Main role of fractal-like nature of conformational space in subdiffusion in proteins

Luca Maggi[1,*] and Modesto Orozco[1,2]

[1]*Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Baldiri Reixac 10, Barcelona 08028, Spain*

[2]*Departament de Bioquímica i Biomedicina. Facultat de Biologia, Universitat de Barcelona, Avgda Diagonal 647, Barcelona 08028, Spain*

(Received 13 September 2023; revised 26 January 2024; accepted 5 February 2024; published 1 March 2024)

Protein dynamics involves a myriad of mechanical movements happening at different time and space scales, which make it highly complex. One of the less understood features of protein dynamics is subdiffusivity, defined as sublinear dependence between displacement and time. Here, we use all-atoms molecular dynamics (MD) simulations to directly interrogate an already well-established theory and demonstrate that subdiffusivity arises from the fractal nature of the network of metastable conformations over which the dynamics, thought of as a diffusion process, takes place.

DOI: 10.1103/PhysRevE.109.034402

I. INTRODUCTION

Protein dynamics is widely being recognized as a pivotal element to understand protein function [1,2]. In fact, enzymatic reactions [3], signal transduction [4], molecular motors [5], or transport across membrane [6] could not be understood ignoring dynamics, which is defined as the set of time-dependent protein conformational changes occurring during the exploration of the protein multidimensional energy landscape. Experimental and computational studies highlighted the "roughness" of the protein energy landscape, with many metastable states separated by energy barriers [1,2,7], which implies that protein dynamics can be thought of as a diffusion process among metastable states [8–10]. Previous investigations showed [9,11–15] that protein dynamics exhibits a poorly understood subdiffusive behavior [16], which implies a sublinear relationship between the mean square displacement (MSD) and time (t). More formally speaking,

$$\mathrm{MSD} = \langle |\boldsymbol{X}(0) - \boldsymbol{X}(t)|^2 \rangle \sim t^{\alpha}, \tag{1}$$

where $\langle \cdots \rangle$ represents an ensemble average and $\boldsymbol{X}(t) = \{x_1(t), \ldots, x_N(t)\}$ is a single protein conformation at time t. This is a set of N time-dependent variables, $x_i(t)$, that are the protein degrees of freedom. In this work, these correspond to the Cα carbon atoms coordinates. The exponent α is equal to 1 for a normal diffusive problem and less than 1 in the subdiffusivity regime. The microscopic origin of subdiffusion is still poorly understood even though several explanations have been put forward: as the intrinsic "viscoelastic" behavior of protein [12,17]; the trapping due to the energy barrier [11]; multiple relaxation processes featured dynamics [16] or the fractal nature of the conformational space [15,18,19]. Here, we will show that the latter should be considered as the main origin of subdiffusion in protein by directly questioning the theory of diffusion on fractals. According to this theory, the probability (P) to observe a displacement l at time t, which in an N-dimensional homogeneous Euclidean space is $P \sim t^{-N/2}\exp(-\frac{Nl^2}{4Dt})$ [20], where D is the mass diffusivity, is given by [21,22]

$$P(l,t) \sim t^{-d_s/2} \exp\left[-A \left(\frac{l^{2d_f/d_s}}{t} \right)^{1/(2d_f/d_s)-1} \right], \tag{2}$$

in which d_f is associated to geometrical properties of the fractal-like space, d_s is related to dynamical features of its exploration process [22,23], and A is constant. Since $\mathrm{MSD} = \int l^2 P(l,t) dl$, exploiting Eq. (2), one gets [21–23]

$$\mathrm{MSD} \sim t^{d_s/d_f}. \tag{3}$$

Therefore, a simple comparison with Eq. (1) provides the definition of the general α exponent within the theory of the diffusion on fractals, being

$$\alpha = \frac{d_s}{d_f}. \tag{4}$$

Since α, d_s, and d_f can be computed independently from atomistic molecular dynamics (MD) simulations, we can use Eq. (4) to validate whether the fractal hypothesis stands for a representative set of small proteins whose dynamics can be well treated by atomistic MD simulations (see Appendix A for details). Particularly, we explore the equilibrium dynamics of the Villin headpiece [Villin; PDB ID: 1VII [24]; Fig. 1(a)], the N-terminal of the human histone H4 tail (H4); and a PDZ domain (PDZ; PDB ID: 1D5G [25]). They differ in the number of residues, 25, 32, and 96 for H4, Villin, and PDZ respectively, and its secondary and tertiary structure [12,17]. For the sake of clarity, we will show in the main text all the details of our analysis for Villin only, but will include H4 and PDZ in the presentation of the main results (see Appendies B and E for the detailed results of the analysis for H4 and PDZ).

*luca.maggi@irbbarcelona.org

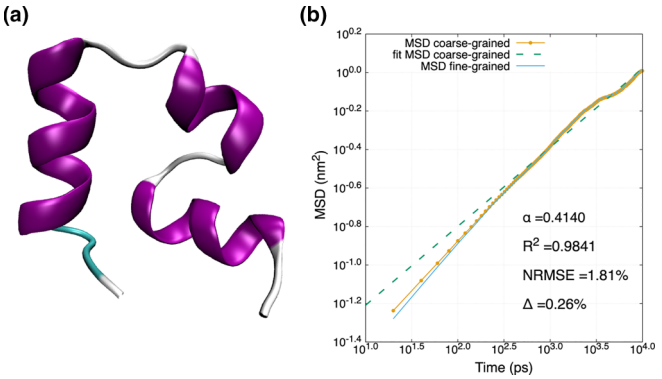

FIG. 1. Snapshot extracted from MD simulations of (a) Villin and (b) the associated MSD profile linearly fitted. Plot shows the R^2 and the NRMSE relative to the linear regression computed on coarse-grained MSD.

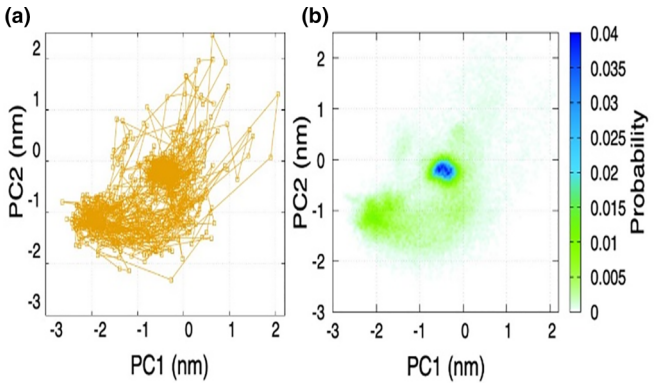

FIG. 2. (a) Villin trajectory in the two-sub-dimensional conformational space. (b) Probability surface obtained from the trajectory.

