Searching the optimal folding routes of a Complex Lasso protein

Claudio Perego[1*], Raffaello Potestio[2]

1 Polymer Theory Department, Max Planck Institute for Polymer Research, Mainz, Germany
2 Physics Department, University of Trento, Trento, Italy

* perego@mpip-mainz.mpg.de

Abstract

Understanding how polypeptides can efficiently and reproducibly attain a self-entangled conformation is a compelling biophysical challenge, which 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, as 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 the present 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.

Author summary

We have investigated, by means of numerical methods, the folding mechanism of Granulocyte-macrophage colony-stimulating factor, a glycoprotein that handles a variety of functions in the human body. Our interest in this protein focuses on the self-entangled native state, which is classified as a so-called complex lasso. Complex lasso structures contain a backbone loop, closed by a cysteine bridge, which is pierced one or more times by the protein chain, resulting in an entangled conformation. Understanding how a polypeptide can encode into its sequence the capability of tying itself into such kind of structures would represent a major advancement in the comprehension of a crucial biological process such as protein folding.

To study this folding mechanism we have employed molecular dynamics simulations, adopting both a well-known minimalistic model of the protein, and an alternative model, that was specifically proposed for unveiling 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 the 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 the present work we address the following question: how can the amino acid chain fold reproducibly and efficiently into a specific, nontrivial topology? Many experiments were conducted to answer this question, showing e.g. that these proteins can spontaneously tie themselves to 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 constituted 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 non-trivial 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 about 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, since the cysteine bridge is stable in an oxidizing solution, while 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, electing CLs as 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 (GōM) (see e.g. [26, 27]), named after the pioneering work of Gō [28]. In GōMs the protein is described as an hetero-polymer chain that encodes its native fold in the interaction potential. This kind of description stems out from the established Energy Landscape Theory (ELT), according to which proteins have evolved to fold along a smooth, funneled free energy landscape. Such "folding funnel" determines the efficient and reproducible collapse of the denatured polymer chain to its compact and functional 3D 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 while 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], also in the framework of lasso folding [20, 22, 23]. However, the underlying theory clashes with the presence of knots in proteins, as the formation of entanglements implies a high degree of coordination at different length scales which can hardly be encoded in native contact potentials. For example, the mandatory passage through the *specific, non-alternative* 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 non-local moves required for the tying of the backbone [46]. Moreover, in agreement with ELT, 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 Ref. [48] an alternative CG description for the study of knotted folded proteins has been proposed. This model, dubbed 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 Ref. [24].

In the present 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 a newly implemented 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 to monitor 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.

Results

System Structure

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 an helical cytokine formed by 127 residues, of which 121 are resolved in the PDB structure, shown in Fig. 1. As highlighted in 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 L_2 lasso structure, where 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.

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 Methods section, 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, both in reducing and in oxidizing conditions. As discussed in the Methods section, the stability of cysteine bridges in oxidizing environment is modeled in the EFM by means of native contact potentials between the cysteine pairs, and in the GōM by rescaling the existent native contacts. We shall employ natural units, indicating energies in units of ϵ, temperatures in units of ϵ/k_{B}, lengths in units of σ, and time-lengths in units of $\tau_{MD} = \sigma\sqrt{m/\epsilon}$, m being the bead mass.

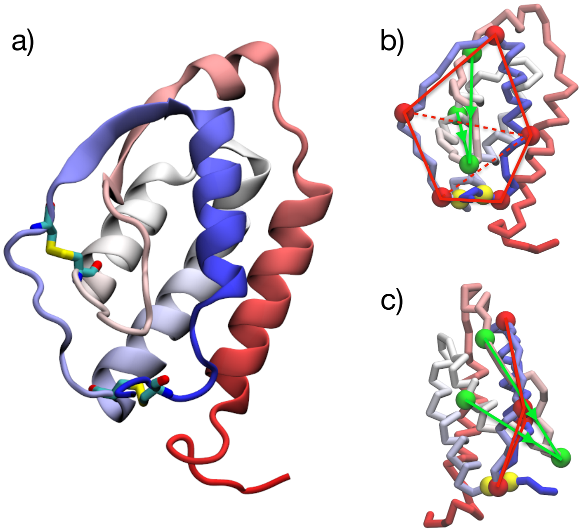


Fig 1. 2GMF Protein structure and geometry of topological descriptors. Panel a): Cartoon representation of Chain A of 2GMF in its native fold. The cysteine bridges are shown with atomistic resolution. Panel b): view of 2GMF native structure, showing only the C_α residues. Cysteines 88 and 121, forming b_1, are represented as yellow beads. The structure reduction employed for the definition of the topological variables L and G (see Methods section) is also displayed: the 5 residues chosen to represent the loop are highlighted as red circles connected by red lines, while the 3 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 3 triangles for computing L. The green arrows represent the integration verse along the hairpin segments used for the calculation of G. Panel c): same as b) but rotated. The coloring of backbone residues depends on their index along the chain, going from red (N-terminal) to blue (C-terminal).

Homogeneous EFM

We first report the folding behavior of 2GMF described by an homogeneous EFM, in which the angular potentials (see Eq. 1 in Sec. Methods) are parametrized 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 Homogeneous Model (HM). The order of magnitude of k_b and k_t is consistent with the settings used in Ref. [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 Ref. [48], we have studied the model at $T = 0.1$, that 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, both in reducing and oxidizing conditions. Eq. 9 was used to model the bridge in oxidizing environment.

In order to define the folding criterion we monitored two variables, the Root Mean Square Displacement (RMSD) $\mathcal{F}^{1/2}$ from the native state, and the lasso variable L, indicating the formation of the CL topology (see the Methods section 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 to point out few false positives, and to distinguish successful folding trajectories with better accuracy. Once the success criterion has been defined, the probability of folding was estimated as $P_f = n_{\mathrm{f}}/n_{\mathrm{tot}}$ 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. Since 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 HM in reducing conditions is equal to 55%, while in oxidizing conditions the folded configuration is reached by the 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 “non-equilibrium free-energy surface” [51, 52]. We also introduce the “successful folding landscape” $F_s = -\log f_s$, which considers only those trajectories that reach the native state.

In Figs. 2A 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 Figs. 2C 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 the roughness of the free-energy surface. Since, 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. 2C), we observe a metastable state at RMSD~ 2.0, presumably connected to the native basin by an open-bridge pathwhay, with $d_{b_1} \sim 4.0$. In oxidizing conditions (Fig. 2D) 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.

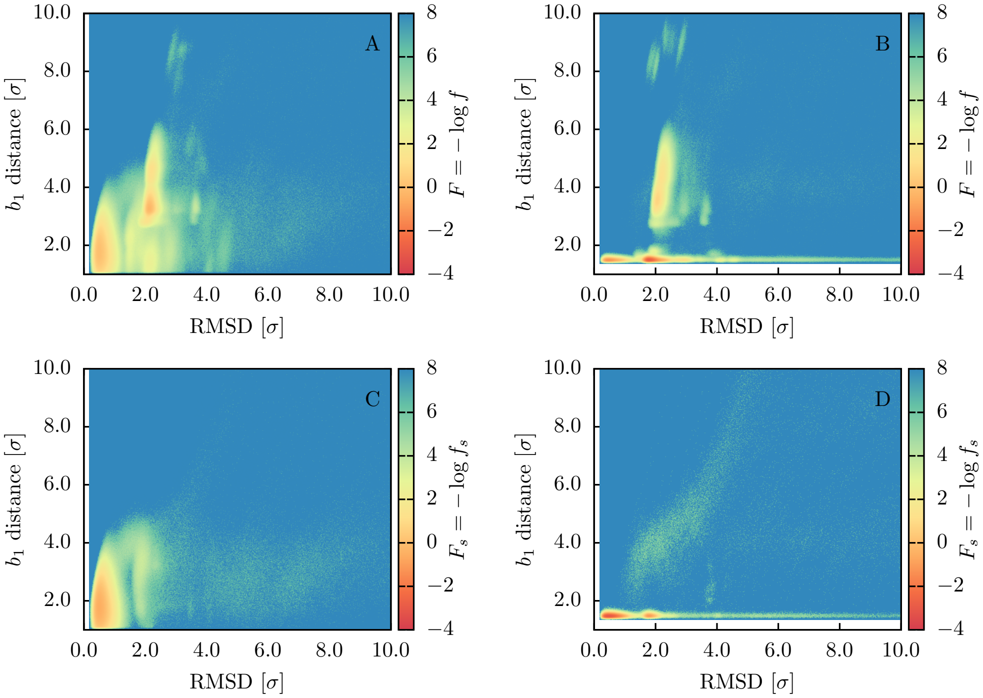


Fig 2. Folding landscapes F and F_s of the HM as a function of the RMSD from the native structure and of the b_1 bridge distance. A panel: $N = 2048$ trajectories in reducing conditions. B panel: $N = 2048$ trajectories in oxidizing conditions. C panel: successful trajectories ($N = 1133$) in reducing conditions. D panel: successful trajectories ($N = 350$) in oxidizing conditions.

In order to extract valuable information about the folding pathways, we have defined two topological descriptors: the aforementioned lasso variable L, and the Gauss' linking number G, which monitors the topology of the backbone by quantifying the intertwining between the covalent loop and the threading hairpin sections of the chain. The definition and implementation of these descriptors is reported in the Methods section. Both L and G are useful to monitor the topological state of the protein along the trajectory but, since they exhibit a different behavior, we employ them for different purposes. Since 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. 3, 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, 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

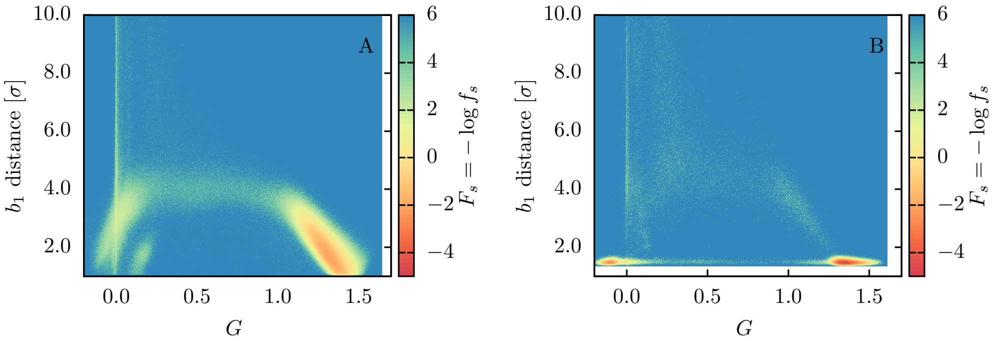


Fig 3. Successful Folding landscape F_s of the HM as a function of the Gaussian linking number G and of the b_1 bridge distance. A panel: successful trajectories ($N = 1133$) in reducing conditions. B panel: successful trajectories ($N = 350$) in oxidizing conditions.

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 graphical illustration of these three processes is provided in Fig. 4. 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: a) the bridge formation time t_b, namely the first time at which C88 and C121 approach at a distance $d_{b_1} < 1.5\sigma_{b_1} = 1.992$, b) the time of first topology formation t_k, when $L > 0.9$, and c) the folding time t_f, that 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 a), b) and c) remain valid for $\Delta t = 10$ for the transition to be completed. Then, by comparing the measured t_b, t_k and t_f 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.

In Fig. 5 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 Tab. 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 pathway, while a relatively small number of threading trajectories is detected. Instead, in oxidizing conditions the reopening is prevented by the action of the cysteine bridge potential and, while the model mostly relies on the open-loop route, threading events are significant.

