


Estimating ruggedness of free-energy landscapes of small globular proteins from principal component analysis of molecular dynamics trajectories

Andreas Volkhardt and Helmut Grubmüller 

Max Planck Institute for Multidisciplinary Sciences, Theoretical and Computational Biophysics Department,
Am Fassberg 11, 37077 Göttingen, Germany



(Received 18 August 2021; accepted 28 February 2022; published 15 April 2022)

The internal dynamics of biomolecules, and hence their function, is governed by the structure of their free-energy landscape. Early flash-photolysis experiments on myoglobin suggested that the free-energy landscapes of proteins are hierarchically structured, with a characteristic distribution of free-energy barriers which gives rise to anomalous diffusion. Analytical results have been derived for one-dimensional or high-dimensional hierarchical free-energy landscapes. Recent improvements in methods and computer performance enable generating sufficiently long molecular dynamics (MD) trajectories to extract dynamics information covering many orders of magnitude, such that the broad distributions of energy barriers of proteins become accessible to quantitative studies of intermediate dimensions. In this work, we present a nonequilibrium method to estimate barrier height distributions from microsecond-long MD simulations. It infers barrier height distributions from anomalous diffusion exponents derived from principal component analysis and by comparison to simple hierarchical lattice models. These models are d -dimensional lattices of states separated by free-energy barriers, the heights of which are distributed as $p(\Delta G) = 1/\gamma \exp(-\Delta G/\gamma)$. The parameter γ quantifies the “ruggedness” of the free-energy landscape in such models. We show that both parameters, i.e., ruggedness and effective dimensionality d , can be inferred from anomalous diffusion exponents. Assuming a similar dependency of anomalous diffusion exponents on γ and d for proteins, we estimate the ruggedness of the free-energy landscapes of 500 small, single-domain globular proteins between 15 and 20 kT per dimension with an estimated accuracy of 4.2 kT and dimensionality between 40 and 60 with an estimated accuracy of 10 dimensions. Remarkably, neither effective dimensionality nor the ruggedness correlates with protein size and both ruggedness and effective dimensionality are much smaller than the scatter of protein sizes. From this finding, we conclude that these two properties of the free-energy landscape of a protein are rather adapted to the particular function of each single protein and that, quite generally, are functionally relevant for globular proteins.

DOI: [10.1103/PhysRevE.105.044404](https://doi.org/10.1103/PhysRevE.105.044404)

I. INTRODUCTION

Most processes in life are governed by proteins, macromolecules consisting of a chain of amino acids. Their biophysical function in living cells is intimately linked to their structure and, in particular, to their remarkably complex internal dynamics on timescales ranging from picoseconds to hours. These thermally activated internal motions are governed by a diffusion process on a free-energy landscape [1]. Moessbauer spectroscopy and neutron scattering experiments showed that protein free-energy landscapes with conformational coordinates as its arguments [2] are characterized by a large number of nearly isoenergetic minima. Free-energy barriers between these minima are structured hierarchically, as shown by flash-photolysis experiments on myoglobin [3] (see Fig. 1).

Recent progress in methods and performance of computational hardware allows generating molecular dynamics (MD) trajectories ranging over multiple orders of magnitude in simulation length up to multiple microseconds as a standard routine. This development allows us to study the hierarchical structure of protein free-energy landscapes *in silico* [4]. However, due to the large number of possible protein configurations, available MD trajectories are still not long enough to reach thermal equilibrium, which constitutes the well-known sampling problem of MD simulations. Slowest relaxation times that correspond to folding and unfolding times are on the timescale of minutes or even hours, which is still beyond what MD simulations are capable of simulating on a reasonable computing timescale.

To circumvent this problem, we use nonequilibrium methods that, rather than finding a model for a protein’s equilibrium dynamics, model its dynamics as a diffusion process within its free-energy landscape. Diffusion processes in hierarchical free-energy landscapes models were explored for simple one-dimensional [5] as well as several many-dimensional models [6,7]. It has been shown that diffusion in such models is anomalous, i.e., the variance of trajectories increases with a power law in time [5,8]. In particular, it was analytically shown how exponents depend on the

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI. Open access publication funded by the Max Planck Society.

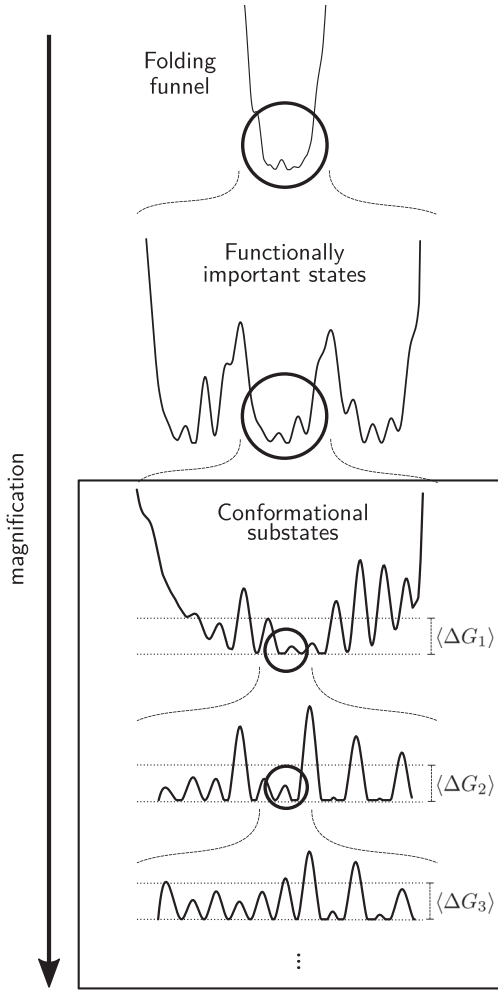


FIG. 1. Hierarchical structure of protein free-energy landscapes proposed by Frauenfelder [3]. Top: On a large scale, the folding funnel dominates free-energy landscapes of globular proteins [1]. Within this folding, on a smaller scale, functionally important states are found. Each of these contains a hierarchy of conformational substates, where hierarchy tiers i are characterized by mean barrier heights $\langle \Delta G_i \rangle$. Due to the multitude of different isoenergetic conformational substates, the free-energy landscape of proteins on this scale is best described in statistical terms, i.e., in terms of barrier distributions [12].

barrier-height distribution [5]. Similarly, anomalous diffusion behavior would be expected for hierarchical protein free-energy landscapes and was indeed observed in MD simulations of small peptides [9] and small globular proteins [10] (see Fig. 2). Assuming that the observed anomalous diffusion behavior arises from the hierarchical structure of protein free-energy landscapes, it should be possible to estimate barrier height distributions from anomalous diffusion exponents obtained from MD trajectories.

In this work, we estimated barrier height distributions of 500 small globular proteins selected to cover known folds and functions from anomalous diffusion exponents observed in MD simulations. To this end, we generated for each of these proteins 1- μ s molecular dynamics trajectories and carried out trajectory-length-dependent principal component

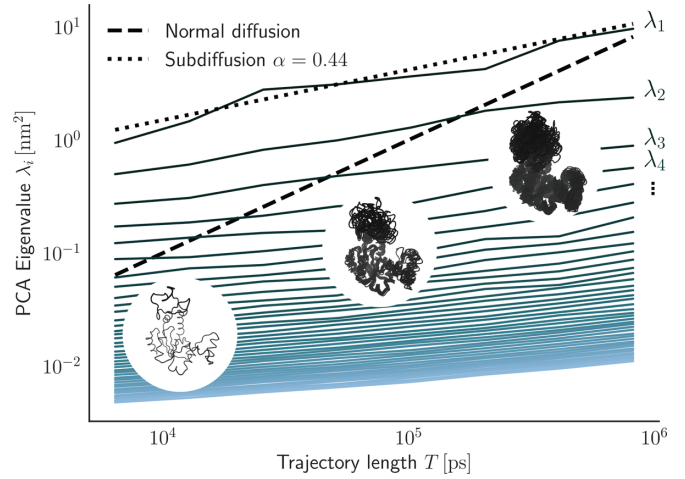


FIG. 2. Variance of molecular dynamics trajectories along collective coordinates shows a power-law-like scaling behavior in dependence of trajectory length. The figure shows the variance of a 5- μ s molecular dynamics trajectory of adenylate kinase from *Escherichia coli* (PDB code: 1AKE) along orthogonal collective coordinates, i.e., principal components (PC) in dependence of trajectory length T . These collective coordinates are (PCA) eigenvectors of the covariance matrix of the trajectory and are typically ordered according to the magnitude of their corresponding (PCA) eigenvalues λ_i , which represent the variance of the trajectory along the eigenvector. PCA eigenvalues λ_i of MD trajectories approximately increase with a power law depending on trajectory length T . The scaling exponents α_i (slopes in a log-log plot) of these power laws show subdiffusive behavior as $\alpha_i < 1$.

