Comparison between slow, anisotropic LE4PD fluctuations and the Principal Component Analysis modes of Ubiquitin

E. R. Beyerle[1] and M. G. Guenza[1, a)]
Institute for Fundamental Science and Department of Chemistry and Biochemistry, University of Oregon, Eugene, Oregon 97403, USA

(Dated: 16 February 2021)

Proteins' biological function and folding mechanisms are often guided by large-scale, slow motions, which involve crossing high energy barriers. In a simulation trajectory, these slow fluctuations are commonly identified using a principal component analysis (PCA). Despite the popularity of this method, a complete analysis of its predictions based on the physics of protein motion has been so far limited. This study formally connects the PCA to a Langevin model of protein dynamics and analyzes the contributions of energy barriers and hydrodynamic interactions to the slow PCA modes of motion. To do so, we introduce an anisotropic extension of the Langevin Equation for Protein Dynamics, called the LE4PD-XYZ, which formally connects to the PCA 'essential dynamics'. The LE4PD-XYZ is an accurate coarse-grained diffusive method to model protein motion, which describes anisotropic fluctuations in the protein's alpha-carbons. The LE4PD accounts for hydrodynamic effects and mode-dependent free-energy barriers. This study compares large-scale anisotropic fluctuations identified by the LE4PD-XYZ to the mode-dependent PCA's predictions, starting from a microsecond-long alpha-carbon molecular dynamics atomistic trajectory of the protein ubiquitin. We observe that the inclusion of free-energy barriers and hydrodynamic interactions has important effects on the identification and timescales of ubiquitin's slow modes.

I. INTRODUCTION

Proteins are semi-flexible objects whose function is determined by the combined effect of their three-dimensional structure and local fluctuations[1]. Large-scale, slow motions relevant to the protein's biological function typically involve crossing high energy barriers and transitions between minima on the protein's free energy surface (FES). The complexity of the FES renders those large-scale fluctuations both anharmonic and anisotropic[2–5]. A common technique used to determine the slow, functional motions of proteins is the Principal Component Analysis (PCA)[3,5–8]. PCA is a dimensionality reduction procedure commonly used in signal analysis to highlight important persistent features underlying noisy data.[9] When used to process atomistically-detailed molecular dynamics (MD) simulations, PCA reduces the dimensionality of the observed FES by identifying a few essential collective fluctuations ordered by decreasing eigenvalues.[10] While PCA is computationally convenient, conceptually simple, and widely used, it lacks a physical basis beyond the empirical observations that it describes some slow, functional motions of a protein[7,11,12].

In this study, we make a formal connection between the PCA expressed in the cartesian coordinates of a protein's alpha-carbons and an approach we have developed to analyze the slow modes in protein dynamics, called the Langevin Equation for Protein Dynamics (LE4PD)[13–18]. The LE4PD theory, initially formalized as an isotropic equation of motion, is extended here to describe the anisotropic dynamics of proteins in the LE4PD-XYZ method. Like the PCA, the LE4PD-XYZ decomposes the protein's motion into an orthogonal set of collective coordinates or modes, and, it captures the anisotropic, slow fluctuations of the protein, starting from the analysis of the atomistic MD trajectory. However, unlike the PCA, the LE4PD-XYZ is based on a Langevin equation of motion, which directly connects the large-scale fluctuations to the physical forces acting on the system. Because of its formulation, the LE4PD-XYZ allows for a detailed examination of the mode-dependent kinetics and fluctuation pathways.

Since the PCA is not directly related to an equation of motion, the amplitudes and timescales of fluctuations may be calculated by different procedures. For example, timescales have been calculated by integrating the autocorrelation function of the principal components.[4,19] The direction and magnitude of the anisotropic fluctuations have been described using either the so-called 'porcupine plots'[19–22] or a simple linear interpolation between the extreme structures in the simulation trajectory[23]. In this manuscript, we compare the predictions of the PCA approach for the fluctuations with the largest amplitude to the results of the LE4PD theory applied to the same trajectory. By this analysis we quantify the importance of the different physical contributions to the PCA fluctuations, starting from the LE4PD-XYZ equation of motion.

The LE4PD-XYZ projects the simulation trajectory onto a mode-dependent free-energy surface. For each Langevin diffusive mode, our theory determines the three-dimensional energy landscape and the kinetic pathways of barrier crossing using a variant of the string method.[24,25] Originally, the LE4PD measured the related kinetic times of barrier crossing by a simple Kramer's rescaling of the friction coefficient, where we calculated the correction term from the height of the barrier, defined by the median absolute deviation (MAD)[16]. More recently, we paired the LE4PD with a Markov state model (MSM) analysis of the mode-dependent barrier crossing events. We then evaluated the mode-dependent kinetic times using the eigenspectrum of the slowest process predicted by the MSM analysis, and interpreted the results using the associated committor function.[18,26–28] A similar analysis

[a)]Electronic mail: mguenza@uoregon.edu.

can be performed for the LE4PD-XYZ modes, and its comparison with the PCA modes is one of this study's primary goals.

This study illustrates the advantages and limitations of the PCA normal mode decomposition compared with the Langevin formalism of the anisotropic LE4PD. PCA does not provide information on the timescales of the fluctuations. However, the LE4PD-XYZ equation, when hydrodynamic interactions are neglected, has forces acting between the amino acids that are directly related to the covariance matrix, and thus to PCA. The test system is a 1-μs MD trajectory of atomistic simulation of ubiquitin in an aqueous solution and physiological salt concentration. From its analysis we calculate the mode-dependent FES, its distinct pathways for protein fluctuations, and the related timescales using both the LE4PD-XYZ with and without hydrodynamics interactions. Then, we directly compare PCA's linear fluctuations to the non-linear fluctuation pathway of LE4PD-XYZ. This analysis identifies the implications of neglecting hydrodynamic interactions and free energy barriers when PCA is extended to treat protein dynamics, by mapping the covariance matrix into the intramolecular matrix of the forces leading to the Langevin equation of motion.

II. THEORY: THE LE4PD-XYZ EQUATION OF MOTION

In this section, we introduce the anisotropic Langevin equation for protein dynamics, or LE4PD-XYZ. The LE4PD-XYZ is a coarse-grained Langevin equation describing the fluctuations of the ith alpha-carbon in a protein composed of N residues, and hence N alpha-carbons, from its equilibrium position, $\Delta\vec{R}_i(t) = \vec{R}_i(t) - \langle\vec{R}_i\rangle = [x_i(t) - \langle x_i\rangle, y_i(t) - \langle y_i\rangle, z_i(t) - \langle z_i\rangle]^T = [\Delta x_i, \Delta y_i, \Delta z_i]^T$. The equilibrium positions are defined as the time average over an MD trajectory consisting of M configuration points, $\langle\vec{R}_i\rangle = \frac{1}{M}\sum_{k=1}^{M}\vec{R}_i(k)$, with $x_i(t)$, $y_i(t)$, $z_i(t)$ the distance of the ith alpha-carbon from the origin of the simulation box at time t in the x-, y-, or z-direction, respectively. For a protein with N alpha-carbons there are a total of $3N$ degrees of freedom in the analysis, which is represented in the LE4PD-XYZ by the $3N$-dimensional vector $\Delta\vec{R}(t)$:

$$\Delta R(t) = [x_1(t) - \langle x_1\rangle, y_1(t) - \langle y_1\rangle, z_1(t) - \langle z_1\rangle, x_2(t) - \langle x_2\rangle, \ldots, z_N(t) - \langle z_N\rangle]^T \ .$$

In the LE4PD-XYZ model, the time-evolution of a $\Delta\vec{R}_i(t)$ along the α direction, $\Delta\vec{R}_i^{\alpha}(t)$, is given by the Langevin equation of motion

$$\frac{d\Delta\vec{R}_i^{\alpha}(t)}{dt} = -\frac{k_B T}{\overline{\zeta}} \sum_{\beta,\gamma\in\{x,y,z\}} \sum_{i=1}^{N}\sum_{j=1}^{N} H_{ij}^{\alpha\beta} A_{jk}^{\beta\gamma} \Delta\vec{R}_k^{\gamma}(t) + \vec{\Delta v}_i^{\alpha}(t), \tag{1}$$

with $\alpha,\beta,\gamma \in \{x,y,z\}$ the coordinates in the three spatial dimensions. The equation is solved by applying the fluctuation-dissipation condition, as described in the Supplementary Material.

Here k_B is the Boltzmann constant, T is the temperature in Kelvin, $\overline{\zeta} = \frac{1}{N}\sum_i \zeta_i$ denotes the average residue friction coefficient, $\vec{v}_i(t)$ is a stochastic velocity, $H_{ij}^{\alpha\beta}$ denotes the hydrodynamic interaction (HI) between the α component on bead i and the β component on bead j and $A_{jk}^{\beta\gamma}$ denotes the connectivity between the β component on residue j and the γ component on residue k. In Eq. 1, the hydrodynamic interaction matrix $H_{ij}^{\alpha\beta}$ is given by

$$H_{ij}^{\alpha\beta} = \frac{\overline{\zeta}}{\zeta_i}\delta_{ij}\delta_{\alpha\beta} + (1-\delta_{ij})\delta_{\alpha\beta}\bar{r}_w \left\langle \frac{1}{r_{ij}} \right\rangle , \tag{2}$$

where ζ_i the friction coefficient of residue i, $\langle\frac{1}{r_{ij}}\rangle$ is the average inverse distance between residues i and j, and $\bar{r}_w = \frac{1}{N}\sum_i r_{w,i}$ is the average residue radius exposed to the solvent.

The structural matrix, related to the mean-force potential, $A_{jk}^{\beta\gamma}$ is defined as

$$A_{jk}^{\beta\gamma} = \left(\left[a \otimes \widehat{I}\right]^T \mathbf{U} \left[a \otimes \widehat{I}\right] \right)_{jk}^{\beta\gamma}, \tag{3}$$

where $\mathbf{U}^{-1} = \langle \Delta\vec{l}(t)\ \Delta\vec{l}(t)^T \rangle$ is the matrix of bond-bond correlations in Cartesian coordinates with $\Delta\vec{l}(t) = \left(a \otimes \widehat{I}\right)\Delta\vec{R}(t)$, $\Delta\vec{l}_i^{\alpha}(t) = \sum_j \left(a \otimes \widehat{I}\right)_{ij} \delta_{\alpha\beta}\Delta\vec{R}_j^{\beta}(t)$, and a the $N-1 \times N$ the incidence matrix that defines the connectivity between residues in the protein,

$$a_{ij} = \begin{cases} 1, i = j-1 \\ 0, i = j \end{cases} .$$

Here, $\delta_{\alpha\beta}$ is the Kronecker delta, and the '$\otimes$' symbol denotes the Kronecker product.[29] A detailed derivation of **H** as well as a formal connection of the **U** matrix defined here to the **U** in the previously developed, isotropic version of the LE4PD[13–16] are given in the Supplementary Material.

It should be noted that the Langevin equation given in Eq.1 is identical in form to the optimized Rouse-Zimm equation for describing polymer dynamics derived by Zwanzig,[30] excepting the detailed form of the **H** and **A** matrices, which here account for the chemical details of each residue and the semi-flexibility of the peptide bonds connecting the alpha-carbons.[13–16,18,31] The Rouse-Zimm equation, without hydrodynamic interaction, can be derived from the Liouville equation, i.e. from the Hamiltonian of the system, by projecting the dynamics of the whole system onto the slow coordinates of the alpha carbons.[32,33] The Rouse-Zimm equation is equivalent to a Fokker-Planck-Smoluchowski for polymer dynamics.[34,35] In this respect, the LE4PD-XYZ equation presented here is founded on well-established first-principles approaches.

A. Connecting the LE4PD-XYZ to PCA

The link between a PCA of the alpha-carbons and the LE4PD-XYZ method outlined above is as follows. For a given protein, an element of the covariance matrix in the Cartesian coordinates of the alpha-carbons is given by

$$C_{ij}^{\alpha\beta} = \langle \Delta\vec{R}_i^{\alpha} \Delta\vec{R}_j^{\beta} \rangle \tag{4}$$

where $\langle \cdots \rangle$ denotes, as above, an average over frames in the simulation trajectory. Using the definition of $\mathbf{A}$ given in Eq. 3 and that of $\mathbf{C}$ given in Eq. 4 it follows that

$$\mathbf{A} = \mathbf{C}^{-1}, \tag{5}$$

Thus, $\mathbf{A}$ and $\mathbf{C}$ possess the same set of eigenvectors and their eigenvalues are inverses of each other (provided $\mathbf{C}$ has full rank, which is always the case for a sufficiently long MD simulation).

The set of coupled Langevin equations, in Eq. 1, is diagonalized by the eigenvector $\mathbf{Q}$, as $\mathbf{Q}^{-1}\mathbf{HAQ} = \Lambda$, with Λ the diagonal matrix of eigenvalues of $\mathbf{HA}$. By applying the eigenvector transformation, Eq.1 can be written in terms of its normal modes, $\Delta\vec{\xi}_a(t) = \sum_i \left(Q^{-1}\right)_{ai} \Delta\vec{R}_i(t)$ as

$$\frac{d\Delta\vec{\xi}_a(t)}{dt} = -\sigma_a \Delta\vec{\xi}_a(t) + \Delta\vec{v}_a(t), \tag{6}$$

where $\sigma_a = \frac{k_B T \lambda_a}{\overline{\zeta}}$ is the characteristic diffusive rate of mode a,[36] with $\lambda_a = (\Lambda)_{aa}$ the eigenvalue of mode a, and $\tau_a = \sigma_a^{-1} = \frac{\overline{\zeta}}{k_B T \lambda_a}$ is the corresponding diffusive, barrier-free timescale of mode a. Finally, $\vec{v}_a(t) = \sum_i \left(Q^{-1}\right)_{ai} \vec{v}_i(t)$ is the random velocity projected into mode coordinates. Note that $< \vec{\xi}_a(t) >= 0$, so $\Delta\vec{\xi}_a(t) = \vec{\xi}_a(t)$.