II. METHODS AND RESULTS

Even though we focused only on Cα's atoms coordinates, the analysis presented in the following requires a further space dimensionality reduction to avoid their otherwise prohibitive computational costs. This choice is purely technical and the subdiffusive behavior is maintained. For this purpose, we used principal component analysis [26,27]. Briefly, this consists in the diagonalization of the N by N correlation of Cα carbon atoms coordinates, where rigid translations and rotations are removed by superimposing each conformation on an average one. The kth principal component (PC_k) is defined as the projection of $\boldsymbol{X}(t)$ onto the kth eigenvector ($\boldsymbol{\varphi}_k$), $\mathrm{PC}_k(t) = \boldsymbol{\varphi}_k \cdot \boldsymbol{X}(t)$. The first few eigenvectors are associated with the largest movements sampled by the protein, so, in the absence of specific prior knowledge, the associated PCs are extremely useful for studying the internal dynamics of protein. Therefore, we chose the first two PCs as descriptors of the conformational space and projected the whole protein dynamics onto this two sub-dimensional space where the diffusion process will take place and a single conformation is defined as $\hat{\boldsymbol{X}}(t) = \{\mathrm{PC1}(t),\ \mathrm{PC2}(t)\}$.

It should be noted that the results presented below are independent of the descriptors chosen and that different descriptors can be used to obtain the same results. This is shown in Appendix C for the case of Villin, where we repeated the analysis carried out in the main text using different variables and still arrived at the same conclusions.

The identification of metastable states has been done over the two-dimensional conformational space by means of agglomerative hierarchical clustering [28–30]. This method uses a bottom-up approach. In the first iteration, each conformation is considered as an individual cluster. They are then merged together according to a similarity criterion based on a linkage method that is a function of the Euclidean distance between the conformations. Here we used the so-called Ward linkage method [29,30]. The similarity criterion consists in defining a cut-off value (ε) for the linkage method to determine whether two clusters should be merged. This parameter thus controls the average size of the clusters.

The reasons leading to employing this clustering method are twofold. Firstly, the intrinsic hierarchy of all metastable states seems reasonable as structural differences among conformations can be naturally classified as subsets of decreasing size that subdivide the entire space. This idea is supported by previous works highlighting this feature [2,8]. On the other hand, this clustering method presents technical advantages as it does not require one to set a fixed number of clusters (as K-means methods [31]) employing the adjustable parameter ε and it does not produce any outlier conformations, difficult to be included in the theoretical picture. The dynamics including each single conformation is, thus, replaced by a coarse-grained one involving only the cluster centroids, which correspond to the representative conformations of each single metastable state. The average cluster size is selected to best replicate the MSD obtained from fine-grained conformational subspace, while still ensuring a reliable sampling of each cluster. The MSD calculations are performed using a moving average to cancel out the dependence from the initial conditions and it reads

$$\mathrm{MSD} = \frac{1}{T-t}\int_0^{T-t} d\tau |\hat{\boldsymbol{X}}(t+\tau) - \hat{\boldsymbol{X}}(\tau)|^2, \qquad (5)$$

where $t < 0.01T$, where T is maximum simulation time (1 μs). α is extracted from a linear regression of the log-log plot of MSD against time [Fig. 1(b)]. We found that setting ε within a range from 1.0 to 0.2, produces converged α values as defined by the relative difference $\Delta = \frac{|\alpha - \alpha_{\mathrm{fine}}|}{\alpha_{\mathrm{fine}}}$ where α and α_{fine} are the exponents calculated for coarse- and fine-grained conformational space, respectively. The small Δ value [Fig. 1(b)] indicates that ignoring the conformational oscillations around the individual centroids, as a consequence of the coarse-graining, leads to a negligible error in the MSD. To quantify the reliability of the linear fits in this study, two statistical measures were calculated: (1) the coefficient of determination (R^2) and (2) the mean square error normalized over the difference between the maximum and minimum values of the interested quantity (NRMSE). Regarding the MSD, it turned out that R^2 is very small, very close to unity, as well as the NRMSE, below 2% [Fig. 1(b)], indicating the accuracy of the linear fits. Furthermore, as a visual example, we showed the trajectory of the Villin and its related probability surface in

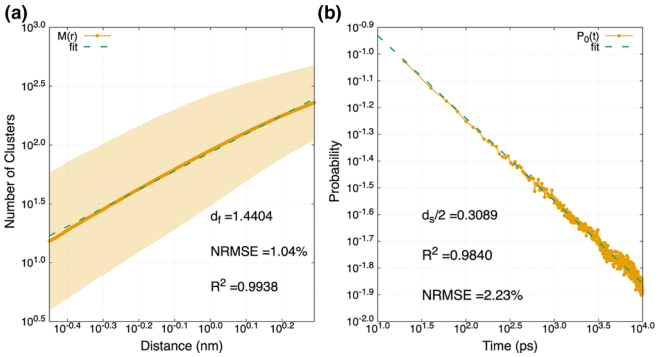

FIG. 3. $M(r)$ and $P_0(t)$ profiles and the associated linear fit for Villin. The yellow-shaded bands in the $M(r)$ plot indicate the standard deviation from the mean values.

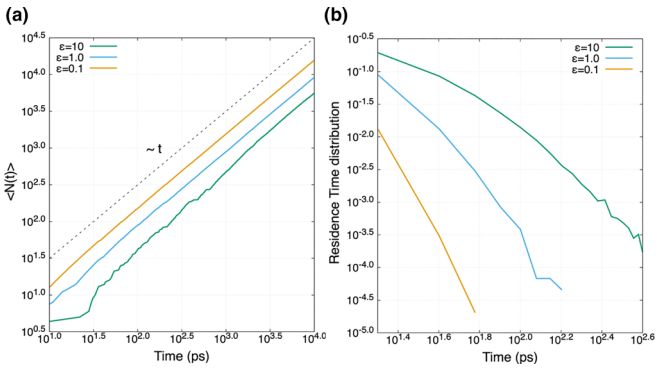

FIG. 4. (a) $\langle N(t)\rangle$ and (b) residence time distribution for Villin with three different average cluster sizes. Any regions of the residence time distribution show an independence from the cluster size, which would be a clear sign of the existence of a power-law profile. H4 and PDZ plots exhibit the same characteristics (Figs. 13 and 14).

a reduced two-dimensional conformational space [Figs. 2(a) and 2(b)].

