Thermal unfolding of myoglobin in the Landau-Ginzburg-Wilson approach

Xubiao Peng,[1,*] Adam K. Sieradzan,[2,†] and Antti J. Niemi[3,4,5,‡]

[1]*Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia V6T1Z4, Canada*

[2]*Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdańsk, Poland*

[3]*Department of Physics and Astronomy, Uppsala University, P. O. Box 803, S-75108, Uppsala, Sweden*

[4]*Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France*

[5]*Department of Physics, Beijing Institute of Technology, Haidian District, Beijing 100081, People's Republic of China*

(Received 17 July 2015; revised manuscript received 6 November 2015; published 16 December 2016)

The Landau-Ginzburg-Wilson paradigm is applied to model the low-temperature crystallographic Cα backbone structure of sperm whale myoglobin. The Glauber protocol is employed to simulate its response to an increase in ambient temperature. The myoglobin is found to unfold from its native state by a succession of α-helical intermediates, fully in line with the observed folding and unfolding patterns in denaturation experiments. In particular, a molten globule intermediate is identified with experimentally correct attributes. A detailed, experimentally testable contact map is constructed to characterize the specifics of the unfolding pathway, including the formation of long-range interactions. The results reveal how the unfolding process of a protein is driven by the interplay between, and a successive melting of, its modular secondary structure components.

DOI: 10.1103/PhysRevE.94.062405

I. INTRODUCTION

According to a paradigm by Anfinsen [1], under isothermal physiological conditions the native structure of a protein relates to the global minimum of Helmholtz free energy F,

$$F = U - TS, \tag{1}$$

where U is the internal energy, S is the entropy, and T is the temperature. The Landau-Ginzburg-Wilson (LGW) approach [2–8] is a systematic method to approximate (1), in terms of the symmetry properties of the underlying physical system. The approach was originally conceived to describe the static properties of phase transitions and critical phenomena. There, it has found numerous applications, for example, in ordinary and quantum fluids, magnetic materials, and superconductors. Subsequently the LGW approach has been expanded to describe time-dependent critical phenomena. It has also been extended to model, e.g., pattern formation in nonequilibrium statistical systems and chaotic behavior in nonlinear dynamics [7]. Even aspects of fundamental string theory, singularity theory, and proof of existence of solutions to certain nonlinear partial differential equations relate to the LGW approach [8].

In the present article we develop and apply the Landau-Ginzburg-Wilson approach to model protein dynamics. As an example, we consider the way myoglobin folds and unfolds when the ambient temperature increases. Myoglobin is the first protein to have its stable three-dimensional structure determined by x-ray crystallography [9]. It is one of the most widely studied protein structures [10].

The *protein folding problem* remains under active scrutiny [11–13]. Many theoretical proposals have been presented to explain how the folding of a protein might proceed [14–17]. Recently, a soliton-based method which is built on the LGW approach has been presented to describe both static folded proteins [18,19] and aspects of protein dynamics [20–22]. This method has been tested and validated computationally by comparing its predictions both with coarse-grained [23,24] and all-atom force fields [25,26]. The results confirm that the folding of a simple protein proceeds by a formation of soliton and in a manner that can be modeled using the Landau-Ginzburg-Wilson paradigm; see, e.g., Ref. [20] in the case of a villin headpiece.

Solitons are collective oscillations, and they emerge and dissolve when a physical system undergoes a large scale structural self-organization. Accordingly, the involvement of solitons in the protein (un)folding process is fully in line with the expected "cooperative" character of the protein folding process [27–29].

Here we combine the LGW approach with Glauber dynamics to study in detail how myoglobin folds and unfolds. Glauber dynamics is a Markov chain Monte Carlo (MC) method that is widely used to describe near-equilibrium relaxation dynamics of a statistical system towards an equilibrium Gibbsian state [30–32].

In the case of myoglobin, instead of temperature variations, most experiments have thus far utilized denaturants to study the unfolding and folding dynamics. However, in a computational approach, it is more convenient to use the ambient temperature as the variable. Moreover, the experiments have mainly concentrated on the heme-free apomyoglobin [33–45]. The heme containing myoglobin has also been investigated [46–51] but, due to apparent complications with the binding of the heme, the studies have been limited to the unfolding process. In both cases, the unfolding of the native state proceeds in stages with several folding intermediates. In the case of apomyoglobin, the folding appears to proceed inversely to the unfolding. The dynamics is also very similar in both cases, except that in the apomyoglobin the F helix [10] is initially disordered [37] while in the heme containing myoglobin the F helix is initially stable but the first to become disordered

*xubiaopeng@gmail.com

†adams86@wp.pl

‡Antti.Niemi@physics.uu.se

when temperature and/or denaturation increases [46–51]. The unfolding of the F helix is followed by an intermediate molten globule in both cases [33–51]. When denaturation and/or temperature increases further, the overall helicity of the molten globule rapidly decreases as the helices B, C, D, and E start to unfold. Finally, the remaining helices A, G, and H lose their stability and the structure becomes a random chain [46–51].

Here we show that the experimentally observed unfolding pattern of myoglobin can be accurately reproduced by a combination of the Landau-Ginzburg-Wilson approach with Glauber dynamics. In line with the apomyoglobin experiments, we also show that the folding proceeds inversely to the unfolding. Moreover, for future experimental scrutiny, we propose a detailed contact map that predicts the way helix formation proceeds and how long-range contacts disappear and appear between the different helical segments during the unfolding and the folding processes, respectively. An experimental analysis of the contact map we propose should reveal the extent of validity of the LGW approach in the case of myoglobin.

Our experimental reference conformation is the Protein Data Bank (PDB) [52] structure 1ABS [53] of wild-type sperm whale heme containing myoglobin. There are 154 amino acids, indexed $i = 0 \ldots 153$ in the PDB file. The structure has been measured at a very low temperature ∼20 K with very small thermal B factors; in our approach a high experimental accuracy is desirable since the model we develop can describe the folded protein structure with sub-Ångström precision.

II. METHODS

A. Continuous curves

For completeness, we start with a review of basic relations in curve geometry [54,55]. We consider a space curve $\mathbf{x}(s)$: $[0,L] \to \mathbb{R}^3$, where L is the total length of the curve and $s \in [0,L]$ measures its proper length so

$$||\dot{\mathbf{x}}|| = 1. \tag{2}$$

The unit tangent vector is

$$\mathbf{t} = \dot{\mathbf{x}} \equiv \frac{d\mathbf{x}(s)}{ds}. \tag{3}$$

The unit binormal vector is

$$\mathbf{b} = \frac{\dot{\mathbf{x}} \times \ddot{\mathbf{x}}}{||\dot{\mathbf{x}} \times \ddot{\mathbf{x}}||} \tag{4}$$

and the unit normal vector is

$$\mathbf{n} = \mathbf{b} \times \mathbf{t}. \tag{5}$$

The orthonormal triplet $(\mathbf{n},\mathbf{b},\mathbf{t})$ defines a framing of the curve that is subject to the Frenet equation [54,55],

$$\frac{d}{ds}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix} = \begin{pmatrix}0 & \tau & -\kappa\\ -\tau & 0 & 0\\ \kappa & 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix}. \tag{6}$$

Here

$$\kappa(s) = \frac{||\dot{\mathbf{x}} \times \ddot{\mathbf{x}}||}{||\dot{\mathbf{x}}||^3} \tag{7}$$

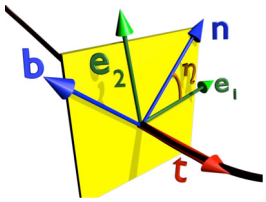

FIG. 1. The (blue) Frenet frame $(\mathbf{n},\mathbf{b})$ and a generic (green) orthogonal frame $(\mathbf{e}_1,\mathbf{e}_2)$ on the normal plane of $\mathbf{t}$, the tangent vector of the curve.

is the curvature and

$$\tau(s) = \frac{(\dot{\mathbf{x}} \times \ddot{\mathbf{x}}) \cdot \dddot{\mathbf{x}}}{||\dot{\mathbf{x}} \times \ddot{\mathbf{x}}||^2} \tag{8}$$

is the torsion. The fundamental theorem of space curves states that the shape of every sufficiently regular curve in three-dimensional space is completely determined by its curvature and torsion; the extrinsic and intrinsic geometries of a curve coincide. Thus, whenever $\kappa(s)$ and $\tau(s)$ are known, we can compute the Frenet framing from (6) and we can then proceed to compute the shape of the curve by integrating (3). Accordingly, the curvature and the torsion are the (only) natural variables for constructing an energy function of a structureless continuous curve. In particular, the shape of a static curve should be computable, as a minimum of the pertinent energy function.

Whenever (2) is valid, the tangent vector is given by (3). But when there is an inflection point, i.e., a parameter value $s = s_0$ so the curvature vanishes,

$$\kappa(s_0) = ||\ddot{\mathbf{x}}(s_0)|| = 0, \tag{9}$$

the vectors $\mathbf{n}$ and $\mathbf{b}$ are not determined and the Frenet framing cannot be introduced. However, there are other ways to frame a curve in a manner that extends continuously through an inflection point and, more generally, through straight segments of the curve. An example is the Bishop (parallel transport) framing [56].

The Landau-Ginzburg-Wilson paradigm states that the energy function must be built so it respects the symmetries of the physical system. For this we consider a generic orthonormal framing $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{t})$. As shown in Fig. 1 whenever the curvature is nonvanishing it can be related to the Frenet framing by a local SO(2) rotation around the tangent vector $\mathbf{t}(s)$,

$$\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\end{pmatrix} \to \begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{pmatrix} = \begin{bmatrix}\cos\eta(s) & -\sin\eta(s)\\ \sin\eta(s) & \cos\eta(s)\end{bmatrix}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\end{pmatrix}. \tag{10}$$

The ensuing generalization of the Frenet equation is

$$\frac{d}{ds}\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix} = \begin{bmatrix}0 & (\tau-\dot{\eta}) & -\kappa\cos\eta\\ -(\tau-\dot{\eta}) & 0 & -\kappa\sin\eta\\ \kappa\cos\eta & \kappa\sin\eta & 0\end{bmatrix}\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{t}\end{pmatrix}. \tag{11}$$

We deduce that the torsion transforms under frame rotations as follows:

$$\tau \to \tau_r \equiv \tau - \dot{\eta}. \tag{12}$$

For the curvature, the effect of the frame rotation is summarized in terms of the complex valued quantity,

$$\kappa \rightarrow \kappa_{\pm} = \kappa e^{\pm i\eta} = \kappa_g \pm i\kappa_n. \tag{13}$$

The (generalized) Frenet equation can be represented as follows:

$$\left(\frac{d}{ds} \pm i\tau_r\right)(\mathbf{e}_1 \pm i\mathbf{e}_2) \equiv \left(\frac{d}{ds} \pm i\tau_r\right)\mathbf{e}_{\pm} = -\kappa_{\pm}\mathbf{t}, \tag{14}$$

$$\frac{d}{ds}\mathbf{t} = 2(\kappa_+\mathbf{e}_+ + \kappa_-\mathbf{e}_-). \tag{15}$$

The real part κ_g of the complex curvature $\kappa_{\pm}$ is called the geodesic curvature, and the imaginary part κ_n is called the normal curvature. These two quantities refer to the extrinsic geometry of a surface that osculates the curve. The osculating surface is not uniquely determined, and different choices of η correspond to different osculating surfaces. The choice $\eta = 0$ specifies the Frenet frame (Frenet gauge), and the choice

$$\eta(s) = \int_0^s \tau(s')ds'$$

specifies Bishop's frames [54–56] that can be defined continuously and unambiguously through an inflection point.

