Multiple scales and phases in discrete chains with application to folded proteins

A. Sinelnikova,[1,*] A. J. Niemi,[1,2,3,4,†] Johan Nilsson,[1,‡] and M. Ulybyshev[5,§]

[1]*Department of Physics and Astronomy, Uppsala University, P.O. Box 516, S-75120 Uppsala, Sweden*
[2]*Nordita, Stockholm University, Roslagstullsbacken 23, SE-106 91 Stockholm, Sweden*
[3]*Laboratory of Physics of Living Matter, School of Biomedicine, Far Eastern Federal University, Vladivostok, Russia*
[4]*Department of Physics, Beijing Institute of Technology, Haidian District, Beijing 100081, People's Republic of China*
[5]*Institute of Theoretical Physics, University of Regensburg, Universitätsstraße 31, D-93053 Regensburg, Germany*

(Received 23 May 2017; revised manuscript received 5 March 2018; published 8 May 2018)

Chiral heteropolymers such as large globular proteins can simultaneously support multiple length scales. The interplay between the different scales brings about conformational diversity, determines the phase properties of the polymer chain, and governs the structure of the energy landscape. Most importantly, multiple scales produce complex dynamics that enable proteins to sustain live matter. However, at the moment there is incomplete understanding of how to identify and distinguish the various scales that determine the structure and dynamics of a complex protein. Here we address this impending problem. We develop a methodology with the potential to systematically identify different length scales, in the general case of a linear polymer chain. For this we introduce and analyze the properties of an order parameter that can both reveal the presence of different length scales and can also probe the phase structure. We first develop our concepts in the case of chiral homopolymers. We introduce a variant of Kadanoff's block-spin transformation to coarse grain piecewise linear chains, such as the Cα backbone of a protein. We derive analytically, and then verify numerically, a number of properties that the order parameter can display, in the case of a chiral polymer chain. In particular, we propose that in the case of a chiral heteropolymer the order parameter can reveal traits of several different phases, contingent on the length scale at which it is scrutinized. We confirm that this is the case with crystallographic protein structures in the Protein Data Bank. Thus our results suggest relations between the scales, the phases, and the complexity of folding pathways.

DOI: 10.1103/PhysRevE.97.052107

I. INTRODUCTION

A linearly conjugated polymer is conventionally viewed as a piecewise linear polygonal chain, connecting a sequence of vertices that coincide with the locations of its skeletal atoms [1–10]. For example, in the case of a protein the vertices coincide with the positions of the Cα atoms, and the connecting line segments concur with the diagonals of the peptide planes. Conventionally, a phase is then assigned to the polymer by inspecting the fractal geometry of the chain. For this let $(\mathbf{x}_0,\ldots,\mathbf{x}_N)$ denote the $N+1$ vertices of a given discrete chain Γ. When N becomes large the radius of gyration

$$R_{\rm gyr} = \sqrt{\frac{1}{2(N+1)^2}\sum_{i,j=0}^{N}(\mathbf{x}_i-\mathbf{x}_j)^2} \tag{1}$$

admits an asymptotic expansion of the form [9,12–14]

$$R_{\rm gyr}^2 \approx L_0^2 N^{2\nu}(1+R_1 N^{-\Delta_1}+\cdots) \overset{N \text{ large}}{\longrightarrow} L_0^2 N^{2\nu}. \tag{2}$$

The prefactor L_0 is an effective segment length (Kuhn length). In the case of a polymer its value depends on the atomic level details of the chain and the environment; it is not a universal quantity. The scaling exponent ν coincides with the inverse Hausdorff dimension of Γ. It is a universal quantity [9–14] with a numerical value that does not depend on the atomic level details of the chain. It acts as an order parameter that can detect the phase of the chain. The exponents $\Delta_1,\Delta_2,\ldots$ that characterize the finite-size corrections are similarly universal [13].

In the case of a discrete chain with a homogeneous structure, the scaling exponent ν is commonly presumed to have only four possible values, corresponding to the four different phases of a homopolymer. At the level of classical mean field theory [1–3,9–14]

$$\nu = \begin{cases} 1/3, & \text{collapsed},\\ 1/2, & \text{random walk (RW)},\\ 3/5, & \text{self-avoiding random walk (SARW)},\\ 1, & \text{straight rod}. \end{cases} \tag{3}$$

The value of ν is determined using (2), by successively increasing the number of vertices $\mathbf{x}_i$ and by observing how the radius of gyration scales when N becomes large. This procedure can work in the case of a homopolymer, when the number of vertices can be increased in an unambiguous manner. Unfortunately, it does not work in the case of a heteropolymer such as a given protein structure, where the amino acid assignment is fixed: There is no unambiguous way to extend the length of a protein, to determine the scaling

[*] Anna.Sinelnikova@physics.uu.se
[†] Antti.Niemi@physics.uu.se; http://www.folding-protein.org
[‡] Johan.Nilsson@physics.uu.se
[§] Maksim.Ulybyshev@physik.uni-regensburg.de

2470-0045/2018/97(5)/052107(21)

of its radius of gyration when N grows, as the number of vertices cannot be systematically increased in a unique manner as required by (2). Instead one can try to deduce the value of ν statistically, by comparing the radius of gyration of a given protein to a statistical pool of different lengths but similar kinds of protein structures such as those classified as α-helical or β-stranded in the Protein Data Bank (PDB) [15]. This procedure has some merits, but it lacks rigor. Moreover, it brings about intriguing but difficult-to-verify proposals, including a suggestion that since, e.g., α-helical and β-stranded proteins have different ν values, the ensuing chains reside in different phases [16–19]. To clarify issues like these, there is need to introduce alternative order parameters, that can directly and unambiguously probe the phase of a given heteropolymer with no need to artificially extend its length.

Here we propose such an order parameter. It builds on the properties of an observable that we introduce, a variant of Kadanoff's block-spin transformation [20–23] that we design to coarse grain a fixed chain in an effective manner. We show how the observable detects different scales, as the coarse graining proceeds. For a homopolymer, we confirm the universal phase structure in line with (3). However, in the case of a heterogeneous chain such as the Cα backbone of a folded protein we find that our observable is *variable*. Its value depends on the scale and it oscillates, apparently between different phases, as the coarse graining proceeds. We interpret this variable character of the observable in terms of a multiphase structure: Depending on the distance scale at which a heteropolymer chain is inspected, it can display different phase properties. For this we recall the following: In the case of N pointlike, chemically independent components there are *a priori* N different dimensionful parameters. The Gibbs phase rule states that in the presence of F intensive thermodynamical variables the number P of coexisting phases is limited by

$$P \leqslant N - F + 2. \tag{4}$$

When the elemental constituents are not pointlike but chainlike, this rule can change. If a relation between the number of dimensionful parameters and the number of thermodynamical phases persists, even a single chain might exhibit different phase characteristics when we inspect it at different length scales. For example, consider a crystallographic globular protein in a collapsed phase. At the same time, at distance scales that are short in comparison to the radius of gyration, its structure can be dominated, e.g., by α helices or β strands which are both in the phase of a straight rod. Thus there is an intermediate length scale, at which the character of the protein structure transits from the straight rod phase to the collapsed phase. The scaling exponent (3) is unsuitable for detecting how such a transition from a regime dominated by a straight rod phase to a collapsed phase regime takes place. But our observable can detect the presence of a scale that produces a transition.

We start by describing the generic theoretical properties of our observable. We then proceed to investigate these properties numerically. Our goal is to develop the observable into a tool that can detect the phase of the chain. For this we devise a variant of the Kadanoff block-spin transformation, specially tailored to inspect chainlike objects. We analyze the ensuing transformation properties of the observable in terms of Monte Carlo simulations using a homopolymer model. We show that the phase structure is in line with the classification (3).

We then continue to apply the observable to inspect crystallographic folded protein structures. We find that for a globular protein with an apparently complex structure, the scaling properties of the observable become highly nontrivial and much more elaborate than in the case of a homopolymer. We propose that the scaling properties of the observable reflect the presence of different phase structures in a complex protein, when inspected at different length scales. Such a presence of a multiple, scale-dependent phase structure should have profound effects on the folding and unfolding transitions, and on other dynamical and structural properties of a protein. In particular, we propose that a protein that displays an elaborated, scale-dependent structure in terms of our observable cannot be a simple two-state folder, but should display a much more complex folding pattern.

II. OBSERVABLES AND PHASE DIAGRAMS

A. Our observable

Let $\mathbf{t}_i$ denote the segment from the vertex $\mathbf{x}_{i-1}$ to the subsequent vertex $\mathbf{x}_i$ along a discrete linear chain Γ, with a total of $N+1$ vertices $(\mathbf{x}_0,\ldots,\mathbf{x}_N)$:

$$\mathbf{t}_i = \mathbf{x}_i - \mathbf{x}_{i-1}. \tag{5}$$

We introduce the following observable:

$$\mathcal{P}_\Gamma(N) = \sum_{1\leqslant i<j\leqslant N} \frac{\mathbf{t}_i\cdot\mathbf{t}_j}{|\mathbf{t}_i||\mathbf{t}_j|} \equiv \sum_{1\leqslant i<j\leqslant N} \cos\kappa_{ij}, \tag{6}$$

where in line with (2) we expect the following asymptotic expansion:

$$\approx \mathcal{P}_\Gamma N^\sigma (1 + \mathcal{Q}_1 N^{-\delta_1} + \cdots). \tag{7}$$

Here both the N-independent factor $\mathcal{P}_\Gamma$ and the scaling exponent σ are the quantities that are of interest to us in the sequel. We shall also consider the finite-size corrections specified by $\mathcal{Q}_1,\delta_1$, etc. The quantity (6), (7) bears resemblance to the radius of gyration (1), except that (6), (7) is dimensionless. We also note that (6), (7) relates to, but is quite different from, the concept of folding angle introduced in [24].

In the sequel we shall introduce a chain-specific coarse graining transformation of (6), (7) akin to the Kadanoff block-spin transformation [20–22] of renormalization group equations [11,14]. We follow how *both* $\mathcal{P}_\Gamma$ and σ evolve during the ensuing flow, and deduce the phase properties of Γ. Even though the numerical value of $\mathcal{P}_\Gamma$ apparently lacks universality, the sign of $\mathcal{P}_\Gamma$ and the numerical value of σ are both specific to the phase where the chain resides.

As an example, consider the straight rod phase where $\nu = 1$. Take a chain that has a linear structure, such that all the vertices lie in the vicinity of a given straight line. Then, in the limit of a large number of vertices,

$$\mathcal{P}_\Gamma(N) \xrightarrow{N\gg 1} C\,\frac{N(N-1)}{2} = \frac{C}{2}(N^2 - N). \tag{8}$$

Here C characterizes the average value of the cosine of the angle between two vectors $\mathbf{t}_i$ and $\mathbf{t}_j$. When the chain becomes straight so that the vectors $\mathbf{t}_k$ are close to parallel, we have

$C \to 1$. This example makes it clear that (6), (7) can never grow faster than N^2. We also note the finite-size correction which is proportional to N in (8); it coincides with the number of nearest neighbor segments along the chain. Finally, in the case of regular protein structures the bond angle $\kappa_{i,i+1}$ between two neighboring Cα atoms along a β strand has the value $\kappa_{i,i+1} \approx 1$ (rad) while along α helices the value is $\kappa_{i,i+1} \approx \pi/2$ (rad) [25]. Thus, in the case of β-stranded proteins we expect a positive-valued finite-size correction $O(N)$ due to nearest neighbor vertices, while in the case of α-helical proteins the finite-size correction due to nearest neighbors should be tiny.

B. Statistical ensembles

We proceed to develop (6), (7) into an order parameter that can probe the phase structure of chains. For this we analyze the statistical ensemble average

$$\langle \mathcal{P}_\Gamma(N) \rangle = \mathrm{Tr}\{\mathcal{P}_\Gamma(N)\rho(\Gamma)\} \tag{9}$$

of the observable (7) in the different phases (3). Here $\rho(\Gamma)$ is a density matrix that determines the thermodynamical ensemble. In our numerical simulations we assume that the system is in a thermodynamical equilibrium state, with $\rho(\Gamma)$ admitting the Gibbsian form

$$\rho(\Gamma) \propto e^{-\beta H} \tag{10}$$

with β the temperature factor and H the Hamiltonian of the chain. In general, the Hamiltonian can depend on multiple length scales $H = H(L_1, \ldots, L_N)$. Accordingly, the canonical partition function can engage various effective thermal de Broglie distances β/L_i^2,

$$\mathrm{Tr}\{\rho(\Gamma)\} = Z\big(\beta L_i^2, L_j/L_k\big), \tag{11}$$

in addition of the dimensionless rations of the various length scales. As a consequence the ensuing phase structure can be quite involved, with various putative critical temperatures $1/\beta_i \propto L_i^2$,

$$\frac{1}{\beta_i^{\mathrm{crit}}} = kT_i^{\mathrm{crit}} \propto L_i^2. \tag{12}$$

1. The random walk

In a random walk the vertices along the chain are mutually independent. The density matrix has the form [9,11]

$$\rho_0(\Gamma) = \delta(\mathbf{x}_0) \prod_{i=1}^{N-1} g(\mathbf{x}_{i-1} - \mathbf{x}_i), \tag{13}$$

where $g(\mathbf{x} - \mathbf{x}')$ is a probability distribution with Gaussian pairwise probabilities,

$$g(\mathbf{x} - \mathbf{x}') = \left(\frac{1}{2\pi a^2}\right)^{3/2} \exp\left[-\frac{1}{2a^2}(\mathbf{x} - \mathbf{x}')^2\right]. \tag{14}$$

We fix the initial point $\mathbf{x}_0$ to the origin, in order to eliminate the space volume as an (infinite) overall normalization factor. We find

$$\begin{aligned}\langle \mathbf{t}_i \cdot \mathbf{t}_j \rangle &= \mathrm{Tr}\{(\mathbf{t}_i \cdot \mathbf{t}_j)\rho_0(\Gamma)\} \\ &= \int d^{N+1}\mathbf{x}\, (\mathbf{t}_i \cdot \mathbf{t}_j)\delta(\mathbf{x}_0) \prod_{k=1}^{N-1} g(\mathbf{x}_k - \mathbf{x}_{k-1}) = 0 \ \ (i \neq j).\end{aligned} \tag{15}$$

Thus the ensemble average of the observable (6), (7) vanishes in the random walk phase.