The distribution of metastable states over the conformational space is connected to d_f, since this exponent relates the number of cluster centroids (M) within a sphere to its radius r as $M \sim r^{d_f}$ [22,23]. Therefore, we counted the number of clusters comprised by a sphere centered on a cluster centroid with a radius r. We repeated this procedure for all the centroids and for different values of r. Averaging over all the centroids, we obtained the profile shown in Fig. 3(a) as log-log plots whose linear fit provides the d_f values. It should be noted that the above power-law relationship does not hold to the entire range of r values. In fact, since the distribution of metastable states is a discrete and finite set of points, for both small and large values of r the $M(r)$ profiles show a "plateau," which invalidates the power-law relationship. To avoid these artifacts, we performed the linear fit neglecting these extreme value ranges. The interval of interest is chosen to include as many clusters as possible while minimizing the largest absolute error between the $M(r)$ profile and the linear fit. Setting the range of r values to exclude the first 5% and include 70% of the total number of clusters proves to be the best choice, as it produces a small error, which, normalized over the number of included clusters, is around 5% for all the systems, while including a significant number of clusters (see Appendix D for details). A homogeneous object in a two-dimensional space should have d_f equal to the dimension of the space, namely, 2. Unlikely, it turns out that the distributions of metastable states always have values less than 2, indicating their fractal nature [22,23].

During its dynamics, the protein jumps among these states whose accessibility is dictated by the potential energy barriers separating them. According to the theory of diffusion on fractals, after a time t the probability that the protein jumps back to the starting state (P_0) is related to d_s as $P_0 \sim t^{-d_s/2}$ [22,23,32]. We exploited this relation to evaluate this exponent. In our case the "starting point" coincides with the starting metastable states (i.e., starting cluster). Hence, we introduced $C(t+\tau, \tau)$, which is a function equal to 1 if the clusters visited at $t+\tau$ and at τ are the same and 0 otherwise. P_0 is calculated using a moving average as follows:

$$P_0(t) = \frac{1}{T-t}\int_0^{T-t} d\tau C(t+\tau,\tau). \quad (6)$$

The log-log plots of P_0 show a very good agreement with a power-law relation [Fig. 3(b)], and a linear regression of those profiles provides the values of d_s. It should be noted that the calculation of P_0 is independent from the metastable states distribution as it is only affected by their relative accessibility, which is regulated by the potential energy barriers. However, the nature of these barriers is still unclear. Indeed, they can originate from potential wells, related to each individual state, that hinder the dynamics regardless of the directions the protein explores in the space. This scenario is associated with the well-known continuous-time random walk (CTRW) model, which prescribes that the distribution of residence time (i.e., the probability the protein spends a particular time span in the same state) follows a power law. This produces an average number of jumps between states ($\langle N(t)\rangle$) that exhibits a sublinear relation with time, $\langle N(t)\rangle \sim t^{\alpha}$, which determines the subdiffusion [33]. Conversely, the barrier heights might depend on the couple of metastable states involved in each protein jump. In this case, barriers resemble walls that force protein to follow windy paths and slow down the exploration process. Therefore, it is important to evaluate the contribution of these two cases for providing a more detailed description of subdiffusion.

The role played by the potential wells related to each single state is evaluated by verifying whether CTRW can adequately model protein dynamics. This is done by directly calculating $\langle N(t)\rangle$ as follows:

$$\langle N(t)\rangle = \frac{1}{T-t}\int_0^{T-t} dt' \int_{t'}^{t+t'} dt'' |C(\tau_s + t'', t'') - 1|, \quad (7)$$

where τ_s is the minimum time step we can observe a jump between two states, which corresponds to the sampling time of MD simulations, which is 20 ps (see Appendix A). This quantity exhibits a time linear dependence as shown by the log-log plots in Fig. 4(a), independently from the size of clusters. Moreover, the residence time distribution does not exhibit a power-law relation as shown by its dependency from the average cluster size, which more resembles an exponential decay [Fig. 4(b)]. Both of those findings are in contrast with the CTRW model prescription and rule out the possibility that single states stability can fully explain the subdiffusive

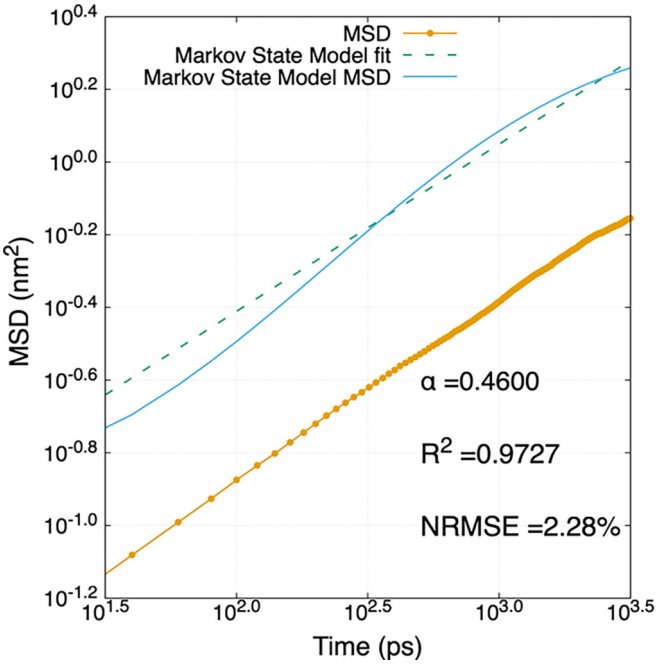

FIG. 5. Comparison between Markov state model MSD and the one calculated from MD simulations for Villin.

behavior indicating, thus, the exploration is mainly driven by path-dependent energy barriers.