Another aspect that emerges from Fig. 5 concerns the folding time-scales 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 re-opening events are distributed also beyond this time-scale, while the few threading events were faster. This is somehow counter-intuitive, as 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

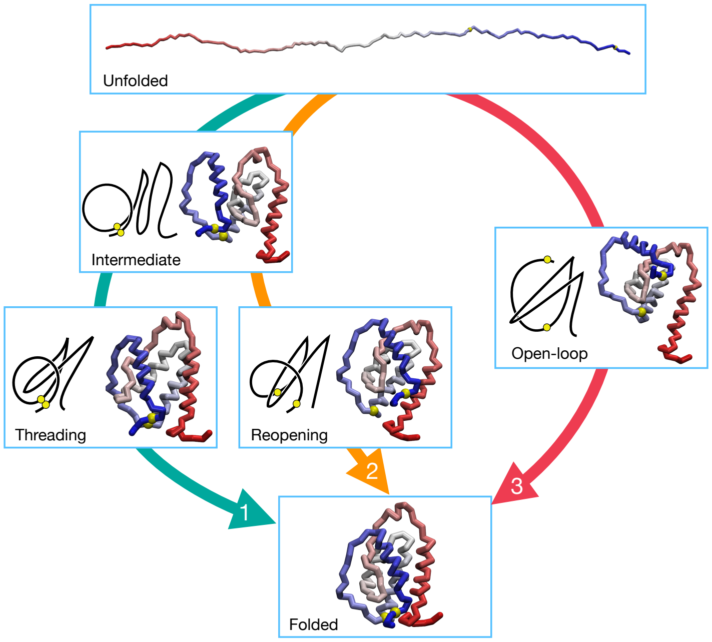


Fig 4. Illustration of the three folding pathways revealed by 2GMF EFM simulations. Each box contains the snapshot of a representative configuration along the corresponding folding route (represented by a colored arrow, numbered according to the pathway definition in the text). For further clarity, intermediate configurations are provided with a schematic diagram of the structure.

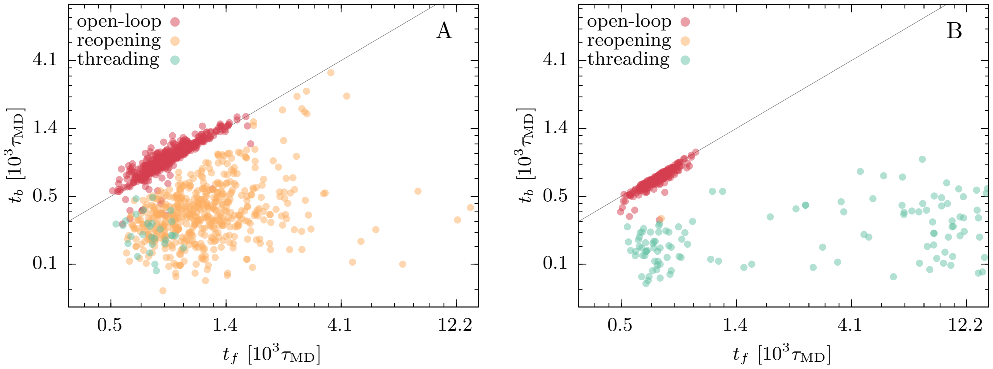


Fig 5. t_b **versus** t_f **for the successful trajectories of the HM.** Panel A shows the results of the $N = 1133$ successful trajectories in reducing conditions and panel B shows the results of the $N = 350$ successful trajectories in oxidizing conditions. The color of the circles indicates the folding pathway, following the classification indicated in the text. The black line corresponds to $t_b = t_f$.

threadings are suppressed as the reopening of the bridge occurs over faster timescales.

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 underline here the importance of the defined topological diagnostics, L and G, in clarifying the folding pathway scenario.

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 $\tau_{\text{run}} = 1.5 \times 10^4$.

Optimized EFM

In this section we report the folding behavior of the EFM when optimized with MFFO, the evolutionary algorithm introduced in Sec. Methods. As most of the lasso structures, 2GMF is a secreted protein, and its folding occurs in the endoplasmic reticulum, that is an oxidizing environment. For this reason the MFFO has been performed in oxidizing conditions. The settings and parameters of the optimization procedure are reported in the Methods section.

The progress of the optimization procedure is displayed in Fig. 6, where 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. 6 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 Ref. [48], outlined in the Methods section). 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, undergoing even more complex folding processes.

After 30 MFFO cycles we chose the top ranked force-field and tested it over 2048 folding trajectories, both in 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's obtained for the OM are reported in Tab. 1. We notice how the OM reaches high probabilities both in 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 Ref. [52]. We then investigated the successful folding landscape F_s associated to the OM, reported in Fig. 7 as a function of $\mathcal{F}^{1/2}$ and d_{b_1}, and in Fig. 8 as a function of G and d_{b_1}. The landscapes look qualitatively different to those of Figs. 2 and 3, 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. 9, reveal that the bridge formation and folding times are on average slower than in the HM model. The

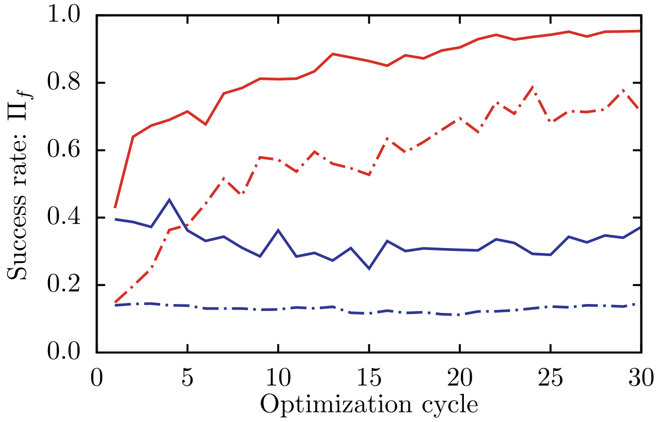


Fig 6. Average success rates of the folding trajectories performed during the optimization procedure. The reported quantity, Π_f, is a proxy of the folding probability, the definition of which is provided in the Methods section. The red curves correspond to MFFO combining $N_K = 16$ force-fields, the 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, while dot-dashed line indicates the average rate of the N_K concurrent force-fields.

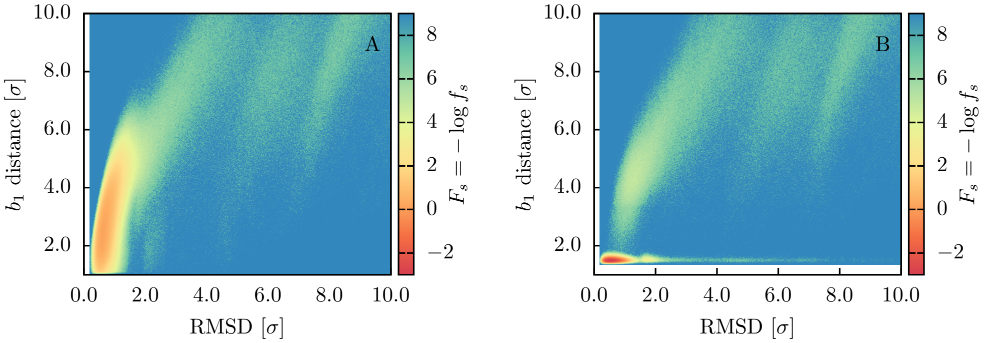


Fig 7. Successful Folding landscape F_s of the OM as a function of the RMSD from the native structure and of the b_1 bridge distance. A panel: successful trajectories ($N = 1970$) in reducing conditions. B panel: successful trajectories ($N = 1946$) in oxidizing conditions.

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, as indicated in Tab. 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].

Gō Model

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], that has already been used by Haglund et al. to study lasso proteins [20, 22, 23].

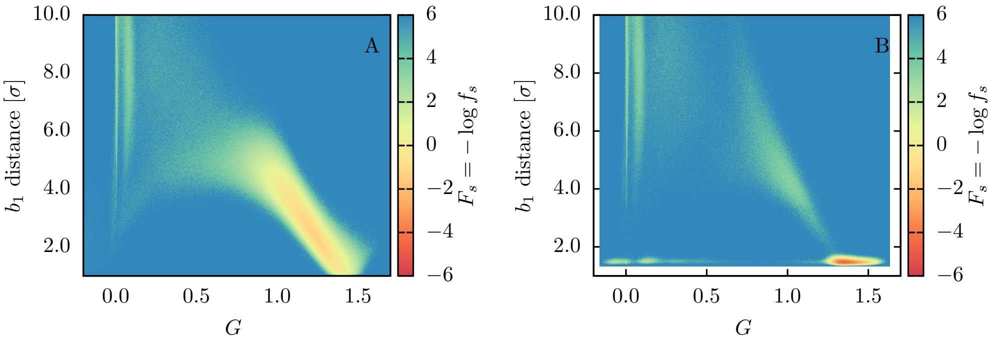


Fig 8. Successful Folding landscape F_s of the HM as a function of the Gauss linking number G and of the b_1 bridge distance. A panel: successful trajectories ($N = 1970$) in reducing conditions. B panel: successful trajectories ($N = 1946$) in oxidizing conditions.

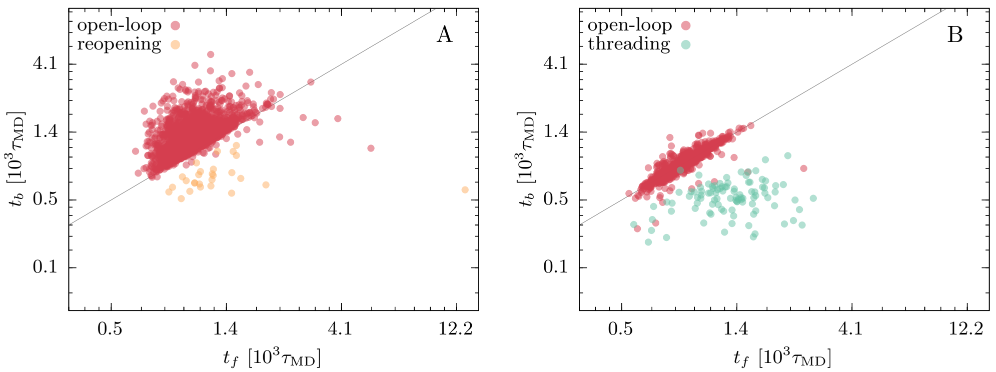


Fig 9. t_b versus t_f for the successful trajectories of the OM. Panel A shows the results of the $N = 1969$ successful trajectories in reducing conditions and panel B shows the results of the $N = 1946$ successful trajectories in oxidizing conditions. The color of the circles indicates the folding pathway, following the classification indicated in the text. The black line corresponds to $t_b = t_f$.

For details on the description we refer to the cited references, here we just underline that the native contacts establish through a 12-10 Lennard Jones potential, which is the main driving force of the folding. As mentioned in Sec. Methods, also this description models the backbone stiffness with the angular potentials of Eq. 6. The stiffness coefficients are homogeneous, set to $k^{\text{bend}} = 40.0$, $k_1^{\text{tor}} = 1.0$, $k_3^{\text{tor}} = 0.5$. Following Ref. [53] 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 Information (SI), we have chosen to simulate this model at a temperature $T = 0.7$, at which the folding is 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, since the dihedral angles are substantially less stiff than in the EFM, the computation of L must involve a larger number of residues (see the SI for further details). The folding success rates resulting from 2048 simulations in both reducing and oxidizing conditions are reported in Tab. 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 Tab. 1, and represented in the t_b versus t_f plots of Figs. 10A and B. The data indicate that the distribution of 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. This mechanism selection is the main reason why the model folds with a similar success rate in both environmental conditions. Moreover, we have found that the behavior of GōM, at the temperature of fastest folding, is in qualitative agreement with that of OM. Indeed both descriptions select a folding pathway characterized by the formation of 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 ($t < 10^3$), while only a minor fraction of trajectories folds in the remaining simulation length. This indicates that the non-successful 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, non-local motions of the backbone, as those driven by EFM angular potentials, are crucial for reproducible and efficient folding of self-entangled proteins. A model (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 studied the GōM at $T = 1.1$, both in reducing and oxidizing conditions. Again, we have collected 2048 runs of length $\tau_{\mathrm{run}} = 1.5 \times 10^4$, to detect the fast folding events. As reported in Tab. 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. 10, 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 symmetric anymore between reduced and oxidized model.

