Protein folding on rugged energy landscapes: Conformational diffusion on fractal networks

Gregg Lois,[1,2] Jerzy Blawzdziewicz,[2] and Corey S. O'Hern[2]

[1]*Areté Associates, P.O. Box 2607, Winnetka, California 91396, USA*

[2]*Department of Mechanical Engineering and Department of Physics, Yale University, New Haven, Connecticut 06520-8286, USA*

We perform simulations of model proteins to study folding on rugged energy landscapes. We construct "first-passage" networks as the system transitions from unfolded to native states. The nodes and bonds in these networks correspond to basins and transitions between them in the energy landscape. We find power-law relations between the folding time and the number of nodes and bonds. We show that these scalings are determined by the fractal properties of first-passage networks. Thus, we have identified a possible mechanism—the small fractal dimension of first passage networks—which can give rise to reliable folding in proteins with rugged energy landscapes.

Understanding how proteins reliably fold to their native conformations despite frustration in the form of non-native interactions between residues is an important, open question. Advances in experimental techniques, such as single-molecule fluorescence [1] and fast thermal quenching methods [2], have enabled a quantitative characterization of the dynamics that occur during folding of single proteins. For example, we now know that a large number of metastable conformations are sampled during the folding and unfolding processes, as observed in folding stability [3] and mechanical denaturation [4] studies.

How does a protein fold reliably to its native conformation even though a large number of metastable states exist? For over twenty years the answer to this question has been the principle of minimal frustration [5]. Within this framework, one recognizes that metastable states are present, but assumes that the barriers separating local energy minima are sufficiently low that there is still a large thermodynamic force driving folding to the native state [6]. This idea is illustrated by the funneled energy landscape in Fig. 1(a), where the roughness scale δE is much smaller than depth of the energy minimum ΔE that drives folding ($\delta E \ll \Delta E$). While the funneled energy landscape may explain how some proteins fold reliably [7], a different picture, i.e., rugged energy landscapes may describe folding and conformational dynamics in metastable [8] and intrinsically disordered [9] proteins, as well as misfolding [10]. Rugged energy landscapes, as shown in Fig. 1(b), possess a roughness scale that is comparable to that of the smooth funnel $\delta E \sim \Delta E$. In this limit, the thermodynamic drive to fold is absent on biological time scales, and protein conformational dynamics proceed via activation over energy barriers with only local knowledge of the landscape.

What physical observables differentiate proteins with funneled versus rugged landscapes? Recent studies indicate that proteins with rugged landscapes exhibit a crossover between single exponential folding at large temperature and stretched exponential folding at low temperature [11], caused by metastable states in the energy landscape that become increasingly important at low temperature [12]. Moreover, computational studies have identified the network of states populated by proteins during folding [13], and it has been postulated that the network topology might provide a basis for understanding the heterogeneity of the transition states in proteins with rugged energy landscapes [14]. Further, simulations of diffusion [15] and return-time probabilities [16] show that the set of all local minima in the energy landscape of model proteins exhibit fractal properties.

Here we study the network topology of the states and transitions sampled by a model protein *during its folding transition*. This is a subset of all local minima and all possible transitions, and is dependent on temperature. The model protein has a rugged energy landscape and explores 10^2–10^4 distinct states during its folding transition to a unique native state. We find that a statistical ensemble of pathways to the native state exists, with large fluctuations in folding times. In fact, the folding time and number of distinct states sampled during folding scale as a power law, which suggests that reliable folding on rugged landscapes can be described as conformational diffusion on a *fractal* network of basins. We calculate the scaling relations expected for diffusion on fractal networks and find excellent agreement with the measured power laws.