Because of Eq. 5, it is straightforward to see that, when the hydrodynamic interaction matrix is approximated as an identity matrix, the PCA modes directly map onto the LE4PD-XYZ modes. Approximating the hydrodynamic interaction matrix in this way corresponds to assuming that i) the friction coefficient of each residue is set equal to the average friction coefficient, i.e., $\zeta_i := \overline{\zeta}$, and that ii) the dynamical correlation due to hydrodynamics is neglected, i.e., $\mathbf{H} := \mathbf{I}$, with $\mathbf{I}$ a 3N×3N identity matrix. Under those approximations, the eigenvalues $(\mathbf{Q}^{-1}\mathbf{AQ})_{aa} = \mu_a$ where $\mu_a = \lambda_{PCA,a}^{-1}$ and $\lambda_{PCA,a}$ is the ath eigenvalue of the covariance matrix. In this manuscript, we analyze the effect of these two approximations in the predicted timescale and amplitude of PCA fluctuations.

Furthermore, the LE4PD-XYZ method, even when neglecting hydrodynamic effects and simplifying the treatment of the residue-specific friction coefficients, accounts for the mode-dependent free energy barriers to transport in the mode space. This step is essential for an accurate description of the kinetics and transition mechanisms of the fluctuation dynamics in the mode coordinates,[14–16,18] and is not part of the conventional PCA.

III. METHODS

A. Molecular Dynamics Simulations

We generated the MD simulation of ubiquitin using GROMACS version 5.0.4[37] with the AMBER99SB-ILDN atomistic force field[38] on the Comet supercomputer at the San Diego Supercomputing Center. The starting structure was selected from the Protein Databank, PDB ID: 1UBQ.[39] We solvated the protein with spc/e water and minimized the energy using the steepest descent algorithm. We added Na^+ and Cl^- ions until the ion concentration was 45 mM, with the concentration of ions selected to match the one used in nuclear magnetic resonance experiments of ubiquitin.[40] Previously, we used those experimental data to test the accuracy of the LE4PD model against NMR data of T_1, T_2, and NOE relaxation experiments, with which the LE4PD approach show quantitative agreement.[14–16] We subjected the protein-solvent system to two rounds of equilibration: first, a 50-ps equilibration in the NVT ensemble at 300 K, with a Nosé-Hoover thermostat controlling the temperature; then, a 450-ps NPT equilibration at 300 K, with the same thermostat and a Berendsen barostat set to 1 bar.

Following the NPT equilibration, we performed a 10-ns 'burnout' simulation at 300 K with the Nosé-Hoover thermostat to maintain the temperature constant. The last frame obtained in this procedure is adopted as initial configuration for the 1-μs production run, which is performed using the same simulation parameters as the burnout simulation. Based on a manual inspection of the root-mean-squared deviation (RMSD) of the alpha-carbons from this first frame, we saw that the entire trajectory was fluctuating around an equilibrium value, and we used the entire 1-μs of trajectory for the simulation analysis. We used the LINCS algorithm[41] to constrain all hydrogen-to-heavy-atom bonds in the system and adopted an integration timestep of 2 fs during both the equilibration and the simulation run. The trajectory was recorded to file every 100 integration steps (every 0.2 ps), yielding a total of $(10^6 \text{ ps})/(0.2 \text{ ps/frame}) = 5 \times 10^6$ frames for analysis.

B. Building the coarse-grained dynamical model of the anisoptropic Langevin equation, the LE4PD-XYZ

The LE4PD equation is a coarse-grained (CG) model for the dynamics of proteins. Each CG unit represents an entire amino acid in the protein's primary sequence, with the center located at the position of the residue's center-of-mass. The equilibrium configuration of the protein gives the equilibrium length of the connecting bonds between CG sites, while the site-specific friction coefficient of each amino acid, ζ_i in Eq.2, is calculated from an extended Stoke's law as[13]

$$\zeta_i = 6\pi(\eta_w r_i^w + \eta_p r_i^p) \ . \tag{7}$$

Here η_w is the solvent's viscosity, and η_p is the viscosity in the hydrophobic core of the protein; r_i^w is the hydrodynamic radius of the amino acid for the solvent-exposed surface area,

and r_i^p is the hydrodynamic radius calculated from the area exposed to the hydrophobic core. The internal viscosity η_p is approximated as related to the water viscosity rescaled by the local energy barrier,[42] $\eta_p = \eta_w \exp[< E_{int} > /k_B T]$ with $< E_{int} > \approx k_B T$ the minimal free energy barrier to the local internal motion of the protein.

Before performing the LE4PD-XYZ analysis, we processed the 'raw' MD trajectory to remove the rigid-body rotational and translational motions. This step is performed by first selecting the first frame of the simulation as the reference frame and then centering it at the simulation box's origin. Concurrently, all the frames where the protein is broken across the periodic boundaries are made whole. Finally, all subsequent simulation frames are centered and superimposed to the reference structure by minimizing the mean-square difference between atomic positions. This procedure guarantees that six eigenvalues of **C** and **A**, which correspond to the rigid-body translational and rotational dynamics, are numerically indistinguishable from zero.

Because the spatial coordinates, which describe fluctuations around the mean value, have zero-mean and because the rigid-body rotational and translational motions of the protein are removed from the MD trajectory prior to analysis, the dynamics in the mode space are decomposed into a set of $3N-6$ internal modes, plus 6 rigid-body modes corresponding to the rigid-body rotational and translational motions, which are associated with eigenvalues equal to zero.[43] Because the protein conserves its globular shape during fluctuations, the removal of translation and rotation is a reasonable approximation, given that the coupling between rotation and fluctuations is minimized.[44] The eigenvalues of the $3N-6$ modes are ordered by ascending magnitude, $\lambda_1 < \lambda_2 < \lambda_3 < \ldots < \lambda_{3N-6}$, with the smallest eigenvalues different from zero, corresponding to the largest amino acid fluctuation. Finally, since **C** is at least positive semi-definite, and **H** is also a positive definite matrix, both **A** and **HA** are at least positive semi-definite,[45] which implies that $\mu_a, \lambda_a \geq 0$, $\forall a$.

C. Validation of the LE4PD-XYZ theory

In Section II A we show how the eigenvalues of the PCA model relate to the eigenvalues and eigenvectors of the LE4PD-XYZ equation when hydrodynamic interactions and residue-specific friction coefficients are neglected and energy barriers are not included. Under those approximations, Eq. 2 is formally consistent with a Fokker-Planck-Smoluchowski equation, where the dynamics are expressed as a function of the probability density function.[34,46] In reference[35], Hinsen et al. analyze the anharmonicity of protein fluctuations starting from the Fokker-Planck-Smoluchowski equation, and by modeling with this equation the time decay of the density fluctuations as measured in neutron scattering experiments. The theoretical predictions in that study are consistent with the time decay observed in their simulations. The parameters entering the equation of motion were obtained by direct comparison with the simulation trajectories. The length of the simulation is limited to 1.5 ns, during which crossing of large energy barriers is unlikely to occur.

Figure 1 shows a direct comparison between the predictions of the LE4PD-XYZ for the time decay of the residue fluctuations and the same properties directly calculated from the trajectory. The parameters entering the LE4PD-XYZ equation are the amino acid friction coefficients and the energy barriers in the mode representation, which are calculated as described in Sections III B and IV. The mode-dependent timescale, which defines the decay of the local fluctuations, is simply rescaled using Kramers' theory of reaction kinetics and the height of the energy barrier (see Section IV for details).[47] The agreement is close to quantitative for the time correlation functions (tcfs) shown in Figure 1. The observed good agreement between theory and simulations depends on including the mode-dependent energy barriers and the hydrodynamic interactions. This observation may seem to be at odds with the good agreement observed in reference[35] between simulations and the Fokker-Planck-Smoluchowski equation, where hydrodynamics and energy barriers are not included. However, in reference[35] the simulation is limited in length to 1.5 ns, while our study simulations are 1 μs long, and during that time several crossing of high energy barriers may occur.

In Figure 1 we report as an example the decay of the fluctuations for six individual amino acids along the protein primary structure. The figure shows the LE4PD-XYZ predictions, with and without hydrodynamic interactions, while the correction due to the internal energy barriers is included. The agreement between theory and experiments is close to quantitative. It also shows the predictions of the LE4PD-XYZ theory without hydrodynamic interactions and without the correction due to the mode-dependent internal energy barriers; those are the predictions obtained if one calculates the timescale directly from the eigenvalues of the covariance matrix, i.e. from PCA. Including hydrodynamics interactions and energy barriers when modeling the dynamics of polymers is important for the good agreement with the data. The inclusion or exclusion of the HI contribution modifies the final decay of the tcfs. In fact, hydrodynamic ensures the correct scaling exponent with time of the polymer's long-time dynamics as observed experimentally, for example, in neutron scattering.[34]

IV. CONTRIBUTION OF THE HIGH ENERGY BARRIERS IN THE FLUCTUATION DYNAMICS OF PROTEINS DETECTED BY THE PCA AND LE4PD-XYZ METHODS

The Langevin equation expresses the time evolution of the protein's motion by identifying the forces that act on each amino acid. Those forces define how the protein's dynamics evolve in time and include forces between amino acids due to the intramolecular potential of mean force and long-range interactions mediated by the solvent. When the Langevin equation represents the dynamics of a protein in solution, the solvent's effect enters through a residue-dependent friction coefficient and the hydrodynamic interaction matrix (see Eq. 2).

When the hydrodynamic interaction matrix reduces to an identity matrix because residue-specific friction coefficients

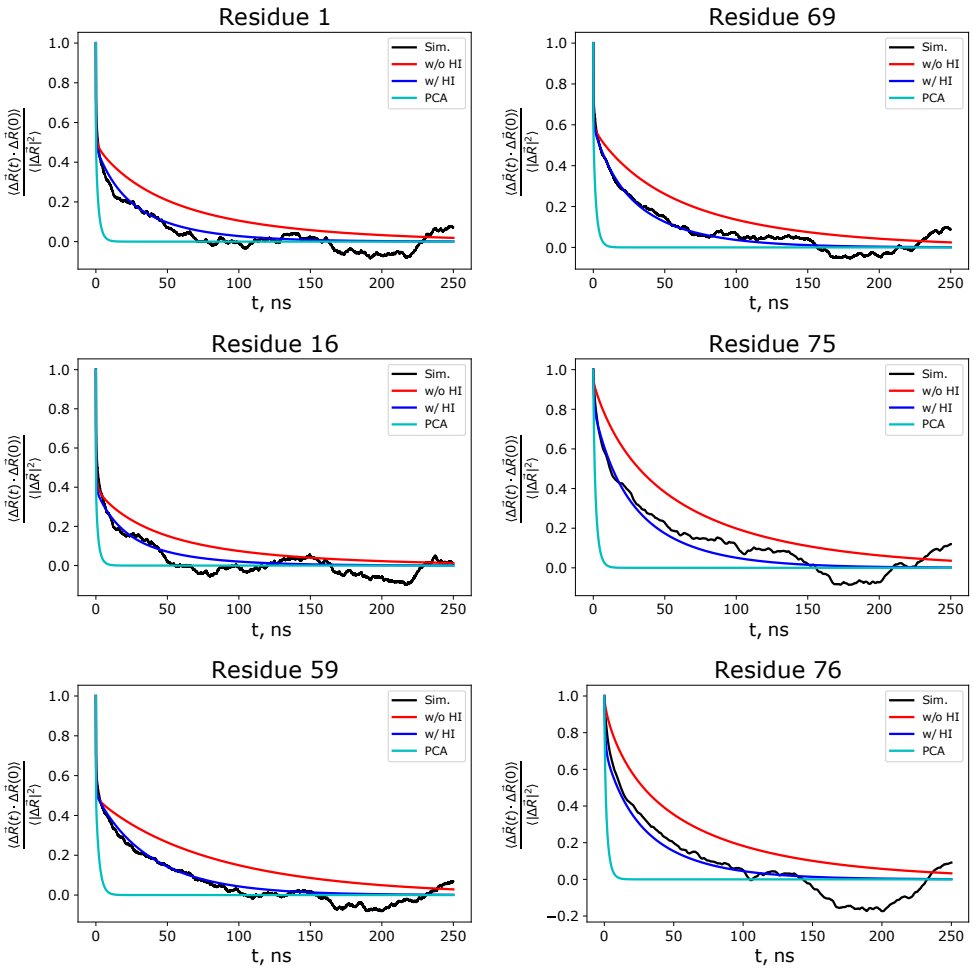

FIG. 1. Time decay of the position fluctuation for different residues in the ubiquitin. The predictions of the LE4PD-XYZ theory with hydrodynamic interactions included (blue) are compared with the theoretical predictions without hydrodynamics (red) and with simulations (black). The predictions of the LE4PD-XYZ theory when hydrodynamic interactions and the mode-dependent internal energy barriers are both excluded (cyan) correspond to the decay of the fluctuations given by the PCA eigenvalues. The agreement between the LE4PD-XYZ data, when energy barriers and hydrodynamic interactions are included, is close to quantitative.

and long-range forces mediated by the solvent are neglected, the time evolution of the protein's motion follows the covariance matrix. The latter describes harmonic fluctuations of the amino acids away from their equilibrium position. Under this approximation, the covariance matrix's eigenvalues are the inverse of the ones from the LE4PD-XYZ approach, and define identical timescales of the dynamics. Thus one may conclude that the dynamics described by the PCA's eigenvalues follows a simplified protein's equation of motion, which neglects the specificity of the amino acid friction coefficient and the solvent-mediated amino acid interactions.