At $T = 1.1$ the model is outside 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, that 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. 10 C), where the 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. 10 D) the situation is different, since 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 towards the threading pathway. The promotion of a single folding route has here a different nature than in EFM results, where it 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.

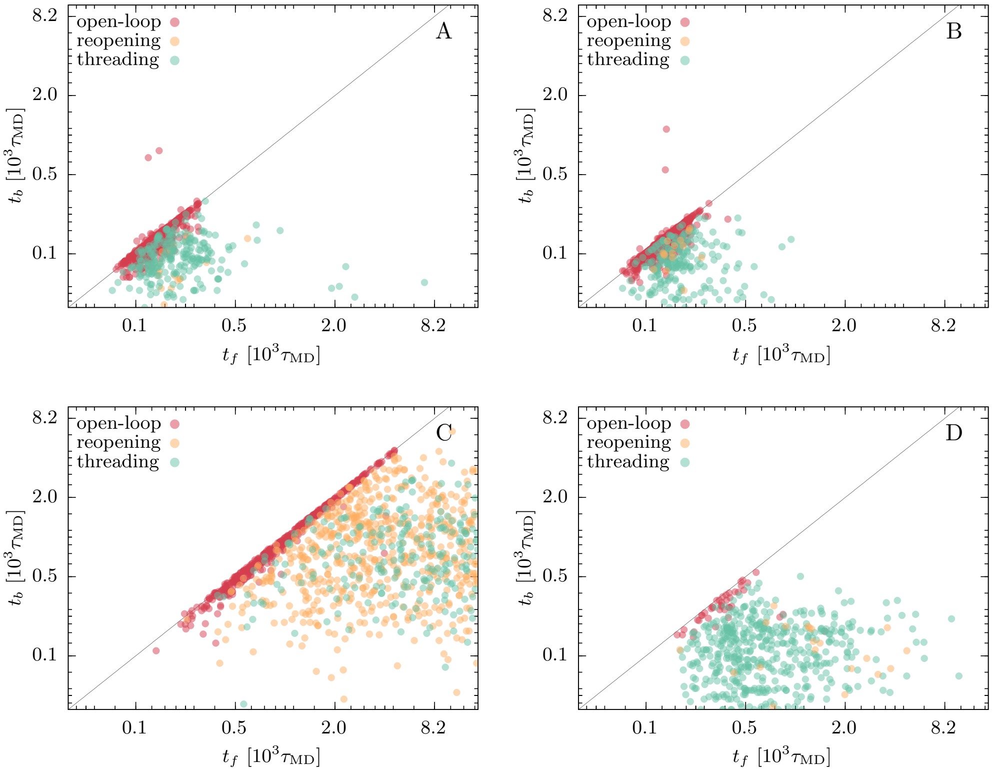


Fig 10. t_b versus t_f for the successful trajectories of GōM simulations. Panels A and B display the results of the GōM at $T = 0.7$: the $N = 1228$ successful trajectories in reducing conditions are shown in panel A and the $N = 1130$ successful trajectories in oxidizing conditions are shown in panel B. Panels C and D display the results of the GōM at $T = 1.1$: the $N = 1291$ successful trajectories in reducing conditions are shown in panel C and the $N = 549$ successful trajectories in oxidizing conditions are shown in panel D. The color of the circles indicates the folding pathway, following the classification indicated in the text. The black line corresponds to $t_b = t_f$.

Discussion

In this work we have presented an investigation on the folding of the glycoprotein Granulocyte-macrophage colony-stimulating factor (2GMF), which presents a Complex Lasso native structure. The study is performed by means of MD simulations, employing both the widely used Gō Model, proposed by Clementi et al. [26] and the EFM, a Coarse Grained, minimalistic description proposed in Ref. [48] for investigating the folding mechanisms of knotted proteins. We here extended the original models by implementing the formation of native cysteine bridges, in order 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 towards 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 optimized model 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 Gō-model 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 towards 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, as 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 towards a single mechanism. However, differently from the optimal kinetic scenario, 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. [54–56]).

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.

Methods

Elastic Folder Model Simulations

The EFM introduced in Ref. [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, parametrized 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 [57] is used for the steric term:

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

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

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

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 [58], 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 maximum bond length. The length scale σ is chosen equal to the steric diameter of Eq. (2), that 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 which has to be described. As mentioned, the folding is guided by the angular potential, that generates the dynamics of the chain bending and torsion angles:

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

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:

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

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

In 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 Gō models as well [26, 27], in order to bias the formation of proper backbone chirality [59].

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:

$$U_{sLJ} = \begin{cases} U_{LJ}(r;\epsilon,\sigma) - U_{LJ}(r_c;\epsilon,\sigma) - (r - r_c)\frac{dU_{LJ}}{dr}\big|_{r=r_c} & \text{if } r < r_c \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

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 inhouse software.

Gō Model Simulations

The employed Gō model is that introduced by Clementi et al. in Ref. [26]. The system setup was generated using the SMOG web server (http://smog-server.org) [60]. 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 [61] is used for the definition of native contacts. As mentioned before, also this description models the backbone stiffness with the angular potentials of Eq. 6. The stiffness coefficients are here homogeneous, set to $k^{\text{bend}} = 40.0$, $k_1^{\text{tor}} = 1.0$, $k_3^{\text{tor}} = 0.5$. The formation of cysteine bridges in oxidizing condition is modeled by increasing the amplitude ϵ_{ij} of the native contact potential associated to the cysteine-cysteine contacts. The value was set to $\epsilon_{ij} = 10\,k_{\text{B}}T$, 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 [62, 63] 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 SI.

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 Ref. [48] this optimization is performed through a stochastic search procedure, which we recall here. Let us first define:

$$\begin{aligned} K &= \{k_1^{\text{bend}}, \ldots, k_{N-2}^{\text{bend}}, k_1^{\text{tor}}, \ldots, k_{N-3}^{\text{tor}}\} = & (11)\\ &= \{k_1^{\text{ang}}, \ldots, k_{2N-5}^{\text{ang}}\}\,, & (12) \end{aligned}$$

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 in two operations: first, a mutated force-field K' is generated:

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

in which the j-th coefficient is modified by adding δk. j is randomly chosen among the $2N-5$ coefficients, while δk is generated with a prescribed probability distribution (e.g. in our calculation it is normally distributed, with standard deviation 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 $|\cdot|$ 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 Ref. [48], chosen according to the the convergence of the MSD value along the trajectory. In the present work we have selected $\tau = \tau_{\mathrm{min}}$, namely the time at which the MSD reaches its minimum value during the test run. The probability of acceptance of the new force-field K' is then:

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

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, that we name Single Force Field Optimization (SFFO), is displayed in Fig. 11.

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 Ref. [48] as well as in the present work, neighboring pairs of coefficients are constrained, that is:

$$K = \{k_1^{\mathrm{bend}} = k_2^{\mathrm{bend}}, k_3^{\mathrm{bend}} = k_4^{\mathrm{bend}}, \ldots, k_1^{\mathrm{tor}} = k_2^{\mathrm{tor}}, k_3^{\mathrm{tor}} = k_4^{\mathrm{tor}}, \ldots\}\,. \tag{17}$$

These local constraints reduce the dimensionality of the stochastic search, but also the generality of the model. In the present 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. 11. An initial set, or population, of force-fields $\{K_j\}_{j=1}^{N_K}$ is chosen, and each of them undergoes m SFFO steps

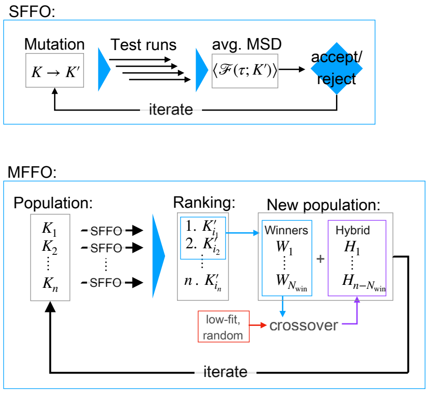


Fig 11. Optimization schemes. Schematic illustration of the SFFO and MFFO optimization algorithms.

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 $N_{\rm win}$ 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:

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

in which W indicates the winners, and H indicates a set of $N_K - N_{\rm win}$ 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. Ref. [64]). More in detail, the H_j are generated by combining fragments of force-fields, randomly picked from a set of parent force-fields, as displayed in Fig. 12). The parent set is formed by the $N_{\rm win}$ winners together with $N_{\rm low}$ "low-fit" candidates, that ensure 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 SI. Once the new population is set the optimization cycle is completed and the algorithm is re-iterated. 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 the present 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 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:

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

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. As mentioned, the ranking operation has 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 Π_f.

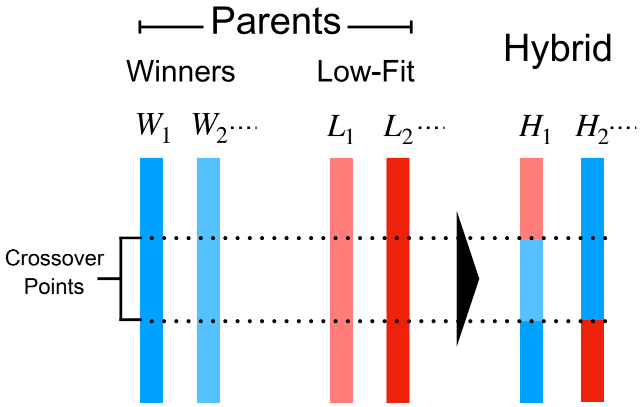


Fig 12. Crossover operation. Schematic representation of the crossover operation generating the hybrid force-fields. The color bars indicate the sets of k^{ang} coefficients associated to the Winners and the Low fit force-fields. These are mixed randomly in the hybrid force-fields.

In the optimization results presented in Sec. Results, 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 point-wise. 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, in order 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 thresold 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, that provided an on-the-fly estimate of the optimization progress. After the force-fields were ranked, the 6 best were chosen as winners, and continued the optimization. The remaining 10 force-fields were constructed combining the winners and 4 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 SI).

Topology Analysis

Minimalistic, CG models make it possible to collect a large statistics of folding trajectories, even in complex folding processes as those of self-entangled proteins. However, in order 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. Refs. [65] 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 [66, 67], so that the topological state is mathematically well-defined, and second by simplifying this structured curve without changing its topology [8, 68]. The resulting curve is then analyzed by computing topological invariants [69, 70] and its entangled state is classified.

Although this is the typical approach used to analyze knotted proteins, the non-trivial topology recognized in CL structures, is not yet classified from a mathematical point of view [2]. In Ref. [19] an approach specifically aimed at detecting CLs is presented. This technique, named Minimal Surface Analysis, uses triangulation algorithms mutuated 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 the present 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 thus 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. This operation is similar to the smoothing performed for protein knot detection [71], 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 SI) 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 a $|\mathbf{R}_{t'+1} - \mathbf{R}_{t'}|$ 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 *lasso-variable*. The approximated surface formed by the $M_l - 2$ triangles is not the minimal area surface of Ref. [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 Refs. [49, 72]. The idea developed in these works is that of employing the Gauss linking number [73], namely the double line integral:

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

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 [73]. This property can been then exploited to assess the linking in protein dimers [49], or the self-entanglement of folded proteins [72] 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, since 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 to 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 sub-chains are integrated. In the present work we have not fixed any conventional orientation, since 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.

Supporting information

SI Appendix Searching the optimal routes to the folding of a Complex Lasso protein: Supporting Information.

Figure S1 CG representation of 2GMF native structure and reduced structures employed for the calculation of L and G.

