Biophysical Journal
Article

Biophysical Society

Searching the Optimal Folding Routes of a Complex Lasso Protein

Claudio Perego[1,*] and Raffaello Potestio[2,3]
[1]Polymer Theory Department, Max Planck Institute for Polymer Research, Mainz, Germany; [2]Department of Physics, University of Trento, Trento, Italy; and [3]INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, Trento, Italy

ABSTRACT Understanding how polypeptides can efficiently and reproducibly attain a self-entangled conformation is a compelling biophysical challenge that might shed new light on our general knowledge of protein folding. Complex lassos, namely self-entangled protein structures characterized by a covalent loop sealed by a cysteine bridge, represent an ideal test system in the framework of entangled folding. Indeed, because cysteine bridges form in oxidizing conditions, they can be used as on/off switches of the structure topology to investigate the role played by the backbone entanglement in the process. In this work, we have used molecular dynamics to simulate the folding of a complex lasso glycoprotein, granulocyte-macrophage colony-stimulating factor, modeling both reducing and oxidizing conditions. Together with a well-established Gō-like description, we have employed the elastic folder model, a coarse-grained, minimalistic representation of the polypeptide chain driven by a structure-based angular potential. The purpose of this study is to assess the kinetically optimal pathways in relation to the formation of the native topology. To this end, we have implemented an evolutionary strategy that tunes the elastic folder model potentials to maximize the folding probability within the early stages of the dynamics. The resulting protein model is capable of folding with high success rate, avoiding the kinetic traps that hamper the efficient folding in the other tested models. Employing specifically designed topological descriptors, we could observe that the selected folding routes avoid the topological bottleneck by locking the cysteine bridge after the topology is formed. These results provide valuable insights on the selection of mechanisms in self-entangled protein folding while, at the same time, the proposed methodology can complement the usage of established minimalistic models and draw useful guidelines for more detailed simulations.

SIGNIFICANCE We have investigated the folding mechanism of granulocyte-macrophage colony-stimulating factor, a glycoprotein that handles diverse functions in the human body. This protein folds in a rather common self-entangled conformation named complex lasso. Understanding how a polypeptide encodes into its sequence the capability of tying itself into such kinds of self-entangled structures would represent a major advancement in the comprehension of protein folding. To study this folding mechanism, we have employed molecular dynamics simulations, using both a well-known minimalistic model of the protein and an alternative model specifically designed to highlight the preferential pathways of entangled folding. Our calculations show how the protein can avoid the kinetic traps related to self-entanglement, managing to fold in a reproducible and efficient way.

INTRODUCTION

Almost a quarter of a century of research has been dedicated to the study of proteins that exhibit a self-entangled native fold. Nowadays, up to 6% of the structures deposited in the Protein Data Bank (PDB) (1) are self-entangled proteins (2,3). Since the first natively knotted protein was discovered in 1994 (4), the existence of such topologically complex folds has represented a new challenge in the understanding of protein folding, fostering a wide range of studies. A number of reviews addressing the topic of self-entangled proteins can be found in the literature (see, e.g., (2,5–7), just to name the most recent), each addressing a different aspect of this variegated research field. The discovery of self-entangled protein structures has raised a few crucial questions related to their scarcity (8,9), their conservation along evolution (10,11), and their possible biological function (2,7,12–14).

In this work, we address the following question: how can the amino-acid chain fold reproducibly and efficiently into a

Submitted January 9, 2019, and accepted for publication May 30, 2019.

*Correspondence: perego@mpip-mainz.mpg.de

Editor: Margaret Cheung.

https://doi.org/10.1016/j.bpj.2019.05.025

specific, nontrivial topology? Many experiments were conducted to answer this question, showing, e.g., that these proteins can spontaneously tie themselves into the native topology (15); that the formation of the entanglement is a rate-limiting step (15–17); and that one or few folding routes happen to be dominant, presumably representing the most efficient and reliable mechanisms (18). These crucial results demonstrate that self-entangled folding clearly differentiates from the simple picture of two-state folding of small, non-entangled proteins, but it is evident that efforts are still needed to reach a comprehensive and sound picture of this phenomenon.

In this framework, an interesting class of proteins is represented by complex lassos (CLs) (19), entangled structures that exhibit a covalent loop closed by a disulphide bridge. The surface of this loop is pierced one or more times by the polypeptide chain, forming a nontrivial topology. Since leptin was classified as the first CL protein (20), this topological state has been found to be widespread in the known PDB structures, characterizing ~18% of the proteins containing a cysteine bridge (21). Most of the CLs are secreted proteins with signaling functions, and their topology is believed to have a crucial role in their biological activity (22,23). Moreover, the topology of CLs can be controlled externally because the cysteine bridge is stable in an oxidizing solution, whereas it does not form in a reducing environment. This feature allows one to directly study the effect of the topological barrier on the folding mechanism, making CLs ideal test systems for a deeper understanding of entangled folding.

As for simple proteins, the experimental probe of folding pathways in self-entangled proteins such as CLs can only provide indirect indications. For this reason, molecular dynamics (MD) simulation represents an essential complementary tool for the study of the process. We must, however, stress that the time duration of self-entangled folding typically exceeds the range accessible by all-atom simulations employing realistic interactions. This is the reason why, except in two notable cases (24,25), the available computational results have been obtained using simplified, minimalistic protein models, which allow for a thorough sampling of the conformational space while at the same time providing indications on the theoretical principles of the folding.

By, far the most used methods are the so-called Gō models (GōMs) (see, e.g., (26,27)), named after the pioneering work of Gō (28). In GōMs, the protein is described as a heteropolymer chain that encodes its native fold in the interaction potential. This kind of description stems from the established energy landscape theory, according to which proteins have evolved to fold along a smooth, funneled free-energy landscape. Such a "folding funnel" determines the efficient and reproducible collapse of the denatured polymer chain to its compact and functional three-dimensional structure (29). The majority of GōMs employ a coarse-grained (CG) representation of the protein, in which each residue is mapped onto a sphere centered at the position of the C_α atoms. The residues in contact in the native state interact via attractive pair potentials, defined so that the energy minimum of the model corresponds to the native fold. This picture assumes that folding is dominated by native contact interactions, whereas non-native interactions play a minor role (30). GōMs have been validated using both experimental data and more detailed simulation models (31–36), and their predictions are considered reliable in the framework of small protein folding.

GōMs have been widely used to study the folding of entangled proteins, providing valuable indications on their thermodynamics and kinetics (37–41) as well as in the framework of lasso folding (20,22,23). However, the underlying theory clashes with the presence of knots in proteins because the formation of entanglements implies a high degree of coordination at different length scales that can hardly be encoded in native contact potentials. For example, the mandatory passage through the specific, nonalternative folding intermediates imposed by topological barriers can trigger the untimely formation of native contacts, which can entrap the molecule in misfolded states. When this happens, the protein has to break such contacts and retrace the proper folding route. On the one hand, this "backtracking" process can explain the longer folding times measured for knotted proteins; on the other hand, it lowers the capability of GōMs to fold reproducibly, resulting in very low success rates (40).

For this reason, the possibility of including non-native interactions within GōMs has been explored, obtaining significant improvements in the folding efficiency (42–45). This suggests that non-native interactions can play a crucial role in topologically complex folding, regulating the timing of native contacts formation, and guiding the concerted nonlocal moves required for the tying of the backbone (46). Moreover, in agreement with energy landscape theory, the folding of GōMs exhibits multiple pathways reaching the folded state (38,42,47) differently from the indications of all-atom MD (24) and experiments (18), which suggest the reproducible selection of a single route.

The presence of a dominant pathway can indicate that evolution has optimized knotted proteins in their folding behavior, minimizing the probability of misfolding, and promoting the most reliable and fast folding routes. Building on this optimality principle, in (48), an alternative CG description for the study of knotted folded proteins has been proposed. This model, dubbed the elastic folder model (EFM), is a CG, minimalistic description in which the folding of the polypeptide is driven exclusively by backbone bending and torsion potentials. EFM embodies the idea that the folding process has been kinetically optimized by evolution in that it promotes the most efficient pathways of the backbone across the topological bottlenecks of knotted folding. To attain this optimality, once a specific protein is

chosen, the relative magnitudes of its angular forces are tuned via a stochastic process aimed at maximizing the folding success rate. The heterogeneous force-field obtained through this optimization procedure represents a sort of mean-field approximation of the cooperation between native and non-native interactions and can provide valuable information on the folding mechanisms of the system under examination. This model has been used to investigate the folding of two small knotted proteins (48), observing a qualitative agreement with the all-atom simulations results of (24).

In this work, we have employed EFM simulations to study the folding of a glycoprotein, granulocyte-macrophage colony-stimulating factor, that exhibits a CL native state. We have extended the original EFM, introducing contact interactions between those cysteines that form a disulfide bridge in the native conformation. This allowed us to simulate the folding in oxidizing conditions, assessing the differences with respect to the process in a reducing environment. The angular potentials of this protein model have been optimized with an evolutionary strategy that could tune the model to fold reproducibly and rapidly, avoiding kinetic traps and efficiently surpassing the topological bottleneck associated to the formation of the lasso. The resulting dynamics has been compared with that of a well-established GōM (26) with the purpose of enlightening the most efficient folding pathways in relation with the topological state of the protein. To this aim, we have also introduced and employed two topological variables that, building on the minimal surface analysis (19) and the Gauss linking number (49) methods, allow for monitoring the evolution of the CL topology along the MD trajectory.

As a result, we could outline a detailed picture of the folding scenario, demonstrating that the same, kinetically optimal mechanism dominates in both reducing and oxidizing conditions. This folding route, characterized by the formation of the cysteine bridge after the lasso topology, is supported both by the GōM simulations at the fastest folding temperature and by the optimized EFM. These results show how the principle of kinetic optimality can determine the selection of a single folding mechanism among the possible ones and qualify the considered protein as an interesting testing ground for all-atom simulations or experimental study.

METHODS

In this work, we have studied the folding of granulocyte-macrophage colony-stimulating factor, a monomeric glycoprotein that acts as growth factor for white blood cells. We shall refer to the protein by using the PDB code of its crystal structure, 2GMF (50). 2GMF is a helical cytokine formed by 127 residues, of which 121 are resolved in the PDB structure, shown in Fig. 1. As highlighted in the figure, 2GMF forms two cysteine bridges, which we name b_1, connecting residues 88 and 121, and b_2, connecting residues 54 and 96. 2GMF is classified as an L_2 lasso structure, in which the covalent loop formed by b_1 is threaded by a 12-residue hairpin from residue 43 to residue 53. Instead, b_2 does not determine any lasso topology.

