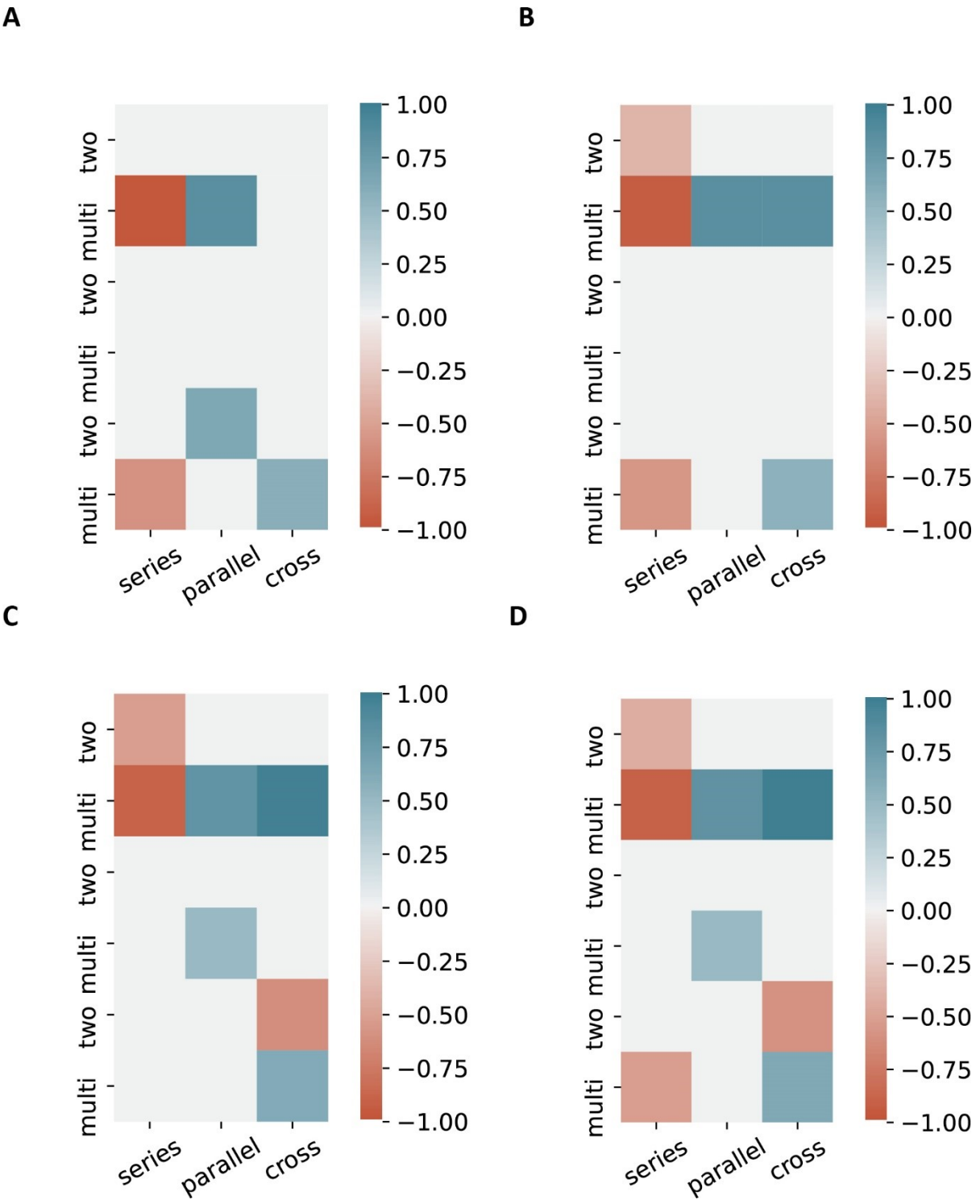


Figure S8. Residue-based CT parameters display differential patterns of correlation with folding rate, which can be highlighted by CO classification. A Folding rate correlation map for residue-based CT, with CO classification, calculated for distance cutoff r=4.0 Å. **B** Folding rate correlation map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

