Topological origin of protein folding transition

Loris Di Cairano*
Computational Biomedicine, Institute for Advanced Simulation IAS-5, and Institute of Neuroscience and Medicine INM-9, Forschungszentrum Jülich, 52425 Jülich, Germany and Physics and Materials Science Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg

Riccardo Capelli†
Department of Biosciences, Università degli Studi di Milano, Via Celoria 26, 20133 Milano, Italy

Ghofrane Bel-Hadj-Aissa‡
University of Siena, Via Roma 56, 53100 Siena, Italy Aix-Marseille Univ, Université de Toulon, CNRS, France and Centre de Physique Théorique, UMR 7332, Marseille, France

Marco Pettini§
Aix-Marseille Univ, Université de Toulon, CNRS, France and Centre de Physique Théorique, UMR 7332, Marseille, France

In this paper, a geometrical and thermodynamical analysis of the global properties of the potential energy landscape of a minimalistic model of a polypeptide is presented. The global geometry of the potential energy landscape is supposed to contain relevant information about the properties of a given sequence of amino acids, that is, to discriminate between a random heteropolymer and a protein. By considering the SH3 and PYP protein-sequences and their randomized versions it turns out that in addition to the standard signatures of the folding transition - discriminating between protein sequences of amino acids and random heteropolymer sequences - also peculiar geometric signatures of the equipotential hypersurfaces in configuration space can discriminate between proteins and random heteropolymers. Interestingly, these geometric signatures are the "shadows" of deeper topological changes that take place in correspondence with the protein folding transition. The protein folding transition takes place in systems with a small number of degrees of freedom (very far from the Avogadro number) and in the absence of a symmetry-breaking phenomenon. Nevertheless, seen from the deepest level of topology changes of equipotential submanifolds of phase space, the protein folding transition fully qualifies as a phase transition.

PACS numbers: 87.15.-v; 02.40.-k

I. INTRODUCTION

The study of the Hamiltonian dynamical counterpart of phase transitions (PTs) combined with the geometrization of Hamiltonian dynamics (where the natural motions are identified with geodesics of suitable Riemannian manifolds) led to find that at the roots of the PTs phenomena there are some peculiar changes of the topology of certain submanifolds of phase space. More precisely, the relevant mathematical objects [1] are the potential level sets (PLSs) $\Sigma_v^{V_N} := \{V_N(q_1,\dots,q_N) = v \in \mathbb{R}\}$ in configuration space, and, equivalently, the balls $\{M_v^{V_N} = V_N^{-1}((-\infty, v])\}_{v\in\mathbb{R}}$ bounded by the $\Sigma_v^{V_N}$. Both geometry and topology of these objects can affect microscopic dynamics and macroscopic thermodynamics of the modeled physical system. In fact, when the ball $M_{v=E}^{V_N} = \{(q_1,\dots,q_N) \in \mathbb{R} | V_N(q_1,\dots,q_N) < E\}$ is endowed with the metric tensor $g_J = 2[E - V(q)]dq^i \otimes dq^k$, then its geodesics are the natural motions given by $\ddot{q}^i = -\nabla^i V(q)$, and the geometry of the manifold $(M_E^{V_N}, g_J)$ determines the properties of order and chaos of the microscopic dynamics [1–4]. On the other hand, a relationship also exists between macroscopic thermodynamics and the topology of the same objects, $M_v^{V_N}$ and $\Sigma_v^{V_N}$. For the latter objects this relationship is expressed by [1]

$$S(v) = \frac{k_B}{N}\log\frac{1}{N!}\int_{\Sigma_v^{V_N}}\frac{d\sigma}{\|\nabla V\|} \approx \frac{k_{\rm B}}{N}\log\left[\mathrm{vol}(\mathbb{S}_1^{N-1})\sum_{i=0}^{N} b_i(\Sigma_v^{V_N}) + \int_{\Sigma_v^{V_N}} d\sigma\frac{\widetilde{t}(v)}{N!}\right] + \frac{1}{N}\log R(v)\ , \qquad (1)$$

* loris.dicairano@uni.lu
† riccardo.capelli@unimi.it
‡ ghofrane.belhadjaissa@gmail.com
§ marco.pettini@cpt.univ-mrs.fr

where S is the configurational entropy, v is the potential energy, and the $b_i(\Sigma_v^{V_N})$ are the Betti numbers (in one-to-one correspondence with topology) of the manifolds $\Sigma_v^{V_N}$; thus in square brackets the first term is of topological meaning and the second term is a smooth function of v as well as the term $R(v)$. On the basis of Eq.(1) one can infer that major topology changes with v of the submanifolds $\Sigma_v^{V_N}$, associated with sharp changes of the potential energy pattern of at least some of the $b_i(\Sigma_v^{V_N})$, can affect the v-dependence of the entropy $S_N(v)$ and thus of its derivatives.

Therefore, at least for a broad class of physical systems, it has been hypothesized that PTs stem from a suitable change of the topology of the PLSs, $\Sigma_v^{V_N}$, and, equivalently, of the manifolds $M_v^{V_N}$, when v, playing the role of the control parameter, crosses a critical value v_c. This hypothesis is at the ground of a theoretical framework composed of exactly solvable models [1, 5] and two theorems [6–9] stating that equilibrium PTs are *necessarily* induced by suitable topological transitions in configuration space. Actually, the main advantage of the topological approach is to provide a deeper insight than those proposed so far in the literature. More specifically, after Landau theory [10, 11], the emergence of PTs has been associated to a symmetry-breaking mechanism. However, there are many examples of systems undergoing PTs in the absence of a symmetry breaking and thus lacking an order parameter. Some relevant examples are: Kosterlitz–Thouless transitions (after the Mermin–Wagner theorem) [12], systems with local gauge symmetries (after Elitzur theorem) [13], liquid–gas transitions, transitions in supercooled glasses and liquids, transitions in amorphous and disordered systems, folding transitions in homopolymers and proteins. Another important limitation of the existing theories consists in the difficulty of providing a coherent definition of PTs in small-size systems, that is, very far from the thermodynamic limit. In fact, the difficulty is due to the so-called thermodynamic limit dogma, a consequence of the Yang-Lee theorems [14, 15] showing that the loss-of-analyticity of thermodynamic observables - which characterizes PTs - is only possible in the thermodynamic limit ($N \to \infty$).

Therefore, transitional phenomena in systems with a fixed number of constituents, *i.e.,* intrinsically lacking a thermodynamic limit, can hardly be given a definition consistent with the one for the systems admitting a $N \to \infty$ limit. This is for example the case of filament to globule transition in homopolymers and of the protein folding transition. Only recently, M. Bachmann proposed a consistent and powerful definition of PTs in the microcanonical ensemble for finite-size systems [16–18]. A complementary microcanonical classification of PTs for systems admitting the thermodynamic limit has been proposed in Refs. [13, 19, 20]. These classifications are very useful tools for investigating the thermodynamic properties of systems in the microcanonical ensemble independently of the size of a system. However these are essentially phenomenological approaches seemingly calling for further explanatory steps. In this context, the topological approach aims at providing a possible explanatory step forward also on this point. In fact, phase transitions that are experimentally observed in finite/small systems are theoretically at odds with the thermodynamic limit dogma, but while thermodynamic observables cannot display non-analytic energy, or temperature, patterns at finite N, this is not true from the topological viewpoint. This is well evident in the case of systems for which unequivocally sharp signatures of a phase transition are displayed by an analytically computed topological invariant (the Euler characteristic), as in the case of the XY-mf model [21] and of the p-trig model [22].

The aim of the present work is twofold. On the one side, we aim at applying the topological approach to PTs occurring in systems with a constitutively small number of degrees of freedom, that is, much smaller than the Avogadro number. In fact, PTs are *experimentally* observed also in nanoscopic and mesoscopic systems, that is, at very small numbers of degrees of freedom, a circumstance which is *theoretically* at odds with the thermodynamic limit dogma stemming from the Yang-Lee theory. A representative example of PT in a small N system is the protein folding transition. Therefore this is a reason to tackle it as a tested for the topological description of the origin of PTs in the case of a small number of degrees of freedom.
On the other side protein folding is a very important and challenging open question in molecular biology, another reason for applying to this phenomenon the new approach. Even though the present work has no pretense to contribute yet the protein folding problem with significant advancement - given also the simplistic model used - the way of looking at the protein folding transition proposed in our prospective work could provide an interesting complementary method to existing ones, worthy of further attention.

The well-known Anfinsen's dogma [23] states that for small globular proteins the sequence of amino acids uniquely determines the native state (*i.e.*, the compact configuration the protein assumes in physiological conditions). For this reason, understanding how the information contained in the sequence is translated into the three-dimensional native structure is at the core of the protein folding problem. All the naturally selected proteins generally fold to a uniquely determined native state, but a generic polypeptide does not, and is considered a random heteropolymer.

Following the line of Refs. [24, 25], instead of linking the folding properties to the energy landscape by locating the energy minima and the saddles joining them, or by undertaking the folding funnel approach [26], we focus on global properties of the energy landscape which can be easily numerically computed through time averages along dynamical trajectories.

In Section II we define the simplified model adopted to describe the protein dynamics and provide information about the numerical simulations carried on for two

different proteins. In Section III we specify the kind of observables computed by means of Molecular Dynamics (MD) simulations. In Sections IV and V we discuss why the signatures of the folding transition detected via geometrical observables probe deeper topological changes of submanifolds of configuration space. In Sections VI and VII the results of numerical simulations and their meaning are discussed.