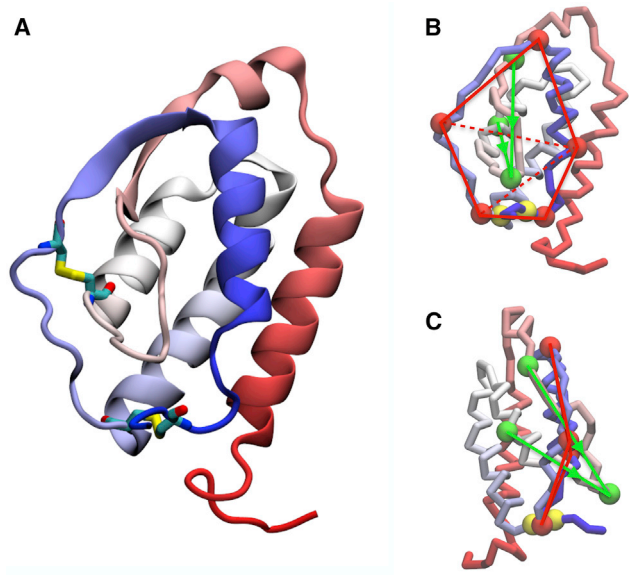

FIGURE 1 2GMF protein structure and geometry of topological descriptors. (*a*) A cartoon representation of chain A of 2GMF in its native fold is shown. The cysteine bridges are shown with atomistic resolution. (*b*) A view of 2GMF native structure, showing only the C_α residues, is given. Cysteines 88 and 121, forming b_1, are represented as yellow beads. The structure reduction employed for the definition of the topological variables L and G (see the text) is also displayed: the five residues chosen to represent the loop are highlighted as red circles connected by red lines, and the three residues representing the threading hairpin are highlighted as green circles connected by green lines. The red dashed lines indicate how the loop surface is divided in three triangles for computing L. The green arrows represent the integration verse along the hairpin segments used for the calculation of G. (*c*) shows the same as (*b*) but rotated. The color of backbone residues depends on their index along the chain, going from red (N-terminal) to blue (C-terminal). VMD was employed for the protein visualization (75). To see this figure in color, go online.

Three different MD models of the protein were employed: a non-optimized EFM with homogeneous stiffness coefficients, an optimized EFM obtained with the MFFO procedure presented in the following, and a GōM constructed using the native-contact-based description proposed by Clementi et al. (26). The folding of 2GMF was studied by performing sets of MD runs starting from random stretched configurations in both reducing and oxidizing conditions. As discussed in the following, the stability of cysteine bridges in an oxidizing environment is modeled in the EFM by means of native contact potentials between the cysteine pairs and in the GōM by rescaling the existing native contacts. We shall employ natural units, indicating energies in units of E, temperatures in units of E/k_B, lengths in units of σ, and time-lengths in units of $\tau_{MD} = \sigma\sqrt{m/\epsilon}$, m being the bead mass.

Elastic folder model simulations

The EFM introduced in (48) is here reviewed in detail. The model describes an N-residues polypeptide chain by means of a CG representation, in which only the C_α atom positions are retained, resulting in a chain of N identical beads connected by stiff bonds. The steric hindrance of each residue is represented by a short-range excluded volume interaction. As said, the driving force of the folding is modeled by bending and torsion potentials, parameterized so that the energy minimum is attained for a chosen reference

configuration. In principle, this reference corresponds to the native PDB structure; however, other choices can be convenient as well (48).

The total potential energy is

$$U_{\text{tot}} = U_{\text{steric}} + U_{\text{bonds}} + U_{\text{angular}} + U_{\text{bridges}}. \tag{1}$$

Weeks-Chandler-Anderson interaction (51) is used for the steric term:

$$U_{\text{steric}} = \sum_{i<j}^{N} U_{\text{WCA}}\left(r_{i,j}\right), \tag{2}$$

where $r_{i,j} = |\mathbf{r}_i - \mathbf{r}_j|$ and the pair potential is given by

$$U_{\text{WCA}} = \begin{cases} U_{LJ}(r;\epsilon,\sigma) + \epsilon & \text{if } r < 2^{1/6} \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

$$U_{\text{torsion}}\left(\phi;\phi^0,k_1,k_3\right) = k_1\cos\left(\phi-\phi^0\right) + k_3\cos\left(3\phi-3\phi^0\right). \tag{8}$$

In the EFM, we consider a single torsion coefficient, imposing $k_3 = k_1/3$. Angular interactions such as Eqs. 7 and 8 (or analogous chiral potentials) have been included in GōMs as well (26,27) to bias the formation of proper backbone chirality (53).

In this work, we have modeled the formation of disulfide bridges by introducing an attractive potential term U_{bridge} between those n_{B} cysteine pairs $\{c_1, c_2\}$ that form a bridge in the reference state:

$$U_{\text{bridge}} = \sum_{\{c_1,c_2\}} U_{sLJ}(r_{c_1c_2};\epsilon_b,\sigma_b), \tag{9}$$

in which $r_{c_1c_2}$ is the distance between the cysteines and U_{sLJ} is a truncated and force-shifted LJ potential, given by

$$U_{sLJ} = \begin{cases} U_{LJ}(r) - U_{LJ}(r_c) - (r-r_c)\dfrac{dU_{LJ}}{dr}\Big|_{r=r_c} & \text{if } r < r_c \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

in which U_{LJ} is the Lennard-Jones potential:

$$U_{LJ}(r;\epsilon,\sigma) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \tag{4}$$

The chain beads are connected via finitely extensible nonlinear elastic (FENE) bonds (52), namely

$$U_{\text{bonds}} = -\sum_{i=0}^{N-2}\frac{k_{\text{FENE}}}{2}\left(\frac{R_0}{\sigma}\right)^2 \ln\left[1-\left(\frac{r_{i,i+1}}{R_0}\right)^2\right], \tag{5}$$

in which k_{FENE} is the interaction strength parameter and R_0 is the maximal bond length. The length scale σ is chosen equal to the steric diameter of Eq. 2, which corresponds to the separation between two consecutive C_α, i.e., roughly 3.8 Å.

The remaining terms of Eq. 1 contain specific structural information of the protein that has to be described. As mentioned, the folding is guided by the angular potential, which generates the dynamics of the chain bending and torsion angles:

$$\begin{aligned} U_{\text{angular}} = &\sum_{i+1}^{N-2} U_{\text{bending}}\left(\theta_i;\theta_i^0,k_i^{\text{bend}}\right) \\ &+ \sum_{i+1}^{N-3} U_{\text{torsion}}\left(\phi_i;\phi_i^0,k_{1i}^{\text{tor}},k_{3i}^{\text{tor}}\right), \end{aligned} \tag{6}$$

in which θ_i^0 and ϕ_i^0 are, respectively, the i-th bending and torsion angles of the reference conformation. k_i^{bend} and k_i^{tor} are the stiffness coefficients associated to the angular potentials, which are given by

$$U_{\text{bending}}\left(\theta;\theta^0,k\right) = k\left(\theta-\theta^0\right)^2 \tag{7}$$

and

The scale length σ_b is chosen so that the minimum of U_{bridge} is located at the reference distance between the residues in the considered pair.

The folding of this protein model is studied, simulating its Langevin dynamics starting from a stretched (i.e., end-to-end distance $\sim N\sigma$), randomly generated configuration. The potential parameters, as well as the MD settings, were chosen following the previous work on EFM (48). The FENE parameters had the typical values $k_{\text{FENE}} = 30$ and $R_0 = 1.5$, the friction time of Langevin equation was $\tau_{\text{frict}} = 1.0$, and the integration time step was $\Delta t = 5 \times 10^{-4}$. The EFM dynamics was integrated by means of an in-house software.

Single force-field optimization

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 the folding. In (48), this optimization is performed through a stochastic search procedure, which we recall here. Let us first define

$$\mathrm{K} = \left\{k_1^{\text{bend}},\ldots,k_{N-2}^{\text{bend}},k_1^{\text{tor}},\ldots,k_{N-3}^{\text{tor}}\right\} = \tag{11}$$

$$= \left\{k_1^{\text{ang}},\ldots,k_{2N-5}^{\text{ang}}\right\}, \tag{12}$$

in which k_i^{ang} is used for both bending and torsion stiffnesses. We shall refer to K as the force-field of the model. The optimization step consists of two operations. First, a mutated force-field K' is generated:

$$K' = \left\{k_1^{\text{ang}},\ldots,k_j^{\text{ang}}+\delta k,\ldots,k_{2N-5}^{\text{ang}}\right\}, \tag{13}$$

in which the j-th coefficient is modified by adding δk. j is randomly chosen among the $2N-5$ coefficients, whereas δk is generated with a prescribed probability distribution (e.g., in our calculation, it is normally distributed, with SD equal to 2.5). Second, the mutation to K' is accepted or rejected according to a Metropolis-like criterion: K' is tested by performing a set of n parallel folding simulations, starting from a randomly generated stretched configuration and running for some properly chosen

length τ_{run}. The outcome of the n test trajectories is then assessed by measuring $\mathcal{F}$, namely the mean-square displacement (MSD) from the target configuration $\mathbf{R}^0$, defined as

$$\mathcal{F}(t;K') = \frac{1}{N\sigma^2}\left|\mathbf{R}(t) - \mathbf{R}^0\right|^2, \tag{14}$$

where $\mathbf{R}(t)$ is the configuration vector of the protein model and $|.|$ is the Euclidean distance. We then define

$$\langle\mathcal{F}(\tau;K')\rangle = \frac{1}{n}\sum_{i=1}^{n}\mathcal{F}^i(\tau;K'), \tag{15}$$

which is the average MSD computed at $t=\tau$ over the n test runs. τ is chosen so that Eq. 15 provides a measure of the folding success of the test runs. It can be set, e.g., equal to τ_{run} or, as in (48), chosen according to the convergence of the MSD value along the trajectory. In this work, we have selected $\tau = \tau_{\text{min}}$, namely the time at which the MSD reaches its minimal value during the test run. The probability of acceptance of the new force-field K' is then

$$P(K'\,|\,K) = \min\{1, \exp[\langle\mathcal{F}(\tau;K)\rangle - \langle\mathcal{F}(\tau;K')\rangle]\}. \tag{16}$$

The operation just described is then iterated to minimize $\langle\mathcal{F}\rangle$, enhancing the average success rate of the folding trajectories. A schematic representation of this procedure, which we name single force-field optimization (SFFO), is displayed in Fig. 2.