Yet, this description raises another question on whether the paths proteins follow depend on their "past." More technically, how much the past visited states at $t_0 < t_1 < t_2 \cdots < t$ determine the protein state at t. This question is thus focused on the memory effect protein dynamic might exhibit [17] and it becomes even more relevant noticing that the theory of diffusion on fractals is entirely focused on Markov processes, which are memory-free processes [22]. To answer this question, we built a Markov state model (MSM), directly from MD simulations [34]. A MSM is uniquely defined by the transition matrix $\boldsymbol{T}$ whose entries, T_{ij} representing transition probabilities between states i and j, can be computed as

$$T_{ij} = \frac{s_{ij}}{\sum_{i=1}^{N_c} s_{ij}} \tag{8}$$

where s_{ij} is the number of jumps between the ith and jth metastable states and N_c is the total number of clusters. Through MSM we described a discrete time Markov process. The time resolution of this MSM corresponds to our MD trajectories sampling time, namely, 20 ps. After n steps the probability to find the protein in a particular state is included in the N_c entries vector $\boldsymbol{P}(n)$, which is propagated as usual, $\boldsymbol{P}(n) = \boldsymbol{T}^n \cdot \boldsymbol{P}(0)$. In this case the MSD is calculated as

TABLE I. Comparison between α_{Markov} and α_{fit} calculated for the different systems under investigations.

System	α_{Markov}	α_{fit}
H4	0.4664	0.4877
Villin	0.4600	0.4140
PDZ	0.5169	0.4681

TABLE II. Summarizing table showing all the evaluated exponents for all the systems under investigation and comparing the α coming from the theory and extracted directly from the fit with MSD (α_{fit}).

System	d_f	d_s	$\alpha = d_s/d_f$	α_{fit}
H4	1.4404	0.7050	0.4894	0.4877
Villin	1.5099	0.6178	0.4092	0.4140
PDZ	1.2220	0.5856	0.4792	0.4681

follows:

$$\text{MSD}_{\text{Markov}}(n) = \sum_{i=1}^{N_c}\sum_{j=1}^{N_c} d_{ij}^2 (\boldsymbol{T}^n)_{ij} P_i(\infty), \tag{9}$$

where d_{ij} is the distance in the conformational space between the ith and jth cluster centroid and $P_i(\infty)$ is the ith entry of the stationary distribution $\boldsymbol{P}(\infty)$. A linear fit of the log-log plot of $\text{MSD}_{\text{Markov}}$ against time provides the α exponent associated to MSM (α_{Markov}) (Fig. 5 and Table I). It turned out that for all the investigated systems these values are very close to those calculated for the MD simulation (Table I) showing that even a simple memory-free model such as MSM can exhibit the "correct" subdiffusive behavior. The fact that $\text{MSD}_{\text{Markov}}$ and MSD profiles do not overlap is mainly associated to already known discretization errors [35–37] (see Appendix F for details).

The values of d_s and d_f are summarized in Table II where we compared the subdiffusion exponent α_{fit} with the ratio d_s/d_f. Clearly, protein dynamics follows the subdiffusion paradigm as indicated by a values in the range 0.4–0.5, and the comparison between $\alpha = d_s/d_f$ and α_{fit} values shows excellent agreement.

III. CONCLUSION

Therefore, in conclusion, we provided compelling evidence that subdiffusive protein dynamics, as described by MD simulations, originates from the fractal nature of the conformational space. The high-dimensional and rough potential energy landscape gives rise to separated basins of attractions, namely, metastable states whose distribution in the conformational space, regulated by d_f, resembles a fractal structure. The paths among these states, followed by the protein during its dynamics, are shaped by the energy barriers, which determine the relative states connectivity and, thus, the d_s values. Therefore, as prescribed by the theory of diffusion on fractal structure, the subdiffusion in protein can be described by the exponents d_f and d_s, which are directly connected to the landscape of the potential energy surface.

ACKNOWLEDGMENTS

L.M. and M.O. acknowledge support by the Spanish "Ministerio de Ciencia e Innovación" (PID2020-116620GB-I00, RTI2018-096704-B-100, PID2021-122478NB-I00); the Center of Excellence for HPC H2020 European Commission; "BioExcel-3. Centre of Excellence for Computational Biomolecular Research" (823830); Catalan SGR and the Instituto de Salud Carlos III–Instituto Nacional de Bioin-

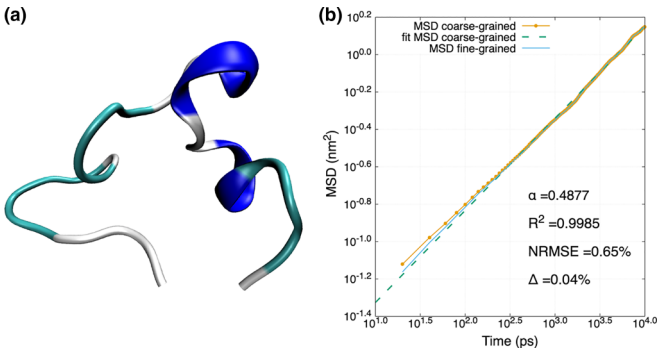

FIG. 6. Snapshot extracted from MD simulations of (a) H4 and (b) the associated MSD profile linearly fitted. Plot shows the R^2 and the NRMSE relative to the linear regression computed on coarse-grained MSD.

formatica (ISCIII PT 17/0009/0007, cofunded by the Fondo Europeo de Desarrollo Regional); the European Regional Development Fund under the framework of the ERFD Operative Programme for Catalunya, the Catalan Government AGAUR (SGR2017-134); and the European Union "MDDB: Molecular Dynamics Data Bank. The European Repository for Biosimulation Data" (101094651).