II. DEFINITION OF THE MODEL AND MD CALCULATIONS

Two different proteins have been considered in this work: SH3 and PYP. For both proteins we generated a C_α-based Gō-model [27] via the SMOG2 [28] implementation, starting from the experimental structures obtained from the Protein Data Bank (1FMK [29] for SH3 and 3PHY [30] for PYP). In this model, only the C_α atom of every amino acid is considered and the model potential is given by

$$U(\Gamma,\Gamma_0)=\sum_{\text{bonds}} K_r(r-r_0)^2+\sum_{\text{angles}} K_\theta(\theta-\theta_0)^2+\sum_{\text{dihedrals}} K_\varphi^{(n)}(1+\cos(n(\varphi-\varphi_0)))+ \\ +\sum_{i<j-3}\varepsilon_{ij}^{\text{native}}\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right]+\sum_{i<j}\varepsilon^{\text{n-nat}}\left(\frac{\sigma_{nn}}{r_{ij}}\right)^{12} \tag{2}$$

where Γ_0 is the initial experimental structure, and Γ is the current system conformation; similarly, r_0, θ_0, and φ_0 are the reference values for all the bonds, angles and dihedrals in the model, while r, θ, and φ are their value in the conformation Γ. In our implementation the dihedral potential is a sum of 2 terms for every 4 adjacent C_α atoms, with periods $n=1$ and $n=3$. The force constants for bonded interactions in our implementation are $K_r=200\varepsilon/\text{Å}^2$, $K_\theta=40\varepsilon/\text{rad}^2$, $K_\varphi=\varepsilon$, and $\varepsilon=1$ kJ/mol. In non-bonded interaction, native contacts are defined as all the C_α pairs that have a mutual distance smaller than a threshold (here defined as 10 Å) in the reference configuration Γ_0, and a distance along the chain of 3 amino acids. All the pairs that do not satisfy these conditions are considered as non-native contacts and their interaction is given only by a repulsive term [last term in Eq. (2)]. σ_{ij} is chosen so that the minimum of the potential is at the distance r_{ij} measured in the reference conformation Γ_0, while $\sigma_{nn}=4$Å. Energy terms for non-bonded interaction are $\varepsilon_{ij}^{\text{native}}=\varepsilon$ and $\varepsilon^{\text{n-nat}}=\varepsilon$.

We emphasize here that we chose two small globular proteins consisting of a single domain to obtain with a Gō-like model a satisfactory approximation of their folding transition. For more complex systems (disordered and/or multi-domain proteins) a finer-grained and more accurate representation would be needed.

To compare this protein-like model with a polymer model that does not have a well-defined folding minimum, we generated 2 random heteropolymer models starting from the initial Gō models. We removed from the original potential almost all the bonded interaction (keeping only the bonds between the residues), and we scrambled the non-bonded interaction matrices, namely

$$U_{\text{RMD}}(\Gamma,\Gamma_0)=\sum_{\text{bonds}} K_r(r-r_0)^2+\sum_{i<j-3}\tilde{\varepsilon}_{ij}^{\text{native}}\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right]+\sum_{i<j}\varepsilon^{\text{n-nat}}\left(\frac{\sigma_{nn}}{r_{ij}}\right)^{12} \tag{3}$$

where $\tilde{\varepsilon}_{ij}^{\text{native}}$ is the scrambled interaction matrix.

We named the 2 systems obtained from the initial SH3 and PYP models RMDa and RMDb, respectively.

All the molecular dynamics simulations were then performed using GROMACS [31] version 2019.6 (compiled in double precision), with a Langevin integrator, with $\gamma=1$ ps^{-1}, and a time step of 0.5 fs. We initially performed a short equilibration run (10 ns) to relax and thermalize the structure at the target temperature. After this initial equilibration, we performed a 100 ns-long simulation with the same parameters. To exhaustively explore the folding curve, we performed a large number of simulations at different temperatures (note that in a Gō model energy units, and consequently temperature units, are arbitrary), namely:

- For SH3 we performed 1 simulation every 0.25 K between 135 and 161 K; every 1 K between 75 K and 135 K and from 161 to 200 K; and every 2 K

from 200 K to 250 K for a total of 229 simulations.

- For PYP we performed 1 simulation every 0.25 K between 145 and 160 K; every 1 K between 75 K and 145 K and from 160 to 200 K; and every 2 K from 200 K to 250 K for a total of 196 simulations.

- For the 2 random energy models, we performed 1 simulation every 5 K from 75 K to 250 K, for a total of 36 simulations.

From these production runs we computed the gyration radius using PLUMED 2.5 [32, 33], and all the other observables needed using the GROMACS suite. From the potential energies at different temperatures we computed the system heat capacity (C_v) with a multiple histogram method [34].

III. GEOMETRICAL SIGNATURES OF TOPOLOGICAL CHANGES

In order to get information on the topology of the manifolds of interest one has to resort to theorems in differential topology relating total geometric quantities of a given manifold with its topology. With "total" it is meant the integral of a given quantity over the whole manifold. One of the theorems in differential topology that can be constructively used is Pinkall's theorem which states that [35]

$$\int_{\Sigma_V^v} (\sigma^2(k_i))^n \, d\eta \geq Vol(\mathbb{S}^n) \sum_{i=1}^{n} \left(\frac{i}{n-i}\right)^{n/2-i} b_i(\Sigma_V^v), \tag{4}$$

where $d\eta := d\mu / \int_{\Sigma_V^v} d\mu$ and $Vol(\mathbb{S}^n)$ is the volume of the unit n-sphere and, given the potential function of a system, $\sigma^2(k_i)$ can be easily computed (see Appendix C) as

$$\begin{aligned}\sigma^2(k_i) = \frac{1}{(n-1)^2}\left(\frac{Tr[(Hess\, V)^2]}{\|\nabla V\|^2} + \frac{\langle \nabla V, Hess\, V\ \nabla V\rangle^2}{\|\nabla V\|^6} - 2\frac{\|Hess\, V\ \nabla V\|^2}{\|\nabla V\|^4}\right) \\ - \frac{1}{n-1}\left(\frac{\triangle V}{\|\nabla V\|} - \frac{\langle \nabla V, Hess\, V\ \nabla V\rangle}{\|\nabla V\|^3}\right)^2\end{aligned} \tag{5}$$

where $\triangle V$ and $Hess\, V$ are, respectively, the Laplacian and the Hessian of the potential function V.

Then, exploiting the equality in Ref. [12], we obtain:

$$\langle \sigma^2(k_i)\rangle_\eta = \left[Vol(\mathbb{S}^n) \sum_{i=1}^{n} \left(\frac{i}{n-i}\right)^{n/2-i} b_i(\Sigma_V^v)\right]^{\frac{2}{n}} r(\Sigma_V^v) \tag{6}$$

where $r(\Sigma_V^v)$ is a small remainder, we notice that the dispersion of principal curvature is related to the sum of Betti numbers.

Another theorem that can be used is Overholt's theorem which states that the range of variability of the scalar curvature can be used to estimate the range of variability of the sectional curvatures and it is given by [36]:

$$\Delta(sectional) \geq \left[\frac{vol(\mathbb{S}_1^N) \sum_{k=0}^{N} b_k(\Sigma_V^v)}{2\ vol(\Sigma_V^v)}\right]^{2/N}. \tag{7}$$

Hence, it turns out that the variations of the topology of Σ_V^v detected by the Betti numbers can shape the potential energy profile of $\Delta(sectional)$. By being the scalar curvature of a manifold, the sum of all the sectional curvatures, thereby, the variance of the scalar curvature $R(\Sigma_V^v)$ is

$$\Delta^2(scal) = \frac{\langle R^2(\Sigma_V^v)\rangle - \langle R(\Sigma_V^v)\rangle^2}{N(N-1)} \simeq \Delta(sectional), \tag{8}$$

where $\langle\cdot\rangle$ is the geometric average over the PLS and it will be defined in the upcoming section. The scalar curvature of any PLS can be written in terms of derivative of the potential function, i.e.:

$$\begin{aligned}R(\Sigma_V^v) = &\left(\frac{\triangle V}{\|\nabla V\|} - \frac{\langle \nabla V, Hess\, V\ \nabla V\rangle}{\|\nabla V\|^3}\right)^2 + \\ &- \left(\frac{Tr[(Hess\, V)^2]}{\|\nabla V\|^2} + \frac{\langle \nabla V, Hess\, V\ \nabla V\rangle^2}{\|\nabla V\|^6} - 2\frac{\|Hess\, V\ \nabla V\|^2}{\|\nabla V\|^4}\right)\end{aligned} \tag{9}$$

In the upcoming section, we discuss how to compute, through molecular dynamics simulations, the dispersion

of the principal curvature and the variance of the scalar curvature.