The first approximation that one needs to enforce to recover the PCA dynamics from the Langevin equation is the assumption that all the amino acids in the protein have identical friction coefficients. The friction coefficient is proportional to the hydrodynamic radius of the residue, as discussed in Section III B. This radius may vary with the residue's chemical nature and its location inside the protein's three-dimensional structure. It depends on the extent of the surface exposed to the solvent, which can change dramatically for different amino acids along the protein's primary sequence.[48,49] Thus, there is no physical motivation to support the adoption of an identical friction coefficient for all the amino acids in a protein. We show how this approximation affects the dynamics in Section V.

The second approximation assumes neglecting the hydrodynamic interaction. The hydrodynamic interaction matrix represents the long-ranged interactions between amino acids, mediated by the solvent described as a continuum medium.[50] When describing the rotational and translational dynamics of proteins, it is common practice to include hydrodynamics effects. However, hydrodynamic contributions to proteins' internal motion are, sometimes, neglected. While this may be a reasonable approximation for very localized motion,

such as local vibrations, in general, hydrodynamic effects are not negligible. The non-local hydrodynamic coupling of amino acids' dynamics can alter the time scale of the large-amplitude fluctuations. We will study in more detail the effect of hydrodynamic interactions on ubiquitin's mode fluctuations in Section V.

A. Building Free Energy Surfaces

An important contribution to the timescale of protein's fluctuations is the crossing of high energy barriers in the FES. Note that this contribution is present even when the hydrodynamic interaction is neglected. The Langevin equation, given in Eq.1, is a diffusive approach that does not explicitly account for energy barriers along the mode coordinates $\vec{\xi}_a(t)$ and, like PCA, describes harmonic fluctuations away from the equilibrium structure. In the absence of hydrodynamics, where the fluctuations in PCA and LE4PD-XYZ are driven by the same covariance matrix, our approach allows one to calculate for each mode, a, the associated free-energy map. From the free energy surface it is possible to quantify the barriers to transport, thus obtaining an accurate determination of the kinetics of the conformational fluctuations in the mode coordinates[16,18] (see Section IV A).

We calculate the mode-dependent FES from the MD trajectory by projecting the position vectors into the modes, using the LE4PD-XYZ eigenvectors. In the absence of hydrodynamic interactions, these are also the PCA eigenvectors. In the three-dimensional description, it is convenient to write each eigenvector as the sum of its components along the x-, y-, and z-directions

$$Q_a^{-1} = Q_{a,x}^{-1} \otimes \hat{x}^T + Q_{a,y}^{-1} \otimes \hat{y}^T + Q_{a,z}^{-1} \otimes \hat{z}^T , \tag{8}$$

with Q_a^{-1} the ath row of the $\mathbf{Q}^{-1}$, which is the matrix of the left eigenvectors of the product $\mathbf{HA}$. Here $Q_{a,\alpha}^{-1}$, with $\alpha \in \{x,y,z\}$ is an element of the $3N \times N$ matrix describing the projection of the x-, y-, and z-coordinates of each alpha-carbon onto mode a, while $\hat{x}$, $\hat{y}$, $\hat{z}$ are the basis vectors in the x-, y-, and z-directions, respectively, e.g., $\hat{x} = (1,0,0)^T$. The projection of the simulation trajectory using the eigenvector matrix defined above leads to the mode coordinates along the three spatial directions

$$\xi_{a,x}(t) = (Q_{a,x}^{-1} \otimes \hat{x}^T)\Delta\vec{R}(t) ,$$

$$\xi_{a,y}(t) = (Q_{a,y}^{-1} \otimes \hat{y}^T)\Delta\vec{R}(t) ,$$

$$\xi_{a,z}(t) = (Q_{a,z}^{-1} \otimes \hat{z}^T)\Delta\vec{R}(t) .$$

From these mode vectors, it is possible to build an FES and more easily visualize the FES by calculating the probability in spherical coordinates. With the mode definition in hand, the polar and azimuthal angles of $\vec{\xi}_a$ are

$$\theta_a(t) = \arccos\left(\xi_{a,z}(t)/|\vec{\xi}_a(t)|\right)$$

$$\phi_a(t) = \arctan\left(\xi_{a,y}(t)/\xi_{a,x}(t)\right),$$

with $|\vec{\xi}_a|$ the magnitude of $\vec{\xi}_a$. For each mode, we derive the FES by binning into a 2D-histogram the probability of occupying a given value of θ_a and ϕ_a, $P(\theta_a,\phi_a) = \int P\left(|\vec{\xi}_a|,\theta_a,\phi_a\right) d|\vec{\xi}_a|$ and then performing a Boltzmann inversion of the probability. The probability distribution in all theoretical calculations is a function of the three spherical coordinates. However, the graphical representation of the free energy surface is simplified when omitting the probability as a function of $|\vec{\xi}_a|$, as the energy plot reduces to three dimensions. This step is possible because the free energy as a function of $|\vec{\xi}_a|$ does not present any remarkable feature. Thus, the total probability can be averaged over the values of $|\vec{\xi}_a|$ without losing important information. Instead, distinct dynamical pathways are visible in the FES as a function of the polar and azimuthal angles. Thus, the mode-dependent free energy surface is conveniently described by the free energy as a function of the θ_a and ϕ_a angles, while we average over the surface's fourth dimension, $|\vec{\xi}_a|$. The free-energy for a given (θ_a,ϕ_a) pair, as sampled by mode a, reduces to

$$F(\theta_a,\phi_a) = -k_B T \ln\left[P(\theta_a,\phi_a)\right] , \tag{9}$$

averaged over all the values of $|\vec{\xi}_a|$. A further discussion of this step is presented in the Supplementary Material.

To account for the effects of energy barriers in the decay of the correlations of the residue fluctuations, such as those shown in Figure 1, the friction coefficient in Eq. 6 is re-normalized using a Kramers-type approach[17,51]: $\overline{\zeta} \to \overline{\zeta}\exp\left[F_{\mathrm{MAD},a}/K_B T\right]$, where $F_{\mathrm{MAD},a} = \mathrm{median}\left(|F(\theta_a,\phi_a) - \min\left(F(\theta_a,\phi_a)\right)|\right)$ is an average free-energy barrier calculated for LE4PD-XYZ mode a using the median absolute deviation (MAD)[52,53] from the minimum of energy on the surface. This approach rescales the diffusive mode timescales as $\tau_a = \frac{\overline{\zeta}}{k_B T \lambda_a} \to \frac{\overline{\zeta}}{k_B T \lambda_a}\exp\left[F_{\mathrm{MAD},a}/K_B T\right]$. Using the MAD statistic removes any poorly sampled regions of the energy surface when the calculating the barrier, and using the MAD to rescale the friction coefficients has previously been shown to be effective in describing the slow-down in the decay of the M_1(t) time correlation function calculated from the LE4PD theory at the 2.5 ns timescale for ubiquitin,[14] the same protein under study here.

We obtain a free energy map for each mode, which presents a complex landscape, with minima, maxima, and complex dynamical pathways (see for example Figure 2). In the following, we will study the energetic pathways that emerge from the LE4PD-XYZ when hydrodynamic interactions are neglected. As mentioned earlier, the LE4PD-XYZ directly connects with PCA in the 'free-draining' limit, where Eq.1 is solved while assuming $\mathbf{H} := \mathbf{I}$. In that case, the mode solutions and the free energy maps from both approaches are identical.

B. Comparing Pathways through the Free-Energy Surface

Figure 2 illustrates the free energy surface (FES) and the kinetic pathways of the first LE4PD-XYZ/PCA mode, which corresponds to the first non-zero eigenvalue. In the top two panels (a) and (b) the FES is identical, while the pathways are different. The FES presents two well-defined minima in energy, one at $\theta_a \approx 75\ (deg)$ and $\phi_a \approx 60\ (deg)$ and the second at $\theta_a \approx 100\ (deg)$ and $\phi_a \approx 270\ (deg)$. Superimposed to the FES are the kinetic pathways of crossing the free energy barriers. Because the angular coordinate ϕ_a, is periodic, there are two possible pathways to connect the two minima in the free energy surface. Those are calculated using a variant of the string method,[18,24] and are reported in the Figure 2a and Figure 2b, top panels. For the PCA, instead, the path is defined as the linear interpolation between the most extreme configurations sampled by the simulation[23] and is identical in the two panels. Interestingly, these paths crossing the energy barriers and the PCA linear interpolation do not coincide.

The PCA interpolation method gives a pathway that does not quite begin or end in the minima of the free-energy surface. The path's extrema capture some less-likely configurations populated by the fluctuations of the mode around the most probable configurations. We also observe that the PCA linear interpolation does not follow the pathway of the energetically-favored barrier crossing. The intermediate states cross through a low-probability region of the surface and do not travel through the 'valley' between the two minima.

Figure 2c presents the superimposed configurations that populate the three kinetic paths just mentioned. The real-space structures that the molecule experiences while moving along the two most probable paths of barrier crossing in the FES are depicted in the left and right panels. The central panel, instead, shows the set of conformations populating the interpolation path of PCA. The PCA structures, in the central panel of Figure 2c, show large deformations along the entire alpha-carbon sequence of ubiquitin. Instead, the two pathways predicted by the LE4PD-XYZ show that the motion is concentrated in the tail region of ubiquitin.

A reason for the inconsistency with LE4PD of the interpolation method is its reliance on the extreme values of $\xi_a(t) = \left(\mathbf{Q}^{-1}\Delta R\right)_a(t)$. Thus, it utilizes only two samples of $\xi_a(t)$ to generate a fictitious trajectory for visualization, while the pathway method used in the LE4PD-XYZ utilizes the entire $\xi_a(t)$ trajectory to create the FES and hence the pathways between minima on the FES. And while minima on the FES and extreme values of $\xi_a(t)$ tend to be correlated, the extrema of $\xi_a(t)$ are by no means guaranteed to represent the configurations of the energetic minima. For example, Figure 3 shows the projection of $\xi_a(t)$ onto the FES's for LE4PD-XYZ modes 1 and 7 without HI. While the lowest and highest values of $\xi_a(t)$ are situated in the FES's deepest minima, the absolute minimum and maximum of $\xi_a(t)$ may not be of the lowest energy. If they were, the locations marked by the colored stars (maximum and minimum projections of $\xi_a(t)$) and triangles (minima of energy) would superimpose in both cases. Figure 3 also shows that the most extreme projections of $\xi_a(t)$ tend to lie outside the deepest minima of the FES. The displacement can be small, as in the extreme positive projection of $\xi_a(t)$ for the first mode and in both the extreme projections for mode 7, or it can be different by a more substantial amount, as is the case for the extreme negative projection of mode 1. The minima in the FES and the extrema of the PCA fluctuations may coincide. However, even in this case, there is no guarantee that the intermediate states in PCA, found using the interpolation procedure, define an energetically favorable pathway.

V. HOW INCLUDING HYDRODYNAMICS MODIFIES EIGENVALUES, EIGENVECTORS AND RELATED QUANTITIES: A COMPARISON OF PCA AND THE DIFFUSIVE LANGEVIN APPROACH OF THE LE4PD-XYZ

The study reported in the previous section shows that the LE4PD-XYZ approach identically maps onto the PCA formalism when the hydrodynamic interaction is neglected because the LE4PD-XYZ eigenvalues and eigenvectors map directly onto the PCA eigenvalues and eigenvectors. It also shows that the simple interpolation procedure of PCA portrays an approximate representation of the slow motion, as the path of the fluctuation may not follow the kinetic pathway of minimum energy between two energetic minima. Thus, the PCA's amplitude and pathway of slow fluctuations can be somewhat inaccurate in representing the most-probable and, likely, biologically-relevant fluctuations in the protein.

Modifying the forces acting on the protein by including hydrodynamics interactions modifies the eigenvalues and eigenvectors of the **HA** matrix product. This may change the timescale and amplitudes of the slow fluctuations around the equilibrium configuration. As a first step, we focus on comparing eigenvalues and eigenvectors with and without hydrodynamic interactions. Note that when we include hydrodynamic interactions, we also assume residue-dependent friction coefficients, which are calculated following the procedure described in the Methods section.

