The Journal of Chemical Physics

Protein self-entanglement modulates successful folding to the native state: A multi-scale modeling study

Cite as: J. Chem. Phys. **155**, 115101 (2021); https://doi.org/10.1063/5.0063254
Submitted: 13 July 2021 . Accepted: 30 August 2021 . Published Online: 17 September 2021

Lorenzo Federico Signorini, Claudio Perego, and Raffaello Potestio

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Protein self-entanglement modulates successful folding to the native state: A multi-scale modeling study

Cite as: J. Chem. Phys. 155, 115101 (2021); doi: 10.1063/5.0063254
Submitted: 13 July 2021 • Accepted: 30 August 2021 •
Published Online: 17 September 2021

Lorenzo Federico Signorini,[1] Claudio Perego,[2] and Raffaello Potestio[3,a)]

AFFILIATIONS

[1] The George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel and Department of Physics, University of Trento, Trento, Italy
[2] Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Manno, Switzerland and Polymer Theory Department, Max Planck Institute for Polymer Research, Mainz, Germany
[3] Department of Physics, University of Trento, Trento, Italy and INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, Trento, Italy

[a)] **Author to whom correspondence should be addressed:** raffaello.potestio@unitn.it

ABSTRACT

The computer-aided investigation of protein folding has greatly benefited from coarse-grained models, that is, simplified representations at a resolution level lower than atomistic, providing access to qualitative and quantitative details of the folding process that would be hardly attainable, via all-atom descriptions, for medium to long molecules. Nonetheless, the effectiveness of low-resolution models is itself hampered by the presence, in a small but significant number of proteins, of nontrivial topological self-entanglements. Features such as native state knots or slipknots introduce conformational bottlenecks, affecting the probability to fold into the correct conformation; this limitation is particularly severe in the context of coarse-grained models. In this work, we tackle the relationship between folding probability, protein folding pathway, and protein topology in a set of proteins with a nontrivial degree of topological complexity. To avoid or mitigate the risk of incurring in kinetic traps, we make use of the elastic folder model, a coarse-grained model based on angular potentials optimized toward successful folding via a genetic procedure. This light-weight representation allows us to estimate *in silico* folding probabilities, which we find to anti-correlate with a measure of topological complexity as well as to correlate remarkably well with experimental measurements of the folding rate. These results strengthen the hypothesis that the topological complexity of the native state decreases the folding probability and that the force-field optimization mimics the evolutionary process these proteins have undergone to avoid kinetic traps.

I. INTRODUCTION

Since the discovery of the first knotted native structure in 1994,[1] a large number of proteins has been found to entail some degree of topological complexity.[2–5] According to KnotProt,[6] at present over 1600 proteins are known that feature one of the various kinds of possible topological motifs:[2,6,7] these can be knots,[1,3,4,8,9] slipknots,[4,10] complex lassos,[11,12] or links.[2] These proteins need to follow a very specific sequence of steps to achieve the knotted native conformation; otherwise, they risk falling into a misfolded state.[5] It has, however, been noted that even the simplest proteins, usually two-state folders, can present more subtle topological features that play a role in the folding event and affect folding efficiency. Several descriptors of the native conformation of known proteins were found to be correlated with their folding rate and efficiency.[13] Examples of these are the contact order,[14,15] relative effective contact order,[16] native contact number,[17,18] the cliquishness (or clustering coefficient),[19] the long range order,[20] the content of local secondary structures,[21] or the native interaction between the polypeptide termini.[22] These descriptors build on the network of residues that are in contact and

interact in the native conformation. However, given the relevance of self-entanglement in the folding of a growing number of structures, we here focus on the *backbone topology* of the polypeptide chain, considering it as a more or less self-entangled curve in three-dimensional space, regardless of the contact network among non-consecutive residues. We are thus interested in non-local descriptors that can quantify the degree of self-entanglement of the protein backbone. To this purpose, several descriptors based on topological invariants have been proposed;[7] we here focus on the work of 2017 by Baiesi *et al.*, which introduced the concept of "maximum intrachain contact entanglement" $|G'|_c$,[23] a proxy for the topological complexity of the native state of any given protein. $|G'|_c$ measures the Gaussian entanglement between any looped portion of a protein with any other non-overlapping subchain. Testing on a set of proteins, they observed that their degree of backbone self-entanglement anticorrelates with experimental folding rates.[23]

This seminal work motivated us to analyze the folding path of those proteins from the perspective of topological complexity by means of molecular dynamics (MD) simulations. The characterization of a protein's free energy landscape and the search for its global minimum are central topics in computational biophysics research[24,25] and are often carried out using coarse-grained (CG) models, which project higher-dimensional, fine-grained degrees of freedom onto lower-dimensional descriptions, thus reducing the computational overhead and increasing efficiency.[26–28] One popular set of CG models are native structure-based CG models, also called Gō models, the simplest implementation of which represents each amino acid (AA) by a single site centered on the C_α. These systems are *minimally frustrated* on the native contacts, meaning that the attractive interactions between residues that are in contact in the native state are explicitly enforced so that the known reference conformation minimizes the potential.[29–31] These models have remarkable computational efficiency while retaining the ability to drive a molecule to its folded conformation, thus widely employed for studying folding, fluctuations, and interactions of proteins with known native structures;[26,32–34] however, they do not perform as well with models of high topological complexity[7,35–37] because the premature formation of native contacts can push those molecules into kinetic traps, preventing them from folding properly. This is in conflict with the fact that the folding pathway for such systems must be polarized toward the correct native state, as a result of natural selection for optimal folding efficiency. To avoid these kinetic traps and form the correct topology, non-native interactions must play a key role,[38] which is neglected by Gō models.

Building on the latter observations, some of us[38] developed the elastic folder model (EFM), a CG model where each AA is represented by the position of its C_α, the only non-bonded interaction is excluded volume, and the whole complexity of the real system's intra-molecular interactions is projected onto angular potentials between neighbors in sequence, built to have a minimum in the target conformation. In this way, the model has no bias toward native contacts in the potential function, and local rearrangements of the chain are the only drivers of collapse to the target state.

Moreover, based on the assumption that topologically complex proteins have evolved an optimized folding pathway, the EFM undergoes an optimization procedure aimed at maximizing folding success by tuning the force-field parameters to efficiently overcome topological bottlenecks. The heterogeneous force-field thus obtained represents a sort of mean-field approximation of the interplay between native and non-native interactions and can give important insights into the underlying mechanisms of a particular folding event.[39,40]

In this work, we aim at answering the following question: given the observed anticorrelation between experimental folding rates and topological complexity, to what extent is the decrease in folding rate ascribable to native structure and topology? In other words, is the native structure's topology alone enough to justify a lower folding rate or are there other elements? Since the EFM conjugates the conceptual simplicity of a native-centric Gō model (its sole initial information is the native structure) with having features tailored to the task at hand (local structural potentials driving the global folding, with parameters evolving through optimization), this model is ideal to answer such questions. We have thus employed the EFM to correctly fold 12 two-state folder proteins with a complex self-entangled topology,[23] and we have investigated the relationship between topology and folding rates. These proteins are a subset of the two-state folders studied in Ref. 23, covering the whole range of $|G'|_c$ values. The little computational overhead of the EFM allowed us to run a large amount of simulations for each protein and thoroughly explore the conformational space to gather data in order to statistically estimate an *in silico* proxy for the folding rate and test its reliability against the experimental folding rates and its correlation with the topological complexity of the proteins, represented by $|G'|_c$. We observe a non-trivial correlation between topology and folding rates obtained by the model, and we also demonstrate that the measures obtained *in silico* correlate with experimental data. For each protein model, the force-field parameters were tuned following a genetic optimization strategy.[40] To showcase to what extent the optimization step, with the correct strategy, affects the overall process, we run simulations using both optimized and unoptimized force-fields and observe that the simulation success rate increases and, more importantly, correlations improve after optimization.