The invariance of the curve under frame rotations, *per se*, constitutes a symmetry that can be exploited to construct LGW energy functions. We follow standard field theory [57] and identify in $(\kappa_{\pm},\tau_r)$ a SO(2)$\sim U(1)$ gauge multiplet (Abelian Higgs multiplet) [58]. The change (12) in τ_r is akin a SO(2)$\sim U(1)$ gauge transformation of a one-dimensional gauge vector, while $\kappa_{\pm}$ transforms like a complex scalar field.

Finally, we observe that the complex valued Hashimoto variable [59]

$$\xi(s) = \kappa_+(s)\exp\left(i\int_0^s \tau_r\, ds'\right) \equiv \kappa(s)\exp\left(i\int_0^s \tau\, ds'\right) \tag{16}$$

is gauge invariant, i.e., independent of the choice of framing.

B. Landau-Ginzburg-Wilson free energy

The Landau-Ginzburg-Wilson approach instructs us to exploit a symmetry to construct an invariant energy function of a curve in the limit of slow spatial variations.

We start with a generic Helmholtz free energy (1) in the limit of slow spatial variations; we follow Ref. [60]. We assume a theory with a single scalar order parameter field $\varphi(x)$, i.e., with no specific symmetry. The free energy (1) may be expanded in powers of the order parameter,

$$F = \sum_n \frac{1}{n!}\int d^Dx_1 \cdots d^Dx_n F^{(n)}(x_1 \cdots x_n)\varphi(x_1)\cdots\varphi(x_n). \tag{17}$$

The coefficients $F^{(n)}$ are the n-point Green's functions, and they are commonly evaluated perturbatively in terms of Feynman diagrams [57,60].

There is an alternative way to expand the free energy [57,60] in powers of derivatives (momentum) about the point where all external derivatives (momenta) vanish. More specifically, we inspect the physical system over a distance scale L such that the spatial variations of $\varphi(x)$ over this scale are small. The derivatives of φ can then be employed as small expansion parameters, and [57,60]

$$F = \int d^Dx\left[V(\varphi) + \frac{1}{2}Z(\varphi)(\partial_\mu\varphi)^2\cdots\right]. \tag{18}$$

This is the expansion in terms of slowly varying variables. Note that the coefficients $V(\varphi), Z(\varphi), \ldots$ are ordinary functions, not functionals. To the leading order $V(\varphi)$ coincides with the classical potential in the Hamiltonian, generically [57,60]

$$Z(\varphi) = 1 + a\varphi^2 + \ldots.$$

We now specify to the case of a regular curve: The geometry of a structureless curve cannot depend on the way it is framed, thus we propose to exploit invariance under local frame rotations as the guiding symmetry; the functional form of the ensuing Helmholtz free energy (18) should remain intact under local frame rotations (10).

In the case of a regular structureless curve, the shape is completely determined by the generalized torsion (12) and curvature (13). Accordingly, these two local quantities constitute a complete set of order parameter variables to specify the Helmholtz free energy (18) of the curve. Since the free energy should be independent of the way the curve is framed, it can only depend on gauge-invariant, i.e., frame-rotation-invariant, combinations of (12) and (13). Thus, in the leading large distance (infrared) order, the pertinent expansion (18) engages the Hamiltonian of the Abelian Higgs model [58,61],

$$F = \int ds[\lambda\left(|\kappa_+|^2 - m^2\right)^2 + |(\partial_s + i\tau_r)\kappa_+|^2 + \sigma\tau_r + \cdots]. \tag{19}$$

To the leading order, this is the most general nonlocal functional of $(\tau_r,\kappa_{\pm})$ which is manifestly invariant under the local frame rotation (10).

The last term in (19) is the *helicity*; it is a one-dimensional version of the Chern-Simons term that breaks the chirality. The helicity is not $U(1)$ invariant, but its $U(1)$ transformation is a surface term. Any surface term should become irrelevant in the thermodynamic limit.

The variables $(\tau_r,\kappa_{\pm})$ can be eliminated in favor of the gauge-invariant, geometric quantities (7) and (8). This corresponds to the unitary gauge [57]: We use (13) and

$$\tau = -\frac{i}{2\kappa^2}[\kappa_-(\partial_s + i\tau_r)\kappa_+ - \text{c.c.}].$$

We substitute in (19), and we obtain

$$F = \int ds[(\partial_s\kappa)^2 + \kappa^2\tau^2 + \lambda(\kappa^2 - m^2)^2 + \sigma\tau]. \tag{20}$$

Specifically, the validity of the approximation (19) and (20) assumes that if κ_0 sets a scale of curvature and when L is a (large) distance scale of interest, then

$$|\partial_s\kappa| \ll \frac{\kappa_0}{L}.$$

Thus, over distance scales which are comparable to L or larger, and subject to the frame rotation invariance, the Helmholtz free energy of a curve is approximated by the LGW free energy, with leading-order expansion (20) in derivatives of $\kappa(s)$.

C. Integrable hierarchy

Besides the frame rotation symmetry (10)–(13) there are other symmetry principles that may be utilized as a guiding principle in the construction of a Landau-Ginzburg-Wilson energy function. As an example, we consider the (infinite) symmetry which is associated with the concept of an integrable model [62,63]:

We start with the observation that the (manifestly frame-independent) Hasimoto variable (16) converts the Hamiltonian of the Abelian Higgs Model into the Hamiltonian of the nonlinear Schrödinger (NLS) equation. Specifically, in terms of the variables (20)

$$\lambda\kappa^4 + \kappa^2\tau^2 + (\partial_s\kappa)^2 = \lambda(\bar{\xi}\xi)^2 + \partial_s\bar{\xi}\partial_s\xi, \tag{21}$$

which is the NLS Hamiltonian [62–64]. The NLS Hamiltonian is the paradigm integrable model. It admits an infinite number of conserved quantities, each associated with a symmetry of (21). The last term in (20), the helicity, is an example of a conserved quantity in the NLS model. The number density

$$\tfrac{1}{2}\bar{\xi}\xi = \tfrac{1}{2}\kappa^2 \tag{22}$$

is another example, and so is the momentum density

$$-\frac{i}{2}\bar{\xi}\partial_s\xi = \frac{1}{2}\kappa^2\tau. \tag{23}$$

Note that, like helicity, momentum breaks chirality.

As such, (22) is the Hamiltonian of the Worm Like Chain (Kratky-Porod) model [65], widely used in modeling aspects of polymers. In terms of the tangent vector,

$$\tfrac{1}{2}\kappa^2 = \tfrac{1}{2}|\partial_s\mathbf{t}|^2.$$

This is the Hamiltonian of the Heisenberg σ model [62].

The LGW paradigm, in combination with the symmetry structure of the NLS model, proposes that the Helmholtz free energy (1) can be systematically expanded in terms of the conserved charges of the NLS hierarchy. In this way we arrive at the following (slight) generalization of (20)

$$F = \int ds\Big[(\partial_s\kappa)^2 + \lambda(\kappa^2 - m^2)^2 + \frac{d}{2}\kappa^2\tau^2 - b\kappa^2\tau - a\tau + \frac{c}{2}\tau^2\Big]. \tag{24}$$

Here the last term is called the Proca mass in gauge theory, and we include it for completeness [66,67].

The remaining conserved quantities of the NLS model involve higher orders of derivatives of $\kappa(s)$. As such, they are higher-order corrections in the expansion (18). We do not include them, in our infrared limit.

D. Topological solitons

Solitons are the paradigm structural self-organizers in nature and the NLS equation is the paradigm equation that supports solitons [62–64]; depending on the sign of λ, the soliton is either dark ($\lambda > 0$) or bright ($\lambda < 0$). Moreover, the torsion-independent contribution to (24) and (20),

$$\int_{-\infty}^{\infty} ds\,\{(\partial_s\kappa)^2 + \lambda\,(\kappa^2 - m^2)^2\}, \tag{25}$$

supports the double well *topological* soliton [68]: When m^2 is positive and when κ can take both positive and negative values, the equation of motion

$$\partial_{ss}\kappa = 2\lambda\kappa(\kappa^2 - m^2)$$

is solved by

$$\kappa(s) = m\,\tanh[m\sqrt{\lambda}(s - s_0)]. \tag{26}$$

Note that this soliton engages an inflection point (9); following [69,70], we use the convention that when a curve passes a simple inflection point, the curvature changes its sign.

The energy function (24) is quadratic in the torsion. Thus we can eliminate τ using its equation of motion,

$$\tau[\kappa] = \frac{a + b\kappa^2}{c + d\kappa^2} \equiv \frac{a}{c}\,\frac{1 + (b/a)\kappa^2}{1 + (d/c)\kappa^2}, \tag{27}$$

and we obtain the following equation of motion for curvature:

$$\kappa_{ss} = V_\kappa[\kappa], \tag{28}$$

where

$$V[\kappa] = -\left(\frac{bc - ad}{d}\right)\frac{1}{c + d\kappa^2} - \left(\frac{b^2 + 8\lambda m^2}{2b}\right)\kappa^2 + \lambda\,\kappa^4. \tag{29}$$

This shares the same large-κ asymptotics, with the potential in (25). With properly chosen parameters, we expect that (28) and (29) continue to support topological solitons. But we do not know their explicit profile, in terms of elementary functions.

Once we have the soliton of (28), we evaluate $\tau(s)$ from (27). We substitute the ensuing (κ,τ) profiles in the Frenet equation (6) and solve for $\mathbf{t}(s)$. We then integrate (3) to obtain the curve $\mathbf{x}(s)$ that corresponds to the soliton. A generic soliton curve looks like a helix-loop-helix motif (more generally, a regular secondary structure-a loop-a regular secondary structure), familiar from crystallographic protein structures. Note that depending on the parameter values, the torsion can fluctuate substantially along a soliton profile even when curvature is slowly varying.

E. Discrete Frenet equation

Proteins cannot be modeled by regular space curves. Proteins are like piecewise linear polygonal chains. In order to construct the LGW energy function for a protein, we need to understand the structure and symmetry of such a chain [71].

In the present article we shall only address the backbone; the folded structure is known, thus there are no issues with potential steric clashes in the *folded* structure.

Let $\mathbf{r}_i$ with $i = 1,\ldots,N$ be the vertices of a piecewise linear discrete chain; in the case of a protein, the vertices correspond to the Cα atoms. At each vertex we introduce the unit tangent vector

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}, \tag{30}$$

the unit binormal vector

$$\mathbf{b}_i = \frac{\mathbf{t}_{i_1} - \mathbf{t}_i}{|\mathbf{t}_{i_1} - \mathbf{t}_i|}, \tag{31}$$

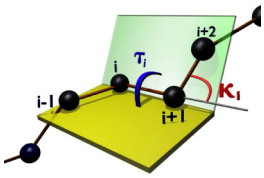

FIG. 2. Definition of bond (κ_i) and torsion (τ_i) angles along a piecewise linear discrete chain.

and the unit normal vector

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i . \tag{32}$$

The orthonormal triplet $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ constitutes a discrete version of the Frenet frames.

In lieu of the curvature and torsion, we have the bond angles and torsion angles, defined as in Fig. 2.