For a polypeptide such as the smallest knotted protein MJ0366, with $N = 82$ residues, the parameter space is quite large, and the SFFO algorithm can explore only a minimal portion of it in reasonable computation time. The situation can be partially improved by constraining the k^{ang} to be locally equal. For example, in (48) as well as in this work, neighboring pairs of coefficients are constrained, that is,

$$\begin{aligned} K = \{&k_1^{\text{bend}} = k_2^{\text{bend}}, k_3^{\text{bend}} = k_4^{\text{bend}}, \ldots, k_1^{\text{tor}} \\ &= k_2^{\text{tor}}, k_3^{\text{tor}} = k_4^{\text{tor}}, \ldots\}. \end{aligned} \tag{17}$$

FIGURE 2 Schematic illustration of the SFFO and MFFO algorithms. To see this figure in color, go online.

These local constraints reduce the dimensionality of the stochastic search but also the generality of the model. In this work, we have employed Eq. 17, pairing neighboring angular coefficients.

Multiple force-field optimization

In this work, we have employed a development of the SFFO strategy, aiming at a more efficient exploration of the K-space. The basic idea is to apply SFFO for the parallel optimization of several force-fields and then combine the results with an evolutionary strategy, as graphically illustrated in Fig. 2. An initial set or population of force-fields $\{K_j\}_{j=1}^{N_K}$ is chosen, and each of them undergoes m SFFO steps independently from the others. The resulting N_K mutated force-fields are then ranked according to their capability of folding. The specific ranking criterion is discussed in detail later on. The N_{win} top-ranked force-fields, which we shall call "winners," are selected to build the new population $\{K'_j\}_{j=1}^{N_K}$ while the remaining, low-ranked candidates are discarded. The new force-field population is given by

$$\{K'_k\}_{k=1}^{N_K} = \left(\{W_i\}_{i=1}^{N_{\text{win}}}, \{H_j\}_{j=1}^{N_K - N_{\text{win}}}\right), \tag{18}$$

in which W indicates the winners and H indicates a set of $N_K - N_{\text{win}}$ newly generated force-fields, which we shall refer to as "hybrid." The latter ones are obtained by means of a crossover operation, typical of genetic algorithms (see, e.g., (54)). In more detail, the H_j values are generated by combining fragments of force-fields randomly picked from a set of parent force-fields, as displayed in Fig. 3. The parent set is formed by the N_{win} winners together with N_{low} "low-fit" candidates, which ensures diversity among the population. The latter can be selected among the worst-ranked force-fields or, otherwise, generated with randomly distributed angular coefficients. Further details about the crossover operation are provided in the Supporting Materials and Methods. Once the new population is set, the optimization cycle is completed, and the algorithm is reiterated. We name this procedure multiple force-field optimization (MFFO).

We now discuss the criterion for the force-field ranking, which naturally builds on the outcomes of the folding tests gathered during the SFFO steps. As explained, each SFFO mutation is tested via n folding simulations. The resulting n trajectories can provide indications on the folding propensities of the N_K force-fields. One can, e.g., compare the average MSD (Eq. 15) attained by each force-field. Another possibility, which we have adopted in this work, is to rank the N_K candidates according to P_f, namely the folding probability along the test runs. More precisely, we have defined an estimate π_f of the folding probability based on the measurement of the

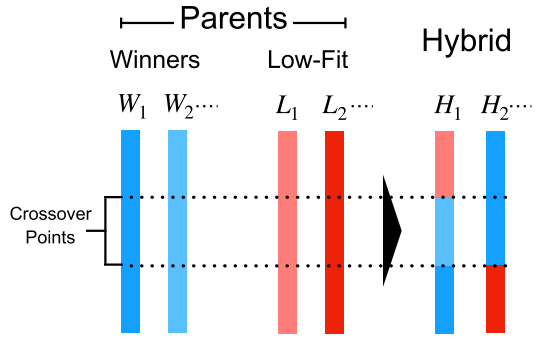

FIGURE 3 Schematic representation of the crossover operation generating the hybrid force-fields. The color bars indicate the sets of k^{ang} coefficients associated to the winners and the low-fit force-fields. These are mixed randomly in the hybrid force-fields. To see this figure in color, go online.

MSD along the test runs. A threshold value $\mathcal{F}_0$ has been set, below which the protein is considered to be in the native state.

Then, for a set of n test runs, we have defined

$$\pi_f(\mathcal{F}_0, \tau) = \frac{1}{n}\sum_{i=1}^{n}\theta[\mathcal{F}_0 - \mathcal{F}(\tau, K)], \tag{19}$$

where θ is a function that switches from 0 to 1 when its argument becomes positive. In particular, we used a Fermi function

$$\theta(z) = \left[1 + \exp\left(-\frac{z}{w}\right)\right]^{-1} \tag{20}$$

that switches continuously with length scale w. Clearly, π_f represents only a proxy of the real folding probability, on the one hand, because the sole MSD is not always reliable in discriminating between the native basin and misfolded configurations, and on the other hand, because it depends on a limited number of finite trajectories. Nonetheless, we have verified that the value of $\mathcal{F}$ represents a suitable descriptor to identify the great majority of folding trajectories for the chosen test case (2GMF), and this positively affected the outcome of the MFFO. In general, the evaluation of π_f should involve proper variables that differentiate the folded conformations, such as the fraction of native contacts or the topological variables introduced in the following.

As mentioned, the ranking operation has been performed every m SFFO iterations. Therefore, the trajectories employed in computing Eq. 19 come from the m-th iteration. However, we can assume that the local mutations tested along each SFFO step have a relatively small effect on the force-field folding propensity. It is thus convenient to include in the ranking also the information from the previous $m - 1$ SFFO steps. To achieve this, we have employed an exponential moving average, defined by the iterative formula

$$\Pi_f^{(i)} = \alpha\pi_f^{(i)} + (1-\alpha)\Pi_f^{(i-1)}, \tag{21}$$

where $\pi_f^{(i)}$ is the folding probability relative to the i-th SFFO iteration and α is the smoothing factor $0 < \alpha < 1$. The final value, i.e., $\Pi_f \equiv \Pi_f^{(m)}$, includes the contribution of all m SFFO iterations, assigning them a weight that increases exponentially with i. Thus, the N_K force-field candidates have been ranked by increasing values of Πf.

In the optimization presented in this work, the MFFO strategy has been applied to a population of $N_K = 16$ force-fields, initially having homogeneous angular coefficients $k_i^{\text{bend}} = k_b$ and $k_i^{\text{tor}} = k_t$, where k_b and k_t were chosen among the possible combinations of 20.0, 40.0, 60.0, or 80.0. Each force-field was optimized via SFFO, during which it mutated pointwise. The local mutations were accepted via a Metropolis criterion, based on the average MSD of 16 parallel folding trajectories (Eqs. 15 and 16) of length $\tau_{\text{run}} = 3.5 \times 10^3$. This trajectory time length has been chosen based on the folding times measured for the HM to promote only the faster folding routes. Every $m = 50$ steps, the force-fields were ranked according to the value of Π_f, as given by Eqs. 19 and 21, where the threshold MSD was $\mathcal{F}_0 = 0.9$, the switching length scale $w = 0.2$, and the smoothing factor $\alpha = 0.03$, corresponding to a decay time of $\tau_\alpha = 33$ steps of the exponential moving average weight. As mentioned, the resulting Π_f is a proxy of the success probability P_f, which provided an on-the-fly estimate of the optimization progress. After the force-fields were ranked, the six best were chosen as winners and continued the optimization. The remaining 10 force-fields were constructed combining the winners and four randomly generated forcefields, with k_i^{bend} and k_i^{tor} uniformly distributed between 30.0 and 60.0 (more details are reported in the Supporting Materials and Methods).

Gō model simulations

The employed GōM is that introduced by Clementi et al. in (26). The system setup was generated using the SMOG web server (http://smog-server.org) (55). Details on the interaction potential, which is based on 12-10 Lennard-Jones native contacts, can be found in the cited references. The shadow contact map (56) is used for the definition of native contacts. As mentioned before, this description also models the backbone stiffness with the angular potentials of Eq. 6. The stiffness coefficients are here homogeneous, set to $k^{\text{bend}} = 40.0$, $k_1^{\text{tor}} = 1.0$, and $k_3^{\text{tor}} = 0.5$. The formation of cysteine bridges in oxidizing condition is modeled by increasing the amplitude E_{ij} of the native contact potential associated to the cysteine-cysteine contacts. The value was set to $E_{ij} = 10\ k_{\text{B}}T$ so that thermal fluctuations would hardly break the bridge once formed.

As for the EFM, the GōM folding is studied by means of Langevin dynamics, starting from random stretched configurations. GROMACS 2018.3 package (57,58) was used for integrating the motion. The MD parameters were chosen consistently with the EFM simulations, with time step $\Delta t = 5 \times 10^{-4}$ and friction time $\tau_{\text{frict}} = 1.0$. To select the simulation temperature, we have performed a study of folding times and probabilities at different values of T; the results are presented in the Supporting Materials and Methods.

Topology analysis

Minimalistic CG models make it possible to collect a large statistics of folding trajectories, even in complex folding processes like those of self-entangled proteins. However, to gather useful information on the folding dynamics, the analysis of these trajectories strongly benefits from the definition of proper topological descriptors. Many methods for detecting the entangled state of a polymer chain have been proposed (see, e.g., (59) for further details) and extensively applied. For example, in the framework of knotted proteins, knot searching algorithms have been used to classify the topology of known native structures, gathering a comprehensive database (3). In general, these techniques operate on the three-dimensional structure of a polymer chain, first, by associating it to an equivalent closed curve (60,61) so that the topological state is mathematically well-defined, and second, by simplifying this structured curve without changing its topology (8,62). The resulting curve is then analyzed by computing topological invariants (63,64), and its entangled state is classified.