This paper is organized as follows: in the Sec. II, we describe the EFM, the genetic optimization algorithm, the topological descriptor, and the simulations strategy. In Sec. III, we compare simulation outcomes for all proteins and show correlations of predicted folding rates with topology and experimental folding rates for both optimized and unoptimized force-fields; we observe non-trivial correlations with the optimized force-fields. We then review in detail two case studies, with varying degree of topological complexity ($|G'|_c$): sperm whale myoglobin (PDB code: 1BZP) and the RNA-binding domain of U1A spliceosomal protein (PDB code: 1URN) to gain insights from single systems and subsequently discuss their folding processes and optimization strategies. We conclude by discussing how $|G'|_c$ affects the free energy landscape and the folding process and how a model of minimal complexity, such as the EFM, is able to not only give valuable kinetics insight into folding paths but also preserve the trends of folding rates of experimental data.

II. MATERIALS AND METHODS

A. Protein dataset

We analyzed 12 proteins with a two-state folding transition. This dataset of molecules, listed in Table I, was derived from the

TABLE I. Results of folding simulations for each protein in the dataset. N is the number of beads, S is the total number of folding simulations, F is the folding success rate (number of folded simulations/total), NC is the total number of native contacts, and RMSD′ (in units of σ) is the threshold value at which the protein is considered to be folded. 1URN was optimized via NC-based optimization, while the rest was optimized with an MSD-based optimization.

				Unoptimized force-fields		Optimized force-fields	
PDB	RMSD′	NC	N	S	F	S	F
1APS	0.359	178	98	1 127	0.028	1 059	0.147
1BNZ_a	0.443	61	64	1 037	0.362	1 013	0.584
1BZP	0.779	66	153	1 135	0.085	1 025	0.926
1FKB	0.571	198	107	1 235	0.145	1 052	0.287
1HRC	0.426	113	104	933	0.549	1 175	0.826
1PSF	0.519	96	69	946	0.172	1 019	0.622
1TEN	0.377	168	89	942	0.115	1 068	0.507
1UBQ	0.224	85	76	1 042	0.186	1 017	0.802
1URN	0.310	121	96	1 319	0.006	1 009	0.459
2ABD	0.308	81	86	936	0.949	1 079	0.948
2CI2	0.201	16	64	947	0.486	1 270	0.943
2VIK	0.776	63	126	1 133	0.417	1 028	0.742
Total				27 094		12 814	

work of Baiesi and co-workers.[23] From this work, we also took the reported values of the logarithm of the experimental folding rate $F_{\exp}$ (see, e.g., Ref. 41). These values were, in turn, obtained from previous literature, in particular, Refs. 15, 16, and 42 and references therein; these rates will be employed as the benchmark against which we will compare the performance of our model. The notion of nontrivial topology employed throughout this work, as well as the observable employed to quantify the topological entanglement, follows Ref. 23. For simplicity, herein we refer to each protein using their PDB codes.

B. Elastic folder model

The elastic folder model (EFM)[38] is a native-structure-based CG representation. In the EFM, the protein is modeled as a chain of beads, each representative of an AA and centered on the C_α atom. The potential energy function associated with the model has the following general form:

$$\mathcal{V} = U_{\mathrm{WCA}} + U_{\mathrm{FENE}} + U_{\mathrm{bending}} + U_{\mathrm{torsion}}. \tag{1}$$

U_{WCA} is the Weeks–Chandler–Andersen (WCA) repulsive potential, the only non-bonded interaction of the model,

$$U_{\mathrm{WCA}} = \frac{1}{2}\sum_{(i,j),j\neq i}^{N} V(d_{i,j}),$$
$$V(r) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} + \frac{1}{4}\right] & \text{for } r \le 2^{\frac{1}{6}}\sigma, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where σ is the diameter of the beads, taken as length unit, and equal to 3.8 Å and $d_{i,j} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the centers of the i and j beads. ϵ is the energy scale parameter set as the energy unit for the rest of the present work; assuming a temperature of ~ 300 K, the numerical value of this energy scale is $\epsilon \sim 25$ kJ/mol. $r = 2^{\frac{1}{6}}\sigma$ is the distance at which $U_{\mathrm{WCA}} = 0$.

The remaining components account for bonded interactions. Peptide bonds are modeled via the finite extensible nonlinear elastic (FENE) potential,[43] U_{FENE}, given by

$$U_{\mathrm{FENE}} = \sum_{i=0}^{N-2} \frac{k_{\mathrm{FENE}}}{2}\left(\frac{R_0}{\sigma}\right)^2 \ln\left[1 - \left(\frac{d_{i,i+1}}{R_0}\right)^2\right], \tag{3}$$

where R_0 is the maximum bond length and k_{FENE} is the FENE interaction strength. U_{bending} and U_{torsion} are employed as a basis set of functions on which the whole complexity of the intra-molecular interactions of the chain is projected. U_{bending} is given by

$$U_{\mathrm{bending}} = \sum_{i=1}^{N-2} k_i^{\mathrm{bend}}(\theta_i - \theta_i^0)^2, \tag{4}$$

where θ_i^0 is the bending angle centered on the ith bead in the reference state and k_i^{bend} is the bending stiffness. U_{torsion} is

$$U_{\mathrm{torsion}} = \sum_{i=1}^{N-3} U_i^{\mathrm{tor}},$$
$$U_i^{\mathrm{tor}} = k_i^{tor}\left[\cos(\phi_i - \phi_i^0) + \frac{1}{3}\cos(3(\phi_i - \phi_i^0))\right], \tag{5}$$

where ϕ_i^0 is the torsion angle of the ith bead in the reference state and k_i^{tor} is the torsion stiffness. The reference angles in Eqs. (4) and (5), setting the minimum of the potential, are chosen from a target conformation, i.e., the PDB crystal structure. We note that this model has no bias toward the formation of native contacts, and the collapse

toward the target conformation is driven by the angular potentials only.

The dynamics of the beads is governed by the overdamped Langevin equations of motion,

$$\frac{\partial \mathcal{V}(t)}{\partial \mathbf{r}_i} + m\gamma \mathbf{v}_i(t) + \mathbf{R}_i(t) = 0, \tag{6}$$

where $\mathcal{V}$ is the potential energy function of Eq. (1); $m, \mathbf{v}_i$, and $\mathbf{r}_i$ are the mass, velocity, and coordinate of the ith bead; γ is the friction coefficient; and R_i is a random force acting on i. Equation (6) is integrated with a symplectic, first order algorithm.[38]

The EFM is based on a criterion of optimality: we assume that topologically complex proteins have evolved to fold along an efficient and reproducible pathway in the free energy landscape. To implement this optimality, the stiffnesses of the angular potentials are tuned by means of an optimization process aimed at maximizing the successful folding rate. In this work, optimization was performed via a genetic algorithm (see Sec. II C). These guidelines yield a model of minimal complexity that can provide useful information about the most efficient folding pathways followed by the protein.

Parameter values for the model employed are reported in Table II, where we report the bending and torsion coefficients chosen for the unoptimized, uniform models, i.e., $k_i^{\text{bend}} = k_i^{\text{tor}} = 50$. Because of the CG representation, the EFM protein models cannot be quantitatively compared to the physical features of realistic proteins; thus, the predictivity of the model is limited only to the qualitative aspect of topology formation and to compare the scaling of characteristic times.

