Coexisting origins of subdiffusion in internal dynamics of proteins

Yasmine Meroz, Victor Ovchinnikov, and Martin Karplus

Phys. Rev. E **95**, 062403 — Published 6 June 2017

DOI: 10.1103/PhysRevE.95.062403

Coexisting Origins of Subdiffusion in Internal Dynamics of Proteins

Yasmine Meroz,[1,*] Victor Ovchinnikov,[2,†] and Martin Karplus*[2,3,‡]

[1] *Harvard University, School of Engineering and Applied Sciences, Cambridge, MA, 02138*
[2] *Harvard University, Department of Chemistry and Chemical Biology, Cambridge, MA, 02138*
[3] *Laboratoire de Chimie Biophysique, ISIS, Université de Strasbourg, 67000 Strasbourg, France*

Subdiffusion in conformational dynamics of proteins is observed both experimentally and in simulations. Although its origin has been attributed to multiple mechanisms, including trapping on a rugged energy landscape, fractional Brownian noise, or a fractal topology of the energy landscape, it is unclear which of these, if any, is most relevant. To obtain insights into the actual mechanism, we introduce an analytically tractable hierachical trapping model, and apply it to molecular dynamics simulation trajectories of several proteins in solution. The analysis of the simulations introduces a subdiffusive exponent that varies with time, and associates plateaus in the mean-squared displacement with traps on the energy landscape. This analysis permits us to separate the component of subdiffusion due to a trapping mechanism from that due to an underlying fluctuating process, such as fractional Brownian motion. The present results thus provide new insights concerning the physical origin of subdiffusion in the dynamics of proteins.

Proteins are dynamic structures [1, 2] whose internal motions evolve on a rugged potential energy suface, with minima separated by barriers of varying heights [3]. The conformational dynamics are *subdiffusive*, characterized by the slow non-exponential relaxation of dynamical observables such as IR spectra [4], fluorescence fluctuations [5], and principal component displacements computed from MD simulations [6]. The mean-squared displacement (MSD) of a subdiffusive process evolves sublinearly in time:

$$\langle r^2(t)\rangle \equiv \langle |\mathbf{r}(t)-\mathbf{r}(0)|^2\rangle \sim t^{\alpha}, \quad 0<\alpha<1, \qquad (1)$$

in contrast to normal diffusion for which α=1. Here $\langle\cdot\rangle$ denotes either an ensemble or temporal average, and $\mathbf{r}$ represents the system coordinates). Protein dynamics has been found to exhibit values of α in the range of $0.1-0.4$ [7, 8]. Although it is generally agreed that subdiffusion can arise from the ruggedness of the protein energy landscape [1–3, 6], the high dimensionality of the latter ($3N-6$ degrees of freedom for a protein with N atoms) makes its detailed characterization difficult [5–10]. Conceptual models of protein subdiffusion generally fall into two categories. The first involves trapping on a rugged energy landscape [5, 6, 10, 11], and the second includes models that are fractal in nature, such as fractional noise [12–16] or a fractal topology of the energy landscape [17, 18]; in the second category, trapping *per se* does not play a role.

The motivation behind trapping models dates back to the studies of Frauenfelder and coworkers [9] on the rebinding of CO to myoglobin (Mb) at low temperatures. Based on their experiments, the authors proposed that Mb is organized into conformational *macrostates* composed of clustered sub-states, and postulated that the energy landscape is organized into a hierarchy of *tiers*, i.e. valleys separated by progressively higher barriers. Further support for the hierarchical landscape model was provided by subsequent analyses of MD simulations using principal components [19], inherent structural basins obtained from quenching of trajectories [10], and conformational transition networks [17, 18]. It has been argued that trapping models are not fully consistent with existing observations of protein dynamics under equilibrium conditions, and non-equilibrium behavior has been proposed [18] . In this communication, we first consider the properties of hierarchical trapping models, and then relate the analysis to molecular dynamics simulations of proteins. The results are used to determine whether the observed subdiffusive behavior requires an intrinsic, possibly fractal mechanism, in addition to trapping.

We performed 1μs-long MD simulations of three proteins to analyze their dynamics. Figure 1a–c illustrates the range of coordinates observed for the three proteins considered here. To make the analysis of MD trajectory data tractable without losing essential details, the evolution of $3N$ atomic positions is projected onto the first two principal components (PCs), which contain most of the slow diffusive dynamics; higher modes tend to correspond to harmonic motions [19–21], and are not treated explicitly [42]. The projected trajectory of the Fre-FAD complex in Fig. 2a illustrates trapping on the energy landscape. The dynamics are transiently confined to different configurational regions with different sizes. The associated 2D free energy landscape in Fig. 2b displays four main minima, which enclose smaller-sized nested minima.

I. HIERARCHICAL TRAPPING MODEL

To model the influence of trapping on the dynamics, we begin by introducing a hierearchy of tiers. The i-th tier is composed of valleys with a characteristic size L_i separated by energy barriers with a characteristic value

*Electronic address: ymeroz@seas.harvard.edu
†Electronic address: ovchinnv@georgetown.edu
‡Electronic address: marci@tammy.harvard.edu

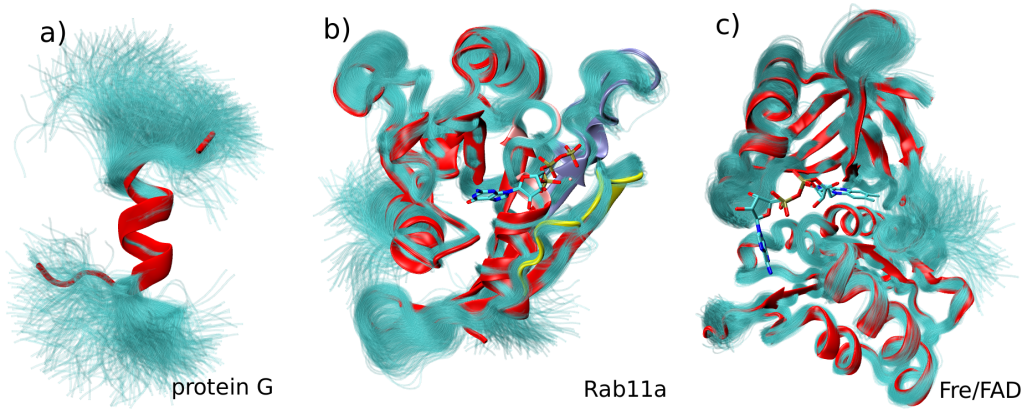

FIG. 1: Illustration of MD trajectories of the proteins considered in this study. (a) protein G [23] modeled in α-helical conformation; (b) Rab11a [24]; (c) Fre-FAD complex [25]. Initial structures are shown in ribbon representation, and backbone structures sampled from MD simulations every 1ns are overlaid as cyan traces.

