POLITECNICO DI TORINO
Repository ISTITUZIONALE

Native state of natural proteins optimizes local entropy

Original
Native state of natural proteins optimizes local entropy / Negri, M., Tiana, G., Zecchina, R.. - In: PHYSICAL REVIEW. E. - ISSN 2470-0045. - ELETTRONICO. - 104:6(2021), p. 064117. [10.1103/PhysRevE.104.064117]

Availability:
This version is available at: 11583/2954904 since: 2022-02-09T13:49:46Z

Publisher:
American Physical Society

Published
DOI:10.1103/PhysRevE.104.064117

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

29 November 2026

Native state of natural proteins optimizes local entropy

M. Negri [1,*] G. Tiana [2,†] and R. Zecchina [3,‖]

[1]*Department Applied Science and Technology, Politecnico di Torino, CorsoDuca degli Abruzzi 24, I-10129 Turin, Italy*

[2]*Department of Physics and Center for Complexity and Biosystems, Università degli Studi di Milano and INFN, via Celoria 16, 20133 Milan, Italy*

[3]*Artificial Intelligence Lab, Bocconi University, Via Sarfatti, 25, 20136 Milan, Italy*

(Received 23 July 2021; revised 17 November 2021; accepted 24 November 2021; published 13 December 2021)

The differing ability of polypeptide conformations to act as the native state of proteins has long been rationalized in terms of differing kinetic accessibility or thermodynamic stability. Building on the successful applications of physical concepts and sampling algorithms recently introduced in the study of disordered systems, in particular artificial neural networks, we quantitatively explore how well a quantity known as the local entropy describes the native state of model proteins. In lattice models and all-atom representations of proteins, we are able to efficiently sample high local entropy states and to provide a proof of concept of enhanced stability and folding rate. Our methods are based on simple and general statistical-mechanics arguments, and thus we expect that they are of very general use.

DOI: 10.1103/PhysRevE.104.064117

I. INTRODUCTION

Proteins are the machinery of life. In order to perform their functions, the majority of proteins fold into a compact native state that is intimately linked with their polypeptide conformation, as has been known for 30 years [1]. The goal of this study was to rationalize the observation that the number of protein sequences is largely more abundant than the number of protein conformations, causing substantial degeneracy in the map between sequence and structure. The problem has regained importance in recent years, when it has become feasible to design proteins *de novo* with custom functions, and thus it has become critical to understand which conformations can be best designed [2].

Most theoretical results point to the conclusion that we can observe a small subset of all possible protein conformations because they exhibit physical properties that make them biologically more fit rather than just because they are poorly sampled by evolution. Some conformations therefore appear to be more "designable" than others [3]. Some works justify the better designability of existing conformations by their greater kinetic accessibility during the folding process, often associated with conformational symmetries such as secondary structures [4–6]. An optimal balance between local and nonlocal contacts would make the folding rate of some native conformations particularly fast. This hypothesis is supported by the correlation observed in proteins between the average linear separation between residues in contact in the native state and the rate of folding [7]. Other works relate the design of a native conformation to its thermodynamic stability, thus focusing on its equilibrium rather than kinetic characteristics. The basic idea is that stable proteins exhibit a large gap between the energy of the native state and those of competing conformations [8], and this gap can accommodate a large number of sequences that fold to the same native state [9,10]. Different conformational properties may contribute to this enhanced thermodynamic stability. More compact conformations are more stable because they can exhibit more attractive interactions and because they protect more efficiently hydrophobic residues from the solvent. Daisy-like conformations that maximize the trace of the eighth power of the contact matrix have also been shown to be particularly stable [11]; in fact, this quantity has been shown to correlate with the evolutionary age of the proteins [12]. Similarly, the presence of loops of specific sizes has been found to improve thermodynamic stability and justified by [13] energy arguments.

Still, the fact that specific protein conformations can be particularly stable at equilibrium is usually associated with the property of amino acid sequences to exhibit markedly low potential energy in those conformations, thus more efficiently minimizing system frustration. On the other hand, although the native state of proteins is usually considered macroscopically unique, its entropy is not negligible compared to the competing denatured state [14]. This entropy arises from the constellation of conformations, structurally similar to the ground state, that lie beyond the transition state. They certainly include several vibrational states, conformational substates [15], and perhaps other conformations whose contribution to the partition function cannot be separated from that of the ground state.

The hypothesis we wish to further investigate in the present work is whether the entropy of the native state of natural proteins, and not just the potential energy, is particularly optimized compared to that of random conformations. This

*Corresponding author: matteo.negri@polito.it
†guido.tiana@unimi.it
‖riccardo.zecchina@unibocconi.it

 ©2021 American Physical Society

hypothesis was suggested several years ago [16] by a qualitative computational analysis in which some of the most mobile dihedrals of small proteins were changed and the number of conformations within 4 Å from native conformations and devoid of steric clashes was estimated. It was found that natural proteins exhibit more neighboring conformations than random decoys. More recently, a knowledge-based local-entropy parametrization was used to predict contact changes between amino acids during protein conformational changes [17].

From a statistical-physics point of view, this quantity is captured by the so-called *local entropy*, i.e., the log of the number of low-energy configurations within a given distance from a reference configuration (for continuous systems the definition can be generalized straightforwardly). Recently, it has been seen that the notion of local entropy plays a central role in systems that exhibit a potentially very complex energy landscape and at the same time possess highly accessible states that correspond to high local entropy minima [18,19]. The latter turn out to be accessible by a multitude of dynamical processes which are not designed to have the Gibbs distribution as a stationary probability measure, due, e.g., to nonthermal external perturbations. Systems of this type are nonconvex models of artificial neural networks (including deep neural networks), in which entropic phenomena play essential roles [20] for the (unexpected) success of the current learning processes, largely based on nonequilibrium variants of gradient descent.