Figure S2 Angular coefficients of the optimized and homogeneous models, OM and HM

Figure S3 Temperature range of fastest folding for the SBM.

Acknowledgments

The authors thank O. Valsson and Y. Zhao for a critical reading of the manuscript. The authors would like to acknowledge the contribution of the COST Action CA17139. Computational resources were provided by The Max Planck Computing and Data Facility.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000;28(1):235–242. doi:10.1093/nar/28.1.235.

2. Dabrowski-Tumanski P, Sułkowska JI. To Tie or Not to Tie? That Is the Question. Polymers. 2017;9(9). doi:10.3390/polym9090454.

3. Jamroz M, Niemyska W, Rawdon EJ, Stasiak A, Millett KC, Sułkowski P, et al. KnotProt: a database of proteins with knots and slipknots. Nucleic Acids Res. 2015;43(Database issue):D306–D314. doi:10.1093/nar/gku1059.

4. Mansfield ML. Are there knots in proteins? Nat Struct Mol Biol. 1994;1(4):213–214. doi:10.1038/nsb0494-213.

5. Jackson SE, Suma A, Micheletti C. How to fold intricately: using theory and experiments to unravel the properties of knotted proteins. Current Opinion in Structural Biology. 2017;42:6 – 14. doi:https://doi.org/10.1016/j.sbi.2016.10.002.

6. Faísca PF. Knotted proteins: A tangled tale of Structural Biology. Comput Struct Biotechnol J. 2015;13:459–468. doi:10.1016/j.csbj.2015.08.003.

7. Lim NCH, Jackson SE. Molecular knots in biology and chemistry. Journal of Physics: Condensed Matter. 2015;27(35):354101.

8. Lua RC. PyKnot: a PyMOL tool for the discovery and analysis of knots in proteins. Bioinformatics. 2012;28(15):2069. doi:10.1093/bioinformatics/bts299.

9. Wüst T, Reith D, Virnau P. Sequence Determines Degree of Knottedness in a Coarse-Grained Protein Model. Phys Rev Lett. 2015;114:028102. doi:10.1103/PhysRevLett.114.028102.

10. Potestio R, Micheletti C, Orland H. Knotted vs. Unknotted Proteins: Evidence of Knot-Promoting Loops. PLOS Computational Biology. 2010;6(7):1–10. doi:10.1371/journal.pcbi.1000864.

11. Sułkowska JI, Rawdon EJ, Millett KC, Onuchic JN, Stasiak A. Conservation of complex knotting and slipknotting patterns in proteins. Proc Natl Acad Sci U S A. 2012;109(26):E1715–E1723. doi:10.1073/pnas.1205918109.

12. Sułkowska JI, Sułkowski P, Szymczak P, Cieplak M. Stabilizing effect of knots on proteins. Proceedings of the National Academy of Sciences. 2008;105(50):19714–19719. doi:10.1073/pnas.0805468105.

13. Christian T, Sakaguchi R, Perlinska AP, Lahoud G, Ito T, Taylor EA, et al. Methyl transfer by substrate signaling from a knotted protein fold. Nature structural & molecular biology. 2016;23(10):941.

14. Dabrowski-Tumanski P, Stasiak A, Sułkowska JI. In Search of Functional Advantages of Knots in Proteins. PLOS ONE. 2016;11(11):1–14. doi:10.1371/journal.pone.0165986.

15. Mallam AL, Jackson SE. Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins. Nature chemical biology. 2012;8(2):147–153.

16. King NP, Jacobitz AW, Sawaya MR, Goldschmidt L, Yeates TO. Structure and folding of a designed knotted protein. Proceedings of the National Academy of Sciences. 2010;107(48):20732–20737. doi:10.1073/pnas.1007602107.

17. Wang I, Chen SY, Hsu STD. Unraveling the Folding Mechanism of the Smallest Knotted Protein, MJ0366. The Journal of Physical Chemistry B. 2015;119(12):4359–4370. doi:10.1021/jp511029s.

18. Lim NCH, Jackson SE. Mechanistic Insights into the Folding of Knotted Proteins In Vitro and In Vivo. Journal of Molecular Biology. 2015;427(2):248 – 258. doi:http://dx.doi.org/10.1016/j.jmb.2014.09.007.

19. Niemyska W, Dabrowski-Tumanski P, Kadlof M, Haglund E, Sułkowski P, Sułkowska JI. Complex lasso: new entangled motifs in proteins. Sci Rep. 2016;6:36895. doi:10.1038/srep36895.

20. Haglund E, Sułkowska JI, He Z, Feng GS, Jennings PA, Onuchic JN. The Unique Cysteine Knot Regulates the Pleotropic Hormone Leptin. PLOS ONE. 2012;7(9):1–13. doi:10.1371/journal.pone.0045654.

21. Dabrowski-Tumanski P, Niemyska W, Pasznik P, Sułkowska JI. LassoProt: server to analyze biopolymers with lassos. Nucleic Acids Research. 2016;44(W1):W383. doi:10.1093/nar/gkw308.

22. Haglund E, Sułkowska JI, Noel JK, Lammert H, Onuchic JN, Jennings PA. Pierced Lasso Bundles Are a New Class of Knot-like Motifs. PLOS Computational Biology. 2014;10(6):1–11. doi:10.1371/journal.pcbi.1003613.

23. Haglund E, Pilko A, Wollman R, Jennings PA, Onuchic JN. Pierced Lasso Topology Controls Function in Leptin. The Journal of Physical Chemistry B. 2017;121(4):706–718. doi:10.1021/acs.jpcb.6b11506.

24. a Beccara S, Škrbić T, Covino R, Micheletti C, Faccioli P. Folding Pathways of a Knotted Protein with a Realistic Atomistic Force Field. PLOS Computational Biology. 2013;9(3):1–9. doi:10.1371/journal.pcbi.1003002.

25. Noel JK, Onuchic JN, Sułkowska JI. Knotting a Protein in Explicit Solvent. The Journal of Physical Chemistry Letters. 2013;4(21):3570–3573. doi:10.1021/jz401842f.

26. Clementi C, Nymeyer H, Onuchic JN. 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 proteins11Edited by F. E. Cohen. Journal of Molecular Biology. 2000;298(5):937 – 953. doi:https://doi.org/10.1006/jmbi.2000.3693.

27. Cieplak M, Hoang TX. Universality Classes in Folding Times of Proteins. Biophysical Journal. 2003;84(1):475 – 488. doi:https://doi.org/10.1016/S0006-3495(03)74867-X.

28. Go N, Taketomi H. Respective roles of short- and long-range interactions in protein folding. Proceedings of the National Academy of Sciences. 1978;75(2):559–563.

29. Onuchic JN, Wolynes PG. Theory of protein folding. Current Opinion in Structural Biology. 2004;14(1):70 – 75. doi:https://doi.org/10.1016/j.sbi.2004.01.009.

30. Best RB, Hummer G, Eaton WA. Native contacts determine protein folding mechanisms in atomistic simulations. Proceedings of the National Academy of Sciences. 2013;110(44):17874–17879. doi:10.1073/pnas.1311599110.

31. Hoang TX, Cieplak M. Sequencing of folding events in Go-type proteins. The Journal of Chemical Physics. 2000;113(18):8319–8328. doi:10.1063/1.1314868.

32. Shea JE, Onuchic JN, Brooks CL. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. Proceedings of the National Academy of Sciences. 1999;96(22):12512–12517. doi:10.1073/pnas.96.22.12512.

33. Shea JE, Onuchic JN, Brooks CL. Probing the folding free energy landscape of the src-SH3 protein domain. Proceedings of the National Academy of Sciences. 2002;99(25):16064–16068. doi:10.1073/pnas.242293099.

34. Sułkowska JI, Cieplak M. Mechanical stretching of proteins—a theoretical survey of the Protein Data Bank. Journal of Physics: Condensed Matter. 2007;19(28):283201.

35. Sułkowska JI, Cieplak M. Selection of Optimal Variants of Gō-Like Models of Proteins through Studies of Stretching. Biophysical Journal. 2008;95(7):3174 – 3191. doi:https://doi.org/10.1529/biophysj.107.127233.

36. Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. Proteins: Structure, Function, and Bioinformatics. 2009;75(2):430–441. doi:10.1002/prot.22253.

37. Sułkowska JI, Sułkowski P, Onuchic J. Dodging the crisis of folding proteins with knots. Proceedings of the National Academy of Sciences. 2009;106(9):3119–3124. doi:10.1073/pnas.0811147106.

38. Noel JK, Sułkowska JI, Onuchic JN. Slipknotting upon native-like loop formation in a trefoil knot protein. Proceedings of the National Academy of Sciences. 2010;107(35):15403–15408. doi:10.1073/pnas.1009522107.

39. Sułkowska JI, Noel JK, Onuchic JN. Energy landscape of knotted protein folding. Proceedings of the National Academy of Sciences. 2012;109(44):17783–17788. doi:10.1073/pnas.1201804109.

40. Chwastyk M, Cieplak M. Cotranslational folding of deeply knotted proteins. Journal of Physics: Condensed Matter. 2015;27(35):354105.

41. Zhao Y, Dabrowski-Tumanski P, Niewieczerzal S, Sułkowska JI. The exclusive effects of chaperonin on the behavior of proteins with 52 knot. PLOS Computational Biology. 2018;14(3):1–20. doi:10.1371/journal.pcbi.1005970.

42. Wallin S, Zeldovich KB, Shakhnovich EI. The Folding Mechanics of a Knotted Protein. Journal of Molecular Biology. 2007;368(3):884 – 893. doi:http://dx.doi.org/10.1016/j.jmb.2007.02.035.

43. Škrbić T, Micheletti C, Faccioli P. The Role of Non-Native Interactions in the Folding of Knotted Proteins. PLOS Computational Biology. 2012;8(6):1–12. doi:10.1371/journal.pcbi.1002504.

44. Soler MA, Nunes A, Faísca PFN. Effects of knot type in the folding of topologically complex lattice proteins. The Journal of Chemical Physics. 2014;141(2):025101. doi:10.1063/1.4886401.

45. Dabrowski-Tumanski P, Jarmolinska AI, Sułkowska JI. Prediction of the optimal set of contacts to fold the smallest knotted protein. Journal of Physics: Condensed Matter. 2015;27(35):354109.

46. Covino R, Škrbić T, Beccara Sa, Faccioli P, Micheletti C. The Role of Non-Native Interactions in the Folding of Knotted Proteins: Insights from Molecular Dynamics Simulations. Biomolecules. 2014;4(1):1–19. doi:10.3390/biom4010001.

47. Chwastyk M, Cieplak M. Multiple folding pathways of proteins with shallow knots and co-translational folding. The Journal of Chemical Physics. 2015;143(4):045101. doi:10.1063/1.4927153.

48. Najafi S, Potestio R. Folding of small knotted proteins: Insights from a mean field coarse-grained model. The Journal of Chemical Physics. 2015;143(24):243121. doi:10.1063/1.4934541.

49. Baiesi M, Orlandini E, Trovato A, Seno F. Linking in domain-swapped protein dimers. Scientific reports. 2016;6:33872.

50. Rozwarski DA, Diederichs K, Hecht R, Boone T, Karplus PA. Refined crystal structure and mutagenesis of human granulocyte-macrophage colony-stimulating factor. Proteins: Structure, Function, and Bioinformatics. 1996;26(3):304–313. doi:10.1002/(SICI)1097-0134(199611)26:3¡304::AID-PROT6¿3.0.CO;2-D.

51. Zhang Y, Zhang J, Wang W. Atomistic Analysis of Pseudoknotted RNA Unfolding. Journal of the American Chemical Society. 2011;133(18):6882–6885. doi:10.1021/ja1109425.

52. Li W, Terakawa T, Wang W, Takada S. Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. Proceedings of the National Academy of Sciences. 2012;109(44):17789–17794. doi:10.1073/pnas.1201807109.