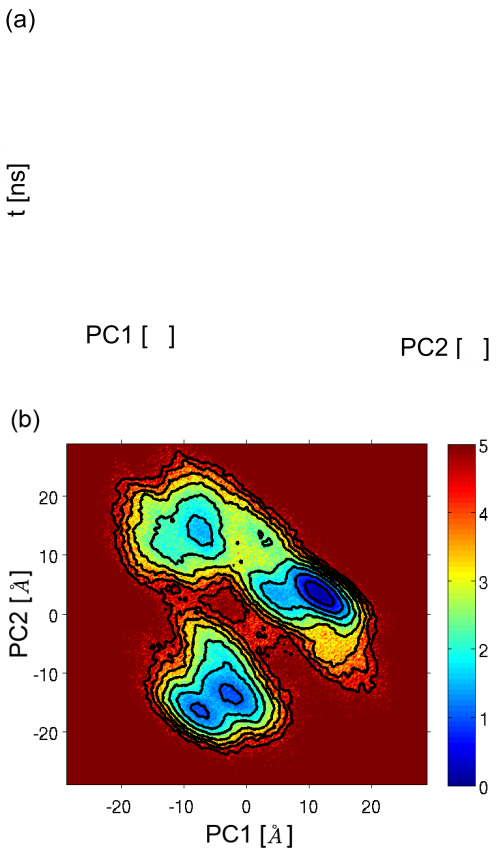

FIG. 2: (a) Evolution of the 1μs MD trajectory of Fre-FAD complex projected onto the first two PCs (see text). The vertical axis corresponds to time, with the markers colored chronologically from blue to red. (b) Free energy landscape of the Fre-FAD complex computed from the last 500ns of the projected MD trajectory shown in Fig. 1(c) as $E/k_BT = -\log H + c$, where k_BT is the thermal energy, $H(\mathbf{r})$ is 2D histogram of trajectory values, and c is an arbitrary constant chosen so that $\min\{E\} = 0$. The contours are drawn corresponding to eight equispaced energy values, at 0.6, 1.2, 1.8, 2.4, 3.0, 3.6, 4.2 and 4.8 k_BT).

E_i. These valleys in turn contain smaller valleys that belong to tier i-1, with $L_{i-1} < L_i$, separated by lower characteristic barriers $E_{i-1} < E_i$, and so on. L_1 and E_1 represent the lowest tier. For simplicity, the potential energy is set to zero everywhere except at the barriers. Because dynamical memory effects are known to vary between proteins, the dynamics on the landscape are left unspecified at this stage. They will be identified from fitting the model to MD simulation data [43]. For illustration, a Brownian dynamics simulation over a two-tier hierarchy with square valleys is presented in Fig. 3.

To show how such a hierarchy can lead to effective subdiffusion, we first consider the case of a single tier with L_1=L and E_1=E. The protein trajectory is assumed to be a random walk (RW) initiated somewhere in the valley, undergoing unhindered diffusion until it reaches the barrier after some characteristic valley *crossing time* τ^c. The (Arrhenius) probability of crossing the barrier is $p \propto \exp[-E]$ (E is nondimensional in units of k_BT, where k_B is Boltzmann's constant and T the temperature). At some characteristic *escape time* τ^e, the RW overcomes the barrier and crosses to a neighboring valley. Assuming that successive escape attempts are uncorrelated, τ^e is approximately

$$\tau^e \simeq \frac{1}{p}\tau^c. \tag{2}$$

For times in the range $\tau^c < t < \tau^e$ the RW is effectively confined to the valley, and the MSD is at a plateau, the value of which is proportional to the square of the valley size L;

$$\langle r^2 \rangle^{\rm plt} \propto L^2, \tag{3}$$

with a proportionality constant that depends on the geometry and dimensionality of the system. In the case of a 2D square lattice, shown in Fig. 3, we have $\langle r^2 \rangle^{\rm plt} \equiv \frac{1}{L^4}\int |x_2 - x_1|^2 dx_1 dx_2 = L^2/3$, which is a Boltzmann average over the valley $x_1, x_2 \in [0, L] \times [0, L]$ (recall that the energy in the valley is assumed to be constant). We

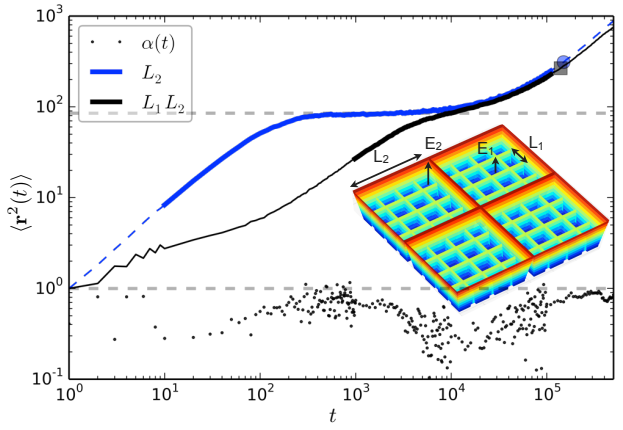

FIG. 3: Brownian dynamics simulations on a model two-tier energy landscape (inset); $L_1 = 4$, $L_2 = 16$, $p_1 = 0.01$ and $p_2 = 0.0001$, where p_i=exp($-E_i$); (black) MSD of a random walk simulation, exhibiting two plateaus associated with each tier; (blue) MSD for a single tier with valleys of size L_2 (e.g. with p_1=1); (dotted) variable exponent $\alpha(t)$ computed using finite differences of $\Delta \log \langle r^2(t) \rangle / \Delta \log t$ whose minima identify the plateaus. Plateau intervals $\tau^c < t < \tau^e$ predicted from random walk theory and indentified from simulations (see Computational Methods) are highlighted in bold. The theoretical plateau location for second tier ($L_2^2/3 \simeq 85$; see text) is drawn as a dashed line.

note that the relation is independent of the nature of the dynamics within the valley; i.e. Eq. (3) is a purely thermodynamic relation.

To relate the temporal scales τ to the spatial scale L, we first recall that for any purely diffusing system, the average length l_c of a classical RW trajectory from entry to first exit for an arbitrary spatial domain of dimensionality d depends only on the ratio of the domain volume L^d and the enclosing surface L^{d-1} [22], i.e. l_c=CL, where C is a geometry-dependent constant; in the special case of a RW on a square system with isotropic incidence, C=$\pi/4$ [22]. Assuming an average velocity v, the average crossing time is $\tau^c = l_c/v = Cv^{-1}L$. For this single-tier case, the dynamics within the domain will be diffusive unless the underlying dynamical process is intrinsically subdiffusive, e.g. due to a fractal nature of the bath [14, 15]. In the latter case, we would have $\langle r^2(t) \rangle \sim t^{\alpha_f}$, with $\alpha_f < 1$, and the relation between time and space is effectively rescaled, so that

$$\tau^c = \frac{C}{v} L^{1/\alpha_f}. \tag{4}$$

Note that the relation corresponding to normal diffusion is recovered with $\alpha_f = 1$. To ensure dimensional consistency a generalized velocity with fractional distance units is used, i.e. $v \sim \delta r^{1/\alpha_f}/\delta t$.

