Topological origin of protein folding transition

Loris Di Cairano, Riccardo Capelli, Ghofrane Bel-Hadj-Aissa, Pettini Marco

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
Center for Computational Engineering Science,
Department of Mathematics, RWTH Aachen University, Germany

Riccardo Capelli†

Department of Applied Science and Technology (DISAT),
Politecnico di Torino, 10129 Torino, 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

Abstract

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

* l.di.cairano@fz-juelich.de

† riccardo.capelli@polito.it

‡ ghofrane.belhadjaissa@gmail.com

§ marco.pettini@cpt.univ-mrs.fr

I. INTRODUCTION

The study of the Hamiltonian dynamical counterpart of phase transitions combined with the geometrization of Hamiltonian dynamics (where the natural motions are identified with geodesics of suitable Riemannian manifolds) has led to find that at the roots of the phase transitions 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 $\Sigma_v^{V_N} := \{V_N(q_1, \ldots, q_N) = v \in \mathbb{R}\}$ 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 modelled physical system. In fact, when the ball $M_{v=E}^{V_N} = \{(q_1, \ldots, q_N) \in \mathbb{R} | V_N(q_1, \ldots, q_N) < E\}$ 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, 2].

On the other hand, a relationship also exists between macroscopic thermodynamics and the topology of the same objects, $M_v^{V_N}$, which is expressed by [1]

$$S_N(v) = (k_B/N) \log \left[\int_{M_v^{V_N}} d^N q \right] = \frac{k_B}{N} \log \left[vol[M_v^{V_N} \setminus \bigcup_{i=1}^{\mathcal{N}(v)} \Gamma(x_c^{(i)})] \; + \sum_{i=0}^{N} w_i \; \mu_i(M_v^{V_N}) + r(N, v) \right], \tag{1}$$

where S is the configurational entropy, v is the potential energy, and the $\mu_i(M_v^{V_N})$ are the Morse indexes (in one-to-one correspondence with topology) of the manifolds $M_{v=E}^{V_N}$; in square brackets: the first term is the result of the excision of certain neighborhoods of the critical points of the interaction potential from $M_v^{V_N}$; the second term is a weighed sum of the Morse indexes, and the third term is a smooth function of N and v. On the basis of Eq.(1) one can infer that major topology changes with v of the submanifolds $M_v^{V_N}$, associated with sharp changes of the potential energy pattern of at least some of the $\mu_i(M_v^{V_N})$, 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 phase transitions stem from a suitable change of the topology of the potential level sets $\Sigma_v^{V_N}$ and, equivalently, of the manifolds $M_v^{V_N}$, when v, playing the role of the control parameter, takes a critical value v_c. This hypothesis is at the ground of a theoretical framework composed of exactly solvable models [1, 3] and two theorems [4–6] stating that an equilibrium

phase transitions are *necessarily* induced by suitable topological transitions in configuration space. Apart from its purely theoretical meaning, this topological approach has been proved interesting to investigate the origin of phase transitions in the absence of a symmetry breaking mechanism, thus in the absence of an order parameter, as is the case of a gauge model [7] and of the Kosterlitz-Thouless transition in the 2D-XY model [8].

In the present work we aim at applying the topological approach to the phase transition occurring in systems with a constitutively small number of degrees of freedom, that is, much smaller than the Avogadro number. In fact, the physical phenomenon of phase transitions is observed also in nanoscopic and mesoscopic systems, that is, also at very small numbers of degrees of freedom; this circumstance is at odds with the thermodynamic limit dogma stemming from the Yang-Lee theory [9]. The transition phenomenon tackled in what follows is the protein folding transition.

Protein folding is a very important and challenging open question in molecular biology. In fact, it is well-known that the sequence of amino acids uniquely determines the native state (*i.e.*, the compact configuration the protein assumes in physiological conditions) and understanding how the information contained in the sequence is translated into the three-dimensional native structure is 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 [10, 11], instead of linking the folding properties to the energy landscape by locating the minima of the landscape and the saddles joining them, or by undertaking the folding funnel approach [12], we focus on global property of the energy landscape which can be easily numerically computed through time averages along dynamical trajectories.

II. DEFINITION OF THE MODEL AND MD CALCULATIONS

For both the proteins studied in this system (SH3 and PYP) we generated a C_α-based Gō-model [13] via the SMOG2 [14] implementation, starting from the experimental structures obtained from the Protein Data Bank (1FMK [15] for SH3 and 3PHY [16] for PYP). In this model, only the C_α atom of every amino acid is considered and the potential is defined as:

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

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, angle 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 $\varepsilon^{\text{n-nat}} = \varepsilon$.

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

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

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 [17] 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 [18, 19], 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 [20].

III. OBSERVABLES AND METHODOLOGY

We sample the value of the characteristic observables along the dynamical trajectories at fixed temperatures of the minimalistic model defined in [21]. Minimalistic models are those where the polymer is described at a coarse-grained level as a chain of N beads, where N is the number of amino acids; no explicit water molecules are considered and the solvent is taken into account only by means of effective interactions among the monomers.

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

heat at constant volume

$$C_V = \frac{\langle E_{tot}^2 \rangle - \langle E_{tot} \rangle^2}{\tilde{K}_B T^2} \ , \tag{4}$$

in function of the temperature, where E_{tot} is the total energy and $\tilde{K}_B$ is the normalized Boltzmann constant defined as $\tilde{K}_B = N N_A K_B$ (being N_A the Avogadro constant and K_B the Boltzmann constant); *(iii)* the dimensionless variance of the Ricci-scalar curvature $\mathcal{R}(q)$ as a function of the temperature

$$\sigma_{\mathcal{R}} = \frac{\sqrt{(\langle \mathcal{R}^2 \rangle - \langle \mathcal{R} \rangle^2)/N}}{\langle \mathcal{R} \rangle / N} \ , \tag{5}$$

where

$$\mathcal{R}(q) = \frac{Lap\, V}{\|\nabla V\|^2} - \frac{Tr[(Hess\, V)^2]}{\|\nabla V\|^2} + 2\frac{\|Hess\, V \nabla V\|^2}{\|\nabla V\|^4} - 2\frac{\langle \nabla V, Hess\, V \nabla V \rangle}{\|\nabla V\|^4}. \tag{6}$$