analysis [10] for the selected proteins. To translate the observed anomalous diffusion exponents into barrier height distributions, we used a d -dimensional hierarchical model. This model consists of a lattice of states: Transition rates between adjacent states are governed by static free-energy barriers ΔG , which are randomly distributed according to $p(\Delta G) = \frac{1}{\gamma} e^{-\Delta G/\gamma}$, where γ quantifies the “ruggedness” of the hierarchical free-energy landscape. The relation between anomalous diffusion exponents and γ is analytically known only for one-dimensional (1D) models and, in a mean-field approximation, for high dimensions $d \rightarrow \infty$ [8]. However, because the essential configurational subspace of proteins is assumed to be $\sim 10 < d < \sim 100$ [11], we had to resort to a numerical approach by simulating random walks in models with 3–200 dimensions. Indeed, we observed large deviations from the mean-field approximation. By cross validation, we showed that, based on this numerically obtained relation, barrier height distributions can be estimated with an accuracy of ~ 5 kT.

Applying the same approach to 1- μ s MD trajectories, we determined γ for protein free-energy landscapes. We found that most ruggedness coefficients of the proteins fall within $\gamma \approx 15 - 20$ kT/ d with an estimated essential subspace dimensionality $d \approx 40 - 60$. This result provides evidence that the dynamics of a broad range of protein folds is governed by similar barrier height distributions.

II. THEORY

To explain nonexponential kinetics, e.g., ligand binding experiments [3], Frauenfelder proposed early on that, in the folded state (Fig. 1, top), the underlying intramolecular protein dynamics is governed by a hierarchical (free) energy landscape [3] (Fig. 1, bottom and magnification). Accordingly, the kinetics of larger, typically functional protein motions are governed by higher free-energy barriers located at correspondingly larger distances in configurational space (top sketch in the box of Fig. 1). These barriers separate “taxonomic conformational states.” Within each of these functional states, increasingly smaller and faster motions between “statistical substates” [12] are described by more frequent crossings of increasingly lower barriers separated by correspondingly smaller distances (lower sketches in the box of Fig. 1). Overall, the protein free-energy landscape is thus described by a hierarchy of energy barriers with characteristic barrier heights and separations at each tier. Precisely how the barrier height increases with increasing mutual distance between the barriers is described by the ruggedness γ . Hence, γ determines the subdiffusive dynamics of the protein over many orders of magnitude [5].

A. Simple hierarchical model free-energy landscape

To relate this subdiffusive behavior to ruggedness γ , we used a d -dimensional lattice model inspired by the subdiffusive behavior of the simple one-dimensional model [5]. The near-power-law-type behavior of essential degrees of freedom of many proteins, as illustrated in Fig. 2, suggests using a hierarchy of barriers, the heights γ of which are distributed exponentially,

$$p(\Delta G) = \frac{1}{\gamma} e^{-\Delta G/\gamma}. \quad (1)$$

In this model, shown in Fig. 3(a), the ruggedness γ describes how the height of the barriers ΔG increases with increasing average distance Δx between these barriers. Specifically, for a distance increase by a factor of two, the barrier heights increase by $\gamma/\log(2)$.

Generalizations to d dimensions have been suggested, which may be considered as a model for the high-dimensional protein free-energy landscape, such as in Ref. [6]. Here we rather choose the model proposed in Ref. [7] because we expect it to give more isotropic diffusion than the other. This hierarchical lattice model is a d -dimensional cubic lattice where exponentially distributed barriers heights (see Fig. 3) govern transitions between states. Note that the hierarchical model proposed here is isotropic only on average over disorder. Each specific realization of a hierarchical model is anisotropic.

It has been shown that this model exhibits anomalous diffusion both for one-dimensional and in the limit of high-dimensional models [8].

For the former, the subdiffusion exponent α is

$$\alpha = \frac{2}{1 + \gamma}. \quad (2)$$

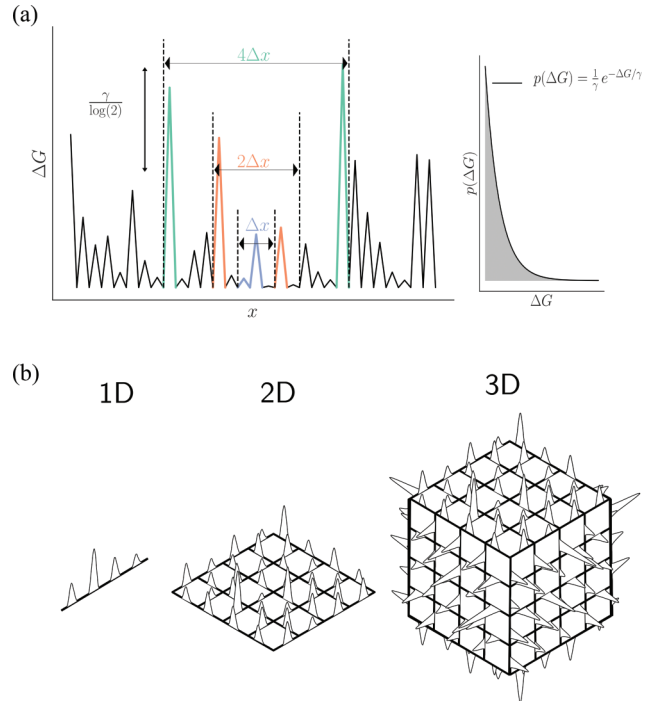


FIG. 3. (a) Sketch of a 1D hierarchical model free-energy landscape, which is characterized by a lattice of free-energy minima with equal free-energy separated by barriers with random heights distributed according to an exponential distribution. With increasing length scales, highest barrier heights increase on average with a characteristic height γ . (b) Sketch of a d -dimensional generalization procedure of a one-dimensional hierarchical model. Whereas the exponential barrier height distribution is kept, the one-dimensional lattice is generalized to d -dimensional lattices (2D and 3D cases are shown).

For high dimensions d , a mean-field approximation [8] yields

$$\alpha = \frac{2d}{\gamma}. \quad (3)$$

This approximation assumes that a random-walk trajectory never visits any state twice, which is strictly only fulfilled for $d \rightarrow \infty$ and is equivalent [8] to continuous-time random walks (CTRW) [13]. Because protein dynamics is well described by typically tens or hundreds of collective coordinates [14], we assumed that neither approximation is sufficiently accurate in this intermediate range and therefore resorted to studying this model numerically.

B. Determining anomalous diffusion exponents from MD trajectories

To this end, we followed common practice in protein dynamics simulations and calculated anomalous diffusion exponents from trajectory length-dependent principal component analysis (tPCA) [9,10]. This approach differs from traditional analyses of subdiffusion in statistical free-energy landscapes in that anomalous diffusion exponents are determined from the trajectory length dependence of the

eigenvalues of the time-averaged covariance matrix,

$$C_{ij}(T) = \frac{1}{T} \left\langle \int_0^T dt [\mathbf{x}_i(t) - \boldsymbol{\mu}_i] \cdot [\mathbf{x}_j(t) - \boldsymbol{\mu}_j] \right\rangle_{\text{ens}}, \quad (4)$$

rather than time dependence of the ensemble-averaged covariance matrix

$$C_{ij}(t) = \langle [\mathbf{x}_i(t) - \boldsymbol{\mu}_i] \cdot [\mathbf{x}_j(t) - \boldsymbol{\mu}_j] \rangle_{\text{ens}}. \quad (5)$$

In the above two equations (4) and (5), \mathbf{x}_i denotes the $3N$ Cartesian coordinates of N selected atoms of a protein and $\boldsymbol{\mu}_i = \int_0^T dt \mathbf{x}_i(t)$ as well as $\boldsymbol{\mu}_i = \langle \mathbf{x}_i(t) \rangle_{\text{ens}}$ their corresponding means. Note that \mathbf{C} is by construction a nonnegative symmetric matrix and is therefore diagonalizable. Its nonnegative eigenvalues (“PCA eigenvalues”) λ_i represent the variance of a trajectory along corresponding eigenvectors \mathbf{v}_i (“PCA eigenvectors”), which represent collective motions. If λ_i follows a power law

$$\lambda_i(T) \propto T^{\alpha_i}, \quad (6)$$

then we define anomalous diffusion exponents as α_i . For Brownian motion in the limit of high-dimensional spaces, it has been shown that this definition is equivalent to the conventional one derived from Eq. (5) [15].