Here we generalize the algorithmic schemes introduced for sampling high local entropy ground states in neural systems [21] and apply them to simple models of 3D protein structures. Considering both lattice model and all-atom representations of proteins, we show that by sampling native states with high local entropy (which are in principle rare compared to states that dominate the Gibbs measure) we find a decrease in the linear separation between contact residues. In addition, the "flatness" of the energy profile in the native state can extend to the transition state, having consequences both on the thermodynamic stability of the protein, lowering the free energy of the native state, and on the folding rate, lowering the free energy of the transition state. The generality of the sampling method would allow it to be used in conjunction with any structure prediction method, such as AlphaFold [22] and RoseTTA fold [23], and could aid in the search for sequences folding onto the most designable structures.

II. LOCAL ENTROPY AND HOW TO CALCULATE IT

To define a probability measure that ignores narrow ground states and enhances the statistical weight of large dense regions of ground states, we can consider the local free entropy

$$S_{\rm loc}(\Gamma,\gamma,\beta)=\log\int d\Gamma'\exp\left[-\beta U(\Gamma')-\gamma d(\Gamma,\Gamma')\right], \tag{1}$$

where $d(\Gamma,\Gamma')$ is any metrics suitable for the model under consideration and γ is its conjugate Lagrange multiplier. We describe the explicit choice of d in Sec. III [see Eq. (7)]. Here and in the following, we shall set Boltzmann's constant to 1. In the limit of $\beta\to\infty$, this expression reduces (up to an additive constant) to a "local entropy": it counts the number of minima Γ' of the energy, weighing them (via the parameter γ) by the distance to a reference configuration Γ. For continuous variables, the local entropy becomes the log of a weighted volume around a reference configuration. We can then define the probability distribution

$$P(\Gamma;y,\gamma,\beta)=\frac{1}{Z(y,\gamma,\beta)}e^{yS_{\rm loc}(\Gamma,\gamma,\beta)}, \tag{2}$$

where y determines the degree of concentration of the probability distribution on high local entropy regions. When y is large, only the configurations Γ that are surrounded by an large number of local minima will have non-negligible weight. By increasing the value of γ, it is possible to focus on tighter neighborhoods around Γ, and at large values of γ the target Γ will also share with high probability the properties of the surrounding minima. From an algorithmic perspective we can use the high local entropy probability distribution as a starting point for designing a Markov chain, in the same way that simulated annealing uses the Gibbs measure. One possibility [19,21] is to observe that if we take y to be a non-negative integer we can rewrite the partition function as a product of identical systems connected by a distance constraint

$$\begin{aligned}Z(y,\gamma,\beta)&=\int d\Gamma e^{yS_{\rm loc}(\Gamma,\gamma,\beta)}\\&=\int d\Gamma_c\prod_{a'=1}^{y}d\Gamma_{a'}\;e^{-\beta\sum_{a=1}^{y}U(\Gamma_a)+\gamma\sum_{a=1}^{y}d(\Gamma_a,\Gamma_c)-\beta U(\Gamma_c)},\end{aligned} \tag{3}$$

where Γ_c is the "central" reference configuration, and $\{\Gamma_a\}$ are the configurations of the replicated systems. This partition function describes a system of $y+1$ interacting replicas of the initial system, one of which acts as the reference system, while the other y systems are subject to a distance constraint with respect to the reference system. This gives us a very simple and general scheme to direct algorithms to sample wide minima of the the energy landscape: replicate the model, add an interaction term with a reference configuration, and run the algorithm on the resulting extended system. In practice we need only to consider the effective system

$$U_{\rm eff}(\Gamma_c,\{\Gamma_a\})=\sum_{a=1}^{y}\left[U(\Gamma_a)-\frac{\gamma}{\beta}d(\Gamma_a,\Gamma_c)\right]+U(\Gamma_c) \tag{4}$$

and run our preferred Monte Carlo (MC) Markov chain update. Replicas are initialized at random, while the center configuration is initialized at the average of the replica configurations. It is worth noting that y controls the value of the local entropy and that relatively small values of y are sufficient to obtain the results we are interested in. By taking the inverse temperature of the individual systems β to be large, we focus the sampling on flat ground states.

III. EFFECTS OF THE LOCAL ENTROPY ON THE EQUILIBRIUM CONFORMATIONS OF POLYMERS

We first tested the effect of controlling the local entropy of a polymer on a simple cubic-lattice model, which offers the advantage of a fast sampling of the conformational space of the system and of defining unambiguously the zero of the

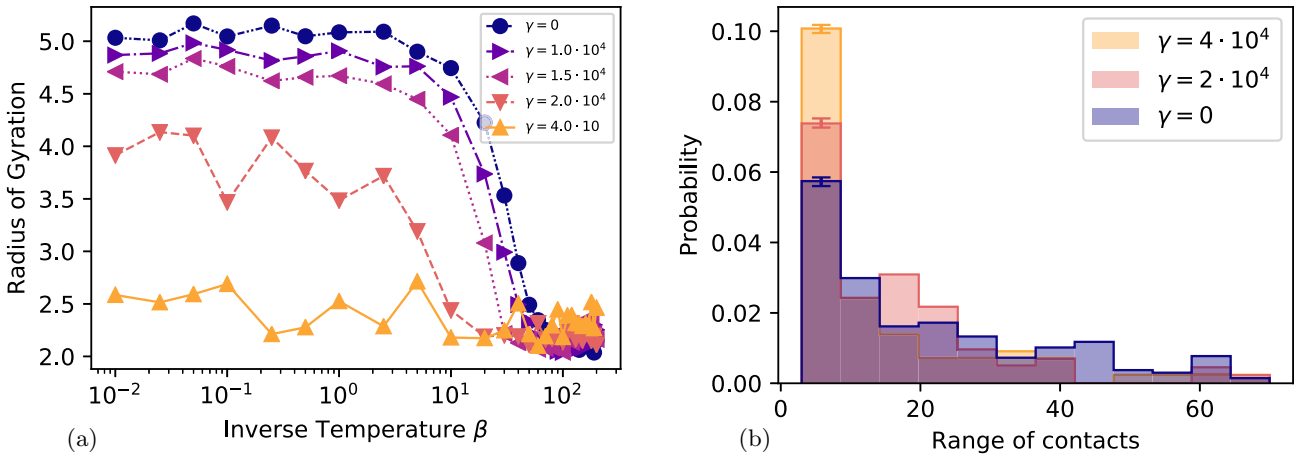


FIG. 1. Local entropy changes the properties of a lattice homopolymer. (a) Radius of gyration as a function of the inverse temperature for increasing values of γ. R_g is dimensionless, since measure lengths in units of the lattice step which is set to 1. The points represent thermal averages. $y = 3 + 1$ replicas were used, the last one being the center. (b) Distribution of the range of contacts $|i - j|$ at $\beta = 100$ for increasing values of γ. Increasing γ suppresses long-range contacts and favors short-range ones. The error bars on the first bin are generated by bootstrapping.