C. Genetic optimization of force-fields

To satisfy the principle of optimality of the folding pathway, the EFM angular force parameters k_i^{bend} and k_i^{tor} are tuned to maximize the success rate of folding. In this work, the strategy of Ref. 40 was followed for the optimization, and a similar approach was also recently employed in Ref. 44. A set of parallel stochastic searches for mutated force-fields [single force-field optimization (SFFO)] is performed. The resulting improved force-fields are then ranked according to a selected criterion and "crossed over" in a genetic step [multiple force-field optimization (MFFO)].

TABLE II. System parameters for the uniform (unoptimized) model: m is the mass of the beads, ϵ is the energy unit, τ_{md} is the time unit, Δt is the time step, R_0 is the FENE bond maximum length, τ_{frict} is the friction coefficient, and T is the temperature ($k_B = 1$).

Parameter	Value
m	1
σ	1
ϵ	1
τ_{md}	$\sigma\sqrt{m/\epsilon} = 1$
Δt	$5 \cdot 10^{-4}\tau_{MD}$
R_0	1.5σ
k_{FENE}	30ϵ
τ_{frict}	$1\tau_{MD}$
k_i^{bend}	50ϵ
k_i^{tor}	50ϵ
T	0.1ϵ

A single force-field K is defined as

$$\begin{aligned} K &= \{k_1^{\text{bend}}, \ldots, k_{N-2}^{\text{bend}}, k_1^{\text{tor}}, \ldots, k_{n-3}^{\text{tor}}\} \\ &= \{k_1^{\text{ang}}, \ldots, k_{2N-5}^{\text{ang}}\}, \end{aligned} \tag{7}$$

where k_i^{ang} is any angular coefficient. To reduce the number of parameters, instead of assigning to each residue independent bending and torsion coefficients, pairs of neighboring residues were enforced to have identical values for torsion and bending parameters, respectively, where possible.

1. Single force-field optimization (SFFO)

In this work, initial values for each pair of coefficients were sampled from a uniform distribution between 15ϵ and 85ϵ so that the mean value is 50ϵ, i.e., the coefficients in the uniform force-field. SFFO starts from the initial force-field K and generates a mutated K',

$$K' = \{k_1^{\text{ang}}, \ldots, k_j^{\text{ang}} + \delta k, \ldots, k_{2N-5}^{\text{ang}}\}, \tag{8}$$

in which the jth coefficient is modified by adding δk. j is randomly chosen among the $2N - 5$ coefficients, while δk is generated from a normal distribution with the standard deviation equal to 2.5. Subsequently, the mutation is accepted or rejected according to a Metropolis-like criterion: K' is tested by performing a set of $n = 16$ parallel folding simulations (the *test runs*), starting from a randomly generated stretched configuration. After $4 \cdot 10^6$ steps, the average Mean Square Displacement (MSD) from the target configuration $\mathbf{R}^0$ is measured. The MSD $\mathscr{F}$ is defined as

$$\mathscr{F}(t; K') = \frac{1}{N\sigma^2}|\mathbf{R}(t) - \mathbf{R}^0|^2, \tag{9}$$

where $\mathbf{R}(t)$ is the coordinates vector of the chain at time t. Then, the average over n test runs is

$$\langle \mathscr{F}(t_{\min}; K') \rangle = \frac{1}{n}\sum_{i=1}^{n} \mathscr{F}^i(t_{\min}; K'), \tag{10}$$

where $t = t_{\min}$ is the time step at which $\mathscr{F}$ is minimum during the ith test run. The probability of accepting the mutation K' is then

$$P(K'|K) = \min\{1, \exp[\langle \mathscr{F}(t_{\min}; K) \rangle - \langle \mathscr{F}(t_{\min}; K') \rangle]\}. \tag{11}$$

The two steps of Eqs. (8)–(11) were repeated for 25 iterations to minimize $\langle \mathscr{F} \rangle$, enhancing the average folding success rate of the trajectories.

2. Multiple force-field optimization (MFFO)

Several SFFOs are run in parallel in the Multiple Force-Field optimization (MFFO). The MFFO is organized in *cycles*. In every *cycle*, an initial population of $N_K = 16$ force-fields $\{K_j\}_{j=1}^{N_K=16}$ is generated; each force-field undergoes $m = 25$ SFFO *iterations* independently from each other. After these iterations, the resulting N_K force-fields are ranked, and a *crossover* step is performed to generate new force-fields, which will be submitted to the next *cycle*.

In the *ranking* step, the N_K mutated force-fields are ranked according to their folding probability. Since this probability is unknown *a priori* we estimate it based on the results of the test runs performed along the optimization. To this end, we compute Π_f, i.e., the exponential moving average of π_f^i, namely, the estimator of the folding probability at the ith iteration step of the SFFO. Π_f can be written as

$$\Pi_f \equiv \Pi_f^{m=25} = \alpha\pi_f^m + (1-\alpha)\Pi_f^{(m-1)}. \tag{12}$$

$0 < \alpha < 1$ is a smoothing factor, and π_f is defined as

$$\pi_f(\xi_0, t_{\min}) = \frac{1}{n}\sum_{i=1}^{n=16} \mathscr{L}[\xi_0 - \xi(t_{\min}, K)]. \tag{13}$$

$\mathscr{L}$ is a Fermi function that switches from 0 to 1 when its argument becomes positive,

$$\mathscr{L}(z) = \left[1 + \exp\left(-\frac{z}{w}\right)\right]^{-1}, \tag{14}$$

where w is a parameter controlling the scale of the switching; ξ is any set of reaction coordinates that can resolve the folding events, and ξ_0 is the vector of threshold values at which the protein is considered to be folded. In this work, either the MSD or the fraction of native contacts, NC, were used as reaction coordinates (see below).

In the *crossover* step, the $N_{win} = 6$ top-ranked force-fields (winners) are kept for the next cycle and are used in combination with new randomly generated force-fields to build a new population $\{K_j'\}_{j=1}^{N_K=16}$ as follows:

$$\{K_j'\}_{j=1}^{N_K} = \left(\{W_i\}_{i=1}^{N_{win}}, \{H_k\}_{k=1}^{N_K - N_{win}}\right), \tag{15}$$

where W indicates the winners and H indicates a set of $N_K - N_{win}$ newly generated hybrid force-fields. The latter ones are obtained by (i) generating six new random force-fields (called "low-fit") with the same uniform distribution of the initial ones, then (ii) splitting both the six newly generated random force-fields and the winner force-fields into six segments each, and finally (iii) randomly selecting among all the force-fields generated by combination of the subsets of winner and low-fit force-fields. This "crossover" operation, typical of genetic algorithms,[45] yields the new set of force-fields $\{K_j'\}_{j=1}^{N_K}$, which will be the initial conditions for the next MFFO *cycle*.

For each of the proteins, the following optimization procedure was employed: 20 MFFO cycles of 16 parallel SFFOs. Each SFFO was run for $m = 25$ iterations, and each iteration had 16 test runs. All proteins were first optimized with an MSD-based MFFO ranking [ξ = MSD in Eq. (13)]. Subsequently, proteins 1APS, 1URN, and 1FKB were also optimized with a NC-based MFFO ranking [ξ = NC in Eq. (13)]. In total, 1 920 000 simulations of $4 \cdot 10^6$ time steps each were run for optimizing all proteins.

D. Simulations scheme

Two kinds of simulations were carried out: (i) *equilibration* simulations, initializing the model from the native conformation (the PDB structure), and (ii) *folding* simulations, starting from a completely unfolded conformation. All simulations lasted $7 \cdot 10^6$ steps.

One equilibration per protein was first performed with the unoptimized force-field to monitor the RMSD from the initial PDB structure under equilibrium conditions. The highest RMSD in this equilibration trajectory (RMSD_p', where the subscript p indicates different proteins) was used as the threshold value below which the protein was considered to be folded (Table I). Subsequently, about 1000 folding simulations per protein (ranging from 933 simulations for 1HRC, to 1319 for 1URN) were carried out.