Once we know the Frenet frames at each vertex, we can compute the angles. The bond angles are

$$\kappa_i \equiv \kappa_{i+1,i} = \arccos(\mathbf{t}_{i+1} \cdot \mathbf{t}_i) \tag{33}$$

and the torsion angles are

$$\tau_i \equiv \tau_{i+1,i} = \mathrm{sgn}\{\mathbf{b}_{i+1} \times \mathbf{b}_i \cdot \mathbf{t}_i\} \arccos(\mathbf{b}_{i+1} \cdot \mathbf{b}_i). \tag{34}$$

Conversely, when the values of the bond and torsion angles are all known, we can use the discrete Frenet equation

$$\begin{pmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos\kappa\cos\tau & \cos\kappa\sin\tau & -\sin\kappa \\ -\sin\tau & \cos\tau & 0 \\ \sin\kappa\cos\tau & \sin\kappa\sin\tau & \cos\kappa \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \tag{35}$$

to compute the frame at vertex $i + i$ from the frame at vertex i. Once all the frames have been constructed, the entire chain is given by

$$\mathbf{r}_k = \sum_{i=0}^{k_1} |\mathbf{r}_{i+1} - \mathbf{r}_i| \mathbf{t}_i . \tag{36}$$

Without any loss of generality we may choose $\mathbf{r}_0 = 0$, choose $\mathbf{t}_0$ to point into the direction of the positive z axis, and let $\mathbf{t}_1$ lie on the y-z plane.

As in the case of a continuum curve, a discrete chain remains intact under frame rotations of the $(\mathbf{n}_i, \mathbf{b}_i)$ zweibein around $\mathbf{t}_i$. This local SO(2) rotation acts on the frames as follows:

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i \to e^{\Delta_i T^3} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i = \begin{pmatrix} \cos\Delta_i & \sin\Delta_i & 0 \\ -\sin\Delta_i & \cos\Delta_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i , \tag{37}$$

where T^3 is one of the SO(3) Lie algebra generators,

$$(T^a)_{bc} = \epsilon^{abc} .$$

In terms of the bond and torsion angles the rotation amounts to

$$\kappa_i \, T^2 \to e^{\Delta_i T^3} (\kappa_i T^2) \, e^{-\Delta_i T^3} , \tag{38}$$

$$\tau_i \to \tau_i + \Delta_{i_1} - \Delta_i , \tag{39}$$

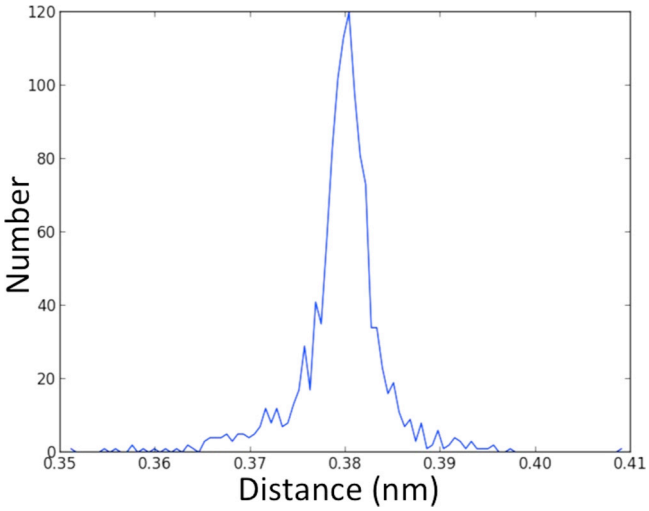

FIG. 3. Distribution of bond length in crystallographic PDB structures; the data set in Ref. [72] has been used.

which is a direct generalization of (12) and (13); following standard field theory [57] the transformation of bond angles is like an adjoint SO(2)∈SO(3) gauge rotation of a Higgs triplet around the Cartan generator T^3, when the Higgs triplet is in the (unitary gauge) direction of T^2. The transformation of torsion angle coincides with that of the SO(2) lattice gauge field.

A priori, the fundamental range of the bond angle is $\kappa_i \in [0,\pi]$ while for the torsion angle the range is $\tau_i \in [-\pi,\pi)$. Thus we may identify (κ_i, τ_i) as the canonical latitude and longitude angles of a two-sphere $\mathbb{S}^2$. However, to account for the presence of putative inflection, it is useful to extend the range of κ_i into negative values $\kappa_i \in [-\pi,\pi]$ $mod(2\pi)$. We compensate for this twofold covering of $\mathbb{S}^2$ by a $\mathbb{Z}_2$ symmetry:

$$\begin{aligned} &\kappa_k \to -\kappa_k \quad \text{for all} \quad k \geqslant i \\ &\tau_i \to \tau_i - \pi . \end{aligned} \tag{40}$$

This is a special case of (38) and (39), with

$$\begin{aligned} &\Delta_k = \pi \quad \text{for} \quad k \geqslant i+1 \\ &\Delta_k = 0 \quad \text{for} \quad k < i+1 . \end{aligned}$$

F. The Cα trace reconstruction

The discrete Frenet equations (35) and (36) discloses that a chain can be constructed from the knowledge of bond and torsion angles and the distances between the vertices. In the case of crystallographic protein structures, the vertices coincide with the positions of the Cα atoms. As shown in Fig. 3, in PDB the virtual Cα-Cα bond lengths are very close to their average value,

$$|\mathbf{r}_{i+1} - \mathbf{r}_i| \sim 3.8 \text{ Å}. \tag{41}$$

Moreover, according to Ref. [72], the Cα backbones of PDB structures can be reliably reconstructed using a combination of the actual bond and torsion angles (33) and (34) and the *average* value (41). Thus the bond and torsion angles constitute a complete set of structural order parameters in the case of crystallographic proteins. The LGW paradigm proposes that

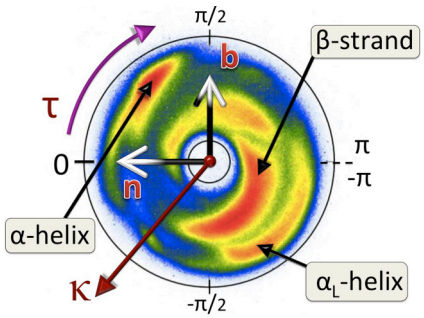

FIG. 4. Distribution of bond and torsion angles in crystallographic protein structures on a stereographically projected two-sphere with κ the latitude and τ the longitude. Red indicates a large number of entries, blue a small number of entries, and white corresponds to no entries. All PDB structures that have been measured with resolution 2.0 Å or better have been used. The major secondary structure regimes are identified. The inner boundary of the annulus has a radius $\kappa \approx 1$ (rad) and the outer boundary has a radius $\kappa \approx 1.6$ (rad).

the leading-order approximation to the Helmholtz free energy is a function of the bond and torsion angles only.

Note: The Ramachandran angles, together with the average value (41), do *not* constitute a complete set of structural order parameters [72].

G. LGW Hamiltonian for proteins

Proteins are commonly modeled using an all-atom force field or a coarse-grained approximation thereof [11–13]. The discretized Newton's equation is solved iteratively, in what *de facto* amounts to a perturbative expansion around a (randomly) chosen initial configuration: The expansion parameter relates to the ratio of the iterative time step length to the time scale of a characteristic atomic oscillation. In an all-atom approach the latter pertains to the frequency of a heavy atom covalent bond oscillation, which makes simulations into an extreme computational challenge. Equation (17) exemplifies a perturbative approach.

Here we follow the Landau-Ginzburg-Wilson paradigm to develop a complementary approach to model proteins and their dynamics. Conceptually, our approach is like the expansion (18) and we need to identify the slowly varying variable: In Fig. 4 we show the distribution of Cα backbone bond and torsion angles in crystallographic PDB protein structures on the stereographically projected two-sphere (κ,τ). The torsion angles

are known to be flexible in proteins; as shown in Fig. 4, their values are distributed over the entire range $\tau \in (-\pi,\pi]$. However, the observed range of variation $\Delta\kappa_{\max}$ in the values of the bond angles is quite constrained. Instead of extending over the entire plane, the angles are largely limited to the annulus between $\kappa \approx 1$ and $\kappa \approx \pi/2$ (radians), shown in Fig. 4. On the original two-sphere, the geometrically allowed range of variations of the bond angle is $\kappa_{\text{tot}} \in [0,\pi]$. Thus we may putatively adopt the ratio

$$\left|\frac{\kappa_{i+1}-\kappa_i}{\kappa_{\text{tot}}}\right| \leqslant \frac{\Delta\kappa_{\max}}{\kappa_{\text{tot}}} \approx \frac{0.6}{\pi} \sim 0.2 \qquad (42)$$

as a slowly varying expansion parameter in the case of crystallographic protein structures. The Landau-Ginzburg-Wilson paradigm then proposes that *if* we adopt the NLS hierarchy as the symmetry principle to guide the construction of the LGW energy function, in the case of a protein backbone, we should adopt a discretized version of (24) as the leading-order LGW approximation of the Helmholtz free energy:

$$H = -\sum_{i=1}^{N_1} 2\,\kappa_{i+1}\kappa_i + \sum_{i=1}^{N}\Big\{2\kappa_i^2 + \lambda\big(\kappa_i^2-m^2\big)^2 + \frac{d}{2}\,\kappa_i^2\tau_i^2 - b\,\kappa_i^2\tau_i - a\,\tau_i + \frac{c}{2}\tau_i^2\Big\} + \ldots. \qquad (43)$$

The approximation (43) should be a valid one, as long as the expansion parameter (42) remains small, i.e., there are no abrupt but only slowly changing bends along the backbone. In particular, long-range interactions are accounted for, as long as they do not cause any sharp localized buckling of the backbone.

In (43) λ, a, b, c, d, and m depend on the atomic level physical properties and the chemical microstructure of the protein and its environment. In principle, these parameters can be computed from this knowledge. In practice, we train the energy function to model a given protein.

H. Topological soliton and protein geometry

The free energy (43) is a *naive* discretization of the NLS hierarchy free energy (24). It is a deformation of the energy function of the integrable discrete nonlinear Schrödinger equation (DNLS) [62–64]. The conventional DNLS equation is known to support solitons. Thus we expect that (43) supports soliton solutions as well.

We follow (27) to eliminate the torsion angle,

$$\tau_i[\kappa] = \frac{a+b\kappa_i^2}{c+d\kappa_i^2} = a\frac{1+(b/a)\kappa_i^2}{c+d\kappa_i^2}. \qquad (44)$$

For bond angles we then have

$$\kappa_{i+1} = 2\kappa_i - \kappa_{i-1} + \frac{dV[\kappa]}{d\kappa_i^2}\kappa_i \quad (i=1,\ldots,N). \qquad (45)$$

We set $\kappa_0 = \kappa_{N+1} = 0$, and $V[\kappa]$ is given by (29). To solve this numerically, we use the iterative equation [19]

$$\kappa_i^{(n+1)} = \kappa_i^{(n)} - \epsilon\big\{\kappa_i^{(n)}V'\big[\kappa_i^{(n)}\big] - \big(\kappa_{i+1}^{(n)} - 2\kappa_i^{(n)} + \kappa_{i-1}^{(n)}\big)\big\}, \qquad (46)$$

where $\{\kappa_i^{(n)}\}_{i\in N}$ is the nth iteration of an initial configuration $\{\kappa_i^{(0)}\}_{i\in N}$ and ϵ is some sufficiently small but otherwise arbitrary numerical constant. We choose $\epsilon = 0.01$, in our simulations. The fixed point of (46) is independent of the value of ϵ and clearly a solution of (45).

Once the fixed point is found, the corresponding torsion angles are obtained from (44). The frames are then constructed from (35), and the entire chain is constructed using (36).