2. The hard-sphere repulsion and SARW

In the self-avoiding random walk (SARW) phase there are repulsive interactions between the vertices. These interactions can have a varying range in terms of the spatial separation, from the short-distance Pauli (steric) repulsion to the extensive reach of Coulomb interaction. But the effect is always that of a long-range interaction, when we measure distance along the chain. In a weak-coupling limit we can try to handle these interactions perturbatively, using a virial expansion in the range of the expansion. For this we assume a homogeneous chain with N vertices, with a short-range interaction potential of the form

$$E = \sum_{1 \leqslant i < j \leqslant N-1} U(\mathbf{x}_i - \mathbf{x}_j) \tag{16}$$

between the vertices; the summation extends over all vertex pairs. The density matrix has the Gibbsian form (10)

$$\rho(\Gamma) = \prod_{1 \leqslant i < j \leqslant N-1} e^{-\beta U(\mathbf{x}_i - \mathbf{x}_j)} \prod_{i=1}^{N-1} g(\mathbf{x}_{i-1} - \mathbf{x}_i). \tag{17}$$

We proceed to an explicit calculation of (6), (7) in the limit where the interaction is entirely due to excluded volume; i.e., we assume there is only a hard-sphere steric repulsion between the vertices:

$$U(\mathbf{x}_i - \mathbf{x}_j) = \begin{cases} \infty, & \text{if } |\mathbf{x}_i - \mathbf{x}_j| \leqslant \Delta, \\ 0, & \text{if } |\mathbf{x}_i - \mathbf{x}_j| > \Delta. \end{cases} \tag{18}$$

Note that in this hard-sphere limit the temperature dependence becomes absent. Thus there are no (temperature-dependent) phase transitions. A single phase prevails and, in the absence of any other interaction, the chain resides in the SARW phase by construction.

We follow [11] and introduce the Mayer function

$$f(\mathbf{x}_i - \mathbf{x}_j) = e^{-\beta U(\mathbf{x}_i - \mathbf{x}_j)} - 1 = \begin{cases} -1, & \text{if } |\mathbf{x}_i - \mathbf{x}_j| \leqslant \Delta, \\ 0, & \text{if } |\mathbf{x}_i - \mathbf{x}_j| > \Delta, \end{cases} \tag{19}$$

and we consider the limit where the hard-sphere radius Δ at the vertex is very small so that

$$f(\mathbf{x}_i - \mathbf{x}_j) = -\tfrac{4}{3}\pi\Delta^3\delta(\mathbf{x}_i - \mathbf{x}_j) \equiv -B\delta(\mathbf{x}_i - \mathbf{x}_j). \tag{20}$$

Here B specifies the excluded volume around a vertex. The virial expansion is

$$\begin{aligned} e^{-\beta E} &= \prod_{1 \leqslant i < j \leqslant N-1} [1 + f(\mathbf{x}_i - \mathbf{x}_j)] \\ &= 1 + \sum_{1 \leqslant i < j \leqslant N-1} f(\mathbf{x}_i - \mathbf{x}_j) \\ &\quad + \sum_{i,j,k,l} f(\mathbf{x}_i - \mathbf{x}_j) f(\mathbf{x}_k - \mathbf{x}_l) + \cdots . \end{aligned} \tag{21}$$

The term which is linear in the Mayer function describes collisions between a pair of vertices; note that the linear term engages interactions that have a long range *along the chain* despite being short range *in space*. The bilinear term describes triple collisions, and so on. In the limit of a dilute chain only pair collisions can be relevant. Thus, in this limit we obtain for our observable the second-order virial approximation

$$\begin{aligned}&\langle \mathcal{P}_\Gamma(N)\rangle \\ &\approx \int \Bigg\{ d\mathbf{x}_0 \cdots d\mathbf{x}_{N-1} \prod_{i=1}^{N-1} g(\mathbf{x}_i - \mathbf{x}_{i-1}) \\ &\quad\times \Bigg(1 - B \sum_{1\leqslant i<j\leqslant N-1} \delta(\mathbf{x}_i - \mathbf{x}_j)\Bigg) \mathcal{P}_\Gamma(\mathbf{x}_0,\ldots,\mathbf{x}_{N-1})\delta(\mathbf{x}_0)\Bigg\}.\end{aligned} \tag{22}$$

We substitute (7) in (22). The integrals are elemental and in the limit where the vertex size is very small in comparison to the segment length $B \ll a^3$, the result is

$$\langle \mathcal{P}_\Gamma(N)\rangle = \left(\frac{3}{2\pi}\right)^{\frac{3}{2}} \frac{B}{2a^3} \sum_{1\leqslant i<j\leqslant N-1} \frac{1}{\sqrt{j-i}} + O\left(\frac{B}{a^3}\right)^2. \tag{23}$$

3. The large-N limit in SARW

For large N we can estimate the sum in (23) using an integral approximation

$$\sum_{1\leqslant i<j\leqslant N-1} (j-i)^{-\frac{1}{2}} \xrightarrow{N\gg 1} \int_0^N dx \int_0^x dy \frac{1}{\sqrt{x-y}} \sim N^{3/2}. \tag{24}$$

We then get for the large-N limit

$$\langle \mathcal{P}_\Gamma(N)\rangle \xrightarrow{N\gg1} D\,\frac{B}{a^3} N^{3/2} \equiv \mathcal{P}_\Gamma\, N^{3/2} > 0 \quad \text{(SARW)} \tag{25}$$

with some chain-specific positive constant D. We note that the observable (25) is proportional to $N^{3/2}$ while the number of terms that contribute in (6), (7) increase like N^2, according to (8). Thus, there must be cancellations of order N^2: We conclude that *in the leading order* there is an equal contribution from terms with positive and negative values of $\cos \kappa_{ij}$, and the result (25) follows due to subleading predominance of positively valued $\cos \kappa_{ij}$. Moreover, since the $x \to y$ singularity in (24) is integrable, in the large-N limit the contribution from small separation values of $|i-j|$ becomes insignificant in comparison to the contribution from large separation values $|i-j|$. Thus the fine details of the interaction potential become increasingly irrelevant. Accordingly, we argue that the dominant scaling exponent $\sigma = 3/2$ in (25) is *universal*, for discrete chains in the SARW phase when N is very large.

We note that the *positive* sign of (25) can be understood as follows: In the SARW phase both the radius of gyration and the end-to-end distance increase faster in N than in the RW phase. This implies that there is a tendency in SARW phase for the different vectors $\mathbf{t}_i, \mathbf{t}_j$ to be more parallel to each other than in the RW phase. In the RW phase the ensemble average of the angle κ_{ij} between any two vectors $\mathbf{t}_i$ and $\mathbf{t}_j$ is $\pi/2$. Thus, in the SARW phase there is an inclination towards $\kappa_{ij} < \pi/2$ implying that (25) is positive.

Finally, the derivation presented here presumes the limit where any interaction between vertices has a very short range in relation to the segment length. Certainly, there are corrections due to higher order terms in the virial expansion, in cases where the interaction has a longer range but the corrections vanish in the limit where the number of vertices becomes very large. The present virial expansion based derivation is appropriate to identify the leading large-N asymptotic of the observable which is a universal characteristic of the SARW phase; the higher order terms in the virial expansion amount to finite-size corrections.

4. Corrections to large-N limit in SARW phase

Since the result (25) reflects a $O(N^2)$ balance between positive and negative values of $\cos \kappa_{ij}$, we can expect that when N is not very large the higher order correction terms in (7) are notable. To analyze these, we observe that oftentimes there is a steric repulsion that enforces a minimum distance between *any* two vertices. For example, in PDB protein structures the minimal distance between any two Cα atoms that do not share a peptide plane is (practically) always larger than the diagonal size ~3.8 Å of a peptide plane. Thus the angle $\kappa_{i,i+1}$ between any two neighboring vectors $\mathbf{t}_i$ and $\mathbf{t}_{i+1}$ is always less than $2\pi/3$. In fact, for most proteins we can confirm that [25]

$$\kappa_{i,i+1} \leqslant \kappa_{\max} \approx \pi/2 (\text{for PDB proteins}).$$

In the general case, the value of $\kappa_{\max} < \pi$ is determined by the ratio between the *effective* hard-sphere radius (18) and the segment length. Accordingly, there must be a finite-N correction to (25) which reflects the local details of steric repulsion. We estimate the correction in the hard-sphere limit, by separating out the effect of very short distance interactions, i.e., contribution from small values of $k = |i-j|$. For this we simply subtract the effect of the ensuing interactions by replacing (24) with

$$\begin{aligned}&\sum_{1\leqslant i<j\leqslant N-1} (j-i)^{-\frac{1}{2}} \xrightarrow{\frac{k}{N}\ll 1} \sum_{\substack{i=0\\ j=i+k}}^{N} (j-i)^{-\frac{1}{2}} \\ &\longrightarrow \int_k^N dx \int_0^{x-k} dy\, \frac{1}{\sqrt{x-y}} \sim \frac{2}{3}N^{3/2} - \sqrt{k}N.\end{aligned} \tag{26}$$

In lieu of (25) we then have an estimate that excludes the short-distance effects of steric repulsion:

$$\mathcal{P}_\Gamma(N) \sim \mathcal{P}_\Gamma\left(N^{3/2} - \frac{3}{2}\sqrt{k}N\right) \quad \left(\frac{k}{N} \ll 1\right). \tag{27}$$

Note that this result is derived using the limit in which the radius of the hard-sphere repulsion becomes small. In the case of most proteins we have noted that mostly $\kappa_{i,i+1} \leqslant \pi/2$ (rad); thus we expect that the $O(N)$ contribution from *nearest* neighbor vertices is often non-negative. Accordingly, in practical scenarios the estimate (27) should apply when k is not very small.

5. The collapsed phase

In the collapsed phase we cannot estimate (7) using a perturbation (virial) expansion around an ideal RW chain. In

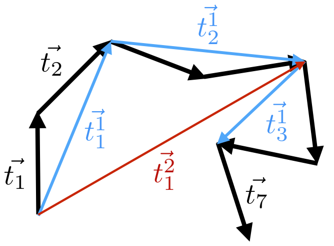


FIG. 1. A scaling akin to Kadanoff's block-spin transformation that successively combines two preceding segments into one following segment and eliminates the middle vertex. The segments in the initial chain are bold black (vectors $\mathbf{t}_1,\ldots,\mathbf{t}_7$), those in the first level of iteration are blue (light gray) (vectors $\mathbf{t}_1^1,\ldots,\mathbf{t}_3^1$), and the thinnest red (dark gray) segment (vector $\mathbf{t}_1^2$) is obtained at the second level of iteration.

the collapsed phase both repulsive and attractive long-distance interactions along a chain are present; the chain properties are ruled by nonperturbative effects.

According to (3) when N increases, for a chain in the collapsed phase the radius of gyration grows slower in N than in the RW phase. Thus, in a statistical ensemble the angle $\kappa_{i,j}$ between any two vectors $\mathbf{t}_i$ and $\mathbf{t}_j$ should have a statistical inclination towards values that are larger than $\pi/2$. Otherwise, a collapsed chain does not curl upon itself at a rate which fills the space faster than in the RW phase. Since the primary contribution to (7) derives from large values of $k = |i-j|$, in accordance with (27) we expect that in the collapsed phase and with small values of k

$$\mathcal{P}_\Gamma(N) \overset{N\gg 1}{\longrightarrow} \mathcal{P}_\Gamma^{\rm coll}(N^\sigma + f(k)N) < 0, \tag{28}$$

where the prefactor $\mathcal{P}_\Gamma^{\rm coll} < 0$. In particular, the exponent $\sigma > 1$. This is because the chain collapse is due to interactions that have a long distance along the chain, and the number of possible vertex pairs increases faster in N than the number of nearest neighbor vertex pairs. Note the small-k near-neighbor correction term that we have included in (28): As in (27) there should be such a term; it includes short-distance repulsion between those vertices that are *very* close to each other along the chain, e.g., nearest neighbors. The $f(k)$ is some function of the short-distance cutoff value k; in general it is model specific.

C. Renormalization group flow

When the number of vertices N is very large we may coarse grain the chain by repeating a Kadanoff block-spin transformation of the vectors that determine the segments, as shown in Fig. 1. This gives rise to a renormalization group (RG) evolution of $\mathcal{P}_\Gamma(N)$. Figure 2 shows the phase diagram that we expect to find in the case of a homopolymer, for very large values of N; we deduce the phase diagram from our preceding analysis of (7); see also [9,11]. For the moment we overlook the straight rod phase. As shown in the figure, we have found that in the SARW phase the strength of repulsive interaction between vertex pairs (second virial coefficient) evolves towards a nontrivial positive fixed point value [9,11]. Consequently, by repeated action of the block-spin transformation shown in Fig. 1 we expect the observable (9) to flow towards a fixed positive value, in the SARW phase.

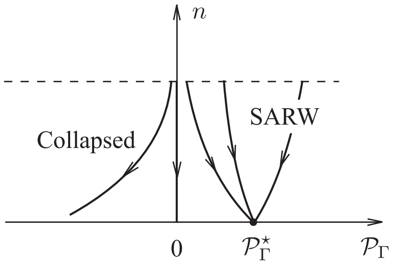


FIG. 2. Expected RG flow of the coefficient $\mathcal{P}_\Gamma$ in (7) in different phases, with $\mathcal{P}_\Gamma = 0$ the Gaussian fixed point and $\mathcal{P}_\Gamma^\star$ the fixed point of SARW.

We expect this value to be universal, quite independently of the underlying energy function (16). On the other hand, since (9) vanishes in the RW phase the ensuing RG evolution defines a vertical basin of attraction towards the Gaussian fixed point, as shown in Fig. 2. This flow separates the SARW phase from the collapsed phase, where the flow is towards a negative value of $\mathcal{P}_\Gamma$. The collapsed phase is commonly assumed to correspond to the space filling fixed point value (3) of the scaling exponent ν, in the case of a homopolymer.

However, we note that there are *numerous* examples of discrete space chains with geometrically nontrivial attractors. A generic, deterministic, and chaotic 3D flow approaches an attractor that can have *a priori* an arbitrary fractal Hausdorff dimension. This opens the possibility for a more complex phase structure also in the case of discrete chains that deserves to be addressed. We conclude that *at this point* we expect the correspondences between the phase of a chain, the sign of (9), and the numerical value of σ shown in Table I.

III. COARSE GRAINING CHAINS

The scaling transformation shown in Fig. 1 is a direct adaptation of Kadanoff's block-spin transformation. It decreases the number of segments at an exponential rate. Thus a chain becomes very rapidly coarse grained. For example, a typical protein backbone with a couple of hundred Cα atoms can support only a handful of block-spin transformations. This is hardly sufficient to define a smooth RG flow, not to mention the identification of distinct length scales that govern the chain properties at intermediate distance scales.

Our scaling procedure for chains

We proceed to develop a chain-specific variant of the block-spin transformation, one that can be iterated a large number of times, comparable to the number of vertices in the chain. With the help of our coarse graining transformation we then hope to

TABLE I. Phases in terms of our observable.