entropy. We employed a standard model on a cubic lattice [8] in which the beads, sitting on the vertices of the lattice, cannot overlap. They interact with a contact potential depending on the conformation $\Gamma = \{\vec{x}_i\}_{i=1}^N$:

$$U(\Gamma) = \frac{1}{2}\sum_{ij=1}^{N} J_{ij}\Delta_{ij}(\Gamma), \tag{5}$$

where the contact function $\Delta_{ij}(\Gamma)$ is 1 if the ith and jth beads are neighbors in space and $|i - j| > 2$ and zero otherwise.

The chain is simulated with a standard Monte Carlo algorithm that includes the corner flip, the crankshaft, and rotations of the ends as elementary moves. At each step, a monomer is chosen at random for it with flat *a priori* probability, and a random move is chosen randomly with uniform probability among those that are possible, accepting it with the standard Metropolis probability. The initial conformation is generated from a random self-avoiding walk in the lattice. The simulations for computing equilibrium quantities consist of 10^7 steps, recording the conformation every 10^4 steps. See Fig. S1 in the Supplemental Material [24] for examples of trajectories that reach equilibrium. Simulations we use to compute the folding time consist of 10^6 steps, recording the conformation every 10^2 steps.

In order to control the local entropy of the system, we performed a Monte Carlo simulations of y replicas of the system starting from independent conformations and interacting with the potential defined in Eq. (4). The Metropolis acceptance rate now reads $p_{\text{accept}} = \min(1, e^{-\Delta})$, where $\Delta = \beta\Delta E + \gamma\Delta d$ and

$$\Delta d = \begin{cases} d\left(\Gamma_a^{\text{new}}, \Gamma_c\right) - d\left(\Gamma_a^{\text{old}}, \Gamma_c\right), & \text{if we are moving a replica } a \\ \frac{1}{y}\sum_y \left[d\left(\Gamma_a, \Gamma_c^{\text{new}}\right) - d\left(\Gamma_a, \Gamma_c^{\text{old}}\right)\right], & \text{if we are moving the center } c. \end{cases} \tag{6}$$

A. Increased local entropy depletes long-range contacts in homopolymers

The simplest polymer model that can be studied is the lattice homopolymer, obtained setting $J_{ij} = -1$ in Eq. (5) for each pair i, j. We compared the results of a standard Monte Carlo sampling of the Boltzmann distribution with a replicated Monte Carlo, which controls the local entropy through the parameter γ, as described in Sec. II. To define the local free entropy we adopt for this system a distance function defined by

$$d(\Gamma_1, \Gamma_2) = 1 - \frac{1}{N_c}\sum_{i<j}^{N} \Delta_{ij}(\Gamma_1)\Delta_{ij}(\Gamma_2), \tag{7}$$

which is the fraction of different contacts between the two conformations Γ_1 and Γ_2; here N_c is the maximum number of contacts that the chain can build. As displayed in Fig. S2 in [24], we have checked that the results are robust with respect to the distance function (e.g., by comparison with root-mean-square distance) and to a different coupling scheme of the replicas. What might be an optimal definition of distance is an interesting problem that goes beyond the scope of our study and that might deserve further study.

In a homopolymer the only relevant equilibrium effect is the coil-globule transition. A comparison between the transition described by a Boltzmann sampling and that obtained controlling the local entropy (cf. Sec. II) is shown in Fig. 1(a) for a polymer with $N = 70$ (see Fig. S3 in [24] for two examples of configurations). In the case of Boltzmann sampling ($\gamma = 0$) the coil-globule transition is marked by a jump of the radius of gyration (R_g) at an inverse temperature $\beta \approx 30$. In the lattice model we measure lengths

in units of the lattice step, which is set to 1 so that R_g is dimensionless. Increasing the bias associated with the local entropy (i.e., by increasing γ) leads to a stabilization of the globule, the transition temperature increasing with γ. At the same time, we observe a compaction of the coil state, associated with the fact that at short distances more compact conformations have certainly a larger number of neighboring other conformations [cf. Eq. (7)] and thus a larger local entropy. The globular phase, in which the polymer is maximally compact, is weakly affected.

In Fig. 1(b) we compared the distributions of contact range (i.e., of the values of $|i-j|$ when i and j are in contact) for fixed inverse temperature $\beta = 100$, at which the system is in the globule phase at all values of γ we simulated. A homopolymeric globule is expected to display an initial decrease of the contact probability up to the distance corresponding to the diameter of the globule, followed by a flat distribution [25]. At $\gamma = 0$ the distribution displays overall the expected character but is quite irregular because of the constraints imposed by the cubic lattice on the R_g of highly compact conformations (i.e., there is a nonmonotonic relation between R_g and energy). Increasing γ, we observed a regular increase of short-range contacts at the expense of long-range contacts.

B. Increased local entropy simultaneously stabilizes and decreases folding time of model proteins

To obtain a clear picture of the effects of the local entropy on the folding properties of model proteins, we calculated the equilibrium stability and the folding rate of a Go model [26] on a lattice. This is defined choosing a target conformation Γ_0 and setting

$$J_{ij} = -J\Delta_{ij}(\Gamma_0) \tag{8}$$

in Eq. (5), where J defines the energy scale and was set to 1. With this choice of J_{ij}, the target conformation is by definition the ground state of the system, and thus the equilibrium state at low temperature. The reason for using a Go model is to decouple the effect of protein sequence from that of protein structure, focusing our attention only on the latter. Here the protein sequence is described effectively, assuming that evolution has minimized energetic frustration to the maximum degree [27].