IV. AVERAGES OF GEOMETRIC OBSERVABLES

In this section, we show how to observe a topology change computing numerically the geometric average of the dispersion of the principal curvature (5) and scalar curvature (9) so as to prove Pinkall's (6) and Overholt's (7) theorems. After Pinkall theorem, the relevant geometric quantity is

$$\langle \sigma^2(k_i)\rangle_{geo} = \frac{\displaystyle\int_{\Sigma_V^v} \sigma^2(k_i)\, d\mu}{\displaystyle\int_{\Sigma_V^v} d\mu}\ , \tag{10}$$

that is, the average of the dispersion of the principal curvatures on a given PLS, Σ_V^v. Instead, Overholt theorem invokes the following geometric quantity

$$\langle R^n(\Sigma_V^v)\rangle = \frac{\displaystyle\int_{\Sigma_V^v} R^n(\Sigma_V^v)\frac{d\mu}{\|\nabla V\|}}{\displaystyle\int_{\Sigma_V^v}\frac{d\mu}{\|\nabla V\|}}. \tag{11}$$

A proper combination of this quantity with $n = 1, 2$, allows to compute $\Delta^2(scal)$ which is a probe of the topological variations of the Σ_V^v. At a first glance, it is apparent that these quantities can be directly computed in the microcanonical ensemble since $d\mu/\|\nabla V\|$ is the microcanonical statistical measure, *i.e.,* it is the natural ergodic invariant measure for the microscopic Hamiltonian dynamics. Thus, for any phase space-valued function, A, invoking the ergodic theorem, the average in Eq. (11) rewrites [19]:

$$\langle A\rangle = \lim_{t\to\infty}\frac{1}{t}\int_0^t A(\tau)\, d\tau. \tag{12}$$

However, our MD simulations have been performed through GROMACS software in the canonical ensemble. This means that the numerical computations of averages of the geometric observables should be performed evaluating the observables along the solutions of the simulations which is equivalent to compute canonical averages. Hence, from our MD simulations, we have access to the following average

$$\langle A\rangle_C(n,T) = \frac{1}{\mathcal{Z}(n,T)}\int_0^\infty n\, e^{-n\bar v/k_BT}\left(\int_{\Sigma_V^v}\frac{A(\Sigma_V^v)}{\|\nabla V\|}\, d\mu\right) d\bar v, \tag{13}$$

where the configurational canonical partition function has been rewritten as follows:

$$\mathcal{Z}(n,T) = n\int_0^\infty e^{-n\bar v/k_BT}\left(\int_{\Sigma_V^v}\frac{d\mu}{\|\nabla V\|}\right) d\bar v, \tag{14}$$

where $d\mu$ is the induced metric on the PLS, Σ_V^v, and $\bar V = n\bar v$, i.e., $\bar v$ is the value of the potential per degrees of freedom. Let $\langle\sigma^2[k_i, \boldsymbol{X}(t), \boldsymbol{P}(t)]\rangle_t$ be the time average of the dispersion of the principal curvatures obtained evaluating Eq. (5) along the numerical trajectories in our canonical simulations. Then, $\langle\sigma^2\rangle_t$, in turn, coincides with the canonical average defined in Eq. (13) due to the ergodic theorem. Hence, the best approximation of Eq. (10) in terms of $\langle\sigma^2\rangle_t$ can be obtained recasting the microcanonical measure in Eq. (13) into the geometric measure. In practice, the numerator in Eq. (10) is approximated by

$$\langle\|\nabla V\|\,\sigma^2\rangle_C = \frac{1}{\mathcal{Z}(n,T)}\int_0^\infty n\, e^{-n\bar v/k_BT}\left(\int_{\Sigma_V^v}\sigma^2\, d\mu\right) d\bar v, \tag{15}$$

similarly, for the denominator we have

$$\langle\|\nabla V\|\rangle_C = \frac{1}{\mathcal{Z}(n,T)}\int_0^\infty n\, e^{-n\bar v/k_BT}\left(\int_{\Sigma_V^v} d\mu\right) d\bar v, \tag{16}$$

Now, dividing Eq. (15) by (16), we get:

$$\frac{\langle\|\nabla V\|\,\Lambda^2\rangle_C}{\langle\|\nabla V\|\rangle_C} = \frac{\displaystyle\int_0^\infty n\, e^{-n\bar v/k_BT}\left(\int_{\Sigma_V^v}\Lambda^2\, d\mu\right) d\bar v}{\displaystyle\int_0^\infty n\, e^{-n\bar v/k_BT}\left(\int_{\Sigma_V^v} d\mu\right) d\bar v}, \tag{17}$$

At this step, it is worth noting that, for large values of n, the canonical measure concentrates around a given potential level set $\Sigma_{\bar v(T)}$, where $\bar v(T)$ is the average potential function per degree of freedom and so the largest contribution to the canonical partition function is given by $\Sigma_{\bar v(T)}$ which is nothing but the equivalence of ensembles.

Hence, this means that, heuristically, in the thermody-

namic limit, the partition function reduces to

$$\mathcal{Z}(n,T) \approx n\, e^{-n\bar{v}(T)/k_BT} \int_{\Sigma_{\bar{v}(T)}} \frac{d\mu^{\Sigma_{\bar{v}(T)}}}{\|\nabla V\|}, \qquad (18)$$

and the average in Eq. (17) reads:

$$\frac{\langle \|\nabla V\|\, \Lambda^2 \rangle_C}{\langle \|\nabla V\| \rangle_C} \overset{n\to\infty}{\longrightarrow} \frac{\displaystyle\int_{\Sigma_{\bar{v}(T)}} \Lambda^2\, d\mu^{\Sigma_{\bar{v}(T)}}}{\displaystyle\int_{\Sigma_{\bar{v}(T)}} d\mu^{\Sigma_{\bar{v}(T)}}} \approx \langle \sigma^2(k_i) \rangle_{geo}, \qquad (19)$$

Of course, when the number of degrees of freedom is not very large, the support of the measure in Eq. (17) is not well concentrated, and we can expect some blurring of the curves $\langle \sigma^2(k_i) \rangle(v)$. It should be stressed that this occurs since we are evaluating "microcanonical" observables (see Eqs. (10) and (11)) with "canonical" trajectories. This effect can be simply eliminated performing molecular dynamics simulations in microcanonical ensemble. As we shall see in the following section, the concentration around the average potential value $\bar{v}(T)$ can be observed comparing Figures 7 and 8 relative to the SH3 protein and PYP protein, respectively, considering that the number of degrees of freedom are $n_{SH3} = 171$ for the SH3 protein and $n_{PYP} = 375$ for the PYP protein. A further contribution to the mentioned blurring can also come from the quantity $\langle \|\nabla V\| \rangle_C$ at the denominator.

Instead, the averages of the scalar curvature in Eq. (11) are given by

$$\langle R \rangle_C(n,T) = \frac{\displaystyle\int_0^\infty n\, e^{-n\bar{v}/k_BT} \left(\int_{\Sigma_V^v} \frac{R}{\|\nabla V\|}\, d\mu \right) d\bar{v}}{\displaystyle\int_0^\infty n\, e^{-n\bar{v}/k_BT} \left(\int_{\Sigma_V^v} \frac{d\mu}{\|\nabla V\|} \right) d\bar{v}}, \qquad (20)$$

and for a large number of particles, we have:

$$\langle R \rangle_C(n,T) \overset{n\gg 1}{\longrightarrow} \frac{\displaystyle\int_{\Sigma_V^v} \frac{R}{\|\nabla V\|}\, d\mu}{\displaystyle\int_{\Sigma_V^v} \frac{d\mu}{\|\nabla V\|}} \approx \langle R \rangle \qquad (21)$$

As we shall see in the next section, the outcomes of $\langle R \rangle(v)$ are not blurred, and since the denominator $\langle \|\nabla V\| \rangle_C$ is here absent, this seems to confirm the above comment on the role of this term in blurring the outcomes of the geometric averages, rather than attributing the blurring to an insufficient concentration of the statistical measure.

V. THERMODYNAMIC OBSERVABLES AND METHODOLOGY

The thermodynamic and geometrical observables are evaluated along the trajectories run at fixed temperatures. Indicating with $\langle \cdot \rangle$ time averages along the trajectories, we analyze: *(i)* the radius of gyration R_{gyr} as a function of the temperature T; *(ii)* the specific heat at constant volume

$$C_V = \frac{\langle E_{tot}^2 \rangle - \langle E_{tot} \rangle^2}{N^2 k_B T^2}\ , \qquad (22)$$

as a function of the temperature, where E_{tot} is the total energy and k_B is the Boltzmann constant; *(iii)* the relation between the temperature T and the energy density is

$$\epsilon = k_B T/2 + \langle V \rangle / N\ , \qquad (23)$$

where we recall that V is the total potential field. The units are the standard GROMACS ones, *i.e.*, $[T]$ = K, $[E_{tot}] = [V]$ = kJ/mol, $[R_{gyr}]$ = nm and $[\tilde{k}_B]$ = kJ/mol K. We analyze the src-Src homology 3 protein domain (SH3, PDB code 1FMK) (see left-hand panel of Figure 1), of 57 amino acids; 2 random sequences of the same 57 amino acids (RDMa,b); and the photoactive yellow protein (PYP, PDB code 2PYP) (see right-hand panel of Figure 1) composed of 125 amino acids. We remark that the simulations are run also for several random sequences yielding very similar results and only two of them are reported here for the sake of simplicity. The randomization is implemented using the SH3 coarse grained potential described in [37] and randomly permuting the parameters involved in the model: this way, we can get a sort of random heteropolymer starting from the good folding sequence of SH3.

The simulation are performed using the GROMACS software [38–43]. Averages and fluctuations are evaluated over 2000 frames for each fixed temperature simulation. The run temperatures are taken, after some tests, in the folding range with an interval of 5K between each trajectory.