We do not know of an analytical expression of the soliton solution to Eq. (45). But an *excellent* approximative solution can be obtained by discretizing the topological soliton (26) [18,69–71,73]:

$$\kappa_i \approx \frac{m_1 e^{c_1(i-s)} - m_2 e^{-c_2(i-s)}}{e^{c_1(i-s)} + e^{-c_2(i-s)}}. \qquad (47)$$

Here $(c_1,c_2,,m_1,m_2,s)$ are parameters. The m_1 and m_2 specify the asymptotic κ_i values of the soliton. Thus, these parameters are entirely determined by the character of the regular, constant bond and torsion angle structures that are adjacent to the soliton. In particular, these parameters are not specific to the soliton *per se* but to the adjoining regular structures. The parameter s defines the location of the soliton along the string. This leaves us with only two loop specific parameter, the c_1 and c_2. These parameters quantify the length of the bond angle profile that describes the soliton.

For the torsion angle, (44) involves one parameter (a) that we have factored out as the overall relative scale between the bond angle and torsion angle contributions to the energy; this parameter determines the relative flexibility of the torsion angles, with respect to the bond angles. Then, there are three additional parameters $(b/a,c/a,d/a)$ in the remainder $\hat{\tau}[\kappa]$. Two of these are again determined by the character of the regular structures that are adjacent to the soliton. As such, these parameters are not specific to the soliton. The remaining single parameter specifies the size of the regime where the torsion angle fluctuates.

On the regions adjacent to a soliton, we have constant values of (κ_i,τ_i). In the case of a protein, these are the regions that correspond to the standard regular secondary structures. For example, the standard right-handed α helix is obtained by setting

$$\alpha \text{ helix}: \quad \begin{cases} \kappa \approx \frac{\pi}{2} \\ \tau \approx 1 \end{cases}. \tag{48}$$

and for the standard β strand

$$\beta \text{ strand}: \quad \begin{cases} \kappa \approx 1 \\ \tau \approx \pi \end{cases}. \tag{49}$$

All the other standard regular secondary structures of proteins such as 3/10 helices, left-handed helices, etc., are similarly modeled by definite constant values of κ_i and τ_i. Protein loops correspond to solitons, the regions where the values of (κ_i,τ_i) are variable.

The presence of solitons significantly reduces the number of parameters in (43), increasing the predictive power. In particular, the number of parameters is far smaller than the number of amino acids, along the protein backbone.

I. Cooperativity and first-order phase transition

In Refs. [27–29] it has been argued on general grounds that protein folding should have a "cooperative" character and that the folding process should relate to a first-order phase transition. The present model is fully in line with these proposals:

Solitons are the paradigm structural self-organizers in physical systems, and solitons are collective oscillations that emerge and dissolve when a physical system undergoes a large scale structural self-organization.

Moreover, in the case of a protein the bond angles are rigid while the torsion angles are flexible. From the point of view of a Born-Oppenheimer approximation, the torsion angles can then be considered as "fast" variables in the background of bond angles that can be considered as "slow" variables. In the limit where variations of κ_i along the backbone become very small in comparison to the variations in τ_i, we may proceed and solve for the mean value $\kappa_i \sim \kappa$ in terms of $\tau_i \sim \tau$,

$$\kappa^2 \approx m^2 + \frac{b}{2\lambda}\tau + \frac{d}{4\lambda}\tau^2.$$

When we substitute this in (43) we obtain an effective free energy for the mean value of the torsion angle, this energy has the form

$$F(\tau) = A + B\tau + C\tau^2 + D\tau^3 + E\tau^4.$$

This coincides with a Landau free energy that models a first-order phase transition when $D < 0$ and $E > 0$: There is both a metastable and a stable minimum energy state, and an abrupt first-order transition takes place when the parameter values change.

Accordingly, the model is in line with and incorporates the proposals made in Refs. [27–29].

J. Proteins out of thermal equilibrium

When a protein folds towards its native state, it is out of thermal equilibrium. Several studies propose, that in the case of a small protein which is not too far away from thermal equilibrium, the folding takes place in a manner which is consistent with Arrhenius' law [74]. This law states that the reaction rate depends exponentially on the ratio of activation energy E_A and physical temperature factor,

$$r \propto \exp\left\{-\frac{E_A}{k_B\theta}\right\},$$

with k_B the Boltzmann constant and θ the temperature measured in Kelvin.

On the other hand, in the case of a simple spin chain, Glauber dynamics [30–32] is known to describe the approach to thermal equilibrium, in a manner which resembles Arrhenius's law. Glauber dynamics evaluates the transition probability from a conformational state a to another conformational state b as follows:

$$\mathcal{P}(a \to b) = \frac{1}{1 + e^{\Delta F_{ba}/T}}.$$

Here $\Delta H_{ba} = H_b - H_a$ is the activation energy and we compute it from (43). The parameter T is the Monte Carlo temperature factor. Note that in general the Monte Carlo temperature factor T does not coincide with the physical temperature factor $k_B\theta$. Instead, we expect [20] that T relates to $k_B\theta$ approximatively as follows:

$$T \sim k_B\theta\, e^{\alpha k_B\theta - \beta}, \tag{50}$$

where α and β are protein specific factors.

K. Simulation details

In all our simulations, at each MC step we perturb either one of the bond angles or torsion angles according to the following prescription:

$$\begin{aligned} \kappa_i &\to \kappa_i + 0.015r \\ \tau_i &\to \tau_i + 1.5r, \end{aligned}$$

where r is a random number with Gaussian distribution with expectation value 0 and variation 1. The different scales on

κ_i and τ_i reflect the different stiffness between the bond and torsion angles in real proteins. We have tested various other values of r to confirm that our results do not essentially depend on the choice of r.

L. Steric constraints

Finally, proteins are subject to long-range interactions along the chain. At the level of the Landau approach this can be accounted for in terms of steric constraints; see Ref. [75], where a detailed analysis on the effect of steric constraints has been carried out in the present model.

In PDB, a statistical analysis shows that two Cα atoms which are *not* nearest neighbors along the backbone, are constrained by

$$|\mathbf{r}_i - \mathbf{r}_k| > 3.8\ \text{Å} \quad \text{for} \quad |i-k| \geqslant 2. \tag{51}$$

Here this constraint is implemented as a criterion in the Monte Carlo algorithm: whether to accept or reject a given Monte Carlo step during simulation. For a given Cα backbone, different side chain libraries can then used to to reconstruct the all-atom structure, as need be.

III. RESULTS

We have performed extensive numerical simulations to analyze the way myoglobin *unfolds* when ambient temperature increases. The motivation to consider in detail the *unfolding* process in the case of a myoglobin relates to the experimental issues due to the binding of heme: It is very difficult to control the process of heme binding in a folding experiment, and thus the myoglobin experiments [46–51] all address the unfolding process.

Results from both unfolding and folding experiments are available, in the case of the heme-free apomyoglobin [33–45]. These experiments reveal that unfolding and folding processes are very similar.

We have confirmed that in our heating and cooling simulations, the unfolding and folding pathways are essentially identical.

Comparisons of experiments with heme containing myoglobin and heme-free apomyoglobin show that the unfolding proceeds very similarly, in the two cases. The only real exception is that, in the case of apomyoglobin, the F-helix is disordered at low temperatures [37]. Accordingly, there are no crystallographic data available, in the case of apomyoglobin, that we could use to construct a high-precision LGW free energy (43). However, the structural effects of heme during the unfolding process are apparently minor. Thus our results are likewise applicable, both in the case of heme-free apomyoglobin and heme containing myoglobin.

A. Multisoliton

We start with the construction of the multisoliton solution of (45) and (44) that models the Cα backbone of 1ABS [53]. We used a combination of the `GaugeIT` and `Propro` packages [76].

The analysis starts with the inspection of the bond and torsion angle spectrum with the help of the $\mathbb{Z}_2$ symmetry (40) to identify the individual solitons. In Fig. 5 we show the (κ_i, τ_i)

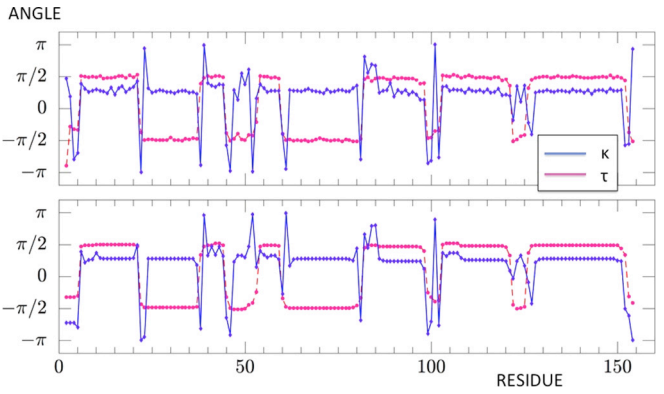

FIG. 5. Top: The bond (κ) and torsion (τ) angle spectrum of the PDB structure 1ABS. Bottom: The bond (κ) and torsion (τ) angle spectrum of the multisoliton. Note that the angles are defined modulo 2π.

spectrum both for 1ABS and for the multisoliton we have constructed; the Cα RMS distance between the two is around 0.8 Å. In Table I we show the parameter values that we have found; there are 92 parameters that describe the 154 different amino acids.

B. Stability and reversibility

We have tested the stability of the multisoliton by subjecting it to repeated heating and cooling simulations using the Glauber algorithm: We start from the low Monte Carlo temperature factor value $T = 10^{-17}$ where we observe no thermal fluctuations. We increase the temperature factor linearly on a logarithmic scale, so we reach the value $T = 10^{-4}$ after 5 million MC steps. We then fully thermalize the configuration at $T = 10^{-4}$ during another 5 million MC steps. Finally, we cool it down, back to the original low temperature value, during 5 million steps.

For production, we have performed 100 full heating-cooling cycles. The Fig. 6 shows the evolution of RMS distance to the low-temperature multisoliton, during the heating and cooling cycle; we find that the heating and cooling proceed very symmetrically, and the configuration returns to the original low temperature structure at the end of the cycle.

C. Heating myoglobin

In our production runs for heating simulations, described in the sequel, we have increased the temperature factor from $T = 10^{-17}$ to $T = 10^{-4}$ during 6 million Monte Carlo steps; the results do not depend on the number of steps, as long as this number is not very small. We have performed 100 independent full-length heating simulations. We have in particular confirmed that the heating process is fully reversible: On cooling the system from the high temperature value back to the original low-temperature value the structure folds back to the native conformation. The unfolding and folding pathways are essentially identical.

Our simulations are extremely time efficient. For example, with a MacPro workstation a complete heating and cooling cycle takes around 10 s of *in silico* time, in a single processor; this is the time it takes us to perform a cycle. By comparison,

TABLE I. The parameters in the energy function for 1ABS. The ensuing profile is shown in Fig. 5 bottom. There are a total of 12 individual soliton profiles along the entire myoglobin backbone. In each soliton, we divide the parameters c and m into c_1,c_2 and m_1,m_2 to reflect the asymmetry of the soliton around its center.

Soliton	s	λ_1	λ_2	m_1	m_2	d	a	c	b
1	4	12.078	3.9172	1.0115	1.5417	1.2567 e-7	−3.1072 e-8	8.3858 e-8	2.1701 e-6
2	20	3.4359	2.029	1.58	1.5138	2.0144 e-8	−7.2358 e-8	2.501 e-8	−1.081 e-6
3	36	7.3192	0.8146	1.5064	1.543	3.7 e-9	−1.0164 e-7	5.4087 e-10	4.8281 e-8
4	43	2.1377	0.657	1.6558	1.6022	5.7539 e-9	−9.0514 e-8	5.1017 e-11	1.2023 e-6
5	52	0.8854	5.9719	1.3645	1.5369	7.6765 e-9	−2.3414 e-7	2.362 4 e-8	−3.301 e-7
6	58	8.7118	0.8337	1.55	1.537	4.8756 e-9	−9.6428 e-8	1.0224 e-10	4.778 e-7
7	80	0.97324	2.1401	1.4617	1.5462	1.9022 e-14	−7.401 e-9	6.9463 e-10	3.8355 e-9
8	97	1.3258	2.9105	1.4771	1.0199	5.4505 e-14	−1.3745 e-13	3.492 e-14	5.6029 e-13
9	101	10.4862	4.2438	1.2225	1.6532	1.22576 e-8	−1.2136 e-7	9.9151 e-11	1.3718 e-6
10	120	0.80042	1.2897	1.5154	1.6028	7.8271 e-8	−2.0349 e-7	1.46 e-11	1.13574 e-6
11	124	3.1526	0.91475	1.5583	1.5515	7.7364 e-9	−1.0781 e-7	7.4957 e-11	1.0277 e-6
12	151	1.0122	1.0637	1.4001	1.3282	1.1378 e-8	−1.1176 e-7	4.3856 e-10	8.6209 e-7

the experimentally observed folding time of apomyoglobin is around 2.5 s [34].