(iv) the temperature T in function of the canonical ensemble energy defines as

$$\epsilon = \tilde{K}_B T/2 + \langle V \rangle / N \ , \tag{7}$$

where we recall that V is the total potential filed; *(v)* the entropy of the system evaluated using

$$S = \int \frac{2}{\tilde{K}_B T(\epsilon)} \, d\epsilon \ , \tag{8}$$

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) (see left-hand panel of FIG.1), of 57 amino acids; 2 random sequences of the same 57 amino acids (RDMa,b); and the photoactive yellow protein (PYP) (see right-hand panel of FIG.1) composed by 125 amino acids. We stress how the simulations are run for several random sequences outlining similar results and we here report only two of them for the sake of simplicity. The randomization is implemented using the SH3 coarse grained potential described in [21] 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 [22–27] and each trajectory is run until 1μs with 0.01ps temporal steps. Averages and fluctuations are evaluated over 1000 outputs 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.

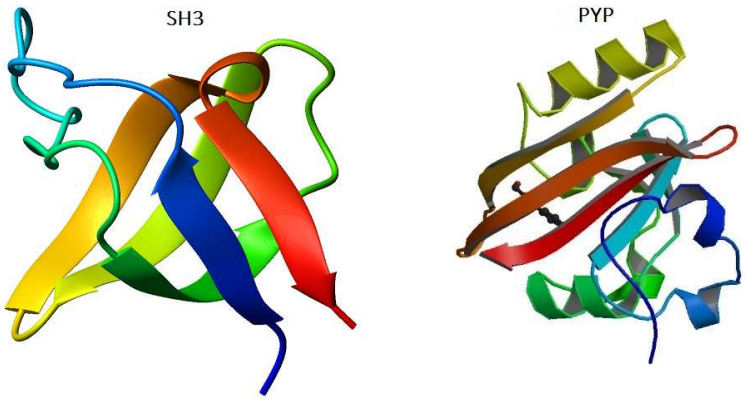

FIG. 1. Cartoon representation of SH3 (left-hand panel) and PYP (right-hand panel).

IV. GEOMETRICAL SIGNATURES OF TOPOLOGICAL CHANGES

Let us associate to the protein a configuration space, M, which coincides with the Euclidean space $\mathbb{R}^{3N}$ where N is the number of the system particles. Thus, by setting $n = 3N$, we have $M = \mathbb{R}^{3N}$ and, therefore, M can be equipped with the Euclidean metric tensor, g_Γ. More precisely, by introducing an orthonormal basis $\{\partial_{q^i}\}_{i=1}^n \subset M$ whose dual basis $\{dq^i\}_{i=1}^n$ is defined by the relation $dq^i(\partial_{q^j}) = \delta^i_j$, the Euclidean metric tensor is:

$$g_M := \delta_{ij}\, dq^i \otimes dq^j. \tag{9}$$

Given a regular Hamiltonian function $V : M \longrightarrow \mathbb{R}$, one defines the potential level sets to be the following collection of sets:

$$\Sigma_{\bar{V}} := \left\{ \boldsymbol{q} \in M \,|\, V(\boldsymbol{q}) = \bar{V} \right\}. \tag{10}$$

Therefore, the metric tensor on M takes the form:

$$\widetilde{g}_M = d\bar{V} \otimes d\bar{V} + (g_{\Sigma_{\bar{V}}})_{ij}\, dy^i dy^j \tag{11}$$

so that M can be reordered as follows:

$$M = \bigcup_{\bar{V}} \Sigma_{\bar{V}}, \tag{12}$$

where the symbol $\bigcup$ is meant as the formal union on all possible level sets and $\{y^1, \ldots, y^{3N-1}\}_{\Sigma_{\bar{V}}}$ is the frame of coordinates on the potential level set $\Sigma_{\bar{V}}$.

In order to study the geometry of the level set $\Sigma_{\bar{V}}$, we introduce the unit normal vector to $\Sigma_{\bar{V}}$:

$$\boldsymbol{\nu} := \frac{\nabla^{g_M} V}{\|\nabla^{g_M} V\|_{g_M}} \tag{13}$$

where $\nabla^{g_M} := (\partial_{q^1}, \ldots, \partial_{q^n})$ and $\|\cdot\|_{g_M}$ are, respectively, the gradient operator and the norm with respect to the metric tensor given in Eq. (9) and, from now on, we will omit to specify the metric unless necessary.

Thus, the curvature properties of the level sets can be analyzed through the introduction of the Weingarten operator [8]:

$$\mathcal{W}_{\boldsymbol{\nu}}(\boldsymbol{X}) := -\nabla_{\boldsymbol{X}} \boldsymbol{\nu}, \qquad \forall\, \boldsymbol{X} \in T_x \Sigma_{\bar{V}} \tag{14}$$

where $T_x \Sigma_{\bar{V}}$ is the tangent space to the level set $\Sigma_{\bar{V}}$ on the point $\boldsymbol{x} \in \Gamma$.

The geometric observables to which we are interested in since they can be directly connected to the topology properties of $\Sigma_{\bar{V}}$ are the dispersion of the principal curvatures $\sigma(k_i)^2$ and the scalar curvature of the potential level sets.

The dispersion of principal curvature $\sigma(k_i)^2$ is defined by [8]:

$$\sigma(k_i)^2 := \frac{Tr[\mathcal{W}_{\boldsymbol{\nu}}^2]}{n-1} - \frac{(Tr[\mathcal{W}_{\boldsymbol{\nu}}])^2}{(n-1)^2} \tag{15}$$

whereas the scalar curvature of $\Sigma_{\bar{V}}$ is [28]:

$$\mathcal{R}_{\Sigma_{\bar{V}}} := Tr[\mathcal{W}_{\boldsymbol{\nu}}]^2 - Tr[\mathcal{W}_{\boldsymbol{\nu}}^2] \tag{16}$$

In order to have an expression for such observables we need to explicitly compute the traces of the Weingarten operator and of its squared.