VI. RESULTS

In Figure 2 the radius of gyration is reported for the different sequences of the SH3 and PYP proteins, respectively. It is evident that only the sequences of the good folders SH3 and PYP exhibit the bifurcation pattern typical of the folding transition. In Figure 3 the specific heat and the caloric curve are reported for the SH3 protein and display the typical patterns of a phase transition. Bachmann's criterion [16, 17] identifies a phase transition point with the inflection point of the caloric curve. In our case, the caloric curves are obtained by averaging the total energy of the system, and, being the temperature an error-free input parameter, after a sufficiently long integration time, the error on the averaged value of energy can be made arbitrarily small. In so doing, the inflection point is very well located. The caloric curve in the upper panel of Fig. 3 displays an inflection point which is absent in the lower panel, again of Fig. 3, reporting the caloric curve of the randomized sequence of the SH3 protein. In particular the inflection point of the

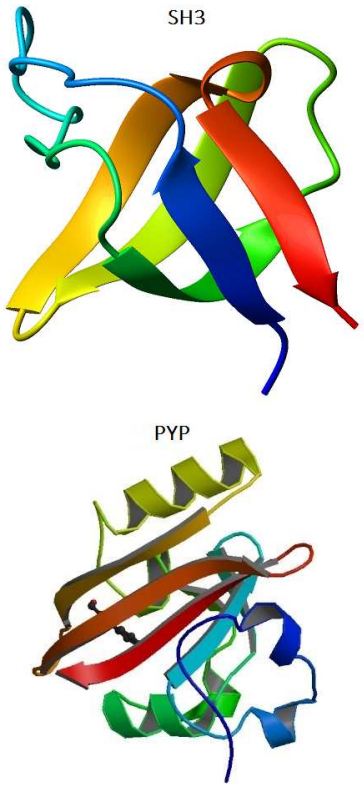

FIG. 1: Cartoon representation of SH3 and PYP.

caloric curve is typical of a second order phase transition as discussed in Refs. [16, 17]. In Figure 4 the specific heat and the caloric curve are reported for the PYP protein and also in this case they display the typical patterns of a phase transition. However, the pattern of the caloric curve - in concordance with the sharp drop of the gyration radius shown by Fig. 2 - could be compatible with a first-order phase transition [16–18]. Remarkably, the thermodynamic signatures of a phase transition, independently of its order, are lost in the case of the randomized sequences of amino acids as shown in the same figures.

In Fig. 5 the total scalar curvature and the total variance of the scalar curvature of the equipotential level sets in configuration space are reported as functions of the temperature, normalized to the folding transition temperature, for the SH3 protein. Both quantities show a kink in correspondence to the folding transition which disappears in the randomized sequence. The same phenomenology is shown in Figure 6 for the PYP protein. By looking at Figures 5 and 6 where the geometrical quantities $\langle \sigma_R(\Sigma_v) \rangle$ and $\langle R(\Sigma_v) \rangle$ are reported, one is tempted to identify inflection points for the SH3 and jumps for the PYP. Thus a second order and a first order transition, respectively. However, even though the temperature patterns of these quantities are neat, these are numerical outcomes and it is hard to make a clearcut assessment. From a theoretical viewpoint, of course there would be the possibility of inferring the order of the transition: Eq.(1) suggests that the behaviour of the sum of the Betti numbers as a function of v affects the order of the derivative of $S(v)$ becoming singular, and thus the order of the transition. But one would need the analytic computation of quantities of topological meaning.

Finally, Figures 7 and 8 show the dispersions of the principal curvatures of the equipotential level sets in the configuration space for SH3 and PYP proteins and randomized sequences, respectively. This quantity shows peculiar patterns that are well evident when plotted as a function of the value of potential energy per degree of freedom. These patterns are less clear when plotted as a function of temperature, although the presence of cusps can be guessed by means of several polynomial fits of the points below and above the folding transition temperature, respectively.

In order to understand what do we learn from the patterns of the geometrical quantities reported as functions of the potential energy and of temperature, let us first consider that the shape of the specific heat depends on the shape of the entropy according to the relation $C_v = -(\partial S/\partial E)^2(\partial^2 S/\partial E^2)^{-1}$ stemming from $C_v = (\partial T(E)/\partial E)^{-1}$ with $T(E) = (\partial S/\partial E)^{-1}$. Then, related with the formula reported in Eq. (1), we also have [1]

$$S_N(E) = \frac{k_{\rm B}}{N} \log \int_{\Sigma_E^N} \frac{d\mu}{\|\nabla H\|} \qquad (24)$$
$$\simeq \frac{k_{\rm B}}{N} \log \left[{\rm vol}(\mathbb{S}_1^{N-1}) \sum_{i=0}^{N} b_i(\Sigma_E^N) + r_1(E) \right] + r_2(E) \ ,$$

where $r_1(E)$, and $r_2(E)$ are smooth functions, $b_i(\Sigma_E^N)$ are the Betti numbers of the energy level sets, $d\mu$ is the measure on the level set, and $\mathbb{S}_1^{N-1}$ stands for a hypersphere of unit radius. From this formula it can be understood that some "abrupt" change in the topology of the energy level sets can affect both the shape of the caloric curve $T = T(E)$ and of the specific heat through the energy variation of $S_N(E)$. Now, the scalar curvature R is the sum of sectional curvatures so that its variance σ_R contains the variance of the sectional curvatures [13], so that the quantity

$$\Delta(sec) > \left[\frac{vol(\mathbb{S}_1^N) \sum_{k=0}^{N} b_k(\Sigma_E)}{2 \ vol(\Sigma_E)} \right]^{2/N} \qquad (25)$$

in strict analogy with Eqs.(7) and (8), detects topology changes of the energy level sets in phase space. Therefore, the jumps in the patterns of the total scalar curvature and the total variance of the scalar curvature reported in Figures 5 and 6 just probe some kind of "abrupt" change in the topology of the energy level sets. Similarly, and complementary to this, the potential energy patterns of the dispersion of the principal curvatures of the equipotential level sets reported in Figures 7 and 8 probe some

kind of "abrupt" change in the topology of these submanifolds of configuration space, and thus also of phase space, after Pinkall's theorem relating the dispersion of the principal curvatures of a manifold with a weighted sum of its Betti numbers as given in Eq. (6).

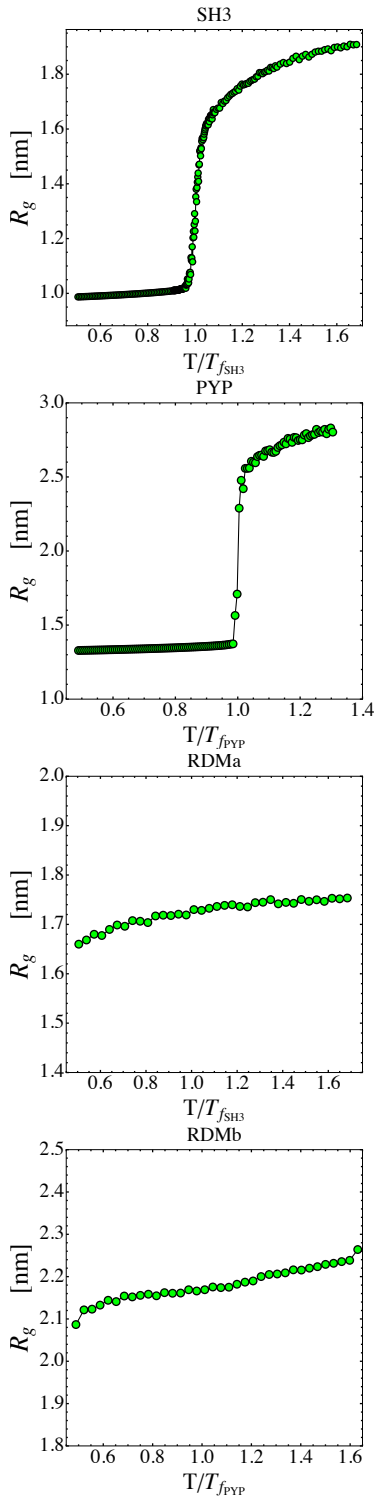

FIG. 2: (Color online) Plots of the gyration radius for the different sequences. It is evident that only good folders as SH3 and PYP show a temperature dependence typical of the folding transition (upper panels) that is lost for randomized sequences (lower panels). $T_{f_{SH3}}$ and $T_{f_{PYP}}$ identify the folding transition of the SH3 and PYP proteins, respectively.

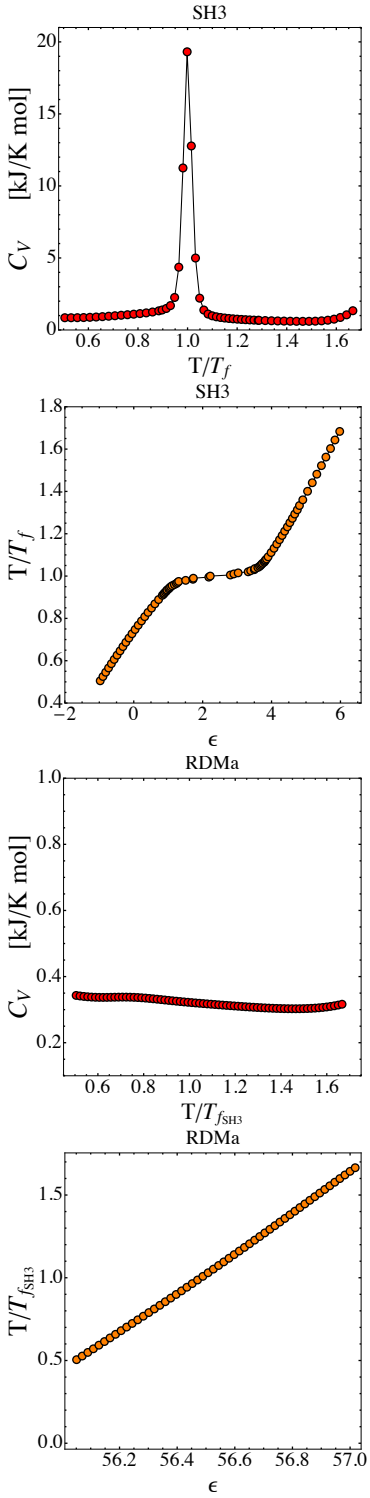

FIG. 3: (Color online) The specific heat and the caloric curve for the SH3 protein show patterns typical of a phase transition (upper panels). These features are lost in the case of the randomized version of the correct sequence of the SH3 protein (lower panels).