APPENDIX A: MOLECULAR DYNAMICS DETAILS

All the simulations presented here have been carried out employing the GROMACS 2020.2 code [38]. The initial protein structures of PDZ and Villin are taken from the Protein Data bank whereas the H4 has been manually reconstructed using Avogadro [39]. They are placed in simulation boxes of $41 \times 41 \times 41$, $47 \times 47 \times 47$, and $70 \times 70 \times 70\,\mathrm{A}^3$ for H4, Villin, and PDZ, respectively. All the systems are filled with TIP3P water molecules. Na^+ and Cl^- ions were added to neutralize the system and bring the salt concentration to a physiological level. All the simulations are performed using Amber ff19SB force field [40] and LINCS [41] to constrain all the bonds involving hydrogens allowing us to employ a 2-fs step to integrate the Newton equations. The mesh Ewald method has been used to account for long-range interactions with a real-space cutoff of 12 A. All the systems are equilibrated running a 50-ns-long simulation of an *NVT* ensemble

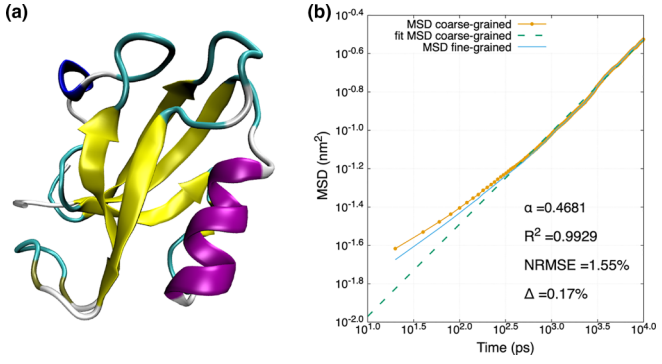

FIG. 7. Snapshot extracted from MD simulations of (a) PDZ and (b) the associated MSD profile linearly fitted. Plot shows the R^2 and the NRMSE relative to the linear regression computed on coarse-grained MSD.

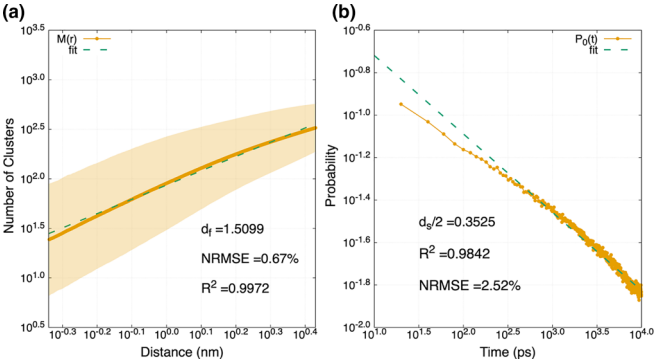

FIG. 8. $\boldsymbol{M(r)}$ and $\boldsymbol{P_0(t)}$ profiles and the associated linear fit for H4. The yellow-shaded bands in the $M(r)$ plot indicate the standard deviation from the mean values.

($T = 310$ K) using the velocity-rescaling algorithm to control the temperature with a coupling constant of 0.4 ps, followed by a 50-ns *NPT* ensemble ($T = 310$ K, $P = 1$ atm) employing a Nose-Hoover thermostat [42] and a Berndsen barostat [43], with a coupling constant of 0.4 and 0.6 ps, respectively. Eventually, the results presented were sampled every 20 ps from the 1-μs production simulation, which has been carried out in an isothermal isobaric ensemble ($T = 310$ K, $P = 1$ atm) using a Nose-Hoover thermostat and the Parinello-Raman algorithm to control pressure [44] with a coupling constant of 0.6 ps.

APPENDIX B: MSD, $M(r)$, AND P_0 PROFILES FOR H4 AND PDZ

Figures 6–9.

APPENDIX C: SUBDIFFUSION ANLAYSIS OF VILLIN WITH DIFFERENT DESCRIPTORS

To corroborate the results presented in the main text, we repeated the same analysis on Villin using different descriptors for the conformational space. Here we focused on the *rmC*α carbon root-mean-square-deviation (RMSD) of two segments of the protein calculated from the reference structure shown in Fig. 1(b) in the main text, as done by others to describe its dynamics [45]. The first one, Segment A, comprises the first 17 residues and the rest belong to Segment B. Figure 10 shows the probability distribution extracted from the simulations and

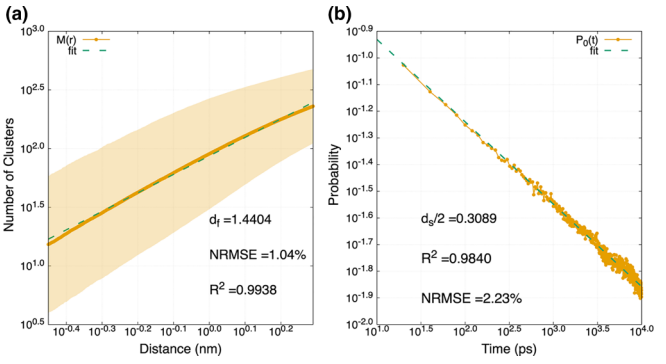

FIG. 9. $\boldsymbol{M(r)}$ and $\boldsymbol{P_0(t)}$ profiles and the associated linear fit for PDZ. The yellow-shaded bands in the $M(r)$ plot indicate the standard deviation from the mean values.

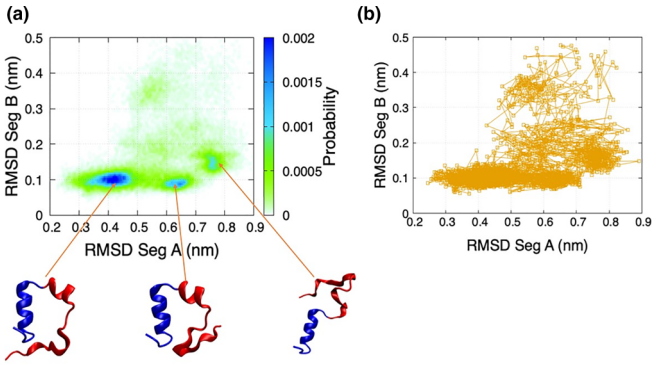

FIG. 10. (a) The probability distribution of Villin conformations studied in RMSD Segments A (red) and B (blue). (b) The Villin dynamics over multiple conformations.

its dynamics within it. Based on this descriptor, we performed the hierarchical agglomerative clustering, setting $\varepsilon = 0.2$, and extracted all the exponents (α, α_{fit}, d_f, and d_s) from the profile of MSD, $M(r)$, and $P_0(t)$ (Fig. 11). We obtained $\alpha = \frac{d_s}{d_f} = 0.3904$ very close to $\alpha_{\text{fit}} = 0.3752$. As stated in the main text, other quantities can be chosen producing the same results, and the advantage of using PCs is purely technical.