For what concerned the trace, we have:

$$Tr[\mathcal{W}_{\boldsymbol{\nu}}] = \frac{Lap\, V}{\|\nabla V\|} - \frac{\langle \nabla V, Hess\, V\ \nabla V \rangle}{\|\nabla V\|^3} \tag{17}$$

where $Lap\, V$ and $Hess\, V$ are, respectively, the Laplacian and the Hessian of the potential function V with respect to g_M. We note that we treat the Hessian as a linear application $Hess\, V : \mathbb{R}^n \longrightarrow \mathbb{R}^n$.

The trace of the squared of the Weingarten operator is [8]:

$$Tr[\mathcal{W}_{\boldsymbol{\nu}}^2] = \frac{Tr[(Hess\, V)^2]}{\|\nabla V\|^2} + \frac{\langle \nabla V, Hess\, V\ \nabla V \rangle^2}{\|\nabla V\|^6} - 2\frac{\|Hess\, V\ \nabla V\|^2}{\|\nabla V\|^4} \tag{18}$$

Thus, by means of the Pinkall's theorem [29], we can connect the dispersion of principal curvatures to the topology [8]:

$$\begin{aligned}\langle\sigma^2(k_i)\rangle_\mu := \frac{\int_{\Sigma_{\bar{V}}} \sigma^2(k_i)\, d\mu}{\int_{\Sigma_{\bar{V}}} d\mu} &= \left[\int_{\Sigma_{\bar{V}}} (\sigma^2(k_i))^{n/2}\, d\eta\right]^{2/n} - r(\Sigma_{\bar{V}}) \\ &= \left[Vol(\mathbb{S}^n)\sum_{i=1}^{n}\left(\frac{i}{n-i}\right)^{n/2-i} b_i(\Sigma_{\bar{V}})\right]^{2/n} - r(\Sigma_{\bar{V}})\end{aligned} \tag{19}$$

where $d\eta$ is the normalized Riemannian measure on the level set $\Sigma_{\bar{V}}$ defined by

$$d\eta := \frac{d\mu}{\int_{\Sigma_{\bar{V}}} d\mu}, \tag{20}$$

and $Vol(\mathbb{S}^n)$ is the volume of the unit n-sphere, $r(\Sigma_{\bar{V}})$ is a small correction and $b_i(\Sigma_{\bar{V}})$ are the Betti numbers.

A further information is given by the scalar curvature of the potential level sets. In fact, the scalar curvature is the sum of all the sectional curvatures and, therefore, the fluctuations of the scalar curvature can be related to the fluctuations of scalar curvature and one can exploit Overholt theorem [30]:

$$\Delta(sectional) > \left[\frac{vol(\mathbb{S}^N_1)\sum_{k=0}^{N} b_k(\Sigma_{\bar{V}})}{2\ vol(\Sigma_{\bar{V}})}\right]^{2/N} \tag{21}$$

The variance of the scalar curvature $\mathcal{R}(\Sigma_{\bar{V}})$ is

$$\Delta^2(scal) = \frac{\langle\mathcal{R}^2(\Sigma_{\bar{V}})\rangle - \langle\mathcal{R}(\Sigma_{\bar{V}})\rangle^2}{N(N-1)} \simeq \Delta(sectional) \tag{22}$$

and

$$\langle\mathcal{R}^n(\Sigma_{\bar{V}})\rangle = \frac{\int_{\Sigma_{\bar{v}}} \mathcal{R}^n(\Sigma_{\bar{V}})\dfrac{d\mu}{\|\nabla V\|}}{\int_{\Sigma_{\bar{v}}} \dfrac{d\mu}{\|\nabla V\|}}, \tag{23}$$

V. AVERAGES OF GEOMETRIC OBSERVABLES

We now discuss how we extrapolated information about the topology associated to the potential landscape of the studied systems from the MD data.

Hence, our aim is that to understand how the phase transition, i.e., the protein folding, is guided by a change of topology of the potential level sets. Therefore, we have to connect the topological properties, namely, average of observables with respect to the geometric measure, with thermodynamic properties, namely, average of observable with respect to a statistical ensemble.

Let us start, then, noticing that the MD simulation have been performed in the canonical ensemble, this implies that the all the thermodynamics average has to be done with respect to the canonical partition function. Thus, we introduce the configurational canonical partition function:

$$\mathcal{Z}(n,T) := \int_M e^{-V(\boldsymbol{q})/T}\, d\boldsymbol{q}. \tag{24}$$

where the Boltzmann constant K_B has been equated to one.

The foliation introduced in Eq. (12) allows to rewrite the integral above as follows:

$$\mathcal{Z}(n,T) = n\int_0^\infty e^{-n\bar{v}/T}\left(\int_{\Sigma_{\bar{v}}} \frac{d\mu}{\|\nabla V\|}\right) d\bar{v}, \tag{25}$$

where $d\mu$ is the induced metric on the potential level set $\Sigma_{\bar{v}}$, $\bar{V} = n\bar{v}$, i.e., $\bar{v}$ is the value of the potential per degrees of freedom.

Thus, the canonical average, $\langle\cdot\rangle_C$ of a generic (geometric) observable, $A : M \longrightarrow \mathbb{R}$ is

$$\langle A\rangle_C(n,T) := \frac{1}{\mathcal{Z}(n,T)}\int_0^\infty n\, e^{-n\bar{v}/T}\left(\int_{\Sigma_{\bar{v}}} \frac{A(\Sigma_{\bar{v}})}{\|\nabla V\|}\, d\mu^{\Sigma_{\bar{v}}}\right) d\bar{v}, \tag{26}$$

The MD simulation that we performed allows to sample any observable along trajectories, therefore, the canonical average can be equivalently measured along the dynamics [33]:

$$\langle A\rangle_C(n,T) = \frac{1}{t}\lim_{t\to\infty}\int_0^t A(\tau)\, d\tau. \tag{27}$$

Let us now go back to the the Pinkall theorem (19) that we recall below:

$$\langle\sigma^2(k_i)\rangle_{geo} := \frac{\displaystyle\int_{\Sigma_{\bar{v}}} \sigma^2(k_i)\, d\mu}{\displaystyle\int_{\Sigma_{\bar{v}}} d\mu} \tag{28}$$

In this case, we should average the dispersion of the principal curvature with respect to a geometric measure but, because of the canonical average, we cannot achieve this expression at a finite number of degrees of freedom. Nevertheless, it is possible to get as closer as possible proceeding as follows.