D. α-helical content

Circular dichroism (CD) spectroscopy can provide information on the helical content of a protein as a function of temperature; it can give an indication how the unfolding proceeds. By comparing how the theoretically evaluated and experimentally observed helical content depends on temperature, we can also determine the relation (50) between the Glauber temperature factor and the physical temperature. Accordingly, we have estimated the temperature dependence of the α-helical content during our heating simulations. By a statistical analysis of PDB structures we have deduced the following criterion: We define that a Cα atom which is centered at $\mathbf{r}_i$ is in an α-helical position when $|\mathbf{r}_{i+4} - \mathbf{r}_i| \approx 6.2 \pm 0.5$ Å and $|\tau_i - \tau_0| < 0.6$ (rad) where τ_0 is the PDB average value of the α-helical torsion angle. From Fig. 5 we observe that folded myoglobin has a substantial α-helical content. According to our criterion, around ∼72% of the Cα atoms in 1ABS are in α-helical position; there is also a small fraction in the closely related 3/10 position.

We have investigated how the α-helical content depends on the Glauber temperature factor T, during the unfolding process. The result is shown in Fig. 7, over the entire temperature range of our simulation.

There is a remarkable similarity with the result shown in Fig. 7, and the experimentally observed circular dichroism data shown in Fig. 2 of Ref. [49]; see also Fig. 8. When we compare the two Figs., we obtain the following relation (50) between the Glauber temperature factor T and the physical temperature T_K (measured in Kelvin),

$$T = T_K \exp\{0.232 T_K - 107.8\}. \tag{52}$$

Figure 8 shows the comparison.

We note that according to Ref. [49] the heme becomes irreversibly damaged at around ∼75–80 °C (the red dashed line in Fig. 7). However, we point out that the structural stability of myoglobin varies between species; the myoglobin in Ref. [49] is from horse heart.

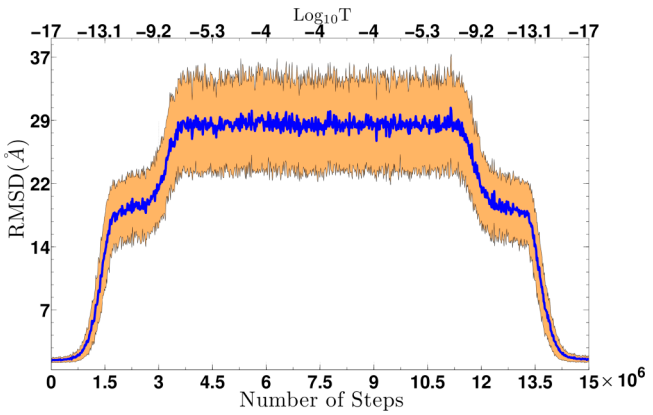

FIG. 6. Evolution of RMS distance between the 1ABS and the multisoliton during the heating-cooling cycle. No further increase in the RMS distance is observed if the Monte Carlo temperature factor is further increased. The blue line is the average and the orange band displays the one standard deviation from the average.

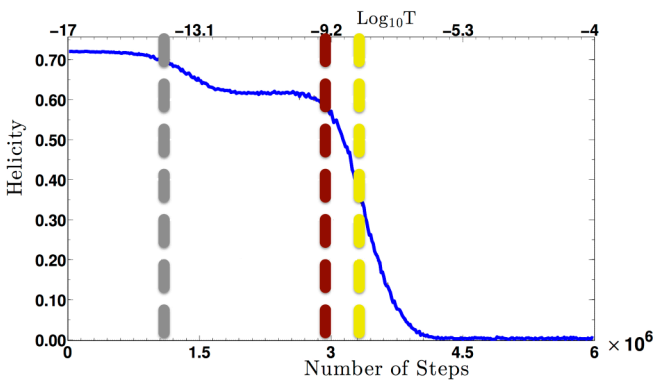

FIG. 7. Simulated α-helical content (in %) as a function of Glauber temperature factor T. The dashed gray line estimates ∼25 °C, the dashed red line estimates ∼75 °C, and the dashed yellow line estimates ∼90 °C in Fig. 2 of Ref. [49]; see also Fig. 8.

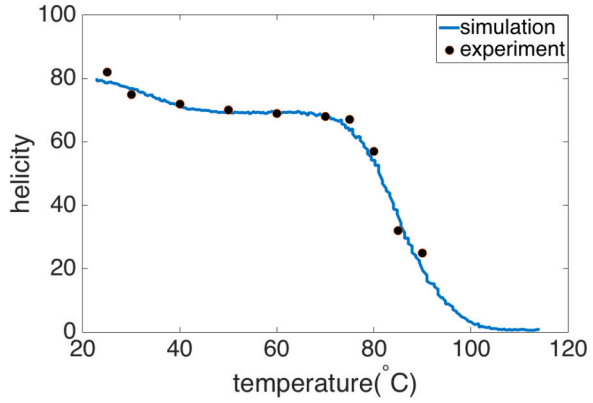

FIG. 8. Comparison between the simulation data in Fig. 7 and data inferred from Fig. 2 of reference [49], with Glauber temperature factor converted into centigrade scale using the functional relation (52). An overall uniform scaling factor of 1.13 has been introduced to account for the apparent scale difference between experimentally observed and computationally defined α-helical structure.

E. Radius of gyration and its susceptibility

In protein unfolding experiments, the radius of gyration R_g and its evolution is used widely to monitor the progress. In Fig. 9 we show how the radius of gyration evolves during our heating (unfolding) simulations.

We observe the presence of a folding intermediate between $\log_{10} T_L \approx -12.8$ and $\log_{10} T_H \approx -9.5$. The ensemble average value $R_g \approx$ 24 Å of the folding intermediate is very close to the experimentally observed value $R_{\rm exp} \sim 23.6$ Å of the molten globule, measured in the case of the apomyoglobin [36,41]. The increase in R_g during the first transition in Fig. 9, from native state to molten globule, is around 9 Å. This is larger than the experimentally observed 1–7 Å low pH values in the apomyoglobin [36,41], but the difference is in line with the observation that at low temperatures the radius of gyration of apomyoglobin is larger than that of heme containing myoglobin [36,41]. Between the molten globule and the fully unfolded state, the ensemble average difference $R_g \sim 10$ Å that we find is very close to the experimentally measured 11 ± 2 Å low pH value, in the case of apomyoglobin [36].

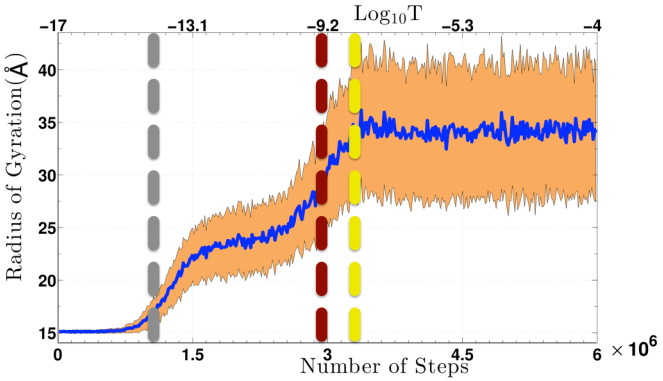

FIG. 9. The dependence of radius of gyration as a function of Glauber temperature factor. As in Fig. 6, the blue line is average value and the orange band denotes the one standard deviations fluctuation distance. A comparison shows that the transition temperatures of R_g are slightly lower than in the case of RMSD in Fig. 6. The dashed gray, red, and yellow lines are as in Fig. 7.

The transition temperatures during the unfolding process can be estimated by evaluating the following *radius of gyration susceptibility*:

$$\chi_g(T) = \frac{dR_g(T)}{d\log_{10} T}. \tag{53}$$

To evaluate this quantity, we introduce a fitting procedure where we first approximate $R_g(T)$ by a function of the form

$$\log_{10}(R_g(T)) \approx R_g^{\rm fit}(\log_{10} T),$$

where we choose

$$\begin{aligned} R_g^{\rm fit}(x) = {} & h_1 + h_2 \arctan[h_3(x - x_1)] \\ & + h_4\, x \arctan[h_5(x - x_2)] - h_6 x. \end{aligned} \tag{54}$$

This function form has been introduced and utilized in Ref. [77], in a related context. The numerical values of the parameters $h_1 \ldots h_6$ and $x_{1,2}$ are determined by a fit to the numerical values of $R_g(T)$ (not shown here). We then use the peaks in (53) to determine the transition values of T. There are two peaks during the heating process, at MC temperature factor values

$$\begin{aligned} T_{Rg,1} &\approx 10^{-13.8} \sim 30\,^\circ{\rm C} \\ T_{Rg,2} &\approx 10^{-9.3} \sim 74\,^\circ{\rm C}, \end{aligned} \tag{55}$$

where the relation (52) has been used.

We conclude that both in terms of α-helical content and radius of gyration, our model appears to correctly describe the observed reversible myoglobin unfolding dynamics below $\log_{10} T_c \approx -8.0$. In particular, in Fig. 9 we observe a folding intermediate between $\log_{10} T_L \approx -12.8$ and $\log_{10} T_H \approx -9.5$. These three temperature factor values correspond to the three dashed lines (gray, red, yellow) which we have also identified in Fig. 7.

F. Energy susceptibility

The various transitions can also be monitored by changes in energy, in terms of *energy susceptibility*. For this we first evaluate the average internal energy $\langle E \rangle$ as a function of the temperature factor T, in thermal equilibrium, using the LGW Hamiltonian (43)

$$\langle E \rangle = -\frac{\partial}{\partial \beta} {\rm Tr} \exp\{-\beta H\} \qquad \left(\beta = \frac{1}{T}\right).$$

The energy susceptibility is akin to the heat capacity,

$$\chi_E = \frac{d\langle E \rangle}{d\log_{10}(T)}. \tag{56}$$

We use a fitting function such as (54) to numerically estimate (56). In Fig. 10 we display the thermal equilibrium state energy susceptibility that we have computed. We observe, in Fig. 10(a), a clear peak, in the high-temperature regime, at

$$\log_{10} T_3 \approx 0.96 \sim 175\,^\circ{\rm C}.$$

It appears that, thus far, this peak has not been observed experimentally; the corresponding physical temperature value is quite high.