I. HETEROPOLYMER MODEL

To study proteins with rugged energy landscapes, simulation models should possess three key features: (1) unique native state, (2) many metastable, local energy minima, and (3) large energy barriers that separate local minima so that
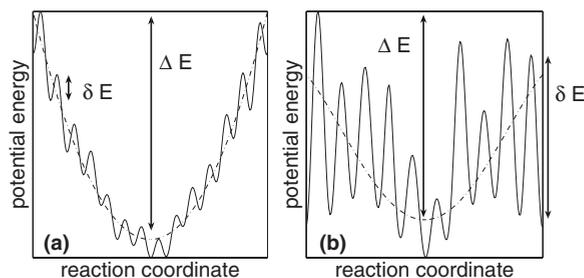
FIG. 1. Schematics of (a) funneled and (b) rugged energy landscapes. In (a), the depth of the energy minimum that drives folding $\Delta E \gg \delta E$, where δE gives the root-mean-square energy fluctuations over the given range of the reaction coordinate. In (b), $\Delta E \sim \delta E$.
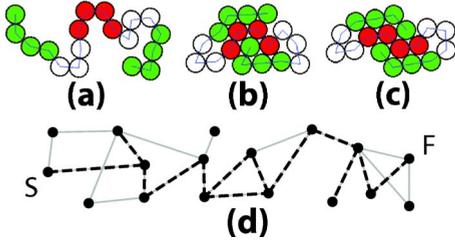
FIG. 2. (Color online) The heteropolymer model in its (a) extended, (b) metastable misfolded, and (c) native states. (d) Schematic of a first-passage network (black dashed lines) from basin "*S*" to "*F*," superimposed on the complete network composed of all basins and transitions between them (gray lines).

$\delta E \sim \Delta E$. Further, we must be able to search configuration space in a reasonable amount of computer time, which at present excludes all-atom simulations. In these studies, we will focus on a model heteropolymer that exhibits features (1)–(3).

We model proteins as heteropolymers composed of equal-sized spherical monomers with hydrophobic and hydrophilic interactions [17]. The model includes hydrophilic monomers (white) and two types of hydrophobic monomers (red and green) as shown in Fig. 2. Green and red monomers interact via a Lennard-Jones potential with minimum energy $-E_{att}$, except the green monomers on both ends of the chain that interact with minimum energy $-2E_{att}$. All other monomer-monomer interactions are purely repulsive [18]. We also include a finitely extensible nonlinear elastic (FENE) potential [19] between adjacent monomers to maintain the polymer constraint. We simulate the 18–mer sequence ggggwwwr-rrwwwgggg, where g, w, and r represent green, white, and red monomers, respectively. This model displays a complex energy landscape with $\sim 10^5$ distinct local energy minima. For simplicity, local minima are defined by the list of contacting green and red monomers [20]. The native conformation of this heteropolymer is given by the particular set of 14 green-red contacts shown in Fig. 2(c). All of the results presented here are for two-dimensional (2D) systems, however, we expect similar results in 3D.

Thermal fluctuations of the heteropolymer are studied using Brownian dynamics, where the temperature T is reported in units of the attractive energy, e.g. $T=1/3$ corresponds to thermal energy $E_{att}/3$. To compare results for rugged and funneled energy landscapes, we also simulated the same heteropolymer with Go-interactions [21], where attractive interactions are only included between monomers that form contacts in the native state. Simple measures of kinetics are the folding and unfolding times shown in Fig. 3. The folding time τ_f is calculated by preparing the heteropolymer in an ensemble of extended states and measuring the average folding time to the native state. τ_u is the average unfolding time from the native state to any extended state with zero red-green contacts. For temperature $T < T^*=0.8$, $\tau_f < \tau_u$, and the extended conformation is significantly less stable than the native state. The increase in τ_f as T decreases, as shown in Fig. 3, has been observed in experimental studies of proteins [22] and is a general feature of materials quenched below the
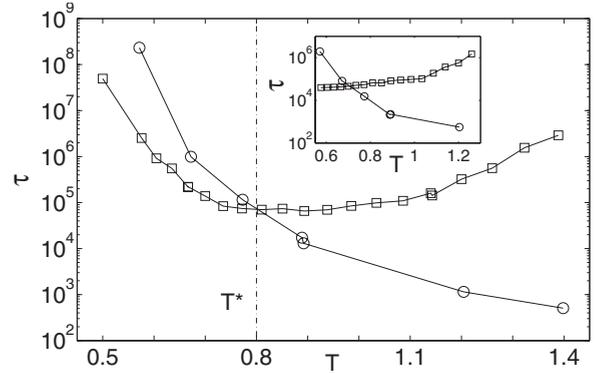
FIG. 3. Ensemble-averaged folding τ_f (squares) and unfolding τ_u (circles) times vs temperature for the heteropolymer (main figure) and Go (inset) models. The vertical line at $T^*=0.8$ indicates the folding temperature.

glass transition temperature [23] when energy barriers become large compared to T. An important feature of the heteropolymer model is that folding only occurs for temperatures where $d\tau_f/dT < 0$. In contrast, folding simulations of the Go model yield $d\tau_f/dT > 0$ for all T, as shown in the inset to Fig. 3.