Let us denote by Λ^2 the dispersion of the principal curvature obtained evaluating Eq. (15) along the trajectory, namely, the data obtained from the right-hand side of Eq. (27). Theoretically, they are associated to the canonical average, thus, the best approximation of the numerator in Eq. (28) is given by:

$$\langle \|\nabla V\| \, \Lambda^2 \rangle_C = \frac{1}{\mathcal{Z}(n,T)} \int_0^\infty n \, e^{-n\bar{v}/T} \left(\int_{\Sigma_{\bar{v}}} \Lambda^2 \, d\mu \right) d\bar{v}, \tag{29}$$

whereas for the denominator, we have:

$$\langle \|\nabla V\| \rangle_C = \frac{1}{\mathcal{Z}(n,T)} \int_0^\infty n \, e^{-n\bar{v}/T} \left(\int_{\Sigma_{\bar{v}}} d\mu \right) d\bar{v}, \tag{30}$$

Now, by dividing Eq. (29) by (30), we get:

$$\frac{\langle \|\nabla V\| \, \Lambda^2 \rangle_C}{\langle \|\nabla V\| \rangle_C} = \frac{\displaystyle\int_0^\infty n \, e^{-n\bar{v}/T} \left(\int_{\Sigma_{\bar{v}}} \Lambda^2 \, d\mu \right) d\bar{v}}{\displaystyle\int_0^\infty n \, e^{-n\bar{v}/T} \left(\int_{\Sigma_{\bar{v}}} d\mu \right) d\bar{v}}, \tag{31}$$

At this step, it is worth noting that, for large values of n, the canonical measure narrows around the 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 thermodynamic limit, the partition function reduces to

$$\mathcal{Z}(n,T) \approx n \, e^{-n\bar{v}(T)/T} \int_{\Sigma_{\bar{v}(T)}} \frac{d\mu^{\Sigma_{\bar{v}(T)}}}{\|\nabla V\|}, \tag{32}$$

and the average in Eq. (31) reads:

$$\frac{\langle \|\nabla V\| \, \Lambda^2 \rangle_C}{\langle \|\nabla V\| \rangle_C} \overset{n\to\infty}{\longrightarrow} \frac{\displaystyle\int_{\Sigma_{\bar{v}(T)}} \Lambda^2 \, d\mu^{\Sigma_{\bar{v}(T)}}}{\displaystyle\int_{\Sigma_{\bar{v}(T)}} d\mu^{\Sigma_{\bar{v}(T)}}} \approx \langle \sigma^2(k_i) \rangle_{geo}, \tag{33}$$

By looking at Eq. (31), we conclude that, at relatively low number of degrees of freedom, there exists a collection of potential level sets which contribute to the canonical measure and the exponential function $e^{-n\bar{v}/T}$ weights all these level sets. The resulting effect is a *dispersion* of the points of the representative curve of $\sigma^2(k_i)$ which is evident in the SH3 protein since its degrees of freedom are $n_{SH3} = 171$. As soon as n increases, such a dispersion shrinks as one can perceive by comparing the plots of $\sigma^2(k_i)$ associated to SH3 with that

of PYP ($n_{PYP} = 375$) in Figs. [?]. A further contribution to such a dispersion can be attributed to the presence of the quantity $\langle \|\nabla V\| \rangle_C$.

For what concerned the measure of the scalar curvature in Eq. (23), this can be directly achieved by applying the definition (26), that is:

$$\langle \mathcal{R} \rangle_C(n,T) = \frac{\displaystyle\int_0^\infty n\, e^{-n\bar{v}/T} \left(\int_{\Sigma_{\bar{v}}} \frac{\mathcal{R}}{\|\nabla V\|}\, d\mu \right) d\bar{v}}{\displaystyle\int_0^\infty n\, e^{-n\bar{v}/T} \left(\int_{\Sigma_{\bar{v}}} \frac{d\mu}{\|\nabla V\|} \right) d\bar{v}}, \tag{34}$$

and for a large number of particles, we have:

$$\langle \mathcal{R} \rangle_C(n,T) \xrightarrow{n \gg 1} \frac{\displaystyle\int_{\Sigma_{\bar{v}}} \frac{\mathcal{R}}{\|\nabla V\|}\, d\mu}{\displaystyle\int_{\Sigma_{\bar{v}}} \frac{d\mu}{\|\nabla V\|}} \approx \langle \mathcal{R} \rangle \tag{35}$$

For such an observable the dispersion of the representative points does not appear and, as anticipated above, this is due to the fact that we do not divide the average of the scalar curvature by $\langle \|\nabla V\| \rangle_C$.

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. In particular the inflection point of the caloric curve is typical of a second order phase transition [31, 32]. In Figure 4 the specific heat and the caloric curve are reported for the PYP protein and also in this case 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 Figure 2 - could be compatible with a first order phase transition [31, 32]. 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 Figure 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.

Finally, in Figures 7 and 8 the dispersions of principal curvatures of the equipotential level sets in configuration space are reported for the SH3 and PYP proteins, respectively, and for the randomized sequences. This quantity displays well evident peculiar patterns when plotted as a function of the potential energy value per degree of freedom. These patterns are less clear when plotted against temperature, even though the presence of cusps can be guessed by means of different polynomial fittings 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]

$$\begin{aligned} S_N(E) &= \frac{k_{\rm B}}{N}\log\int_{\Sigma_E^N}\frac{d\mu}{\|\nabla H\|} \\ &\simeq \frac{k_{\rm B}}{N}\log\left[\mathrm{vol}(\mathbb{S}_1^{N-1})\sum_{i=0}^{N} b_i(\Sigma_E^N) + r_1(E)\right] + r_2(E)\ , \end{aligned} \tag{36}$$

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 is the sum of sectional curvatures so that its variance contains the variance of the sectional curvatures [7], thus the quantity

$$\Delta(sec) > \left[\frac{vol(\mathbb{S}_1^N)\sum_{k=0}^{N} b_k(\Sigma_E)}{2\ vol(\Sigma_E)}\right]^{2/N} \tag{37}$$

in strict analogy with Eqs.(21) and (22), 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.(19).