In the next section, we show that this equivalence also holds for anomalous diffusion in high-dimensional hierarchical lattice models. To that end we first show that the hierarchical lattice model is equivalent to the well-known quenched trap model [16] in the limit of high dimensions.

III. SCALING OF PCA EIGENVALUES FOR HIGH-DIMENSIONAL HIERARCHICAL MODELS

In the limit of high dimensionality $d \rightarrow \infty$, the probability of crossing a particular barrier vanishes as the number of possible transitions $2d$ becomes infinite. Hence, the probability of returning to a previously visited state which requires crossing a particular barrier vanishes for infinite-dimensional hierarchical models.

Due to the vanishing probability of returning to a previously visited state, random walks in high-dimensional models are equivalent to random walks where after every step a new set of $2d$ free-energy barriers is encountered. Such random walks are equivalent to continuous time random walks [13], for which energy barriers determine the waiting time t in the current state until a barrier crossing event occurs. In the following we derive a scaling relation for PCA eigenvalues of random walks that is equivalent to the scaling relation of the well-studied quenched trap model [16]. To that end, we decompose the CTRW into two parts, a random walk which determines the states visited and a waiting time in each of these states.

It has been shown [15] that PCA eigenvalues of random walks scale linearly with the number of steps n ,

$$\lambda(T) \propto n(T). \quad (7)$$

The total time $T = \sum_{i=0}^n t_i$ is given by the sum of all waiting times in the individual states, which are random variables. The waiting time distribution $p(t)$ within a state depends on ruggedness γ and dimensionality d and is obtained directly by

a probability transformation of the barrier height distribution which yields

$$p(t) \propto t^{1-\frac{\gamma}{2d}}. \quad (8)$$

For $\gamma/d < 1$, waiting times t_i have a well-defined expectation value, such that for $n \rightarrow \infty$

$$T \approx n \langle t_i \rangle, \quad (9)$$

$$\lambda(T) \propto \frac{T}{\langle t_i \rangle}, \quad (10)$$

which leads to normal diffusion behavior. Anomalous diffusion behavior emerges for $\gamma/2d > 1$, because the expectation value for the waiting times $\langle t \rangle$ diverges for $n \rightarrow \infty$. However, for a finite number of transitions n , this expectation value is also finite and depends asymptotically on T as

$$\langle t_i \rangle \propto T^{1-\frac{2d}{\gamma}}. \quad (11)$$

Inserting this result into Eq. (9) yields the scaling behavior of PCA eigenvalues in the approximation of high-dimensional random walks

$$\lambda(T) \propto T^{\frac{2d}{\gamma}}. \quad (12)$$

For $\gamma/2d > 1$ the expectation value for the waiting times diverges.

IV. METHODS

A. Trajectory length-dependent principal component analysis

Anomalous diffusion exponents α_i were determined in a similar way from both MD trajectories and random-walk trajectories in hierarchical model free-energy landscapes. To estimate α_i via Eqs. (4) and (6), each trajectory of length T_0 was split into 100 windows (nT_0) of length T . On each of these windows, PCA was performed and the resulting set of PCA eigenvalues was averaged. Ten different time window lengths T were chosen, distributed exponentially from 100 ps to 300 ns, such that the overlap of consecutive windows was below 10% to maximize information content. See the supplementary material [17] for an example distribution of the PCA eigenvalues for different time window lengths T . For these data, anomalous diffusion exponents α_i were estimated from the slopes of linear least-squares fits to the logarithm of window length T and PCA eigenvalues λ_i .

In case of random-walk trajectories, the number n of PCA eigenvalues that are sufficiently large depends on the length of a random walk or trajectory. Due to the limited sampling in random walks at hand, only the $n = 6$ largest PCA eigenvalues were larger than 10^{-5} for the smallest time windows, which allowed for numerically stable fits. We, therefore, used only the first six PCA eigenvalues to determine anomalous diffusion exponents in all of the generated random-walk trajectories.

B. Random-walk generation

We generated 40 000 random walks, each in a separately generated but static hierarchical free-energy landscape on a d -dimensional grid. Random walks were generated using the Gilliespie algorithm [18] with parameter ranges summarized

TABLE I. Parameter range of random-walk simulations of the hierarchical free-energy landscape model.

T_0	$10^{5..12}$ (trajectory length)
d	3..200
γ/d	3..30 (kT)

in Table I. This algorithm uses Kramers rule [19] to calculate transition rates r from free-energy barrier heights ΔG as

$$r = A \exp(-\Delta G/kT), \quad (13)$$

where T is the temperature, from which a time-step length t , following an exponential distribution p ,

$$p dt = k \exp(-kt) dt, \quad (14)$$

is randomly chosen. The prefactor A depends on the shape of the barriers and wells but can be absorbed into the barrier height ΔG . Therefore, in our random-walk simulations, we set this prefactor to 1. This also fixes the lattice constant. We also set diffusion constant and time increments dt to 1. These choices do not affect the scaling exponents as the absolute time and length scales determined by those quantities do not contribute to the scaling behavior.

Due to the high dimensionality of the energy landscapes, sections of these were generated dynamically on demand. To this end, for each visited state, adjacent barriers to previously visited states were recovered from memory, whereas new adjacent barriers were chosen randomly from the exponential distribution $p(\Delta G) = \frac{1}{\gamma} e^{-\Delta G/\gamma}$ and stored [Eq. (1) and Fig. 1]. The next barrier crossing was chosen as described in the Gillespie algorithm [18] and the time was advanced accordingly. The temperature was set to 1 such that barrier heights and ruggedness γ is given in units of kT in the following. For intermediate-dimensional models ($d > 10$) with high ruggedness ($\gamma/d > 20$ kT), an enhanced sampling algorithm (see Appendix) was used to generate random-walk trajectories that were long enough to sample a sufficiently large PCA subspace.

C. Estimation of the dependence of anomalous diffusion exponents on ruggedness

To determine which functional form f describes best how the obtained anomalous diffusion exponents α_i decrease with ruggedness γ , we considered three different plausible functional forms, each pointing to a different underlying process—an exponential $\alpha = f_1(\gamma/d) = \beta_1 \exp[-(\gamma/d)/\beta_2]$, a power law $\alpha = f_2(\gamma/d) = \beta_1 / (\gamma/d)^{\beta_2}$, and a linear $\alpha = f_3(\gamma/d) = \beta_1 \gamma/d + \beta_2$ dependence. Here we have used the argument γ/d rather than γ , because, as will be shown in the results part, the data suggested normalizing the ruggedness γ by dimension d . To assess which of the three functions fits the data best, we calculated the posterior probability $P(f, \beta_1, \beta_2 | \{\alpha_i\})$ for each of these functional forms f , given $\{\alpha_i\}$, via a Bayesian approach,

$$P(f, \beta_1, \beta_2 | \{\alpha_i\}) \propto P(\{\alpha_i\} | f, \beta_1, \beta_2) P(f, \beta_1, \beta_2).$$

Here the likelihood $P(\{\alpha_i\} | f, \beta_1, \beta_2)$ of observing a set of scaling exponents at a given ruggedness value was described

by a Gaussian distribution

$$P(\{\alpha_i\} | f, \beta_1, \beta_2) \propto \exp \left\{ - \sum_{\gamma/d} \sum_i \frac{[\alpha_i - f(\gamma/d)]^2}{\sigma^2(\gamma/d)} \right\}, \quad (15)$$

and $f(\gamma/d)$ was chosen as the linear, exponential or power-law function as described above. A constant prior $P(f, \beta_1, \beta_2)$ was assumed for each parameter of the likelihood function. The variance $\sigma^2(\gamma/d)$ of the Gaussian probability distributions was approximated by $\sigma^2(\gamma/d) = a \cdot \gamma/d + b$, because the observed distributions of anomalous diffusion exponents are well described by a Gaussian distribution, as will be shown in Sec. VA. We also observed an increase of variance in anomalous diffusion exponents with increasing normalized ruggedness γ/d . To determine the two parameters a and b , we generated 1000 trajectories each for $\gamma/d \in 5, 15, 17$, and 20 kT, respectively, and calculated the respective variances. To these, the above linear function was fitted. Posterior probabilities for the function and their two parameters were determined using Gibbs sampling [20] with 100 000 steps. As starting points we used a least-squares fit of the respective function to the observed anomalous diffusion exponents. We discarded the first 10 000 steps from the sampling to reduce the influence of the starting parameters.