Having determined the basic scaling relations, we consider a multi-tiered hierarchical landscape. The hierarchy assumption (i.e. $E_i > E_{i-1}$; $L_i > L_{i-1}$) requires only that E increase monotonically with L, but does not provide the functional relation between E_i and L_i. However, the relation can be determined from the MD simulations. To compute the subdiffusional exponent of the MSD, we recall that each tier in the hierarchy will have a characteristic escape time τ_i^e determined by L_i and E_i=$-\log p_i$. Combining Eqs. (2), (3), and (4) we obtain

$$\langle r^2(\tau_i^e) \rangle \simeq C_2 \, (\tau_i^e p_i)^{2\alpha_f}. \tag{5}$$

The reason that Eq. (5) holds approximately for each tier i, independently of the inner tier structure, is that the inner tier structure does not significantly impact the average number of barrier crossing attempts. The essential effect of the inner barrier hierarchy is to slow the diffusion within the outer valley, which does not change the probability of being located near the outer boundary, or the average outward flux across this boundary. Eqs. (2) and (3) yield values for L_i and p_i (and therefore E_i) provided that $\langle r^2(\tau_i^e) \rangle$, τ_i^c and τ_i^e are known. These are obtained from MD simulations using a trajectory postprocessing analysis (see Hierarchical Plateau Analysis in Computational Methods). We find that the relation E_i *vs.* L_i is logarithmic (see Fig. 4b in Results), i.e.,

$$E_i = E_0 + \gamma \log(L_i/L_0), \tag{6}$$

with the *hierarchy* parameter $\gamma \sim$ 2-4.5 [44]. To compute the subdiffusional exponent of the MSD, we use Eqs. (2), (4), and (6) and write p_i in Eq. (5) as a function of τ_i^e, i.e., p_i=$C_3(\tau_i^e)^{-\gamma\alpha_f/(1+\gamma\alpha_f)}$. This leads to the effective subdiffusive power-law, as also verified in Brownian simulations shown in Fig. 6 in the Appendix.

$$\langle r^2(t) \rangle \propto t^{\frac{2}{1/\alpha_f+\gamma}} = t^{\alpha}, \tag{7}$$

where we have defined the effective subdiffusive exponent α=$2/({\alpha_f}^{-1}+\gamma)$ and replaced τ^e by t (i.e. interpolating the power law between the discrete times τ_i^e). In the special case of normal diffusion, where α_f=1, the model predicts the subdiffusive exponent due to trapping alone with

$$\alpha_t \equiv 2/(1+\gamma). \tag{8}$$

We note that the derivation of Eq. (7) requires the existence of well-defined plateaus (so that Eq. (2) is valid), which in turn implies the existence of significant energy barriers ($E_i \gg 0$). For this reason, setting $\gamma = 0$ is not permissible in Eq. (7), and does not lead to the correct scaling for normal diffusion in Eq. (7). Numerical tests of Eq. (7) on model hierarchical landscapes (see Fig. 6 in the Appendix) show that Eq. (7) is accurate for $\gamma >$1. Additional subdiffusion arising from the effects of the bath, reflected in α_f, will act to decrease the subdiffusive exponent.

II. APPLICATION TO PROTEINS

We use the hierarchical trapping model to interpret molecular dynamics (MD) trajectories of three proteins

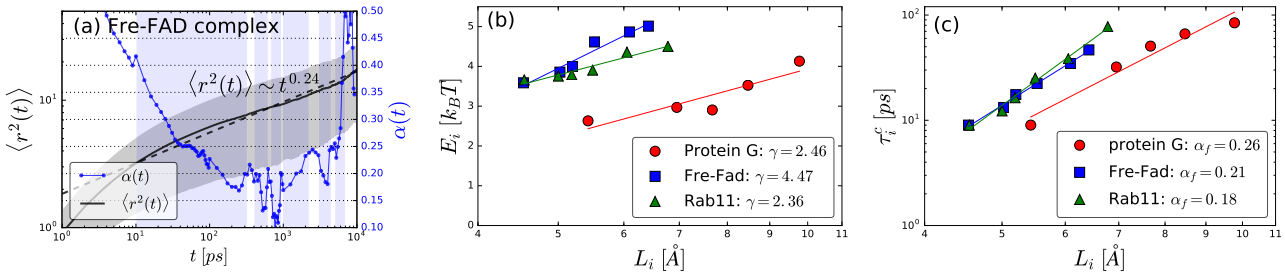

FIG. 4: (a) Illustration of Hierachical Plateau Analysis applied to the Fre-FAD complex (see Fig 8 for plots corresponding to Rab11a and protein G). (black solid) MSD for the projected trajectory; (black dashes) fit to a constant power law as in Eq. (1), with $\alpha = 0.24$; the standard deviation of the MSD is shown in light gray; (dotted blue curve) the time-varying exponent $\alpha(t)=d\log\langle r^2(t)\rangle/d\log t$ (see Computational Methods); plateaus extracted from the MSD curve (see Computational Methods) are highlighted in light blue; (b) Characteristic energy barriers E_i (units of k_BT) *vs.* L_i fitted to Eq. 6, yielding the corresponding values for γ. (c) Average crossing time τ_i^c *vs.* valley size L_i fit to Eq. 4, yielding the corresponding values for the exponent α_f.

of varying size and complexity (see Fig. 1): a 16-amino-acid fragment of protein G [23] (247 atoms), a signaling protein from the Ras family Rab11 [24] (2725 atoms) and the Flavin reductase enzyme (Fre) complexed with Flavin adenine dinucleotide (FAD) [25] studied previously [5, 6] (4064 atoms). The simulations are summarized in Computational Methods and in the Appendix.