VII. DISCUSSION

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.

The global geometry/topology of both total energy and potential energy landscapes is found 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.

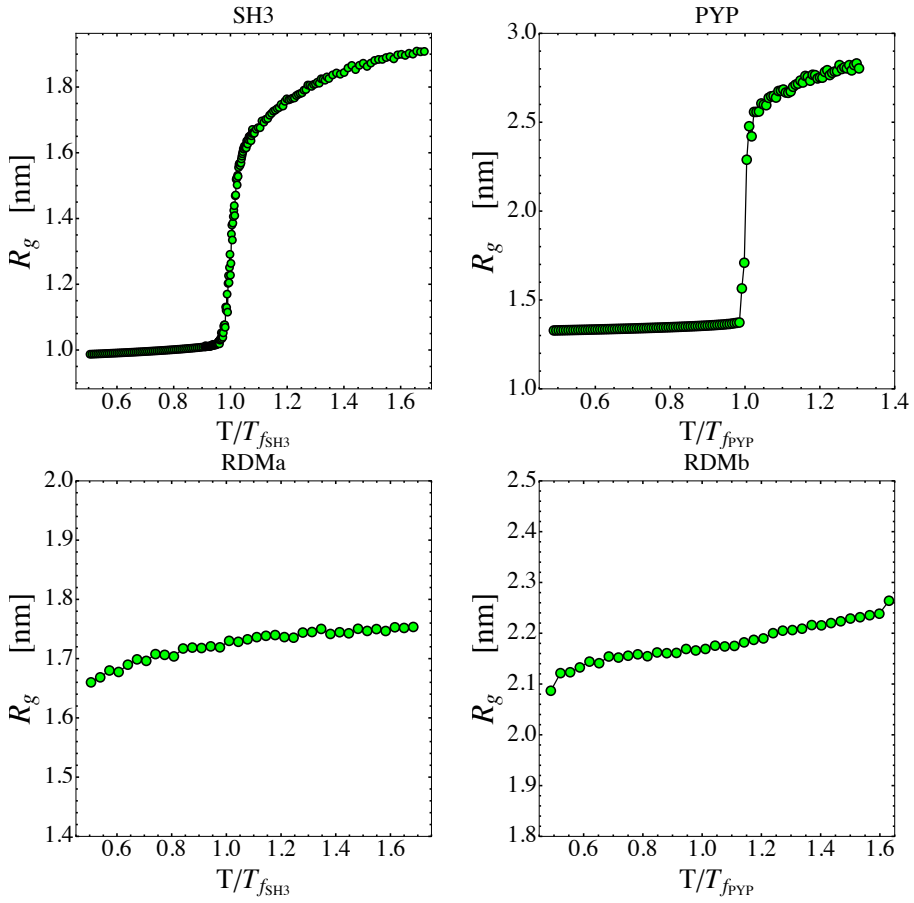

FIG. 2. Plots of the gyration radius for the different sequences. It is evident how only the good folders SH3 and PYP exhibit a temperature dependence typical of the folding transition (upper panels) which are lost for the randomized sequences (lower panels). $T_{f_{SH3}}$ and $T_{f_{PYP}}$ identify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

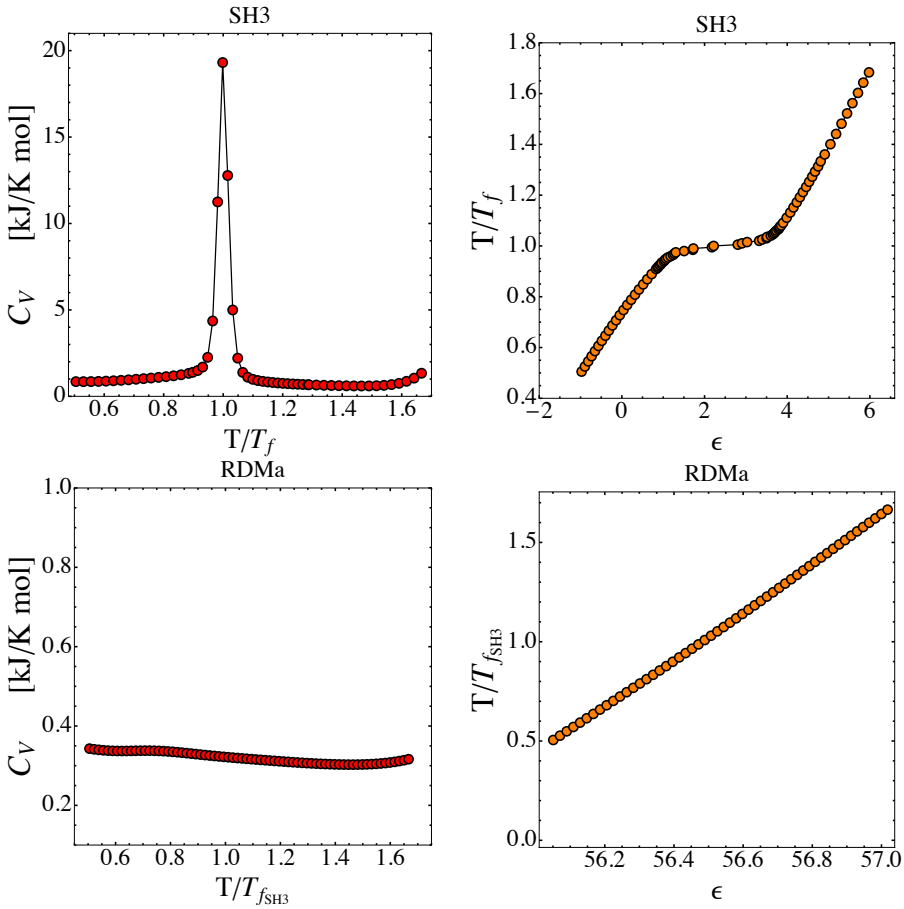

FIG. 3. The specific heat and the caloric curve for the SH3 protein show patterns typical of a phase transition (upper panels). These features are lost in the case of the randomized version of the correct sequence of the SH3 protein (lower panels).