Figure 4 shows how the eigenvalues are modified by the inclusion of residue-specific friction coefficients in the hydrodynamic interaction matrix, and by the inclusion of the full HI matrix with long-ranged cross interactions and residue-specific friction coefficients. Given that the mode-dependent timescales are defined as the inverse of the eigenvalues of the matrix product **HA**, $\tau_a = \overline{\zeta}/(k_B T \lambda_a)$, one can see that including the HI 'flattens' the eigenvalues, decreasing the timescale of the lowest-index modes, making them faster, and increasing the timescale of the highest-index modes, making them slower. The theory of polymer dynamics predicts a similar effect: the scaling exponent of the Rouse modes is modified by the inclusion of the hydrodynamic interaction leading to the Rouse-Zimm approach, from which the LE4PD is derived. Including hydrodynamic effects 'softens' the eigenvalue spectrum by lowering the magnitude of the dynamic scaling exponent .[15,34]

Modifying the description of the hydrodynamic interaction is likely to affect the values of the eigenvectors of the matrix product **HA** as well. Figure 5 compares the eigenvector projections into the x-, y-, and z-coordinates, with and with-

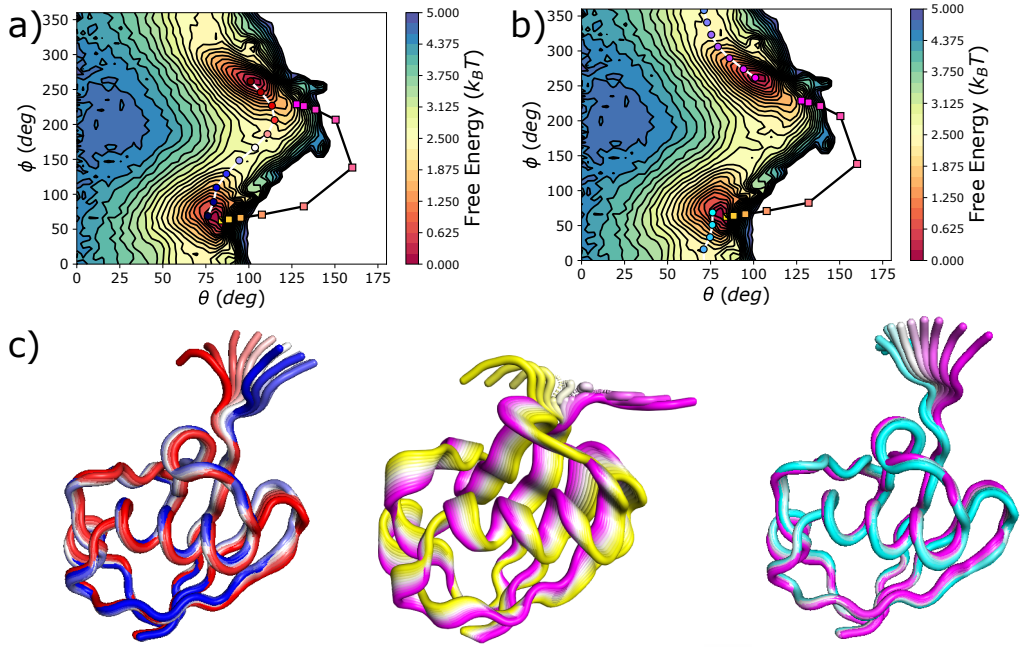

FIG. 2. Free-energy surface for the first LE4PD-XYZ mode, solved for the case where $\mathbf{H} := \mathbf{I}$, so that the LE4PD-XYZ mode solutions are identical to the calculated PCA modes. The FES in the a) and b) panes are identical. a) The blue-white-red path is a minimum energy pathway between the two minima, found using the string method. The yellow-white-magenta path shows the PCA trajectory of the linear interpolation between the two extreme structures. b) The cyan-white-magenta path follows a second minimum energy pathway that crosses the periodic boundary at $\phi = 0\ (deg) = 360\ (deg)$. The PCA linear-interpolation trajectory is identical to the one in the a) panel. c) The real-space, 3D fluctuations corresponding to each of these pathways are depicted by the superimposed ubiquitin structures. Each structure is colored corresponding to the analogously colored image along the pathway.

out HI, for the five slowest LE4PD-XYZ modes. Note that the timescales of these modes are given by the inverse of the LE4PD eigenvalues: including the energy barriers may modify the modes' order.

For the three slowest modes, the eigenvectors are essentially indistinguishable whether hydrodynamic interactions are included or not; however, differences become more apparent for modes 4 and 5. Because for a given mode the eigenvector determines the position and the amplitude of the fluctuations along the primary sequence of the protein, we expect that the inclusion of hydrodynamics will not affect the location of the slow fluctuations, but may modify their amplitude.

The direct comparison of the eigenvectors may be affected by the different ordering of the eigenvalues in the complete (with HI) and approximated (without HI) formalism, because the ordering of the eigenvectors depends on the ordering of the eigenvalues. Including the HI can modify the frequency of some mode fluctuations, thus changing the ordering of those modes. For example, the fluctuation of a loop could become slower due to the presence of long-ranged interactions once the hydrodynamics is included. To study possible cross-correlation between modes of different number, we calculated the overlap matrix, O_{ab}, between the eigenvector of mode a calculated without hydrodynamics (i.e. in PCA, Q_a), and the eigenvector of mode b calculated with HI included,(i.e. in LE4PD-XYZ, Q'_b), as

$$O_{ab} = Q_a \cdot Q'_b \,. \tag{10}$$

Since each eigenvector is normalized, the overlap matrix has the dimension of the number of internal modes, squared. For ubiquitin, a protein composed of 76 residues, 3N - 6 = 222 and O is a 222×222 matrix. A more in-depth explanation of the significance of O is given in the Supplementary Material.

Figure 6 shows O, which, overall, has a weakly diagonal structure, indicating that there is a similarity between the dynamical processes described by both types of LE4PD-XYZ treatments. However, the trace of O is only 34.8, while a perfect correspondence between processes would yield a trace of 3N - 6 = 222. For the ten slowest modes, which are represented by the insert in the up right corner of Figure 6, the overlap between eigenvectors from the two treatments is large, especially for the three slowest modes and the fifth slowest mode, each of which possesses an overlap greater than 0.91. Furthermore, from the trace of the overlap matrix we can calculate the average overlap of the first ten modes, which is 0.75. Subtracting this contribution from the total trace of 34.8, the remaining 212 internal modes have an average overlap of $\frac{34.8-7.5}{212} = 0.13$, which is small, indicating little similarity between the higher-index, fast modes from the two approaches. Thus, including the hydrodynamic interactions in the equation of motion produces a significant alteration of the high-

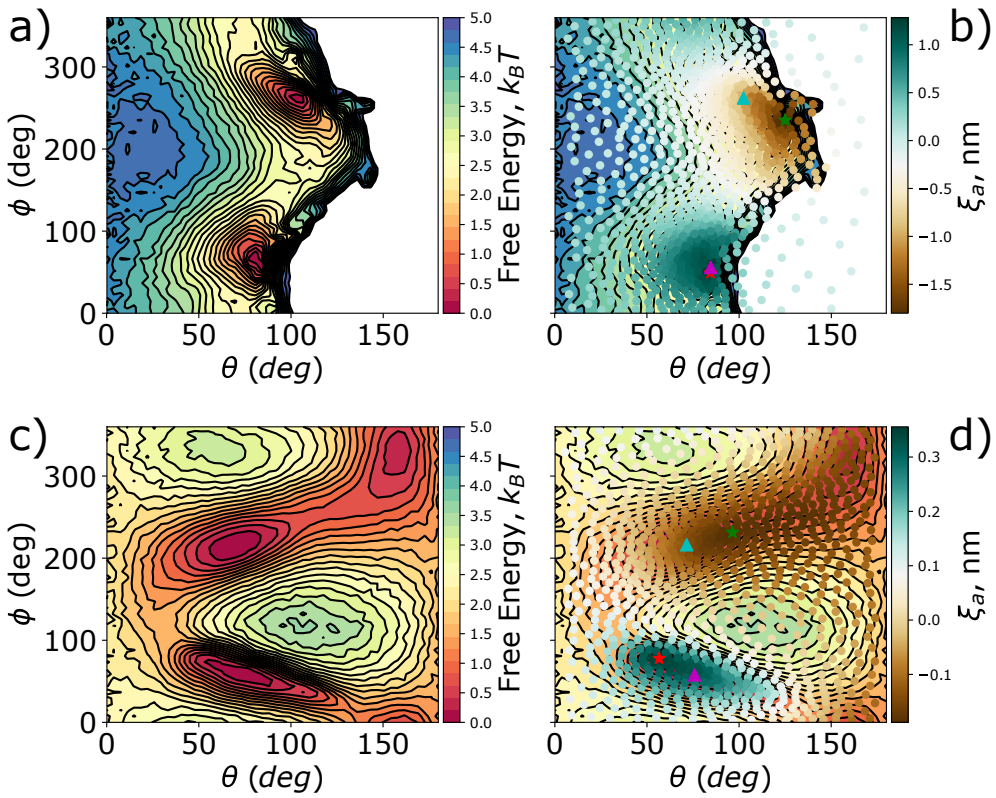

FIG. 3. a) FES for the first internal LE4PD-XYZ mode. b) Projection of ξ_a (green-white-brown markers) onto the two-dimensional FES for a=1. c) FES for the seventh internal LE4PD-XYZ mode. d) Projection of ξ_a (green-white-brown markers) onto the two-dimensional FES for a=7. Green and red stars mark the locations with the lowest and highest projection of ξ_a, respectively. Cyan and magenta triangles mark the locations with the lowest free-energy, subject to the constraint of having either a negative or positive projection along ξ_a, respectively.

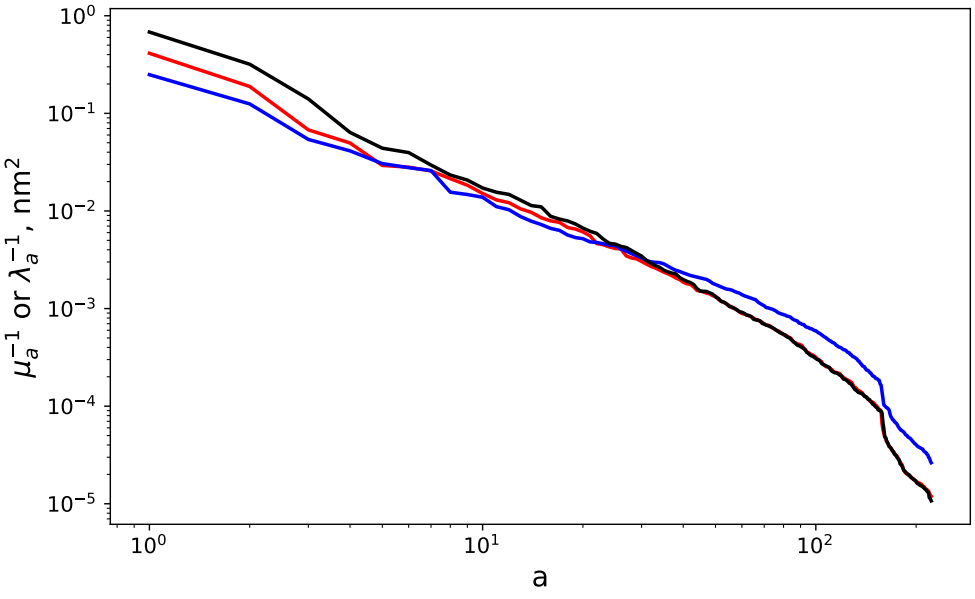

FIG. 4. Comparison of the eigenvalue spectrum without hydrodynamic interaction (black curve), with the ones where the reside-specific friction coefficients are included, to account for the chemical specificity of each residue (red curve), and with the eigenvalues resulting from the diagonalization of the product of matrices containing the full hydrodynamic interaction matrix (blue curve).