II. FIRST-PASSAGE NETWORKS

For each heteropolymer conformation, we determine the list of contacting green and red monomers and uniquely associate this list of contacts with a basin that surrounds the associated local energy minimum. For rugged landscapes, the system will sample a large number of basins as folding proceeds from the extended to the native state. The trajectory of the model protein as it folds can be viewed as a network of connected nodes in configuration space. The nodes represent the basin of a local energy minimum sampled by the system, and bonds that join two nodes represent transitions from one basin to another. These networks are termed "first-passage networks" since they are formed as the protein makes its first passage from an initial to the native conformation. Note that each first-passage network is a subset of the static network of all basins and transitions, as illustrated in Fig. 2(d).

We compiled $\sim 10^6$ first-passage networks originating from the non-native conformation in Fig. 2(b) and ending at the native state over a range of $T \leq 0.8$. We map the conformation of the heteropolymer to its associated basin every q time steps to construct first-passage networks. We assume that the features of the first-passage networks depend on T but are independent of the initial state since the first-passage networks are composed of a large number of nodes.

The simplest properties of first-passage networks are the number of distinct basins sampled (nodes) N_i and bonds N_b. Nodes and bonds are only counted once, even if multiple transitions are made between a given set of basins. We also measure the total number of transitions $N_t \propto \tau_f \geq N_b$. Figure 4 shows raw data for the number of bonds N_b and transitions N_t plotted versus the number of nodes N_i using $q=1000$. There are 850 data points for each temperature, each taken
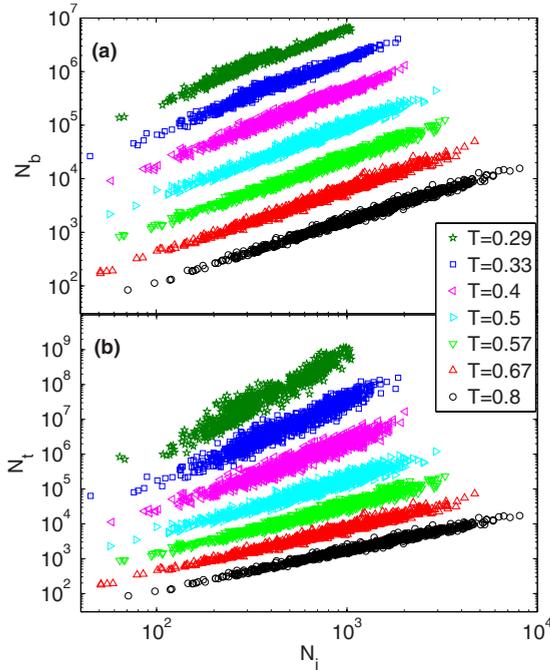
FIG. 4. (Color online) Number of (a) bonds N_b and (b) transitions N_t in first-passage networks vs the number of nodes N_i over a range of temperature T. For each T, N_b, and N_t have been multiplied by constant factors (shifted vertically) for clarity.

from a distinct first-passage network. For all T the number of sampled basins, N_i, fluctuates between 10^2 and 10^4, which indicates that the model protein adopts a large number of conformations before arriving at the native state. The wide range of N_i indicates that there is not a single folding pathway, but rather a statistical ensemble of pathways.

In Fig. 4, N_b, N_t, and N_i show strong fluctuations from one realization to the next; however, the fluctuations obey power-law scaling

$$N_b \propto N_i^{\Lambda} \quad \text{and} \quad N_t \propto N_i^{\Gamma}. \tag{1}$$

This correlation is nontrivial and depends on global properties of first-passage networks. We find that distributions of local features of the network, such as single-jump activation times and distances, and the number of bonds per node, are exponential. (See Fig. 7.) Thus, local properties of first-passage networks cannot be responsible for the power-law scaling.