D. Ruggedness and dimensionality estimates

We proceeded in two steps to estimate ruggedness and dimensionality for a given α_i in the absence of an analytical expression. First, we estimated the likelihood of observing anomalous diffusion exponents in random-walk trajectories which were generated in models with given ruggedness and dimensionality.

To that end, an eight-dimensional joint kernel density using a standard kernel density estimator [21] was estimated from ruggedness γ and dimensionality d parameters as well as 6 PCA eigenvalue anomalous diffusion exponents α_i . A multivariate Gaussian kernel is placed on each observed data point, and the normalized sum of all kernels serves as an estimate for the probability density of observing a set of anomalous diffusion at given ruggedness and dimensionality values. The bandwidth of the Gaussian kernels is equal for all data points and determined such that the variance of the resulting kernel density has the same value as the sample variance of the data points.

In a second step, a likelihood $p(\gamma/d, d | \{\alpha_i\})$ was estimated for a range of ruggedness (4 – 30 kT/ d with 1 kT steps) and dimensionality values (3 – 200 dimensions with one dimension steps) from the joint kernel density as the marginal at the values of scaling exponents $\{\alpha_i\}$ (here we used the scaling exponents of the first 6 PCA eigenvalues). We estimate dimensionality and normalized ruggedness from the marginal distribution using the most likely combination of the considered parameter ranges.

E. Protein selection

The 500 proteins were selected using the protocol found in Ref. [22]. In this protocol, nonhomologous proteins were

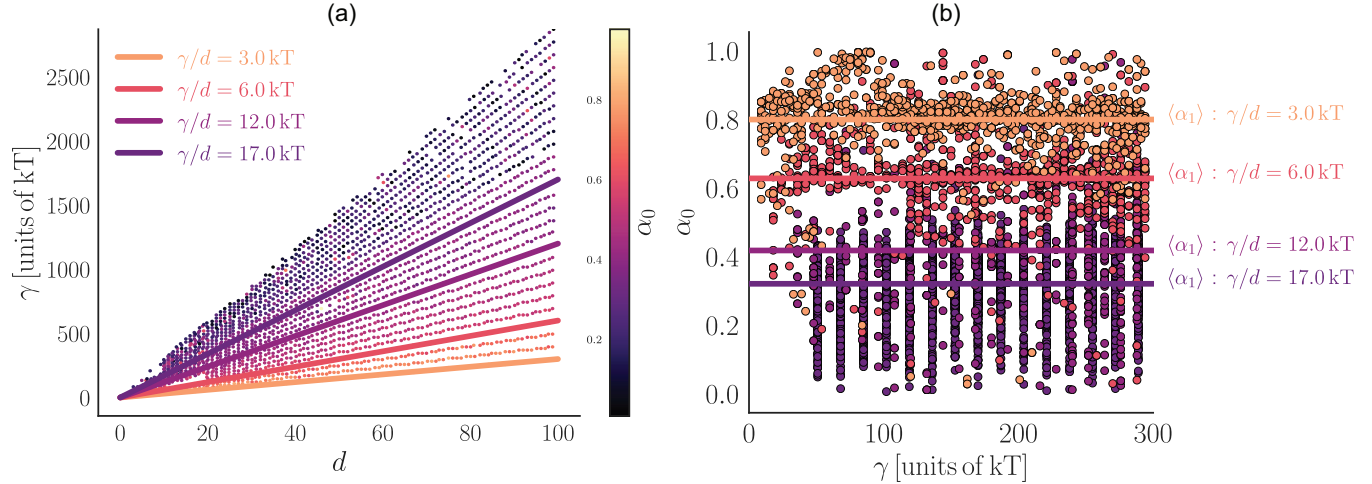


FIG. 4. (a) Dependence of the scaling exponent of the largest PCA eigenvalue α_1 on ruggedness γ/d and dimensionality d . The color of each point represents an average scaling exponent averaged over all simulations with corresponding ruggedness and dimensionality values. (b) Dependence of anomalous diffusion exponents on ruggedness γ for different ratios of γ/d . Because an exhaustive sampling for all possible parameter combinations was not possible, we generated for $\gamma/d = 17$ kT multiple random walks for specific ruggedness and dimensionality values to sample the distribution of α_1 rather than generating trajectories equally distributed over the whole range of ruggedness and dimensionality values. The results of these simulations appear as “vertical” lines in the plot.

selected from the protein data bank (PDB) [23] such that a large range of small globular proteins with less than 90% sequence identity was retrieved. Only monomeric structures without gaps consisting of only standard residues were used. From the remaining protein structures, those containing polymeric or nonconstitutive ligands were excluded. See Supplemental Material [17] for a distribution of the sizes of the selected proteins.

Among the 500 selected proteins, 100 enzymes and nonenzymes using the classification available on the protein data bank were selected to perform three additional 1- μ s MD simulations.

F. Generation of MD trajectories

For each of the 500 selected proteins, MD simulations were performed using the simulation package software GROMACS 2018 [24]. Starting structures were obtained as described above in Sec. IV E. Solvent (TIP4P-Ew water model [25]) and ions (Na^+ and Cl^-) were added, establishing a salt concentration of 0.15 mol l^{-1} and neutralizing the overall system charge. A triclinic box with periodic boundary conditions was used with a 1.5-nm distance between solute and box boundary. Prior to each simulation run, energy minimization was performed using the GROMACS steepest descent algorithm until convergence was reached. This energy minimization was followed by a 1-ns (NPT) MD simulation to equilibrate the system. After energy minimization and equilibration a 1- μ s MD trajectory was generated for each protein using Amber99*ildn force field [26] with a 2.5-fs time step with virtual sites [27]. All bond lengths were constrained, using the Settle algorithm [28] for the solvent and Lincs algorithm [29] for the solute, with a Lincs order of 4 during energy minimization and equilibration and 6 in the production run. Van-der-Waals forces were ignored for distances > 1 nm and Coulomb forces were calculated using the particle mesh Ewald method [30]

with a real-space cutoff of 1 nm, PME order of four, and a Fourier grid spacing of 1.2 Å.

For 200 of the selected proteins comprising 100 enzymes and 100 nonenzymes, three additional microsecond trajectories were calculated following the same protocol to estimate the statistical uncertainty of the determined ruggedness and dimensionality.

V. RESULTS AND DISCUSSION

A. Anomalous diffusion in intermediate-dimensional hierarchical models

Using random-walk trajectories, generated as described in Sec. IV B, we first determined how anomalous diffusion exponents depend on the ruggedness and dimensionality of the hierarchical lattice model shown in Fig. 4. Almost normal diffusion ($\alpha = 1$) is seen for small ruggedness parameters γ , with increasingly strong subdiffusion ($\alpha < 0.1$) for larger γ as shown in Fig. 4(a). Notably, similar α are seen for regions of similar γ/d ratios as shown in Fig. 4(b).