9. Correlation between folding rate and CT parameters, with distance filtering

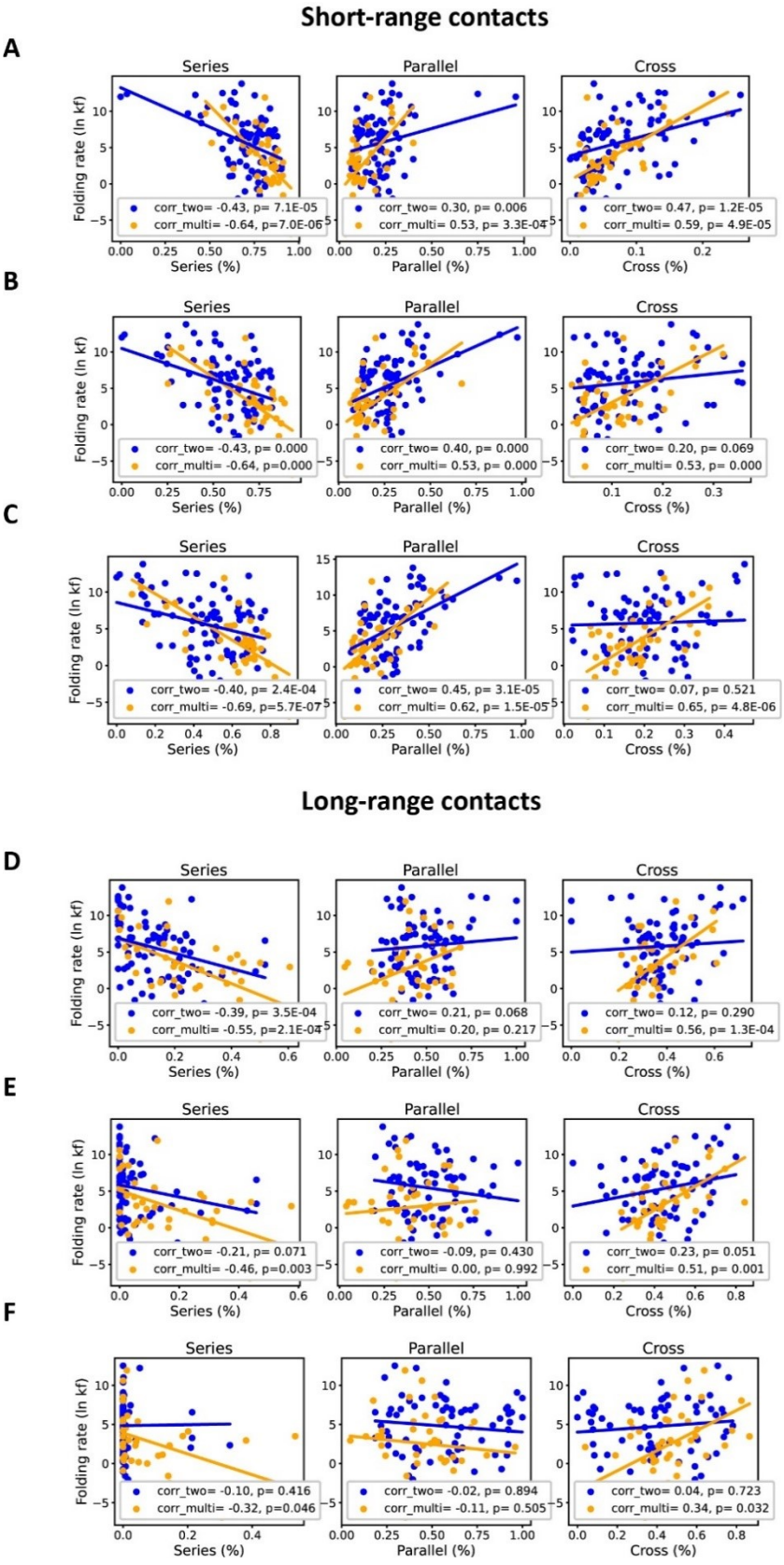


Figure S9. The topology of local and non-local contacts impacts folding rate in different measures, with short-range contacts displaying overall higher correlations. A Scatterplot of topological fractions (Series, Parallel and Cross) versus Folding rate (*ln kf*), for short-range residue-based contacts, 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.