In Fig. 5, we plot the scaling exponents Γ and Λ at different temperatures T. While Λ reaches a plateau at ≈ 1.4 at small T, Γ continues to increase with decreasing T. The increase of Γ is a signature of temperature-dependent exploration of configuration space in systems with rugged landscapes. A system with a rugged energy landscape at energy E only samples a small temperature-dependent fraction of conformations at that energy due to large activation barriers. In contrast, $\Gamma \approx 1.5$ at all T for the same heteropolymer model with Go interactions. In systems with funneled energy landscapes (i.e., the Go model), a protein with energy E samples conformations with that energy more uniformly.
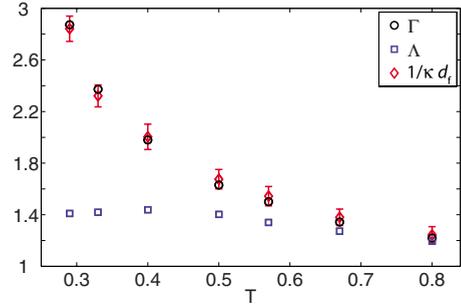
FIG. 5. (Color online) The scaling exponents Γ and Λ and the prediction $1/\kappa d_{\mathrm{f}}$ for Γ from Eq. (4). Error bars for Γ and Λ are smaller than the symbol sizes.

The data shown in Fig. 4 are obtained by identifying basins every $q = 1000$ time steps. We have also performed simulations in the range $1 < q < 10^4$ and observe that the exponents Γ and Λ are independent of q. These results further indicate that first-passage networks are self-similar and fractal.

III. ORIGIN OF POWER LAWS

If we assume that first-passage networks are fractal, we can predict the exponent Γ from the fractal scaling exponents of the network. This assumption will be verified *a posteriori*.

On any network we can define the chemical distance Δc given by the shortest path between two nodes of the network. This distance is useful because it depends only on network connectivity and is independent of the embedding space [24]. For a fractal network, we expect [25]

$$\Delta c \propto t^{\kappa}, \tag{2}$$

$$N(\Delta c) \propto \Delta c^{d_{\mathrm{f}}}, \tag{3}$$

where $N(\Delta c)$ is the number of distinct basins sampled within chemical distance Δc and time interval t, d_{f} is the chemical fractal dimension, and the exponent κ characterizes the scaling of Δc with time.

Given these relations, the correlation between N_i and N_t can be explained as follows. A single first-passage network is formed over folding time $\tau_f \propto N_t$, during which the system explores average chemical distance $\Delta c \propto N_t^{\kappa}$ [Eq. (2)]. Moreover, for a given chemical distance Δc, the number of sampled basins on the first passage network scales as $N_i \propto N(\Delta c) \propto \Delta c^{d_{\mathrm{f}}}$ [Eq. (3)]. Thus, both N_i and N_t are related to Δc, and we find $N_t \propto N_i^{1/\kappa d_{\mathrm{f}}}$, or

$$\Gamma = \frac{1}{\kappa d_{\mathrm{f}}}. \tag{4}$$

The prediction for Γ relies on the first-passage networks being fractal. In Fig. 6(a), we test Eq. (2) and observe that Δc grows as a power law at large t for all temperatures studied. We average Δc over 1 500 first-passage networks and only include $t < \tau_f$ for each realization. The exponent κ decreases with T, which implies that colder systems explore chemical distance more slowly.
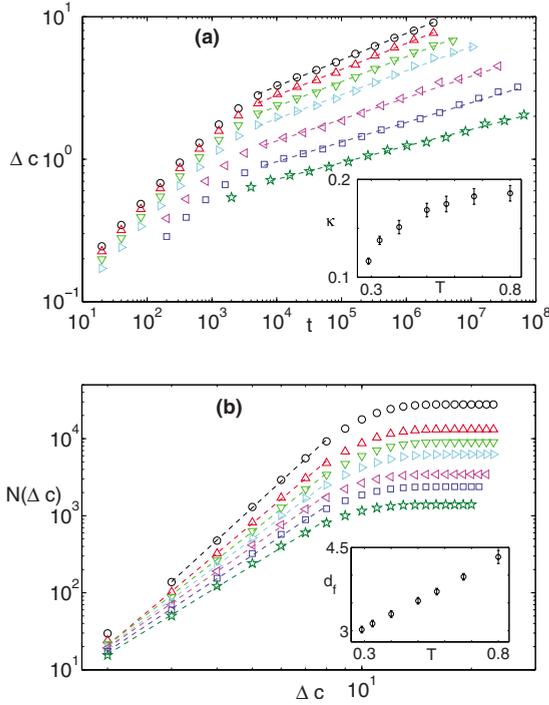
FIG. 6. (Color online) (a) The chemical distance Δc sampled in the time interval t by the heteropolymer and (b) the mean number of basins $N(\Delta c)$ within Δc at different temperatures. In (a) and (b), the symbols are the same as in Fig. 4, and the insets display the scaling exponents used to fit the data (dotted lines) for different temperatures.