For an explanation of this behavior, note that a trajectory is dominated by crossings of the lowest barriers ΔG_{\min} . For the hierarchical lattice model, it follows from Eq. (1) that for each visited state, the lowest of the $2d$ adjacent barriers is distributed as

$$p(\Delta G_{\min}) = \frac{2d}{\gamma} \exp\left(-\frac{2d \Delta G_{\min}}{\gamma}\right). \quad (16)$$

As this distribution is a function of γ/d , similar anomalous diffusion exponents are expected for equal ratios of γ/d . This idea also explains why strong subdiffusion is only observed for unexpectedly high γ , particularly for large dimensionalities d . This finding motivates the use of this ratio or “normalized” ruggedness as an argument for the functional dependence $\alpha(\gamma/d)$ further below.

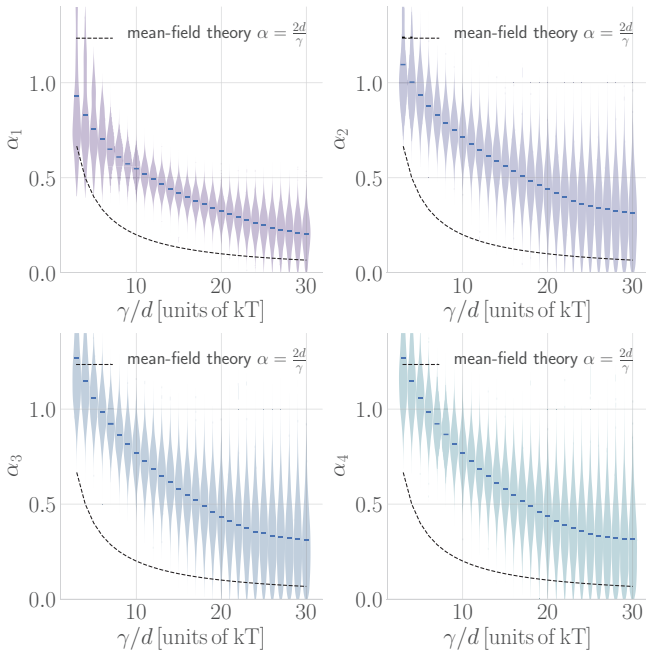


FIG. 5. Distributions of scaling exponents of the first four PCA eigenvalues α_1 (top left), α_2 (top right), α_3 (bottom left), and α_4 (bottom right) in dependence of ruggedness γ/d . Probabilities of observing an anomalous diffusion exponent at a given ruggedness value are represented by violins. The dashed black line indicates the mean-field approximation.

Next, we compared our numerically obtained anomalous diffusion exponents α to the mean-field approximation and asked how accurately γ/d can be estimated from α . To that aim, Fig. 5 shows, as a violin plot for the largest four PCA eigenvalues, how much scaling exponents α scatter when derived from single trajectories for different landscapes as a function of γ/d . Note that the considerable width of these distributions results not only from the stochastic nature of the individual trajectories and the underlying energy landscapes

but also from their different ruggedness and dimension for given γ/d . As expected, increasing subdiffusion is seen for increasing γ/d . For the smaller eigenvalues and small γ/d , some superdiffusion is seen as was already explained in terms of ballistic motion [15]. Overall, much weaker subdiffusion is seen compared to the mean field approximation (dashed lines), with decreasing discrepancy for larger dimensions, as also expected. Scaling exponents of large PCA eigenvalues decrease faster with increasing γ/d and show a lower variance, mainly due to better sampling of these coordinates. Notably, the shown scatters generally exhibit large overlaps for adjacent γ/d values, particularly for larger γ/d , which suggests that reconstructions of ruggedness and dimension from subdiffusion exponents α involve considerable uncertainties. These will be explored further below.

As our observed anomalous diffusion exponents deviate considerably from the functional relation 3 derived in a mean-field approximation (black dashed line in Fig. 5), we asked which function $f(\gamma/d)$ describes the observed mean anomalous diffusion exponents best. As functional forms, we considered a power law (see caption of Fig. 6) as the generalization of the mean-field result, an exponential decay as a plausible alternative, and a linear function for comparison. For these three functional forms, posterior probabilities obtained via Gibbs sampling (see methods) are shown in Fig. 6(a). As an example, Fig. 6(b) shows the posterior distributions for the respective parameters (β_1, β_2) for the exponential function. As can be seen in the figure, the highest posterior probability is obtained for the exponential function and in that sense describes our numerical results the best among the three considered functions. Remarkably, the power law, for which the mean-field theory Eq. (3) is a special case for $\beta_1 = 2, \beta_2 = 1$, turns out to be the least probable, which suggests that a simple modification of the mean-field theory will most likely not suffice for a quantitative explanation of the anomalous diffusion exponents. We conclude that the underlying assumption that no trajectory visits any state twice most likely does not provide a good approximation for intermediate-dimensional models.

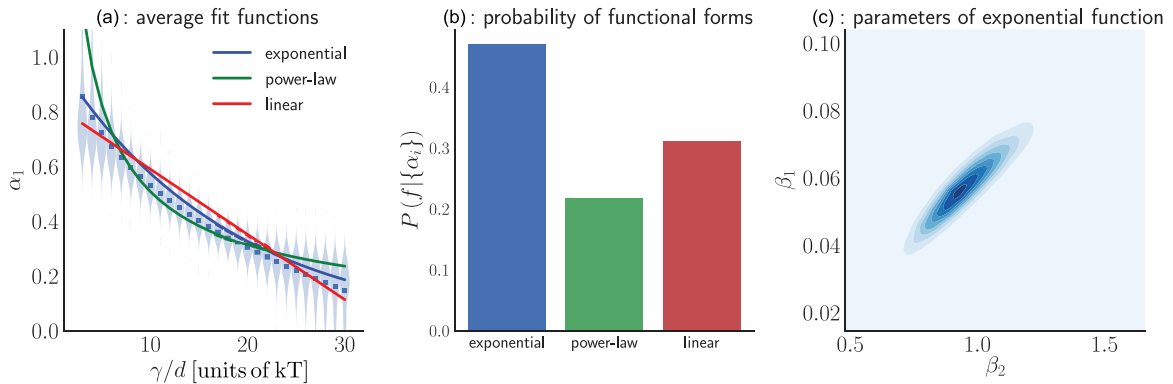


FIG. 6. (a) Distributions of anomalous diffusion exponents (raw data) shown as violin histograms depending on γ . Average scaling exponents are shown as thick blue lines. Colored lines show the average fit function for three selected functional forms, i.e., exponential (blue), power law (green) and linear (red), of the relation between mean scaling exponent (α_1) of the first PCA eigenvalue and γ/d based on the scaling exponent distributions of generated random walks. (b) Box plots of posterior probabilities of exponential $\alpha = f_1(\gamma/d) = \beta_1 \exp[-(\gamma/d)/\beta_2]$ (blue), power law $\alpha = f_2(\gamma/d) = \beta_1 / (\gamma/d)^{\beta_2}$ (green), and linear $\alpha = f_3(\gamma/d) = \beta_1 \gamma/d + \beta_2$ (red) functions obtained from a Gibbs sampling. (c) Kernel density plot of the distribution of posterior probability of the two parameters of the exponential function.