The force-fields were optimized with the MFFO algorithm, and new simulations were performed using the resulting force-fields. With these optimized force-fields, for each of the 12 proteins, about 1000 folding simulations per protein (ranging from 1009 for 1URN to 1270 for 2CI2) were carried out. For 1URN, 1FKB, and 1APS, we tested both the MSD-based and NC-based ranking criteria in MFFO optimization. RMSD_p' values and total number of simulations are shown in Table I. The conformations sampled by all the equilibration or folding trajectories are then gathered in ensembles that are employed for the calculations of the properties of each protein model, such as folding rates (see Sec. II F) and free energy surfaces (see. Sec. II of the supplementary material for further details).

E. Topological descriptor for self-linked proteins

In order to describe the topology of a protein, one can look at the C_α backbone and think of it as a single piece of string that folds itself in the three-dimensional space.[8] Herein, we consider proteins with a self-entangled topology, i.e., where a part of the chain forms a topological link with another part of the chain. The topological complexity of these intrachain links is measured via *maximum intrachain contact entanglement* $|G'|_c$ (defined below), a descriptor based on Gauss's linking number,[23,46,47] namely, a double integral computed on two closed curves γ_1 and γ_2,

$$G = \frac{1}{4\pi}\oint_{\gamma_1}\oint_{\gamma_2} \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3}(d\mathbf{r}_1 \times d\mathbf{r}_2), \tag{16}$$

where $\mathbf{r}_1 \in \gamma_1$ and $\mathbf{r}_2 \in \gamma_2$. $G = l$, where l is the number of times the loop γ_1 threads through γ_2.[46,47] This value is reciprocal (it remains the same if the curves are interchanged) and is a topological invariant, meaning that it does not depend on the shape of the two curves.[46–48] If one or both the curves γ_1 and γ_2 are open, then G is no longer an integer but remains a proxy for their level of entanglement.[48] G can be thus computed along the backbone of one or more proteins, integrating over the curves traced in the space by its subchains, obtaining a measure of the topological complexity of the polypeptide conformation. Knowing this, $|G'|_c$ is calculated as follows: a pair of non-overlapping subchains γ_i and γ_j are selected from the backbone of a protein. γ_i is essentially a closed loop, meaning that its first and last residues ($\mathbf{r}_{i1}$ and $\mathbf{r}_{i2}$) form a *contact*, i.e., their native positions are closer than 9 Å. Instead, γ_j is not constrained this way and can therefore be an open loop. Figure 1 provides an example of a protein chain subdivided in a closed γ_i and an open γ_j loop.

The integral can be calculated via discretizing the chain over the number of residues ($i = 1, \ldots, N$) and by defining the average positions $R_i = \frac{1}{2}(r_i + r_{i+1})$ and the bond vectors $dR_i = r_{i+1} - r_i$; hence,

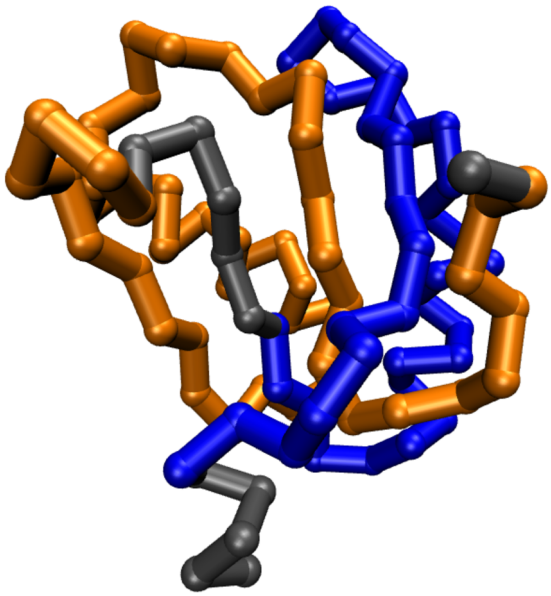

FIG. 1. Example of protein backbone (protein 1URN), highlighting the closed subchain γ_i (residues 55–90, blue) and the open subchain γ_j (residues 1–49, orange), which yield the maximum intrachain contact entanglement.

$$G'_{ij} = \frac{1}{4\pi}\sum_{i=i_1}^{i_2-1}\sum_{j=j_1}^{j_2-1}\frac{\mathbf{R}_i-\mathbf{R}_j}{|\mathbf{R}_i-\mathbf{R}_j|^3}(d\mathbf{R}_i\times d\mathbf{R}_j), \tag{17}$$

where the prime in G'_{ij} is to point out the fact that the calculation is on an open chain. Then, calculating G'_{ij} for every possible combination of non-overlapping subchain couples $\{\gamma_i, \gamma_j\}$ and taking the maximum of the absolute value (the sign depends only on the relative directions of the two subchains), one obtains the *maximum intrachain contact entanglement*:[23]

$$|G'|_c = \max_{[i_1,i_2],[j_1,j_2]}|G'_{ij}|. \tag{18}$$

$|G'|_c$ was calculated for 12 small two-state folder proteins (see Table III), during one equilibration simulation (see Sec. II D), sampled every 5000 steps in the trajectory, thus obtaining a distribution of values of $|G'|_c$ for each protein.

F. Folding rate estimators

In order to assess the folding efficiency of the proteins under examination, we computed the folding frequencies F_p for the proteins as $F_p = S_p^f/S_p$, where S_p is the total number of simulations and S_p^f is the number of simulations for protein p whose RMSD has fallen below RMSD'_p at any point during the simulation. These were calculated both with the unoptimized and the optimized force-fields for each protein and correlated with experimental quantities from Ref. 23.

Additionally, we computed a quantity more akin to a folding rate, $\widetilde{R}_p$, based on the median folding time for every protein p, as

$$\widetilde{R}_p = \frac{1}{median\left[\sum_{t=0}^{T}\theta(t_p - t'_p)\right]_p\Delta t}, \tag{19}$$

where t is the current time step, t'_p is the time step at which $\text{RMSD}(t'_p) = \text{RMSD}'_p$, RMSD'_p is the benchmark value of RMSD, $T = 7\cdot 10^6$ is the maximum time step, Δt is the length of the time step, and θ is the Heaviside function. The rates are expressed in τ_{MD}^{-1}, which is an arbitrary time unit derived from the model constants (Table II).

III. RESULTS

A. Optimization results

A first batch of 20 cycles of MFFO optimizations was run for all the proteins of Table I, where the criterion for successful folding was that MSD < 0.9. Subsequently, since the optimized model of protein

TABLE III. $|G'|_c$ comparison table: $|G'|_c$ for every protein, as calculated by Baiesi *et al.*,[23] compared to average $\langle|G'|_c\rangle$ over one equilibrium simulation of $7\cdot 10^6$ steps with the elastic folder model and their relative standard deviation, max and min. i_1, i_2, j_1, j_2 are residue indexes that identify subchains $\gamma_i\gamma_j$.