We apply the Hierarchical Plateau Analysis to MD simulations extracting the quantities $\langle r^2\rangle_i^{\rm plt}$, τ_i^c and τ_i^e, and apply Eqs. (2), (3), and (4) to compute estimates for the characteristic valley sizes L_i, the energy barriers E_i, and the subdiffusion exponent intrinsic to the dynamics, α_f (see Fig. 4 and Table II in the Appendix). The barrier energies are in the range $\sim 2-5\ k_BT$, and the valley sizes shown are between ∼4Å and ∼11Å (Fig. 4b). The model estimates are in rough accord with the free energy landscape in Fig. 2b for the Fre-FAD complex. Further, Rab11a and Fre-FAD have higher barriers and smaller valley sizes compared to protein G. This is consistent with a visual examination of the MD trajectories, which show that protein G appears more flexible than Rab11a and Fre-FAD (see Fig. 1), and with the average RMS deviations of the protein backbones from the average MD simulation structures, which were 2.32Å, 0.97Å, and 0.98Å for protein G, Rab11a, and Fre-FAD, respectively; in part, the larger apparent flexibility of protein G arises from the unfolding of the helix at the termini (Fig. 1a). The plot of E_i *vs.* L_i (Fig. 4b) shows the extracted values for the hierarchy parameter γ in Eq. (6), obtained from a least-squares fit to the data for protein G, Rab11a, and Fre-FAD. Substituting γ into Eq. (8) yields α_t, the subdiffusive exponent resulting from the contribution of trapping alone; i.e. the exponent that would be observed if the underlying diffusive process were Brownian. The values for the different proteins can be compared to the subdiffusive exponents $\alpha_{\rm fit}$, obtained by a least-squares fit to the MSDs (Table I). α_t is significantly larger than $\alpha_{\rm fit}$ in all three cases, indicating that trapping alone cannot account fully for the observed subdiffusion. Figure 4c confirms the power-law relation between τ_i^c and L_i in Eq. (4), and provides values for α_f, the subdiffusive exponent corresponding to the inherent dynamics. It is noteworthy that $\alpha_f < 1$ for all three proteins, indicating that the inherent dynamical process is not Brownian, but rather subdiffusive. Further, α_f is also significantly larger than $\alpha_{\rm fit}$, indicating that neither can the inherent process alone account for the measured subdiffusion, as found for the trapping mechanism. Only the effective exponent α brought in Eq. 7, that combines the two sources of subdiffusion, shows excellent agreement with $\alpha_{\rm fit}$, demonstrating that both mechanisms are critical for explaining the subdiffusive dynamics.

	$\alpha_t = \frac{2}{1+\gamma}$	α_f	$\alpha = \frac{2}{1/\alpha_f+\gamma}$	$\alpha_{\rm fit}$
protein G	0.58	0.26	0.31	0.30
Rab11a	0.60	0.18	0.25	0.25
Fre-FAD	0.37	0.21	0.22	0.24

TABLE I: Model predictions of the subdiffusive exponents representing, from left to right: (i) the contribution due to trapping alone, α_t in Eq. (8); (ii) the underlying fluctuations alone, α_f; (iii) the combined contribution of both mechanisms α in Eq. (7). For comparison, the subdiffusive exponent $\alpha_{\rm fit}$ is computed by the least squares fit to the MSD.

III. DISCUSSION

We present an analytically tractable hierarchical trapping model, consistent with the postulates of Frauenfelder et al. [9], which shows how a particular hierarchical structure of the energy landscape provides a source of subdiffusion due to trapping. The model is general with respect to the geometry of the energy wells and dimensionality of the energy landscape, which influence the proportionality constants but not the functional form of the derived subdiffusive power law (Eq. (7)). An im-

portant distinction between this model and traditional continuous-time random walk (CTRW) models is that the subdiffusion is not due to ageing, which is essentially the lack of ergodicity in the dynamics [18]. In the present model, the subdiffusion arises naturally in the ergodic setting from the hierarchical arrangement of energy wells and barriers.

While the model uses a single value for the energy and length scale in each tier of the hierarchy, E_i and L_i, realistic energy landscapes are expected to exhibit distributions of energy barriers and valley sizes. However, our finding of *distinct* plateaus in the MSD computed from the protein simulations suggests that the distributions of the energy barriers and the valley sizes are relatively compact. Otherwise, the plateaus would be smeared out by the temporal averaging inherent in the MSD computation. The identified plateau regions were robust with respect to the choice of the threshold ϵ (see Fig. 9 in the Appendix), projecting the dynamics onto different combinations of principal component vectors (see Fig. 10 in the Appendix), as well as to repeating the calculation with non-overlapping trajectory segments (see Fig. 11 in the Appendix), indicating that they are not artifacts of dimensionality reduction [26], or noise. These results permit us to conclude that the distributions of energies and length scales can be characterized by single values E_i and L_i for each tier i, representing the most probable or average values. The logarithmic relationship between energy barrier heights and valley sizes found for all three proteins considered here, as described in Eq. (6), provides evidence that the energy landscape is, in fact, hierarchical.

The dominant conformational macrostates of the larger proteins, Rab11a and Fre, are shown by clustering the MD trajectories on the basis of the free energy landscape of the first three PCs (see Fig. 5). The macrostates are seen to differ mainly in the conformations of flexible loop regions (Fig. 5a,b) and in the relative positions of secondary structure motifs, primarily α-helices, which do not undergo significant internal structural changes (Fig. 5c) [27]. For the Flavin reductase, the differences between the clusters appear largest in the vicinity of the FAD binding pocket (Fig. 5b,c). These differences could explain the experimental observation of dynamic disorder in the distance between the isoalloxazine moiety of FAD and the active site residue Tyr35 found by Yang et al. [5], since transitions between the macrostates affect the position of the isoalloxazine ring, as was suggested on the basis of short MD simulations [6]. Further, if the rates of the Flavin reduction reaction by Fre are substantially different for each macrostate, one would expect to see dynamic disorder of reaction rates, and nonexponential relaxation of dynamical observables, as described by Frauenfelder et al. [9] for the rebinding of CO to myoglobin, or by Lu et al. [28] for enzymatic turnovers of cholesterol oxidase molecules.

We emphasize the conclusion that hierarchical trapping alone cannot account fully for the observed subdiffusion. We have consistently found that parametrizing the trapping model from MD simulation data predicts subdiffusion that is faster than what is observed by fitting the calculated MSD directly. For example, projecting the MD trajectory data onto different principal components (see Fig. 10 in the Appendix) did not change significantly the value of the trapping exponent. Further, we found that the size of the valley L does not scale linearly with the time needed to cross it τ_c, as would be expected for Brownian diffusion. This non-linear scaling law is consistent across all tiers, for all three proteins. We therefore assumed that the fluctuating process itself (i.e., independently of energy barriers) is itself subdiffusive, as captured by the parameter α_f in Eq. (4).

The origin of the inherent subdiffusion could be the fractal topology of proteins, fractional noise of the unresolved ("bath") degrees of freedom, or another, as yet unspecified, source. This finding is also consistent with the fact that observations from both experiments and simulations show that autocorrelation functions calculated for protein dynamics typically exhibit slow power-law decay [4–6], whereas trapping models with a truncated distribution of finite energy barriers lead to a truncated distribution of waiting times, producing exponentially decaying autocorrelation functions [40].

The present study thus demonstrates that subdiffusion in protein dynamics originates from multiple physical phenomena. Given that internal protein motions are intimately related to biological function, the results are expected to be of general interest in the study of proteins.