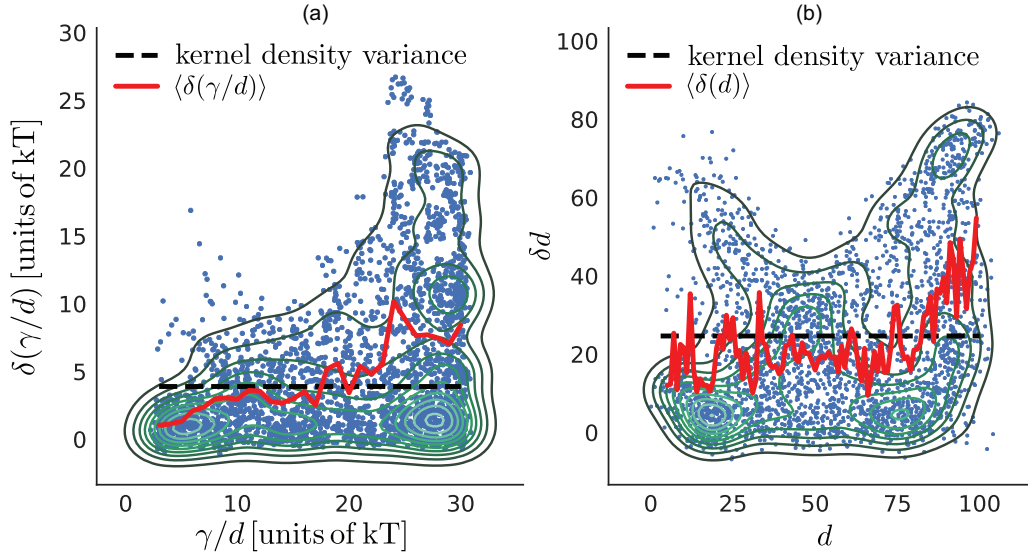


FIG. 7. Error estimates of ruggedness and dimensionality using cross validation. For each ruggedness value, 100 random-walk trajectories were randomly chosen, from which ruggedness and dimensionality of the corresponding intermediate-dimensional hierarchical model were estimated using a maximum likelihood estimator. (a) Dependence of the error in ruggedness estimates on ruggedness γ/d . Errors in ruggedness estimates for individual random-walk simulations are shown as blue dots; their density is shown as green contours. The mean error (red line) of the ruggedness estimates increases with higher ruggedness values, but is similar to the error estimated from the kernel density (black dashed line). (b) Dependence of the error in dimensionality estimates on dimensionality. Colors and symbols are as in (a). The mean error in dimensionality estimates shows no clear dependence on dimensionality. Gaussian noise with 1% variance was added to shown points in both plots for visualization purposes.

However, although the exponential function fits the data best, it is only a slightly better description of the numerical anomalous diffusion exponents (see colored lines in Fig. 6). Therefore, we did not use any of these three functions to extract ruggedness and dimensionality from the anomalous diffusion exponents extracted from protein MD simulations further below, but rather resort to a probabilistic approach. To that end, we used a kernel density estimator to model the joint probability density of anomalous diffusion exponents $\alpha_1, \dots, \alpha_4, \gamma/d$, and d as observed in a total of 40 000 random walks (see Sec. IV A). This probability density will serve to obtain distributions of γ/d and d for given $\alpha_1, \dots, \alpha_4$ by marginalization.

Before discussing the obtained γ/d and d , we used 2200 trajectories to estimate the uncertainty of these values by two independent approaches, from the variance of the marginalized distribution of γ/d and d via cross validation. Figure 7 shows for each of the trajectories (blue dots) the actual error of the estimate, i.e., the deviation of the estimated normalized ruggedness [Fig. 7(a)] and dimensionality [Fig. 7(b)] from their known values that were used to build the respective underlying energy landscapes. For comparison, the average error estimate obtained from the marginalized distributions is shown as a black dashed line. In addition, the red line shows the cross validation in terms of the average actual error for 100 trajectories with the same ruggedness (or dimensionality) values, which have not been used for the training of the kernel density. We obtained an overall mean error of 4.2 kT for the ruggedness estimate and 10 dimensions for the dimensionality estimate. The mean relative error of ruggedness estimation is moderate and increases with increasing ruggedness as expected, whereas the mean relative error of dimensionality

estimates is substantially larger and independent of dimensionality. Overall, the hierarchical lattice model suggests that it should be possible to estimate the ruggedness of proteins rather reliably from anomalous diffusion exponents that were obtained via trajectory length-dependent principal component analysis of atomistic simulations.

B. Anomalous diffusion in realistic protein free-energy landscapes

To this end, we used the above probabilistic model to explore ruggedness and dimensionality of the free-energy landscapes of 500 small globular proteins selected to cover known folds and functions as described under Sec. IV E. We carried out a 1- μ s MD simulation for each of these 500 proteins and performed a tPCA for each of the trajectories as described under Sec. IV A. From least-squares fits to the trajectory length-dependent largest eigenvalue scaling exponents α_1 were obtained. Figure 8(a) shows the distribution of scaling exponents jointly as a function of protein size (number of C_α atoms N) as described by the dimension $d_{\text{conf}} = 3N$ by the respective configurational space. As can be seen, almost all of the obtained scaling exponents show subdiffusion ($\alpha_1 < 1$). In fact, ca. 90% of the scaling exponents are smaller than 0.6 and the mean anomalous diffusion exponent is 0.3. Remarkably, no significant correlation between α_1 and configuration space dimensionality d_{conf} is seen (Pearson correlation coefficient $c = 0.12$).

Using these anomalous diffusion exponents, we estimated the ruggedness γ and effective dimensionality d of the 500 protein free-energy landscapes via the above probabilistic model. Figure 8(c) shows the distribution of dimensionality

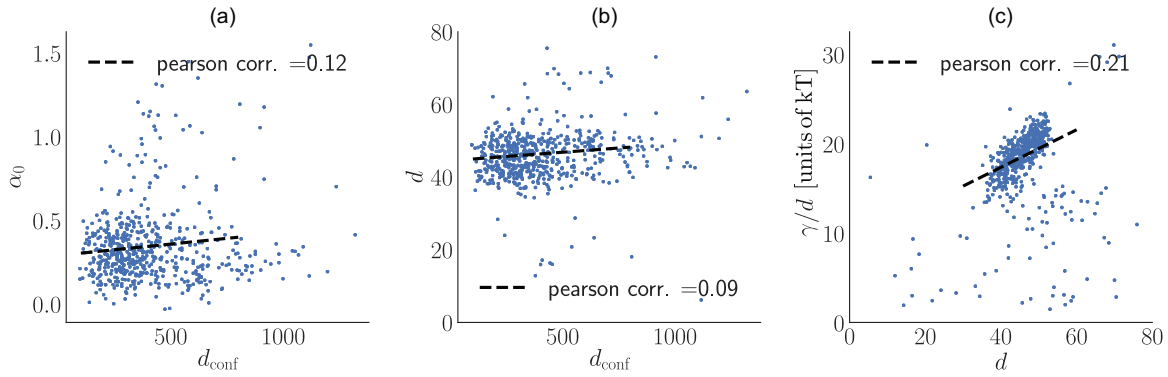


FIG. 8. (a) Subdiffusion exponents α_1 as a function of configuration space dimensionality obtained from MD trajectories of 500 small globular proteins with configuration space dimensionality $d_{\text{conf}} = 3N$, where N is the number of C_α atoms. (b) Estimated effective dimensionality d and configuration space dimensionality. (c) Frequency of dimensionality and ruggedness estimates of 500 small globular proteins obtained from microsecond molecular dynamics simulation.

and ruggedness estimates of γ/d and d , where the ruggedness has been normalized by the effective dimension as suggested for the simple hierarchical grid model and shown in Fig. 4.

As can be seen, ruggedness values between 15 and 20 kT per dimension dominate, as well as effective dimensionalities d between 40 and 60. This result is reproducible for four independent sets of MD simulations of a test set of 200 proteins as described under Sec. IV E. For this test set, an average standard deviation of 1.1 kT for ruggedness and 4.8 dimensions for the dimensionality was obtained, which is lower than the expected error from our random-walk simulations. We attribute these low errors to the fact that MD simulations were started from the same starting structure and therefore explore similar regions in their free-energy landscape, whereas in the case of the hierarchical lattice model, each trajectory is simulated in a different free-energy landscape.

The high estimated ruggedness values suggest that most free-energy barriers are not crossed. Indeed, Fig. 9 shows that

the crossed barriers in random walks in hierarchical model free energy landscapes with protein-typical ruggedness and dimensionality values range between 5 and 6 kT. Values in this range were also observed in flash photolysis experiments on myoglobin.