We chose as target conformation Γ_0 either conformation sampled by the homopolymeric model according to Boltzmann distribution ($\beta = 100$ and $\gamma = 0$) or biasing their local entropy ($\beta = 100$ and $\gamma > 0$), in all cases taking care to select only globular conformations (our choice for the cutoff is $R_g < 2.5$). We thus obtained different potentials that depend on γ through the choice of Γ_0.

We then simulated at low temperature ($\beta = 120$) the dynamics of the system starting from a random, high-temperature, coil state with the standard Metropolis scheme, that at fixed temperature approximates the Smoluchowski equation and thus reports realistic trajectories of the system [28]. From each trajectory we obtained the fraction of native contacts $f_N(t)$ (that for the Go model is $= U[\Gamma(t)]/U(\Gamma_0)$, where $\Gamma(t)$ is the conformation of the chain at time t), we calculated the average $\overline{f_N(t)}$ over 40 simulations and fitted them by a two-state kinetics

$$f_N(t) = f_{\rm eq}(1 - e^{-t/\tau}) + f_0, \tag{9}$$

where τ is the mean folding time, f_0 is the residual fraction of native contacts in the initial conformation, and $f_{\rm eq}$ is the equilibrium similarity to the target conformation, that in the two-state approximation is equal to the equilibrium probability of the native state (cf. Fig. S4 in [24]). We show in Table S1 in [24] the average and standard deviation of the mean square error for each fitted curve.

In Fig. 2(a) we plot the values of τ (in MC steps) and $f_{\rm eq}$ as a function of the parameter γ that controls the local entropy of the target conformation Γ_0. The values are medians over 60 realizations of Γ_0. The corresponding standard deviations are shown in Table S1 in [24]. It is shown that the stability of the native state increases with γ, while the folding time displays a nonmonotonic behavior with a minimum at $\gamma \approx 1.2 \times 10^4$.

To rationalize these results, we estimated what is the radius of the neighborhood of the target conformation that is affected by the increase in local entropy at varying γ. In Fig. 2(b) we plot the average interconformation distance obtained from the replica simulation from which we obtained the conformations Γ_0 [cf. Eq. (4)] as a function of γ. For large γ ($> 1.5 \times 10^4$) the average distance displays a plateau at $d \approx 0.6$; decreasing γ the average distance increases, overcoming 0.9 for plain Monte Carlo simulations. In the plot is also marked the transition state at $d = 0.7$, which separates the denatured state from the native basin.

Both the stabilization and the kinetic effects of the local entropy can be clarified noting what part of the free-energy profile $F(d) = E(d) - TS(d)$ is affected by the local entropy. At large γ the local entropy affects only the close neighborhood of the Γ_0 ($d \lesssim 0.4$), decreasing its free energy and thus stabilizing it with respect to the denatured state, which is unaffected. The transition state is not affected as well, so the folding rate, which according to Kramers' theory depends on the free energy of the transition state calculated with respect to the denatured state, is similar to that at $\gamma = 0$. When γ is decreased, the neighborhood of Γ_0 affected by the increase in local entropy reaches the transition state ($d \simeq 0.7$) and lowers it, decreasing the folding time. The nonmonotonicity of the folding time arises from the fact that making γ even smaller, the denatured state also is affected and then the folding barrier grows again.

The degree of cooperativity of the folding transition decreases slightly with γ (see Table S2 in [24]). This is not unexpected because $\gamma > 0$ leads to the stabilization of conformations with varying distance from the native one, thus decreasing the two-state character of the folding transition.

IV. THE LOCAL ENTROPY OF NATURAL PROTEINS IS LARGER THAN THAT OF RANDOM DECOYS

We then tested the hypothesis that the native state of natural proteins displays a larger local entropy than random polypeptidic conformations with the same density. We studied seven natural proteins and a stable protein designed *de novo* (HHH) [2].

To estimate the local entropy associated with the native state we need to describe the energy of the protein in the

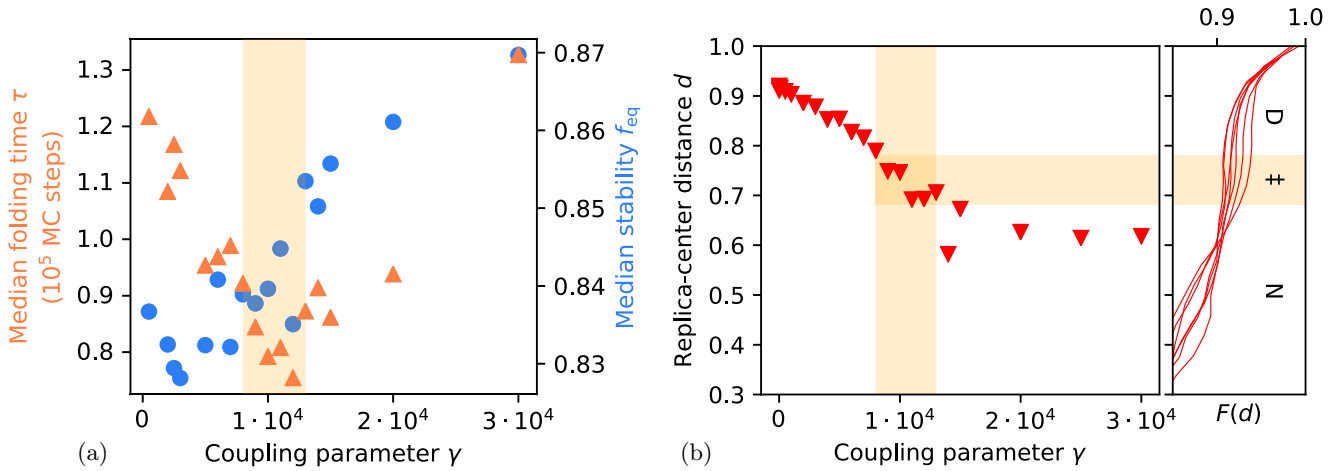


FIG. 2. Structures with high local entropy are more stable and fold faster. (a) The average folding time (in MC steps, orange triangles, left y-axis) for target conformations Γ_0 obtained at different values of γ and the average stability of the structure (blue circles, right y-axis). The orange bar indicates the transition state ‡ calculated in the right panel. (b) The average distance between the central conformation and the replicas at different values of γ obtained from the simulation used to generate the Γ_0. Aligned with the right axis we show the free energy of a Go model as function of the distance from the native state for various structures at low γ. Both the distance and the free energy are dimensionless given the definition of the model. We can see that the region around $\gamma = 1 \times 10^4$, marked with an orange bar, corresponds to the typical distance between native and transition states.

neighborhood of the crystallographic structure. For this purpose we made use of an all-atom Go model [26], which is expected to be particularly realistic in the native basin. At the same time, it allows one to decouple the effect of the sequence from that of the native conformation in the calculation of the local entropy.