53. Zhao Y, Cieplak M. Stability of structurally entangled protein dimers. Proteins: Structure, Function, and Bioinformatics. 2018;86(9):945–955. doi:10.1002/prot.25526.

54. Baldwin RL. Clash between energy landscape theory and foldon-dependent protein folding. Proceedings of the National Academy of Sciences. 2017;114(32):8442–8443. doi:10.1073/pnas.1709133114.

55. Eaton WA, Wolynes PG. Theory, simulations, and experiments show that proteins fold by multiple pathways. Proceedings of the National Academy of Sciences. 2017;doi:10.1073/pnas.1716444114.

56. Englander SW, Mayne L. Reply to Eaton and Wolynes: How do proteins fold? Proceedings of the National Academy of Sciences. 2017;114(46):E9761–E9762. doi:10.1073/pnas.1716929114.

57. Weeks JD, Chandler D, Andersen HC. Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. The Journal of Chemical Physics. 1971;54(12):5237–5247. doi:10.1063/1.1674820.

58. Grest GS, Kremer K. Molecular dynamics simulation for polymers in the presence of a heat bath. Phys Rev A. 1986;33:3628–3631. doi:10.1103/PhysRevA.33.3628.

59. Kwiecińska JI, Cieplak M. Chirality and protein folding. Journal of Physics: Condensed Matter. 2005;17(18):S1565.

60. Noel JK, Levi M, Raghunathan M, Lammert H, Hayes RL, Onuchic JN, et al. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. PLOS Computational Biology. 2016;12(3):1–14. doi:10.1371/journal.pcbi.1004794.

61. Noel JK, Whitford PC, Onuchic JN. The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function. The Journal of Physical Chemistry B. 2012;116(29):8692–8702. doi:10.1021/jp300852d.

62. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. Journal of Computational Chemistry. 2005;26(16):1701–1718. doi:10.1002/jcc.20291.

63. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1-2:19 – 25. doi:https://doi.org/10.1016/j.softx.2015.06.001.

64. Huang C, Yang X, He Z. Protein folding simulations of 2D HP model by the genetic algorithm based on optimal secondary structures. Computational Biology and Chemistry. 2010;34(3):137 – 142. doi:https://doi.org/10.1016/j.compbiolchem.2010.04.002.

65. Micheletti C, Marenduzzo D, Orlandini E. Polymers with spatial or topological constraints: Theoretical and computational results. Physics Reports. 2011;504(1):1 – 73. doi:https://doi.org/10.1016/j.physrep.2011.03.003.

66. Millett K, Dobay A, Stasiak A. Linear Random Knots and Their Scaling Behavior. Macromolecules. 2005;38(2):601–606. doi:10.1021/ma048779a.

67. Tubiana L, Orlandini E, Micheletti C. Probing the Entanglement and Locating Knots in Ring Polymers: A Comparative Study of Different Arc Closure Schemes. Progress of Theoretical Physics Supplement. 2011;191:192. doi:10.1143/PTPS.191.192.

68. Kolesov G, Virnau P, Kardar M, Mirny LA. Protein knot server: detection of knots in protein structures. Nucleic Acids Res. 2007;35(Web Server issue):W425–W428. doi:10.1093/nar/gkm312.

69. Wu FY. Knot theory and statistical mechanics. Rev Mod Phys. 1992;64:1099–1131. doi:10.1103/RevModPhys.64.1099.

70. Cromwell PR. Knots and links. Cambridge University Press; 2004.

71. Koniaris K, Muthukumar M. Self-entanglement in ring polymers. The Journal of Chemical Physics. 1991;95(4):2873–2881. doi:10.1063/1.460889.

72. Baiesi M, Orlandini E, Seno F, Trovato A. Exploring the correlation between the folding rates of proteins and the entanglement of their native states. Journal of Physics A: Mathematical and Theoretical. 2017;50(50):504001.

73. Panagiotou E, Kröger M, Millett KC. Writhe and mutual entanglement combine to give the entanglement length. Phys Rev E. 2013;88:062604. doi:10.1103/PhysRevE.88.062604.

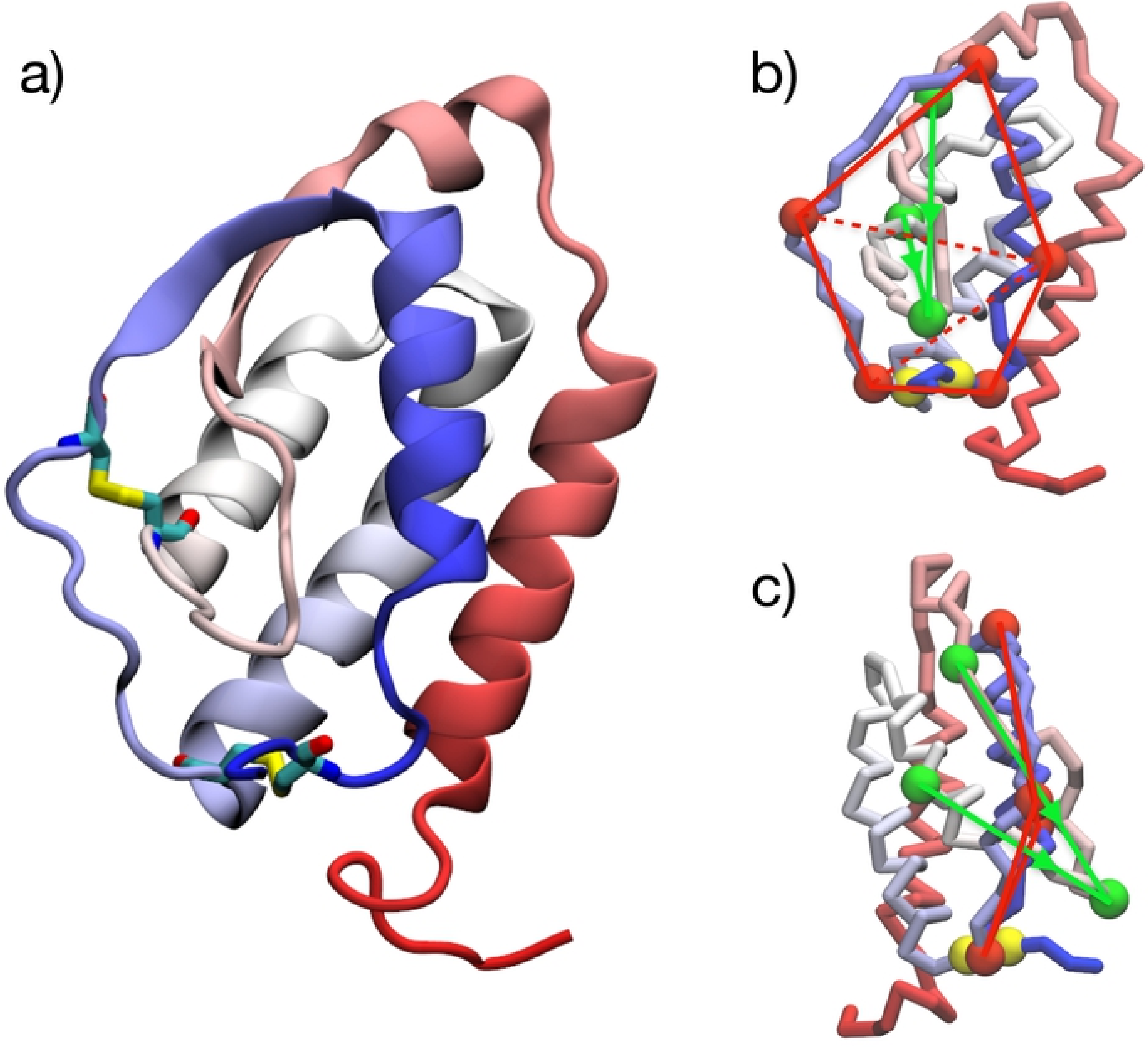


Fig. 1

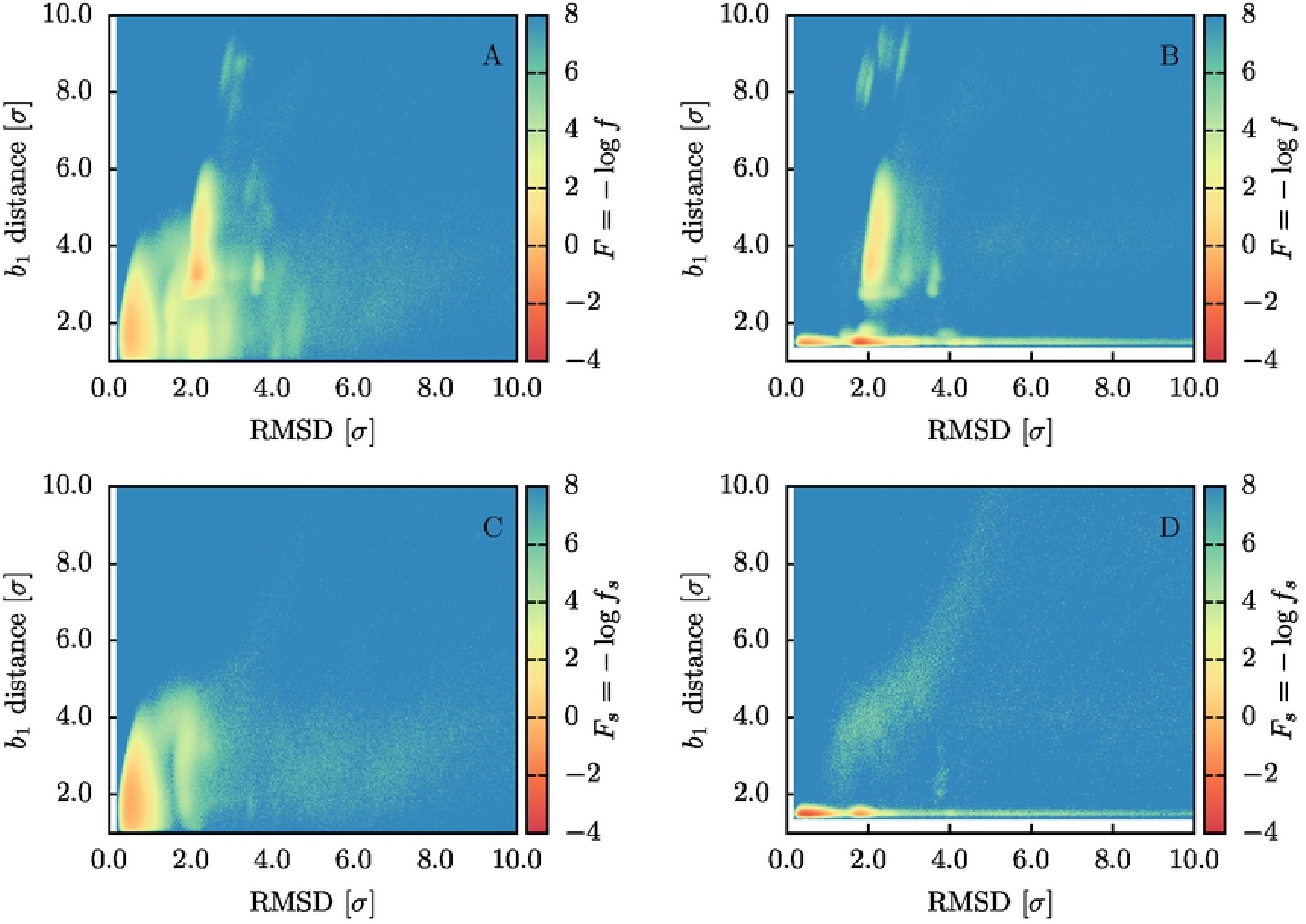
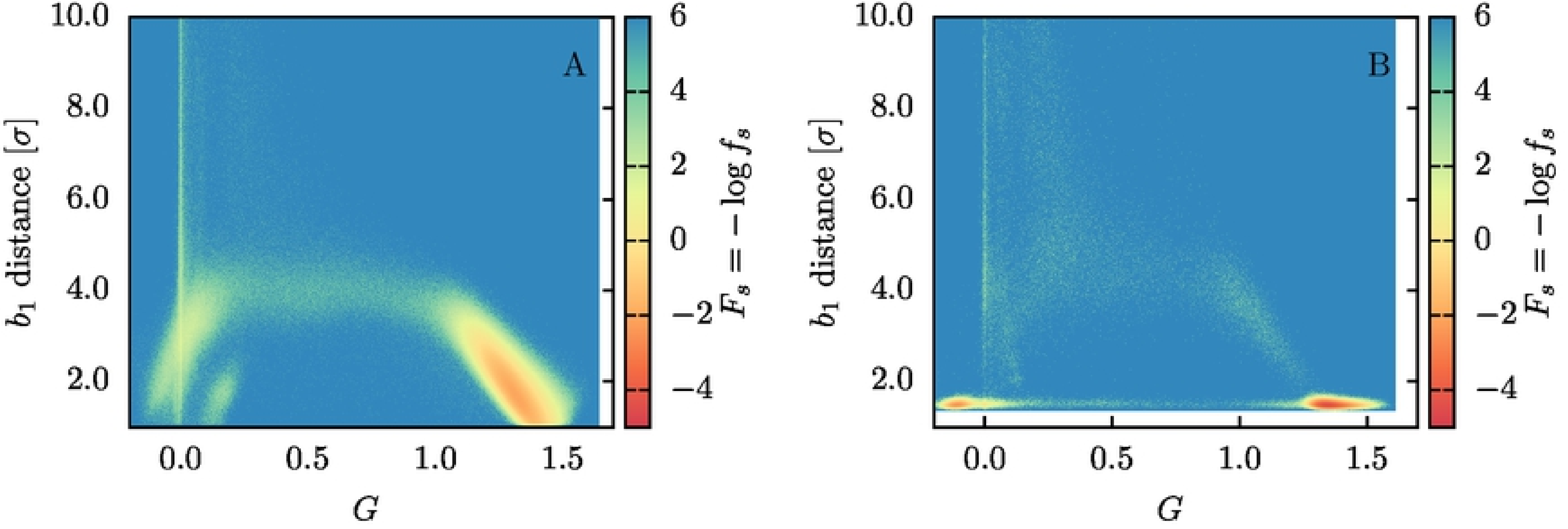