Unexpectedly, despite the fact that there is no correlation between the ruggedness coefficient and the protein size, Fig. 8(c) shows a strong correlation between normalized ruggedness γ/d and effective dimensionality d (Pearson correlation coefficient $c = 0.21$). We asked if this correlation is due to a possible correlation between protein size and effective dimensionality. However, no such correlation is seen in the respective scatter plot [Fig. 8(b), Pearson correlation coefficient $c = 0.09$]. Taken together, these results suggest that both the effective dimensionality and normalized ruggedness of a protein do not depend on its size and rather are adapted to the particular function of each single protein. Furthermore, it is remarkable that the ranges of both normalized ruggedness

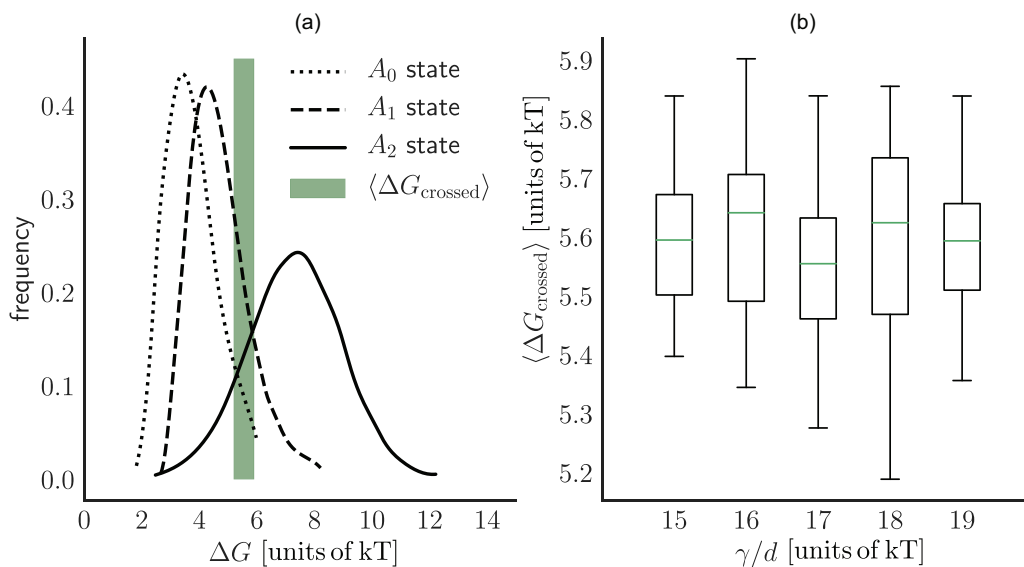


FIG. 9. (a) Distributions of barrier heights of three states of myoglobin (A_0 , A_1 , A_2) estimated from flash photolysis experiments [12] (b). Boxplots of distributions of barrier heights $\delta G_{\text{crossed}}$ that were crossed in random walks in intermediate-dimensional hierarchical lattice models with typical ruggedness and dimensionality as estimated for the selected proteins (40 – 60 and 15 – 20 kT per dimension).

and effective dimensionality (by a factor of about 1.5) are much smaller than the scatter of protein sizes (by a factor of ca. 5) among the selected 500 proteins. This finding suggests that, quite generally, these narrow ranges are optimal for the function of essentially every protein.

VI. CONCLUSION

In this work, we developed a method to connect simple hierarchical free-energy lattice models to atomistic simulations of biological macromolecules. To this end, we have characterized the high-dimensional free-energy landscape of 500 small globular proteins in terms of effective dimensionality and distribution of free-energy barrier heights. These quantities have been obtained from anomalous diffusion exponents observed in microsecond molecular dynamics trajectories of these proteins.

For the hierarchical free-energy lattice model, we assumed an exponential distribution $p(\Delta G) \propto \exp(-\Delta G/\gamma)$ of static barrier heights, where γ denotes the “ruggedness” of the energy landscape, similar to a disorder temperature. While analytic expressions have been derived for one-dimensional [6] and high-dimensional lattices [8], we are not aware of any result for the intermediate effective dimensions relevant for biomolecules; we therefore resorted to a numerical approach.

To this end, we carried out random-walk simulations and found indeed anomalous diffusion exponents that deviate from both the one-dimensional and high-dimensional limiting cases. A Bayesian analysis showed that, overall, anomalous diffusion exponents decrease less steeply with increasing ruggedness and most likely not by the inverse-law of the two limiting cases. These significant deviations suggest different mechanisms from which anomalous diffusion behavior arises.

In the limit of infinite dimensions, random walks are equivalent to high-dimensional CTRW, for which no state is visited more than once and therefore the mean-field description is accurate. In particular, the probability of returning to a previously visited state approaches zero. As a result, anomalous diffusion in CTRWs is caused by waiting time distributions that are heavy-tailed (i.e., diverging averages) [31] and dominated by a few extreme waiting times in states enclosed by high barriers. In contrast, for the intermediate-dimensional models discussed here, recrossings were observed with consequences, the analysis of which is beyond the scope of this work. Based on our observations (data not shown), we speculate that the anomalous diffusion in intermediate-dimensional models most likely originates—similarly to the 1D case—from high free-energy barriers confining random-walk trajectories to a region in conformational space. However, anomalous diffusion in intermediate-dimensional models differs from that of 1D models in that, due to the higher dimensionality of these regions, high free-energy barriers are circumvented. This effect results in a fractallike topology of the subregion actually accessed by trajectories [9]. This scenario is also supported by our observation that the height distribution of actually crossed barriers is much lower than the overall barrier distribution of the free-energy landscape.

In that sense, this study reveals a connection of the two main conceptual frameworks explaining anomalous diffusion in protein dynamics, diffusion on fractal geometries and hier-

archical free-energy landscapes. Specifically, we have shown that, for intermediate dimensionality, fractallike topologies of accessible configurational space arise necessarily from dynamics in hierarchical energy landscapes with very high ruggedness.

Using our numerical results, we asked how accurately ruggedness and dimensionality can be estimated based on anomalous diffusion exponents obtained from nonequilibrium trajectories. For the hierarchical lattice model, we showed via cross validation that a maximum likelihood estimate yields an accuracy of 4.2 kT for ruggedness and 10 dimensions for the effective dimensionality.

This result enabled us to use our method to estimate ruggedness and dimensionality based on anomalous diffusion exponents which we calculated from MD trajectories of a total of 500 small globular proteins. We obtained typical ruggedness estimates in the range of 15 – 20 kT per dimension and effective dimensionality of 40–60. The robustness of the ruggedness and dimensionality estimates for three independent MD simulations shows that the intermediate-dimensional hierarchical model indeed captures features of protein free-energy landscapes that govern the anomalous diffusion behavior in protein dynamics. The range of heights of crossed barriers in hierarchical models with protein-typical ruggedness and dimensionality values agrees with experimentally observed barrier distributions. See the Supplemental Material [17] for an example of ruggedness and dimensionality estimates of a 50- μ s trajectory, indicating that our results are also robust with respect to increasing trajectory lengths.

There are two main assumptions in applying our hierarchical model to protein dynamics. The first assumption is that the free energies of all minima are equal, such that, except for the barriers, the protein free-energy landscape is essentially flat. This assumption only holds for the short timescale protein dynamics studied here by atomistic simulations, and also only for the essential and slow degrees of freedom identified via PCA. For longer timescales, the “walls” of the folding funnel, which stabilizes the folded structure of a protein, restrict the dynamics, which would result in a levelling off of the largest PCA eigenvalues for these long timescales. Because we did not observe such this effect for the largest PCA eigenvalues that we used to compare with the hierarchical model (and only for smaller PCA eigenvalues of the smallest considered proteins), we think that, for the microseconds dynamics studied here, this assumption is justified. On much longer timescales, where the folding funnel does affect protein dynamics along large PCA modes, the levelling off of the PCA values would result in a smaller anomalous diffusion exponent and, hence, an overestimate of the ruggedness of proteins.

The second assumption is that the protein dynamics within the essential conformational subspace that is spanned by the largest eigenvectors is governed by a hierarchical free-energy landscapes that has, on average, an isotropic barrier height distribution, whereas it would certainly be possible to include anisotropic barrier height distributions, we do not consider it necessary. The reason is that anisotropic barrier height distributions would imply different diffusion exponents for different principal components, which we did not observe at the level of sensitivity achieved by our analysis of our microseconds atomistic simulations.

It is remarkable that neither the effective dimensionality nor the ruggedness correlates with protein size, whereas there is a significant correlation between effective dimension and ruggedness. Further, the ranges of both normalized ruggedness and effective dimensionality are much smaller than the scatter of protein sizes (by factors of about 1.5 and 5, respectively) among the selected 500 proteins.