In order to evaluate the local entropy, we performed molecular dynamics (MD) simulations with Gromacs 2020.4 [29] using the all-atom Go model obtained by Smog2 [30]. The native conformations of the proteins were the following: protein G (pdb code 1pgb, 56 residues), ACBP (pdb code 2abd, 87 residues), CI2 (pdb code 2ci2, 83 residues), src-SH3 (pdb code 1srl, 64 residues), villin headpiece (pdb code 5vnt, 63 residues), barnase (pdb code 1bnr, 110 residues), and HHH (artificial protein, pdb code 5uoi, 43 residues). MD simulations were performed in the range of temperature from 1 to 200 (in energy units) for 2×10^5 steps of time step 5×10^{-4} ps with stochastic dynamics [29]. The Go potential of each decoy, whose parameters in Smog2 depend on the number of residues and of native contacts in the native conformation, is rescaled to that of the corresponding native protein in order to facilitate the comparison among them. The microcanonic entropy $S(E)$ is extracted from all simulations performed at different temperatures for the same protein with the maximum-likelihood code developed in Ref. [31].

As a control model we generated putative native conformations from a homopolymeric model, derived from the original models as follows. Starting from the crystallographic structure of each protein, we generated an all-atom model similar to that described above (i.e., with the same atomic structure), but where each pair of atoms interacts instead in the same way with the Lennard-Jones potential

$$V_{\mathrm{LJ}}(r_{ij}) = C_{ij}^{(12)}/r_{ij}^{12} - C_{ij}^{(6)}/r_{ij}^{6}, \tag{10}$$

where for all i, j we set $C_{ij}^{(6)} = 1.4 \times 10^{-2}\,\mathrm{kJ\,mol^{-1}\,nm^6}$ and $C_{ij}^{(12)} = 1.0 \times 10^{-4}\,\mathrm{kJ\,mol^{-1}\,nm^6}$. No potential is applied to the dihedrals at this stage. The parameters $C^{(6)}$ and $C^{(12)}$ are chosen with a grid search so that, after an annealing MD of the chain, the number of contacts in the putative conformations is not smaller than in the original model (see Fig. S5 in [24]) to rule out trivial effects in the calculation of the local entropy. The resulting conformation is used as putative Γ_0 (examples of these configurations can be found in Fig. S6 in [24]).

The entropy $S(E)$ for four proteins is displayed in Figs. 3(c)–(f) (see also Fig. S7 in [24]) and compared with the random decoys of each of them, matching the different curves at infinite temperature (cf. the insets). In most cases, proteins display in the native energy region (below the transition state) an entropy that is larger than that of the random decoys. This effect is summarized in Fig. 3(a), which displays the density of local entropy

$$s_{\mathrm{loc}} = \frac{1}{N} \log \sum_{E=E_{\mathrm{N}}}^{E_\ddagger} e^{-\beta E + S(E)}, \tag{11}$$

where E_{N} is the minimum energy of the system and $E_\ddagger$ is the energy of the transition state, approximated as the average energy at the transition temperature (cf. the caption of Table S3 in [24]). The local entropy is calculated at low temperature ($\beta = 10^{-1}$) at which all proteins and decoys are stable (cf. Fig. S8 in [24]). Equation (11) is the microcanonical counterpart of Eq. (1). The density of local entropy of native proteins is always larger than the average $\bar{s}$ of the decoys, in five cases out of eight for more than one standard deviation σ_s. From $\bar{s}$ and σ_s we estimated the P values associated with the null hypothesis that the entropy density of the native protein is smaller than that of the decoys within a Gaussian approximation, that, is $p = \{1 - \mathrm{erf}([s - \bar{s})/\sqrt{2}\sigma_s]\}/2$, where erf is the error func-