PDB code	Number of residues	i_1, i_2	j_1, j_2	$\lvert G'\rvert_c$ as in Ref. 23	$\langle\lvert G'\rvert_c\rangle$ with EFM	$std(\lvert G'\rvert_c)$ with EFM	Max $\lvert G'\rvert_c$ with EFM	Min $\lvert G'\rvert_c$ with EFM
1APS	98	41, 97	1, 40	1.62	1.487	0.054	1.655	1.159
1BNZ_a	64	19, 6	37, 55	0.27	0.256	0.024	0.362	0.172
1BZP	153	95, 149	35, 94	0.47	0.429	0.034	0.546	0.335
1FKB	107	45, 104	4, 31	0.96	0.839	0.041	0.988	0.666
1HRC	104	1, 89	90, 104	0.56	0.429	0.059	0.598	0.181
1PSF	69	21, 66	1, 14	0.47	0.392	0.059	0.578	0.203
1TEN	89	38, 86	3, 37	0.67	0.593	0.030	0.715	0.495
1UBQ	76	12, 66	1, 11	0.47	0.410	0.031	0.546	0.293
1URN	96	55, 90	1, 49	1.15	1.038	0.032	1.127	0.835
2ABD	86	22, 53	54, 84	0.60	0.502	0.040	0.647	0.379
2CI2	64	3, 44	45, 55	0.68	0.588	0.032	0.680	0.468
2VIK	126	66, 119	14, 65	0.86	0.550	0.059	0.801	0.416

1URN failed to fold consistently, a new optimization run using the number of native contacts (NCs) as the proxy for folding success was performed, increasing the folding success rate of the model, *vide infra*; finally, the NC-optimized force-field was retained for 1URN.

The values of the proxy success rate Π_i (calculated over the test runs of the best performing force-field per each MFFO iteration) are reported as a function of the optimization cycle in the supplementary material (see Fig. S1), showcasing a general increase in folding success after every cycle. In Fig. 2, the values of the folding rate calculated over ~ 1000 runs before and after the optimization for each of the 12 proteins under examination are reported. The optimization increased the folding success rate in practically all cases.

Around 1000 folding runs of 7×10^6 time steps were run for each of the 12 proteins using the best overall force-field from the optimization. Moreover, 64 equilibration simulations were performed per each protein model to collect statistics about the equilibrium state.

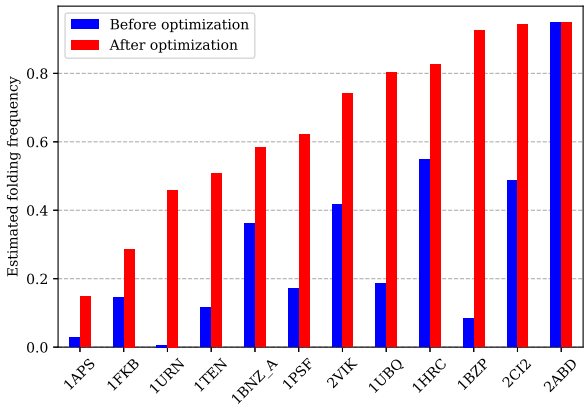

FIG. 2. Estimated folding frequencies F before and after 20 cycles of optimization.

B. Correlations

As it is commonly the case in the context of CG models, quantitative measures of time are of difficult interpretation due to the characteristic "telescoping" of time scales;[49] we thus resorted to a definition of *in silico* folding rates as the frequency of successful folding events. These frequencies were then correlated with topological descriptors as well as experimentally measured folding rates. Figure 3 shows how the estimated folding frequency (for

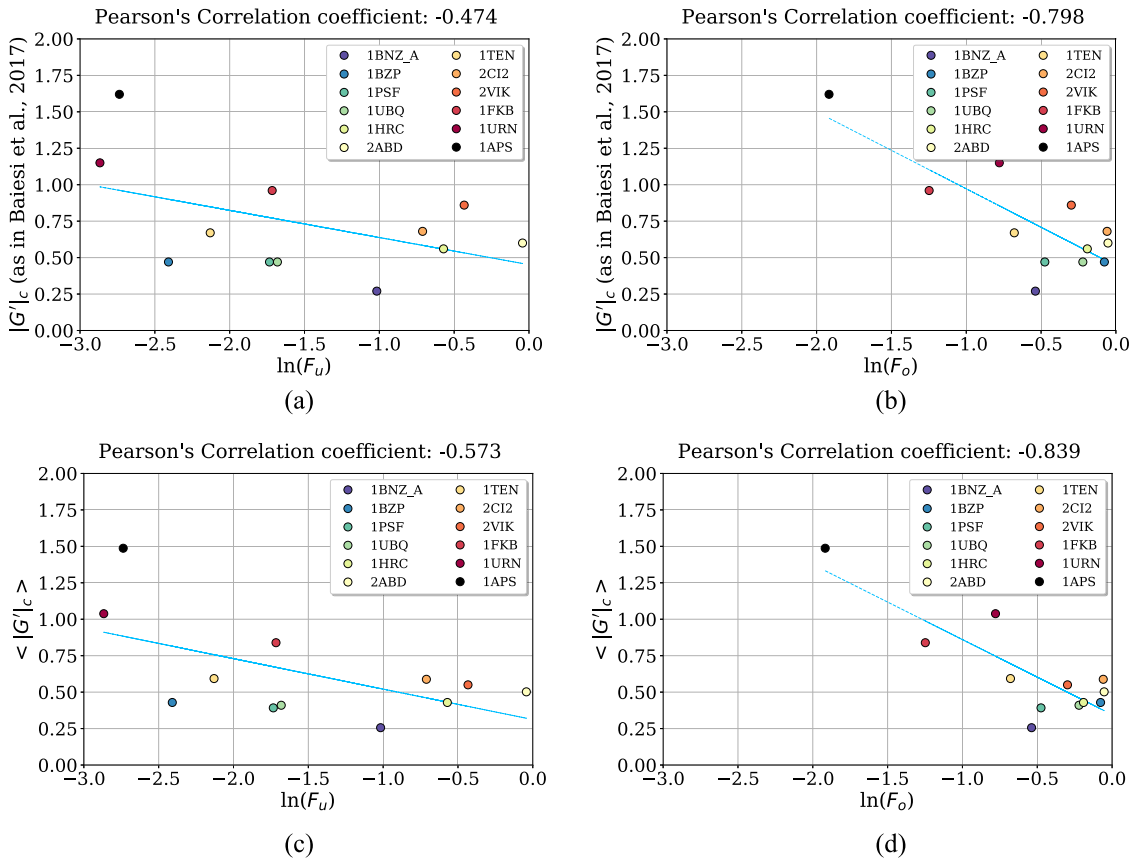

FIG. 3. Correlations of estimated folding frequency for *unoptimized* (left column) and *optimized* (right column) force-fields vs $|G'|_c$ (top row) and $\langle|G'|_c\rangle$ (bottom row). F is the estimated folding frequency. Force-field optimization improves all correlations. Proteins are colored sequentially according to their ordering given by $\langle|G'|_c\rangle$ going from purple (lowest value) to black (highest value).

non-optimized and optimized force-fields) correlates with (i) the Gauss linking number $|G'|_c$ computed on the native structure [panels (a) and (b)] and the Gauss linking number $\langle|G'|_c\rangle$ averaged over an equilibrium simulation starting in the native state [panels (c) and (d)]. Even before optimization, all proteins reached their native structure. After the optimization step, the folding frequency was higher in all cases but one; the most significant difference was found with protein 1BZP, which increased its folding frequency by 0.877 (0.085–0.962), and the least difference was found with 1FKB (0.142, from 0.145 to 0.287). The only force-field that did not improve was that of 2ABD, which featured a rate as high as 0.949 before the optimization and changed to 0.948 after the optimization. Our predicted folding frequencies strongly anticorrelate with topology [see Figs. 3(b)–3(d)].