Phase	$\langle \mathcal{P}_\Gamma \rangle$	σ
Rod	>0	2
SARW	>0	≈3/2
RW	=0	
Collapsed	<0	>1

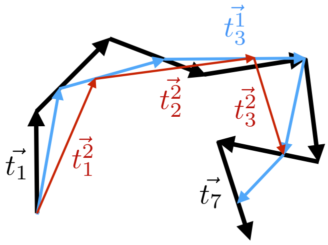


FIG. 3. Coarse graining procedure for scaling parameter $s = 4/3$. The initial chain is bold black (vectors $\mathbf{t}_1, \ldots, \mathbf{t}_7$), the first step of the coarse graining procedure is thin blue (light gray) arrows (vectors $\mathbf{t}_1^1, \ldots, \mathbf{t}_5^1$), and the second step is the thinnest red (dark gray) lines (vectors $\mathbf{t}_1^2, \ldots, \mathbf{t}_3^2$).

detect and identify the different length scales that characterize a given chain, even when there are only a relatively few vertices such as in the case of a generic protein backbone.

We start by introducing a scaling parameter s. We define it to be the number of old segments which are connected by the new one, during a coarse graining process. In the case of the conventional (Kadanoff) block-spin transformation s is always an integer. For example, in Fig. 1 we have $s = 2$. For $s = 3$ we connect every third vertex, while for $s = 1$ we simply repeat the original chain.

Canonically, in the case of a spin system the parameter s can only have integer values. But in the case of a chain, it turns out that we can promote s into an *a priori* arbitrary number and here we are particularly interested in values $s \in (1,2]$. For this we introduce a new coarse graining procedure, and in Fig. 3 we show how it proceeds when $s = 4/3$: We initiate the coarse graining with the vectors $\mathbf{t}_i$ that determine the segments of the chain, at the current iteration level. We then define the vector $\mathbf{t}_1^{\text{new}}$ which determines the first segment of the following iteration step by

$$\mathbf{t}_1^{\text{new}} = \mathbf{t}_1 + \tfrac{1}{3}\mathbf{t}_2.$$

To construct the second segment $\mathbf{t}_2^{\text{new}}$ in the chain of the following iteration step, we add the remaining two thirds of $\mathbf{t}_2$ together with two thirds of $\mathbf{t}_3$,

$$\mathbf{t}_2^{\text{new}} = \tfrac{2}{3}\mathbf{t}_2 + \tfrac{2}{3}\mathbf{t}_3.$$

Finally, for $\mathbf{t}_3^{\text{new}}$ we add one third of $\mathbf{t}_3$ and $\mathbf{t}_4$, so that

$$\mathbf{t}_3^{\text{new}} = \tfrac{1}{3}\mathbf{t}_3 + \mathbf{t}_4.$$

The third vertex of the following iteration step then coincides with the fourth vertex of the preceding iteration step. The process is repeated with $\mathbf{t}_5$ and so forth, until the entire chain becomes covered. Note that as shown in Fig. 3, the last vertex of the preceding chain is not necessarily reached by the last vector of the following chain. The figure shows this in the case when the preceding chain has seven vertices and we have chosen $s = 4/3$. By repeating the coarse graining, at the end of the second iteration (red line in the figure) we again miss part of the end in the preceding chain. This loss of structure at the end of the chain can be avoided by choosing the scaling parameter s_p at iteration step p so that

$$N_p = s_p q, \tag{29}$$

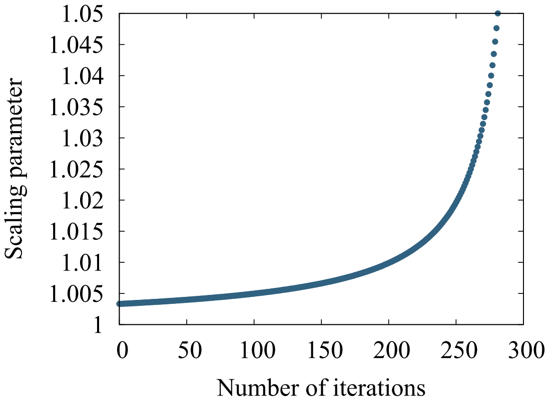


FIG. 4. Dependence of the optimal scaling parameter s_p^{opt} value on the number of iteration steps p, for a chain with 300 initial vertices.

where N_p is the number of vertices at the iteration level p and q is some integer. Thus, the smallest value we can choose for s_p is

$$s_p^{\text{opt}} = \frac{N_p}{N_p - 1} = 1 + \frac{1}{N_p - 1}. \tag{30}$$

Now the end points of the chain do not move, but the scaling parameter varies with the iteration step. However, this variation is quite small. As an example, for a chain with 300 vertices which is quite typical in the case of a protein backbone, we estimate that after ∼200 iteration steps the optimal value s_p^{opt} becomes changed by less than 0.7% as shown in Fig. 4.

Figure 5 shows the effect of coarse graining on the chain geometry. The effect is to suppress any abrupt short-wavelength oscillation in the geometry; those sections of the chain with many twists and turns become more regular as shown in the figure: A chain becomes visibly smoother while preserving its overall shape, as the coarse graining advances.

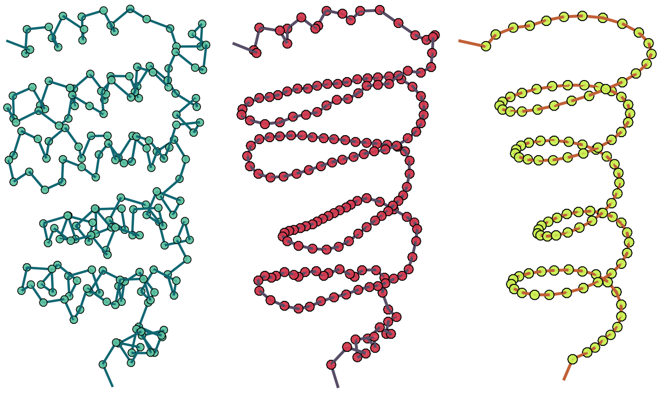


FIG. 5. Illustration of the effect of the coarse graining procedure for a helical chain; the example is from the Cα backbone of PDB structure 5DN7. On the left is the PDB structure, in the middle the structure after 10 coarse graining steps, and on the right after 80 coarse graining steps.

IV. HOMOPOLYMER MODEL

We shall employ (23) in combination with our coarse graining procedure to investigate the homopolymer phase structure numerically, using the universal energy function introduced in [26,27]. This energy function has been found to model the folding trajectories of several globular proteins in a realistic manner [28] with very few computational resources. In the sequel we shall present results from several thousand full folding and unfolding simulations, for homopolymer chains with up to 1000 residues and over extended temperature ranges. Using the present model, a full folding and unfolding simulation, even with very long chains, takes only a few minutes in a single processor in our approach. At the same time such simulations would not be even possible with more conventional coarse grained force fields designed for describing protein folding, using presently available (super)computers.

A. Frenet frames

To fully describe chain geometry, we need to introduce a framing. For this we consider four generic consecutive vertices $\mathbf{x}_{i-1},\mathbf{x}_i,\mathbf{x}_{i+1},\mathbf{x}_{i+2}$ along a piecewise linear stringlike chain. Let $\mathbf{t}_i,\mathbf{t}_{i+1},\mathbf{t}_{i+2}$ be the three segments that connect these four vertices. For each vertex we evaluate the ensuing bond (κ) and torsion (τ) angle as follows: The bond angle is obtained directly in terms of the segments,

$$\kappa_i \equiv \kappa_{i+1,i} = \arccos\left(\frac{\mathbf{t}_{i+1}\cdot\mathbf{t}_i}{|\mathbf{t}_{i+1}||\mathbf{t}_i|}\right). \qquad (31)$$

For the torsion angles we first introduce the normal vector

$$\mathbf{b}_i = \mathbf{t}_{i-1}\times\mathbf{t}_i \qquad (32)$$

of the $(\mathbf{x}_{i-2},\mathbf{x}_{i-1},\mathbf{x}_i)$ plane. The torsion angle τ_i is then

$$\tau_i \equiv \tau_{i+1,i} = \mathrm{sgn}[(\mathbf{b}_{i-1}\times\mathbf{b}_i)\cdot\mathbf{t}_i]\times\arccos(\mathbf{b}_{i+1}\cdot\mathbf{b}_i). \qquad (33)$$

Note that a torsion angle can be introduced whenever $\mathbf{t}_{i-1}$ and $\mathbf{t}_i$ are linearly independent. We also note that (κ_i,τ_i) yield spherical coordinates around $\mathbf{x}_i$, and that the three vectors $(\mathbf{t}_i,\ \mathbf{b}_i,\ \mathbf{n}_i = \mathbf{b}_i\times\mathbf{t}_i)$ constitute an orthogonal frame at this vertex.

Inversely, once the bond and torsion angles and in addition the segment lengths are known, we recover the chain as follows: From the angles we first compute the frames $(\mathbf{t}_i,\mathbf{b}_i,\mathbf{n}_i)$ using the discrete Frenet equation, as described in [25]. The entire chain is then given by the solution of the discrete Frenet equation

$$\mathbf{x}_i = \sum_{k=1}^{i}\mathbf{t}_k.$$

B. Landau free energy

We deduce the Landau free energy of a chain using a symmetry principle [26,27]: The energy function of a structureless, piecewise linear discrete chain should not depend on the way the chain is framed. Thus the energy function must remain intact under frame rotations around the segment vectors $\mathbf{t}_i$. Let $(\mathbf{e}_i^1,\mathbf{e}_i^2)$ denote two orthogonal unit vectors that relate to $(\mathbf{b}_i,\mathbf{n}_i)$ by a generic SO(2) rotation around $\mathbf{t}_i$. The orthogonal basis $(\mathbf{t}_i,\mathbf{e}_i^1,\mathbf{e}_i^2)$ could then be used instead of the Frenet basis $(\mathbf{t}_i,\mathbf{b}_i,\mathbf{n}_i)$, to construct the energy function. Mathematically, this determines a local SO(2) gauge structure.

The Cα backbone of a protein is akin our piecewise linear discrete chain, with an average ~3.8 Å distance between vertices; from this perspective, the only influence of side chains is to introduce a heterogeneous interaction between the Cα atoms. Therefore, the Cα backbone must employ an SO(2)-invariant energy function of the (virtual) backbone bond and torsion angles. The bond angles κ_i transform like a two-component SO(2) scalar field and the torsion angles τ_i transform like an SO(2) gauge field under a local frame rotation [26,27]. *Universality* then implies that the leading order Cα energy function for a protein backbone with N residues (vertices) must relate to the lattice Abelian Higgs model (AHM) Hamiltonian. This follows directly because the AHM Hamiltonian is the most general SO(2) gauge invariant Hamiltonian there is. In the unitary gauge the AHM Hamiltonian coincides with the following discrete nonlinear Schrödinger (DNLS) Hamiltonian with a spontaneously broken symmetry [25–27,29–32] (see [33] for a review with applications to proteins):

$$E(\kappa,\tau) = \sum_{i=1}^{N-1}(\kappa_{i+1}-\kappa_i)^2 + \sum_{i=1}^{N}\left\{\lambda\left(\kappa_i^2-m^2\right)^2 + \frac{d}{2}\kappa_i^2\tau_i^2 - b\kappa_i^2\tau_i - a\tau_i + \frac{c}{2}\tau_i^2\right\} + \sum_{i\neq j}U(\mathbf{x}_i-\mathbf{x}_j). \qquad (34)$$

Here (λ,m,a,b,c,d) are parameters; they are specific to a given amino acid sequence in the case of a protein. The terms in the first line coincide with a *naive* discretization of the continuum nonlinear Schrödinger equation. In the second line, the first term (b) is the conserved momentum in the DNLS model, the second (a) is the Chern-Simons term, and the third (c) is the Proca mass term; see [29–33] for detailed analysis and interpretation of the various terms in the context of proteins. We note that both momentum and Chern-Simons terms are chiral; with positive parameters a,b these two terms ensure that the backbone is right-handed chiral in line with protein structures.

Besides the terms that we have displayed explicitly in (34) there are also two-body interactions (16) that have a long range along the chain, and are governed by the last term in (34). These interactions include Pauli exclusion, electromagnetic, and van der Waals interactions between the various atoms. Here we consider a simple homogeneous variant of $U(\mathbf{x}_i-\mathbf{x}_j)$ that in addition of the hard sphere (Pauli) repulsion (18) has a spatially short range attractive component,

$$U(r) = \begin{cases} +\infty, & 0 < r < R_0, \\ U_0\{\tanh(r-R_0)-1\}, & R_0 < r < +\infty. \end{cases} \qquad (35)$$

For $r < R_0$ there is a hard-core repulsion but for $r > R_0$ there is a spatially short range attractive interaction with strength determined by the parameter U_0. In the case of proteins we choose $R_0 \sim \Delta = 3.8$ Å, which is the distance between two neighboring Cα atoms. We refer to [34] for a detailed analysis of the effects of long-range (along chain) interactions in (34).

1. Cooperativity and first-order phase transition

The free energy (34) can be validated by verifying its compatibility with Privalov's criterion [35–37]. It states that protein folding is a cooperative process which in the case of a short two-stage folding protein resembles a first-order phase transition.

For (34) cooperativity is due to solitons that are supported by the DNLS equation [38]; solitons are the paradigm cooperative organizers in numerous physical scenarios. Here a soliton emerges when we first eliminate the torsion angles using their equation of motion,

$$\tau_i[\kappa] = \frac{a + b\kappa_i^2}{c + d\kappa_i^2}. \tag{36}$$

For bond angles we then obtain

$$\kappa_{i+1} = 2\kappa_i - \kappa_{i-1} + \frac{dV[\kappa]}{d\kappa_i^2}, \tag{37}$$

where

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

The difference equation (37) can be solved iteratively using the algorithm developed in [39]. A soliton solution models a super-secondary protein structure such as a helix-loop-helix motif, with the loop corresponding to the soliton proper.

To identify the putative first-order transition character we observe that in the case of a protein, the bond angles are rigid and slowly varying while the torsion angles are highly flexible. Thus, over sufficiently large distance scales we may try to proceed self-consistently in a Born-Oppenheimer approximation, using a mean field $\kappa_i \sim \kappa$ and then solving for κ in terms of torsion angles $\tau_i \sim \tau$. From (34)

$$\frac{\delta E}{\delta \kappa} = 0 \Rightarrow \kappa^2 = m^2 + \frac{b}{2\lambda}\tau - \frac{d}{4\lambda}\tau^2. \tag{38}$$

In those cases that are of interest to us, this equation always has a solution: Both κ and τ are multivalued angular variables, and for proteins the parameters b and d are small in comparison with m^2 and λ. We substitute the solution into (34) which gives for the energy

$$-\frac{d^2}{16\lambda}\tau^4 + \frac{bd}{4\lambda}\tau^3 - \left(\frac{b^2}{\lambda} - 2dm^2 - 2c\right)\tau^2 + (a + bm^2)\tau. \tag{39}$$

We identify here the canonical form of the Landau–de Gennes free energy of a first-order phase transition [40], originally introduced in the context of liquid crystals. This completes our qualitative validation of (34) in line with Privalov's criterion [35–37], at the level of mean field theory.