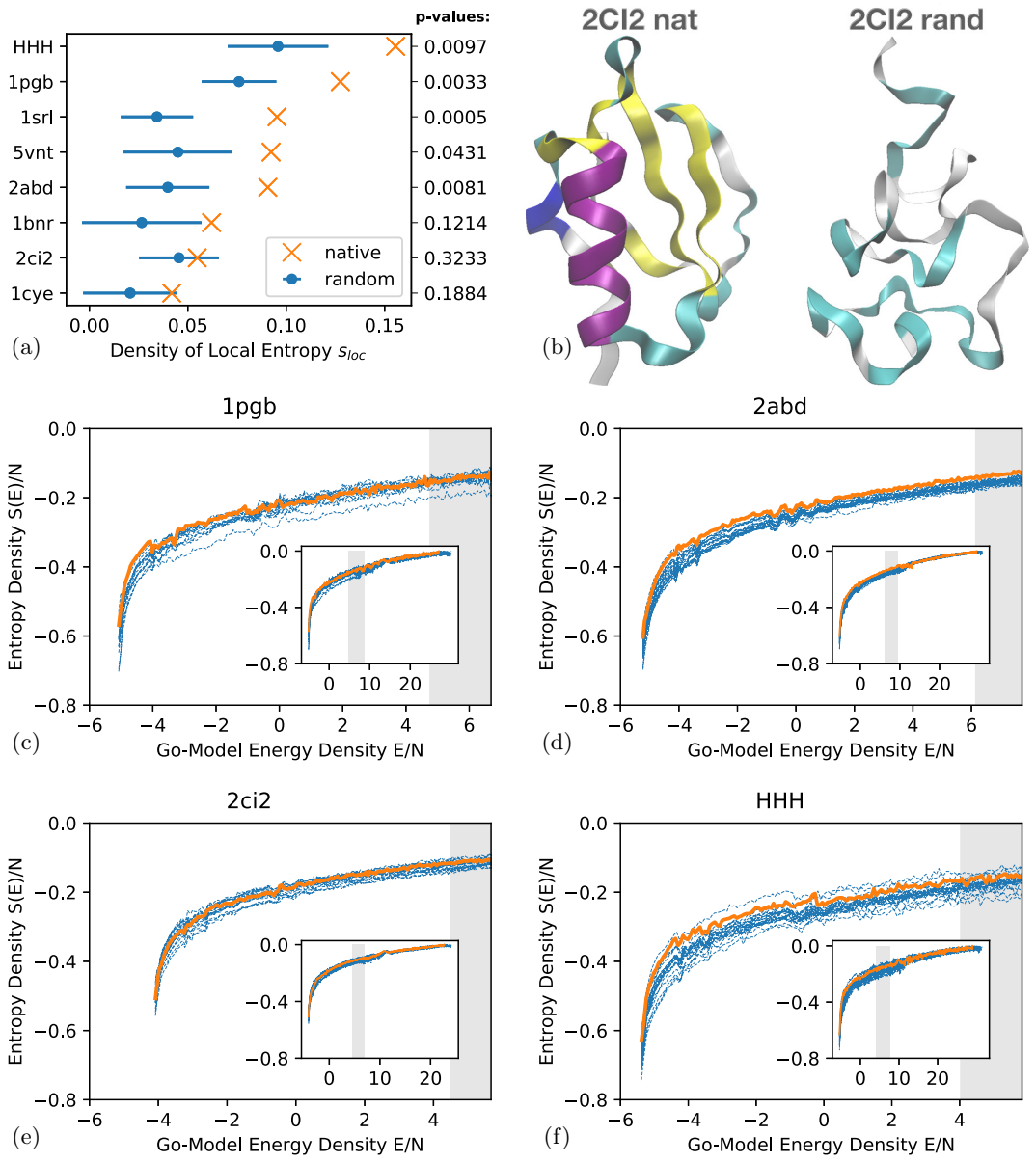


FIG. 3. Natural structures have higher local entropy than random ones at low energy. (a) The density of local entropy of the native state (orange crosses), the average density of local entropies of the decoys (blue circles) and the associated standard deviation (blue bars) for each protein. On the right axis are shown the associated P values. (b) The native conformation of CI2 (2ci2) and an example of decoy. (c–e) The entropy density calculated as a function of energy for four proteins (solid orange curves) and for the associated decoys (dashed blue curves). The gray vertical bars indicate the region where the transition states are (the band is centered on the average over the protein and the decoys, its width is twice a standard deviation). In the inset is displayed the whole entropy density, whereas in the main figures only the region of the native state. Note the energy density has been rescaled for each structure as described in Sec. IV, so the energy units are arbitrary.

tion. The P values are rather low, except for CI2 (2ci2) and for the CHE-Y (1cye).

The protein HHH, designed *de novo* and not shaped by natural evolution, seems to display the same features of natural proteins. However, one has to consider that the scaffold they used for the design is a helix bundle typical of natural proteins.

Finally (cf. Fig. S8 in [24]), we observed that the peaks in the specific heat of the four proteins tend to be wider and less pronounced, in agreement with what we found in Sec. III B: the wider the specific heat, the less cooperative the transition [32] (a measure of cooperativity of the transition can be found in Table S2 in [24]). At the same time, the majority of native proteins display a larger folding temperature (cf. the main peak in the specific heat in Fig. S8 in [24]), in agreement with the fact that a larger local entropy stabilizes the native state.

V. THE LOCAL ENTROPY DEPENDS ON THE TOPOLOGY OF CONTACTS

Finally, one can investigate whether there is an elementary way to reinterpret why natural proteins display a larger local entropy than random conformations. This aspect can be stud-

ied easily in the context of a Go model in which the frustration associated with the sequence is minimized [27].

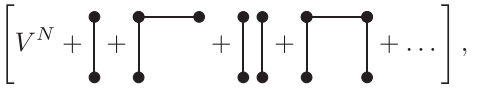
In fact, in this case one can insert in the definition of local entropy of Eq. (1) the expressions of Eqs. (5), (7), and (8), obtaining

$$S_{loc}(\Gamma_0) = \log \int d\Gamma \exp\left[\left(\beta J + \frac{\gamma}{N_c}\right)\sum_{i<j}^{N} \Delta_{ij}(\Gamma_0)\Delta_{ij}(\Gamma)\right], \tag{12}$$

where $d\Gamma \equiv dx_1 dx_2 \cdots$ and an immaterial constant has been disregarded. This is essentially the free energy of the Go model at a rescaled temperature. The exponential in the integrand can be expanded in series:

$$\begin{aligned} e^{\cdots} = 1 &+ \left(\beta J + \frac{\gamma}{N_c}\right)\sum_{i<j}^{N} \Delta_{ij}(\Gamma_0)\Delta_{ij}(\Gamma) \\ &+ \frac{1}{2}\left(\beta J + \frac{\gamma}{N_c}\right)^2 \sum_{\substack{i<j \\ k<l}}^{N} \Delta_{ij}(\Gamma_0)\Delta_{kl}(\Gamma_0)\Delta_{ij}(\Gamma)\Delta_{kl}(\Gamma) \\ &+ \cdots . \end{aligned} \tag{13}$$

Moving the $\Delta_{ij}(\Gamma_0)$ out of the integral and defining the interaction volume as $v \equiv \int d\Gamma \Delta_{ij}(\Gamma)$ and the total volume available to each degree of freedom as V, one obtains

$S_{\rm loc}(\Gamma_0) = \log$  , (14)

where the graphs indicate the terms of the expansions. For example,

 $= \left(\beta J + \frac{\gamma}{N_c}\right) V^{N-1} v \sum_{i<j}^{N} \Delta_{ij}(\Gamma_0)$ (15)

is the contribution of a single contact, that depends on the number $\sum \Delta_{ij}(\Gamma_0)$ of contacts of the native conformation. Thus, the larger is the number of contacts on the native conformation, the larger is the local entropy. Similarly,