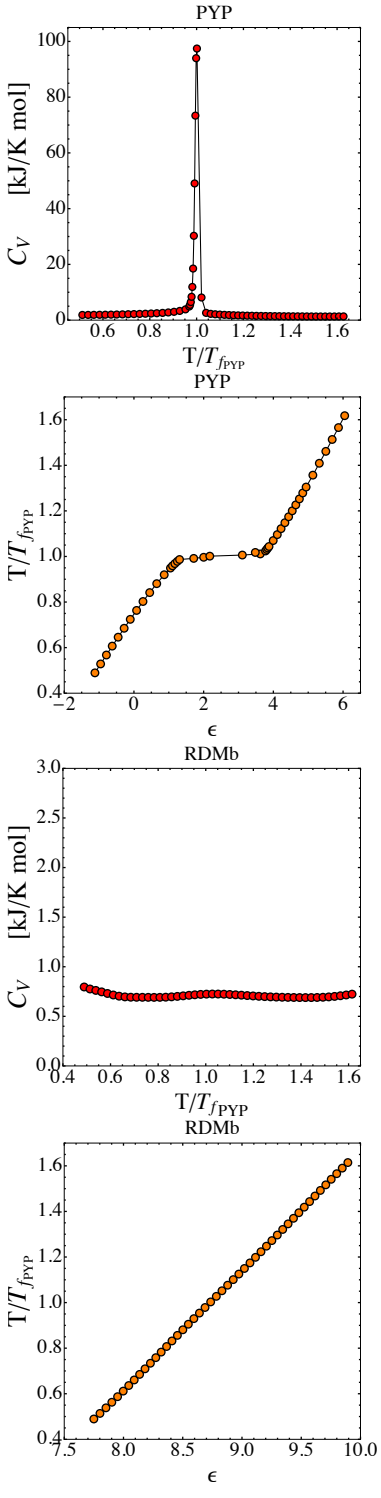

FIG. 4: (Color online) The specific heat and the caloric curve for the PYP protein show patterns typical of a phase transition (upper panels). These features are lost in the case of the randomized version of the correct sequence of the PYP protein (lower panels).

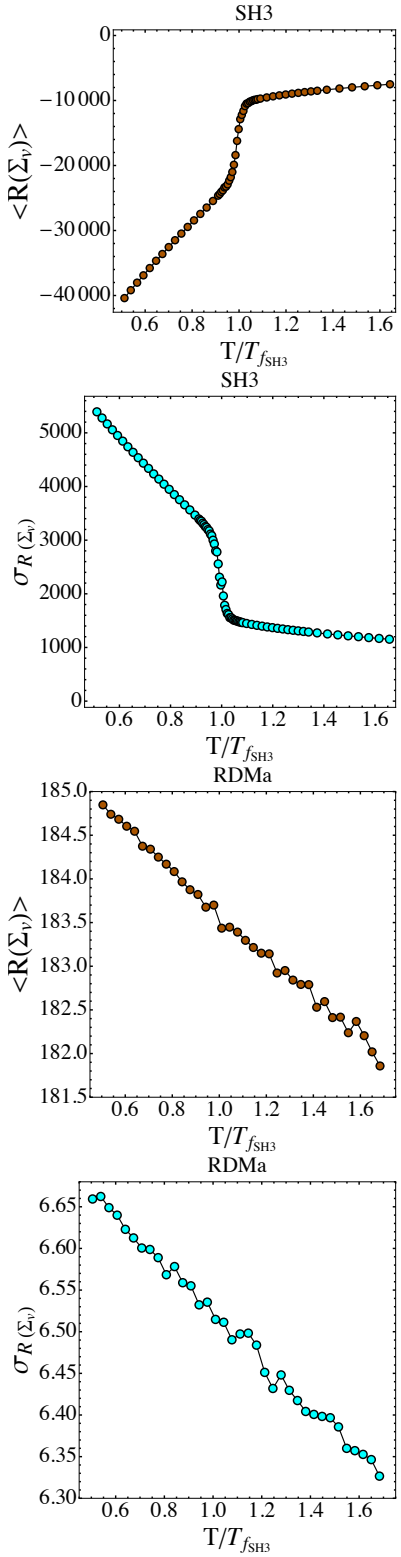

FIG. 5: (Color online) The total scalar curvature and its variance, of equipotential level sets, are reported as functions of temperature for the SH3 protein (upper panels) and for its randomized sequence of amino acids (lower panels).

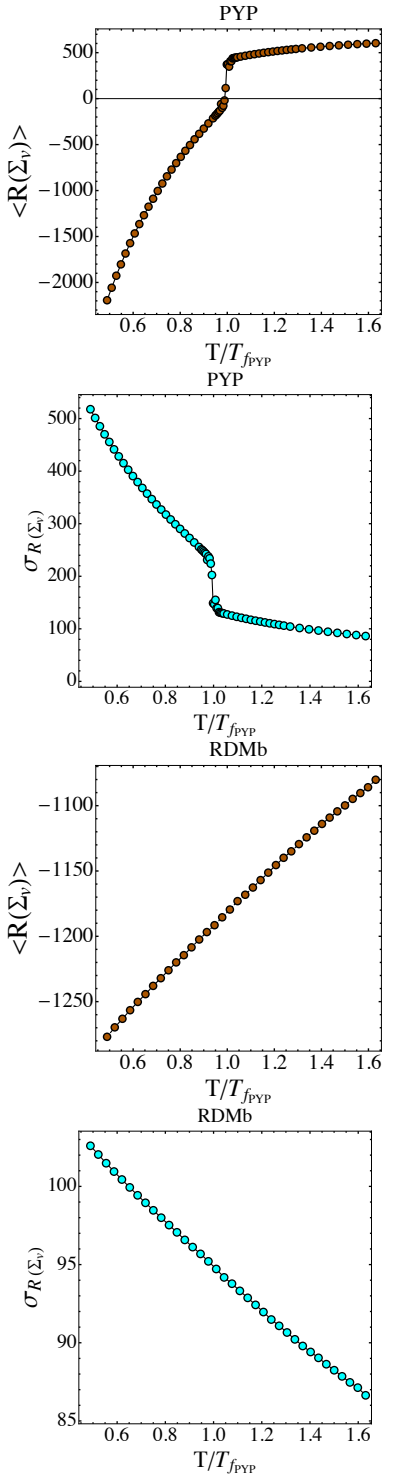

FIG. 6: (Color online) The total scalar curvature and its variance, of equipotential level sets, are reported as functions of temperature for the PYP protein (upper panels) and for its randomized sequence of amino acids (lower panels).

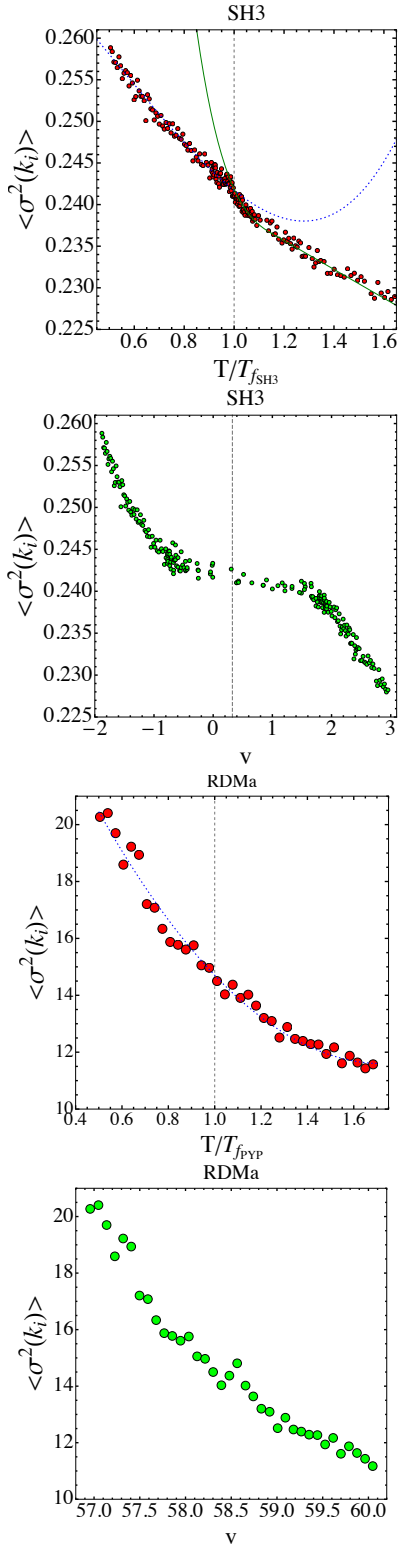

FIG. 7: (Color online) The variance of the principal curvatures of the equipotential level sets is reported as a function of temperature (left panel) and of the potential energy per degree of freedom v for both the SH3 protein (upper panels), and for its randomized version (lower panels).