We conclude with the following comment: Despite the suggestive analogy between (39) and the de Gennes free energy of a first-order transition, a chain collapse from the SARW phase to a space-filling phase proceeds through an intermediate that includes the random walk phase; see Fig. 2. The intermediate can be either a tricritical θ point [1–4,9] in which case we encounter the characteristics of a first-order phase transition in line with Privalov's criterion, or it can be an extended θ regime possibly with its own internal structure, possibly including molten globule folding intermediates [41,42]: An analysis at the level of a Landau-Ginsburg theory is suggestive, but not sufficient, in determining the character of a phase transition. Entropic corrections are important for chain collapse, and accounted for in the usual manner of Landau-Ginsburg-Wilson theory [23].

2. Algorithm details

In our numerical simulations we employ the heat bath algorithm that has been introduced in [34]. This algorithm has a very fast rate of convergence towards a thermal equilibrium state, in the case of a single chain. It updates the bond and torsion angles according to the following probability distribution (obeying the detailed balance condition):

$$P(\kappa_{\text{new}},\tau_{\text{new}})\exp\{-\beta E(\kappa_{\text{old}},\tau_{\text{old}})\}$$
$$= P(\kappa_{\text{old}},\tau_{\text{old}})\exp\{-\beta E(\kappa_{\text{new}},\tau_{\text{new}})\},$$

where E is the DNLS Hamiltonian (34) and the update is a "walk" through the entire chain with a provisional revision of each value (κ_i,τ_i). New values $(\kappa_i^{\text{new}},\tau_i^{\text{new}})$ are generated randomly, according to probability distributions

$$P(\kappa_i^{\text{new}}) = \frac{1}{Z_{i,\kappa}}\exp\left\{-\beta E_{i,\kappa}(\kappa_i^{\text{new}})\right\} \tag{40}$$

and

$$P(\tau_i^{\text{new}}) = \frac{1}{Z_{i,\tau}}\exp\left\{-\beta E_{i,\tau}(\tau_i^{\text{new}})\right\}, \tag{41}$$

where $E_{i,\kappa}$ and $E_{i,\tau}$ are the sum of all those terms in the Hamiltonian that contain κ_i and τ_i, respectively, with the given index i and the $Z_{i,\kappa}$ and $Z_{i,\tau}$ are normalization factors. Note that the updated values of κ_i^{new} and τ_i^{new} do not depend on the previous values of κ_i and τ_i.

Explicitly, the probability density for κ_i^{new} has the form

$$P(\kappa_i) \sim \exp\left\{-c_1\kappa_i^4 - c_2\kappa_i^2 - c_3\kappa_i\right\}, \tag{42}$$

where

$$c_1 = \beta\lambda,$$
$$c_2 = \beta\left(2 - 2\lambda m^2 + \frac{d}{4}\tau_i^2 - b\tau_i\right),$$
$$c_3 = \beta(-2(\kappa_{i+1} + \kappa_{i-1})). \tag{43}$$

Thus (42) is non-Gaussian. On the other hand, the probability density $P(\tau_i^{\text{new}})$ has the Gaussian profile

$$P(\tau_i) \sim \exp\left\{-\beta\left(\tfrac{1}{2}\left[d\kappa_i^2 + c\right]\tau_i^2 - a\tau_i\right)\right\}. \tag{44}$$

Note that the Monte Carlo temperature $T = \beta^{-1}$ is *not* equal to the physical temperature factor $k_B\theta$ where k_B is the Boltzmann constant and the temperature θ is measured in kelvins. In the low-temperature collapsed regime general renormalization group arguments [33] imply that the Monte Carlo temperature T is related to the real physical temperature in the following way:

$$\ln T = k_B\theta + \cdots. \tag{45}$$

Finally, the algorithm approaches the canonical Gibbs equilibrium distribution

$$\exp\{-\beta E(\kappa,\tau)\} \tag{46}$$

TABLE II. The parameter values in (34) during our simulations.

λ	m	a	b	c	d	U_0
3.5	1.5	10^{-4}	0	10^{-4}	10^{-4}	0.5

at exponential rate in the updates; see [34] for a detailed description.

3. Radius of gyration vs temperature and two stages of collapse

To scrutinize the details of chain collapse in the homopolymer model, we investigate the temperature dependence of the radius of gyration using numerical simulations. Our parameter values for (34) are shown in Table II where the numerical value of m corresponds to α-helical protein structures; for β-stranded chains we choose $m = 1$. The parameters that relate to torsion angles are relatively small, in comparison to those that relate to bond angles only. This is in line with proteins where bond angles are known to be quite rigid while torsion angles are often found to be highly flexible; see the analysis in connection with Eqs. (38) and (39).

Figure 6 shows how the value of radius of gyration (1) increases with increasing temperature factor, in the case of a homopolymer chain with $N = 300$ vertices. We find that at low Monte Carlo temperatures $\log_{10} T < 0$ the chain is in the collapsed phase, where the radius of gyration is temperature independent. When temperature is (roughly) in the range $0 < \log_{10} T < 0.5$ the chain is in the transient θ region where the radius of gyration rapidly increases as a function of the temperature; the RW phase is located in this θ region. For $\log_{10} T > 0.5$ the chain enters the SARW phase where the radius of gyration value eventually stabilizes into a temperature-independent value. The apparent two-state character with lack of structure in the θ regime is in line with the two-stage folding nature (Privalov's criterion) of the Landau free energy function, in the case of a homopolymer chain. We refer to [34] for additional details of the chiral homopolymer phase structure.

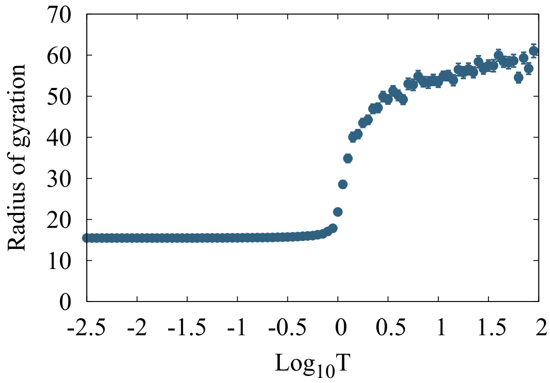


FIG. 6. The evolution of the radius of gyration as a function of temperature factor $T = 1/\beta$, in the homopolymer model with $N = 300$ vertices.

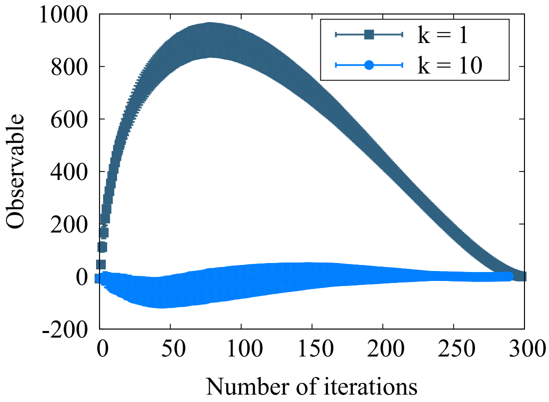


FIG. 7. The evolution of the observable (9) for $k = |i - j| \geqslant 1$ and $k = |i - j| \geqslant 10$ in (47). The figure shows the ensemble average of 128 simulations together with the one σ deviation distance from the average value, in each case; we note that as k increases the observable converges rapidly to vanishing value.

V. RANDOM CHAIN SIMULATIONS

From Fig. 6 we confirm that the RW phase appears in the phase diagram of the homopolymer model (34) in the θ regime, between the high-temperature SARW phase and the low-temperature collapsed phase [34]. The width of the θ regime relates to finite-size effects; in this regime the radius of gyration is very sensitive to temperature variations. Accordingly we find it delicate to try to describe a statistical ensemble of RW phase homopolymer chains with energy function (34), for the exact ideal value $\nu = 1/2$ of the scaling exponent in (2). Instead we proceed to simulate the RW phase directly. For this we promote $\kappa_i \in [0,2\pi)$ and $\tau_i \in [0,2\pi)$ into independent random variables. We fix the segment length to a constant value, e.g., 3.8 (Å). In particular, we ignore all the effects of the energy function (34) in the Gibbsian, including the short-distance Pauli repulsion.

Figure 7 shows the evolution of the observable (9) in the RW model, as a function of coarse graining steps and in the case of a chain with $N = 300$ initial vertices. The lateral axis depicts the progress of the iterative coarse graining procedure. In our simulations we use the value $s = s^{\text{opt}}$ that we determine from (30) for the scaling parameter. This enables us to iterate the coarse graining $n = 300$ times.

In the sequel we investigate the sensitivity of the observable to short-distance corrections (see discussion below and in Sec. II B 4) by modifying (6) using a short-distance cutoff k as follows:

$$\mathcal{P}_\Gamma(N) \to \mathcal{P}_\Gamma^k(N) = \sum_{i=1}^{N} \sum_{j=i+k}^{N} \langle \cos \kappa_{ij} \rangle. \tag{47}$$

Figure 7 shows the evolution of (9) both when we account for all values of the segment separation, i.e., when we have $k = |i - j| \geqslant 1$ in (47), and when we eliminate the short segment distance effects; i.e., we only account for pairs with $k = |i - j| \geqslant 10$ in (47).

In the case when we include all values $k \geqslant 1$ in Fig. 7, the observable initially vanishes in line with (15). When we proceed to coarse grain, the value of the observable starts rapidly increasing. It then decreases, approaching a vanishing value towards the end of the coarse graining process. The intermediate increase in the value of the observable can be understood as follows: Consider the blue segments (arrows) in Fig. 3 that show the outcome of the first coarse graining step. The first blue segment connects the first vertex of the initial chain to the second (black) segment of the initial chain. The second blue segment then connects the second segment of the initial (black) chain to the second vertex of the coarse grained chain, located on the third segment of the initial chain. The fact that both coarse grained segments engage the same (second) segment of the initial chain introduces a correlation between the (blue) coarse grained segments; the nearest neighbor segments along the coarse grained chain are not mutually fully independent. This interdependence, caused by the coarse graining process, implies that the observable (6) does not vanish during the flow. Instead, after initially increasing, the observable decreases towards a vanishing value when the number of coarse graining steps becomes very large. At the end of the flow, when there are only three vertices and two connecting segments left, the observable vanishes: The angle between the two final segments is randomly distributed.

Figure 7 shows also the evolution of the observable, once we remove the contribution of the first 10 nearest neighbor pairs, those with $k = |i - j| \leqslant 10$. This removal of short-distance correlations, caused by the coarse graining procedure, yields an observable that is in line with the RW phase behavior, one that vanishes with one standard deviation precision. Thus the result shown in Fig. 7 with $k = |i - j| \leqslant 10$ suggests that the correlations introduced by the coarse graining procedure have a short range, in terms of segments along the chain.

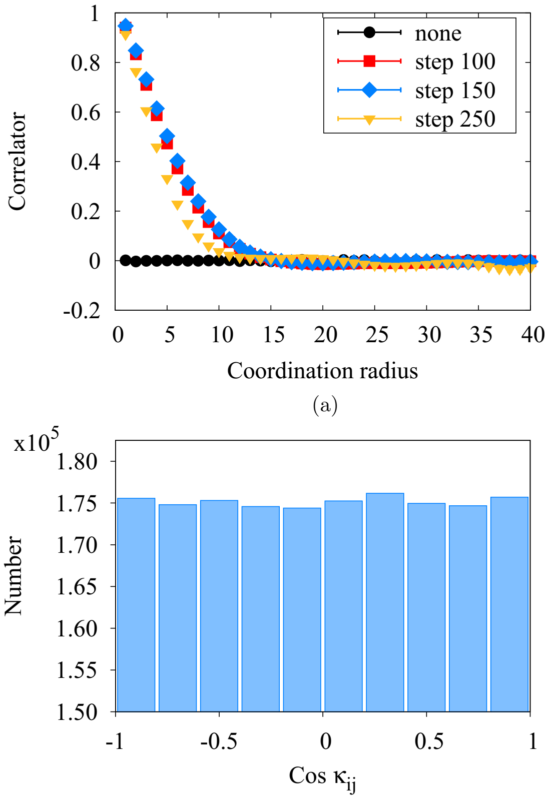
In Fig. 8(a) we show how the correlation length of the cosines in (6)

$$G(k) = \langle \cos \kappa_{ij} \rangle \tag{48}$$

depends on the segment distance $k = |i - j|$ and on the number n of coarse graining steps. We find that quite independently of the number of coarse graining steps, the quantity (48) decays at an (apparently) exponential rate in k, so that after around $k \sim 10$ these correlations are vanishingly small; in Fig. 8(a) we display the correlation length of (48) after $n = 100$, 150, and 250 coarse graining steps.

We conclude that in the RW phase there are finite-size effects due to short-distance correlations between the coarse grained segments. But these correlations have a short range and become vanishingly small beyond $k = |i - j| \sim 10$, in the RW phase. The results shown in Fig. 8(b) confirm this: In this figure we display the distribution of $\cos \kappa_{ij}$ for $k = 10$ and after $n = 150$ coarse graining steps.

We note that the histogram in Fig. 8 is in line with what we can expect in the RW phase: In RW phase, since the value of the observable vanishes, we expect the distribution of $\cos \kappa_{ij}$ to be symmetrical with respect to $\cos \kappa = 0$. Our simulations confirm that the histogram tends to a uniform flat distribution for $k = |i - j| \gg 1$, as expected.



(b)

FIG. 8. The RW phase. (a) The dependence of the RW phase correlator (48) on the coordination radius $k = |i - j|$ for different numbers of coarse graining steps. (b) RW phase histogram for the values of $\cos \kappa_{ij}$ after $n = 150$ iteration steps, $k = 10$ nearest neighbor subtraction.

VI. HOMOPOLYMER SIMULATIONS

We proceed to investigate (9) in combination with our coarse graining, in the SARW and collapsed phases of the homopolymer model (34). We use the parameter values shown in Table II. In our simulations we employ the heat bath algorithm that has been detailed in (40)–(44); see [34] for more details. We study chains with $N = 300$, $N = 700$, and $N = 1000$ initial segments. We control the thermodynamical phase by adjusting the ambient temperature in the heat bath algorithm [34]. We coarse grain the chains using the optimal scaling parameter (30). The number of vertices then decreases slowly, and the number of coarse grain iterations supported by the chains becomes comparable to the number of initial vertices.