A few comments are in order regarding the (anti)correlation between $\langle|G'|_c\rangle$ and the folding frequency. As a first thing, we observe that the correlation coefficient increases (in absolute value) from −0.57 to −0.84 when going from the unoptimized to the optimized force-fields, that is, an increment of ∼ 47%. Second, we note that in the work by Baiesi and co-workers,[23] the correlation found between $|G'|_c$ and the experimental folding rate is −0.64; a higher value, namely, −0.91, is achieved only when the experimental folding rate is correlated with a weighted sum of $|G'|_c$ and the relative contact order (RCO), a quantifier of the local structural packing of the protein in the native state. Furthermore, in the aforementioned linear combination, the RCO accounts for the 67% of the parameter. We thus conclude that in the case of the elastic folder model with optimized force-fields, the $\langle|G'|_c\rangle$ parameter *alone* largely accounts for the impediments that the topological self-entanglement introduces in the folding process of the proteins under examination. The residual lack of correlation suggests that further optimizations are possible: the discrepancy between the correlation coefficient obtained with the optimized EFM force fields and the larger one measured in Ref. 23 making use of a mixed topological/structural observable hints at possible modifications of the optimization procedure and/or of the interactions themselves that, when accounting for structural features of the native state, might boost the folding accuracy of the model.

Regarding the comparison with the experimental folding rate, no significant correlation is observed with the one computed measuring the simulation time required for the proteins to reach their native state, $\tilde{R}$ [see Figs. 4(a) and 4(b)]; this was expected, given the distortion of time scales that is known to affect coarse-grained models. On the contrary, however, a rather strong positive correlation

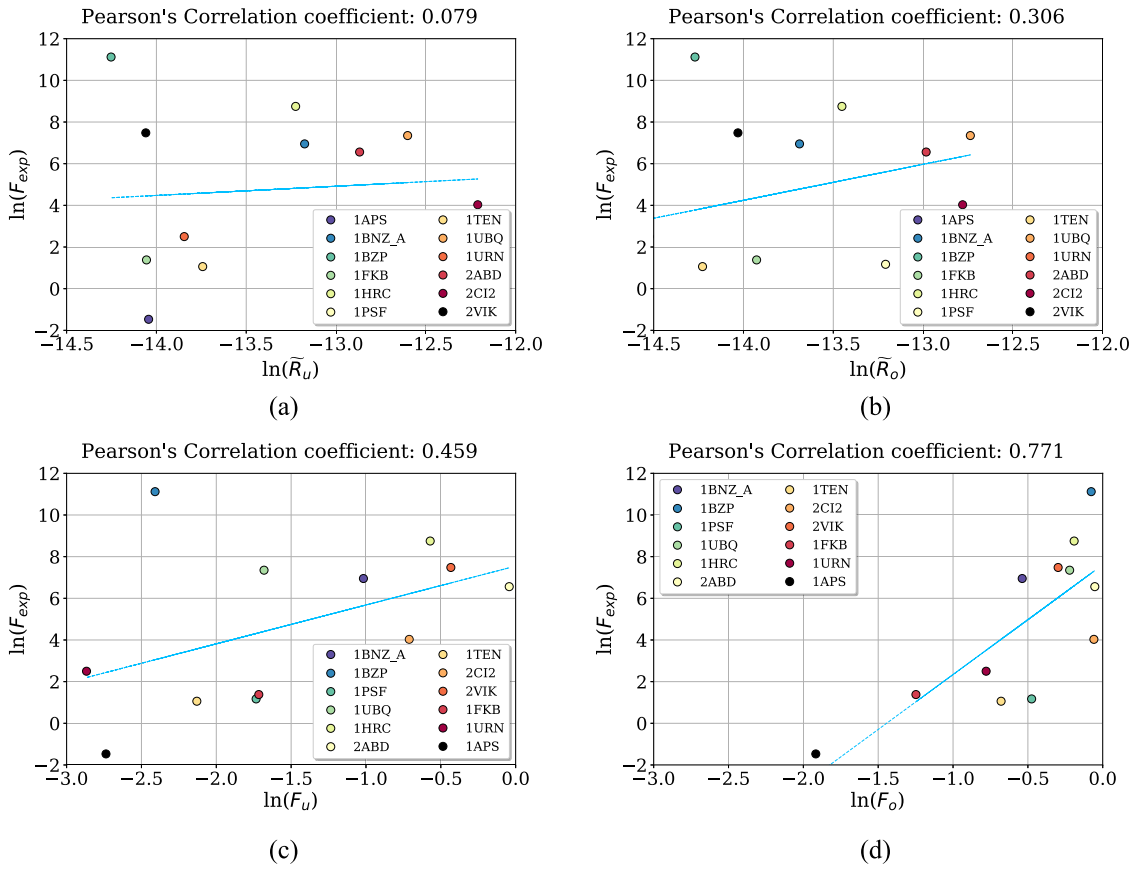

FIG. 4. Correlations of the experimentally estimated folding rate F_{exp}[23] vs the numerical folding rate $\tilde{R}$ (top row) and the folding frequency F (bottom row) for *unoptimized* (left column) and *optimized* (right column) force-fields.

emerged between the experimental estimated folding rate and the *in silico* folding frequency computed after the optimization [Figs. 4(c) and 4(d)]. On the one hand, the data show that the experimentally measured folding rate and the folding frequency computed from our simulations are largely determined by the same properties of the molecules under examination and that the optimization procedure enhances this correlation by endowing the model proteins with a capacity to avoid kinetic traps that is semi-quantitatively in line with that of the real proteins selected by evolution. On the other hand, however, the discrepancy between the two quantities highlights the fact that they indeed measure two related yet distinct properties of the system: while the experimental folding rates entail kinetic information, the folding probability only quantifies the reliability of the folding process, i.e., the capacity of the protein to reach the correct native state. Furthermore, the degree of frustration intrinsic to the proteins is certainly reduced and minimized by the EFM, its interactions, and the optimization process; however, it is possible that a certain amount of frustration remains, whose removal, if possible, would require to modify this coarse-grained model with an extension of its interactions and optimization with the inclusion of more structure-based properties, as previously discussed.

In conclusion, these results strengthen the previous observations and prove that, thanks to the force-field optimization procedure, the impact of topological complexity on the folding process can be captured to great extent by a simple model that employs no other input parameter beyond the native conformation.

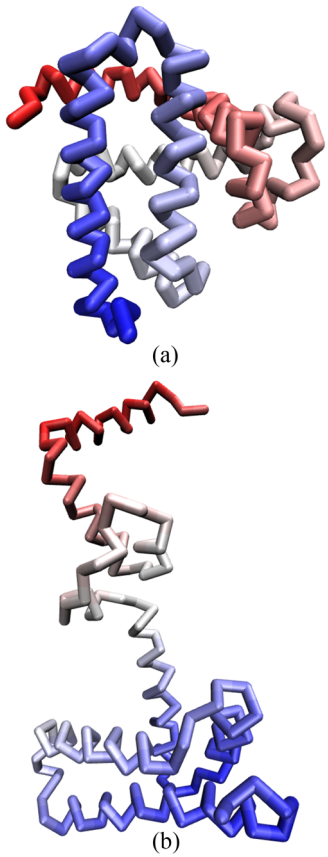

FIG. 5. Folded (a) and misfolded (b) state of 1BZP. Colors go from the C-term (red) to the N-term (blue).

C. Case study 1

1BZP is a myoglobin from sperm whale (*Physeter macrocephalus*), a 153 residue globular protein consisting of 8α helices separated by loops. It has $\langle|G'|_c\rangle = 0.429$ (Tables I and III). This protein showcases the importance of parameter refinement in the EFM: using MSD as a proxy for the successful folding, the force-field optimization greatly improved the folding rate, bringing it from 0.085 to 0.95 after 20 cycles (Table I, Fig. 2). The structure corresponding to a misfolded and a properly folded conformation is reported in Fig. 5, while the free energy surface (FES) of this protein's folding process in various conditions is provided in the supplementary material.