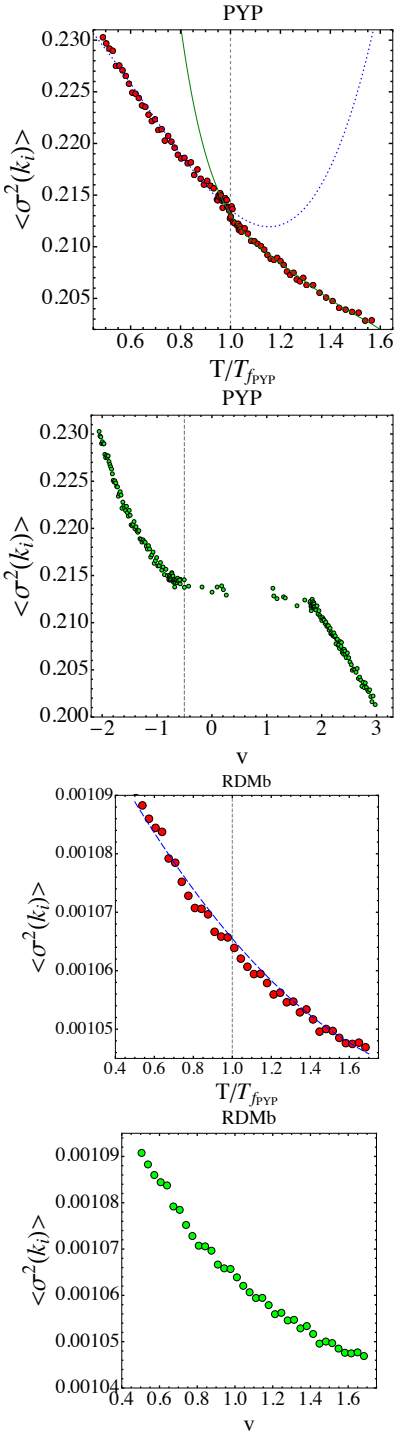

FIG. 8: (Color online) The variance of the principal curvatures of the equipotential level sets is reported as a function of temperature (left panel) and of the potential energy per degree of freedom v for both the PYP protein (upper panels), and for its randomized version (lower panels).

A. Remark

As already mentioned throughout this paper, the precise relationship between geometry and topology is given by theorems in differential topology but only a few number of them can be constructively used (essentially the Gauss–Bonnet–Hopf, the Chern–Lashof, and Pinkall theorems [35]). However, sharp changes of various geometrical observables signaling PTs have been also reported within a purely geometrical theory [20, 44] — based on the work in [45] - also in the absence of any known theorem connecting geometry and topology. Sharp changes of geometry of the leaves of a family of manifolds foliating configuration space can be generically supposed to stem from their topological changes even when this fact cannot be proved. Anyway, the use of simple geometrical observables, easier to compute with respect to the Gauss-Kronecker curvature or with respect to the dispersion of principal curvatures, can be of practical interest to detect PTs in the absence of symmetry-breaking or in small N systems. On the other side, recent constructive methods developed in algebraic topology, namely *persistent homology* [46] in the framework of Topological Data Analysis (TDA) [47], provide a different strategy to constructively detect the topological origin of phase transitions [48]. TDA and persistent homology have recently been used also in the context of protein folding [49] to capture the formation of tertiary structures thus providing a topological approach to the dynamics of protein folding. This is a complementary problem to the "static" description of protein folding, seen as a phase transition at equilibrium that occurs as a control parameter changes.

VII. CONCLUSIONS

By considering a minimalistic model of the SH3 and PYP proteins, besides the standard signatures of the folding transition, the computation of suitable geometric quantities of the equipotential hypersurfaces in configuration space and of the energy hypersurfaces in phase space of these molecules, respectively, allows to probe topological changes of both families of hypersurfaces. The computation of the same geometric quantities for randomized versions of the correct sequences of the SH3 and PYP proteins yielded monotonic patterns as functions of the potential energy density, or of the total energy density, manifestly discriminating between proteins and random heteropolymers. Remarkably, the peculiar geometric signatures found in correspondence with the protein folding transition are the "shadows" of some peculiar and sharp topological change of the mentioned submanifolds of configuration space and of phase space. The protein folding transition takes place in systems with a small number of degrees of freedom (very far from the Avogadro number) and in the absence of a symmetry-breaking phenomenon, however, considered from this topological perspective, the protein folding transition fully qualifies as a phase transition.

ACKNOWLEDGEMENTS

The authors are indebted with Lapo Casetti, Mary Anne Rohrdanz, Lorenzo Mazzoni, Lorenzo Boninsegna, Nakia Carlevaro and Cecilia Clementi for inspiring discussions and suggestions. This work has been done within the framework of the project MOLINT which has received funding from the Excellence Initiative of Aix-Marseille University-A*Midex, a French 'Investissements d'Avenir' Programme. Ghofrane Bel Hadj Aissa thanks the support by the QuantERA, ERA-NET Cofund 731473 (Project Q-CLOCKS), Italy.

Appendix A: Topological theory of phase transitions in a nutshell

Let us sketchily present the basic conceptual origin of the topological theory of phase transitions. The theory stems from the geometrization of Hamiltonian dynamics which proceeds as follows. Given a generic system of N degrees of freedom described by a Hamiltonian $H = \frac{1}{2}\sum_{i=1}^{N} p_i^2 + V(q_1, \ldots, q_N)$, or equivalently by the corresponding Lagrangian function $L = \frac{1}{2}\sum_{i=1}^{N} \dot{q}_i^2 - V(q_1, \ldots, q_N)$, its dynamics can be identified with a geodesic flow of an appropriate Riemannian differentiable manifold. This differential geometric framework is given by configuration space $M_E = \{q \in \mathbb{R}^N | V(q) < E\}$ endowed with the non-Euclidean metric of components [1] $g_{ij} = 2[E - V(q)]\delta_{ij}$, whence the infinitesimal arc element $ds^2 = 2[E - V(q)]^2 dq_i\ dq^i$; then Newton equations are retrieved from the geodesic equations

$$\frac{d^2q^i}{ds^2} + \Gamma^i_{jk}\frac{dq^j}{ds}\frac{dq^k}{ds} = 0\ ,$$

where Γ^i_{jk} are the Christoffel connection coefficients of the manifold. Then, in this context, the natural question is whether the mechanical manifolds (M_E, g) undergo some peculiar geometrical change when E crosses a critical value E_c that corresponds to a phase transition. And it has been discovered that this is actually the case [1]. Moreover, the peculiar geometrical changes associated with phase transitions were discovered to be also the effects of deeper topological changes of the potential level sets $\Sigma_v^{V_N} := \{V_N(q_1, \ldots, q_N) = v \in \mathbb{R}\}$ in configurations space, and, equivalently, of the balls $\{M_v^{V_N} = V_N^{-1}((-\infty, v])\}_{v\in\mathbb{R}}$ bounded by the $\Sigma_v^{V_N}$. In other words, given a Hamiltonian system undergoing a phase transition, let $v_c = v_c(E_c)$ be the average potential energy corresponding to the phase transition point, a topological change means that the manifolds $\Sigma_{v<v_c}^{V_N}$ and $\Sigma_{v>v_c}^{V_N}$ are not diffeomorphic, that is, they cannot be

transformed one into the other with a differentiable application with differentiable inverse. Topological changes of these manifolds are related with the presence of critical points of the potential function $V(q)$ in configuration space. To get an intuitive idea of the relationship between critical points of a function in a given space and the topology of its level sets, let us consider a low dimensional and intuitive case. Given a smooth function f, bounded below, such that $f : \mathbb{R}^N \to \mathbb{R}$. Its level sets $\Sigma_u = f^{-1}(u)$ are diffeomorphically transformed one into the other by the flow [50]

$$\frac{dx}{du} = \frac{\nabla f}{\|\nabla f\|^2} \ ,$$

where $x \in \mathbb{R}^N$, i.e., the points of a hypersurface Σ_{u_0} with $u_0 \in [a, b] \subset \mathbb{R}$, are mapped by this flow to the points of another Σ_{u_1} with $u_1 \in [a, b]$, provided that ∇f never vanishes in the interval $[a, b]$. In other words, if in the interval $[a, b]$ the function f has no critical points, all the level sets $\Sigma_u = f^{-1}(u)$, with $u \in [a, b]$, have the same topology. Conversely, the appearance of critical points of f at some critical value u_c breaks the diffeomorphicity among the $\Sigma_{u<u_c}$ and $\Sigma_{u>u_c}$. This is illustrated by one of the simplest possible examples in Figure 9. A systematic study is developed within *Morse theory* of the relationship between topological properties of a manifold and the critical points of a suitable class of real-valued functions (Morse functions) defined on it. In particular, if $f \equiv V$, Morse theory tells us that the existence of critical points of V is associated with topological changes of the hypersurfaces $\{\Sigma_v\}_{v\in\mathbb{R}}$, and also of the $\{M_v\}_{v\in\mathbb{R}}$, provided that V is a good Morse function (that is, bounded below, with no vanishing eigenvalues of its Hessian matrix). In general, finding either analytically or numerically all the critical points of a potential $V(q)$ is a very hard task, often an unfeasible one. Thus in order to get information on the topology of the manifolds of interest one has to resort to the available theorems in differential topology, like the Chern-Lashof theorem mentioned in the next appendix, or the Pinkall theorem used in the main text. These theorems relate some total (that is integrated over the whole manifold) geometric property of a manifold with some information on its topology. Note that Morse indexes $\mu_k(M)$ of a manifold M count the number of critical points of degree k (the number of negative eigenvalues of the Hessian of the Morse function). Betti numbers are related with Morse indexes by the inequalities $\mu_k(M) \geq b_k(M)$. The $b_k(M)$ are dimensions of some groups (homology and cohomology of M) invariant under diffeomorphisms of M.

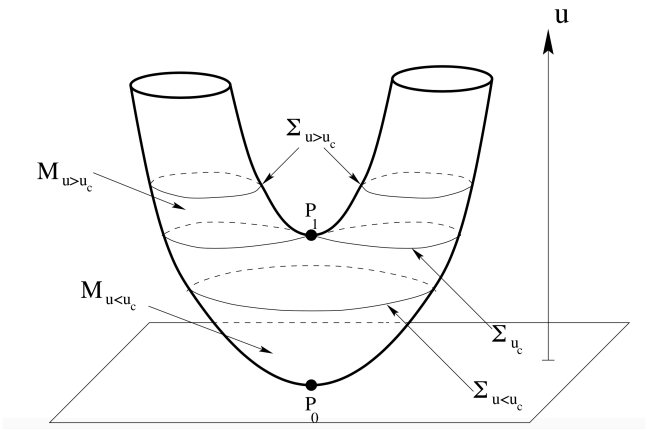

FIG. 9: The function f is here the height of a point of the bended cylinder with respect to the ground. In P_1 it is $df = 0$. The level sets $\Sigma_u = f^{-1}(u)$ below this critical point are circles, whereas above are the union of two circles. The manifolds $M_u = f^{-1}((-\infty, u])$ are disks for $u < u_c$ and cylinders for $u > u_c$

Appendix B: Derivation of Equation (1)

In this appendix, we sketch the proof of formula (1). Any details about the rigorous proof can be found in Ref. [1].