FIG. 7. (Color online) Symbols are same as in Fig. 4. (a) Waiting-time distributions $\psi(\tau)$ at different temperatures. The inset shows $\tau^2 \psi(\tau)$. (b) Average chemical distance Δc as a function of the number of transitions N_t at different temperatures. The inset shows the power-law exponents κ (black circles) and compares them to those shown in Fig. 6(a) (red squares).

In Fig. 6(b) we test Eq. (3) and find that, over the limited range of chemical distance accessible to our small heteropolymer, the chemical fractal dimension d_f is well-defined and depends linearly on temperature. $N(\Delta c)$ is computed by including all sampled basins in 850 different first-passage networks at each T. While power-law scaling of $N(\Delta c)$ only holds for $\Delta c \lesssim 8$, the average chemical distance explored on a first-passage network is always smaller than 8. Therefore, the prediction for Γ based on power-law scaling should hold during the folding process. In Fig. 5, we find excellent agreement between the folding-time exponent Γ and our prediction $1/\kappa d_f$.

The assumption that first-passage networks are fractal has been empirically justified *a posteriori* by comparing the measured exponent Γ with the prediction $1/\kappa d_f$. Another possible cause of anomalous diffusion is kinetic in origin. It is well-known [26] that if the waiting-time distribution between transitions is power law distributed $\psi(\tau) \propto \tau^{-\mu}$ with exponent $1 < \mu < 2$, anomalous diffusion will occur with $2\kappa = \mu - 1$. We test the importance of kinetic effects in this model protein by calculating the waiting (or residence) time distributions $\psi(\tau)$ in Fig. 7(a) as a function of temperature, where τ is the waiting time normalized by the average value. τ is obtained by measuring the time elapsed while the system resides in a given local minimum. As shown in inset to Fig. 7(a), $\psi(\tau)$ shows approximate power-law scaling with exponent two at small τ, and a much faster decay at large τ. Even if the waiting distribution possessed power-law scaling with

$\mu = 2$ over the full range of τ, this would yield weak anomalous diffusion with $\kappa = 0.5$. Since we observe $\kappa < 0.2$ for all T, it is clear that broad waiting time distributions are not the origin of the fractal scaling behavior of the folding dynamics for our model protein. To further test this hypothesis, we measured the dependence of the average chemical distance Δc on N_t in Fig. 7(b) and find that $\Delta c \propto N_t^\kappa \propto t^\kappa$ with κ identical to that in Fig. 6(a). This is further evidence that the fractal 'first-passage' networks, not kinetics from one local minimum to another, give rise to anomalous diffusion in our model protein. While kinetic effects often play an important role in folding [27], we find that they are not relevant to the properties of the model protein studied here.

IV. CONCLUSIONS

We have studied first-passage networks formed by the folding trajectories of a heteropolymer and observed power-law scaling between the folding time ($\propto N_t$) and number of nodes N_i and bonds N_b in first-passage networks. We have also demonstrated that the folding-time exponent Γ can be obtained by measuring the fractal exponents that characterize the structure of first-passage networks in configuration space.

The configuration space of our model protein contains only 10^5 minima, which is small compared to what might be expected for real proteins. Therefore, it is possible for the model protein to fold via an exhaustive search of configura-

tion space, whereas real proteins cannot (Levinthal's paradox [28]). However, we observe that the model protein studied here does not search all of its configuration space. Instead, it saves time by searching a fractal network of the possible states. Since this search mechanism is utilized in our simple model protein, it is possible that it also occurs in heteropolymers with varied sequences and more complex proteins.