10. Correlations between folding rate and number of circuits

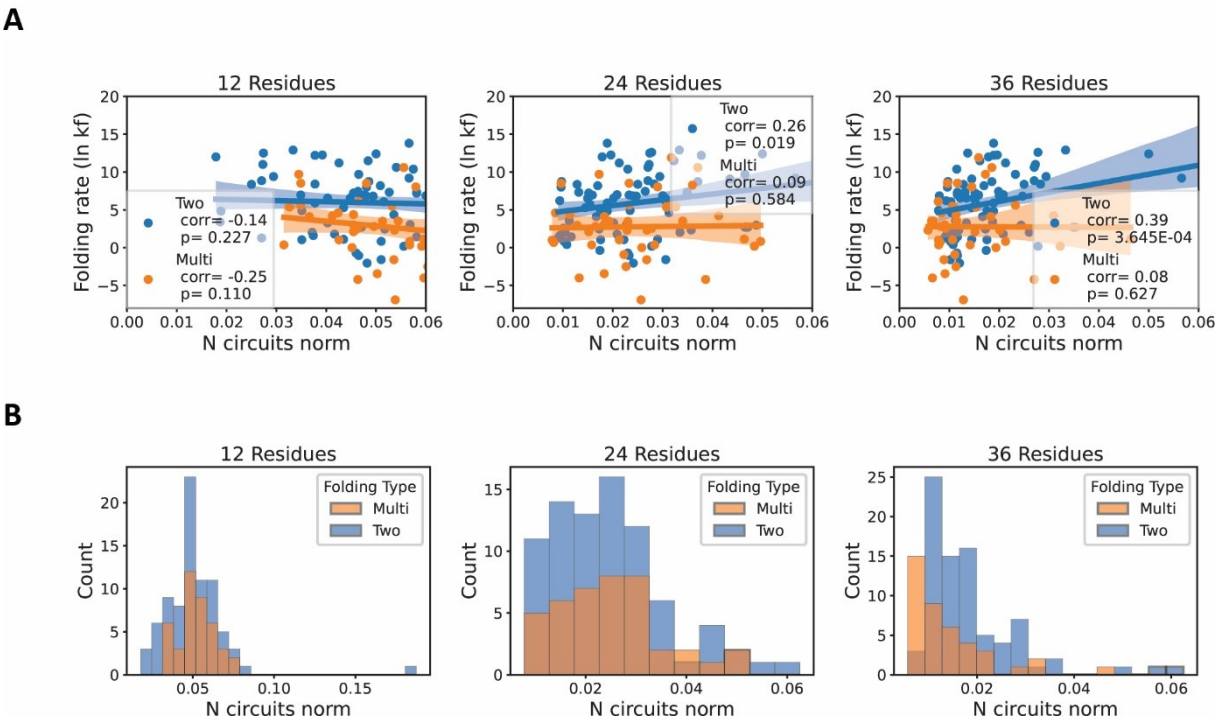


Figure S10. Folding rate correlates positively with the number of topological circuits composing the protein, normalized by size. A Scatterplot of number of circuits normalized by protein length versus folding rate (*ln kf*). Circuits we calculated with a threshold for long-range exclusion equal to 12, 24 and 36 residues. No additional threshold t_l 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)

5.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

5.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

5.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)

6.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

6.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

6.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)

7.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

7.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

7.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)

8.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

8.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

8.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.