A
B
C
D
b_1 distance $[\sigma]$
RMSD $[\sigma]$
$F = -\log f$
$F_s = -\log f_s$
0.0
2.0
4.0
6.0
8.0
10.0
8
6
4
2
0
−2
−4

Fig. 2



10.0
8.0
6.0
4.0
2.0
b_1 distance $[\sigma]$
0.0
0.5
1.0
1.5
G
A
B
6
4
2
0
−2
−4
$F_s = -\log f_s$

Fig. 3

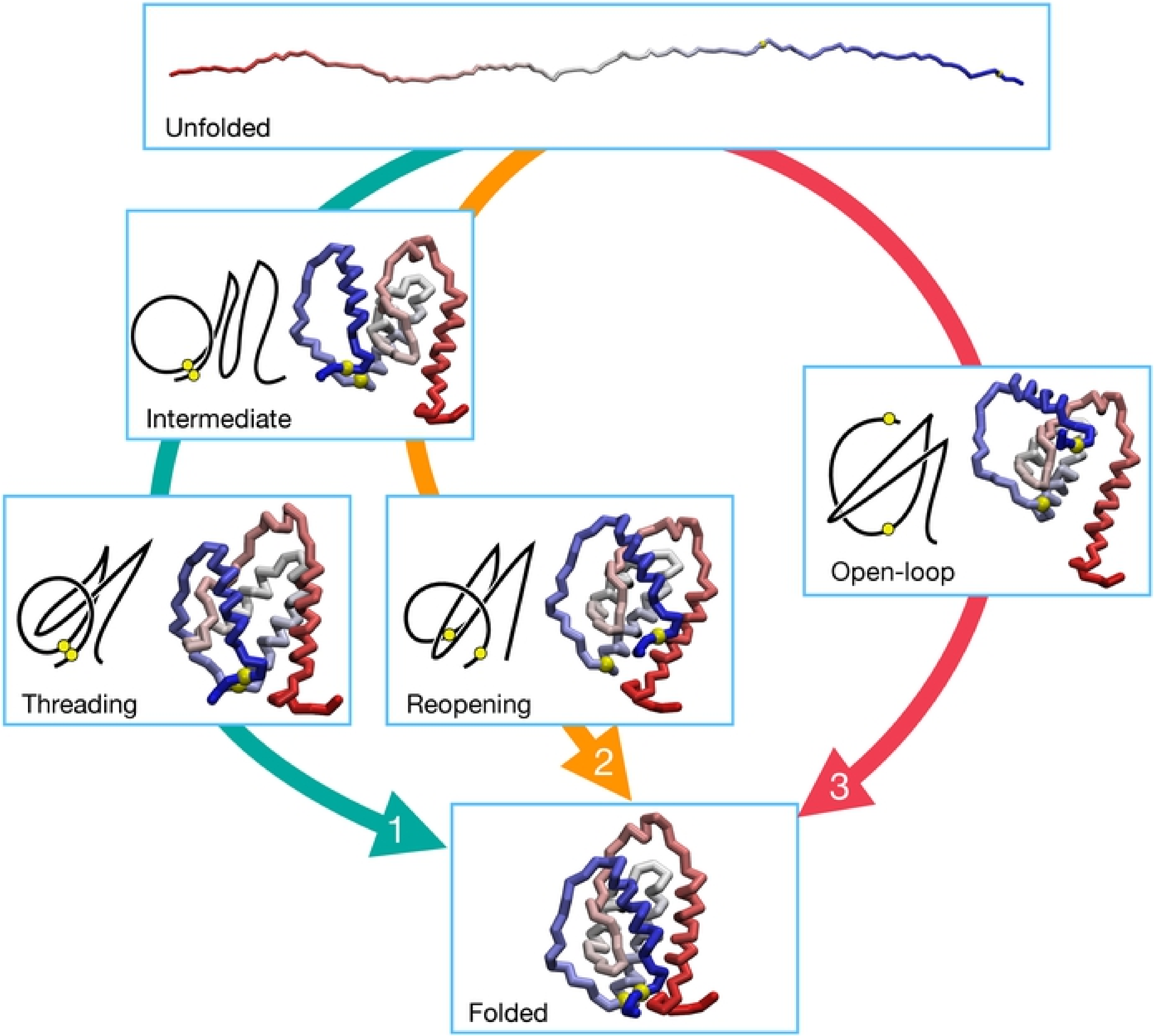
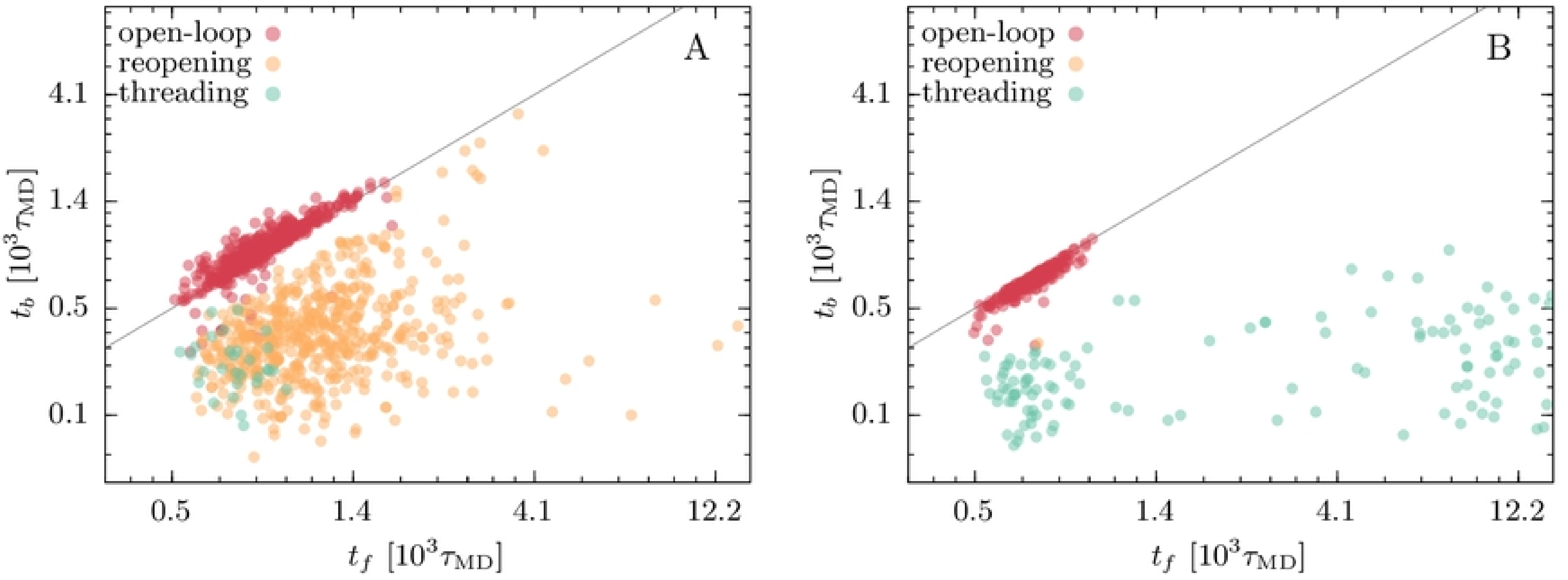
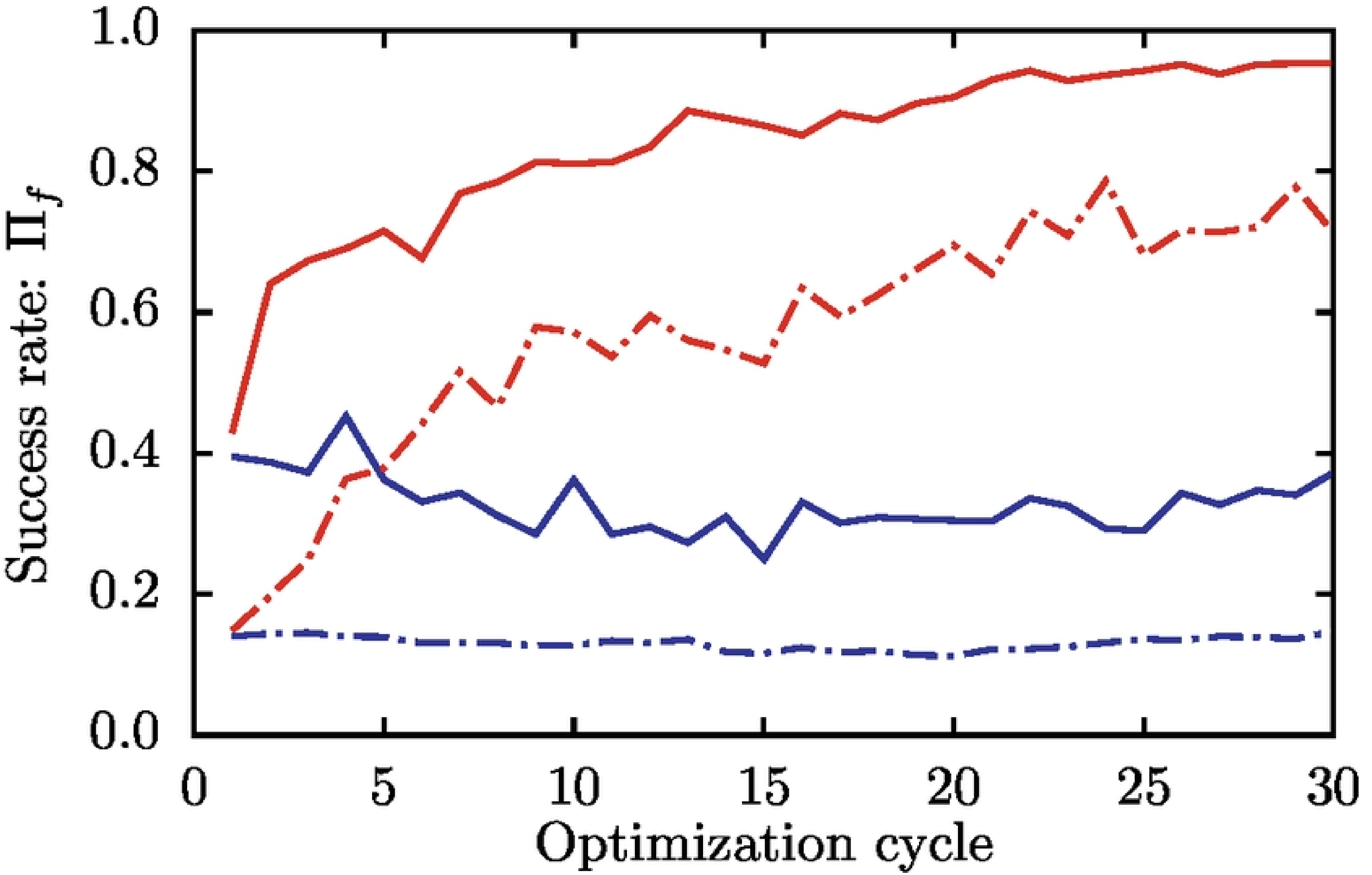