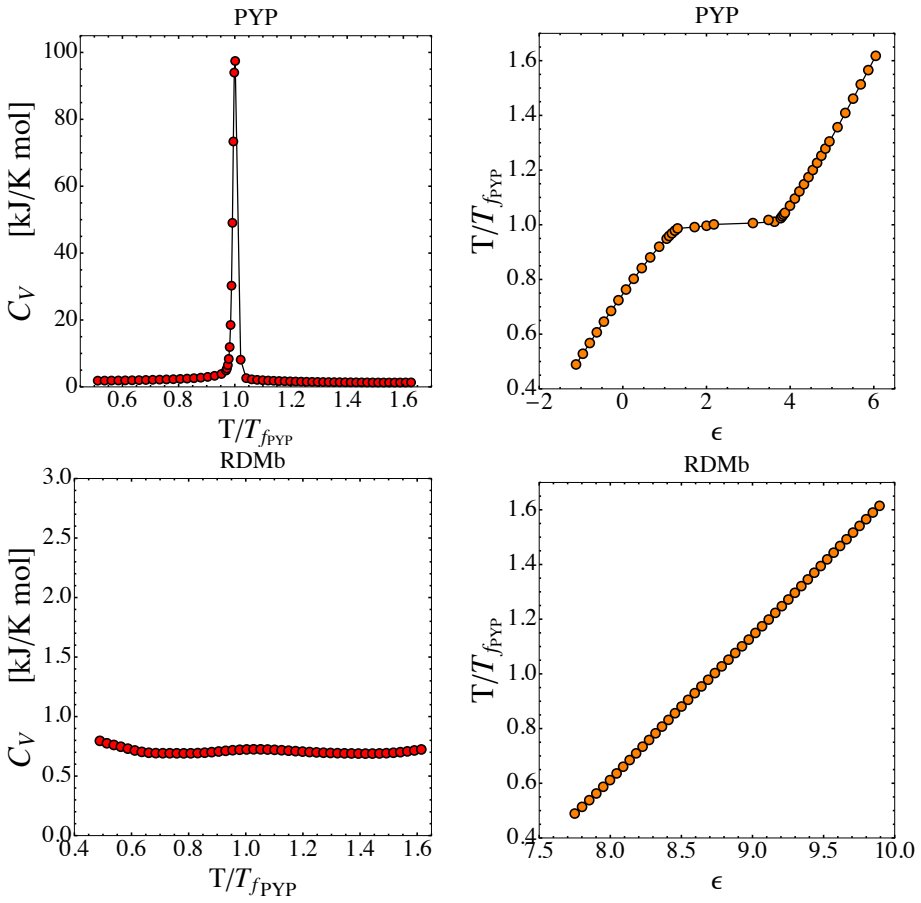

FIG. 4. The specific heat and the caloric curve for the PYP protein show patterns typical of a phase transition (upper panels). These features are lost in the case of the randomized version of the correct sequence of the PYP protein (lower panels).

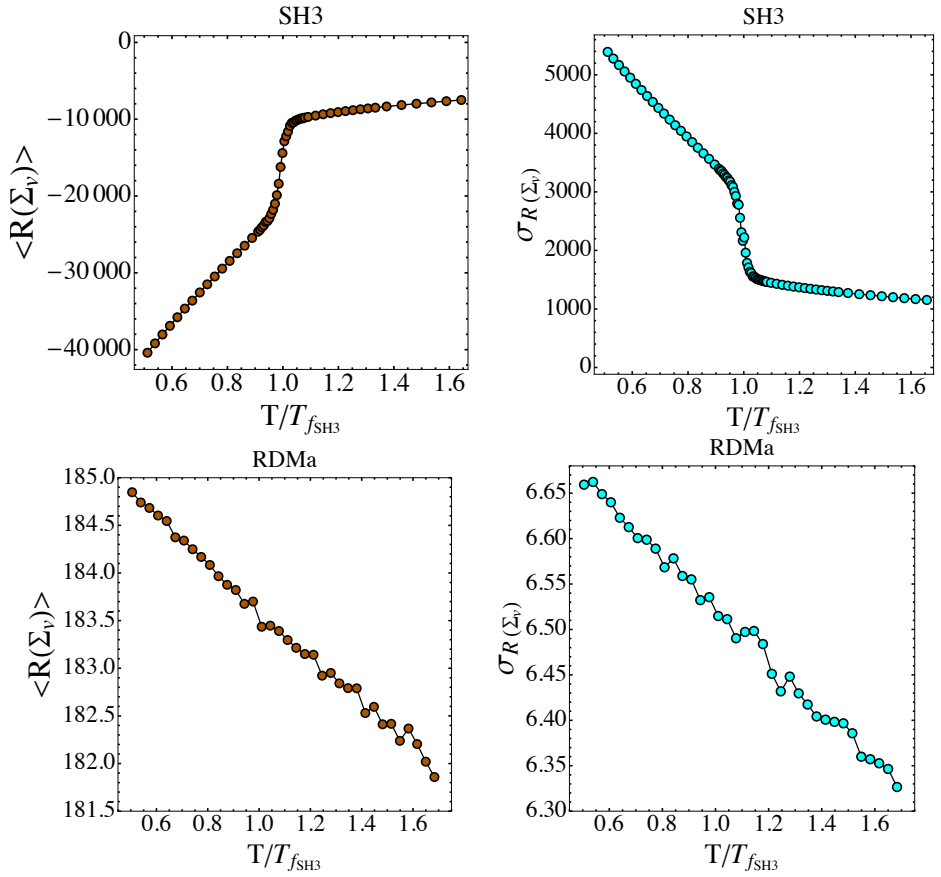

FIG. 5. The total scalar curvature and its variance, of equipotential level sets, are reported as functions of temperature for the SH3 protein (upper panels) and for its randomized sequence of amino acids (lower panels).