D. Case study 2

1URN is the 96-residue-long RNA-binding domain of the U1A spliceosomal protein[50] and has a $\langle|G'|_c\rangle = 1.038$ (Tables III and I).

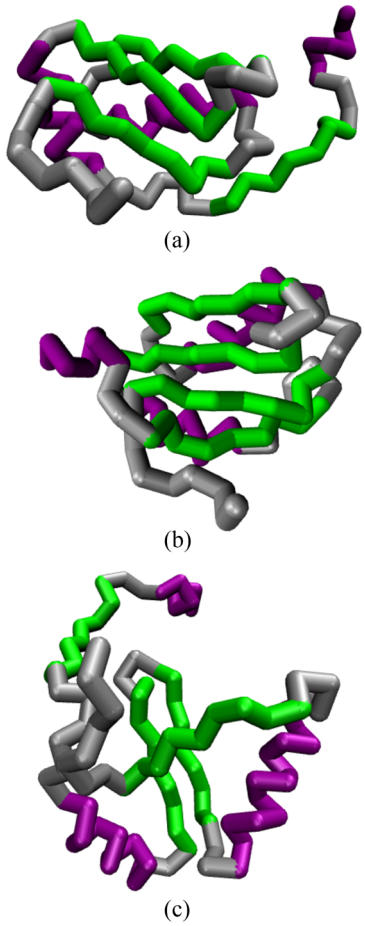

FIG. 6. EFM representation of 1URN. (a) Intermediate folded state: the N-terminal will form a hairpin and arrive in (b) the folded (native) state. (c) Misfolded state. α helices are highlighted in purple and β-sheets are in green. The N-terminal is the one with the small α helix structure.

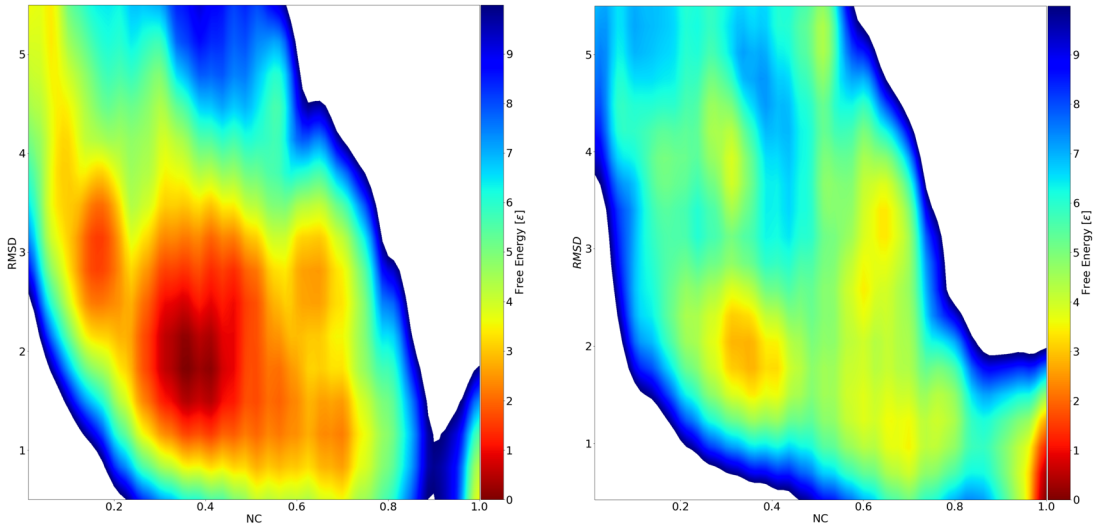

FIG. 7. Free energy surfaces for 1URN. (a) 1319 runs with *unoptimized* force-field; (b) 1009 folding runs with force-field optimized via NC-based ranking.

The native state presents two α-helices and one β-sheet [Fig. 6(b)]. The latter is composed of four antiparallel strands, two of which are on the N-terminal and C-terminal, respectively. In the correctly folded simulations, this structure is formed by initially bringing together three out of four of the β-strands and the 2 α-helices, forming the "bulk" of the structure and leaving the N-terminal unfolded [Fig. 6(a)]. This creates a hairpin between the C-terminal and the β strand immediately downstream. Finally, the N-terminal β-strand folds inside such hairpin and forms the complete β-sheet [Fig. 6(b)]. Simulation of this pathway was possible only after NC-based genetic optimization: the EFM model, in fact, essentially fails to fold 1URN to its native state before optimization. After force-field optimization using MSD as proxy, however, its folding success rate is still extremely low. Inspection of the FES associated with the folding runs of the unoptimized 1URN model shows that a free energy barrier emerges along the native contact reaction coordinate [Fig. 7(a)], separating the correctly folded state (NC ~ 1) from the rest of the conformational space. Since this feature is not visible along the MSD coordinate, this observation led us to change the reaction coordinate used to define the folded state in the genetic optimization from MSD to NC. As a result, after the new optimization, the folding success rate of 1URN dramatically increased (Fig. 2), effectively overcoming the NC barrier [Fig. 7(b)]. This suggests that the native contact formation plays an important role in the correct folding of the β sheet [Fig. 6(c)].

IV. DISCUSSION AND CONCLUSIONS

In this work, we have investigated the relation between structure and folding probabilities on a database of 12 topologically complex proteins via the EFM, a structure-based CG model. In the EFM, the protein is described as a chain of beads connected by bonds, where the non-bonded interactions are limited to excluded volume and the whole system-specific features are embedded in the angular potentials. The EFM can thus be classified as an "angular" Gō model, where the folding is enforced through local bending and rotations. Nonetheless, the absence of a bias on the native contacts (typical of Gō models) and the projection of the folding propensity on the angular interactions make it possible to effectively include non-native interactions at the same level of the native ones: this is a crucial aspect because of the role the former can play in the folding process of topologically complex proteins.[35,38,51,52] We stress once again that given the degree of approximation and physical ingredients retained by the EFM, the predictivity of the model solely concerns the native state topology and its formation.

Herein, we have shown that such a simple representation is sufficient to successfully fold the chain, starting from a stretched configuration, in the vast majority of the cases under examination. However, we have also noted that a nontrivial degree of topological complexity can hinder the folding process. Based on the observation that self-entangled proteins have likely evolved to avoid the kinetic traps introduced by topological constraints,[38] we have maximized the probability to reach the native state by optimizing the force-field coefficient of the model, employing a genetic optimization method.[40] The trends reported in Fig. 2 indeed show that the use of this methodology can increase the probability of folding, providing models that can autonomously and efficiently collapse to the native state as realistic proteins are believed to.

Concerning topological complexity, we found an extremely good agreement between the values of $|G'|_c$ computed by Ref. 23 and the averages $\langle |G'|_c \rangle$ over equilibrium simulations, reported in Table III; however, our results also point out that in the minimum of the potential energy function of a protein (i.e., the native state), the computed value $|G'|_c$ can fluctuate significantly at the equilibrium. This suggests that taking the average $\langle |G'|_c \rangle$ over an ensemble of conformations sampled at the equilibrium might be a more

informative indicator of the self-entanglement degree of a folded protein.

Interestingly, we were able to correlate *in silico* folding probabilities with experimental data on folding rates, highlighting the impact of topological complexity on the latter. It is remarkable how the agreement with experimental data improves after the optimization, thus supporting the core hypothesis behind EFM, according to which optimization recapitulates evolution, thus improving the model and bringing it closer to reality.