A. Scaling effects on radius of gyration

We first analyze how the radius of gyration (1) evolves under coarse graining, in the SARW and collapsed phases. Figure 9 shows the result for a chain with $N = 700$ initial segments. The stability of the radius of gyration during coarse graining proposes that the chain preserves its overall geometry as the coarse graining proceeds. Note that for a renormalization group flow which builds on our coarse graining procedure, the radius

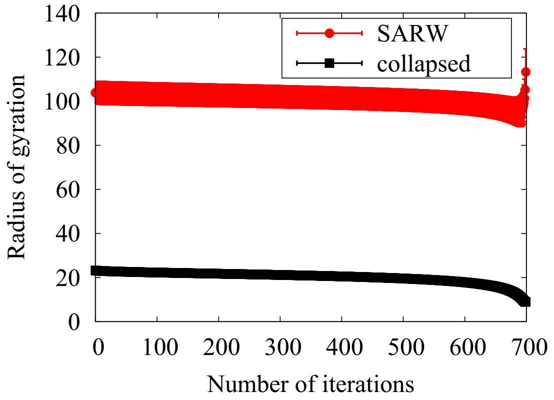


FIG. 9. Variation of radius of gyration vs number of coarse graining steps, for two different phases. The initial chain has 700 vertices; the error bars are for one standard deviation.

of gyration would appear to be akin to a renormalization group invariant quantity.

Figure 10 shows how the effective segment (Kuhn) length varies during the coarse graining process for a chain in the SARW phase, with an initial segment length of 3.8 (Å) and $N = 700$ initial segments. We observe that, with the parameter values in Table II, initially the effective segment decreases and reaches a a minimum value ∼1.9 (Å) after around 200 coarse graining iterations. Subsequently the effective segment length increases, and eventually it becomes comparable to the radius of gyration of the initial chain when the coarse graining terminates. This can be understood so that initially, the effect of coarse graining is to suppress any abrupt short-wavelength oscillation in the geometry; those sections of the chain with many twists and turns become more regular, in line with Fig. 5. This leads to an initial decrease in the segment length. Eventually, when the coarse graining progresses, since $s > 1$, the effective chain length then starts increasing.

B. The observable

We proceed to investigate how the statistical ensemble average (9) evolves during repeated coarse graining, in the SARW and collapsed phase of the homopolymer model.

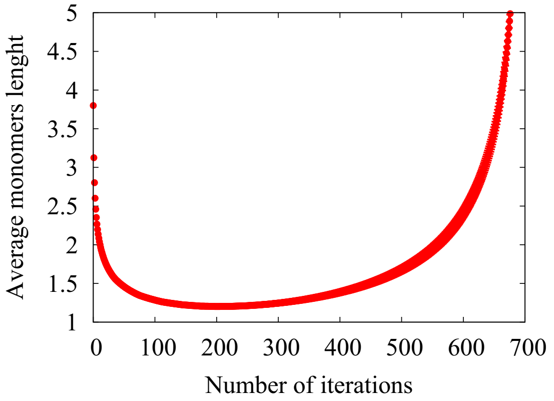


FIG. 10. Dependence of the average segment length on scaling step n for $N = 700$ initial segments in the SARW phase.

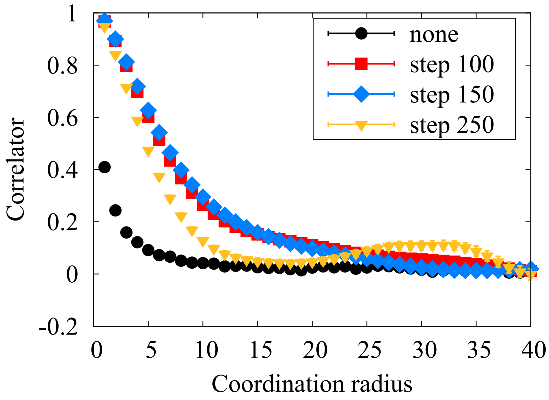


FIG. 11. The dependence of the correlator (48) on the coordination radius $k = |i - j|$ for different numbers of coarse graining steps in the SARW phase of a homopolymer. Note that the small bump in step-250 curve is due to finite-size effects.

1. *Homopolymer in the SARW phase*

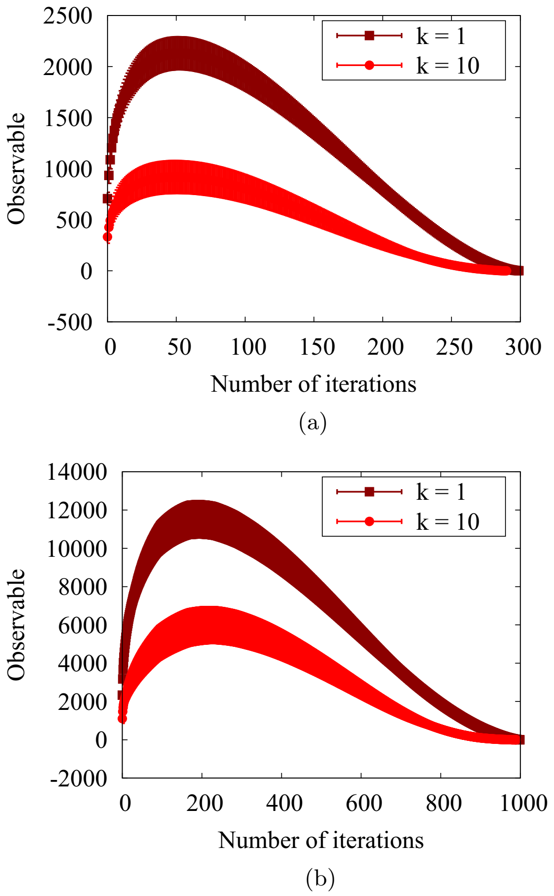
We evaluate the statistical average of the observable (9) using the homopolymer in the SARW phase, with chains that have $N = 300$ and $N = 1000$ initial vertices.

We recall that for a RW chain the correlations between neighboring vertices vanish; see for example (15) and Fig. 8(a). But we have pointed out that in RW phase, coarse graining introduces correlations between neighboring vertices. In Fig. 8(a) we estimate that these correlations have a finite extent in the RW phase; they appear to be effectively vanishing when vertices are a distance of $k \sim 10$ segments apart.

In the SARW phase the correlation length can be expected to be longer; there are native correlations between vertices along the entire chain such as Pauli repulsion that ensure self-avoidance and act between any pair of vertices. We estimate to what extent the *additional* correlations that are introduced by the coarse graining process interfere with the correlations that are native to the SARW phase.

Figure 11 shows our simulations results for the correlation length (48) in the SARW phase homopolymer model, using various levels of coarse graining. We observe that in line with the RW, in the SARW phase the coarse graining introduces short-range correlations between vertices. But these correlations, together with the effect of Pauli repulsion, seem to be observable only up to distances that are $k = |i - j| \sim 20$ segments apart from each other along the chain, in our model. Moreover, already after $k \sim 10$ the influence becomes quite small.

In Fig. 12 we show simulation results for the flow of observable (47) under the coarse graining, in the SARW phase. In this figure we can compare the case $k = 1$ where we sum over all pairs in the observable (47), with the case $k = 10$ where we only consider the contribution from those pairs where the vertices are a minimum segment distance $k = |i - j| \geqslant 10$ apart from each other along the chain. We observe that overall, the profiles in the figure display self-similarity in their shape. The same conclusion persists for larger values of k: For a homopolymer in the SARW phase, the only visible finite-size effect on the observable seems to be that the height of the curve becomes lower as k increases. In particular, each of the



(a)

(b)

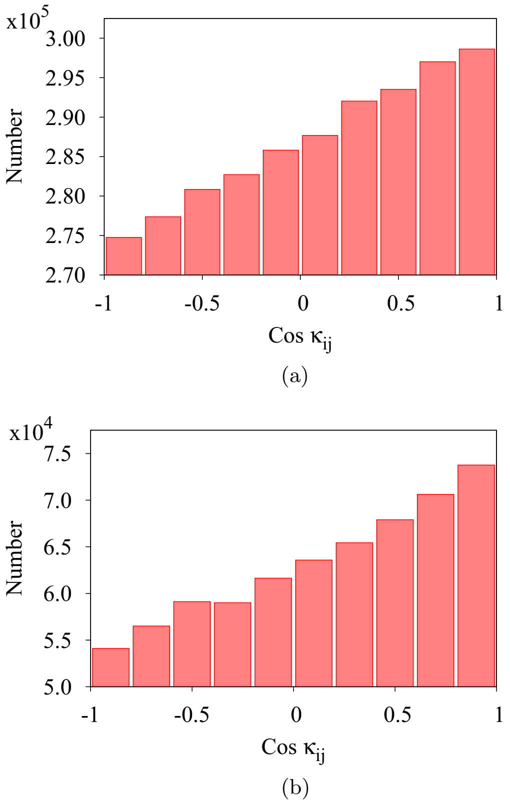
FIG. 12. Dependence of the observable (47) in the SARW phase on the number of scaling steps n for chains with various number of initial vertices, and with different values of subtraction k. The average value is shown, together with the one standard deviation fluctuation regime. (a) $N = 300$ initial vertices. (b) $N = 1000$ initial vertices.

curves in Fig. 12 has initially a positive value; they display convergence towards vanishing values as the coarse graining proceeds, in a self-similar manner.

The qualitative behavior shown in Fig. 12 is a characteristic of the SARW phase. In particular, the value of the observable is positive throughout, in line with Table I.

In Fig. 13 we show the histograms for our statistical ensemble of the $\cos\kappa_{ij}$ of (6) for $N = 300$ initial vertices, in the SARW phase. Figure 13(a) shows the initial SARW distribution of the cosines in (6) with no subtraction for the nearest neighbors, and Fig. 13(b) shows the SARW distribution we obtain after we repeat the coarse graining 150 times and in addition introduce the nearest neighbor subtraction (27) with $k = 10$ segments along the chain. The figure confirms the self-similarity that we already observed in Fig. 12: In the SARW phase, the histogram profile is stable under the coarse graining flow.

Finally, we inquire whether a relation akin to (27) can be introduced, to model how our observable depends on the number of vertices. Instead of considering an ensemble of chains with an increasing number of vertices, which can be very CPU-time consuming, we proceed as follows. We have found that in the SARW phase the observable (47) displays self-similarity, under coarse graining. Thus, we consider a statistical pool of chains with $N = 300$ vertices and inquire how a relation such as (27) can describe the flow of the observable during coarse graining: A large number of coarse graining iterations yields a chain with a small number of vertices. The relevant question to address is then how a relation like (27) models the coarse grained observable (47) when the number of coarse graining iterations increases. For this, let r denote the number of vertices in the coarse grained chain. A small value of r corresponds to a large number of coarse graining iterations, and when r becomes large the number of coarse graining iterations becomes small. The relation (27) instructs us to inquire how the ensuing observable $\mathcal{P}_\Gamma(r)$ depends on r as its value increases. For this we use an ansatz of the form



(a)

(b)

FIG. 13. Histograms for the values of $\cos\kappa_{ij}$ in the SARW phase for $N = 300$ initial vertices. (a) The initial distribution. (b) The distribution obtained after 150 coarse graining steps and with $k = 10$ subtraction.

$$\mathcal{P}_\Gamma(r) \approx a r^b + c r. \tag{49}$$

We use the pool of 1000 chains used in Fig. 12(b) to get the result shown in Fig. 14. We find that

$$\begin{aligned} a &= 0.24 \pm 0.03, \\ b &= 1.61 \pm 0.02, \\ c &= -1.45 \pm 0.12. \end{aligned} \tag{50}$$

Here we use the Levenberg-Marquardt nonlinear least-squares algorithm for fitting, with one-sigma (standard deviation) errors. Note the difference between Figs. 12 and 14. The former

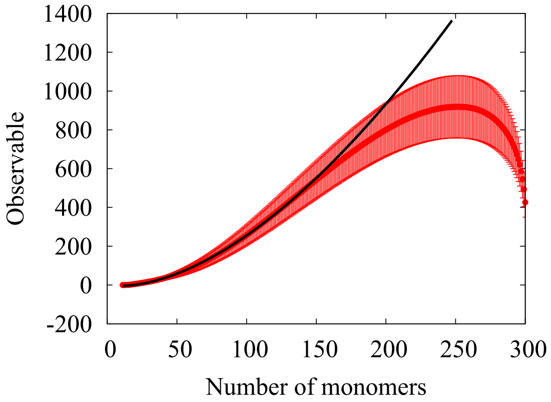


FIG. 14. A fit of the form (49) to the observable in the SARW phase with $k = 10$ subtraction.

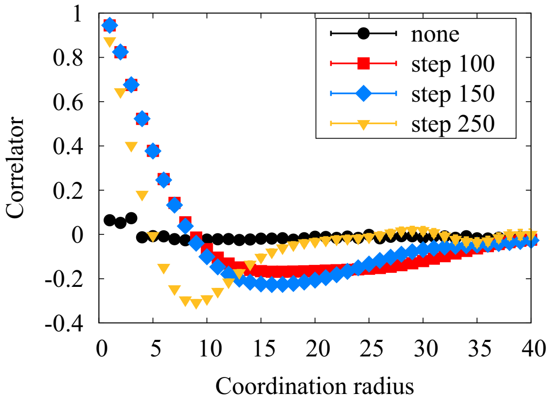


FIG. 15. The dependence of the correlator (48) on the coordination radius $k = |i - j|$ for different numbers of coarse graining steps in the collapsed phase of a homopolymer.

displays the observable when the number of coarse graining steps increases. In the latter the observable is displayed in terms of increasing number of vertices during the coarse graining flow.

We make the following comment: We have arrived at the value $\sigma = 3/2$ in Table I by assuming a chain with a very large number of vertices N, and the value $\sigma = 3/2$ is very close to the value of b we deduce in (49), (50). The relation (27) is derived using the perturbation theory in the vicinity of the RW phase, and our coarse grained chain reproduced this regime, with a large number of iterations. This is because when the number of iterations increases the ensuing segment length also increases. Thus the influence of the self-avoiding condition gradually disappears, with the observable approaching a vanishing value of the RW phase: When the number of iterations grows the perturbation theory works increasingly well. We conclude that our coarse graining method is an efficient way to describe properties of chains with varying lengths, in terms of a pool of fixed length chains.

2. *The collapsed phase*

In the case of RW we have investigated a statistical pool of chains that do not depend on the details of the homopolymer model. In the SARW phase we have used the homopolymer energy function (34). However, as in the case of the RW phase we expect the results to be universal: The SARW phase describes the high-temperature limit of the homopolymer model. In this limit the details of the energy function become irrelevant, as in this limit the temperature factor β in the Gibbsian (10) vanishes; only the hard-core $r < R_0$ Pauli repulsion of (35) survives, and the details of the repulsive interaction become increasingly irrelevant.

The situation is very different in the collapsed phase that occurs at low temperatures in the homopolymer model. Now the temperature factor β becomes large, and the thermodynamics becomes increasingly ruled by the energy function: Unlike in the case of RW and SARW phases where universality is due to the apparent insensitivity of the phase on the details of the ensuing chain Hamiltonian, in the collapsed phase the model-specific details matter most. Indeed, despite the asserted universality of (3) we are not aware of any compelling argument why the low temperature phase properties should be insensitive to dynamical details. Quite to the contrary: Discrete flows towards fractal attractors of all kinds are abundant in three dimensions. Accordingly, we scrutinize the collapsed phase of the homopolymer model (34), with the parameter values in Table II and using the heat bath method described in [34]. Figure 15 shows the correlation length (48) in the collapsed phase, for different values of the coarse graining steps.