Unfolded
Intermediate
Open-loop
Threading
Reopening
1
2
3
Folded

Fig. 4



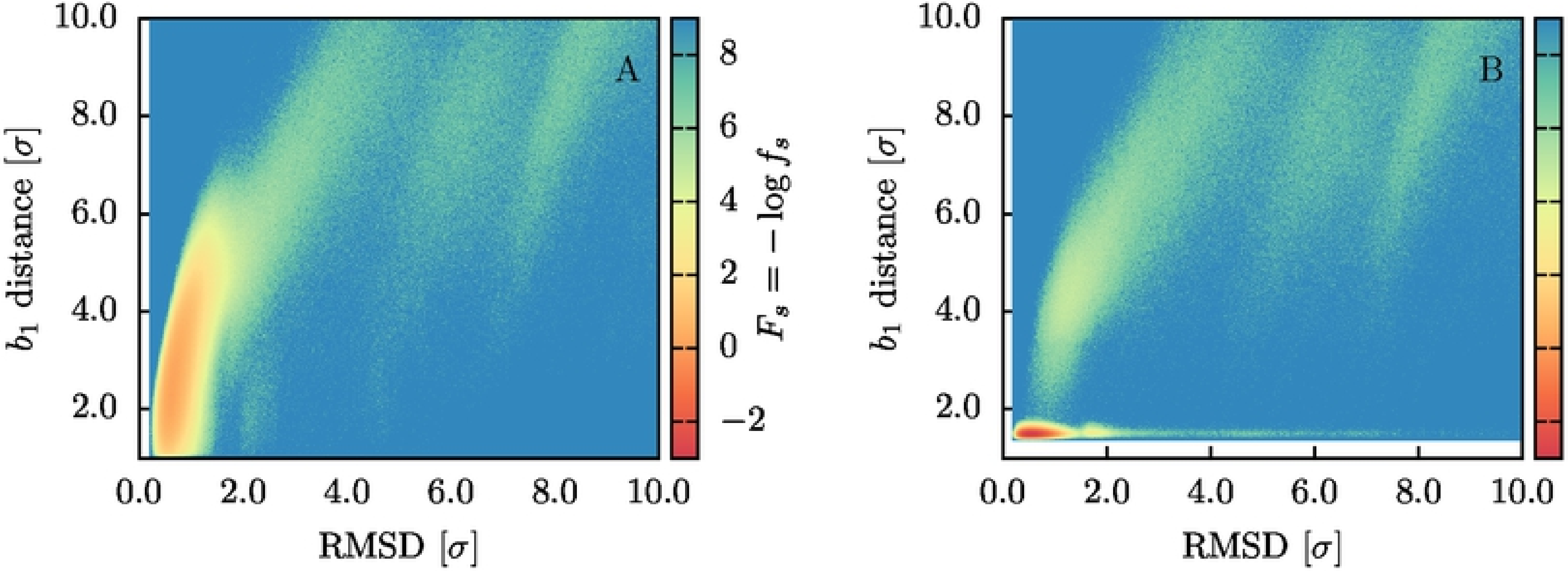
open-loop
reopening
threading
A
B
4.1
1.4
0.5
0.1
t_b $[10^3\tau_{\mathrm{MD}}]$
0.5
1.4
4.1
12.2
t_f $[10^3\tau_{\mathrm{MD}}]$

Fig. 5



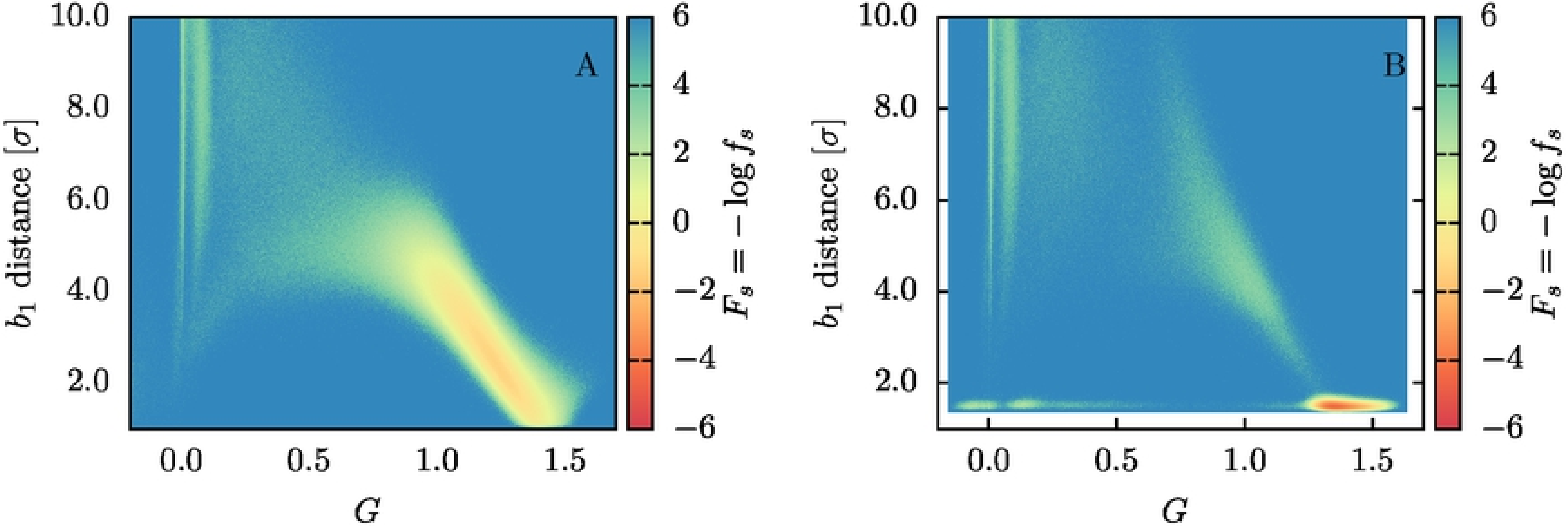
1.0
0.8
0.6
0.4
0.2
0.0
Success rate: Π_f
0
5
10
15
20
25
30
Optimization cycle

Fig. 6



10.0
8.0
6.0
4.0
2.0
b_1 distance $[\sigma]$
A
0.0
2.0
4.0
6.0
8.0
10.0
RMSD $[\sigma]$
8
6
4
2
0
-2
$F_s = -\log f_s$
B

Fig. 7



A
B
10.0
8.0
6.0
4.0
2.0
b_1 distance $[\sigma]$
0.0
0.5
1.0
1.5
G
6
4
2
0
−2
−4
−6
$F_s = -\log f_s$

Fig. 8

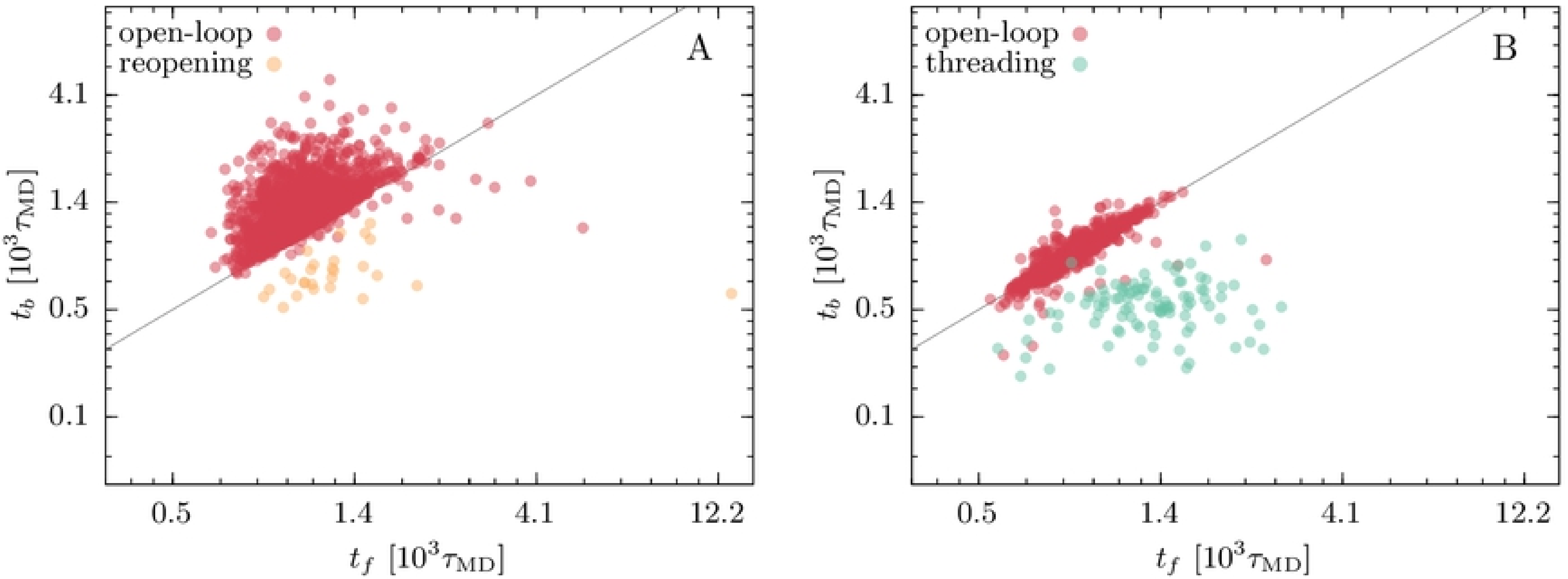

open-loop ●
reopening ●
A
open-loop ●
threading ●
B
t_b $[10^3\tau_{\rm MD}]$
4.1
1.4
0.5
0.1
0.5
1.4
4.1
12.2
t_f $[10^3\tau_{\rm MD}]$