IV. COMPUTATIONAL METHODS

Molecular Dynamics Simulations. The protein structures for *protein G*, *Rab11*, and *Fre* were obtained from Protein Data Bank (PDB) files 1GB1, 1YZK, and 1QFJ, respectively. Unresolved protein coordinates were modeled using the the program CHARMM [29], and coordinates for the FAD ligand were taken from the active site of Flavodoxin reductase/FAD complex (PDB ID 1FDR), which is structurally homologous to Fre. MD simulations in the canonical ensemble were performed with the program ACEMD [32] for 1μs for the three proteins, using the CHARMM energy function [33, 34, 37].

To check that the data selected for analysis were equilibrated, the trajectories were divided into four consecutive segments of 250ns, and the subdiffusive exponent was calculated for each segment. The exponent started from higher values and relaxed to a constant value in less than 500ns for all proteins. Therefore we parametrized the trapping model using only the final 500ns of the MD.

Principal Component Analysis. The coordinates of the C_α atoms were extracted from the MD simulation trajectories at 1ps intervals. Principal components (PCs) were computed using the program CARMA [38]. To obtain the coordinates used in the hierarchical model, the original coordinates were projected into the first three

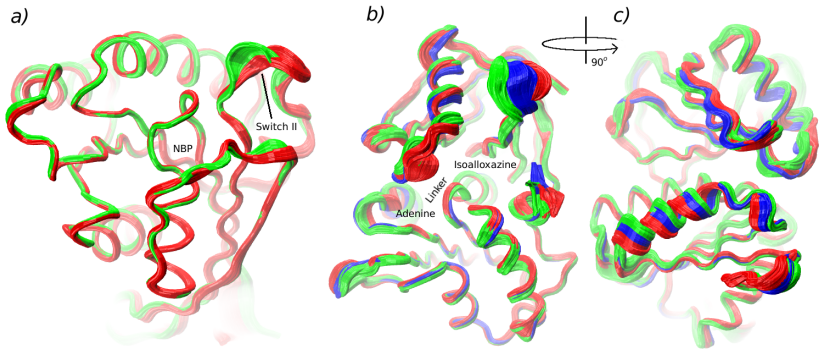

FIG. 5: Conformational ensembles corresponding to the dominant free energy minima obtained from principal component analysis of (a) Rab11a [39], (b) Fre [25] and (c) Fre (side view). The ensembles are drawn for Fre in red, green, and blue, in the order of increasing free energy. For clarity, only two ensembles are shown for Rab11a in red and green. In (a), NBP is the nucleotide binding pocket, which contains GTP (see Fig. 1a).

PCs. Because similar plateau parameters were obtained using projections on various combinations of PCs (see Fig. 10 in the Appendix), the main results are presented in the 2D space of PC1 and PC2 to facilitate visualization of the corresponding free energy landscape in Fig. 1d,e.

Hierarchical Plateau Analysis. Working in projected coordinates, we calculate the MSD using a moving average:

$$\langle r^2(t)\rangle = \frac{1}{T-t}\int_0^{T-t} |\mathbf{r}(\tau+t)-\mathbf{r}(\tau)|^2 d\tau, \qquad (9)$$

where T is the trajectory length. For $t \to T$ sampling becomes poor, and thus we only consider the range $t < 0.1T$). The MSDs were fit to a power law as in Eq. (1), using least squares to obtain the overall exponent $\alpha_{\rm fit}$, displayed in Table I. More generally, we consider the exponent α as variable in time, and compute it using a finite-difference approximation to $\alpha(t)=d\log\langle r^2(t)\rangle/d\log t$. MSD plateaus are identified as local minima of $\alpha(t)$. The onset (τ_i^c) and end (τ_i^e) of the plateaus, highlighted in Fig. 4a, are related to the first time point for which $d\alpha(t)/dt < -\epsilon$, and the last point for which $d\alpha(t)/dt > \epsilon$, respectively, where ϵ is en empirically tuned constant, which we set to 0.015. Additional details are given in Section B in the Appendix.

Acknowledgments

YM received support from the Weizmann Institute of Science, National Postdoctoral Award Program for Advancing Women in Science. VO and MK were supported by NIH grant No. 5R03AI111416. Computer resources were provided by the Faculty of Arts and Sciences Computing Group at Harvard.

Appendix A: Trapping on a model hierarchical energy landscape

In this section we illustrate subdiffusion caused by trapping in a hierarhical structure with an underlying Brownian process, i.e. with α_f=1. We use a model 2D landscape, similar to the one in Fig. 3. We consider a maximum of six tiers, with the valley sizes corresponding to each tier indicated in Fig. 6 (see legend). The hierarchy is specified using Eq. (6) with different values for the parameter γ. To generate the simulation trajec-

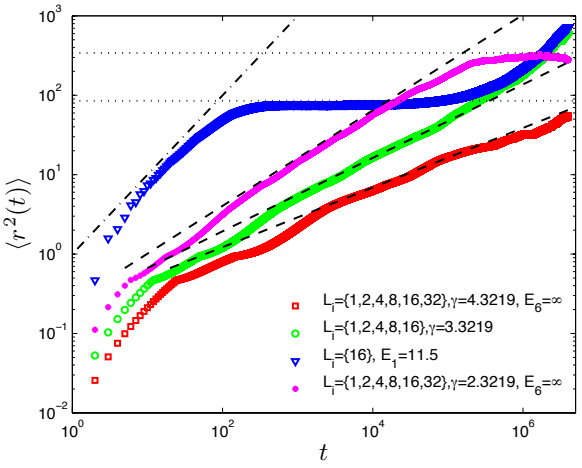

FIG. 6: (color online) Evolution of the MSD of a Brownian random walker for 2D landscapes with different hierarchies. The dashed lines corresponds to the power law $\langle r\rangle \propto t^{\alpha_t}$ predicted by Eq. (9) in the main text, i.e. $\alpha_t = 2/(1+\gamma)$, yielding α_t=0.376, 0.463 and 0.602. Horizontal dotted lines indicate theoretical plateau locations $\langle r\rangle^{plt}$ as calculated from Eq. (3) for $L = 16$ and $L = 32$. Brownian diffusion law $\langle r^2\rangle \propto t$ is shown as a dash-dot line in the case of a single tier (blue triangles). The MSDs were computed from RW simulations with 40 million Monte Carlo steps; time t corresponds to the number of steps. E=∞ corresponds to an impenetrable barrier.

tories, we use a discrete 2D random walk on a square lattice, with stepping probabilities of 0.25 in each of the four available directions (left, up, right, down). Fig 6 shows that the agreement between model predictions and direct calculations is very good. For comparison with a non-hierarchical case, a simulation with a single tier (L_1=16, p_1=0.00001) is used, which results in Brownian diffusion before and after the single plateau.