In fact, the broad range of different protein structures, secondary structure arrangements, and architectures with quite different dynamic properties that is reflected even in the set of 500 small globular proteins we have selected for our analysis, should imply quite different protein dynamics for different proteins. Accordingly, one would expect that particularly their ruggedness and effective dimensionalities show correspondingly broad distributions—which, however, are not seen. Similarly, no correlation between normalized ruggedness and effective dimensionality would be expected *a priori*. These findings suggest, in our view as the most likely explanation, an evolutionary pressure on these observables. In particular, it is plausible that it is evolutionarily favorable for free-energy landscapes of proteins with a higher effective dimensionality to also exhibit a higher effective ruggedness to achieve tightly coordinated dynamics of functionally relevant degrees of freedom. Also, Although the precise functional advantage is unclear to us, we speculate that a more rugged free-energy landscape may kinetically stabilize the structure of larger proteins by slowing down the diffusion process in configuration space, hence rendering the folded state more long-lived and less prone to protein degradation.

Taken together, our results suggest that both, ruggedness and effective dimensionality of protein dynamics free-energy landscapes are adapted to the particular function of each single protein, and that, quite generally, these narrow ranges are optimal for the function of essentially every small globular protein.

APPENDIX: ENHANCED SAMPLING FOR RANDOM WALK GENERATION

For ruggedness values $\gamma/d > 20$ we found that random-walk trajectories are confined to a small number of states such

that only a few degrees of freedom are sampled. In this case, we used an enhanced sampling method, similar to Gilliespie's algorithm [18], to reach timescales to escape from such confinement regions to enable sampling in multiple dimensions. Instead of sampling transitions within a confinement region, we estimate the probability of the number of transitions n until an escape event occurs and use accordingly distributed samples as numbers of transitions. The probability of escaping p_{esc} a confinement region is given by the probability of occupying a boundary state μ_i and the probability p_{ij} of transitioning from state i to a state j which is outside of the confinement region is

$$p_{\text{esc}} = \sum_{i \in \Omega} \mu_i \sum_{j \notin V} p_{ij},$$

where Ω is the set of states at the confinement region boundary, V is the set of all states in the confinement region. If the confinement region is stable enough, the number of transitions before leaving are memoryless,

$$\Pr(n > s + t | n > s) = \Pr(n > t).$$

Therefore, escape attempts are statistically independent events and the number of transitions n within a confinement region until an escape event occurs are exponentially distributed as

$$p(n) \propto (1 - p_{\text{esc}})^{n-1} p_{\text{esc}} \approx e^{\log(1-p_{\text{esc}})n}. \quad (\text{A1})$$

Assuming a sufficient amount of transitions within a confinement region, we estimate μ_i as the fraction of visits of state i and the total amount of transitions. We use random numbers, distributed according to (A1), to estimate the number of transitions within the confinement region. Because time averages are independent of the order in which states are visited, the number of visits to each state $n\mu_i$ is sufficient to calculate the time-averaged covariance matrix. The next state j outside of a confinement region is determined with probability

$$p_j = \frac{\sum_{k \in \Omega} p_{kj}}{\sum_{l \in \bar{\Omega}} \sum_{k \in \Omega} p_{kl}}, \quad (\text{A2})$$

where $\bar{\Omega}$ is the set of states outside of the confinement region with a connection to its boundary.

-
- [1] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Theory of protein folding: The energy landscape perspective, *Annu. Rev. Phys. Chem.* **48**, 545 (1997).
 - [2] H. Frauenfelder, F. Parak, and R. D. Young, Conformational substates in proteins, *Annu. Rev. Biophys. Biophys. Chem.* **17**, 451 (1988).
 - [3] A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, I. E. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young, Protein states and proteinquakes, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 5000 (1985).
 - [4] R. Elber and M. Karplus, Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin, *Science* **235**, 318 (1987).
 - [5] A. Maritan and A. L. Stella, Exact Renormalization Group for Dynamical Phase Transitions in Hierarchical Structures, *Phys. Rev. Lett.* **56**, 1754 (1986).
 - [6] S. Teitel and E. Domany, Dynamical Phase Transitions in Hierarchical Structures, *Phys. Rev. Lett.* **55**, 2176 (1985).
 - [7] K. Kundu and P. Phillips, Hopping transport on site-disordered d-dimensional lattices, *Phys. Rev. A* **35**, 857 (1987).
 - [8] S. Havlin, J. E. Kiefer, and G. H. Weiss, Anomalous diffusion on a random comblike structure, *Phys. Rev. A* **36**, 1403 (1987).
 - [9] T. Neusius, I. Daidone, I. M. Sokolov, and J. C. Smith, Subdiffusion in Peptides Originates from the Fractal-Like Structure of Configuration Space, *Phys. Rev. Lett.* **100**, 188103 (2008).
 - [10] U. Hensen, T. Meyer, J. Haas, R. Rex, G. Vriend, and H. Grubmüller, Exploring protein dynamics space: The dynasome as the missing link between protein structure and function, *PLoS One* **7**, e33931 (2012).
 - [11] C. C. David and D. J. Jacobs, Principal component analysis: A method for determining the essential dynamics

- of proteins, in *Protein Dynamics* (Springer, Berlin, 2014), pp. 193–226
- [12] G. Ulrich Nienhaus, J. D. Müller, B. H. McMahon, and H. Frauenfelder, Exploring the conformational energy landscape of proteins, *Physica D* **107**, 297 (1997).
- [13] E. W. Montroll and G. H. Weiss, Random walks on lattices. ii, *J. Math. Phys.* **6**, 167 (1965).
- [14] I. Daidone and A. Amadei, Essential dynamics: Foundation and applications, *WIREs Comput. Mol. Sci.* **2**, 762 (2012).
- [15] B. Hess, Similarities between principal components of protein dynamics and random diffusion, *Phys. Rev. E* **62**, 8438 (2000).
- [16] J. Bouchaud, A. Comtet, A. Georges, and P. Le Doussal, Anomalous diffusion in random media of any dimensionality, *J. Phys. France* **48**, 1445 (1987).
- [17] See Supplemental Material <http://link.aps.org/supplemental/10.1103/PhysRevE.105.044404> for three figures showing the size distribution of the considered proteins, a distribution of PCA eigenvalues for different time window sizes, and the trajectory length dependence of ruggedness estimates.
- [18] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem.* **81**, 2340 (1977).
- [19] H. Kramers, Brownian motion in a field of force and the diffusion model of chemical reactions, *Physica* **7**, 284 (1940).
- [20] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6**, 721 (1984).
- [21] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley & Sons, New York, 2015).
- [22] T. Meyer, M. D’Abramo, A. Hospital, M. Rueda, C. Ferrer-Costa, A. Pérez, O. Carrillo, J. Camps, C. Fenollosa, D. Repchevsky, J. L. Gelpí, and M. Orozco, Model (molecular dynamics extended library): A database of atomistic molecular dynamics trajectories, *Structure* **18**, 1399 (2010).
- [23] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The protein data bank, *Nucleic Acids Res.* **28**, 235 (2000).
- [24] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX* **1-2**, 19 (2015).
- [25] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew, *J. Chem. Phys.* **120**, 9665 (2004).
- [26] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Improved side-chain torsion potentials for the amber ff99sb protein force field, *Proteins* **78**, 1950 (2010).
- [27] H. Berendsen and W. Van Gunsteren, Molecular dynamics simulations: Techniques and approaches, in *Molecular Liquids* (Springer, Berlin, 1984), pp. 475–500.
- [28] S. Miyamoto and P. A. Kollman, Settle: An analytical version of the shake and rattle algorithm for rigid water models, *J. Comput. Chem.* **13**, 952 (1992).
- [29] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, LINCS: A linear constraint solver for molecular simulations, *J. Comput. Chem.* **18**, 1463 (1997).
- [30] T. E. I. Cheatham, J. L. Miller, T. Fox, T. A. Darden, and P. A. Kollman, Molecular dynamics simulations on solvated biomolecular systems: The particle mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins, *J. Am. Chem. Soc.* **117**, 4193 (1995).
- [31] M. M. Meerschaert and H.-P. Scheffler, Limit theorems for continuous-time random walks with infinite mean waiting times, *J. Appl. Probab.* **41**, 623 (2004).