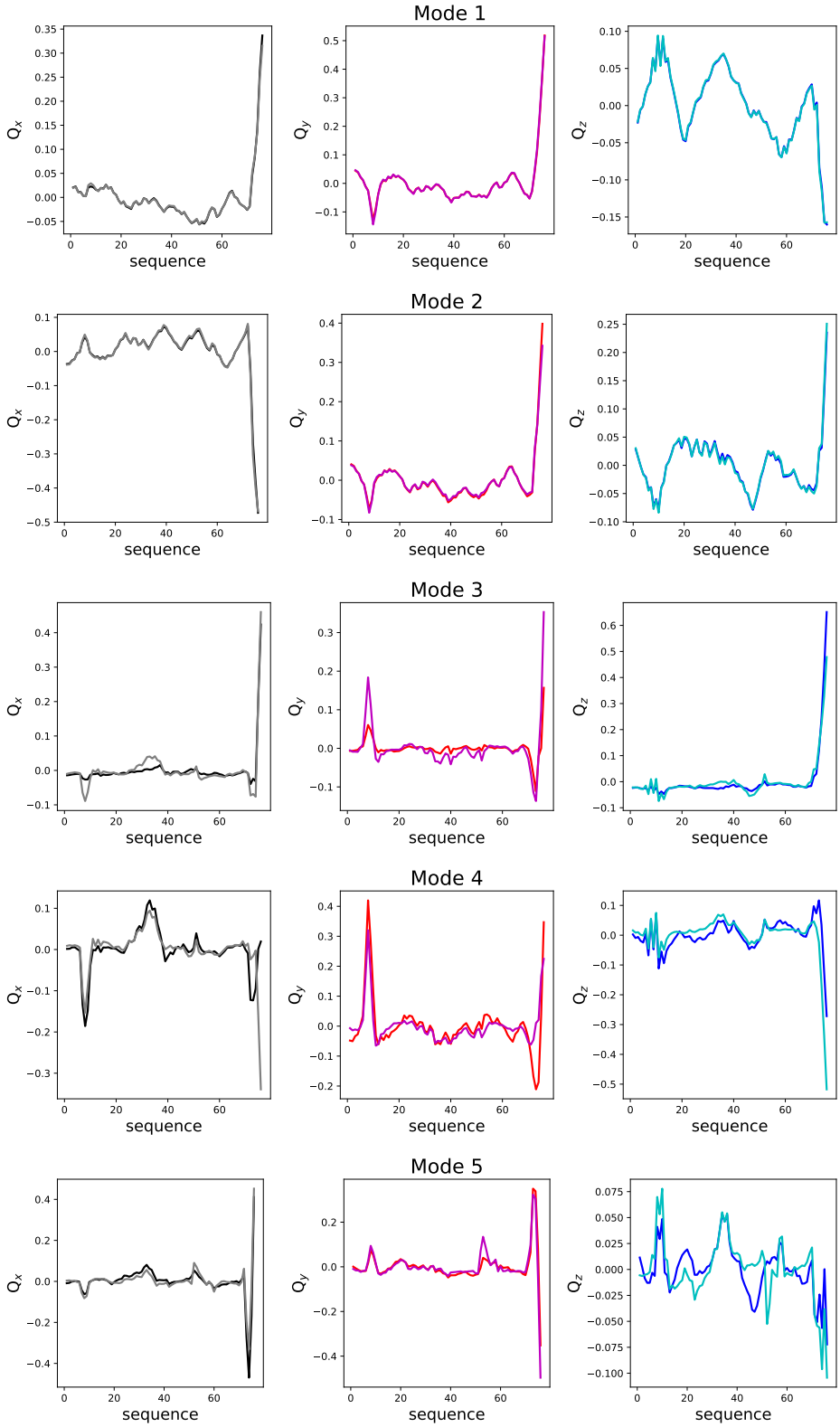

FIG. 5. Comparison of the Q_a^x, Q_a^y, Q_a^z without (black, red, blue) and with (grey, magenta, cyan) hydrodynamic interactions for the first 5 LE4PD-XYZ modes.

est index modes, which are fast and have small amplitude. The changes of the eigenvectors corresponding to the slowest modes appear, instead, to be more contained, at least for the protein ubiquitin, which is the focus of this study. These results suggest that we should expect small changes in the FESs corresponding to the slowest dynamical modes when hydrodynamic interactions are included.

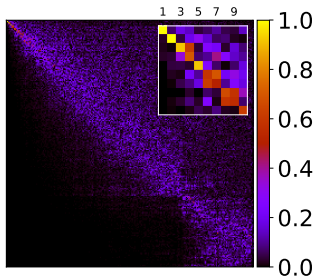

FIG. 6. Overlap matrix O, defined in Eq. 10, between the right eigenvectors of the LE4PD-XYZ without hydrodynamics and the right eigenvectors of the LE4PD-XYZ with hydrodynamic interactions included. Overall, there is weak diagonal trend in this matrix, indicating similarity between the analogous modes in both approaches. Inset: Sub-matrix of O corresponding to the overlap between the first ten modes from each LE4PD-XYZ treatment. Scale bar for the overlap between each mode is given to the right of the plot.

VI. EFFECT OF INCLUDING HYDRODYNAMIC INTERACTIONS ON THE POSITION AND AMPLITUDE OF SLOW MODE FLUCTUATIONS: A COMPARISON OF PCA VERSUS THE DIFFUSIVE LANGEVIN APPROACH OF THE LE4PD-XYZ

The previous section of this paper has shown that there are small but significant differences emerging in the eigenvalues and eigenvectors when one includes hydrodynamic interactions in the equation of motion. Thus, the eigenvectors used to map the simulation trajectory onto the normal modes and build the FES in the PCA are different from the ones in the anisotropic LE4PD with hydrodynamic interaction. So, it is likely that including the hydrodynamic interaction will lead to timescales and location of fluctuations that are different from the ones measured in PCA. In general, one may expect the timescales of processes that include hydrodynamic interactions to be more realistic because their dynamics follow an equation of motion that properly accounts for the effects of the solvent.

However, in the case of ubiquitin, the eigenvectors of the first two modes are almost quantitatively identical (see Figure 5) while the corresponding eigenvalues are modified by the presence of hydrodynamic interactions (see Figure 4). Thus, we expect the free energy maps for those two modes to be very similar in PCA and LE4PD-XYZ, while the timescale of fluctuations may be different.

The FES is calculated by mapping the mode $\vec{\xi}_a(t)$ in polar coordinates, where $\vec{\xi}_a(t) = \sum_i Q^{-1}_{ai}\Delta\vec{R}_i(t) = \sum_i \left(Q^{-1}_{(a,x)i}\Delta x_i(t) + Q^{-1}_{(a,y)i}\Delta y_i(t) + Q^{-1}_{(a,z)i}\Delta z_i(t)\right)$. The resulting free energy surfaces, $F(\theta_a, \phi_a)$, of the first five LE4PD-XYZ modes are displayed in Figures 7 and 8. The ordering of the modes in Figures 7 and 8 is based on the eigenvalues of the **HA** matrix, which do not include the slow-down in the mode-dependent dynamics due to the inclusion of free-energy surfaces.

Each mode with HI included is compared directly to the one without, considering modes that have the highest overlap according to Eq. 10. In agreement with Figure 5, the first two modes and the fifth mode in both approaches have striking similar free-energy surfaces. In contrast, modes 3, 4, and 6 or 7 are quite different, which is also in agreement with the mode-mode overlap matrix O shown in Figure 6. Note that mode 6 with HI (LE4PD-XYZ) has the maximum overlap with mode 7 without HI (PCA), and they will be directly compared.

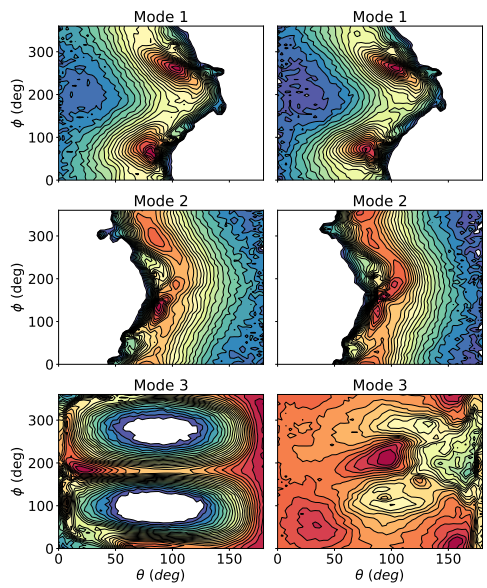

FIG. 7. Comparing the three slowest modes without (left) and with (right) hydrodynamics in the LE4PD-XYZ analysis. Red corresponds to low energy and blue to high energy; all regions with a free-energy above 5 k_BT are 'masked' as white. The scaling of the free energy is the same as that in Figures 2 and 3.

To analyze the position and amplitude of the local fluctuations along the primary sequence of ubiquitin, we calculate the total mean-squared fluctuations of the alpha-carbons, as follows:

$$\sum_i \langle \Delta\vec{R}_i \cdot \Delta\vec{R}_i \rangle = \sum_a \mu_a^{-1} \sum_i Q^2_{ia}, \tag{11}$$

By isolating the element in the first sum corresponding to mode a and the element i in the second sum corresponding to residue i in the protein on the right-hand side of Eq. 11 we obtain the definition of the mean-square fluctuations at residue i due to the process described by mode a, which we will call

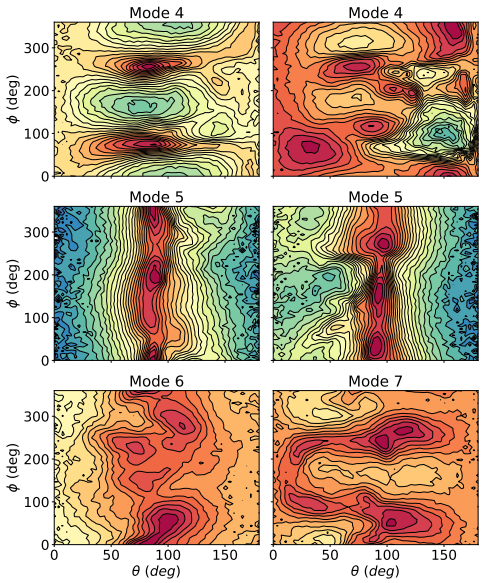

FIG. 8. Comparing the three next slowest modes without (left) and with (right) hydrodynamics in the LE4PD-XYZ analysis. Red corresponds to low energy and blue to high energy; all regions with a free-energy above 5 k_BT are 'masked' as white. Mode 7 is swapped with mode 6 in the hydrodynamics treatment because it overlaps more strongly with mode 6 without hydrodynamics.

the mean-squared local mode lengthscale, LML^2_{ia}:

$$\mathrm{LML}^2_{ia} = Q^2_{ia}\mu_a^{-1} \ . \tag{12}$$

In the anisotropic formalism of LE4PD-XYZ, $\langle \Delta\vec{R}_i \cdot \Delta\vec{R}_i \rangle = \langle \Delta x_i^2 \rangle + \langle \Delta y_i^2 \rangle + \langle \Delta z_i^2 \rangle$, and, by partitioning the Q_a into its $x-$, $y-$, and $z-$components, the LML^2_{ia} can be decomposed into $x-$, $y-$, and $z-$projections:

$$\mathrm{LML}^2_{ia,x} = (Q^x_{ia})^2\,\mu_a^{-1} \tag{13}$$

$$\mathrm{LML}^2_{ia,y} = \left(Q^y_{ia}\right)^2\,\mu_a^{-1} \tag{14}$$

$$\mathrm{LML}^2_{ia,z} = (Q^z_{ia})^2\,\mu_a^{-1}. \tag{15}$$

The use of the anisotropic LML^2_{ia} provides information on location, amplitude, and directionality of the localized fluctuations. The LML for the six slowest LE4PD-XYZ modes are shown in Figures 9 and 10; as previously, the modes with the highest overlap in LE4PD and PCA are compared directly in each of the subplots of the figures. Qualitatively, in most of the slowest modes, more specifically in the first three modes, there is little difference between the mode-dependent *location* of the fluctuations predicted whether hydrodynamic effects are included or neglected, in agreement with what we observed in the eigenvectors (see Figure 5). For these slow modes, the anisotropy of the fluctuations is not changed; the fluctuations in the $x-$, $y-$, and $z-$coordinates are the same regardless of the level of theory chosen, PCA or LE4PD-XYZ.

Note that some modes, such as mode 3, have a FES that is very different if hydrodynamic interaction is included or not. However, the largest fluctuations are localized in the same region of the protein's primary sequence. Although these modes describe fluctuations in a well-defined regions of ubiquitin (according to Figure 5), the *mechanism* of these dynamics maybe quite different, as indicated by the change in the free-energy landscape.

The situation is different for mode 5, where the anisotropic LML shows differences in the localization of the fluctuations when the two theories are compared. When hydrodynamics are included, there are larger y-coordinate fluctuations predicted in the 50 s loop of ubiquitin not seen when hydrodynamics are neglected. This situation is seen as well in mode 6, although there including hydrodynamics also increase the amplitude of the fluctuations in the C-terminal tail in the y- and z-coordinates, and reduces fluctuations of the Lys11 loop region in the x- and y-coordinates.

Additional changes in the conformational fluctuations described by the two approaches are observed when examining the transition pathways between minima on the free-energy surfaces when hydrodynamic effects are either included or neglected. Figure 11 demonstrates how hydrodynamic effects can alter the predicted conformational changes along mode-dependent transition pathways. For the slowest LE4PD-XYZ mode, shown in Figure 11a, there is little change in the free-energy surface when hydrodynamic effects are included and hence little change in either the predicted transition pathway between minima or the corresponding conformational fluctuations undergone by the protein.

However, for higher order modes, where the eigenvectors show disagreement between the two approaches, the free-energy landscape is strongly modified and we observe corresponding alterations of the conformational pathways crossing the barriers between energy minima, and related changes in conformational fluctuations along those pathways. Figure 11b shows this effect for mode 7 without hydrodynamics and mode 6 with hydrodynamics (which is the mode with the highest overlap with mode 7 without hydrodynamics). While the free-energy surfaces look roughly the same, there are significant differences in terms of barriers along the pathway between minima and the number of 'trap states' along the pathway. Furthermore, mode 7 without hydrodynamics predicts large-scale fluctuations in the C-terminal tail, the 50 s loop, and the Lys11 loop of ubiquitin, while mode 6 with hydrodynamics only predicts motion in the C-terminal tail and lower magnitude fluctuations in the 50 s loop. The motion of the C-terminal tail is quite different in the two modes.