APPENDIX D: $M(r)$ PROFILE

In order to avoid artifacts at the extreme range r values, we restricted the linear fit in a value range, which excludes the first 5% and includes 70% of the total number of clusters. Figure 12 shows the $M(r)$ profile over all the range of r. We assessed the discrepancy between the linear fit (M_{fit}) and $M(r)$ computing the maximum absolute errors between these two functions normalized over the total number of clusters comprised by the linear fit ($M_{\text{tot-fit}}$), namely,

$$\frac{\max_r \{|M(r) - M_{\text{fit}}(r)|\}}{M_{\text{tot-fit}}}. \tag{D1}$$

We set a conservative cut-off value for this quantity at around 5% and the range mentioned above provides an error

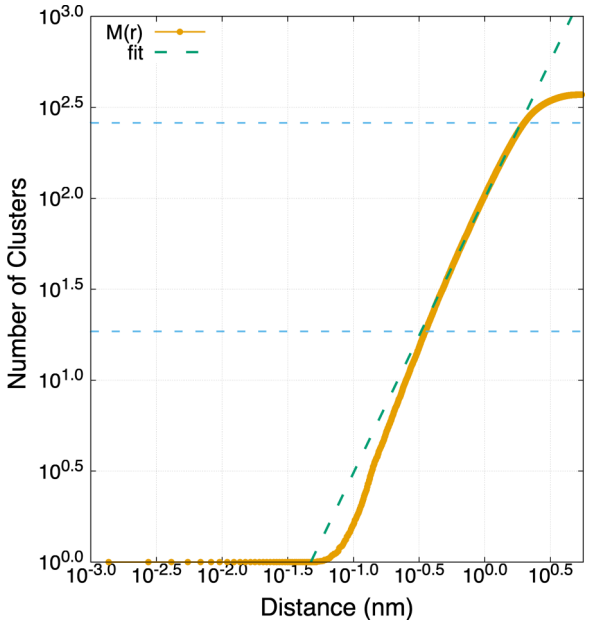

FIG. 12. $M(r)$ Villin profile. The dashed light-blue line represents the 5% and 70% of the total number of clusters.

of this entity for all the systems considered, being 4.5%, 5.1%, and 4.6% for H4, Villin, and PDZ, respectively. Furthermore, it should be noted that the maximum r values we selected (∼2 nm for each system) are always larger than the maximum displacement (∼1 nm), which ensure the validity of the power-law relationship at least in the range defined by the maximum displacement considered.

APPENDIX E: H4 AND PDZ RESIDENCE TIME DISTRIBUTION AND $\langle N(t) \rangle$

Figures 13 and 14.

APPENDIX F: MARKOV STATE MODEL MSD

As stated in the main text, the MSM is totally defined by the transition matrix T from which we calculated

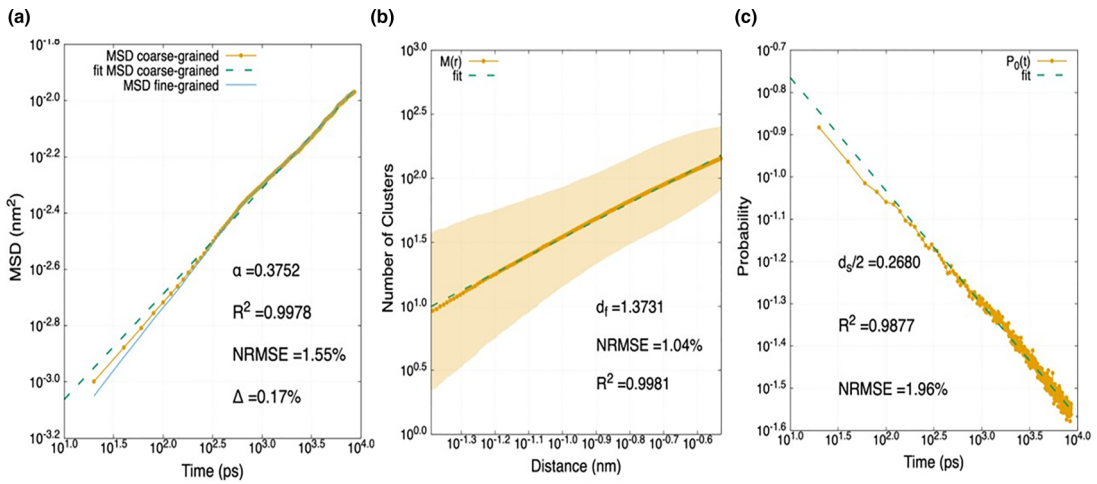

FIG. 11. (a) Coarse- and fine-grained MSD. (b) $M(r)$ and (c) $P_0(t)$ profiles with the associated linear fits for the Villin. Yellow shading on $M(r)$ represents the standard deviation from the mean values.

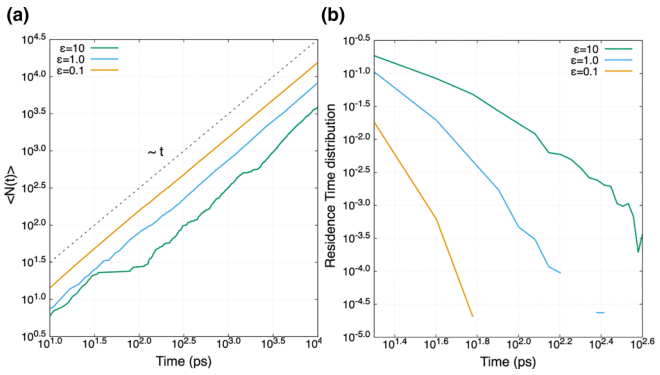

FIG. 13. (a) $\langle N(t)\rangle$ and (b) residence time distributions for H4.

the stationary distribution $P(\infty)$, and using the matrix of the distances between all the metastable states d, we can compute the $\mathrm{MSD}_{\mathrm{Markov}}$. The matrix T can be decomposed by its eigenvalues (σ_k) and left (right) eigenvectors ($\boldsymbol{v}_k^{l(r)}$):

$$T = \sum_k \sigma_k \boldsymbol{v}_k^l \otimes \boldsymbol{v}_k^r, \tag{F1}$$

where k runs over the total number of eigenvalues. Therefore, exploiting the orthonormality between right and left eigenvectors, the MSD can be rewritten as

$$\mathrm{MSD}_{\mathrm{Markov}}(n) = \sum_{i,j} P_i(\infty) d_{ij}^2 \left\{ \sum_k \sigma_k^n \left[\boldsymbol{v}_k^l \otimes \boldsymbol{v}_k^r\right]_{ij} \right\}. \tag{F2}$$