Appendix B: Hierarchical Plateau Analysis (HPA)

Computing crossing times (τ_i^c) and escape times (τ_i^e) in a multi-tier hierarchy. The validity of Eq. (5) rests on the fact that the time to escape from a valley of size L_i is not changed significantly by the presence of smaller inner valleys with lower barriers inside L_i. This can be easily seen from a transition state theory argument, whereby the ratio of the partition function of the outer barrier (assuming a small finite barrier width) to that of the enclosed valley changes only slightly with the addition of inner barriers, provided that the widths of the inner barriers are not too large relative to the valleys. However, the inner barriers affect significantly the kinetics of motion within the outer valley, and in particular, the time required to cross the outer valley, which is needed to extract the valley sizes from the MSD, *via* Eq. (4).

To apply Eq. (4) to the MSD in the presence of the multi-tier hierarchy, we coarse-grain the spatial dynamics inside L_i. Specifically, we define a (possibly fractional) coarse-grained velocity v_i, consistent with the units of Eq. (4), as

$$v_i = L_{i-1}^{1/\alpha_f}/\tau_{i-1}^e, \tag{B1}$$

and use it in Eq. (4) to define a coarse-grained crossing time $\hat{\tau}_i^c$ as,

$$\hat{\tau}_i^c = \frac{C}{v_i} L_i^{1/\alpha_f} = C\tau_i^e \left(\frac{L_i}{L_{i-1}}\right)^{1/\alpha_f} = C\tau_i^e \frac{\tau_i^c}{\tau_{i-1}^c}, \tag{B2}$$

where in the second step we used Eq. (4) to eliminate L_i and L_{i-1}.

The only difference between Eqs. (4) and (B2) is the velocity. Clearly, the inner barriers imply that $v \gg v_i$ (and therefore $\hat{\tau}_i^c \gg \tau_i^c$); in fact, the main effect of using v_i instead of v is to remove the relatively faster equilibration of trajectories within the inner valleys. Otherwise, the apparent valley crossing time τ^c would appear too low. More specifically, it would be "contaminated" by frequent encounters with the same boundary due to transient confinement near the tier-i boundary by the inner barriers, which would prevent estimating L_i from the MSD. This effect is seen in Fig. 3, where, for the two-tier case, the second plateau appears two orders of magnitude later in time than the plateau in the single-tier case. Solving Eq. (B2) for τ^c we obtain

$$\tau_i^c = \frac{1}{C}\hat{\tau}_i^c \frac{\tau_{i-1}^c}{\tau_{i-1}^e}. \tag{B3}$$

$\hat{\tau}_i^c$ is obtained directly from the MSD by analyzing its temporal finite differences, as described below. In practice, Equation (B3) is applied recursively for each tier i, starting at tier 1, for which $\hat{\tau}_1^c$=τ_1^c. The onset ($\hat{\tau}_i^c$) and end of the plateaus (τ_i^e) in the MSD are identified as the first point at which $\frac{\Delta\alpha(t)}{\Delta t} < -\epsilon$, and the last point at which $\frac{\Delta\alpha(t)}{\Delta t} > \epsilon$, respectively, where ϵ is an empirical threshold, taken here to be 0.015, and Δ's indicate that a finite difference approximation to the derivative was used. This method of locating plateaus in the MSD curves was validated by applying it to 2D model potentials (see Fig. 7).

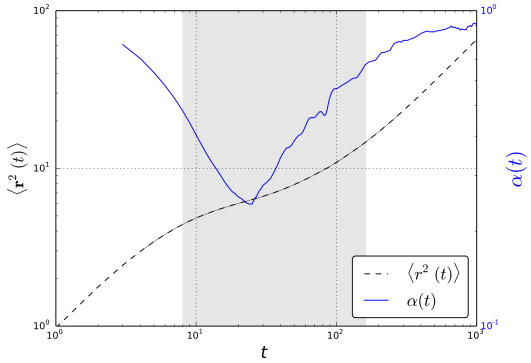

FIG. 7: Application of plateau identification to Brownian motion simulations in a single tier L=4 and $p = \exp(-E)$=0.01. The identified plateu is marked in a light blue block, and the theoretical value for the plateau escape time τ^e is $\sim L/p$=40.

HPA applied to Rab11 and protein G. Figure 8 shows the plateaus identified from the MSD curves of proteins Rab11a and protein G using Hierarchical Plateau Analysis, marked with light blue blocks. The results of HPA applied to each of the three proteins are summarized in Table II.

Sensitivity of HPA to ϵ. To evaluate the sensitivity of HPA on the choice of plateau threshold parameter ϵ, we performed HPA on the Fre-FAD complex using five values of ϵ in the range $[0.01, 0.03]$. Fig. 9 shows that the corresponding values of α_t are essentially unchanged, indicating the robustness of the obtained exponents.

Sensitivity of HPA to the choice of principal components. To assess the sensitivity of HPA to the choice of the principal component vectors (PCVs), we repeated the analysis of the Fre-FAD complex MD trajectories using the three possible pairs of PCVs from the set of the three PCVs corresponding to the largest eigenvalues. HPA was performed on each projection and the values for E_i and L_i were extracted at each tier i. The corresponding trapping exponents α_t, shown in Fig. 10, do

	protein G 247 atoms				Rab11a 2725 atoms				Fre-FAD 4064 atoms			
i	τ_i^c $[ps]$	τ_i^e $[ps]$	E_i $[k_BT]$	L_i $[\AA]$	τ_i^c $[ps]$	τ_i^e $[ps]$	E_i $[k_BT]$	L_i $[\AA]$	τ_i^c $[ps]$	τ_i^e $[ps]$	E_i $[k_BT]$	L_i $[\AA]$
1	9.0	125	2.63	5.43	9.0	350	3.66	4.55	9.0	325	3.59	4.54
2	32.1	625	2.97	6.95	12.3	525	3.76	5.00	13.2	625	3.86	5.02
3	50.7	925	2.91	7.67	16.4	725	3.79	5.19	17.5	950	3.99	5.20
4	66.2	2250	3.53	8.47	25.2	1250	3.91	5.50	22.3	2250	4.61	5.53
5	84.3	5250	4.13	9.77	38.5	3000	4.36	6.05	34.7	4500	4.87	6.09
6					77.5	7000	4.50	6.78	46.6	7000	5.01	6.42

TABLE II: Tier values extracted *via* Hierarchical Plateau Analysis of the MSDs calculated for the projected MD trajectories for protein G, Rab11a, and the Fre-FAD complex. Tier $i = 1$ corresponds to the first plateau in the MSD.

not vary significantly across the three PCV projections, and suggest that the conclusions are robust.

Convergence of simulations. To assess the convergence of the results, we split the equilibrated MD trajectory of the Fre-FAD complex into two subtrajectories of equal lengths, and computed $\alpha(t)$ and the tier values E_i and L_i for each subtrajectory (see Fig. 11). The resulting plateaus were similar, and lead to essentially the same scaling constants.