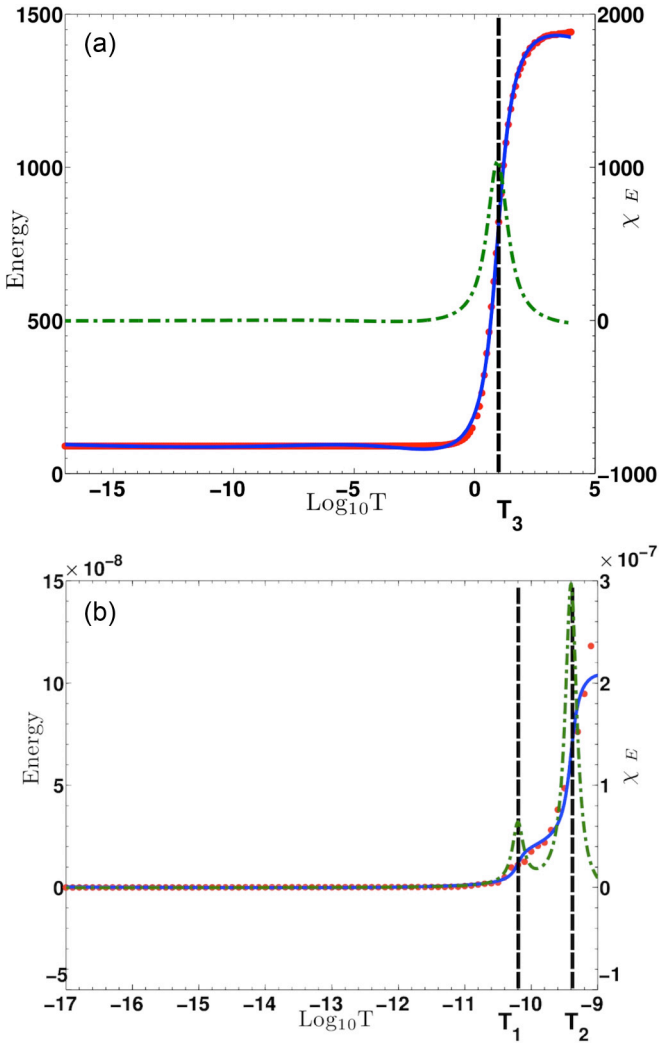

FIG. 10. The average energy and the energy susceptibility plots for different temperature ranges. (A) $T \in [10^{-17}, 10^{4}]$. (B) energy and heat capacity zoomed in $T \in [10^{-17}, 10^{-9}]$, where the scale on the energy axis has been subtracted by the energy at native state [equaling to ∼77.83 in units of (43)]. The red solid dots are the average energy values at corresponding temperature, and the blue line is the energy curve fittings based on (54). The green dashed line is the energy susceptibility calculated from (56). The temperatures of the energy susceptibility peaks are denoted as T1, T2, and T3.

We also observe two peaks at lower temperature values

$$\log_{10} T_1 \approx -10.2 \sim 65\,^{\circ}\mathrm{C}$$
$$\log_{10} T_2 \approx -9.4 \sim 73\,^{\circ}\mathrm{C}.$$

See Fig. 10(b). Note that in Fig. 10(a) these two peaks are not very visible, as their heights are much lower than the height of the high-temperature peak. The normalization of the peak height reflects the relation (50).

Our observation of two lower-temperature peaks, very close to each other, is consistent with the presence of an experimentally measured single wide peak [33] where the experimentally heat capacity peak at pH 5.0 is broad; see Fig. 5 in Ref. [33].

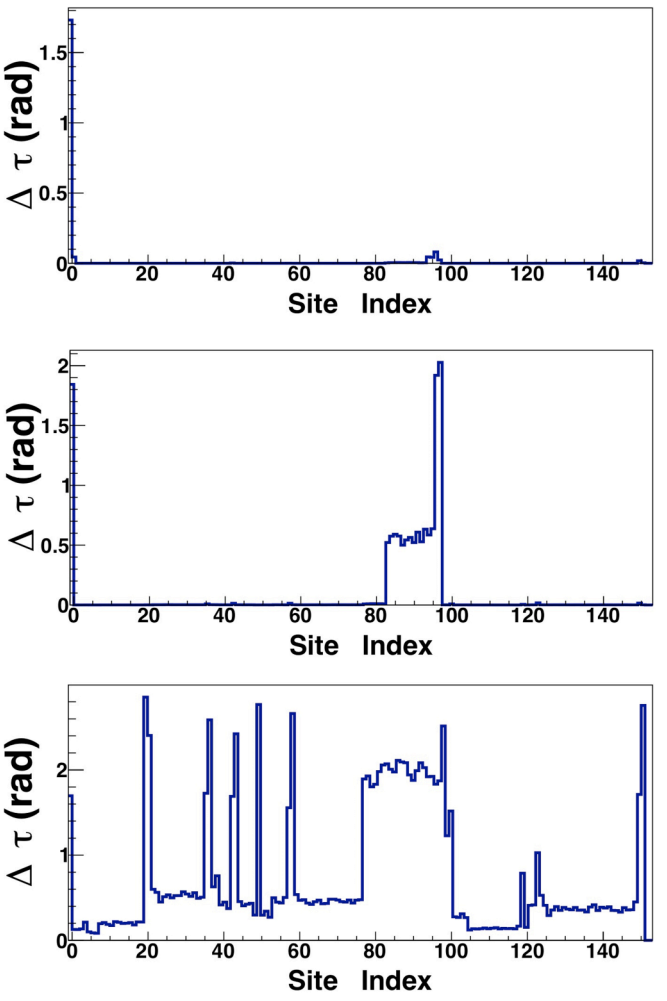

FIG. 11. The average values (57) at three different temperature factor values $\log_{10}(T) = -17, -13.1, -8.46$.

G. α-Helix denucleation

We monitor details of the unfolding process by evaluating the temperature dependence in the fluctuations $\Delta\tau_i$ of the individual backbone torsion angles, defined as follows:

$$\Delta\tau_i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\tau_{i,k} - \bar{\tau}_i)^2}. \tag{57}$$

The index k counts the conformation in a given heating simulation, and the average is over the entire ensemble of $N = 100$ heating simulations. In Figs. 11 and 12 we display the evolution of (57) at six different temperatures during the unfolding process.

We observe in particular how different helices become denucleated at different temperatures, during the unfolding process.

To determine the denucleation temperatures of the eight individual helical segments $X = (A, B, \ldots, H)$ in the natively folded myoglobin, we evaluate the following average values of the τ fluctuations:

$$\Delta\tau_X = \sqrt{\frac{1}{|X|}\sum_{X}(\Delta\tau_i)^2}. \tag{58}$$

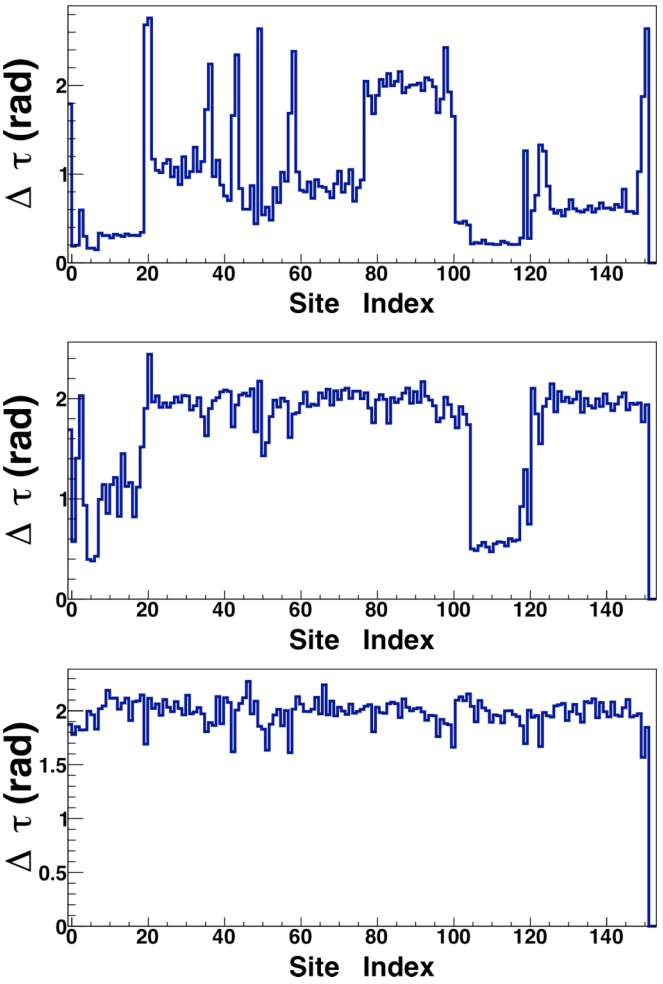

FIG. 12. The average values (57) at three different temperature factor values $\log_{10}(T) = -7.99, -7.06, -5.89$.

Here $|X|$ is the number of residues in the native helical segment X, and both $\Delta\tau_i$ and $\Delta\tau_X$ are evaluated at 1000 different sampling temperatures during the heating.

We introduce a susceptibility akin (53) and (56) to monitor the individual α-helix unfolding,

$$\mathcal{X}_\tau = \frac{d\Delta\tau_X}{d\log_{10} T}. \tag{59}$$

Figure 13 summarizes the results obtained from Eqs. (53), (56), and (59).

H. Contact maps

We monitor long-range interactions between any pair of different backbone segments in terms of a contact map. For this we denote by

$$d_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$$

the distance between *any* two Cα atoms. We define a scoring function for each pair of helical segments $X,Y = (A,B,\ldots,H)$ in the natively folded myoglobin as

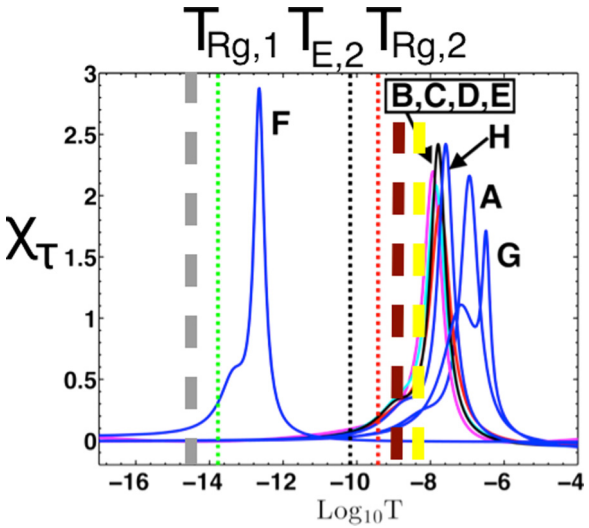

FIG. 13. The susceptibility (59) in myoglobin. Helices B, C, D, and E are identified by the colors magenta, cyan, red, and black, respectively. The vertical black dotted line denotes the maximum of (56) at $\log_{10} T_E \approx -10.2$. The vertical green and red dotted lines denote two R_g susceptibility peaks at $\log_{10} T_{Rg,1} \approx -13.8$ and at $\log_{10} T_{Rg,2} \approx -9.3$. The dashed gray, red, and yellow lines are as in Figs. 7 and 9. Units along ordinate derive from (59).

follows:

$$S_{ij} = \begin{cases} 0 & \text{for} \quad d_{ij} > 12 \\ \frac{12-d_{ij}}{4} & \text{for} \quad 8 \leqslant d_{ij} \leqslant 12. \\ 1 & \text{for} \quad d_{ij} < 8 \end{cases} \tag{60}$$

We define the average contact by

$$\mathcal{S}_{X,Y} = \frac{\sum_{i\in X}\sum_{j\in Y}\overline{S_{ij} - S_{ij0}}}{\min(|X|,|Y|)}. \tag{61}$$

Here $|X|$, $|Y|$ are the lengths of the helical segments X, Y; S_{ij0} is the value of (60) in the native state; and the average is taken over the entire ensemble. A detailed analysis of (57)–(61) confirms that in our simulations the thermal unfolding does indeed proceed sequentially, through helical intermediates, in a manner which is fully in line with the experimental observations: The individual contact maps for each of the eight helices are presented in Figs. 14–17.