VII. KINETICS OF BARRIER CROSSING IN THE MODE-DEPENDENT FREE ENERGY LANDSCAPE, CALCULATED BY MARKOV STATE MODELS: A COMPARISON OF PCA VERSUS THE DIFFUSIVE LANGEVIN APPROACH OF THE LE4PD-XYZ

The examples illustrated in the previous section show how examining the fluctuations predicted from the two different approaches provides important insights on the relevance of hydrodynamic interaction in the mode decomposition of the protein dynamics. Here we evaluate the quantitative timescale of the protein fluctuations by analyzing the dynamics of bar-

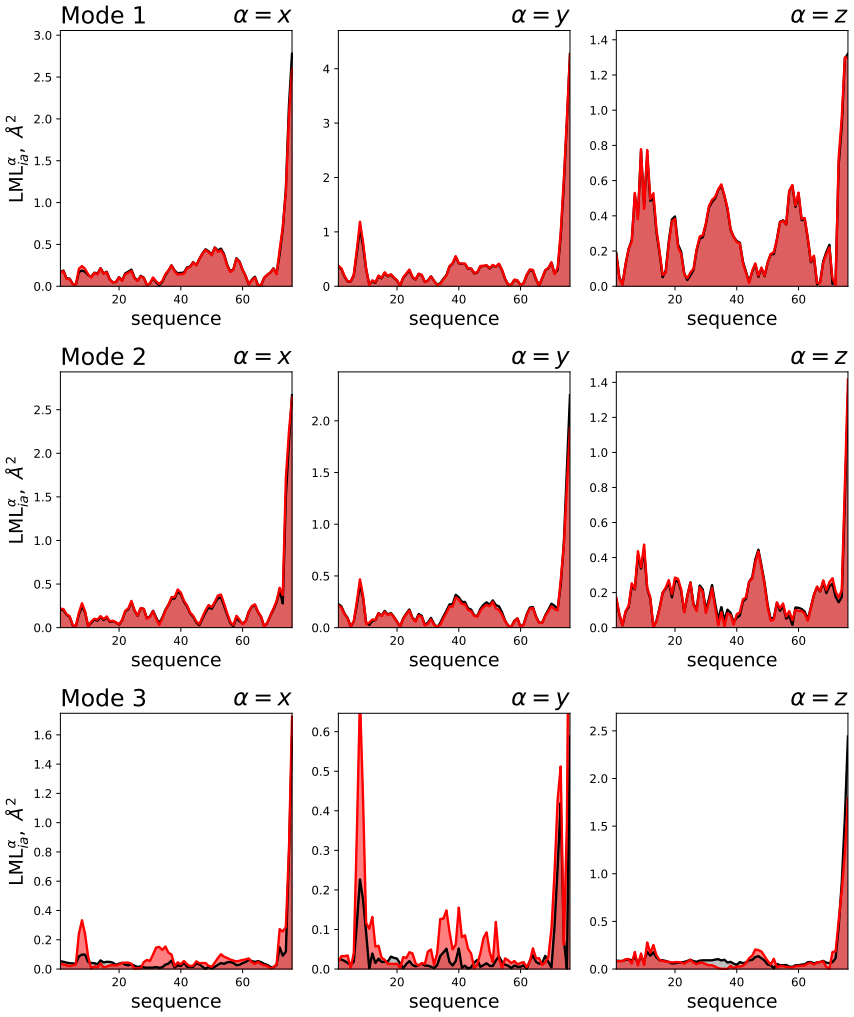

FIG. 9. Anisotropic LML, $\mathrm{LML}_{ia}^{\alpha}$ for the first three LE4PD-XYZ modes, as ordered by the λ_a eigenvalues, for the case where hydrodynamic effects are neglected (black) and with hydrodynamic effects included (red).

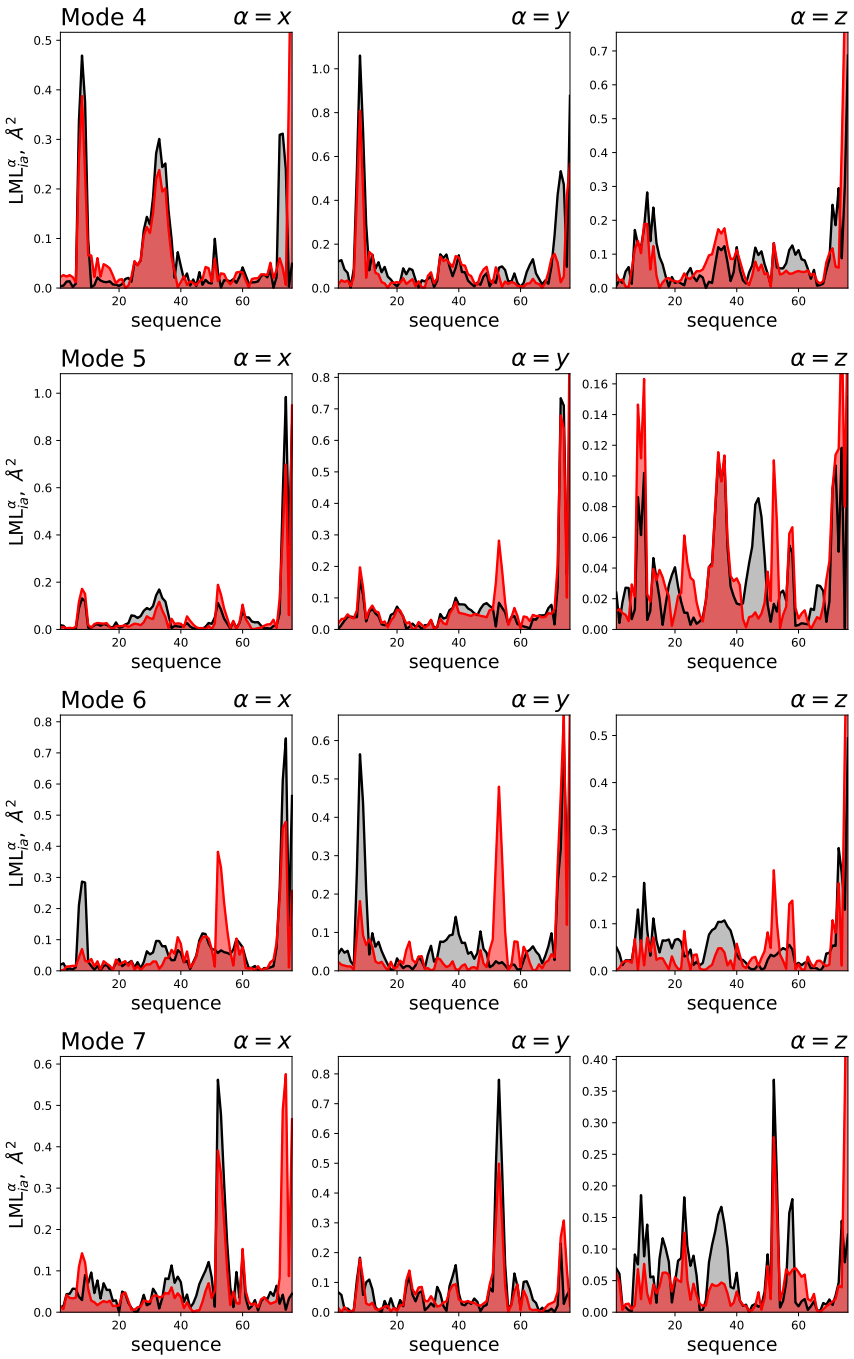

FIG. 10. Anisotropic LML, $\mathrm{LML}_{ia}^{\alpha}$ for the next four slowest LE4PD-XYZ modes, as ordered by the λ_a eigenvalues, for the case where hydrodynamic effects are neglected (black) and with hydrodynamic effects included (red).

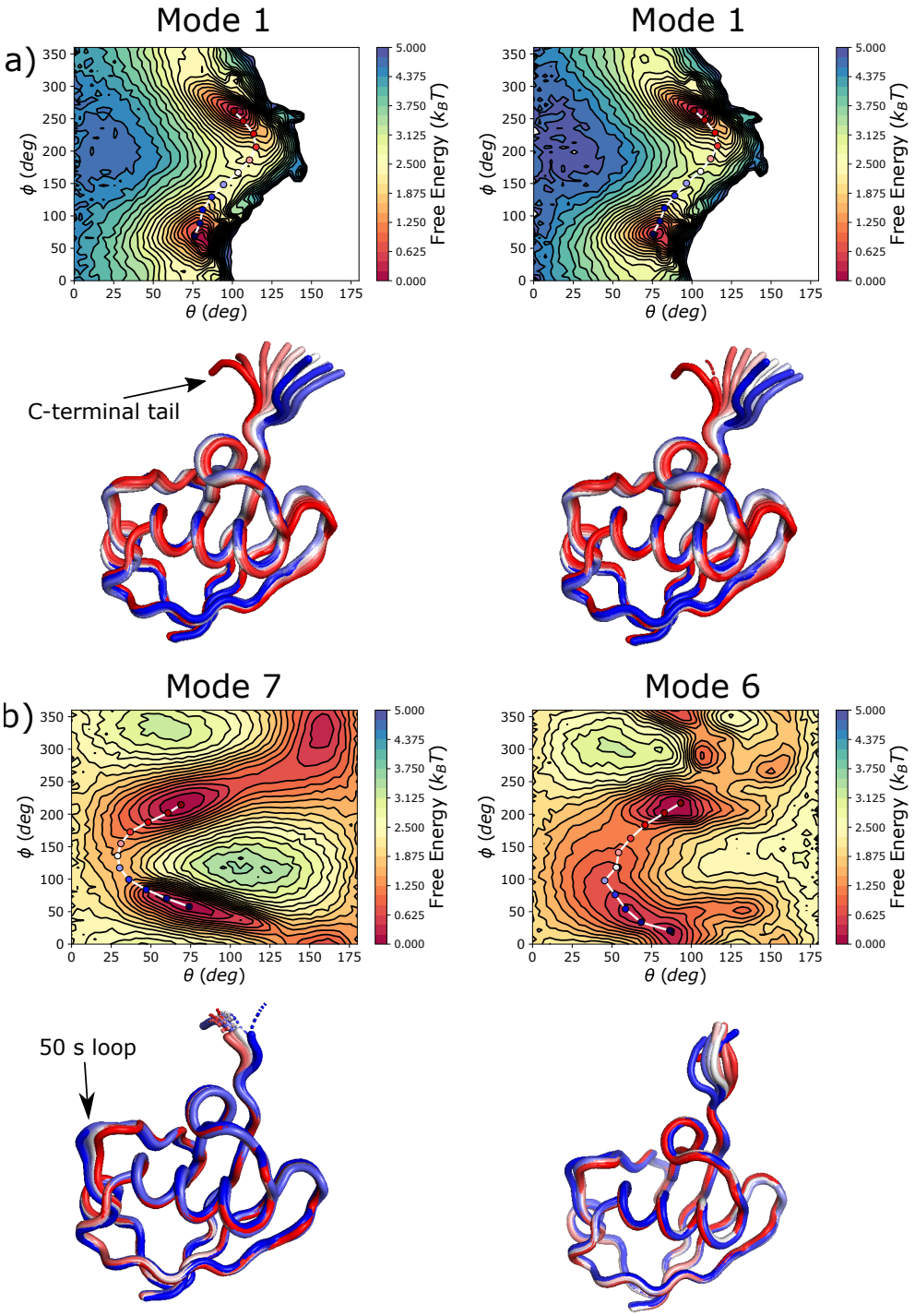

FIG. 11. Comparing the mode-dependent fluctuations along a path in the free-energy surface for the slow LE4PD-XYZ modes when hydrodynamics are neglected (left), which is equivalent to PCA, or included (right) for a) LE4PD-XYZ mode 1 and b) LE4PD-XYZ mode 7 without HI (left) and mode 6 with HI (right). Representative structures of ubiquitin for each image along the pathway are given below the corresponding free-energy surface, with the colors of the structure identical to the corresponding image along the pathway.

rier crossing with a Markov State Model analysis.

Markov state models (MSMs) are discrete-state master equation used to determine the kinetics of processes in multidimensional energy landscapes.[54] It is convenient to simplify the analysis of the free energy surface by mapping the multidimensional free energy landscape into a set of slow variables in the state space. Those slow variables are extracted from a time-ordered set of configurations, usually generated by an MD simulation, using a procedure of dimensionality reduction like PCA[26,54,55]. In the MSM approach, the state space of slow variables from the simulation is broken into a set of L discrete states. The conditional probability of transitioning between two of the given states i and j is calculated from the MD trajectory by sampling it at a properly-selected lagtime, τ. The transition probability is stored in a transition matrix, $\mathbf{T}(\tau)$, as $T_{ij}(\tau)$. The transition matrix is diagonalized to obtain a set of eigenvalues and eigenvectors; because $\mathbf{T}(\tau)$ is a stochastic matrix, its eigenvalues are bounded from above by 1 and all other eigenvalues are of modulus strictly less than 1, according to Perron's theorem.[56] Following the same theorem, the first right eigenvector, ψ_1, of the transition matrix is a vector of "1's", $\psi_1 = (1,1,\ldots,1)^T$ and the first left eigenvector, ϕ_1, is equal to the stationary distribution of the system, π; $\phi_1 = \pi$.[46,56]

Using the relationship of $\mathbf{T}(\tau)$ to the corresponding rate matrix, $\mathbf{K}(\tau)$, $\mathbf{T}(\tau) = e^{\mathbf{K}(\tau)\tau}$, the eigenvalues $\lambda_i^{\mathrm{MSM}}(\tau)$ of the transition matrix can be used to find the timescales, t_i, of the dynamic processes described by the MSM[26,54,55]:

$$t_i = -\frac{\tau}{\ln\left(|\lambda_i^{\mathrm{MSM}(\tau))}|\right)}. \tag{16}$$

This definition of the timescale of the transition relies on the Chapman-Kolmogorov theorem of Markovian statistics.[54] Since the eigenvalues of $\mathbf{T}(\tau)$ are sorted in descending order, $\lambda_1^{\mathrm{MSM}(\tau)} = 1 > \lambda_2^{\mathrm{MSM}(\tau)} > \lambda_3^{\mathrm{MSM}(\tau)} > \ldots > \lambda_L^{\mathrm{MSM}(\tau)}$, and $\lambda_1^{\mathrm{MSM}(\tau)} = 1$ corresponds to the stationary distribution. The slowest non-stationary process described by the MSM corresponds to ψ_2 with a timescale $t_2 = -\frac{\tau}{\ln\left(|\lambda_2^{\mathrm{MSM}(\tau))}|\right)}$. Thus, the slowest MSM mode, t_2, will give the timescale of barrier crossing in the free energy landscape. However, as aforementioned, the MSM analysis is more conveniently applied when the multidimensional free energy landscape is reduced into three dimensional free energy maps. The dimensionality reduction is often performed by using PCA normal modes. However, this study has shown that adopting an anisotropic LE4PD normal mode analysis can give fluctuations that properly include the effect of the solvent and, thus, are more physically sound.