 $= \frac{1}{2}\left(\beta J + \frac{\gamma}{N_c}\right)^2 V^{N-2} v^2 \sum_{i<j<k}^{N} \Delta_{ij}(\Gamma_0)\Delta_{jk}(\Gamma_0),$ (16)

where $\sum \Delta_{ij}(\Gamma_0)\Delta_{jk}(\Gamma_0)$ is the number of triples of nodes (e.g., amino acids) interacting pairwise. For a graph with loops, for example,

 $= \frac{1}{3!}\left(\beta J + \frac{\gamma}{N_c}\right)^3 V^{N-2} A_{\triangleright} v^2 \sum_{i<j<k}^{N} \Delta_{ij}(\Gamma_0)\Delta_{jk}(\Gamma_0)\Delta_{ki}(\Gamma_0),$ (17)

where $A_{\triangleright} = 3/4$ is a parameter that arise in the integration of the contact functions Δ from the constraints given by looped graphs. In fact,

$$\int dx_i\, dx_j\, dx_k \Delta(|x_i - x_j|)\Delta(|x_j - x_k|)\Delta(|x_k - x_i|) = V\frac{3}{4}v^2. \tag{18}$$

For a generic graph, $A \leqslant 1$ and is equal to the unity if the graph does not contain loops because the variables can be integrated sequentially. For a fully connected graph of n nodes, that in the language of network theory is called a clique, $A = n/2^{n-1}$.

By the linked cluster theorem, the sum of graphs in Eq. (13) is the exponential of the sum of connected graphs, so the local entropy results simply the sum of connected graphs. Note that the connected graphs are different from zero only if the associated $\sum \Delta_{ij}(\Gamma_0)\Delta_{jk}(\Gamma_0)\cdots$ is different from zero, that is, the corresponding structure is in the native protein. The goal is to spot the most important graphs. The general form of a graph is

$$\frac{A}{l!}\left(\beta J + \frac{\gamma}{N_c}\right)^l V^{N-n+1} v^{n-1} \times (\text{no. of instances in } \Gamma_0), \tag{19}$$

where n is the number of interacting nodes, l is the number of links (i.e., interactions between nodes), and A depends on how links loop together. Each term is also proportional to the number of instances that the specific graph appears in the native conformation of the protein.

Proteins will display a large local entropy if they are rich with graphs with large values of

$$w \equiv \frac{A}{l!} B^l V^{N-n+1} v^{n-1}, \tag{20}$$

where we defined $B \equiv \beta J + \gamma/N_c$. Thus, local entropy depends on the topology of native contacts. For unlooped graphs ($A = 1$), for fixed n, w has a maximum at $l^* = B$. Typically, for proteins βJ is of the order of 1; γ/N_c for the model of Fig. 2(b) is of the order of $10^4/10^2 = 10^2$, and consequently $l^* \sim 10^2$.

However, for graphs composed of l^* links, fully connected graphs display a larger weight w. In fact, if we compare the value of $w_{\rm clique}$ for a fully connected graph ($l \sim n^2$) with that $w_{\rm unloop}$ of an unlooped graph ($l \sim n$) at $l = B$, one obtains $w_{\rm clique}/w_{\rm unloop} = (V^N/v)^{B/2} B^{-1/2}$, which is large for $B \gg 1$.

It was shown [11] that $\mathrm{Tr}\,[\Delta_{ij}(\Gamma_0)]^n$ for $n \gg 1$ is a good determinant of protein designability. Since this quantity counts the number of closed paths of length n in the interaction network of the protein, it will be correlated with the number of clusters with large w; consequently, the tendency of evolution to maximize $\mathrm{Tr}\,[\Delta_{ij}(\Gamma_0)]^8$ [12] is tantamount to the optimization of the local entropy. Notice that for the lattice-model proteins discussed above, the average values of $\mathrm{Tr}\,[\Delta_{ij}(\Gamma_0)]^n$ with $n = 4, 8$ are increasing with γ, that is with the bias towards having high entropy at short distances (cf. Table S4 in [24]).

VI. DISCUSSION AND CONCLUSIONS

The concept of local entropy was used to explain the properties of complex systems like artificial neural networks [18,19], in which the "energy" landscape is very rough and the system learning dynamics leads to atypical configurations with respect to the Gibbs measure. In particular, states characterized by a large local entropy can be easily accessed in spite of the abundance of competing states that tend to block the learning procedure.

In the present work we claimed that the concept of local entropy plays an important role also in systems that can move in a smoother energy landscape and thus are not in a glassy regime. In particular, we focused our attention on the three-dimensional conformation of proteins, which evolved along the eons through a complex dynamics described by Darwinian evolution. The high diversity of native conformations of known proteins and their redundancy (i.e., the existence of analogous proteins) suggests that evolution explored a large part of conformational space and thus that protein conformations can be regarded as stationary realizations of the evolutive dynamics.

In the present work, we made use of a Go model to estimate the local entropy of proteins, in order to separate the evolutionary problem of sequence design from that of conformational selection, removing in this way frustration from the system [33]. Although it is known [34] that frustration plays an important role in determining the features of the denatured state up to the formation of the transition state, one expects it to be less relevant to determine the local properties of the native state and thus the estimation of its local entropy, in agreement with the principle of minimal frustration of the native state [27]. The Go model is thus particularly suitable for the specific problem we face.

A relevant question is then whether the ensemble of native conformations can be described by Boltzmann statistics or the evolutionary accessibility of native conformations is important to define the fitness of proteins. The latter case would imply that out-of-equilibrium effects affect the set of existing protein conformations. In the cases we studied, the local entropy of proteins is indeed larger than that of random decoys displaying the same length and density. As already mentioned in the introduction, these ideas are not new. However, our aim was to provide a simple proof of concept with a technique that can be generalized to more realistic settings for protein design. While previous estimates of the local entropy are either qualitative [16] or based on a knowledge-based, empirical function [17], the methods we suggest inspired by the study of complex systems [18] make use of a direct calculation of the local entropy. Being based on simple statistical-mechanical concepts, we expect this method to be widely applicable to different class of models.