Note that in the case of helices A and B we observe an increase in mutual contacts in the high temperature (unfolded) state. In the native state there are very few contacts between A and B, and, as a consequence, there is an apparent increase reflecting high-temperature random contact formation.

From the contact maps shown in Figs. 14–17, we estimate the critical temperature values, where a long-range interaction between each pair of helices disappears. The critical values of temperature factors T are shown in Table II for each pair of helices.

I. A summary of observations

According to our simulations, the α-helix unfolding in myoglobin takes place *only* in conjunction with the

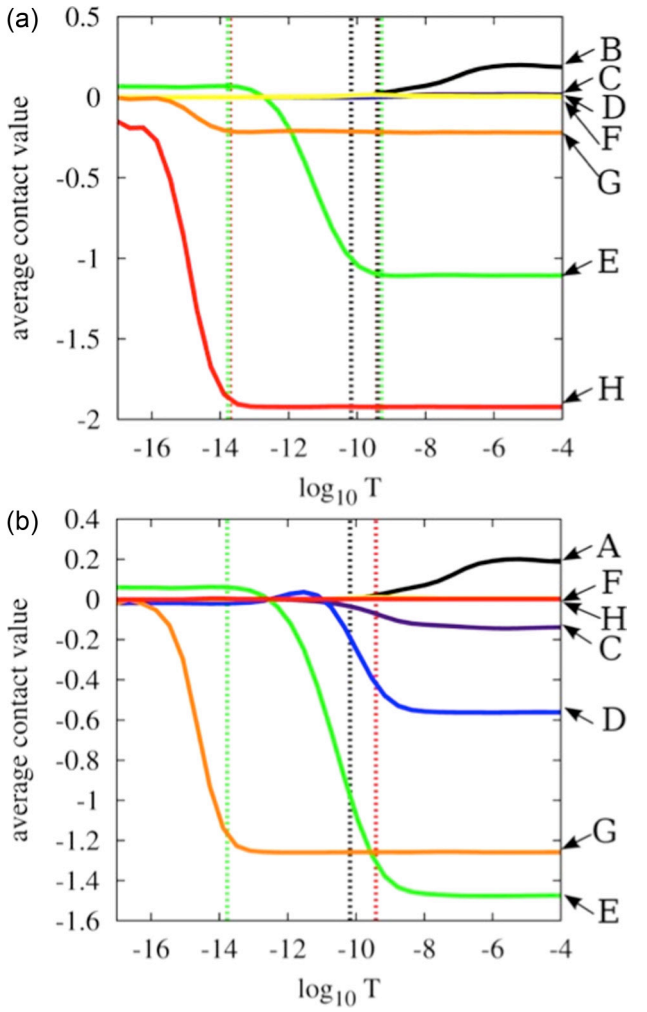

FIG. 14. Top: Contact maps between helix A with all other helices. Bottom: Contact maps between helix B with all other helices. The value of average contacts is defined by (61). The vertical lines indicate major changes in R_g and energy: The vertical red dashed line denotes the high-temperature peak of the susceptibility (53) and the vertical green dashed line denotes its low-temperature peak. The vertical black dashed line denotes the low-temperature energy susceptibility peak.

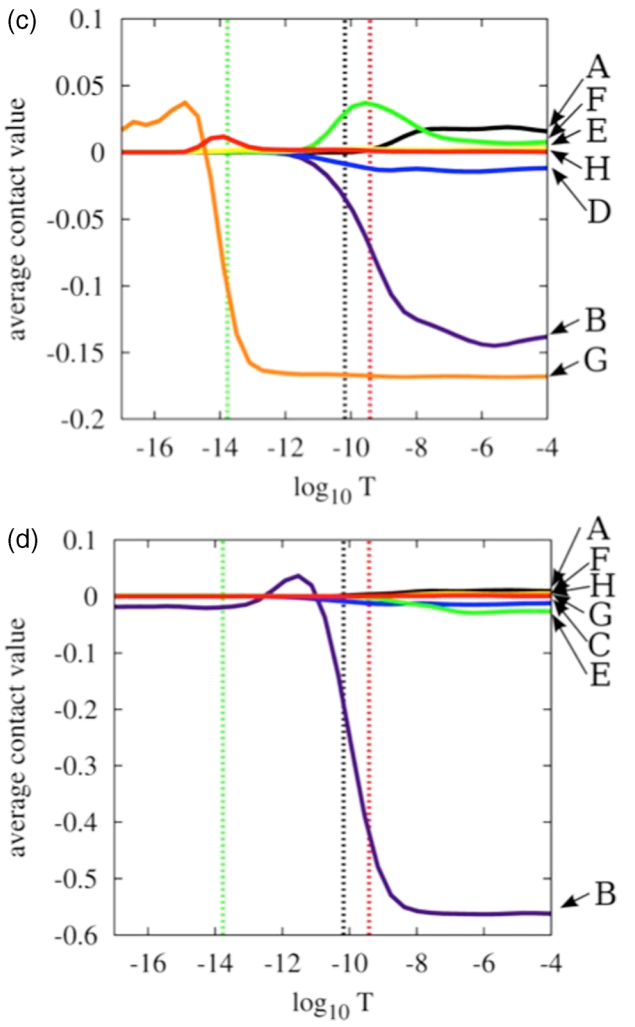

FIG. 15. Same as in Fig. 14 for helices C and D.

disappearance of long-range interactions between the ensuing helix and another helical segment of the backbone. In particular, we observe that the F helix becomes unstable already at relatively low temperatures before the molten globule forms. This is consistent with experiments that are made with heme containing myoglobin and in line with the observed disordered character of the F helix in the case of apomyoglobin. We find that the helices B, C, D, and E remain largely stable during the molten globule phase. Their unfolding coincides with the melting of the molten globule, apparently in conjunction with the irreversible destruction of the heme pocket [49]. Subsequently we observe the unfolding of A and H helices, in line with the original folding prediction in Refs. [78,79]. In our simulations, the G helix is the last to unfold.

The following diagram summaries our observed thermal unfolding of the myoglobin, with increasing temperature:

$$F \longrightarrow \text{molten globule} \longrightarrow B,C,D,E \longrightarrow A,H \longrightarrow G. \tag{62}$$

When all helices have become unfolded, which occurs at around $\log_{10} T \approx -6$ (yellow line in Figs. 7 and 9), the backbone appears to be in the universality class of self-avoiding random walk. In particular, we observe no further increase in R_g even when the ambient temperature becomes substantially increased.

When we adiabatically cool the system down to the original temperature, we observe that the helices form in an order which is opposite to that during the heating process.

IV. CONCLUSIONS

In conclusion, we have combined the Landau-Ginzburg-Wilson approach with nonequilibrium Glauber dynamics to model the way myoglobin unfolds when the ambient temperature increases. All our simulation results appear to be in excellent agreement with available experimental results. In addition, we have proposed new observables, including a detailed contact map between different helical segments. This

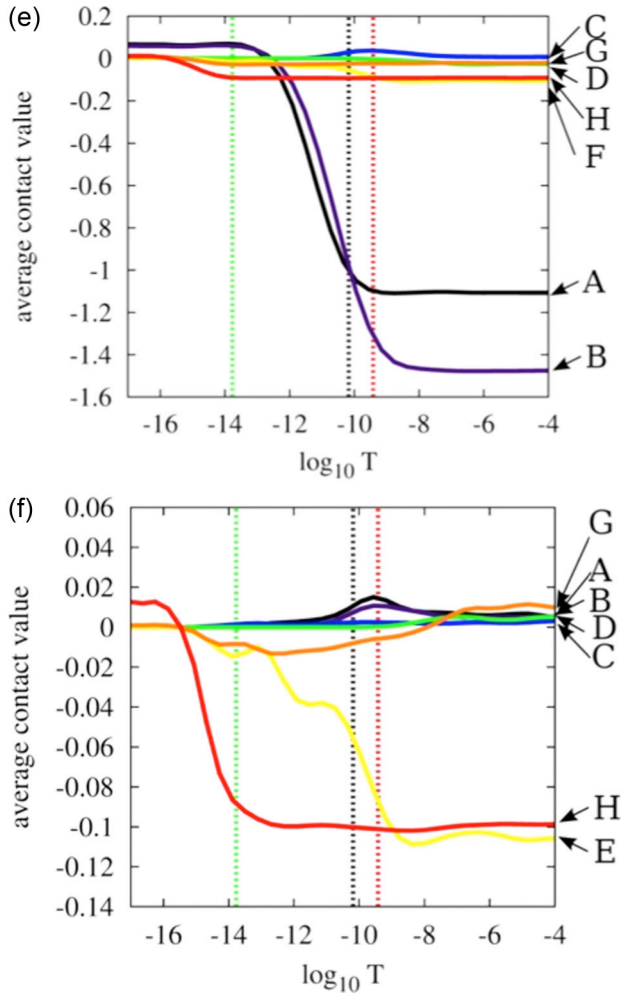

FIG. 16. Same as in Fig. 14 for helices E and F.

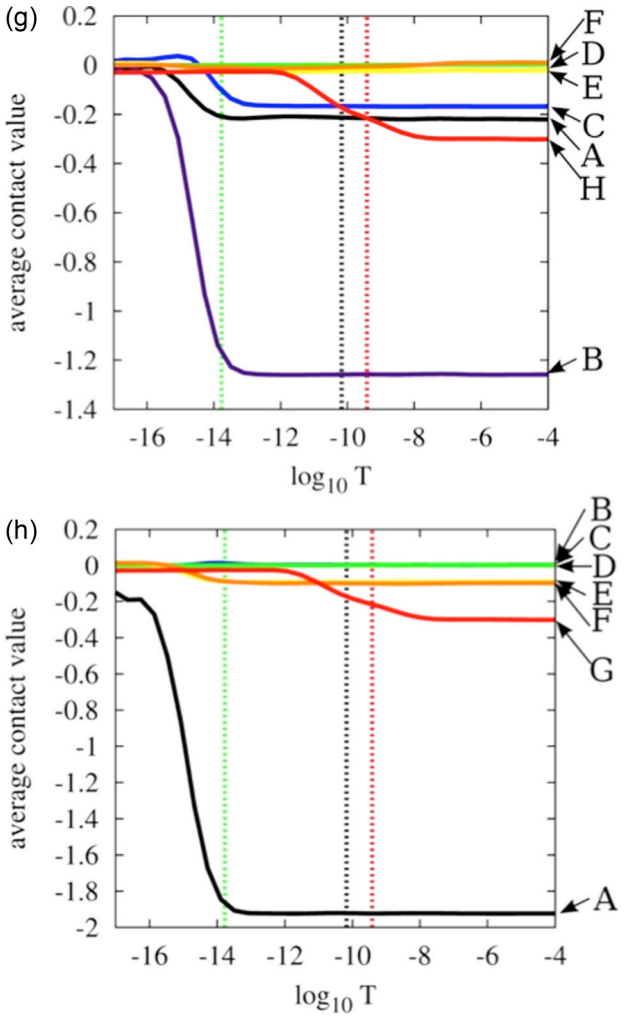

FIG. 17. Same as in Fig. 14 for helices G and H.

could be tested in future experiments to estimate the range of validity of the LGW approach in the case of proteins.

The approach that we have developed models both the natively folded low-temperature structure and its thermally driven unfolding process in terms of a multisoliton solution of the pertinent Landau-Ginzburg-Wilson energy function and its Glauber dynamics. In particular, the multisoliton is a solution of a universal discrete nonlinear Schrödinger equation. It approximates the Cα backbone profile in the limit where the spatial variations along the backbone have a long wavelength. This is the limit where one generally expects the Landau-Ginzburg-Wilson approach to become valid. Moreover, the presence of solitons furnishes the energy function with a substantial predictive power: The number of free parameters is even much less than the number of amino acids in myoglobin.