Changing the order of summation, one has

$$\mathrm{MSD}_{\mathrm{Markov}}(n) = \sum_k \sigma_k^n \left\{ \sum_{i,j} P_i(\infty) d_{ij}^2 \left[\boldsymbol{v}_k^l \otimes \boldsymbol{v}_k^r\right]_{ij} \right\}. \tag{F3}$$

The term in curly brackets does not depend on the number of steps n. On the log-log plots shown in the main text, it produces a translation of the MSD profile. Even though the MSM can reproduce some properties of the modeled system, the discretization of the conformational space generates errors on the eigenvectors $\boldsymbol{v}_k^{l(r)}$ with respect to the continuous eigenfunctions associated to the operator generating the "real dynamics" that originates the MSD profile shown in the main

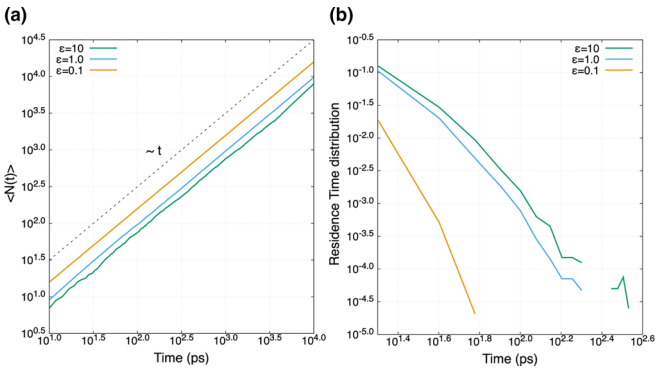

FIG. 14. (a) $\langle N(t)\rangle$ and (b) residence time distributions for PDZ.

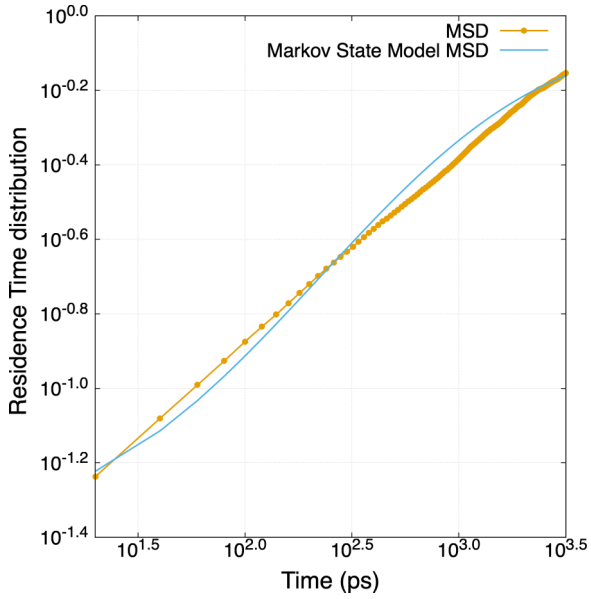

FIG. 15. The MSD profile obtained by a MSM translated downwards of 0.42 on a $\log_{10}$ scale. The overlaps between the two profiles are very good over the whole range of simulation time.

text. These errors are well known and documented elsewhere [35,36]. This explains the two profiles shown in the main text do not overlap. A simple translation of the logarithmic scale $\mathrm{MSD}_{\mathrm{Markov}}$ plots produces, indeed, a very good overlap as shown in Fig. 15. The discrepancy on α_{Markov}'s are mainly due to the finite dimension of T and the discretization errors. Furthermore, Fig. 16 shows the MSD profile for H4 and PDZ.

APPENDIX G: ANALYSIS TOOLS

The SCIKIT package [46] is employed to implement the hierarchical clustering algorithm. All the linear regressions are made using GNUPLOT and the rest of the analyses are the results of in-house scripts.

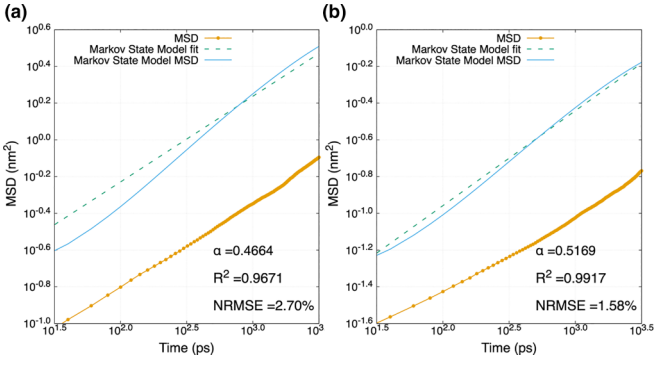

FIG. 16. Comparison between Markov state model MSD and the one calculated form MD simulations for (a) H4 and (b) PDZ.

[1] K. Henzler-Wildman and D. Kern, Dynamic personalities of proteins, Nature (London) **450**, 964 (2007).

[2] 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. USA **82**, 5000 (1985).

[3] S. D. Schwartz, Protein dynamics and enzymatic catalysis, J. Phys. Chem. B **127**, 2649 (2023).

[4] Y. Wang, K. Bugge, B. B. Kragelund, and K. Lindorff-Larsen, Role of protein dynamics in transmembrane receptor signalling, Curr. Opin. Struct. Biol. **48**, 74 (2018).

[5] A. B. Kolomeisky and M. E. Fisher, Molecular motors: A theorist's perspective, Annu. Rev. Phys. Chem. **58**, 675 (2007).

[6] K. Bartels, T. Lasitza-Male, H. Hofmann, and C. Löw, Single-molecule FRET of membrane transport proteins, ChemBioChem **22**, 2657 (2021).

[7] L. Milanesi, J. P. Waltho, C. A. Hunter, D. J. Shaw, G. S. Beddard, G. D. Reid, S. Dev, and M. Volk, Measurement of energy landscape roughness of folded and unfolded proteins, Proc. Natl. Acad. Sci. USA **109**, 19563 (2012).