In fact, we showed that the large local entropy of model proteins has consequences on both their thermodynamic and equilibrium properties. It straightforwardly stabilizes the native state, decreasing its free energy, but extending to the transition state can also improve the folding rate. From this point of view, the concept of evolutionary optimization of the local entropy is related to the dynamical variational principle stated in Ref. [6]. In fact, the mechanistic reason for the increase in the folding rate also is the same: local entropy biases the stabilizing contacts to be closer along the sequences, making their formation kinetically faster [7].

Although this point is not within the scope of the present work, we note that the results we discuss could be used in connection with the recent advancements in the understanding of the relation between protein sequence and structure [22,23] to improve the design of new proteins.

The scripts and the main data used and produced in this work can bee freely downloaded at [35].

[1] A. V. Finkelstein, A. M. Gutun, and A. Ya. Badretdinov, Why are the same protein folds used to perform different functions? FEBS Lett. **325**, 23 (1993).

[2] A. Chevalier, D.-A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K.-H. Lam, G. Yao *et al.*, Massively parallel de *novo protein* design for targeted therapeutics, Nature (London) **550**, 74 (2017).

[3] H. Li, R. Helling, C. Tang, and N. Wingreen, Emergence of preferred structures in a simple model of protein folding, Science **273**, 666 (1996).

[4] S. Govindarajan and R. A. Goldstein, Why are some proteins structures so common? Proc. Natl. Acad. Sci. USA **93**, 3341 (1996).

[5] P. G. Wolynes, Symmetry and the energy landscapes of biomolecules, Proc. Natl. Acad. Sci. USA **93**, 14249 (1996).

[6] A. Maritan, C. Micheletti, and J. R. Banavar, Role of Secondary Motifs in Fast Folding Polymers: A Dynamical Variational Principle, Phys. Rev. Lett. **84**, 3009 (2000).

[7] K. W. Plaxco, K. T. Simons, and D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins, J. Mol. Biol. **277**, 985 (1998).

[8] E. Shakhnovich, Proteins with Selected Sequences Fold Into Unique Native Conformation, Phys. Rev. Lett. **72**, 3907 (1994).

[9] R. A. Broglia, G. Tiana, H. E. Roman, E. Vigezzi, and E. Shakhnovich, Stability of Designed Proteins Against Mutations, Phys. Rev. Lett. **82**, 4727 (1999).

[10] G. Tiana, R. A. Broglia, and D. Provasi, Designability of lattice model heteropolymers, Phys. Rev. E **64**, 011904 (2001).

[11] J. L. England and E. I. Shakhnovich, Structural Determinant of Protein Designability, Phys. Rev. Lett. **90**, 218101 (2003).

[12] G. Tiana, B. E. Shakhnovich, N. V. Dokholyan, and E. I. Shakhnovich, Imprint of evolution on protein structures, Proc. Natl. Acad. Sci. USA **101**, 2846 (2004).

[13] I. N. Berezovsky, A. Y. Grosberg, and E. N. Trifonov, Closed loops of nearly standard size: Common basic element of protein structure, FEBS Lett. **466**, 283 (2000).

[14] M. Karplus, T. Ichiye, and B. M. Pettitt, Configurational entropy of native proteins, Biophys. J. **52**, 1083 (1987).

[15] H. Frauenfelder, F. Parak, and R. D. Young, Conformational substates in proteins, Annu. Rev. Biophys. Biophys. Chem. **17**, 451 (1988).

[16] D. Shortle, K. T. Simons, and D. Baker, Clustering of low-energy conformations near the native structures of small proteins, Proc. Natl. Acad. Sci. USA **95**, 11158 (1998).

[17] K. Sankar, K. Jia, and R. L. Jernigan, Knowledge-based entropies improve the identification of native protein structures, Proc. Natl. Acad. Sci. USA **114**, 2928 (2017).

[18] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses, Phys. Rev. Lett. **115**, 128101 (2015).

[19] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes, Proc. Natl. Acad. Sci. USA **113**, E7655 (2016).

[20] C. Baldassi, F. Pittorino, and R. Zecchina, Shaping the learning landscape in neural networks around wide flat minima, Proc. Natl. Acad. Sci. USA **117**, 161 (2020).

[21] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Local entropy as a measure for sampling solutions in constraint satisfaction problems, J. Stat. Mech.: Theory Exp. (2016) P023301.

[22] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with AlphaFold, Nature (London) **596**, 583 (2021).

[23] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network, Science **373**, 871 (2021).

[24] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.104.064117 for additional details of the analysis and for examples not shown in the main text.

[25] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).

[26] N. Go, Theoretical studies of protein folding, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).

[27] J. D. Bryngelson and P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).

[28] G. Tiana, L. Sutto, and R. A. Broglia, Use of the Metropolis algorithm to simulate the dynamics of protein chains, Phys. A Stat. Mech. Appl. **380**, 241 (2007).

[29] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX **1**, 19 (2015).

[30] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford, Smog 2: A versatile software package for generating structure-based models, PLoS Comput. Biol. **12**, e1004794 (2016).

[31] G. Tiana and L. Sutto, Equilibrium properties of realistic random heteropolymers and their relevance for globular and naturally unfolded proteins, Phys. Rev. E **84**, 061910 (2011).

[32] P. L. Privalov and N. N. Khechinashvili, A Thermodynamic Approach to the problem of stabilization of globular protein structure: A calorimetric study, J. Mol. Biol. **86**, 665 (1974).

[33] P. Das, S. Matysiak, and C. Clementi, Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes, Proc. Natl. Acad. Sci. USA **102**, 10141 (2005).

[34] E. Paci, M. Vendruscolo, C. M. Dobson, and M. Karplus, Determination of a transition state at atomic resolution from protein engineering data, J. Mol. Biol. **324**, 151 (2002).

[35] https://gitlab.com/bocconi-artlab/local-entropy-proteins.