Our results propose that the unfolding process of a myoglobin is primarily driven by collective motions with a relatively long wavelength along the backbone. When the ambient temperature increases, these collective motions

TABLE II. Critical temperatures for disappearing interactions between helices. The symbol "*" indicates that no apparent change in contact is observed, in the contact map of Figs. 14–17.

	A	B	C	D	E	F	G	H
A	*	[−10,−6.5]	[−10,−7]	[−12,−9]	[−12,−8]	*	[−16,−13]	[−15,−14]
B	[−10,−6.5]	*	[−10,−6.5]	[−10,−7]	[−10,−7]	[−15,−12]	[−16,−13]	[−15,−12]
C	[−10,−7]	[−10,−6.5]	*	[−11,−6.5]	[−10,−8]	*	*	*
D	[−12,−9]	[−12,−6.5]	[−11,−6.5]	*	[−10,−6.5]	*	*	*
E	[−12,−8]	[−10,−7]	[−10,−8]	[−10,−6.5]	*	[−15,−12]	*	[−15,−13]
F	*	[−15,−12]	*	*	[−15,−12]	*	[−8,−5.5]	[−11,−8]&[−15,−13]
G	[−16,−13]	[−16,−13]	*	*	*	[−8,−5.5]	*	[−11,−6]
H	[−15,−14]	[−15,−12]	*	*	[−15,−13]	[−11,−8]&[−15,−13]	[−11,−6]	*

cause a stepwise melting of the individual solitons, until the backbone resembles a random chain.

Finally, we note that the present approach is not designed to answer at atomic-level scrutiny, how the protein folding proceeds. For a detailed investigation, in particular over short time periods, the present approach can be combined with all-atom molecular dynamics. We refer to Ref. [26], where such a hybrid approach has been investigated.

ACKNOWLEDGMENTS

A.K.S. was supported in part by National Science Centre, Poland Maestro (NCN, DEC-2012/06/A/ST4/00376), Foundation for Polish Science FNP (Mistrz7./2013), the Foundation for Polish Science (FNP START 100.2014), and a Swedish Institute scholarship. A.J.N. acknowledges support from the Vetenskapsrådet, Carl Trygger's Stiftelse för vetenskaplig forskning and Qian Ren Grant at BIT. We thank H. Scheraga for many discussions and comments. A.J.N. also thanks G. Petsko for communications and suggesting the PDB structure 1ABS, P. Jennings, P. Wright, and F. Wilczek for discussions, and J. Olson and S. Kundu for communications. A.J.N. thanks D. Melnikov and the International Institute of Physics—UFRN for hospitality during completion of this work. Computational resources were provided by the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdansk and by the 184-processor Beowulf cluster at the Faculty of Chemistry, University of Gdansk.

[1] C. Anfinsen and H. Scheraga, Adv. Prot. Chem. **29**, 205 (1975).
[2] L. D. Landau, Zh. Eksp. Teor. Fiz. **7**, 19 (1937).
[3] B. Widom, J. Chem. Phys. **43**, 3892 (1965).
[4] L. P. Kadanoff, Physics **2**, 263 (1966).
[5] K. Wilson, Phys. Rev. B **4**, 3174 (1971).
[6] M. E. Fisher, Rev. Mod. Phys. **46**, 597 (1974).
[7] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, Reading, MA, 1992).
[8] M. R. Douglas and M. Gross, *Dirichet Branes and Mirror Symmetry* (American Mathematical Society, Providence, RI, 2009).
[9] J. C. Kendrew *et al.*, Nature **181**, 662 (1958).
[10] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 6th ed. (Garland Science, New York, 2014).
[11] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, Ann. Rev. Biophys. **37**, 289 (2008).
[12] K. A. Dill and J. L. MacCallum, Science **338**, 1042 (2012).
[13] B. Montgomery Pettitt, J. Biomol. Struct. Dyn. **31**, 1024 (2013).
[14] E. López-Hernéndez and L. Serrano, Fold. Des. **1**, 43 (1996).
[15] R. Zhou, X. Huang, C. J. Margulis, and B. J. Berne, Science **305**, 1605 (2004).
[16] D. Bashford, F. Cohen, M. Karplus, I. Kuntz, and D. Weaver, Proteins **4**, 211 (1988).
[17] M. Karplus and D. L. Weaver, Prot. Sci. **3**, 650 (1994).
[18] M. Chernodub, S. Hu, and A. J. Niemi, Phys. Rev. E **82**, 011916 (2010).
[19] N. Molkenthin, S. Hu, and A. J. Niemi, Phys. Rev. Lett. **106**, 078102 (2011).
[20] A. Krokhotin, M. Lundgren, A. J. Niemi, and X. Peng, J. Phys.: Condens. Matter **25**, 325103 (2013).
[21] A. Krokhotin, M. Lundgren, and A. J. Niemi, Phys. Rev. E **86**, 021923 (2012).
[22] A. Krokhotin, A. Liwo, A. J. Niemi, and H. A. Scheraga, J. Chem. Phys. **137**, 035101 (2012).
[23] A. Krokhotin, A. Liwo, G. G. Maisuradze, A. J. Niemi, and H. A. Scheraga, J. Chem. Phys. **140**, 025101 (2014).
[24] A. K. Sieradzan, A. J. Niemi, and X. Peng, Phys. Rev. E **90**, 062717 (2014).
[25] N. Ilieva, A. J. Niemi, X. Peng, and A. Sieradzan, Biomath Comm. **2**, 1 (2015).
[26] J. Dai, A. J. Niemi, J. He, A. Sieradzan, and N. Ilieva, Phys. Rev. E **93**, 032409 (2016).
[27] P. L. Primalov, Adv. Protein Chem. **33**, 167 (1979).
[28] P. L. Primalov, Ann. Rev. Biophys. Biophys. Chem. **18**, 47 (1989).
[29] E. Shakhnovich, and A. Finkelstein, Biopolymers **28**, 1667 (1989).
[30] R. J. Glauber, Journ. Math. Phys. **4**, 294 (1963).
[31] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, J. Comput. Phys. **17**, 10 (1975).
[32] B. A. Berg, *Markov Chain Monte Carlo Simulations and Their Statistical Analysis* (World Scientific, Singapore, 2004).
[33] Y. V. Griko, P. L. Privalov, S. Y. Venyaminov, and V. P. Kutyshenko, J. Mol. Biol. **202**, 127 (1988).
[34] P. A. Jennings and P. E. Wright, Science **262**, 892 (1993).
[35] H.-C. Shin *et al.*, Biochemistry **32**, 6356 (1993).
[36] D. Eliezer *et al.*, Science **270**, 487 (1995).
[37] D. Eliezer and P. E. Wright, J. Mol. Biol. **263**, 531 (1996).
[38] M. Jamin. R. L. Baldwin, J. Mol. Biol. **276**, 491 (1998).
[39] M. Jamin, S. R. Yeh, D. L. Rousseau, and R. L. Baldwin, J. Mol. Biol. **292**, 731 (1999).
[40] T. Uzawa *et al.*, Proc. Natl. Acad. Sci. USA **101**, 1171 (2004).
[41] C. Nishimura, H. J. Dyson, and P. E. Wright, J. Mol. Biol. **355**, 139 (2006).
[42] T. Uzawa *et al.*, Proc. Natl. Acad. Sci. USA **105**, 13859 (2008).
[43] D. W. Meinhold and P. E. Wright, Proc. Natl. Acad. Sci. USA **108**, 9078 (2011).
[44] C. Nishimura, H. J. Dyson, and P. E. Wright, Journ. Mol. Biol. **411**, 248 (2011).
[45] M. Xu, O. Beresneva, R. Rosario, and H. Roder, J. Phys. Chem. **B116**, 7014 (2012).
[46] M. S. Hargrove, A. J. Wilkinson, and J. S. Olson, Biochemistry **35**, 11300 (1996).
[47] D. S. Culbertson and J. S. Olson, Biochemistry **49**, 6052 (2010).
[48] Y. Ochiai *et al.*, Biosci. Biotech. Biochem. **74**, 1673 (2010).
[49] Y. Moriyama and K. Takeda, J. Phys. Chem. **B114**, 2430 (2010).
[50] Y. Ochiai, World Academy of Science, Engineering and Technology **5**, 02 (2011).
[51] S. Uppal *et al.*, J. Biol. Chem. **290**, 1979 (2015).
[52] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, Nucl. Acids Res. **28**, 235 (2000); http://www.pdb.org.

[53] I. Schlichting, J. Berendzen, G. N. Phillips, Jr., R. M. Sweet, Nature **371**, 808 (1994).
[54] A. J. Hanson, *Visualizing Quaternions* (Morgan Kaufmann Elsevier, London 2006).
[55] J. B. Kuipers, *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace, and Virtual Reality* (Princeton University Press, Princeton, 1999).
[56] R. L. Bishop, Amer. Math. Monthly **82**, 246 (1974).
[57] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, Boston, 1995).
[58] A. J. Niemi, Phys. Rev. D **67**, 106004 (2003).
[59] H. Hasimoto, J. Phys. Soc. Jpn. **31**, 293 (1971).
[60] S. Coleman and E. Weinberg, Phys. Rev. D **7**, 1888 (1973).
[61] U. H. Danielsson, M. Lundgren, and A. J. Niemi, Phys. Rev. E **82**, 021910 (2010).
[62] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer-Verlag, Berlin, 2007).
[63] M. J. Ablowitz, B. Prinari, and A. D. Trubatch, *Discrete and Continuous Nonlinear Schrödinger Systems* (Cambridge University Press, Cambridge, 2004).
[64] P. G. Kevrekidis, *The Discrete Nonlinear Schrödinger Equation: Mathematical Analysis, Numerical Computations and Physical Perspectives* (Springer-Verlag, Berlin, 2009).
[65] O. Kratky and G. Porod, Rec. Trav. Chim. Pays-Bas. **68**, 1106 (1949).
[66] S. Hu, Y. Jiang, and A. J. Niemi, Phys. Rev. D **87**, 105011 (2013).
[67] T. Ioannidou, Y. Jiang, and A. J. Niemi, Phys. Rev. D **90**, 025012 (2014).
[68] N. Manton and P. Sutcliffe, *Topological Solitons* (Cambridge University Press, Cambridge, 2004).
[69] A. J. Niemi, Theor. Math. Phys. **181**, 1235 (2014).
[70] A. J. Niemi, arXiv:1412.8321.
[71] S. Hu, M. Lundgren, and A. J. Niemi, Phys. Rev. E **83**, 061908 (2011).
[72] K. Hinsen, S. Hu, G. R. Kneller, and A. J. Niemi, J. Chem. Phys. **139**, 124115 (2013).
[73] A. Krokhotin, A. J. Niemi, and X. Peng, Phys. Rev. E **85**, 031906 (2012).
[74] M. L. Scalley and D. Baker, Proc. Natl. Acad. Sci. USA **94**, 10636 (1997).
[75] A. Sinelnikova, A. J. Niemi, and M. Ulybyshev, Phys. Rev. E **92**, 032602 (2015).
[76] http://folding-protein.org/.
[77] M. N. Chernodub, M. Lundgren, and A. J. Niemi, Phys. Rev. E **83**, 011126 (2011).
[78] R. R. Matheson, Jr. and H. A. Scheraga, Macromolecules **11**, 819 (1978).
[79] M. Gerritsen, K. C. Chou, G. Némethy, and H. A. Scheraga, Biopolymers **24**, 1271 (1985).