As in the SARW phase, there is a hard-core repulsion between all vertex pairs. There are also interactions that are due to the dynamical details of (34), including solitons that are absent in the high-temperature SARW phase. Finally, we have the correlation between vertices due to the coarse graining process. But in line with the RW and SARW phases, we find that the effect of coarse graining extends only over a relatively short range in the segment distance $k = |i - j|$: From Fig. 15 we deduce that the effects of coarse graining are largely unobservable when k is greater than $k \sim 35$, a somewhat longer segment distance than what we found in the RW and SARW phases.

In Fig. 16 we show the evolution of the observable (47) during the coarse graining, when we increase the number $k = |i - j|$ of finite-size subtractions (28); the initial chain has $N = 700$ vertices.

When there is no subtraction, i.e., $k = 1$, we find that the observable is initially negative, in line with our general arguments in Table I. Then, after around 200 coarse graining steps the observable vanishes; the chain appears to reside in the RW phase. However, this is an apparent short-range effect, due to correlations between nearest neighbor vertices with $k = 1$: When we subtract the nearest neighbor contribution, i.e., we set $k = 2$ in (28), the value of the observable is negative throughout the coarse scaling process, in line with the general arguments of Table I.

In Fig. 16 we show the result also with $k = 10$ and with $k = 30$. Comparison of the profiles proposes self-similarity, in line with what we observed previously in the SARW phase in Fig. 12. Note that for a chain with $N = 700$ vertices, there can be additional finite length effects when k becomes much larger. In Fig. 17 we show three representative histograms of $\cos \kappa_{ij}$ in (6), in the collapsed phase. Figure 17(a) is for the initial chain. Here, we observe a clear accumulation of values

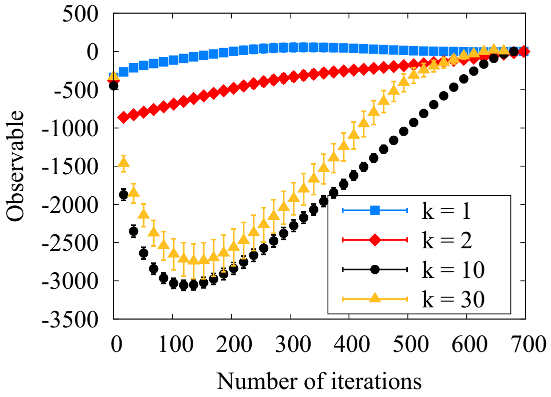


FIG. 16. Dependence of the observable (47) on the number of scaling steps n for chains with $N = 700$ initial vertices, and with different values of subtraction $k = |i - j|$ in the collapsed phase. The average value is shown, together with the one standard deviation fluctuation regime.

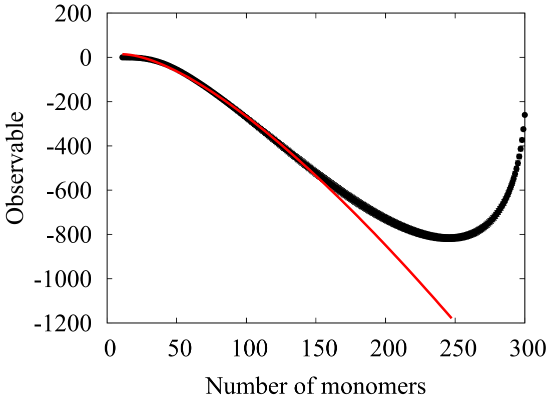


FIG. 18. A fit of the form (49) to the observable in the collapsed phase with $k = 10$ subtraction.

between $0 < \cos\kappa_{ij} < 0.2$. This reflects the effect of local minima for the κ angle in the Hamiltonian (34). Since these minima are located at $\kappa = m = 1.5 \approx \pi/2$, the peaks appear due to values of $\cos\kappa_{ij}$ for $|i - j| = 1$.

In Fig. 17(b) we remove all nearest neighbor contributions with segment distances $k = |i - j| < 10$. Now, there is a clear excess of negative values. Finally, in Fig. 17(c) we introduce $n = 150$ coarse graining steps in the histograms of Fig. 17(b). Now we obtain a monotonic, decreasing distribution of the $\cos\kappa_{ij}$ values. Note that the monotonic character of the distribution is quite in line with that in Fig. 13, except for the sign.

Finally, in analogy with Fig. 14 of the SARW phase, in Fig. 18 we introduce a fitting of the form (49) to the evolution of the collapsed state observable, during the coarse graining process.

We find

$$a = 6.22 \pm 2.05,$$
$$b = 1.16 \pm 0.03,$$
$$c = -10.5 \pm 2.4.$$

C. Summary of homopolymer simulations

Our results show that in the case of a homopolymer, the observable (47) flows in a self-similar manner during repeated coarse graining.

(1) In the SARW phase the observable is positive during the entire coarse graining process, with a self-similar profile akin the one shown in Fig. 12. The histogram profiles shown in Fig. 13 are also qualitatively universal, for chains in this phase.

(2) In the RW phase the observable initially vanishes, in line with (15). The coarse graining introduces correlations between neighboring segments causing the observable to have a small positive value. The observable then flows asymptotically towards a vanishing value, as we proceed and iterate the coarse graining. The qualitative features of the flow are universal with a self-similar (for $k > 1$) profile akin the one shown in Fig. 7, together with an evenly and uniformly distributed histogram as shown in Fig. 8(b).

(3) We have simulated the observable in the collapsed phase of the homopolymer model (34). Unlike in the case of universal RW and SARW phases, we expect that the results are in general model dependent. The observable is negative and—in the case of a homopolymer—its value first decreases but then starts to increase towards a vanishing value as the coarse graining proceeds. Once we remove the effect of very short distance repulsion between neighboring segments, the profile of the flow becomes self-similar as shown in Fig. 16; the histogram in Fig. 17(c) is also self-similar over a wide range of chains.

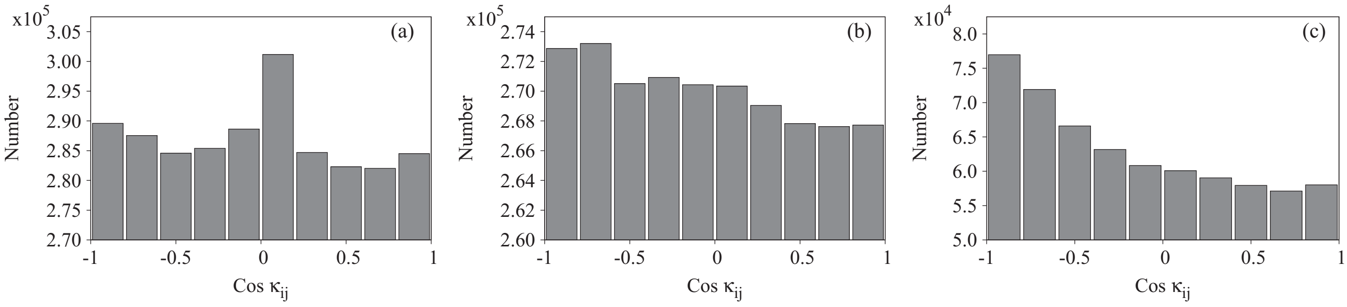


FIG. 17. Histograms for the values of $\cos\kappa_{ij}$ for a chain with $N = 300$ initial vertices in the collapsed phase. (a) The full initial chain. (b) With $k = 10$ nearest neighbor subtractions. (c) After $n = 150$ coarse graining steps and with $k = 10$ nearest neighbor subtractions.

We note that in all cases, the observable converges towards a vanishing value when the number of coarse graining iterations becomes large. This can be understood as follows: When the coarse graining terminates, we are left with only three vertices and two segments. In a statistical ensemble of long chains, the angle between these two remaining segments is randomly distributed with a vanishing average value.

VII. APPLICATIONS TO COLLAPSED PROTEINS

We proceed to investigate the coarse graining flow of the observable (47) with different values of the cutoff distance k, in the case of heteropolymers. For this we analyze the discrete chains that model the Cα backbones of crystallographic PDB proteins [15]. Our analysis is not comprehensive but indicative: We limit to a presentation of certain major features that we commonly observe among PDB proteins, using three representative examples.

In particular, we inquire how the structural heterogeneity of a protein becomes reflected in the flow when we successively increase the value of k: Over a short distance scale, which is sensitive to a small value of k in (47), the makeup of a protein is typically dominated by rodlike segments such as α helices and β strands. Nevertheless, the global protein structure commonly resides in a space filling, collapsed phase. Thus we explore how a transition from the short-distance regime of straight and regular rodlike segments to the large-scale regime of space filling collapsed structures can be detected and described by the observable (47) and its coarse graining flow. As order parameters to expose a transition, we employ the following characteristics of the observable:

(1) The initial value of the observable (47), in particular the way in which the initial value of the observable varies as a function of the cutoff distance k.

(2) The initial stages of the coarse graining flow, in particular the way in which the value of the observable starts to evolve when we initiate the coarse graining process.

(3) The self-similarity of the entire flow pattern of the observable during the entire coarse graining process.

Additional quantitative order parameters could be introduced in a more refined search of universal patterns in the observable during its coarse graining evolution. For example, the exponent b in a fit akin to (49) can be estimated during various stages of the flow. Such additional universal aspects of protein structure will be analyzed in a future publication.

Specifically, we have learned the following from the homopolymer analysis:

(1) In the SARW phase the observable (47) is initially *positive*. Moreover, its numerical value *increases* during the initial steps of the coarse graining flow. In addition, in a stable SARW phase the overall profile of the flow should be self-similar and resemble the profiles in Fig. 12, at least for the initial coarse graining steps, as the value of k is increased.

Note that it is conceivable that the value of the observable starts decreasing during the initial stages of the flow. This is an indication that some kind of phase change is in progress.

(2) The θ regime is tricritical and as such it is sensitive to perturbations. Thus we expect that in the RW phase the observable initially either vanishes or becomes vanishingly small. As the flow progresses and the value of k increases, the observable should then either continue to vanish or fluctuate between small positive and negative values in a self-similar fashion.

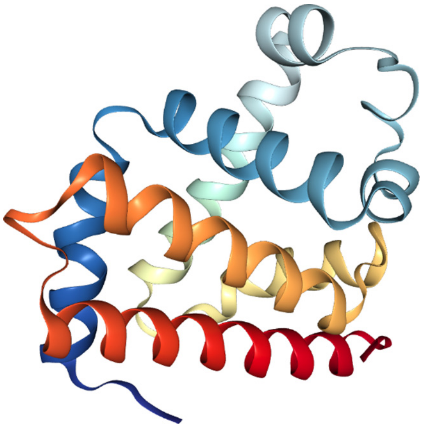


FIG. 19. Cα backbone of PDB structure 1ABS prepared with WebGL [15].

(3) In a collapsed phase the observable (47) starts with a *negative* value. As shown in Fig. 16, in the case of a collapsed homopolymer the value then initially *decreases* during the flow (when $k > 1$, i.e., as the short-distance steric repulsion becomes negligible). Again, we expect the profile of the flow to display self-similarity throughout a given phase, when the value of k is increased.

It is conceivable that the value of the observable initially starts and then continues increasing. In that case the chain might reside in a collapsed phase which can be different from the collapsed phase of a homopolymer.

When a transition between two phases takes place as the short-distance cutoff scale k increases, we expect to detect the presence of the transition as a qualitative change in one or several of these order parameters: There could be a qualitative alteration such as a change in sign in the initial value of (47). There could be a reversal in the initial coarse graining flow pattern; e.g., an initially decreasing flow of (47) turns into an increasing one. There could also be a major modification in the overall self-similarity pattern of the global coarse grained flow profile.

We now proceed to consider three representative examples of protein structures, to exemplify how phase transitions can be observed in terms of our order parameters.

A. α-helical myoglobin

The first example that we consider is myoglobin. It is a relatively short and widely studied example of α-helical proteins. We use the crystallographic structure with PDB code 1ABS shown in Fig. 19. There are 154 amino acids and there are eight α-helical structures that have an average length of around 14 amino acids. A total of 74% of amino acids reside in the helical structures, according to DSSP classification [15,43].

We monitor both the initial value of the observable (47) and the way in which its value starts evolving during the initial stages of the coarse graining flow, when we increase the value of the cutoff k. Figure 20 shows six representative examples of the coarse graining flow profiles that we encounter; the

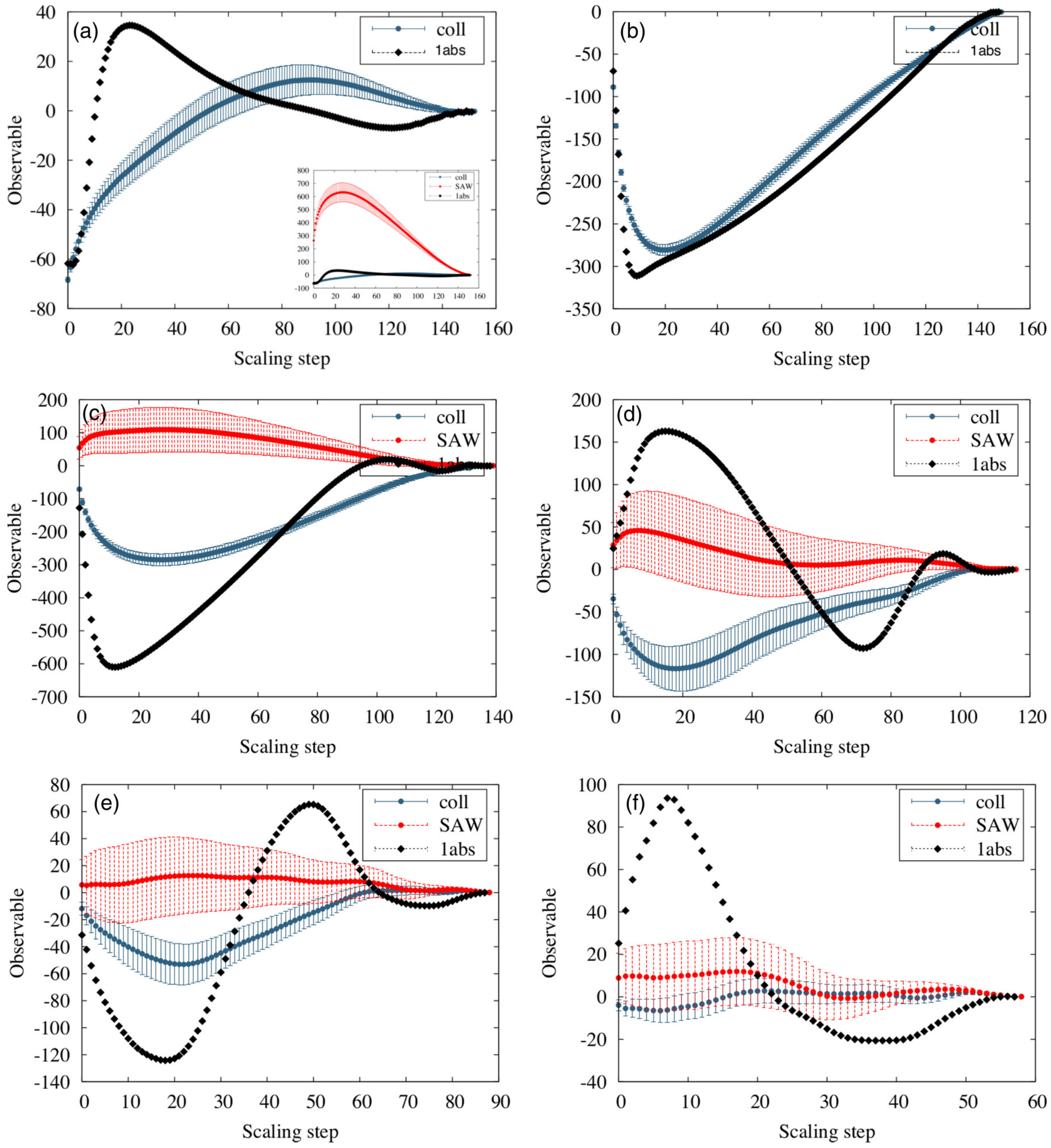


FIG. 20. Coarse graining flow of the observable (47), for myoglobin 1ABS. In panel (a) for the entire chain $k = 1$. In panel (b) with $k = 4$ subtraction, together with the collapsed phase distribution in our homopolymer model. In panel (c) with $k = 14$, in (d) with $k = 37$, in panel (e) with $k = 65$, and in panel (f) with $k = 95$ subtractions, together with both collapsed phase and SARW phase homopolymer distributions. The panels for all values of k are available as Supplemental Material [44].