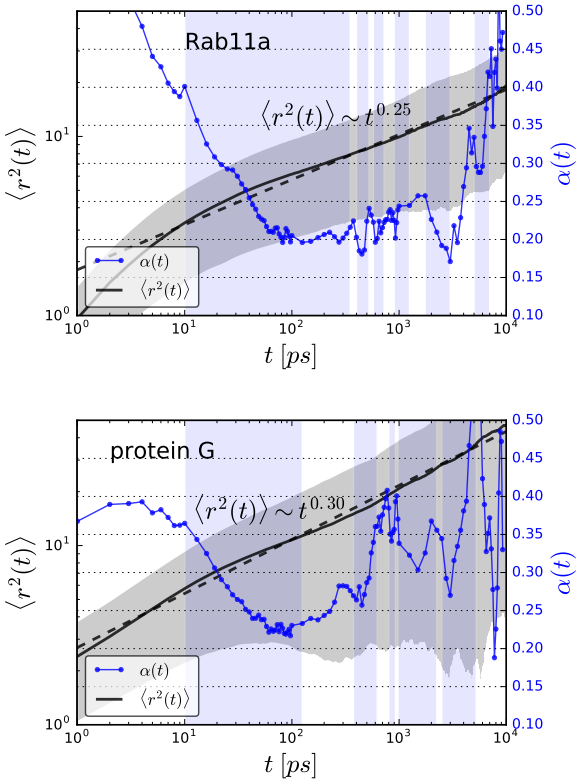

FIG. 8: Illustration of Hierarchical Plateau Analysis applied to (a) Rab11a and (b) protein G ; black solid: MSD for the projected trajectory; black dashes: fit to a constant power law with α=0.25 and 0.30 for Rab11a and protein G, respectively; the standard deviation of the MSD is shown in light gray; blue : the time-varying exponent $\alpha(t)$ (see Computational Methods); plateaus extracted from the MSD curve (see Computational Methods) are highlighted in light blue (color online).

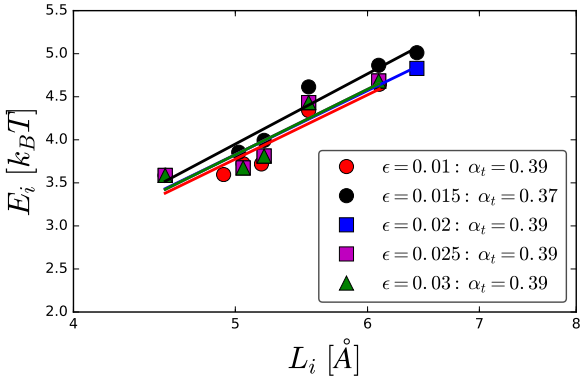

FIG. 9: Barrier energies and valley sizes (E_i vs L_i) for the Fre-FAD complex obtained from HPA using different values of parameter ϵ (see text). The trapping exponent α_t was computed from Eqs. (6) and (8).

Appendix C: Details of MD Simulations

Protonation states of titratable residues were assigned using the program PROPKA [31]. The resulting protonation states were the same as those in pH-neutral solution. The structures were immersed in pre-equilibrated cubic boxes with TIP3 water molecules, ensuring an environment of solvent molecules between the protein and the nearest box boundary of at least 11Å. From the set of water molecules that were at least 5Å away from the proteins, some were replaced with Na^+ and Cl^- ions to achieve a charge-neutral system with a concentration of about 100mM. The CHARMM22 force field with CMAP correction [33] was used for all simulations. Initial force field parameters for FAD ligand were obtained from the

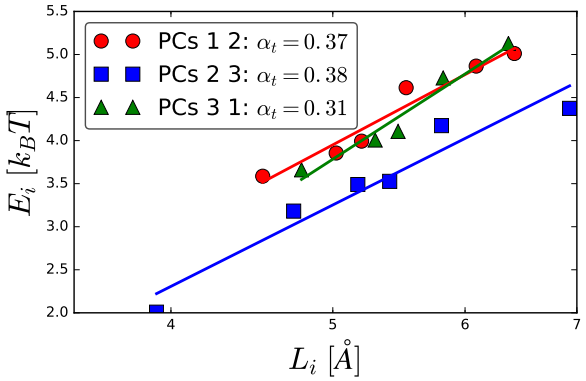

FIG. 10: HPA of Fre-FAD complex simulations projected onto three 2D principal component eigenvector (PCV) spaces, $1-2$, $2-3$ and $1-3$, with the PCVs numbered in the order of decreasing eigenvalue. Barrier energies and valley sizes (E_i vs L_i) are plotted for the different projections, and α_t is computed from Eqs. (6) and (8).

www.paramchem.org server using CGENFF tools [34], and refined using the FFTK software [35] for Visual Molecular Dynamics [36]. Each solvated system was equilibrated in the NPT ensemble at 300K and 1 atm for 1ns with weak harmonic positional restraints acting on the heavy protein atoms, to allow the solvent atoms to relax around the protein, and to adjust the system volume to the imposed pressure. An additional 1ns of equilibration was performed without harmonic restraints. For these (equilibration) and for all subsequent (production) simulations, the Langevin thermostat with a small friction coefficient of 0.1/ps was used. Such a small value ensures that the protein dynamics are only slightly perturbed by the thermostat, while still maintaning a prescribed temperature. The barostat was then turned off (to increase simulation speed), and the systems were simulated in the canonical ensemble for 1000ns for protein G, Rab11 and Fre-FAD. For these production simulations, the following long-range force options were used. The cutoff for the van der Waals (VDW) and short-range electrostatic interactions was 9Å, and the VDW interactions were smoothly scaled to zero for inter-atom distances in the range 7.5Å–9Å using the CHARMM cutoff function [29]. Long-range electrostatics were re-evaluated at every other simulation step using PME with a multiple-step RESPA integrator. The masses of hydrogen atoms were increased to 4 a.m.u.; the masses of the atoms bonded to the hydrogens were decreased to keep the total mass unchanged; all bonds involving hydrogens were constrained using the SHAKE/RATTLE algorithms [41]. These adjustments allow the simulation step to be increased to 4fs. Each simulation required about 5 days on a workstation equipped with a NVIDIA GTX780 graphical processor.