Our results do not describe properties of the complete static network of basins in the energy landscape (which is also fractal [15,16]). Instead, our results suggest that it is not necessary to characterize the complete static network if kinetic effects (such as waiting-time distributions) do not strongly affect dynamics on the network. Just as normal diffusion will trace out a two-dimensional fractal network of sampled nodes, no matter what the dimension of the underlying space is, proteins with rugged energy landscapes can trace out fractal networks that are independent of the complete network. This behavior is not peculiar to proteins with rugged energy landscapes, but can also occur in glass-forming materials at low temperature [29]. Moreover, d_f decreases with temperature, and is always much smaller than the dimension of configuration space D, which implies that $N_i \sim (\Delta c)^{d_f} \ll (\Delta c)^D$. This provides a mechanism by which systems with rugged energy landscapes can fold reliably without kinetic pathways and offers a potential resolution to Levinthal's paradox.

[1] E. A. Lipman, B. Schuler, O. Bakajin, and W. A. Eaton, Science **301**, 1233 (2003).

[2] W. A. Eaton, V. Munoz, P. A. Thompson, C.-K. Chan, and J. Hofrichter, Curr. Opin. Struct. Biol. **7**, 10 (1997).

[3] H. Yang, G. Luo, P. Karnchanaphanurach, T.-M. Louie, I. Rech, S. Cova, L. Xun, and X. S. Xie, Science **302**, 262 (2003); J. Bredenbeck, J. Helbing, J. R. Kumita, G. A. Woolley, and P. Hamm, Proc. Natl. Acad. Sci. U.S.A. **102**, 2379 (2005).

[4] J. Brujić, R. I. Hermans Z., K. A. Walther, and J. M. Fernandez, Nat. Phys. **2**, 282 (2006).

[5] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).

[6] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[7] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).

[8] D. Baker and D. A. Agard, Biochemistry **33**, 7505 (1994); J. C. Whisstock and S. P. Bottomley, Curr. Opin. Struct. Biol. **16**, 761 (2006).

[9] A. Vitalis, X. Wang, and R. V. Pappu, Biophys. J. **93**, 1923 (2007).

[10] C. Soto, Nat. Rev. Neurosci. **4**, 49 (2003).

[11] J. A. Ihalainen, B. Paoli, S. Muff, E. H. G. Backus, J. Bredenbeck, G. A. Woolley, A. Caflisch, and P. Hamm, Proc. Natl. Acad. Sci. U.S.A. **105**, 9588 (2008).

[12] S. V. Krivov, S. Muff, A. Caflisch, and M. Kaplus, J. Phys. Chem. B **112**, 8701 (2008).

[13] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **103**, 17747 (2006).

[14] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).

[15] T. N. Neusius, I. Daidone, I. M. Sokolov, and J. C. Smith, Phys. Rev. Lett. **100**, 188103 (2008).

[16] R. Du, A. Y. Grosberg, and T. Tanaka, Phys. Rev. Lett. **84**, 1828 (2000).

[17] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).

[18] G. Lois, J. Blawzdziewicz, and C. S. O'Hern, Biophys. J. **95**, 2692 (2008).

[19] R. M. Jendrejack, M. D. Graham, and J. J. de Pablo, J. Chem. Phys. **113**, 2894 (2000).

[20] A contact is defined when the distance between two monomers is less than the separation at which the concavity of the Lennard-Jones potential changes sign.

[21] H. Taketomi, Y. Ueda, and N. Go, Int. J. Pept. Protein Res. **7**, 445 (1975).

[22] S.-I. Segawa and M. Sugihara, Biopolymers **23**, 2473 (1984).

[23] W. H. Wang, W. Utsumi, and X.-L. Wang, EPL **71**, 611 (2005).

[24] We use Δc because the Euclidean distance between basins in configuration space is not uniquely defined.

[25] R. Orbach, Science **231**, 814 (1986); T. Nakayama and K. Yakubo, Rev. Mod. Phys. **66**, 381 (1994).

[26] J.-P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990).

[27] L. Bongini, L. Casetti, R. Livi, A. Politi, and A. Torcini, Phys. Rev. E **79**, 061925 (2009).

[28] R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. U.S.A. **89**, 20 (1992).

[29] G. Lois, J. Blawzdziewicz, and C. S. O'Hern, Phys. Rev. Lett. **102**, 015702 (2009).