We note that the relation in Eq. (1) relates thermodynamic entropy, defined in the microcanonical configurational ensemble, with quantities of topological meaning (the Morse indexes) of the configuration-space submanifolds $\mathcal{M}_v = V_N^{-1}((-\infty, v]) = \{q = (q_1, \ldots, q_N) \in \mathbb{R}^N | V_N(q) \leq v\}$.

Let us consider the definition of the configurational microcanonical entropy $S_N(v)$ ($k_B = 1$)

$$S_N(v) = \frac{1}{N} \log \Omega_N(v) \ , \tag{B1}$$

with

$$\Omega_N(v) = \frac{1}{N!} \int_{\Sigma_v} \frac{d\sigma}{\|\nabla V_N\|}, \tag{B2}$$

where Σ_v is the potential level set (PLS) defined by $\Sigma_v^{V_N} := \{q \in \mathbb{R}^N \,|\, V_N(q_1, \ldots, q_N) = v \in \mathbb{R}\}$. By exploiting Federer's derivation formula:

$$\frac{d^k}{dv^k} \int_{\Sigma_v} \alpha \ d\sigma = \int_{\Sigma_v} A^k(\alpha) \ d\sigma, \tag{B3}$$

where α is any configuration space-valued function and

$$A(\alpha) := \frac{1}{\|\nabla V_N\|} \nabla \cdot \left(\alpha \frac{\nabla V_N}{\|\nabla V_N\|} \right), \tag{B4}$$

Eq. (B2) reduces to

$$\frac{d\Omega_N}{dv}(v) = \frac{1}{N!} \int_{\Sigma_v} \frac{M^*}{\|\nabla V_N\|} \frac{d\sigma}{\|\nabla V_N\|} + \mathcal{O}\left(\frac{1}{N}\right). \tag{B5}$$

where we defined $M^* := \nabla(\nabla V_N/\|\nabla V_N\|)$. Now, integrating (B5) and then, at large N, considering that the volume measure $d\mu := d\sigma/\|\nabla V_N\|$ concentrates on the boundary Σ_v, we get:

$$\Omega_N(v) = \frac{1}{N!}\int_{\Sigma_v}\frac{M^*}{\|\nabla V_N\|}d\mu \simeq \frac{\delta v}{N!}\langle\|\nabla V_N\|^{-1}\rangle\int_{\Sigma_v} M^* d\mu \tag{B6}$$

where δv is the length of a small energy interval around the value v and where we have used that $\|\nabla V_N\|$ is positive and only very weakly varying at large N. By means of Hölder's inequality for integrals we get

$$\int_{\Sigma_v} M^* d\mu \leq \left(\int_{\Sigma_v}|M^*|^N\ d\mu\right)^{\frac{1}{N}}\left(\int_{\Sigma_v} d\mu\right)^{\frac{N-1}{N}} \tag{B7}$$

the sign of equality being better approached when M^* is everywhere positive. Then, by making use of Eqs. (B2) and (B6), we have

$$\Omega_N(v) \leq [\Omega_\nu(v)]^{\frac{N-1}{N}}\left(\frac{1}{N!}\int_{\Sigma_v}|M^*|d\mu\right)^{\frac{1}{N}}\frac{\delta v}{\langle\|\nabla V_N\|\rangle}, \tag{B8}$$

and introducing a suitable deformation factor $d(v)$, we can reach the following equality

$$\Omega_N(v) = \frac{[d(v)]^N(\delta v)^N}{\langle\|\nabla V_N\|\rangle^{2N}}\frac{1}{N!}\int_{\Sigma_v}|M^*|^N\ d\mu. \tag{B9}$$

Noticing that M^* is proportional to the mean curvature, and being the latter the sum of the principal curvatures, $\{\kappa\}_{i\in N}$, we can write

$$\begin{aligned}(\kappa_1+\ldots+\kappa_N)^N &= (\epsilon_0|\kappa_1+\ldots+\kappa_N|)^N \\ &= |\kappa_1|^N+\ldots+|\kappa_N|^N + t(v)\end{aligned} \tag{B10}$$

where $t(v)$ contains all the terms of the multinomial expansion that one have passing from the second to the third equality but that one with $n_1 = n_2 = \ldots = n_\rho = 1$. Then, we also defined $\epsilon_0 = \text{sign}(\kappa_1+\ldots+\kappa_N)$. Now, applying the multinomial expansion ($\rho\in\mathbb{N}$):

$$(x_1+\ldots+x_\rho)^\rho = \sum_{\{n_i\},\sum n_k=\rho}\frac{\rho!}{n_1!\ldots n_\rho!}x_1^{n_1}\ldots x_\rho^{n_\rho}, \tag{B11}$$

Recalling that the Gauss-Kronecker curvature of Σ_v is $K_G = \prod_{i=1}^N\kappa_i$ we get

$$|M^*|^N \approx N!|K_G| + \widetilde{t}(v), \tag{B12}$$

and we obtain

$$\Omega_N(v) \approx \frac{[d(v)]^N(\delta v)^N}{\langle\|\nabla V_N\|\rangle^{2N}}\int_{\Sigma_v}\left(|K_G| + \frac{\widetilde{t}(v)}{N!}\right)\ d\sigma, \tag{B13}$$

where, again, we have disregard the term $\|\nabla V_N\|^{-1}$ in the integration measure since it is very weakly varying at large N. Finally, according to the Chern–Lashof theorem, we can rewrite

$$\int_{\Sigma_v}|K_G|\ d\sigma = \frac{1}{2}vol(\mathbb{S}_1^{N-1})\sum_{i=0}^N\mu_i(\Sigma_v), \tag{B14}$$

where $\mu_i(\Sigma_v)$ are the Morse indexes of Σ_v.

Finally, the entropy per degree of freedom reads as

$$\begin{aligned}S(v) &= \frac{k_B}{N}\log\Omega_N(v) \\ &= \frac{1}{N}\log\left[\text{vol}(\mathbb{S}_1^{N-1})\sum_{i=0}^N\mu_i(\Sigma_v) + \int_{\Sigma_v}d\sigma\frac{\widetilde{t}(v)}{N!}\right] \\ &\quad + \frac{1}{N}\log\frac{[d(v)]^N(\delta v)^N}{\langle\|\nabla V\|\rangle^{2N}}\ .\end{aligned} \tag{B15}$$

The meaning of (B15) is better understood if we consider that the Morse indexes $\mu_i(M)$ of a differentiable manifold M are related to the Betti numbers $b_i(M)$ of the same manifold by the inequalities

$$\mu_i(M) \geq b_i(M)\ . \tag{B16}$$

At large dimension we can safely replace (B16) with $\mu_i(M)\approx b_i(M)$.

Equation (B15), rewritten as

$$\begin{aligned}S(v) &\approx \frac{k_B}{N}\log\left[\text{vol}(\mathbb{S}_1^{N-1})\sum_{i=0}^N b_i(\Sigma_v) + \int_{\Sigma_v}d\sigma\frac{\widetilde{t}(v)}{N!}\right] \\ &\quad + \frac{1}{N}\log\frac{[d(v)]^N(\delta v)^N}{\langle\|\nabla V\|\rangle^{2N}}\ ,\end{aligned} \tag{B17}$$

links topological properties of the *microscopic* phase space with the *macroscopic* thermodynamic potential $S(v)$.

Appendix C: Dispersion of principal curvature and scalar curvature

The Weingarten operator is a tensor containing the most relevant information about the extrinsic geometry of a hypersurface such as Σ_V^v and it is defined by [12, 44, 45]:

$$\mathcal{W}_{\boldsymbol{\nu}}(\boldsymbol{X}) = \nabla_{\boldsymbol{X}}\boldsymbol{\nu}, \tag{C1}$$

where $\boldsymbol{\nu}$ is the unit normal vector to the hypersurface

$$\boldsymbol{\nu} = \frac{\nabla V}{\|\nabla V\|} \tag{C2}$$

whereas $\boldsymbol{X}$ is any vector tangent to Σ_V^v and $\nabla := (\partial_{q^1},\ldots,\partial_{q^n})$ is the gradient operator. The topological observables that we want to compute are the *dispersion of the principal curvatures*, $\sigma(k_i)^2$, and the *scalar curvature*, R, of Σ_V^v.