Although this is the typical approach used to analyze knotted proteins, the nontrivial topology recognized in CL structures is not yet classified from a mathematical point of view (2). In (19), an approach specifically aimed at detecting CLs is presented. This technique, named minimal surface analysis, uses triangulation algorithms borrowed from computer graphics to determine the minimal area surface spanned by a protein covalent loop. When this surface is obtained, the lasso type is detected by searching for segments of the backbone that pierce the minimal surface. This is a robust method to assess and classify CL structures, and it has been employed to establish a database of polymeric structures characterized by this topology (19,21). However, in our work, we are interested in descriptors that can monitor the topological state along the folding trajectory of a specific protein. To this purpose, the computation can be expensive, and a faster, less general method could be more effective. We can exploit the fact that proteins fold reproducibly in a well-defined topology, which is known a priori. For this reason, we relax the generality of the topological descriptor and focus on the specific native geometry of the system under consideration. In CL geometries, the main topological feature is a covalent loop closed by a cysteine bridge, pierced by part of the backbone. For simplicity, we limit the discussion to the case of a single loop and a single threading segment; the strategy can be then generalized to more complex topologies. Let $l_1, \ldots, l_{N_l}$ be the indexes of loop residues and $t_1, \ldots, t_{N_t}$ be the indexes of the threading segment residues. We operate a reduction of the structure, selecting

only few crucial residues, namely $l'_1, \ldots, l'_{M_l}$ for the loop and $t'_1, \ldots, t'_{M_t}$ for the threading tail, where $M_l < N_l$ and $M_t < N_t$. The residues l' and t' are chosen so that their position can describe whether the protein is in the native topology or not. This operation is similar to the smoothing performed for protein knot detection (65); however, the procedure is not automated and needs some preliminary analysis of the structure and folding behavior. For clarity, in Fig. 1 (and in the Supporting Materials and Methods), the reduction we adopted for 2GMF is illustrated. The surface spanned by the M_l loop residues is then approximated by $M_l - 2$ triangles, with vertexes corresponding to the l' residues positions. After this, the threading of an $|\mathbf{R}_{t'+1} - \mathbf{R}_{t'}|$ segment through the loop can be verified by computing its intersections with the surface triangles. Once the number and directions of the piercings through the loop surface are determined, it is clear whether the protein has attained its native topology. By means of continuous switching functions (see, e.g., Eq. 20), we can associate this binary information to a continuous value L varying from 0 (non-native topology) to 1 (native topology); we name this quantity the lasso variable. The approximated surface formed by the $M_l - 2$ triangles is not the minimal area surface of (19), which is typically formed by many more triangles. However, in our study, this simplification is convenient to speed up the calculations.

Another interesting approach to topology detection is adopted in (49,66). The idea developed in these works is that of employing the Gauss linking number (67), namely the double line integral:

$$G \equiv \frac{1}{4\pi} \int_{\gamma_1} \int_{\gamma_2} \frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3} \cdot (\mathrm{d}\mathbf{r}_1 \times \mathrm{d}\mathbf{r}_2), \qquad (22)$$

in which the integrals are performed along the two curves γ_1 and γ_2, $\mathbf{r}_1$ and $\mathbf{r}_2$ being the position vectors belonging to γ_1 and γ_2, respectively. If the two curves are closed (in $\mathbb{R}^3$), G takes an integer value that is a topological invariant typically used to define links. By applying a proper closure procedure, G can be therefore employed to detect the entanglement of two chains.

A crucial observation is that when γ_1 and γ_2 are not closed, G is not an integer topological invariant, but it still provides relevant information on the curves' mutual entanglement (67). This property can then be exploited to assess the linking in protein dimers (49) or the self-entanglement of folded proteins (66) without the need to define a closure operation. A strong correlation has been found between the value of G computed over open curves and its "closed counterpart." This indicates that Eq. 22 can be used as a descriptor for the topological state of entangled structures such as CLs. Once again, because we are interested only in a specific topological state, we have simplified the calculation of G in the same way as done for L. Therefore we have computed G by applying Eq. 22 to the polygonal curves defined by the M_t and M_l residues selected by structure reduction. In this case, however, we have adopted the convention that the bridge-forming cysteines are always the ends of the integration along the covalent loop. This way, G depends on the distance between the two cysteines being affected by the opening and closing of the covalent loop.

The cross product in Eq. 22 implies that G depend on the relative orientation of γ_1 and γ_2 curves. Therefore, one has to define an orientation along which the two subchains are integrated. In this work, we have not fixed any conventional orientation because we have not compared different molecules. However, we have computed G for an L_2 lasso structure, in which the tail pierces the loop twice in opposite directions (as shown in Fig. 1). In this case, the contribution to G provided by the threading in one direction is partially compensated (or entirely compensated, if the curves are closed) by the threading in the opposite direction. To adapt G such that it can detect this double piercing, we have separated the threading tail in two parts, assigning two different orientations for the calculation of Eq. 22. As a result, the contributions coming from the two piercings add up, detecting the L_2 state.

RESULTS

Homogeneous EFM

We first report the folding behavior of 2GMF described by an homogeneous EFM, in which the angular potentials (see Eq. 6 in Methods) are parameterized using homogeneous angular coefficients $k_i^{\mathrm{bend}} = k_b$ and $k_i^{\mathrm{tor}} = k_t$, where $k_b = 36.5$ and $k_t = 38.5$. From now on, we shall refer to this representation as the homogeneous model (HM). The order of magnitude of k_b and k_t is consistent with the settings used in (48), but the values were chosen equal to the average of the optimized bending and torsion coefficients presented in the following. This choice allowed us to assess the impact of the force-field heterogeneity introduced by the optimization procedure. Consistently with (48), we have studied the model at $T = 0.1$, which is below the melting point of the model and, as shown in the following, determines a quite frustrated free-energy landscape. An ensemble of 2048 folding trajectories has been collected in both reducing and oxidizing conditions. Equation 9 was used to model the bridge in an oxidizing environment.

To define the successfully folded trajectories, we monitored two variables, the root MSD (RMSD) $\mathcal{F}^{1/2}$ from the native state and the lasso variable L, indicating the formation of the CL topology (see the Methods for the definitions of $\mathcal{F}$ and L). We have selected two threshold values for $\mathcal{F}^{1/2}$ and L, considering the protein as fully folded only if both $\mathcal{F}^{1/2} < 0.9$ and $L > 0.9$. In most of the cases, the RMSD criterion was sufficient to classify the nativeness; however, the measurement of L has allowed for pointing out a few false positives and to distinguish successful folding trajectories with better accuracy. In the Supporting Materials and Methods, we report the comparison between this folding criterion and a more standard one based on the fraction of native contacts Q (68). In all the cases investigated in this manuscript, we have found the criterion employed here to be robust in determining the successfully folding trajectories, achieving a better accuracy than the Q criterion. Once the success criterion has been defined, the probability of folding was estimated as $P_f = n_f/n_{\mathrm{tot}}$, where n_f is the number of trajectories attaining the folding and $n_{\mathrm{tot}} = 2048$ is the total number of runs. This estimate of the success rate depends on the length τ_{run} of the simulated trajectories. Because the EFM focuses on the optimal pathways of folding, we aimed at observing those folding events that occur within the initial stages of the dynamics, not long after the collapse of the polymer chain. We have chosen $\tau_{\mathrm{run}} = 1.5 \times 10^4$, which, as shown in the following, is enough to capture all the fastest folding events, obtaining indications on the timescales of the slower processes as well.

The computed P_f of the HM in reducing conditions is equal to 55%, whereas in oxidizing conditions, the folded configuration is reached by 17% of the trajectories. This shows that the topological barrier introduced by the cysteine

bridge significantly increases the frustration of the model. We define the "folding landscape" as $F = -\log f$, where f is the frequency histogram of some chosen reaction variables (e.g., the RMSD) computed over the ensemble of trajectories. This quantity is sometimes named "nonequilibrium free-energy surface" (69,70). We also introduce the "successful folding landscape" $F_s = -\log f_s$, which considers only those trajectories that reach the native state.

In Fig. 4, *A* and *B*, the folding landscape of the HM in reducing and oxidizing conditions is reported as a function of the RMSD and of d_{b_1}, the distance of the two cysteine residues forming b_1. The corresponding F_s is instead shown in Fig. 4, *C* and *D*. By comparing the successful trajectories to the whole ensemble, we observed that the native basin is located in the region $\mathcal{F}^{1/2} \lesssim 0.9$. In both environmental conditions, the landscapes show a variety of metastable states, testifying to the roughness of the free-energy surface. Because, except from the bridge potential, EFM does not introduce native contacts, this roughness is the result of the topological bottlenecks encountered during the folding trajectories. In particular, if we consider only the successful trajectories in reducing conditions (Fig. 4 *C*), we observe a metastable state at RMSD ~2.0, presumably connected to the native basin by an open-bridge pathway, with $d_{b_1} \sim 4.0$. In oxidizing conditions (Fig. 4 *D*), this metastable state is perturbed by the action of the bridge potential, which restrains part of the trajectories close to its minimum, where the covalent loop is closed. Most of these trajectories remain trapped in this state and cannot overcome the topological barrier to reach the native basin.

To extract valuable information about the folding pathways, we have employed the lasso variable L and the Gauss' linking number G, defined in the Methods. Both L and G are useful to monitor the topological state of the protein along the trajectory, but because they exhibit a different behavior, we employ them for different purposes. Because L switches sharply from 0 to 1 when the native topology is attained, it is used to detect the time of formation of the lasso and, as mentioned before, to assess the folded state. G displays instead a smoother behavior; it is thus employed as reaction variable for computing the folding landscape, as shown in Fig. 5, where $F_s(G, d_{b_1})$ is reported. The plot confirms that in reducing conditions, the model establishes the lasso topology (attaining $G \gtrsim 1$) while the loop is open and that the metastable state preceding the folding can be identified with a populated region without lasso conformation ($G \sim 0$). In oxidizing conditions, the topological barrier is instead surpassed along two separate pathways, either with closed or open loop. We can classify the folding pathways as follows:

1) A "threading" mechanism, in which the contact between C88 and C121 is formed before the topology, and then the closed loop is threaded by the hairpin to reach the native basin.
2) A "bridge reopening" mechanism, in which, again, the covalent loop is closed before the lasso is formed. The topology is then attained in a second moment thanks to a wide fluctuation of the bridge distance and to the subsequent penetration of the loop by the hairpin.
3) An "open-loop" path, in which the lasso is formed before the contact between C88 and C121, with the loop that "wraps around" the hairpin to form the native state, a behavior that is reminiscent of the "embracement" mechanism defined in (47).