All the proteins considered in the present work can be labeled as *two-state folders*,[23] meaning that the transition from a completely denatured conformation to the native state follows a simple two-state process kinetics, without the formation of intermediate metastable states. By looking at the FESs (reported for all proteins in the supplementary material), we can observe the presence of "potential wells" in intermediate portions of the path, at least in some of the proteins, which may host intermediate states. Nonetheless, since all the degrees of freedom of the protein are projected onto CG interactions parameterized according to a top-down procedure, it is not clear *a priori* how much the simulated folding pathways correlate with an all-atom kinetics or an *in vitro* scenario. A comparison of these results with all-atom simulations is thus required, and it is the object of future studies.

In conclusion, we have shown that the reduced accuracy of the CG representation here employed and the lack of chemical detail are balanced by the remarkable computational efficiency, which enables one to generate statistically significant datasets of folding trajectories from which qualitative and semi-quantitative information can be extracted. The reported results thus showcase the effectiveness of simple models, such as the EFM, in tackling questions about the impact of geometry and topology on the folding process of proteins and support the notion that self-entanglement of the polypeptide chain plays a crucial role in the kinetics of protein folding.

SUPPLEMENTARY MATERIAL

See the supplementary material for details on the optimization process, free energy calculations, free energy surfaces, and folding times.

ACKNOWLEDGMENTS

The authors thank Thomas Tarenzi for a critical reading of the manuscript. The authors also acknowledge the contribution of the COST Action CA17139. Computational resources were provided by the Max Planck Computing and Data Facility and the HPC cluster of the University of Trento. C.P. and R.P. acknowledge funding from the European Union's Horizon 2020 research and innovation program under GOKNOT Marie Skłodowska-Curie Grant Agreement No. 796969.

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material and/or from the corresponding author upon reasonable request.

REFERENCES

[1] W. R. Taylor, Nature **406**, 916 (2000).
[2] P. Dabrowski-Tumanski and J. I. Sulkowska, Proc. Natl. Acad. Sci. U. S. A. **114**, 3415 (2017).
[3] S. E. Jackson, A. Suma, and C. Micheletti, Curr. Opin. Struct. Biol. **42**, 6 (2017).
[4] P. F. N. Faísca, Comput. Struct. Biotechnol. J. **13**, 459 (2015).
[5] N. C. H. Lim and S. E. Jackson, J. Mol. Biol. **427**, 248 (2015).
[6] M. Jamroz, W. Niemyska, E. J. Rawdon, A. Stasiak, K. C. Millett, P. Sułkowski, and J. I. Sulkowska, Nucleic Acids Res. **43**, D306 (2014).
[7] C. Perego and R. Potestio, J. Phys.: Condens. Matter **31**, 443001 (2019).
[8] G. M. Crippen, J. Theor. Biol. **45**, 327 (1974).
[9] M. Piejko, S. Niewieczerzal, and J. I. Sulkowska, Isr. J. Chem. **60**, 713 (2020).
[10] A. Begun, S. Liubimov, A. Molochkov, and A. J. Niemi, PLoS One **16**, e0244547 (2021).
[11] W. Niemyska, P. Dabrowski-Tumanski, M. Kadlof, E. Haglund, P. Sułkowski, and J. I. Sulkowska, Sci. Rep. **6**, 36895 (2016).
[12] J. M. Simien and E. Haglund, Trends Biochem. Sci. **46**, 461 (2021).
[13] D. Baker, Nature **405**, 39 (2000).
[14] K. W. Plaxco and D. Baker, Proc. Natl. Acad. Sci. U. S. A. **95**, 13591 (1998).
[15] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[16] P. D. Dixit and T. R. Weikl, Proteins: Struct., Funct., Bioinf. **64**, 193 (2006).
[17] D. E. Makarov and K. W. Plaxco, Protein Sci. **12**, 17 (2003).
[18] S. Wallin and H. S. Chan, Protein Sci. **14**, 1643 (2005).
[19] C. Micheletti, Proteins: Struct., Funct., Bioinf. **51**, 74 (2003).
[20] M. M. Gromiha and S. Selvaraj, J. Mol. Biol. **310**, 27 (2001).
[21] H. Gong, D. G. Isom, R. Srinivasan, and G. D. Rose, J. Mol. Biol. **327**, 1149 (2003).
[22] H. Krobath, A. Rey, and P. F. N. Faísca, Phys. Chem. Chem. Phys. **17**, 3512 (2015).
[23] M. Baiesi, E. Orlandini, F. Seno, and A. Trovato, J. Phys. A: Math. Theor. **50**, 504001 (2017).
[24] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Technical Report (Academic Press, 2002).
[25] M. Compiani and E. Capriotti, Biochemistry **52**, 8601 (2013).
[26] W. Noid, J. Chem. Phys. **139**, 090901 (2013).
[27] D. Fritz, C. R. Herbers, K. Kremer, and N. F. A. van der Vegt, Soft Matter **5**, 4556 (2009).
[28] M. Giulini, M. Rigoli, G. Mattiotti, R. Menichetti, T. Tarenzi, R. Fiorentini, and R. Potestio, Front. Mol. Biosci. **8**, 676976 (2021).
[29] B. C. Gin, J. P. Garrahan, and P. L. Geissler, J. Mol. Biol. **392**, 1303 (2009).
[30] H. Kaya and H. S. Chan, J. Mol. Biol. **326**, 911 (2003).
[31] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **89**, 8721 (1992).
[32] P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, Rep. Prog. Phys. **75**, 076601 (2012).
[33] R. D. Hills, Jr., L. Lu, and G. A. Voth, PLoS Comput. Biol. **6**, e1000827 (2010).
[34] L. L. Chavez, J. N. Onuchic, and C. Clementi, J. Am. Chem. Soc. **126**, 8426 (2004).
[35] S. a Beccara, T. Škrbić, R. Covino, C. Micheletti, and P. Faccioli, PLoS Comput. Biol. **9**, e1003002 (2013).
[36] D. Bölinger, J. I. Sułkowska, H.-P. Hsu, L. A. Mirny, M. Kardar, J. N. Onuchic, and P. Virnau, PLoS Comput. Biol. **6**, e1000731 (2010).
[37] S. Wallin, K. B. Zeldovich, and E. I. Shakhnovich, J. Mol. Biol. **368**, 884 (2007).
[38] S. Najafi and R. Potestio, J. Chem. Phys. **143**, 243121 (2015).
[39] S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. U. S. A. **106**, 434 (2009).
[40] C. Perego and R. Potestio, Biophys. J. **117**, 214 (2019).
[41] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10428 (1991).
[42] V. Grantcharova, E. J. Alm, D. Baker, and A. L. Horwich, Curr. Opin., Struct. Biol **11**, 70 (2001).
[43] K. Kremer and G. S. Grest, J. Chem. Phys. **92**, 5057 (1990).
[44] F. Norbiato, F. Seno, A. Trovato, and M. Baiesi, Int. J. Mol. Sci. **21**, 213 (2020).

[45]C. Huang, X. Yang, and Z. He, Comput. Biol. Chem. **34**, 137 (2010).
[46]C. F. Gauss, *Werke* (Königiche Gesellschaft der Wissenschaften zu Göttingen, Leipzig, Berlin, 1867), p. 605, note dated 22 January 1833.
[47]R. L. Ricca and B. Nipoti, J. Knot Theory Ramifications **20**, 1325 (2011).
[48]S. F. Edwards *et al.*, *The Theory of Polymer Dynamics* (Oxford University Press, 1986).
[49]A. A. Louis, arXiv:1001.1166 (2010).
[50]C. Oubridge, N. Ito, P. R. Evans, C.-H. Teo, and K. Nagai, Nature **372**, 432 (1994).
[51]R. Covino, T. Škrbić, P. Faccioli, C. Micheletti *et al.*, Biomolecules **4**, 1 (2013).
[52]A. Kluber, T. A. Burt, and C. Clementi, Proc. Natl. Acad. Sci. U. S. A. **115**, 9234 (2018).