The dispersion of principal curvatures is defined by [12]:

$$\sigma(k_i)^2 = \frac{Tr[\mathcal{W}_{\boldsymbol{\nu}}^2]}{n-1} - \frac{(Tr[\mathcal{W}_{\boldsymbol{\nu}}])^2}{(n-1)^2} \tag{C3}$$

whereas the scalar curvature is [51]:

$$R_{\Sigma_V^v} = Tr[\mathcal{W}_{\boldsymbol{\nu}}]^2 - Tr[\mathcal{W}_{\boldsymbol{\nu}}^2] \tag{C4}$$

Although, Eqs. (C3) and (C4) seem to be just formal relations, it can be shown [44, 45, 51] that they simply correspond to specific combinations of derivatives of the potential function. The trace is

$$Tr[\mathcal{W}_{\boldsymbol{\nu}}] = \frac{\triangle V}{\|\nabla V\|} - \frac{\langle \nabla V, Hess\, V\ \nabla V\rangle}{\|\nabla V\|^3} \tag{C5}$$

where $\triangle V$ and $Hess\, V$ are, respectively, the Laplacian and the Hessian of the potential function V whereas the trace of the square of the Weingarten operator is [12, 44, 45]:

$$Tr[\mathcal{W}_{\boldsymbol{\nu}}^2] = \frac{Tr[(Hess\, V)^2]}{\|\nabla V\|^2} + \frac{\langle \nabla V, Hess\, V\ \nabla V\rangle^2}{\|\nabla V\|^6} - 2\frac{\|Hess\, V\ \nabla V\|^2}{\|\nabla V\|^4}. \tag{C6}$$

It should be stressed that Eqs. (C5) and (C6) can be easily computed in a molecular dynamics simulation. In fact, it requires to know the forces acting between all the particles composing the system and the Hessian of the potential function. It is apparent that $F_i := \nabla_{q_i} V$ and $Hess\, V_{ij} = \nabla_{q_i}\nabla_{q_j} V$ are well-posed quantities that can be easily defined in a simulation.

[1] M. Pettini, *Geometry and Topology in Hamiltonian Dynamics and Statistical Mechanics*, IAM Series n.33, (Springer, New York, 2007).

[2] Lapo Casetti, Cecilia Clementi, and Marco Pettini, *Riemannian theory of Hamiltonian chaos and Lyapunov exponents*, Phys. Rev. E**54**, 5969 (1996).

[3] Di Cairano, L., Matteo G., and Pettini, M., *Coherent Riemannian-geometric description of Hamiltonian order and chaos with Jacobi metric.*, Chaos: An Interdisciplinary Journal of Nonlinear Science **29**, 123134 (2019).

[4] Di Cairano, L., Gori, M., Pettini, G. and Pettini, M., *Hamiltonian chaos and differential geometry of configuration space–time.*, Physica D: Nonlinear Phenomena **422**, 132909 (2021).

[5] L.Casetti, M. Pettini, E.G.D. Cohen, *Geometric approach to Hamiltonian dynamics and statistical mechanics*, Phys. Rep. **337**, 237-342 (2000).

[6] R. Franzosi, and M. Pettini, *Theorem on the origin of Phase Transitions*, Phys. Rev. Lett. **92**, 060601 (2004).

[7] R. Franzosi, M. Pettini, and L. Spinelli, *Topology and Phase Transitions I. Preliminary results*, Nucl. Phys. B**782** [PM], 189 (2007).

[8] R. Franzosi and M. Pettini, *Topology and Phase Transitions II. Theorem on a necessary relation*, Nucl. Phys. B**782** [PM], 219 (2007).

[9] M. Gori, R. Franzosi, G. Pettini, and M. Pettini, *Topological Theory of Phase Transitions*, J. Phys. A: Math. Theor. **55**, 375002 (2022).

[10] L.D. Landau, *On the theory of phase transitions. I.*, Zh. Eksp. Teor. Fiz., **11**, (1937).

[11] L.D. Landau, *On the theory of phase transitions. II.*, Zh. Eksp. Teor. Fiz., **11**, (1937).

[12] Gh. Bel-Hadj-Aissa, M. Gori, R. Franzosi, and M. Pettini, *Geometrical and topological study of the Kosterlitz-Thouless phase transition in the XY model in two dimensions*, J. Stat. Mech.: Theory and Experiment **2**, 023206 (2021).

[13] G. Pettini, M. Gori, R. Franzosi, C. Clementi, M. Pettini, *On the origin of phase transitions in the absence of symmetry-breaking*, Physica **A** 516, 376 (2019).

[14] C. N. Yang and T. D. Lee, *Statistical theory of equations of state and phase transitions. I. Theory of condensation*, Physical Review, **87**, (1952)

[15] T. D. Lee and C. N. Yang, *Statistical theory of equations of state and phase transitions. II. Lattice gas and Ising model*, Physical Review, **87**, (1952).

[16] M. Bachmann, *Thermodynamics and Statistical Mechanics of Macromolecular Systems*, (Cambridge University Press, New York, 2014).

[17] K. Qi and M. Bachmann, *Classification of Phase Transitions by Microcanonical Inflection-Point Analysis*, Phys. Rev. Lett. **120**,180601 (2018).

[18] T. Koci, and M. Bachmann, *Subphase transitions in first-order aggregation processes*, Physical Review E, **95**, (2017).

[19] Gh. Bel-Hadj-Aissa, M. Gori, V. Penna, G. Pettini, and R. Franzosi, *Geometrical Aspects in the Analysis of Microcanonical Phase-Transitions*, Entropy **22**, 380 2020.

[20] L. Di Cairano, *The geometric theory of phase transitions.*, J. Phys. A: Math. Theor. **55**, 27LT01 (2022).

[21] L. Casetti, M. Pettini and E.G.D. Cohen, *Phase transitions and topology changes in configuration space*, J. Stat. Phys. **111**, 1091 (2003).

[22] L. Angelani, L. Casetti, M. Pettini, G. Ruocco and F. Zamponi, *Topology and Phase Transitions: from an exactly solvable model to a relation between topology and thermodynamics*, Phys. Rev. E**71**, 036152 (2005).

[23] C. B. Anfinsen, *Principles that Govern the Folding of Protein Chains*, Science **181**, 223-230. (1973).

[24] L. N. Mazzoni and L. Casetti, *Curvature of the energy landscape and folding of model proteins*, Phys. Rev. Lett. **97**, 18104 (2006).

[25] L. N. Mazzoni and L. Casetti, *Geometry of the energy landscape and folding transition in a simple model of a protein*, Phys. Rev. E**77**, 051917 (2008).

[26] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, *Toward an outline of the topography of a realistic protein-folding funnel*, Proc. Nat. Acad. Sci. **92**, 3626 (1995).

[27] C. Clementi, H. Nymeyer, and J. N. Onuchic, *Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins*, Journal of Mol. Biology **298**, 937 (2000).

[28] 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 computational biology **12**, e1004794 (2016).

[29] W. Xu, S. C. Harrison, and M. J. Eck, *Three-dimensional structure of the tyrosine kinase c-Src*, Nature **385**, 595 (1997).

[30] P. Düx, G. Rubinstenn, G. W. Vuister, R. Boelens, F. A. A. Mulder, K. Hard, W. D. Hoff, et al., *Solution structure and backbone dynamics of the photoactive yellow protein*, Biochemistry **37**, 12689 (1998).

[31] M. J. Abraham, T. Murtola, R. Schulz, S. Pall, J. C. Smith, B. Hess, and E. Lindahl, *GROMACS: High performance molecular simulations through multilevel parallelism from laptops to supercomputers*, SoftwareX **1-2**, 19 (2015).

[32] M. Bonomi, G. Bussi, C. Camilloni, G. A. Tribello, P. Banas, A. Barducci, M. Bernetti, P. G. Bolhuis, S. Bottaro, D. Branduardi, et al., *Promoting transparency and reproducibility in enhanced molecular simulations*, Nature Methods **16**, 670 (2019).

[33] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, *PLUMED 2: New feathers for an old bird*, Comp. Phys. Comm. **185**, 604 (2014).

[34] 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).

[35] U. Pinkall, *Inequalities of Willmore type for submanifolds*, Math. Zeitschrift, **193**, 241 (1986).

[36] M. Overholt, *Fluctuation of sectional curvature for closed hypersurfaces*, Rocky Mount. J. of Math, 385 (2002).

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

[38] H.J.C. Berendsen, et al., *GROMACS: A message-passing parallel molecular dynamics implementation*, *Comp. Phys. Comm.* **91**, 43 (1995).

[39] E. Lindahl, et al., *GROMACS 3.0: a package for molecular simulation and trajectory analysis*, *J. Mol. Model.* **7**, 306 (2001).

[40] D. Van der Spoel, et al., *GROMACS: fast, flexible, and free*, *J. Comput. Chem.* **26**, 1701 (2005).

[41] B. Hess, et al., *GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation*, *J. Chem. Theory Comput.* **4**, 435 (2008).

[42] S. Pronk, et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit*, *Bioinformatics* **29**, 845 (2013).

[43] S. Páll, et al. *Proc. of EASC 2015 LNCS* **8759**, 3 (2015).

[44] L. Di Cairano, M. Gori, and M. Pettini, *Topology and Phase Transitions: A First Analytical Step towards the Definition of Sufficient Conditions.*, Entropy 23, 1414 (2021).

[45] M. Gori, *Configurational microcanonical statistical mechanics from Riemannian geometry of equipotential level sets.*, arXiv preprint, arXiv:2205.14536 (2022).

[46] G. Carlsson and A. Zomorodian, *Persistent homology - A survey*, Discrete Comput. Geom. **33**, 249 (2005).

[47] E. Munch, *A user's guide to topological data analysis*, Journal of Learning Analytics, **4**, 2 (2017).

[48] I. Donato, M. Gori, M. Pettini, G. Petri, S. De Nigris, R. Franzosi, F. Vaccarino, *Persistent homology analysis of phase transitions*, Phys. Rev. E**93**, 052138 (2016).

[49] T. Ichinomiya, *Topological data analysis gives two folding paths in HP35(nle-nle), double mutant of villin headpiece subdomain.*. Sci Rep 12, 2719 (2022).

[50] M.W. Hirsch, *Differential Topology*, (Springer, New York 1976).

[51] Y. Zhou, *A simple formula for scalar curvature of level sets in Euclidean spaces*, arXiv preprint arXiv:1301.2202, (2013).