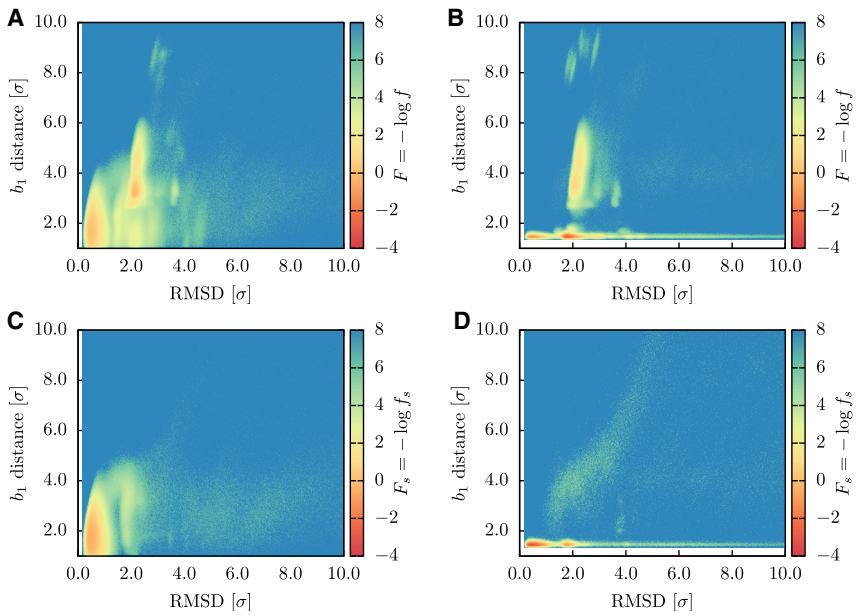

FIGURE 4 Folding landscapes F and F_s of the HM as a function of the RMSD from the native structure and of the b_1 bridge distance. (*A*) $N = 2048$ trajectories in reducing conditions. (*B*) $N = 2048$ trajectories in oxidizing conditions. (*C*) Successful trajectories ($N = 1133$) in reducing conditions. (*D*) Successful trajectories ($N = 350$) in oxidizing conditions. To see this figure in color, go online.

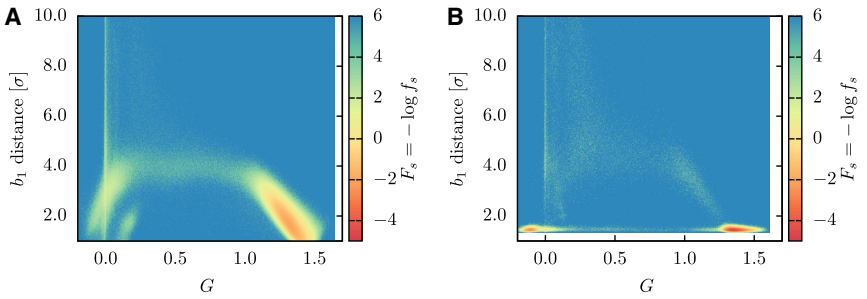

FIGURE 5 Successful folding landscape F_s of the HM as a function of the Gaussian linking number G and of the b_1 bridge distance. (*A*) Successful trajectories ($N = 1133$) in reducing conditions. (*B*) Successful trajectories ($N = 350$) in oxidizing conditions. To see this figure in color, go online.

A graphical illustration of these three processes is provided in Fig. 6. The successful trajectories can be classified according to these three pathways by performing a "kinematic" analysis that compares the timing of the main events in the folding process. For each trajectory, we thus computed three transition times: 1) the bridge formation time t_b, namely the first time at which C88 and C121 approach at a distance $d_{b_1} < 1.5\sigma_{b_1} = 1.992$; 2) the time of first topology formation t_k, when $L > 0.9$; and 3) the folding time t_f, which is when the protein first visits the native basin ($\mathcal{F}^{1/2} < 0.9$ and $L > 0.9$). We required that the conditions for 1), 2), and 3) remain valid for $\Delta t = 10$ for the transition to be completed. Then, by comparing the measured t_b and t_k with the time evolution of d_{b_1}, which signals the closure of the loop, and of L, which indicates the topological state, we could classify the folding routes traveled by the protein in successful simulations.

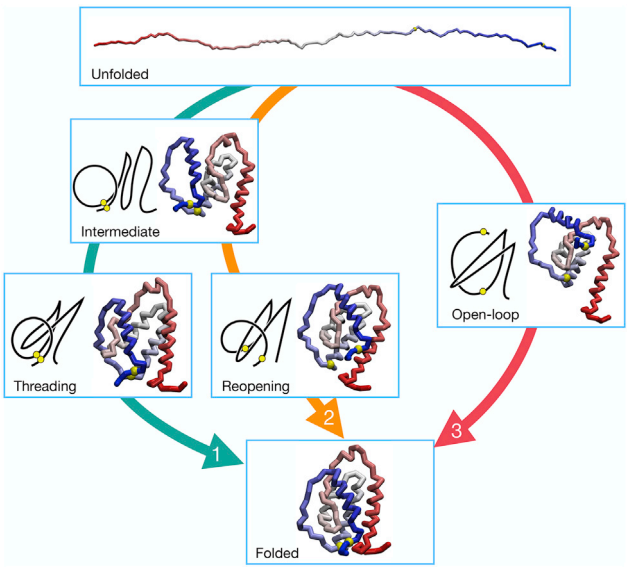

FIGURE 6 Illustration of the three folding pathways revealed by 2GMF simulations. Each box contains the snapshot of a representative configuration along the corresponding folding route (represented by a colored arrow, numbered according to the pathway definition in the text). For further clarity, intermediate configurations are provided with a schematic diagram of the structure. To see this figure in color, go online.

In Fig. 7, the bridge formation times t_b are plotted versus the folding times t_f for each successful HM trajectory. The mechanism associated to each trajectory is indicated by different colors. The fraction of trajectories undertaking different routes is reported in Table 1. The folding mechanisms are differently distributed in reducing and oxidizing simulations. In the first case, the successful folding events are similarly divided between open-loop and reopening pathways, whereas a relatively small number of threading trajectories are detected. Instead, in oxidizing conditions, the reopening is prevented by the action of the cysteine bridge potential, and, although the model mostly relies on the open-loop route, threading events are significant.

Another aspect that emerges from Fig. 7 concerns the folding timescales characterizing the different pathways. Most of the observed folding events occurred for $t < 10^3$, in particular those undergoing open-loop mechanism. In reducing conditions, the reopening events are distributed also beyond this timescale, whereas the few threading events were faster. This is somewhat counterintuitive because we expect that the entropic barrier of piercing the loop is larger when this is closed. If we look at the oxidized model results, we observe that threading events exhibit a bimodal time distribution; this suggests the existence of two possible threading pathways, a fast process taking place for $t < 10^3$ and a slower one that requires a timescale comparable to $\tau_{\text{run}} = 1.5 \times 10^4$. This bimodality disappears in reduced folding, in which the slow threadings are suppressed as the reopening of the bridge occurs over faster timescales. We underline that the defined folded basin allows fluctuations of d_b larger than 1.5σ. This explains the possibility of having $t_f < t_b$, which is evident in Fig. 7 and in the following Figs. 11 and 12. In these cases, a compact conformation that features the native lasso topology is attained, whereas the bridge contact occurs slightly later, after a diffusion phase within the native basin.

Overall, this analysis reveals the main features of the folding of 2GMF as described by the EFM and highlights the role of the topological barrier in selecting the accessible mechanisms to attain the native state. We stress the importance of the defined topological diagnostics, L and G, in clarifying the folding pathway scenario.

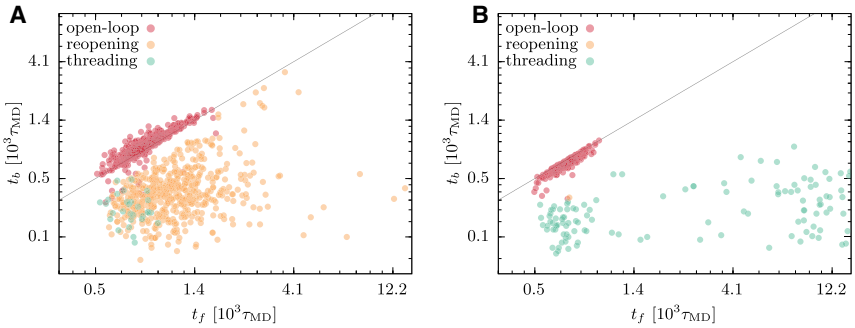

FIGURE 7 t_b versus t_f for the successful trajectories of the HM. (*A*) shows the results of the $N = 1133$ successful trajectories in reducing conditions, and (*B*) shows the results of the $N = 350$ successful trajectories in oxidizing conditions. The color of the circles indicates the folding pathway, following the classification indicated in the text. The line corresponds to $t_b = t_f$. To see this figure in color, go online.

Optimized EFM

In this section, we report the folding behavior of the EFM when optimized with MFFO, the evolutionary algorithm described in the Methods. Like most of the lasso structures, 2GMF is a secreted protein, and its folding occurs in the endoplasmic reticulum, which is an oxidizing environment. For this reason, the MFFO has been performed in oxidizing conditions.

The progress of the optimization procedure is displayed in Fig. 8, in which the evolution of the folding success rate is reported. We observe that as the MFFO introduces heterogeneity in the angular interactions, the rate increases significantly, reaching a value larger than 0.95. In Fig. 8, we also show how the folding success rate evolved when no crossover between different force-fields was operated. This represents the success rate resulting from 16 independent SFFO runs (namely the serial stochastic optimization algorithm of (48)). It is evident that the MFFO approach provides a remarkable boost to the optimization, attaining a strong folding reproducibility, before the independent SFFOs exhibit any significant improvement. This substantial advancement in the optimization strategy opens the possibility of employing the EFM for the study of larger proteins subject to even more complex folding processes.

After 30 MFFO cycles, we chose the top-ranked force-field and tested it over 2048 folding trajectories in both reducing and oxidizing conditions. We shall refer to this optimized model with the acronym OM. As described before, the bending and torsion stiffnesses of the HM have been set equal to the average values of the OM; this way, we could assess the impact of heterogeneity on the folding behavior.

The P_f values obtained for the OM are reported in Table 1. We notice how the OM reaches high probabilities in both reducing and oxidized folding, showing that the heterogeneity of angular forces can be crucial to achieve a nontrivial topology in a reproducible way, in agreement with what found for knotted protein folding in (70). We then investigated the successful folding landscape F_s associated to the OM, reported in Fig. 9 as a function of $\mathcal{F}^{1/2}$ and d_{b_1} and in Fig. 10 as a function of G and d_{b_1}. The landscapes look qualitatively different to those of Figs. 4 and 5, indicating that the OM selects different folding pathways with respect to HM. In particular, we can appreciate how the non-entangled intermediate state is now less populated and how the closure of the cysteine bridge mostly occurs as a late event.