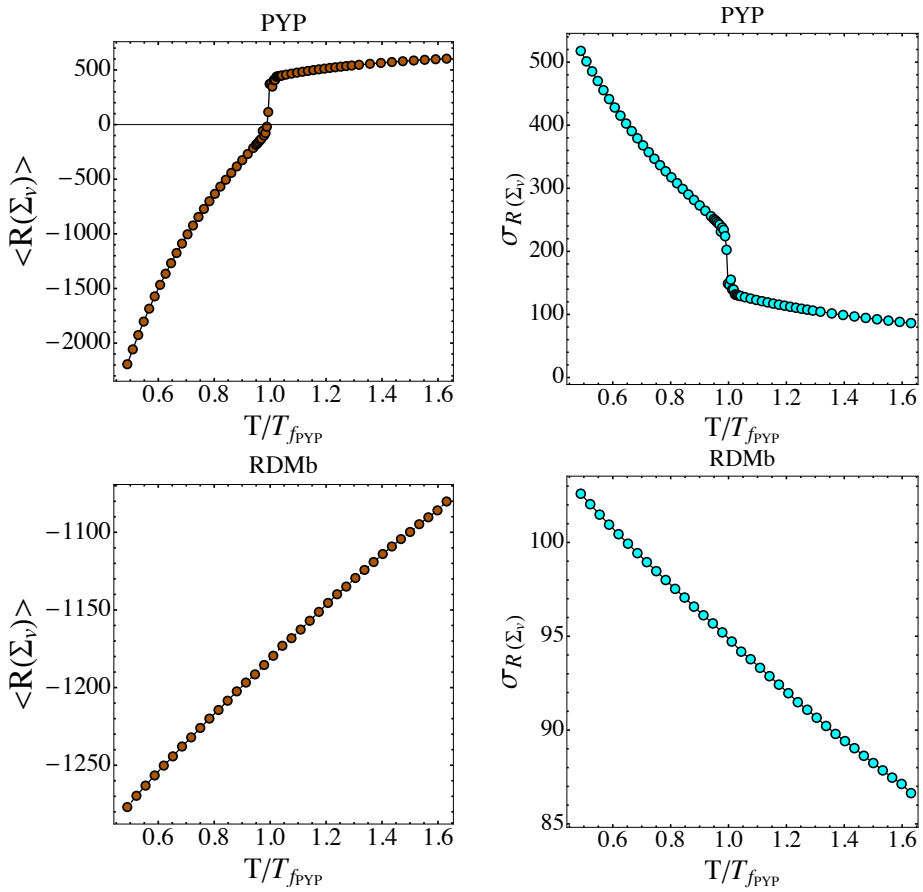

FIG. 6. The total scalar curvature and its variance, of equipotential level sets, are reported as functions of temperature for the PYP protein (upper panels) and for its randomized sequence of amino acids (lower panels).

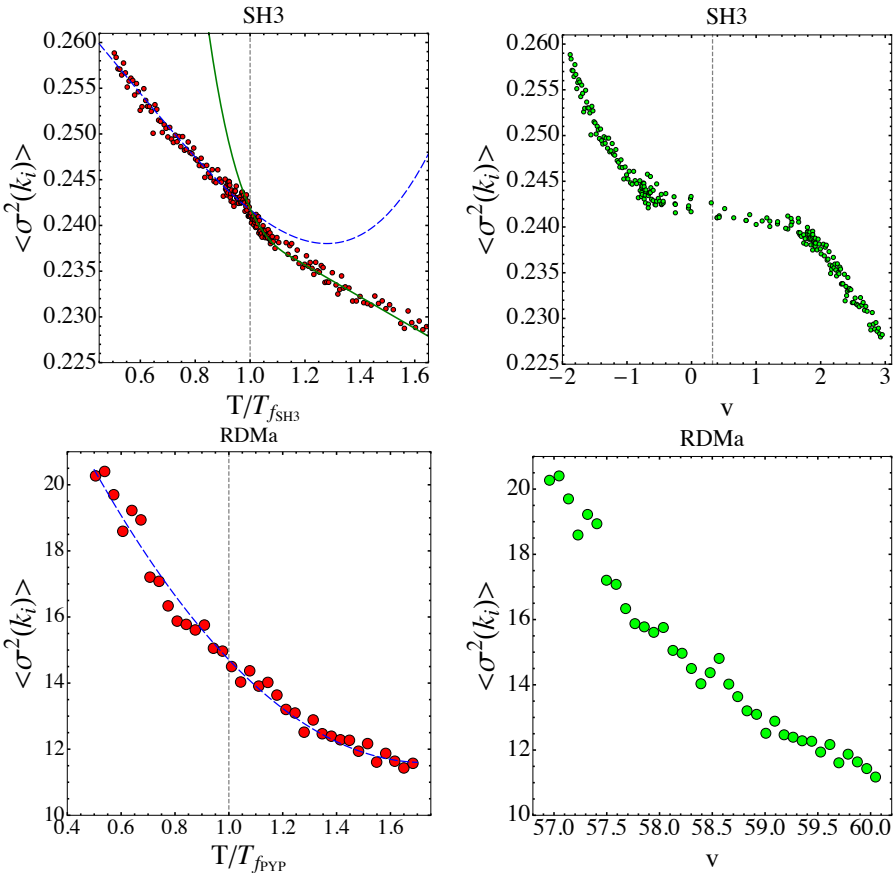

FIG. 7. The variance of the principal curvatures of the equipotential level sets is reported as a function of temperature (left panel) and of the potential energy per degree of freedom v for both the SH3 protein (upper panels), and for its randomized version(lower panels).