entire flow can be found in the Supplemental Material [44]. In Figs. 20(a) and 20(b) we compare the myoglobin flow profile with the flow of collapsed phase homopolymers. The statistical homopolymer distribution is evaluated using a pool of 40 chains that we have constructed using our homopolymer model with 154 vertices. The statistical distributions show both the average value and the one standard deviation distance from the average, in the pool of homopolymer chains. In the inset of Fig. 20(a) and in the remaining panels we show both the collapsed phase and the SARW phase statistical distributions. We conclude that there are five major phases during the flow, exemplified by our choices of representative profiles in Fig. 20. We proceed to analyze the different phases in detail.

1. *Collapsed phase with k below 32*

Figures 20(a)–20(c) with $k = 1,4,$ and 14 show how initially and when the number of iterations remains small the coarse graining flow retains the negative value of the observable. There is a qualitative resemblance to the profile that we find in the collapsed phase of the homopolymer model. Moreover, both in the case of myoglobin and in the case of the homopolymer background, initially for $k = 1$ the observable is negative and then increases when the flow starts; this is due to short-range steric repulsion, between neighboring Cα atoms. But when $k \geqslant 2$ both flows become decreasing, initially. We note that while the initial $k = 1$ profile in Fig. 20(a) deviates visibly from the homopolymer distribution and resembles the RW distribution (see the inset), after only a few steps in k the two flows are very similar as shown in the $k = 4$ Fig. 20(b).

Note that the value $k = 14$ in Fig. 20(c) is equal to the average length of the helical segments in myoglobin. Nevertheless, the ensuing profile is clearly in a collapsed phase. It is even more deeply in the collapsed phase than the homopolymer distribution, as shown in Fig. 20(c). When we inspect the flow profiles with nearby values of k, we observe that there is an apparent self-similarity when k stays within the range between $k = 8$ and $k = 25$. Thus for the ensuing range of distance scales in k, the myoglobin structure displays the stable self-similar characteristics of a collapsed phase.

Starting after around $k = 25$ we observe a transition from the collapsed phase towards an apparent SARW phase in myoglobin. The flow profile begins to convert towards the $k = 37$ profile shown in Fig. 20(d).

2. SARW phase for values of k above 32 but below 57

For this range of cutoff values k the structure is firmly in the SARW phase: The initial value of the observable (47) is positive and increases during the initial stages of the coarse graining flow, as expected in the SARW phase. The flow profile is solidly positive valued and qualitatively akin to the homopolymer SARW profile when the number of coarse graining steps remains below ~50 in the case of $k = 37$, as shown in Fig. 20(d). In fact, as shown in this panel when the number of coarse graining steps is ~20 the flow profile is clearly above the SARW profile; apparently the structure approaches the straight rod phase, for these values of coarse graining. When the coarse graining proceeds beyond ~50 steps the ensuing values of the observable turns negative, i.e., apparently enters the collapsed phase: For a large number of coarse graining steps the flow starts to probe the phases that will eventually dominate the structure at much larger values of k. We also note that there is apparent self-similarity over several k values suggesting that the SARW phase is quite stable.

3. Collapse phase for values of k above 57 but below 83

In this range of k the initial value and the evolution of the observable during the early coarse graining steps has the characteristics of a collapsed phase, as shown in the representative Fig. 20(e). In the panel we observe that the coarse grained observable turns positive valued after around ~38 coarse graining steps; the flow starts to probe the SARW phase that will dominate the structure at values of k above $k = 83$.

4. SARW and eventual RW phases for values of k above 84

In the range from $k = 84$ to around $k = 109$ the profile first returns to the SARW phase as shown in Fig. 20(f) for $k = 95$. It then starts slowly evolving towards the asymptotic RW phase when k exceeds the value ~110, in line with our general arguments.

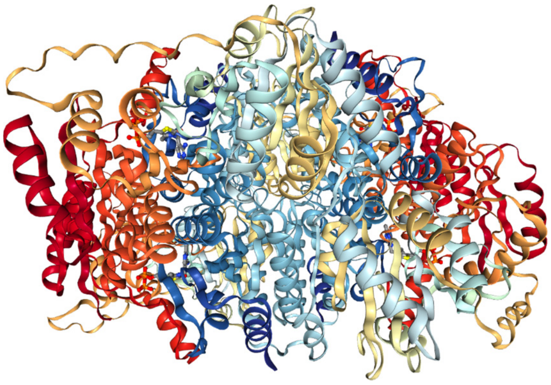


FIG. 21. Cα backbone of PDB structure 1Q6Z prepared with WebGL [15].

5. 1ABS summary

In summary, we find that in the case of the relatively short α-helical myoglobin the analysis of the coarse graining flow displays four different phase regimes: There is an initial collapsed phase which is followed by a SARW phase. Then we have again a collapsed phase, followed by a final SARW phase that slowly evolves towards the terminal RW phase. Remarkably, we do not observe any dramatic transition when the value of k is in the vicinity of the average length of helical structures; at these values of short-distance cutoff scale k the profile is self-similar and resides deeply in the collapsed phase.

We note that in experiments, the myoglobin folding proceeds in two stages from the high-temperature random coil to the low-temperature native collapsed state [45]: The first to fold as the temperature decreases are helices B,C,D, and E. Then, the helices A,G, and H fold. The structure then becomes a molten globule, which is followed by a slow stabilization of the remaining F helix when temperature further decreases. It would be interesting to see whether the length scales that are associated with our two collapsed k regimes somehow relate to the two temperature scales where first the B,C,D,E and then the A,G,H helices are formed.

B. An example of α-β proteins

The second example we choose is *a priori* much more complex: We consider the 528 residue protein with PDB code 1Q6Z shown in Fig. 21. There are 27 helices with around 39% of residues in these helical structures, and 25 strands that contain around 17% of residues according to DSSP classification [15,43]. Thus, the total number of residues located in the regular rodlike structures (~56%) is clearly smaller than in the case of myoglobin (~74%). Moreover, the helices have an average length of around 8 residues (there are ~27 helices in ~208 residues); i.e., the helices in 1Q6Z are on average clearly shorter than in the case of myoglobin. The average length of a strand is even shorter; there are on average only ~4 residues; according to DSSP [15,43] there are 25 strands with a total of 93 residues in 1Q6Z. Thus the average length of a regular rodlike structure is around 6 residues which sets an intrinsic scale. Since this scale is quite smaller from the corresponding scale in myoglobin while 1Q6Z is a much longer

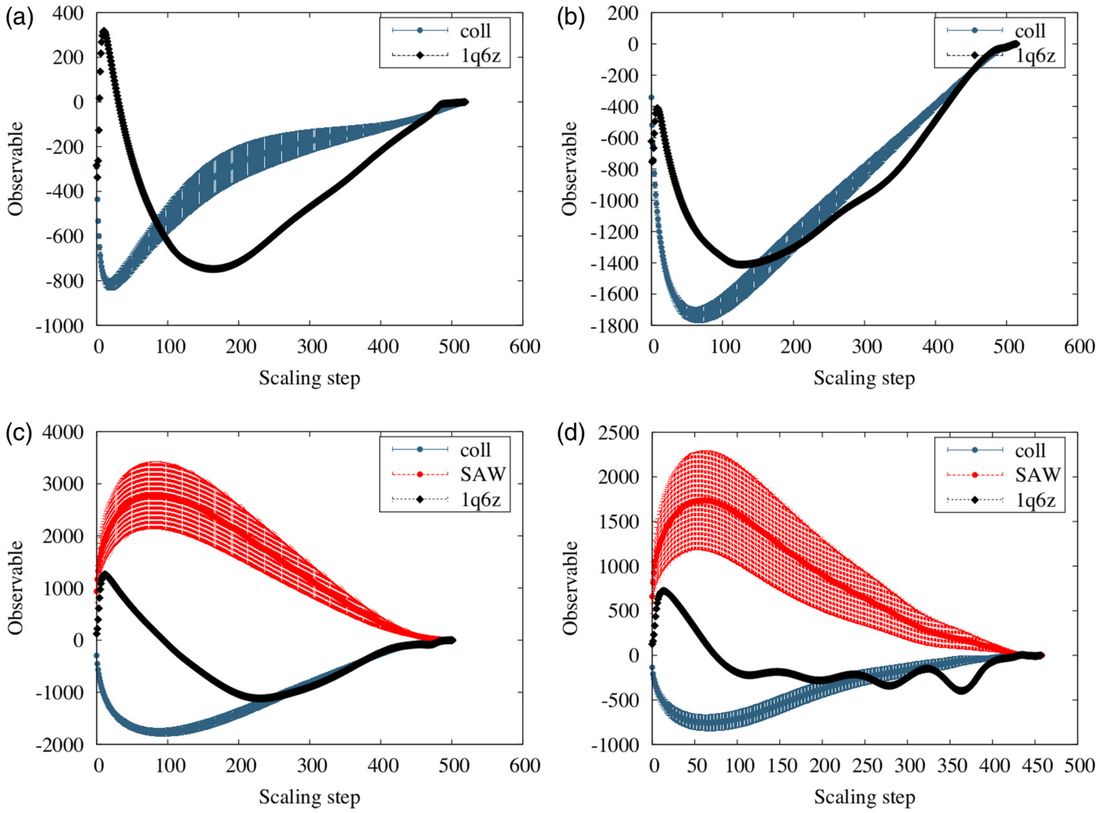


FIG. 22. Coarse graining flow of the observable (47), for 1Q6Z. In panel (a) for $k = 3$ and in panel (b) with $k = 8$ subtraction, both together with the collapsed phase homopolymer distributions. In panel (c) with $k = 21$ and in (d) with $k = 65$ subtractions, together with both collapsed phase and SARW phase homopolymer distributions. The panels for all values of k are available as Supplemental Material [44].

chain, *a priori* the phase structures of 1Q6Z and myoglobin should be different.

1. Initial collapse from $k = 1$ to $k = 18$

When $k \leqslant 18$ the Cα backbone is in a collapsed phase in the sense that the observable (47) vanishes. However, the value of the observable increases during the initial stages of the coarse graining flow, for all values of k in this phase. Thus the profile of the flow is different from that in the case of collapsed-phase homopolymers, suggesting that the collapsed phase of 1Q6Z is also different.

2. The values k from $k = 19$ to $k = 143$

Starting at $k \sim 19$ there is a crossover to a phase that resembles the SARW phase, as the initial value of the observable becomes positive. Moreover, the observable increases during the initial stages of the flow and the profile converges towards the one shown in Fig. 22(c) with with $k = 21$. The flow profile shows very persistent self-similarity over a wide range of k values, slowly converging towards the RW phase.

3. The values k from $k = 143$ to $k = 157$

For the range between $k = 143$ and $k = 157$ there is a short regime where the initial value becomes negative, in line with the collapsed phase. However, the negative values are small and the profile returns to one that resembles the profiles with $k < 143$.

4. The values k above $k = 157$

For large values of k we observe slow convergence towards the asymptotic large-k RW phase profile.

5. 1Q6Z summary

In summary, 1Q6Z does not appear to have any clear and persistent collapsed regime of k values; there is a collapsed regime only for $k < 18$ but with an increasing observable during initial stages of the flow. The profile quickly converges to an apparent SARW profile that persist for a long period, slowly converging towards the large-k RW asymptote. The results propose that despite its apparent structural complexity, 1Q6Z might collapse with very few folding intermediates.

C. An example of β-stranded protein

The third example we consider is the β-stranded 452 residue protein with PDB code 2VK5 shown in Fig. 23. There are 4 helices that contain 3% of residues and 39 strands that contain 48% of residues according DSSP classification [15,43]. The average length of regular rodlike substructures, ~6 residues, is much smaller than in the case of myoglobin but comparable to that in 1Q6Z. The total number of residues in the regular structures of ~51% is also clearly smaller than in the case of myoglobin but comparable to 1Q6Z.

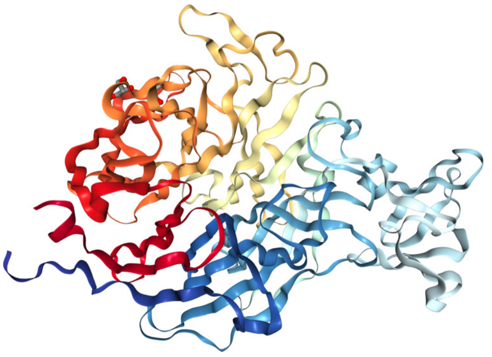


FIG. 23. Cα backbone of PDB structure 2VK5 prepared with WebGL [15].