The LE4PD formalism has the advantage of decomposing the complex dynamics of a protein in modes that are quasi-linearly-independent. They provide free energy landscapes that are dimensionally reduced and whose linear combination reconstructs the complete dynamics of the protein.

However, the application of the MSM requires the presence of high enough energy barriers, so that it is possible to separate fast interconverting motions that occur in the energy wells from the slow transitions due to crossing of barriers between states. We have observed that while the slowest protein modes have high energy barriers, and the most local modes have those as well, the intermediate LE4PD modes present a less structured energy landscape, with not localized high energy barriers.[18] In this study, the MSM analysis is applicable to the ten slowest dynamical modes. Higher index modes show a rough free energy landscape, which is well approximated by a diffusive approach to barrier crossing, such as Kramers equation.[47]

For the slowest LE4PD-XYZ modes, the effective kinetics is determined by performing the MSM analysis in each mode's FES in $(\theta_a(t), \phi_a(t))$ coordinate space. The second MSM eigenvector, ψ_2, separates the FES into macrostates where the protein rapidly interconverts, while transitions between macrostates are slow. The trajectories are sampled at a lag time, τ, that is consistent with the Markovian statistics, as tested using the Chapman-Kolmogorov criteria. In particular, we adopted a method that use the committor function to identify the top of the transition barrier and that we proposed in a recent publication.[18] In our method the lagtime τ in the MSM are selected based on the projection of ψ_2 onto the (θ_a, ϕ_a) surface. The longest lagtime τ for which the maximum and minimum projections of ψ_2 were both located in deep minima on the surface were chosen as the lagtimes for the MSM on the slow LE4PD-XYZ modes, both with and without HI. This method of selecting τ effectively places the node of ψ_2 at the top of the largest barrier on the (θ_a, ϕ_a) surface. So, the t_2 of the MSM corresponds to the timescale it takes the system to move from one minimum on the surface to the other over the barrier. If the kinetics on the surface follow two-state kinetics, then the locations where ψ_2 has a node, discrete states where $\psi_2 = 0$, and there the committor will equal 0.5,[27,28] which is the value for a transition state on the surface.[27,57] In that case, ψ_2 and the committor give the same information.

Furthermore, for all the MSMs presented here, the number of discrete states $L = 1000$, which is selected based on cross-validation analyses[58–60] performed on the presented MSMs. $\mathbf{T}(\tau)$ is constructed using the reversible, maximum likelihood estimator given in[61]. An in-depth example of how the MSM is constructed for LE4PD-XYZ mode 7 is given in the Supplementary Material.

Table I shows the timescales for the slowest-occurring processes in the $(\theta_a(t), \phi_a(t))$-space for the ten slowest LE4PD-XYZ modes, when hydrodynamic effects are included and when the HI contributions are neglected. For the LE4PD-XYZ with HI, the two slowest modes, as predicted by the MSM, are modes 1 and 4, which correspond to high-amplitude motions in the the C-terminal tail and the Lys11 loop of ubiquitin occurring over a timescale of 8.0 and 6.4 ns, respectively.

Table I also shows the lagtimes and the slowest timescales from a MSM constructed on the two-dimensional FES of the ten slowest LE4PD-XYZ modes without HI, which would correspond to PCA. Qualitatively, the timescales for the slow modes without HI are similar to those with HI. One exception is mode 4, which is predicted to occur on a much faster timescale when HI are neglected, likely due to the lack of motion in the C-terminal tail compared with the same index

TABLE I. Lagtimes, τ, and predicted timescales of the slowest process from the MSM, t_2, (both in ns) of the ten slowest LE4PD-XYZ with and without hydrodynamic interaction included.

Mode	(w/ HI) $t_2(\tau)$, ns	(w/o HI) $t_2(\tau)$, ns
1	8.0(3.2)	6.5(2.8)
2	3.7(1.1)	4.6(1.5)
3	4.3(2.5)	4.3(2.0)
4	6.4(4.0)	1.0(0.3)
5	5.8(4.0)	5.5(4.9)
6	3.3(1.0)	3.1(2.0)
7	3.6(2.0)	7.4(3.0)
8	0.6(0.2)	—(—)
9	0.3(0.1)	1.3(0.5)
10	0.4(0.1)	0.4(0.2)

mode when HI is included. Also, mode 8 without HI occurs on a timescale that is too fast to be modeled with an MSM. Finally, without HI, the LE4PD-XYZ method predicts the slowest mode is mode 7 with a timescale of 7.4 ns. Mode 7 describes motion almost exclusively in the 50 s loop of ubiquitin. Although slightly faster than the 10 ns timescale predicted for a similar dynamics in the isotropic LE4PD theory,[18] nevertheless the qualitative result is the same, in that there is a single, slow mode that isolates the slowest dynamics in the 50 s loop of ubiquitin. The LE4PD-XYZ theory predicts that this characteristic motion of the 50 s loop is split between modes 6 and 7, which is why the approach with HI does not select this motion as the single slowest mode (Figure 10).

Thus, it is interesting to notice how the inclusion of the hydrodynamic interaction modifies both the energy maps and the timescales of fluctuations as measured by Markov State Model analysis. Furthermore, the slow modes identified by the anisotropic LE4PD and PCA display a dynamics in the 50 s loop of ubiquitin, which is in agreement with the results of the isotropic LE4PD equation,[18] while the LE4PD-XYZ with hydrodynamics identifies as the slowest fluctuation the dynamics in the C-terminal tail of ubiquitin, which is the second slowest motion, as identified by the isotropic LE4PD.[18]

VIII. COMPARING THE TIMESCALES PREDICTED BY THE DECAY OF THE MODE TIME CORRELATION FUNCTION

A common method used to calculate the timescales for the decay of the PCA modes is the integration of the time correlation function for each mode,[10,19] defined as:

$$\tau_a^{autocorr} = \int_0^\infty \frac{\langle \xi_a(t)\,\xi_a(0)\rangle}{\langle \xi_a(0)^2\rangle} dt. \tag{17}$$

This approach gives the decorrelation time for any arbitrary stochastic process.[62] In practicality, the upper limit of the integral is taken to be the lagtime t where the autocorrelation function hits 0 for the first time.[10,19] If the process is characterized by a single exponential decay, then

$$\frac{\langle \xi_a(t)\,\xi_a(0)\rangle}{\langle \xi_a(0)^2\rangle} = e^{-\frac{t}{\tau_a^{autocorr}}}.$$

However, in general, for PCA (or LE4PD) modes calculated from a long equilibrium MD simulation of a folded protein, such as the 1-microsecond simulation analyzed here, the relaxation spectrum of the mode autocorrelation function will be more complicated than single exponential.[63,64] Thus, $\tau_a^{autocorr}$ will give an averaged value of the timescale, which includes many relaxation processes.

Using the slowest timescale from the MSM constructed on the (θ_a, ϕ_a) surfaces also estimates the slowest timescale process of each $\xi_a(t)$, but makes the assumption that the kinetic process is Markovian. In general, we do not expect for the two timescales to be identical.

Here, the timescale $\tau_a^{autocorr}$ is compared to the previously estimated mode-dependent timescales, namely the t_2 from the MSM (along with the MSM lagtime, τ^{MSM}, used to generate the associated MSM) and the diffusive τ_a timescale from the LE4PD-XYZ equation of motion, for the ten slowest LE4PD-XYZ modes, either without (Table II) or with (Table III) hydrodynamic interactions included. For all the modes shown here, the τ_a calculated from the equation of motion 6 are lower bounds to t_2 and $\tau_a^{autocorr}$, as expected since the τ_a do not account for free-energy barriers along the mode coordinate. In general, for the slowest modes, the timescales calculated using either the MSM or the autocorrelation function are in reasonable agreement, especially for modes 1, 3, and 5 without HI and modes 1, 4, and 5 with HI.

The main discrepancies arise for modes 8 through 10 in both cases, which are the modes where the MSM starts to become less effective and the dynamics approach a regime where the crossing of energy barriers in the (θ_a, ϕ_a) surfaces becomes more diffusive due to the rough free-energy landscape.[18] For mode 7 without HI and modes 6 and 7 with HI, which all describe the slow motion in the 50 s loop of ubiquitin, the autocorrelation function relaxes more slowly than the timescale predicted by the MSM. In this case, the difference is likely due to the methodology used to parameterize the MSMs, where the lagtime of the MSM τ is selected such that the slowest timescale of the MSM describes transitions between the minima on the surface. In fact the mode trajectory samples not only transitions between the minima but also rare dynamics in the highest energy regions of the surface. These rare events may occur over even longer timescales. Since the autocorrelation function of ξ_a accounts for *all* the processes occurring, $\tau_a^{auotcorr}$ inherits this information and reports longer timescales than the MSM, in general.

IX. DISCUSSION AND CONCLUSIONS

Large-scale, anisotropic fluctuations in protein dynamics are important as they lead to rare conformational transitions, which are deemed to be relevant for protein folding, and, more

TABLE II. Timescales for the first ten LE4PD-XYZ mode without hydrodynamics. Barrier-free timescales predicted from the LE4PD-XYZ equation, τ_a; the slowest process of the Markov state model, t_2 (with the lagtime of the Markov state model, τ, given in parentheses next to t_2); and the de-correlation timescale, $\tau_a^{autocorr}$, from intergrating the normalized autocorrelation function of the LE4PD-XYZ modes. All timescales are in ns.

Mode	τ_a, ns	$t_2\,(\tau)$, ns	$\tau_a^{autocorr}$, ns
1	2.51	6.5(2.8)	9.2
2	1.17	4.6(1.5)	10.2
3	0.52	4.3(2.0)	9.0
4	0.24	1.0(0.3)	5.5
5	0.16	5.5(4.9)	8.5
6	0.15	3.1(2.0)	4.2
7	0.11	7.4(3.0)	44.9
8	0.09	–(–)	2.0
9	0.08	1.3(0.5)	5.5
10	0.06	0.4(0.2)	7.1

TABLE III. Timescales for the first ten LE4PD-XYZ mode with hydrodynamics. Barrier-free timescales predicted from the LE4PD-XYZ equation, τ_a; the slowest process of the Markov state model, t_2 (with the lagtime of the Markov state model, τ, given in parentheses next to t_2); and the de-correlation timescale, $\tau_a^{autocorr}$, from intergrating the normalized autocorrelation function of the LE4PD-XYZ modes. All timescales are in ns.

Mode	τ_a, ns	$t_2\,(\tau)$, ns	$\tau_a^{autocorr}$, ns
1	0.92	8.0(3.2)	9.8
2	0.46	3.7(1.1)	13.8
3	0.20	4.3(2.5)	10.7
4	0.15	6.4(4.0)	6.0
5	0.11	5.8(4.0)	11.7
6	0.10	3.3(1.0)	31.1
7	0.096	3.6(2.0)	15.7
8	0.057	0.6(0.2)	3.1
9	0.054	0.3(0.1)	2.0
10	0.051	0.4(0.1)	8.4

generally, for the protein's biological function.[65,66] A popular method to select these large fluctuations is a Principal Component Analysis or PCA.[9] In an MD trajectory, PCA identifies collective fluctuations, which are ordered by their decreasing amplitude, from the most extended to the smallest amplitude. While PCA is both computationally convenient and conceptually simple, it lacks a physical basis beyond the empirical observations that it describes some large-scale, collective motions, functional for the protein.[7,11,12]

In this study, we revisit the PCA formalism and formally connect it to a Langevin equation of motion, which was developed to identify slow dynamical modes and study their kinetics in protein dynamics, called the Langevin equation for protein dynamics, or LE4PD.[13–18] Like the PCA, the LE4PD decomposes the protein's motion into an orthogonal set of collective coordinates or modes.

To make a formal connection with PCA, the original LE4PD was extended in this study to describe the anisotropic fluctuations around an average structure. We call this formalism the LE4PD-XYZ. This equation of motion, which is solved analytically into eigenvalues and eigenvectors, captures the anisotropic slow fluctuations of a protein's alpha carbons, starting from the analysis of the atomistic MD trajectory. The LE4PD-XYZ is a first-principles approach, which allows us to formally connect fluctuations to the different force contributions that model proteins' dynamics. In this way, the LE4PD-XYZ can be viewed as a powerful equation of motion to accurately describe the dynamics of proteins in solutions.