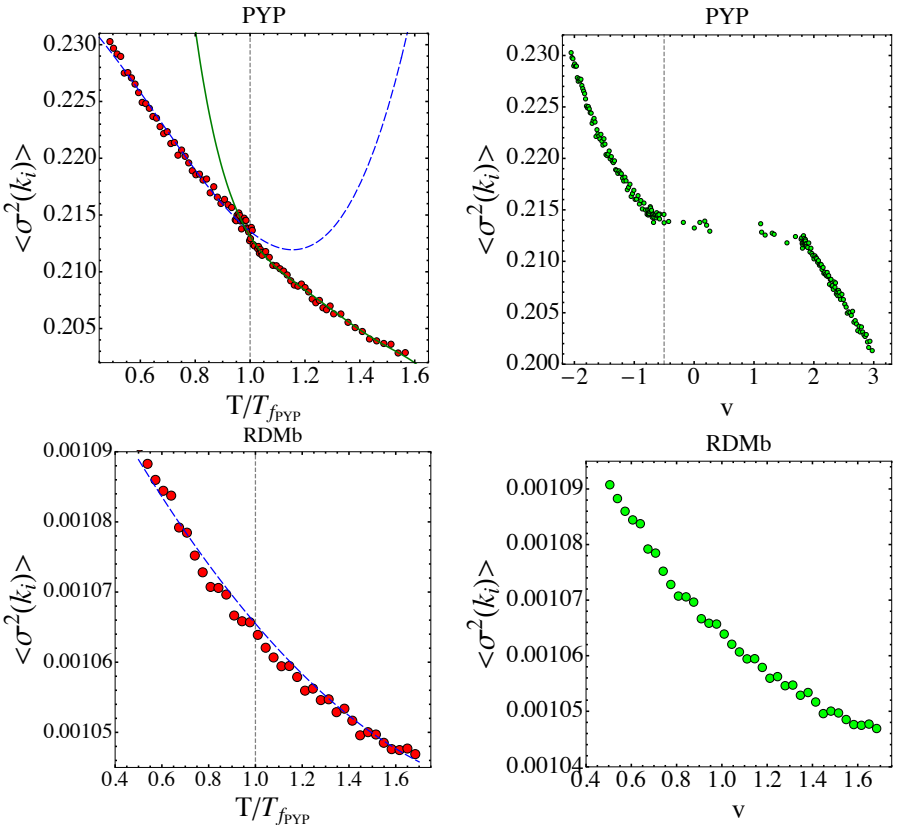

FIG. 8. The variance of the principal curvatures of the equipotential level sets is reported as a function of temperature (left panel) and of the potential energy per degree of freedom v for both the PYP protein (upper panels), and for its randomized version(lower panels).

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 Co-fund 731473 (Project Q-CLOCKS), Italy.

---

[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*, Physical Review E**54**, 5969 (1996).

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

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

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

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

[7] 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).

[8] Ghofrane Bel-Hadj-Aissa, Matteo Gori, Roberto Franzosi, and Marco Pettini, *Geometrical and topological study of the Kosterlitz-Thouless phase transition in the XY model in two dimensions*, Journal of Statistical Mechanics: Theory and Experiment **2**, 023206 (2021).

[9] C.N. Yang, and T.D. Lee, *Statistical theory of equations of state and phase transitions I. Theory of condensation*, Phys. Rev. **87**, 404 - 409 (1952); T.D. Lee, and C.N. Yang, *Statistical theory of equations of state and phase transitions II. Lattice gas and Ising model*, Phys. Rev. **87**, 410 - 419 (1952).

[10] Lorenzo N Mazzoni and Lapo Casetti, *Curvature of the energy landscape and folding of model proteins*, Physical Review Letters **97**, 18104 (2006).

[11] Lorenzo N Mazzoni and Lapo Casetti, *Geometry of the energy landscape and folding transition in a simple model of a protein*, Physical Review E**77**, 051917 (2008).

[12] J Nelson Onuchic, Peter G Wolynes, Z Luthey-Schulten, and Nicholas D Socci, *Toward an outline of the topography of a realistic protein-folding funnel*, Proceedings of the National Academy of Sciences **92**, 3626 (1995).

[13] Cecilia Clementi, Hugh Nymeyer, and José Nelson 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 Molecular Biology **298**, 937 (2000).

[14] Jeffrey K Noel, Mariana Levi, Mohit Raghunathan, Heiko Lammert, Ryan L Hayes, José N Onuchic, and Paul C Whitford, *SMOG 2: a versatile software package for generating structure based models*, PLoS computational biology **12**, e1004794 (2016).

[15] Wenqing Xu, Stephen C Harrison, and Michael J Eck, *Three-dimensional structure of the tyrosine kinase c-Src*, Nature **385**, 595 (1997).

[16] Petra Düx, Gilles Rubinstenn, Geerten W Vuister, Rolf Boelens, Frans A A Mulder, Karl Hard, Wouter D Hoff, et al., *Solution structure and backbone dynamics of the photoactive yellow protein*, Biochemistry **37**, 12689 (1998).

[17] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilard Pall, Jeremy C. Smith, Berk Hess, and Erik Lindahl, *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers*, SoftwareX **1-2**, 19 (2015).

[18] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth A Tribello, Pavel Banas, Alessandro Barducci, Mattia Bernetti, Peter G Bolhuis, Sandro Bottaro, Davide Branduardi, et al., *Promoting transparency and reproducibility in enhanced molecular simulations*, Nature Methods **16**, 670 (2019).

[19] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi, *PLUMED 2: New feathers for an old bird*, Computer Physics Communications **185**, 604 (2014).

[20] Guido Tiana and Ludovico Sutto, *Equilibrium properties of realistic random heteropolymers and their relevance for globular and naturally unfolded proteins*, Physical Review E**84**, 061910

(2011).

[21] P. Das, S. Matysiak, C. Clementi, *Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes*, Proceedings of the National Academy of Sciences **102**, 10141 (2005).

[22] H.J.C. Berendsen, et al. *Comp. Phys. Comm.* **91**, 43 (1995).

[23] E. Lindahl, et al. *J. Mol. Model.* **7**, 306 (2001).

[24] D. Van der Spoel, et al. *J. Comput. Chem.* **26**, 1701 (2005).

[25] B. Hess, et al. *J. Chem. Theory Comput.* **4**, 435 (2008).

[26] S. Pronk, et al. *Bioinformatics* **29**, 845 (2013).

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

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

[29] Ulrich Pinkall, *Inequalities of Willmore type for submanifolds*, Mathematische Zeitschrift, **193**, 241 (1986).

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

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

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

[33] Ghofrane Bel-Hadj-Aissa, Matteo Gori, Vittorio Penna, Giulio Pettini, and Roberto Franzosi, *Geometrical Aspects in the Analysis of Microcanonical Phase-Transitions*, Entropy **22**, 2020.