[8] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, The energy landscapes and motions of proteins, Science **254**, 1598 (1991).

[9] H. Yang, G. Luo, P. Karnchanaphanurach, T.-M. Louie, I. Rech, S. Cova, L. Xun, and X. S. Xie, Protein conformational dynamics probed by single-molecule electron transfer, Science **302**, 262 (2003).

[10] I. Grossman-Haham, G. Rosenblum, T. Namani, and H. Hofmann, Slow domain reconfiguration causes power-law kinetics in a two-state enzyme, Proc. Natl. Acad. Sci. USA **115**, 513 (2018).

[11] G. Luo, I. Andricioaei, X. S. Xie, and M. Karplus, Dynamic distance disorder in proteins is caused by trapping, J. Phys. Chem. B **110**, 9363 (2006).

[12] S. C. Kou and X. S. Xie, Generalized Langevin equation with fractional Gaussian noise: Subdiffusion within a single protein molecule, Phys. Rev. Lett. **93**, 180603 (2004).

[13] Y. Wang and H. P. Lu, Bunching effect in single-molecule T4 lysozyme nonequilibrium conformational dynamics under enzymatic reactions, J. Phys. Chem. B **114**, 6669 (2010).

[14] I. E. T. Iben *et al.*, Glassy behavior of a protein, Phys. Rev. Lett. **62**, 1916 (1989).

[15] 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).

[16] V. Calandrini, D. Abergel, and G. R. Kneller, Fractional protein dynamics seen by nuclear magnetic resonance spectroscopy: Relating molecular dynamics simulation and experiment, J. Chem. Phys. **133**, 145101 (2010).

[17] C. Ayaz, L. Tepper, F. N. Brünig, J. Kappler, J. O. Daldrop, and R. R. Netz, Non-Markovian modeling of protein folding, Proc. Natl. Acad. Sci. USA **118**, e2023856118 (2021).

[18] R. Granek and J. Klafter, Fractons in proteins: Can they lead to anomalously decaying time autocorrelations? Phys. Rev. Lett. **95**, 098106 (2005).

[19] S. Reuveni, R. Granek, and J. Klafter, Anomalies in the vibrational dynamics of proteins are a consequence of fractal-like structure, Proc. Natl. Acad. Sci. USA **107**, 13696 (2010).

[20] G. H. Weiss and R. J. Rubin, Random walks: Theory and selected applications, *Advances in Chemical Physics* (Wiley, New York, 1982), pp. 363–505.

[21] U. Mosco, Invariant field metrics and dynamical scalings on fractals, Phys. Rev. Lett. **79**, 4067 (1997).

[22] M. T. Barlow, Diffusions on fractals, *Lectures on Probability Theory and Statistics. Lecture Notes in Mathematics: Ecole d'Eté de Probabilités de Saint-Flour XXV - 1995* (Springer, New York, 2006).

[23] S. Havlin and D. Ben-Avraham, Diffusion in disordered media, Adv. Phys. **51**, 187 (2002).

[24] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, NMR structure of the 35-residue villin headpiece subdomain, Nat. Struct. Biol. **4**, 180 (1997).

[25] G. Kozlov, D. Banville, K. Gehring, and I. Ekiel, Solution structure of the PDZ2 domain from cytosolic human phosphatase HPTP1E complexed with a peptide reveals contribution of the B2–B3 loop to PDZ domain–ligand interactions, J. Mol. Biol. **320**, 813 (2002).

[26] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Essential dynamics of proteins, Proteins Struct. Funct. Bioinf. **17**, 412 (1993).

[27] C. C. David and D. J. Jacobs, Principal component analysis: A method for determining the essential dynamics of proteins, *Methods in Molecular Biology* (Humana, Totowa, NJ, 2014), Vol. 1084, pp. 193–226.

[28] R. K. Bijral, J. Manhas, and V. Sharma, Hierarchical clustering based characterization of protein database using molecular dynamic simulation, *Recent Innovations in Computing, Lecture Notes in Electrical Engineering* (Springer, Singapore, 2022), Vol. 832, pp. 427–437.

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2009).

[30] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data* (Wiley, New York, 1990).

[31] S. Mannor, X. Jin, J. Han, X. Jin, J. Han, X. Jin, J. Han, and X. Zhang, *K*-means clustering, *Encyclopedia of Machine Learning* (Springer US, Boston, MA, 2011), pp. 563–564.

[32] S. Alexander, J. Bernasconi, W. R. Schneider, and R. Orbach, Excitation dynamics in random one-dimensional systems, Rev. Mod. Phys. **53**, 175 (1981).

[33] I. M. Sokolov, Models of anomalous diffusion in crowded environments, Soft Matter **8**, 9043 (2012).

[34] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, Markov models of molecular kinetics: Generation and validation, J. Chem. Phys. **134**, 174105 (2011).

[35] J. D. Chodera and F. Noé, Markov state models of biomolecular conformational dynamics, Curr. Opin. Struct. Biol. **25**, 135 (2014).

[36] N. Kozlowski and H. Grubmüller, Uncertainties in Markov state models of small proteins, J. Chem. Theory Comput. **19**, 5516 (2023).

[37] E. Suárez, R. P. Wiewiora, C. Wehmeyer, F. Noé, J. D. Chodera, and D. M. Zuckerman, What Markov state models can and cannot do: Correlation versus path-based observables in protein-folding models, J. Chem. Theory Comput. **17**, 3119 (2021).

[38] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindah, Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX **1–2**, 19 (2015).

[39] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, Avogadro: An advanced semantic chemical editor, visualization, and analysis platform, J Cheminform **4**, 17 (2012).

[40] C. Tian *et al.*, Ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution, J. Chem. Theory Comput. **16**, 528 (2020).

[41] 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).

[42] D. J. Evans and B. L. Holian, The Nose-Hoover thermostat, J. Chem. Phys. **83**, 4069 (1985).

[43] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys. **81**, 3684 (1984).

[44] M. Parrinello and A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, J. Appl. Phys. **52**, 7182 (1981).

[45] H. Lei, C. Wu, H. Liu, and Y. Duan, Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations, Proc. Natl. Acad. Sci. USA **104**, 4925 (2007).

[46] F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python, J. Mach. Learn Res. **12**, 2825 (2011).