To assess which folding pathways are more populated, we repeated the kinematic analysis operated for the previous model. The results, shown in Fig. 11, reveal that the bridge

TABLE 1 Probability of Folding P_f and Pathway Distribution

Model	Environment	P_f	$P_{\text{threading}}$	$P_{\text{reopening}}$	$P_{\text{open-loop}}$
HM	Red.	0.55	0.01	0.26	0.28
	Ox.	0.17	0.06	0.0	0.11
OM	Red.	0.96	–	0.01	0.95
	Ox.	0.95	0.05	–	0.90
GōM ($T = 0.7$)	Red.	0.60	0.12	0.01	0.47
	Ox.	0.55	0.11	0.01	0.44
GōM ($T = 1.1$)	Red.	0.63	0.11	0.30	0.23
	Ox.	0.27	0.24	0.01	0.02

Folding probability P_f and probability of undergoing different mechanisms ($P_{\text{threading}}$, $P_{\text{reopening}}$, and $P_{\text{open-loop}}$) for each of the considered models in reducing and oxidizing conditions. The probabilities are estimated as frequency of occurrence over 2048 trajectories of length $\tau_{\text{run}} = 1.5 \times 10^4$.

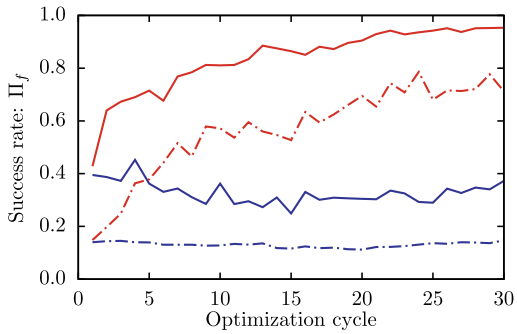

FIGURE 8 Average success rates of the folding trajectories performed during the optimization procedure. The reported quantity, Π_f, is a proxy of the folding probability, the definition of which is provided in the Methods. The red curves correspond to MFFO combining $N_K = 16$ force-fields, and the blue curves correspond to an MFFO without crossover of force-fields, equivalent to N_K parallel SFFOs. Solid lines indicate the success rate of the best-ranked force field, and dot-dashed line indicates the average rate of the N_K concurrent force fields. To see this figure in color, go online.

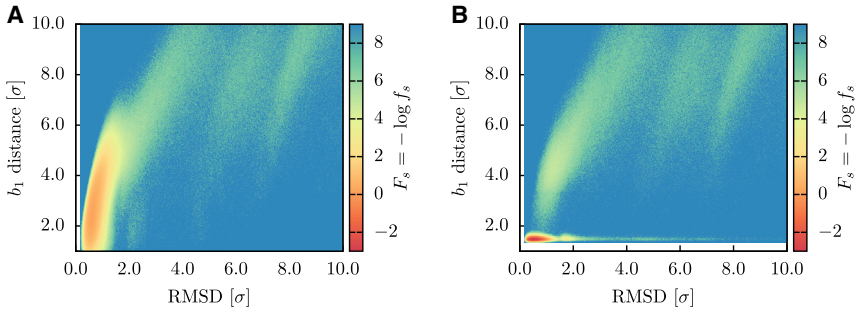

FIGURE 9 Successful folding landscape F_s of the OM as a function of the RMSD from the native structure and of the b_1 bridge distance. (*A*) Successful trajectories ($N = 1970$) in reducing conditions. (*B*) Successful trajectories ($N = 1946$) in oxidizing conditions. To see this figure in color, go online.

formation and folding times are on average slower than in the HM model. The optimization acted on the timescale of the folding events by delaying the closure of the loop. As a result, the open-loop folding mechanism is promoted and characterizes the great majority of the trajectories in both reducing and oxidizing conditions, as indicated in Table 1. In EFM, the open-loop folding turns out to be the optimal route to the formation of the native lasso fold, in agreement with the intuition that the closure of the covalent loop determines an entropic barrier, slowing down the process. The behavior of OM shows how the optimization pressure, building on the requirement of a reproducible and efficient folding, can select a pathway among the possible ones and polarize the mechanism of folding, similarly to what is observed in experiments and simulations of small, knotted protein folding (18,24).

GōM

To complement the picture obtained by means of the EFM, we have performed a set of folding simulations employing the well-established GōM proposed by Clementi et al. (26), which has already been used by Haglund et al. to study lasso proteins (20,22,23). For details on the description, we refer to the cited references; here, we just underline that the native contacts are established through a 12-10 Lennard-Jones potential, which is the main driving force of the folding. As mentioned in Methods, this description also models the backbone stiffness with the angular potentials of Eq. 6. The stiffness coefficients are homogeneous, set to $k^{\text{bend}} = 40.0$, $k_1^{\text{tor}} = 1.0$, and $k_3^{\text{tor}} = 0.5$. Following (71), we model the oxidizing conditions by rescaling the contact potential between the cysteines that form bridges in the native conformation. As a result of the temperature study presented in the Supporting Materials and Methods, we have chosen to simulate this model at a temperature $T = 0.7$, at which the folding is referred to as kinetically optimal or minimally frustrated (27,35).

The folding criterion adopted for this model is the one chosen for EFM, namely requiring that $\mathcal{F}^{1/2} < 0.9$ and $L > 0.9$ simultaneously. However, because the dihedral angles are substantially less stiff than in the EFM, the computation of L must involve a larger number of residues (see the Supporting Materials and Methods for further details). The folding success rates resulting from 2048 simulations in both reducing and oxidizing conditions are reported in Table 1. The measured probability is in both cases above 0.55, with a lower value in oxidized conditions. This similarity in folding propensity suggests that the topological barrier imposed by the formation of the bridge does not have a substantial effect in this model. This is possibly related to the fact that the native contact between the cysteines is present also in the model under reducing conditions, albeit weaker. However, the analysis of the folding pathways provides further indications to explain this similar capability of folding.

Applying the same criteria employed for the EFM, we have analyzed the successful trajectories collected with reduced and oxidized GōMs and assessed the population of different folding mechanisms. The results are reported in Table 1 and represented in the t_b versus t_f plots of Fig. 12, *A* and *B*. The data indicate that the distribution of

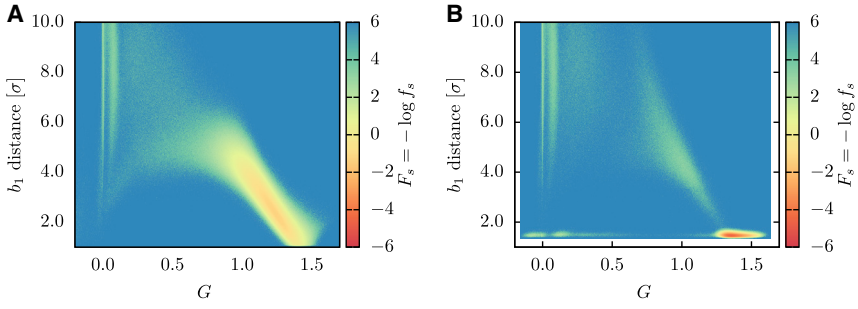

FIGURE 10 Successful folding landscape F_s of the HM as a function of the Gauss linking number G and of the b_1 bridge distance. (*A*) Successful trajectories ($N = 1970$) in reducing conditions. (*B*) Successful trajectories ($N = 1946$) in oxidizing conditions. To see this figure in color, go online.

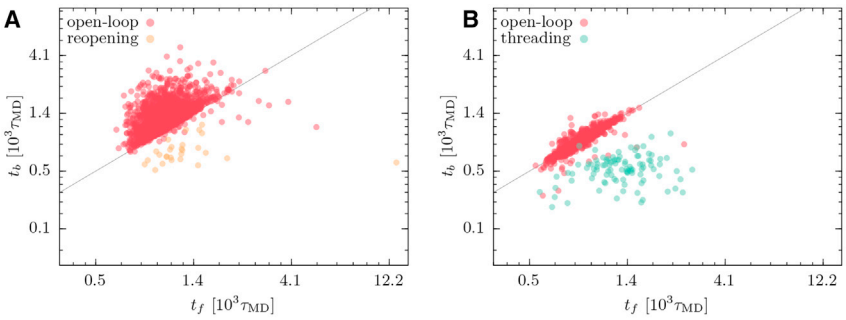

FIGURE 11 t_b versus t_f for the successful trajectories of the OM. (*A*) shows the results of the $N = 1969$ successful trajectories in reducing conditions, and (*B*) shows the results of the $N = 1946$ successful trajectories in oxidizing conditions. The color of the circles indicates the folding pathway, following the classification indicated in the text. The line corresponds to $t_b = t_f$. To see this figure in color, go online.

folding mechanisms is similar in reducing and oxidizing conditions. This symmetry confirms indeed that the successful folding events are not significantly affected by the cysteine bridge potential. However, most of the trajectories adopted an open-loop pathway, in which the topology forms before the contact of cysteine residues. The selection of this mechanism is the main reason why the model folds with a similar success rate in both environmental conditions. In the choice of the pathway, we have found that the GōM, at the temperature of fastest folding, is in qualitative agreement with the OM. Indeed, both descriptions show a rather symmetrical choice of pathway in reducing and oxidizing conditions, in which the open-loop mechanism is selected. This means that both models mostly fold by forming the CL topology before the closure of the covalent loop, pointing at this way of overcoming the topological bottleneck as the most efficient option for the protein.

Moreover, almost all folding events take place at early times ($t < 10^3$), whereas only a minor fraction of trajectories fold in the remaining simulation length. This indicates that the nonsuccessful runs have reached deep, metastable states and would need much longer times to find their way to the native basin. We thus notice that this Gō-like description of 2GMF is prone to kinetic traps, hampering the reproducible folding of the model. Backtracking is here a crucial factor in determining the access to the native state, but at this temperature, it would necessitate much longer timescales than those accessed by our simulations. The optimized EFM model could instead reach a very high probability of folding within the early stages of dynamics. This supports the idea that concerted, nonlocal motions of the backbone, like those driven by EFM angular potentials, are crucial for reproducible and efficient folding of self-entangled proteins. The adopted GōM, (almost) purely driven by native contacts, misses this aspect and thus fails in folding reproducibly.