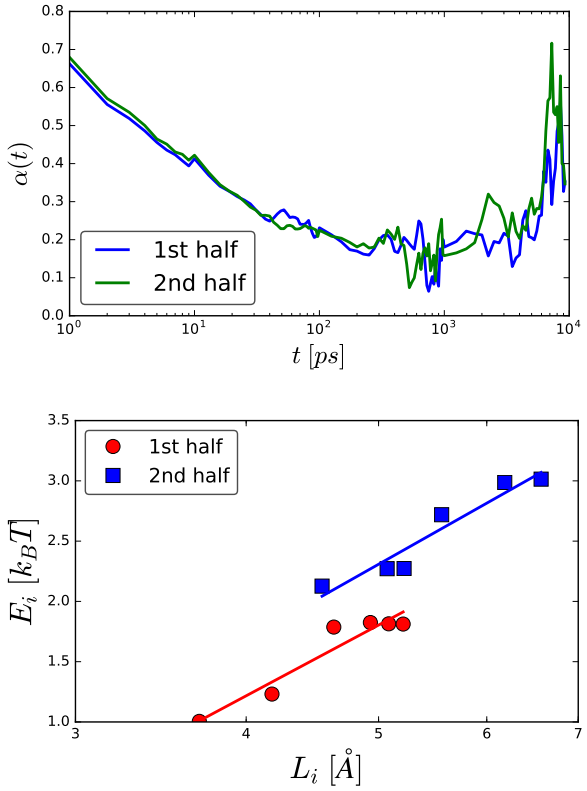

FIG. 11: Analysis of Fre-FAD complex simulation for two consecutive sub-trajectories of $250ns$ each. Top: $\alpha(t)$ (the time-dependent subdiffusive exponent of the MSD) calculated for the first and second half of the full trajectoy, exhibiting similar behavior; bottom: plateau values (E_i vs L_i) extracted from HPA analysis run on the two sub-trajectories. Both trajectories exhibit a similar power-relation.

[1] G. Weber, Adv. Protein Chem. **29**, 1 (1975).
[2] J. McCammon, B. Gelin, and M. Karplus, Nature **267**, 585 (1977).
[3] B. Gelin and M. Karplus, Journal of the American Chemical Society **97**, 6996 (1975).
[4] I. Iben, D. Braunstein, W. Doster, H. Frauenfelder, M. Hong, J. Johnson, S. Luck, P. Ormos, A. Schulte, P. Steinbach, et al., Physical review letters **62**, 1916 (1989).
[5] H. Yang, G. Luo, P. Karnchanaphanurach, T.-M. Louie, I. Rech, S. Cova, L. Xun, and X. S. Xie, Science **302**, 262 (2003).
[6] G. G. Luo, I. I. Andricioaei, X. S. X. Xie, and M. Karplus, J. Phys. Chem. B **110**, 9363 (2006).
[7] P. Senet, G. Maisuradze, C. Foulie, P. Delarue, and H. Scheraga, Proceedings of the National Academy of Sciences of the United States of America **105**, 19708 (2008).
[8] L. Milanesi, J. Waltho, C. Hunter, D. Shaw, G. Beddard, G. Reid, S. Dev, and M. Volk, Proceedings of the National Academy of Sciences of the United States of America **109**, 19563 (2012).
[9] H. Frauenfelder, S. Sligar, and P. Wolynes, Science **254**, 1598 (1991).
[10] F. Rao and M. Karplus, Proceedings of the National Academy of Sciences of the United States of America **107**, 9152 (2010).
[11] C. Monthus and J.-P. Bouchaud, J. Phys. A **29**, 3847 (1996).
[12] B. Mandelbrot and J. V. Ness, SIAM Rev. **10**, 422 (1968).
[13] K. S. Schweizer, J. Chem. Phys. **91**, 5802 (1989).
[14] R. Metzler, E. Barkai, and J. Klafter, Phys. Rev. Lett. **82**, 3563 (1999).
[15] S. Kou and X. Xie, Physical Review Letters **93**, 180603 (2004).
[16] R. Granek and J. Klafter, Phys. Rev. Lett. **95**, 098106 (2005).
[17] T. Neusius, I. Daidone, I. Sokolov, and J. Smith, Physical review letters **100**, 188103 (2008).
[18] X. Hu, L. Hong, M. Smith, T. Neusius, X. Cheng, and J. Smith, Nature Phys. (2015).
[19] A. Kitao, S. Hayward, and N. Go, Proteins **33**, 496 (1998).
[20] T. Ichiye and M. Karplus, Proteins **11**, 205 (1991).
[21] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Structure, Function, and Bioinformatics **17**, 412 (1993), ISSN 1097-0134.
[22] S. Blanco and R. Fournier, EPL **61**, 168 (2003).
[23] A. Gronenborn, D. Filpula, N. Essig, A. Achari, M. Whitlow, P. Wingfield, and G. Clore, Science **253**, 657 (1991).
[24] S. Eathiraj, X. Pan, C. Ritacco, and D. Lambright, Nature **436**, 415 (2005).
[25] M. Ingelman, S. Ramaswamy, V. Nivi'ere, M. Fontecave, and H. Eklund, Biochemistry **38**, 7040 (1999).
[26] S. V. Krivov, PLoS Comput Biol **6**, e1000921 (2010).
[27] R. Elber and M. Karplus, Science (New York, N.Y.) **235**, 318 (1987).
[28] H. Lu, L. Xun, and X. Xie, Science **282**, 1877 (1998).
[29] B. Brooks, C. Brooks III, A. Mackerell Jr., L. Nilsson, R. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al., J. Comput. Chem. **30**, 1545 (2009), pMC2810661.
[30] Y. Zhang and J. Skolnick. . Nucl. Acids Res., 33:2302–2309, 2005.
[31] Hui Li, Andrew D. Robertson, and Jan H. Jensen. . PROTEINS: Struct. Funct. and Bioinf., 61:704–721, 2005.
[32] M. Harvey, G. Giupponi, and G. D. Fabritis, J. Chem. Theory and Comput. **5**, 1632 (2009).
[33] A. MacKerell Jr, M. Feig, and C. Brooks III, J. Comput. Chem. **25**, 1400 (2004).
[34] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, et al., J. of Comput. Chem. **31**, 671 (2010), ISSN 1096-987X.
[35] C.G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J.C. Gumbart, Journal of Computational Chemistry **34**, 32 (2013).
[36] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33–38 (1996).
[37] R. Best, X. Zhu, J. Shim, P. Lopes, J. Mittal, M. Feig, and A. MacKerell Jr., J. Chem. Theor. Comput. **8**, 3257 (2012).
[38] N. Glykos, J. Comput. Chem. **27**, 1765 (2006).
[39] B. J. Sung and A. Yethiraj, Biophysical Journal **97**, 472 (2009).
[40] S. Burov, J.-H. Jeon, R. Metzler, and E. Berkai, Physical Chemistry Chemical Physics **13**, 1800–1812 (2011).
[41] J.-P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen, J. Comput. Phys. **23**, 327–341 (1977).
[42] The unresolved modes influence the dynamics of the resolved PCs implicitly, since they are part of the 'bath' degrees of freedom.
[43] Generally, the dynamics will not be Markovian, and furthermore, could evolve under the influence of fractional noise [12–16].
[44] E_0, and $L_0 > 0$ are parameters related to the intercept that are unimportant for this analysis.