Fig. 9

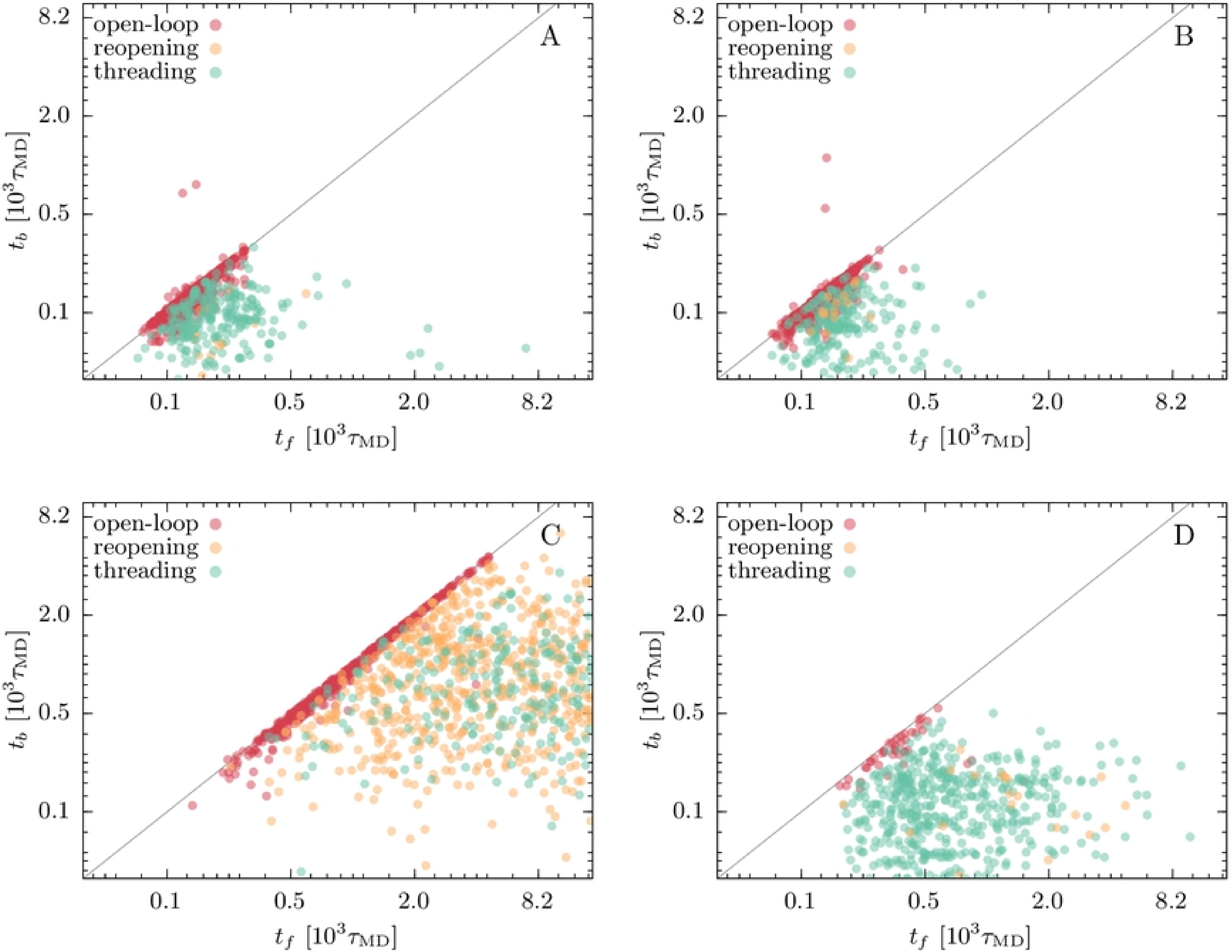

A
B
C
D
open-loop
reopening
threading
8.2
2.0
0.5
0.1
t_b [$10^3\tau_{\mathrm{MD}}$]
t_f [$10^3\tau_{\mathrm{MD}}$]

Fig. 10

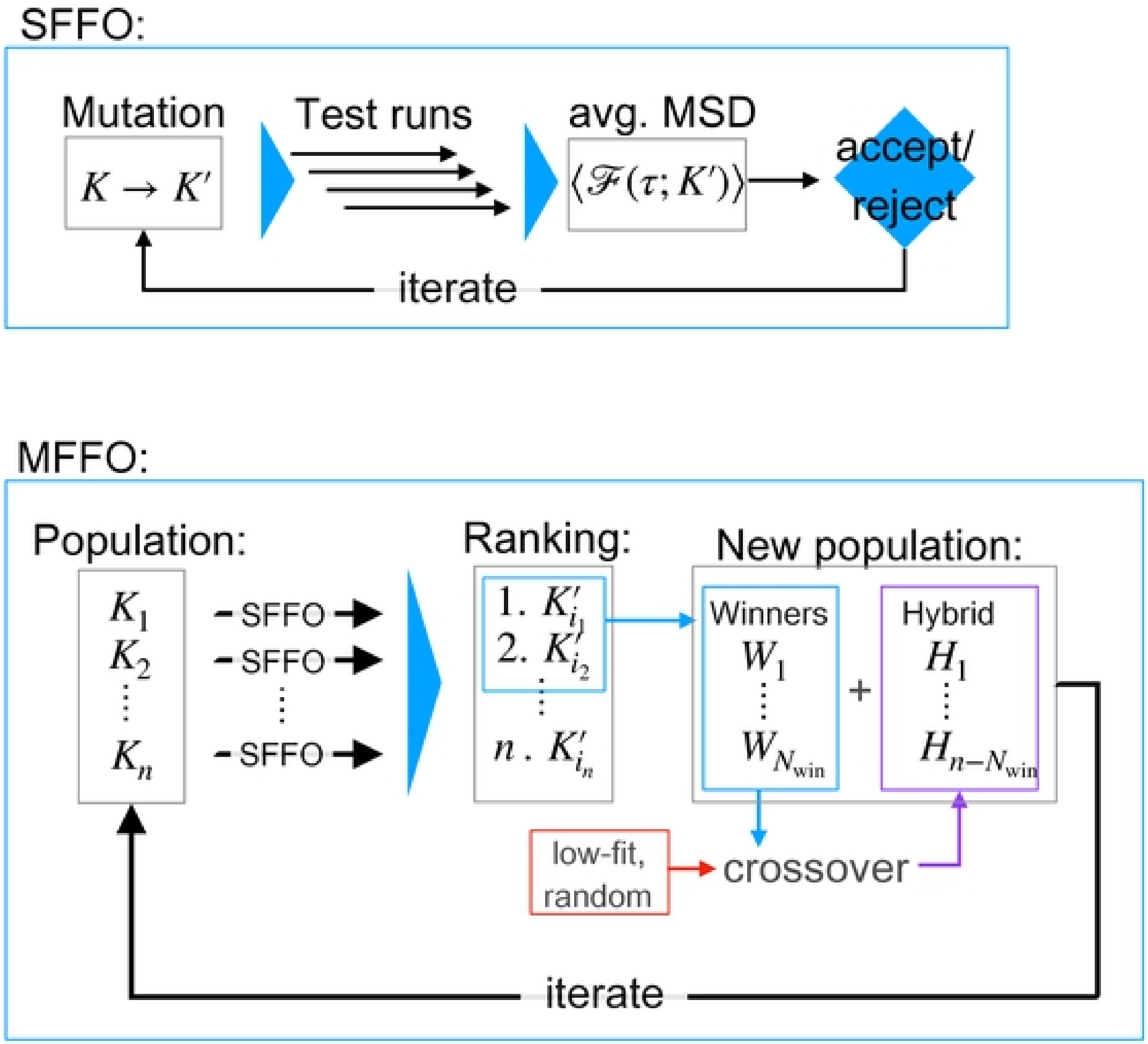

SFFO:
Mutation
$K \to K'$
Test runs
avg. MSD
$\langle \mathscr{F}(\tau; K') \rangle$
accept/ reject
iterate
MFFO:
Population:
K_1
K_2
K_n
SFFO
SFFO
SFFO
Ranking:
1. K'_{i_1}
2. K'_{i_2}
n . K'_{i_n}
New population:
Winners
W_1
$W_{N_{\text{win}}}$
+
Hybrid
H_1
$H_{n-N_{\text{win}}}$
low-fit, random
crossover
iterate

Fig. 11

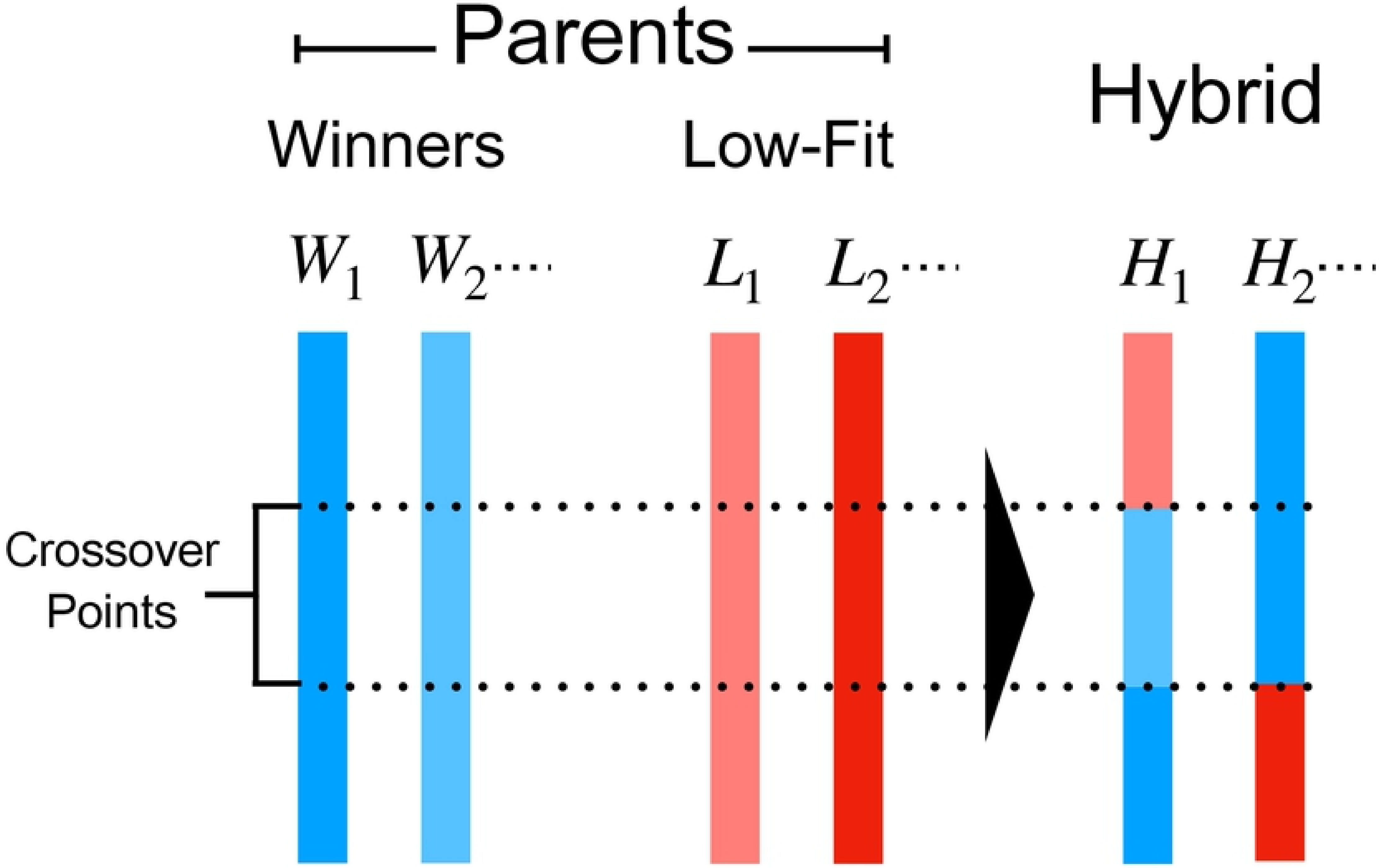

Parents
Hybrid
Winners
Low-Fit
W_1 W_2····
L_1 L_2····
H_1 H_2····
Crossover Points

Fig.12