To further enrich this picture, we show the behavior of the GōM when the folding temperature is increased, facilitating the backtracking mechanism. To this purpose, we have

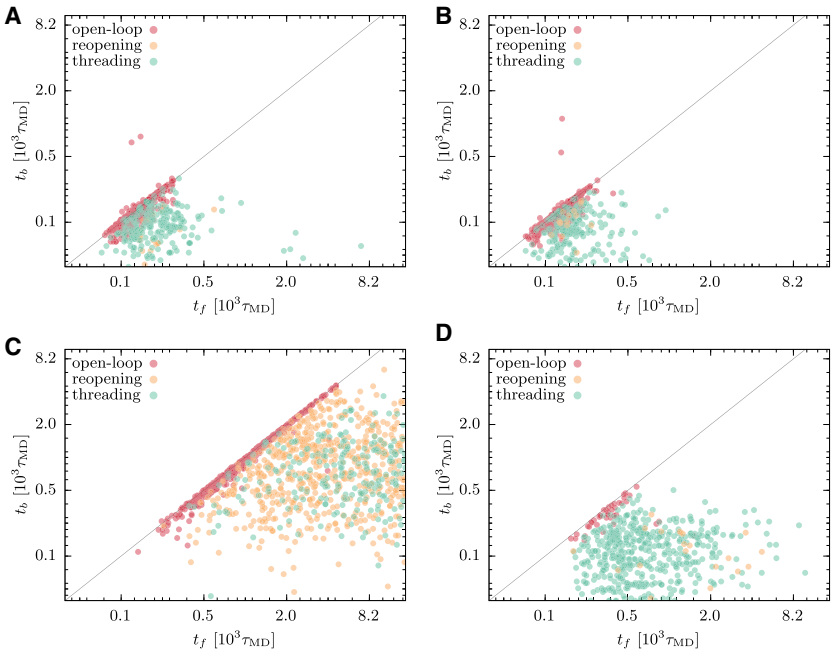

FIGURE 12 t_b versus t_f for the successful trajectories of GōM simulations. (*A*) and (*B*) display the results of the GōM at $T = 0.7$: the $N = 1228$ successful trajectories in reducing conditions are shown in (*A*), and the $N = 1130$ successful trajectories in oxidizing conditions are shown in (*B*). (*C*) and (*D*) display the results of the GōM at $T = 1.1$: the $N = 1291$ successful trajectories in reducing conditions are shown in (*C*), and the $N = 549$ successful trajectories in oxidizing conditions are shown in (*D*). The color of the circles indicates the folding pathway, following the classification indicated in the text. The line corresponds to $t_b = t_f$. To see this figure in color, go online.

studied the GōM at $T = 1.1$ in both reducing and oxidizing conditions. Again, we have collected 2048 runs of length $\tau_{run} = 1.5 \times 10^4$ to detect the fast folding events. As reported in Table 1, the probability of folding within this simulation time is now strongly affected by the environment, with a much lower success rate in oxidizing conditions. To investigate the reason for this difference, we have collected the distribution of folding times, once again distinguishing among the different pathways. The results, displayed in Fig. 12, show that the process at $T = 1.1$ is on average much slower than at $T = 0.7$ and that the population of folding routes is not any more symmetric between the reduced and oxidized models.

At $T = 1.1$, the model is not in the kinetically optimal regime, and slower routes that at $T = 0.7$ are prevented by the roughness of the free-energy surface are made accessible by thermal fluctuations, which allow backtracking and the exploration of the folding funnel across different pathways. This aspect is evident from the behavior of the model in reducing conditions (Fig. 12 *C*), in which all three mechanisms are well populated and the incidence of slower pathways is limited only by the finite sampling time of the trajectories. In the oxidized model (Fig. 12 *D*), the situation is different because the cysteine bridge potential anticipates the closure of the loop, narrowing the conformational space accessible by thermal fluctuations and polarizing the choice of folding mechanism toward the threading pathway. As in the OM results, in this last case (oxidized GōM at $T = 1.1$) as well, a single folding route is promoted. However, the pathway selection has here a different nature than in EFM results. Although in this case, it is the early closure of the bridge that imposes the folding mechanism, in the OM, the choice was determined by the optimality of folding kinetics. It would be therefore of great interest to verify the preferential folding pathway of 2GMF by means of more detailed all-atom MD simulations or with experimental probing. This kind of evidence, on the basis of the results presented here, would indeed provide insights on the nature of folding mechanism selection that is a characterizing feature of self-entangled proteins.

CONCLUSIONS

In this work, we have presented an investigation on the folding of the glycoprotein granulocyte-macrophage colony-stimulating factor (PDB: 2GMF), which presents a CL native structure. The study is performed by means of MD simulations, employing both a widely used GōM variant, proposed by Clementi et al. (26), and the EFM, a CG, minimalistic description proposed in (48), for investigating the folding mechanisms of knotted proteins. We here extended the original models by implementing the formation of native cysteine bridges to assess their effect on the folding process.

The EFM dynamics is based on optimized bending and dihedral potentials, which are tuned to improve the folding capability of the model, with the purpose of enlightening the optimal pathways toward the native structure. In this work, we have introduced the MFFO, an evolutionary approach for the optimization of EFM interaction potentials. The results show that this algorithm significantly outperforms the original stochastic method, allowing the study of more complex systems with EFM. Moreover, this evolutionary strategy is general and can be employed to optimize other minimalistic protein descriptions, such as Gō-like models. Relying on this evolutionary approach, we have built an OM of 2GMF, capable of reaching a very high success rate during the early stages of the folding, avoiding kinetic traps, and providing indications on the pathways that enable efficient and reproducible folding. We have then compared the behavior of this model to the results obtained with the GōM.

In our study, we focused on the capability of folding in relatively short times, that is, without encountering major kinetic traps. The optimized EFM is in this sense more successful, attaining a folding probability of 0.95 against the 0.6 achieved by the considered GōM at the temperature of fastest folding. This demonstrates the importance of force-field heterogeneity and concerted angular motions for efficiently crossing the topological bottlenecks of self-entangled folding.

Besides the capability of reaching the native state, we were also interested in studying the folding pathways of the protein. To this purpose, we have defined two topological descriptors, the lasso variable L and the Gauss linking number G, inspired by successful methodologies for the classification of protein structures. By monitoring the topology of the protein, these variables turned out to be useful tools for the analysis and classification of folding trajectories. As a result, we were able to characterize the folding scenario of 2GMF, outlining three main mechanisms. Building on this picture, we showed that the optimization of EFM can polarize the trajectories toward an open-loop folding route, in which the lasso topology sets in before the cysteine bridge is formed and seals the covalent loop. The selection of this optimal pathway is also confirmed by the GōM that, at the temperature of fastest folding, privileges an open-loop folding route.

By simulating the GōM at a higher temperature to lower the free-energy barriers and allow for backtracking mechanism, we have found that the scenario of folding pathways changes. These temperature conditions fall outside the range of optimal folding kinetics, and the process requires much longer simulation times. Nonetheless, because native contacts can break more easily, the protein can sample a larger portion of the free-energy landscape, populating all possible folding routes. Also, at this temperature regime, under oxidizing conditions, we have observed a polarization of the folding pathway toward a single mechanism. However, differently from the optimal kinetic scenario in which the open-loop mechanism was privileged, these simulations favor a loop-threading route. Indeed, the early formation

of the covalent loop imposes an entropic restraint to the model, restricting the possible routes to the threading one. Starting from this picture, we think that the study of 2GMF folding using further techniques, either more detailed simulations or experimental studies, would be crucial to validate the hypothesis that entangled folding has evolved to privilege optimal pathways. Overall, this discussion can provide a useful viewpoint in the debate on protein folding mechanisms, and their driving principles (see, e.g., (72–74)).

The methodological advancements presented here constitute a useful complement to the existing protein models. They can provide valuable insights on the folding landscape of topologically complex proteins and draw the guidelines for molecular simulations using more detailed physical models. Moreover, by highlighting the most efficient folding routes, the qualitative picture obtained with the EFM can also shed light on the role played by environmental factors that accelerate folding, such as chaperonins or cotranslational folding.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2019.05.025.

AUTHOR CONTRIBUTIONS

C.P. and R.P. designed the research and developed the methodologies. C.P. carried out all simulations and analyzed the data. C.P. and R.P. wrote the article.

ACKNOWLEDGMENTS

The authors thank O. Valsson and Y. Zhao for a critical reading of the manuscript.

The authors acknowledge funding from the European Union's Horizon 2020 research and innovation program under the GOKNOT Marie Skłodowska-Curie Grant Agreement No. 796969 and the contribution of the COST Action CA17139. Computational resources were provided by The Max Planck Computing and Data Facility.