1. The values k from $k = 1$ to $k = 18$

As shown in Fig. 24(a) already with $k = 1$ we observe more diversity in the profile of the flow than in the previous two examples. The flow profile is clearly in a collapsed phase, with quick convergence towards the collapsed phase homopolymer distribution as shown in Fig. 24(b) with $k = 6$: As in the case of myoglobin, when k coincides with the average length of the regular rodlike substructures—in this case mostly strands—the chain is deeply in the collapsed phase. Starting at $k = 9$ the flow profile starts increasing, but the collapsed phase persists until $k = 17$. At that point the initial value of the observable becomes positive valued and increases during the initial stages of the flow, suggesting that a transition to a SARW phase takes place.

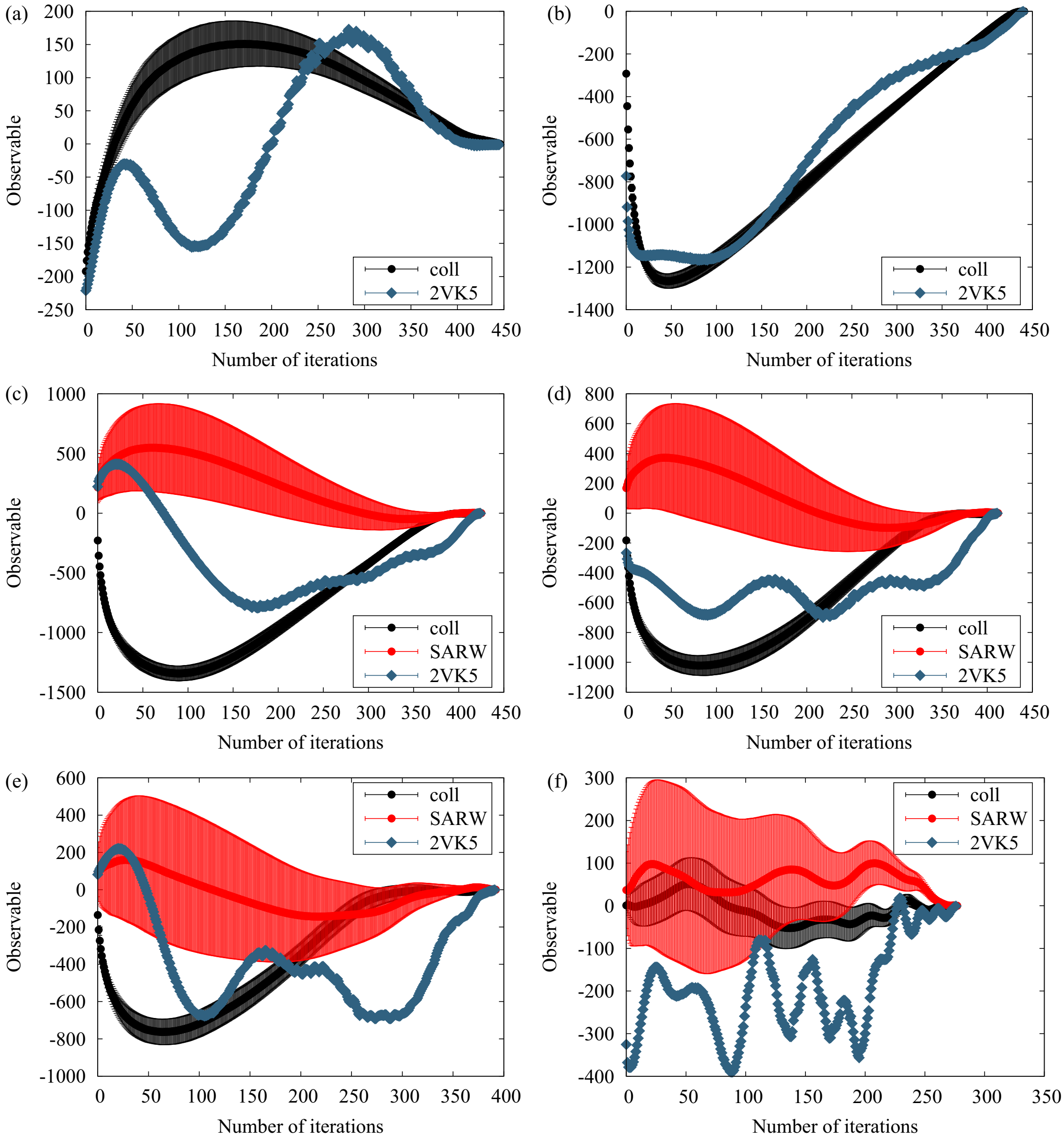


FIG. 24. Coarse graining flow of the observable (47), for PDB structure 2VK5. In panel (a) for $k = 1$ together with the ensuing homopolymer collapsed phase distribution. In panel (b) with $k = 6$ subtraction, together with the collapsed phase homopolymer distributions. In panel (c) with $k = 22$, in panel (d) with $k = 35$, in panel (e) with $k = 55$, and in panel (f) with $k = 170$ subtractions, in all these with both collapsed phase and SARW phase homopolymer distributions. The panels for all values of k are available as Supplemental Material [44].

2. The values k from $k = 19$ to $k = 29$

Between $k = 19$ and 29 we observe a SARW phase and in Fig. 24(c) we show the ensuing flow profile when $k = 22$. It is notable that the flow profile is very similar to that shown in Fig. 22(c). But unlike in the case of 1Q6Z the apparent SARW phase becomes quickly converted back to a collapsed phase.

3. The values k from $k = 30$ to $k = 53$

The flow profile proposes that the chain enters a collapsed phase akin that of the homopolymer, when $k = 30$. In Fig. 24(d) we show a typical flow profile in this range of k values; the panel corresponds to $k = 35$.

4. The values k from $k = 54$ to $k = 61$

There is an apparent SARW phase from $k = 54$ to $k = 61$. A generic flow profile is shown in Fig. 24(e); it is quite similar to the flow profile in Fig. 24(c).

5. The values k with $k = 62$ and above

When k grows above 62 the flow profile will mostly stay in the collapsed phase, except for very short fluctuations into the SARW phase for example when k is between 86 and 90. Figure 24(f) shows a generic flow profile; it is obtained for $k = 170$. This collapsed phase then converges slowly towards the asymptotic RW phase.

6. 2VK5 summary

There is apparent similarity between the evolution of the 2VK5 and myoglobin flow profiles as the cutoff value k increases: In both cases we observe repeated transitions between collapsed and SARW phases. However, while large-k myoglobin appears to reside in the SARW phase and approaches the RW phase when k increases, the large-k limit of 2VK5 appears to reside predominantly in a collapsed phase which asymptotically approaches the RW phase with increasing k.

In both 2VK5 and 1Q6Z the regular rodlike substructures, helices, and strands in the case of 1Q6Z, and strands in the case of 2VK5, have a similar average length ~6 residues. Moreover, in both cases the initial collapsed phase extends only to k values ~13–14. At the same time, in the case of myoglobin where the average helical length is much longer, around ~14 residues, the initial collapsed phase extends all the way to $k \sim 32$.

VIII. CONCLUDING REMARKS

The equilibrium phase diagram of a linear homopolymer chain can be highly elaborate [34], even though *in essence* the chain can only reside in one of the four phases that we have listed in (3). The phase where a particular homopolymer chain resides can often be determined using a single, simple global order parameter such as the radius of gyration (1), even though for a finite chain there can be corrections that are characterized by the universal exponents Δ_i in (2) which can be different for different homopolymer chains.

In the case of a heteropolymer chain such as a protein Cα backbone, it is commonly assumed that the phase structure is akin that of a homopolymer: The space-filling collapsed phase is found to be distinct to biologically active globular proteins, while proteins such as type 1 collagen form fibrils that commonly reside in the rigid rod phase under physiological conditions. When the environmental variables become altered the protein may change its phase; for example when the ambient temperature is very high most protein structures become denatured and start resembling a self-avoiding random chain.

In the case of a heteropolymer it is usually very difficult to pinpoint the phase where a chain resides, beyond visual inspection. Moreover, there are arguments that different folded proteins might actually reside in different collapsed phases [16–19]. The reason is that there is no apparent simple order parameter such as the radius of gyration that can unambiguously detect the phase of a heteropolymer chain: Unlike in the case of a homopolymer, the length of a heteropolymer chain cannot be categorically extended, as is required by a scaling law such as (1). Moreover, the phase where a protein chain appears to reside often depends on the distance scale at which the chain is inspected: At short distance scales a protein chain is often dominated by regular rodlike helices and strands, even though at large distance scales the protein appears to be space filling. The existence of multiple characteristic length scales along the chain is a characteristic feature of biologically active proteins; their biological activity is driven by an interplay of different length scales. This presence of multiple scales brings about the possibility that a folded and an apparently space-filling protein chain might simultaneously support the coexistence of several distinct phases, and with a phase structure that depends on the length scale at which the chain is inspected.

In this article we have addressed the problem of how to detect and identify the various different phases that can be supported by a linear polymer chain, in a manner that circumvents a need to extend its length as required by a scaling law such as (2). For this we have first introduced an observable that relates to the geometry of the chain in terms of local angles instead of length scales. We have also introduced a chain-specific variant of the Kadanoff block-spin transformation, that enables us to investigate the scaling properties of a given chain, without the need to increase the number of vertices and increase the length of the chain as required by the scaling law of radius of curvature. We have developed our methodology by studying the universal properties of homopolymers, including the analysis of the phase structure in terms of a computational model. We have shown how to identify the phase structure of a homopolymer using our observable and its scaling properties, effectively and in an unambiguous manner.

Finally, we have proceeded to analyze the properties of our observable in the case of crystallographic protein structures, using three representative examples. We have shown in terms of the examples how our approach enables us to identify the presence of multiple length scales, and to detect coexisting phases and the way these phases dominate the chain geometry at different length scales: Multiple length and time scales are a prerequisite for the emergence of the kind of complex structures and structural self-organization that takes place in proteins of live matter. However, our understanding of the relevance and physical origin of the diverse scales that are observed in larger globular proteins remains incomplete. The methodology we have introduced and analyzed clearly reveals how a protein chain can support different phase characteristics, in coexistence, at different length scales. Accordingly, we look forward to developing our approach into a systematic method for inspecting the multiple phases and the phenomenon of phase coexistence in the case of proteins and other linear heteropolymers.

IX. CODE

The code which we used for coarse graining and calculating the observable for polymer chains is accessible online [46].

ACKNOWLEDGMENTS

A.S. and J.N. acknowledge funding from the Knut and Alice Wallenberg Foundation and Vetenskapsrådet. The work of A.J.N. is supported by a grant from Vetenskapsrådet and by a Qian Ren grant. The work of M.U. is supported by German Research Fund DFG Grant No. BU 2626/2-1.

[1] M. L. Huggins, J. Chem. Phys. **9**, 440 (1941).

[2] P. J. Flory, J. Chem. Phys. **9**, 660 (1941).

[3] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953).

[4] P. J. Flory, J. Chem. Phys. **17**, 303 (1949).

[5] P. G. de Gennes, J. Chem. Phys. **55**, 572 (1971).

[6] P. G. de Gennes, Phys. Lett. A **38**, 339 (1972).

[7] J. Des Cloizeaux, J. Phys. (Paris) **36**, 281 (1975).

[8] S. F. Edwards, Polymer **6**, 143 (1977).

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

[10] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, New York, 1986).

[11] A. Yu. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules*, AIP Series in Polymers and Complex Materials (AIP, Woodbury, 1994).

[12] T. Nakayama, Y. Kousuke, and R. L. Orbach, Rev. Mod. Phys. **66**, 381 (1994).

[13] B. Li, N. Madras, and A. Sokal, J. Stat. Phys. **80**, 661 (1995).

[14] L. Schäfer, *Excluded Volume Effects in Polymer Solutions, as Explained by the Renormalization Group* (Springer-Verlag, Berlin, 1999).

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucl. Acids Res. **28**, 235 (2000).

[16] T. G. Dewey, J. Chem. Phys. **98**, 2250 (1993).

[17] L. Hong and J. Lei, Polym. Sci. B **47**, 207 (2009).

[18] J. Lei and K. Huang, Europhys. Lett. **88**, 68004 (2009).

[19] N. Rawat and P. Biswas, J. Chem. Phys. **131**, 165104 (2009).

[20] L. P. Kadanoff, Physics **2**, 263 (1966).

[21] K. Wilson, Phys. Rev. B **4**, 3174 (1971).

[22] M. E. Fisher, Rev. Mod. Phys. **46**, 597 (1974).

[23] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, Reading, 1992).

[24] A. Krokhotin, S. Nicolis, and A. J. Niemi, J. Chem. Phys. **140**, 095103 (2014).

[25] S. Hu, M. Lundgren, and A. J. Niemi, Phys. Rev. E **83**, 061908 (2011).

[26] A. J. Niemi, Phys. Rev. D **67**, 106004 (2003).

[27] U. H. Danielsson, M. Lundgren, and A. J. Niemi, Phys. Rev. E **82**, 021910 (2010).

[28] X. Peng, A. K. Sieradzan, and A. J. Niemi, Phys. Rev. E **94**, 062405 (2016).

[29] S. Hu, Y. Jiang, and A. J. Niemi, Phys. Rev. D **87**, 105011 (2013).

[30] T. Ioannidou, Y. Jiang, and A. J. Niemi, Phys. Rev. D **90**, 025012 (2014).

[31] T. Ioannidou and A. J. Niemi, Phys. Lett. A **380**, 333 (2015).

[32] I. Gordeli, D. Melnikov, A. J. Niemi, and A. Sedrakyan, Phys. Rev. D **94**, 021701(R) (2016).

[33] A. J. Niemi, in *Topological Aspects of Condensed Matter Physics*, edited by C. Chamon, M. O. Goerbig, R. Moessner, and L. F. Cugliandolo (Oxford University Press, Oxford, 2017).

[34] A. Sinelnikova, A. J. Niemi, and M. Ulybyshev, Phys. Rev. E **92**, 032602 (2015).

[35] P. L. Primalov, Adv. Protein Chem. **33**, 167 (1979).

[36] P. L. Primalov, Annu. Rev. Biophys. Biophys. Chem. **18**, 47 (1989).

[37] E. Shakhnovich and A. Finkelstein, Biopolymers **28**, 1667 (1989).

[38] M. N. Chernodub, M. Lundgren, and A. J. Niemi, Phys. Rev. E **83**, 011126 (2011).

[39] N. Molkenthin, S. Hu, and A. J. Niemi, Phys. Rev. Lett. **106**, 078102 (2011).

[40] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals* (Clarendon Press, Oxford, 1995).

[41] O. B. Ptitsyn, J. Protein Chem. **6**, 273 (1987).

[42] O. B. Ptitsyn, Curr. Opin. Struct. Biol. **5**, 74 (1995).

[43] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[44] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.97.052107 for detailed analysis of the different protein structures.

[45] H. J. Dyson and P. E. Wright, Acc. Chem. Res. **50**, 105 (2017).

[46] A. Sinelnikova, http://www.doi.org/10.5281/zenodo.581166.