REFERENCES

1. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The protein data bank. *Nucleic Acids Res.* 28:235–242.
2. Dabrowski-Tumanski, P., and J. I. Sułkowska. 2017. To tie or not to tie? That is the question. *Polymers (Basel).* 9:454.
3. Jamroz, M., W. Niemyska, …, J. I. Sułkowska. 2015. KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res.* 43:D306–D314.
4. Mansfield, M. L. 1994. Are there knots in proteins? *Nat. Struct. Mol. Biol.* 1:213–214.
5. Jackson, S. E., A. Suma, and C. Micheletti. 2017. How to fold intricately: using theory and experiments to unravel the properties of knotted proteins. *Curr. Opin. Struct. Biol.* 42:6–14.
6. Faísca, P. F. 2015. Knotted proteins: a tangled tale of structural biology. *Comput. Struct. Biotechnol. J.* 13:459–468.
7. Lim, N. C., and S. E. Jackson. 2015. Molecular knots in biology and chemistry. *J. Phys. Condens. Matter.* 27:354101.
8. Lua, R. C. 2012. PyKnot: a PyMOL tool for the discovery and analysis of knots in proteins. *Bioinformatics.* 28:2069–2071.
9. Wüst, T., D. Reith, and P. Virnau. 2015. Sequence determines degree of knottedness in a coarse-grained protein model. *Phys. Rev. Lett.* 114:028102.
10. Potestio, R., C. Micheletti, and H. Orland. 2010. Knotted versus unknotted proteins: evidence of knot-promoting loops. *PLoS Comput. Biol.* 6:e1000864.
11. Sułkowska, J. I., E. J. Rawdon, …, A. Stasiak. 2012. Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Natl. Acad. Sci. USA.* 109:E1715–E1723.
12. Sułkowska, J. I., P. Sułkowski, …, M. Cieplak. 2008. Stabilizing effect of knots on proteins. *Proc. Natl. Acad. Sci. USA.* 105:19714–19719.
13. Christian, T., R. Sakaguchi, …, Y. M. Hou. 2016. Methyl transfer by substrate signaling from a knotted protein fold. *Nat. Struct. Mol. Biol.* 23:941–948.
14. Dabrowski-Tumanski, P., A. Stasiak, and J. I. Sułkowska. 2016. In search of functional advantages of knots in proteins. *PLoS One.* 11:e0165986.
15. Mallam, A. L., and S. E. Jackson. 2011. Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins. *Nat. Chem. Biol.* 8:147–153.
16. King, N. P., A. W. Jacobitz, …, T. O. Yeates. 2010. Structure and folding of a designed knotted protein. *Proc. Natl. Acad. Sci. USA.* 107:20732–20737.
17. Wang, I., S. Y. Chen, and S. T. Hsu. 2015. Unraveling the folding mechanism of the smallest knotted protein, MJ0366. *J. Phys. Chem. B.* 119:4359–4370.
18. Lim, N. C., and S. E. Jackson. 2015. Mechanistic insights into the folding of knotted proteins in vitro and in vivo. *J. Mol. Biol.* 427:248–258.
19. Niemyska, W., P. Dabrowski-Tumanski, …, J. I. Sułkowska. 2016. Complex lasso: new entangled motifs in proteins. *Sci. Rep.* 6:36895.
20. Haglund, E., J. I. Sułkowska, …, J. N. Onuchic. 2012. The unique cysteine knot regulates the pleotropic hormone leptin. *PLoS One.* 7:e45464.
21. Dabrowski-Tumanski, P., W. Niemyska, …, J. I. Sułkowska. 2016. LassoProt: server to analyze biopolymers with lassos. *Nucleic Acids Res.* 44:W383–W389.
22. Haglund, E., J. I. Sułkowska, …, P. A. Jennings. 2014. Pierced Lasso Bundles are a new class of knot-like motifs. *PLoS Comput. Biol.* 10:e1003613.
23. Haglund, E., A. Pilko, …, J. N. Onuchic. 2017. Pierced lasso topology controls function in leptin. *J. Phys. Chem. B.* 121:706–718.
24. a Beccara, S., T. Škrbić, …, P. Faccioli. 2013. Folding pathways of a knotted protein with a realistic atomistic force field. *PLoS Comput. Biol.* 9:e1003002.
25. Noel, J. K., J. N. Onuchic, and J. I. Sułkowska. 2013. Knotting a protein in explicit solvent. *J. Phys. Chem. Lett.* 4:3570–3573.
26. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.
27. Cieplak, M., and T. X. Hoang. 2003. Universality classes in folding times of proteins. *Biophys. J.* 84:475–488.
28. Go, N., and H. Taketomi. 1978. Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. USA.* 75:559–563.
29. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.

30. Best, R. B., G. Hummer, and W. A. Eaton. 2013. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. USA.* 110:17874–17879.
31. Hoang, T. X., and M. Cieplak. 2000. Sequencing of folding events in go-type proteins. *J. Chem. Phys.* 113:8319–8328.
32. Shea, J. E., J. N. Onuchic, and C. L. Brooks, III. 1999. Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. USA.* 96:12512–12517.
33. Shea, J. E., J. N. Onuchic, and C. L. Brooks, III. 2002. Probing the folding free energy landscape of the Src-SH3 protein domain. *Proc. Natl. Acad. Sci. USA.* 99:16064–16068.
34. Sułkowska, J. I., and M. Cieplak. 2007. Mechanical stretching of proteins-a theoretical survey of the Protein Data Bank. *J. Phys. Condens. Matter.* 19:283201.
35. Sułkowska, J. I., and M. Cieplak. 2008. Selection of optimal variants of Gō-like models of proteins through studies of stretching. *Biophys. J.* 95:3174–3191.
36. Whitford, P. C., J. K. Noel, …, J. N. Onuchic. 2009. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins.* 75:430–441.
37. Sułkowska, J. I., P. Sułkowski, and J. Onuchic. 2009. Dodging the crisis of folding proteins with knots. *Proc. Natl. Acad. Sci. USA.* 106:3119–3124.
38. Noel, J. K., J. I. Sułkowska, and J. N. Onuchic. 2010. Slipknotting upon native-like loop formation in a trefoil knot protein. *Proc. Natl. Acad. Sci. USA.* 107:15403–15408.
39. Sułkowska, J. I., J. K. Noel, and J. N. Onuchic. 2012. Energy landscape of knotted protein folding. *Proc. Natl. Acad. Sci. USA.* 109:17783–17788.
40. Chwastyk, M., and M. Cieplak. 2015. Cotranslational folding of deeply knotted proteins. *J. Phys. Condens. Matter.* 27:354105.
41. Zhao, Y., P. Dabrowski-Tumanski, …, J. I. Sułkowska. 2018. The exclusive effects of chaperonin on the behavior of proteins with 52 knot. *PLoS Comput. Biol.* 14:e1005970.
42. Wallin, S., K. B. Zeldovich, and E. I. Shakhnovich. 2007. The folding mechanics of a knotted protein. *J. Mol. Biol.* 368:884–893.
43. Škrbic, T., C. Micheletti, and P. Faccioli. 2012. The role of non-native interactions in the folding of knotted proteins. *PLoS Comput. Biol.* 8:1–12.
44. Soler, M. A., A. Nunes, and P. F. Faísca. 2014. Effects of knot type in the folding of topologically complex lattice proteins. *J. Chem. Phys.* 141:025101.
45. Dabrowski-Tumanski, P., A. I. Jarmolinska, and J. I. Sułkowska. 2015. Prediction of the optimal set of contacts to fold the smallest knotted protein. *J. Phys. Condens. Matter.* 27:354109.
46. Covino, R., T. Skrbić, …, C. Micheletti. 2013. The role of non-native interactions in the folding of knotted proteins: insights from molecular dynamics simulations. *Biomolecules.* 4:1–19.
47. Chwastyk, M., and M. Cieplak. 2015. Multiple folding pathways of proteins with shallow knots and co-translational folding. *J. Chem. Phys.* 143:045101.
48. Najafi, S., and R. Potestio. 2015. Folding of small knotted proteins: insights from a mean field coarse-grained model. *J. Chem. Phys.* 143:243121.
49. Baiesi, M., E. Orlandini, …, F. Seno. 2016. Linking in domain-swapped protein dimers. *Sci. Rep.* 6:33872.
50. Rozwarski, D. A., K. Diederichs, …, P. A. Karplus. 1996. Refined crystal structure and mutagenesis of human granulocyte-macrophage colony-stimulating factor. *Proteins.* 26:304–313.
51. Weeks, J. D., D. Chandler, and H. C. Andersen. 1971. Role of repulsive forces in forming the equilibrium structure of simple liquids. *J. Chem. Phys.* 54:5237–5247.
52. Grest, G. S., and K. Kremer. 1986. Molecular dynamics simulations for polymers in the presence of a heat bath. *Phys. Rev. A.* 33:3628–3631.
53. Kwiecińska, J. I., and M. Cieplak. 2005. Chirality and proteins folding. *J. Phys. Condens. Matter.* 17:S1565–S1580.
54. Huang, C., X. Yang, and Z. He. 2010. Protein folding simulations of 2D HP model by the genetic algorithm based on optimal secondary structures. *Comput. Biol. Chem.* 34:137–142.
55. Noel, J. K., M. Levi, …, P. C. Whitford. 2016. SMOG 2: a versatile software package for generating structure-based models. *PLoS Comput. Biol.* 12:e1004794.
56. Noel, J. K., P. C. Whitford, and J. N. Onuchic. 2012. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B.* 116:8692–8702.
57. Van Der Spoel, D., E. Lindahl, …, H. J. Berendsen. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.
58. Abraham, M. J., T. Murtola, …, E. Lindahl. 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX.* 1–2:19–25.
59. Micheletti, C., D. Marenduzzo, and E. Orlandini. 2011. Polymers with spatial or topological constraints: theoretical and computational results. *Phys. Rep.* 504:1–73.
60. Millett, K., A. Dobay, and A. Stasiak. 2005. Linear random knots and their scaling behavior. *Macromolecules.* 38:601–606.
61. Tubiana, L., E. Orlandini, and C. Micheletti. 2011. Probing the entanglement and locating knots in ring polymers: a comparative study of different arc closure schemes. *Prog. Theor. Phys.* 191 (*Suppl*):192–204.
62. Kolesov, G., P. Virnau, …, L. A. Mirny. 2007. Protein knot server: detection of knots in protein structures. *Nucleic Acids Res.* 35:W425–W428.
63. Wu, F. Y. 1992. Knot theory and statistical mechanics. *Rev. Mod. Phys.* 64:1099–1131.
64. Cromwell, P. R. 2004. Knots and Links. Cambridge University Press, Cambridge, UK.
65. Koniaris, K., and M. Muthukumar. 1991. Self-entanglement in ring polymers. *J. Chem. Phys.* 95:2873–2881.
66. Baiesi, M., E. Orlandini, …, A. Trovato. 2017. Exploring the correlation between the folding rates of proteins and the entanglement of their native states. *J. Phys. A Math. Theor.* 50:504001.
67. Panagiotou, E., M. Kröger, and K. C. Millett. 2013. Writhe and mutual entanglement combine to give the entanglement length. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 88:062604.
68. Wołek, K., and M. Cieplak. 2016. Criteria for folding in structure-based models of proteins. *J. Chem. Phys.* 144:185102.
69. Zhang, Y., J. Zhang, and W. Wang. 2011. Atomistic analysis of pseudoknotted RNA unfolding. *J. Am. Chem. Soc.* 133:6882–6885.
70. Li, W., T. Terakawa, …, S. Takada. 2012. Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. *Proc. Natl. Acad. Sci. USA.* 109:17789–17794.
71. Zhao, Y., and M. Cieplak. 2018. Stability of structurally entangled protein dimers. *Proteins.* 86:945–955.
72. Baldwin, R. L. 2017. Clash between energy landscape theory and foldon-dependent protein folding. *Proc. Natl. Acad. Sci. USA.* 114:8442–8443.
73. Eaton, W. A., and P. G. Wolynes. 2017. Theory, simulations, and experiments show that proteins fold by multiple pathways. *Proc. Natl. Acad. Sci. USA.* 114:E9759–E9760.
74. Englander, S. W., and L. Mayne. 2017. Reply to Eaton and Wolynes: how do proteins fold? *Proc. Natl. Acad. Sci. USA.* 114:E9761–E9762.
75. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28.