The LE4PD-XYZ is a coarse-grained approach to protein dynamics and describes the slow fluctuations of the alpha-carbons' coordinates for each residue. All the residues are modeled as interacting via a harmonic potential of mean force, which is built using the covariances of each residue, as calculated from an MD simulation. This anisotropic equation of motion for the fluctuations of the amino acids' positions (Eq.1) is diagonalized into a set of equations describing the independent, uncorrelated LE4PD-XYZ normal modes (Eq. 6).

This study shows that the anisotropic Langevin equation for protein dynamics, or LE4PD-XYZ, becomes formally equivalent to an equation of motion guided by the forces in the covariance matrix. Thus the LE4PD-XYZ is equivalent to a Principal Component Analysis approach, but only when two specific approximations are adopted. The first approximation is that the equation-of-motion disregards the hydrodynamic interaction, i.e. that there is no correlation in the dynamics of the amino acids caused by the presence of long-ranged interactions mediated by the solvent. This is the so-called "free-draining" limit. The second approximation is that every amino acid in the protein has identical friction. Only when these two approximations are enforced, the fluctuations identified by the PCA become identical to the ones modeled by a diffusive equation of motion. Unfortunately, these approximations are in general not justified, even if they are frequently adopted. Including solvent-mediated interactions is important when one models protein dynamics in an effective medium. And this study shows, specifically, that hydrodynamic interactions modify the dynamics of the protein, and importantly the timescales of the slow modes. Likewise, the degree of exposure to the solvent of each amino acid, and their unique friction, affects the timescale of the amino acids' dynamics and their fluctuations.

While the dynamics of protein is anharmonic, the Langevin formalism and PCA both rely on the harmonicity of the fluctuations, which is a valid approximation only in the proximity of the folded state. This approximation is common in many structural approaches such as the Gaussian Normal Modes method.[67–69] However, in the LE4PD approach the anharmonic fluctuations are identified through a procedure that maps the simulation trajectory onto its mode-dependent two-dimensional free energy landscapes. This convenient procedure of variable reductions allows one to study the dynamics of the protein in separated modes. For each mode, one can

study the position, amplitude, and timescale of the fluctuations that are important for the biological function.[18]

The eigenvalues of the LE4PD-XYZ define an 'original' timescale, which is corrected by the analysis of the mode-dependent FES to include the slowing down of the dynamics due to the presence of high energy barriers. The corrected timescale is then calculated either from a Markov State Model analysis of the FES, or from the integral of the time-correlation function of the modes, which is a procedure often used in the PCA.[19]

The study applies the LE4PD-XYZ method to a 1-μs simulation of ubiquitin and compares the results with an analysis performed using the Principal Component Analysis method. The comparison between the timescale from the eigenvalues, and the more realistic timescales measured by applying a Markov state model analysis to the mode-dependent free energy maps, after identifying the leading transition pathways for these fluctuations, show the relevance of the energy barriers in measuring the kinetic timescale of fluctuations (see for example Tables I, II, and III).

Interestingly, the decay of the time-correlation function of the mode coordinates, which is the procedure often used to calculate the timescale in PCA, is qualitatively consistent the more elaborate Markov state model analysis of the slow pathways in the LE4PD-XYZ free energy maps. This result confirms the need to include both barrier crossing and the hydrodynamic interaction in a Langevin description of protein dynamics.

Furthermore, when we examined the effect including hydrodynamics on the predicted mode-decomposition of the dynamics, we observed that some, but not all, of the slowest modes are little changed. However, the introduction of hydrodynamic interaction has important effects on the faster modes, which are involved in local-scale processes of the proteins, for example in chemical reactions.

Specifically for ubiquitin, we observe that the slowest LE4PD-XYZ modes predict timescales between 300 ps and 8 ns, roughly in-line with those predicted in an early study of the same system using the isotropic LE4PD.[18] In contrast, the anisotropic LE4PD-XYZ approach is not able to isolate the slow fluctuations of the 50 s loop of ubiquitin, seen in mode 9 by the isotropic approach. The anisotropic description separates into the directional contributions this slow dynamics. However, when the HI is neglected one can again detect the unidirectional, slow fluctuation of the 50 s loop.

For ubiquitin, both the LE4PD-XYZ and PCA identify slow fluctuations and large-amplitude motion in the C-terminal tail region. It is known that the tail of ubiquitin is involved in many of the protein's binding events to substrates, both covalent[70,71] and non-covalent .[72] The wide range of possible conformations that are available for the binding of the tail may be important for the protein to discriminate among different possible reaction substrates. Thus, in line with the conformational selection hypothesis,[73,74] the large number of modes dedicated to describing motion in the C-terminal tail may indicate the opportunity for the protein to follow different transition pathways in the mechanism of binding to different substrates.

In conclusion, we have presented here the formalism for an anisotropic Langevin equation, the LE4PD-XYZ, that describes the motions of a protein in terms of a set of orthogonal normal modes and used it to analyze a 1-μs MD simulation of the protein ubiquitin. This approach coarse-grains the dynamics of the protein at the level of the protein's amino acids and accounts for the hydrodynamic interaction (HI) between amino acids as well as free-energy barriers along each of the LE4PD-XYZ modes. When HI, the specificity of each amino acid's friction coefficient, and free-energy barriers are neglected, the LE4PD-XYZ approach maps *exactly* onto the analogous PCA (in the sense that the dynamics is described by the same set of eigenvalues and eigenvectors). The inability of PCA alone to describe the dynamics correctly (unless hydrodynamics and energy barriers are included in the related equation of motion) is highlighted in Figure 1, where the time correlation function calculated from the simulation is compared with the mode-dependent decay of the PCA eigenvalues and displays a clear disagreement, with the correlation functions predicted from the PCA modes decaying too quickly relative to the correlation functions calculated from the simulation trajectory.

This study shows that the inclusion of the HI modulates the location and amplitude of the predicted fluctuations (Figures 5, 9, 10), eigenvalues (Figure 4), free-energy surfaces (Figure 7, 8, 11), and timescales (Table I). Finally, we have also shown how including free-energy barriers causes the dynamics predicted by the slow LE4PD-XYZ modes without HI to be different from those predicted by the analogous PCA (Figures 2, 3). These results demonstrate the importance of considering both hydrodynamic effects, with specific friction coefficients, and energetic barriers to transport when analyzing the equilibrium dynamics of ubiquitin about its folded state. Only by including these effects the time correlation functions that define the decay of local correlations quantitatively reproduce the decay measured in the atomistic simulations.

SUPPLEMENTARY MATERIAL

The Supplementary Material contains a detailed derivation of the **H** for the LE4PD-XYZ, a formal relationship between the **U** matrices from the isotropic LE4PD and the LE4PD-XYZ, comparisons of the conformational fluctuations predicted along the interpolated and minimum energy pathways for more of the slow LE4PD-XYZ modes, and a detailed example of the MSM construction for the first LE4PD-XYZ mode.

DEDICATION

M.G.G. dedicates this paper to all the women (friends, family, and colleagues) who made her life and her work stronger and more enjoyable with their advice, love, and support.

ACKNOWLEDGMENTS AND FUNDING

E.R.B. was supported by the John Keana Graduate Student Fellowship from the University of Oregon and the National Science Foundation through grants CHE-1665466 and CHE-1362500 to M.G.G. The computational work was performed on the supercomputer Comet at the San Diego Supercomputer Center, with the support of XSEDE[75] allocation TG-CHE100082 to M.G.G. (XSEDE is a program supported by the National Science Foundation under Grant No. ACI-1548562).

DATA AVAILABILITY

The codes used to perform the LE4PD-XYZ analysis described here are available on GitHub (https://github.com/GuenzaLab/LE4PD-XYZ), and the processed MD trajectory and $\left(\theta_a(t), \phi_a(t), |\vec{\xi}_a(t)|\right)$ trajectories for the first 10 LE4PD-XYZ modes both with and without hydrodynamics included are available on Zenodo (https://doi.org/10.5281/zenodo.4312224).

REFERENCES

[1] S. Hayward, A. Kitao, and H. J. C. Berendsen, "Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme," Proteins: Structure, Function and Genetics **27**, 425–437 (1997).

[2] S. H. Northrup, M. R. Pear, J. D. Morgan, J. McCammon, and M. Karplus, "Molecular dynamics of ferrocytochrome c: Magnitude and anisotropy of atomic displacements," Journal of Molecular Biology **153**, 1087 – 1109 (1981).

[3] A. L. Tournier and J. C. Smith, "Principal components of the protein dynamical transition," Phys. Rev. Lett. **91**, 208106 (2003).

[4] J. Chan, K. Takemura, H.-R. Lin, K.-C. Chang, Y.-Y. Chang, Y. Joti, A. Kitao, and L.-W. Yang, "An efficient timer and sizer of biomacromolecular motions," Structure **28**, 259 – 269.e8 (2020).

[5] A. E. García, "Large-amplitude nonlinear motions in proteins," Physical Review Letters **68**, 2696–2699 (1992), arXiv:0000135489.

[6] I. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics (Springer, 2002).

[7] H. Lou and R. I. Cukier, "Molecular dynamics of apo-adenylate kinase: A principal component analysis," The Journal of Physical Chemistry B **110**, 12796–12808 (2006).

[8] H. J. Berendsen and S. Hayward, "Collective protein dynamics in relation to function," Current Opinion in Structural Biology **10**, 165 – 169 (2000).

[9] J. Spiegelberg and J. Rusz, "Can we use pca to detect small signals in noisy data?" Ultramicroscopy **172**, 40 – 46 (2017).

[10] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, "Principal component analysis and long time protein dynamics," The Journal of Physical Chemistry **100**, 2567–2572 (1996), https://doi.org/10.1021/jp9536920.

[11] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," Proteins: Structure, Function, and Bioinformatics **17**, 412–425 (1993).

[12] K. Sankar, J. Liu, Y. Wang, and R. L. Jernigan, "Distributions of experimental protein structures on coarse-grained free energy landscapes," Journal of Chemical Physics **143** (2015), 10.1063/1.4937940.

[13] E. Caballero-Manrique, J. K. Bray, W. A. Deutschman, F. W. Dahlquist, and M. G. Guenza, "A theory of protein dynamics to predict NMR relaxation," Biophysical Journal **93**, 4128–4140 (2007).

[14] J. Copperman and M. G. Guenza, "Coarse-Grained Langevin Equation for Protein Dynamics: Global Anisotropy and a Mode Approach to Local Complexity," Journal of Physical Chemistry B **119**, 9195–9211 (2015).

[15] J. Copperman and M. G. Guenza, "Predicting protein dynamics from structural ensembles," The Journal of Chemical Physics **143**, 243131 (2015).

[16] J. Copperman and M. G. Guenza, "Mode localization in the cooperative dynamics of protein recognition," Journal of Chemical Physics **145**, 015101 (2016).

[17] J. Copperman, M. Dinpajooh, E. R. Beyerle, and M. G. Guenza, "Universality and Specificity in Protein Fluctuation Dynamics," Physical Review Letters **119**, 158101 (2017).

[18] E. R. Beyerle and M. G. Guenza, "Kinetics analysis of ubiquitin local fluctuations with markov state modeling of the le4pd normal modes," The Journal of Chemical Physics **151**, 164119 (2019), https://doi.org/10.1063/1.5123513.

[19] Y. Naritomi and S. Fuchigami, "Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions," Journal of Chemical Physics **134** (2011), 10.1063/1.3554380.

[20] L. W. Yang, E. Eyal, I. Bahar, and A. Kitao, "Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics," Bioinformatics **25**, 606–614 (2009).

[21] L. Skjaerven, A. Martinez, and N. Reuter, "Principal component and normal mode analysis of proteins; a quantitative comparison using the groel subunit," Proteins: Structure, Function, and Bioinformatics **79**, 232–243 (2011).

[22] T. Wu, *A Molecular Dynamics Simulation Based Principal Component Analysis Framework for Computation of Multi-scale Modeling of Protein and Its Interaction with Solvent*, Ph.D. thesis, New Jersey Institute of Technology (2011).

[23] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. D. Caves, "Bio3d: an R package for the comparative analysis of protein structures," Bioinformatics **22**, 2695–2696 (2006), https://academic.oup.com/bioinformatics/article-pdf/22/21/2695/700557/btl461.pdf.

[24] W. E, W. Ren, and E. Vanden-Eijnden, "String method for the study of rare events," Phys. Rev. B **66**, 052301 (2002).

[25] W. E and E. Vanden-Eijnden, "Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events," Annual Review of Physical Chemistry **61**, 391–420 (2010).

[26] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, "Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states," Journal of Chemical Physics **126**, 155102 (2007).

[27] N. V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," Journal of Physical Chemistry B **112**, 6057–6069 (2008).

[28] A. Berezhkovskii and A. Szabo, "Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices," The Journal of Chemical Physics **121**, 9186–9187 (2004), https://doi.org/10.1063/1.1802674.

[29] R. Horn and C. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, 1994).

[30] R. Zwanzig, "Theoretical basis for the rouse-zimm model in polymer solution dynamics," The Journal of Chemical Physics **60**, 2717–2720 (1974), https://doi.org/10.1063/1.1681433.

[31] A. Perico and M. Guenza, "Viscoelastic relaxation of segment orientation in dilute polymer solutions," The Journal of Chemical Physics **83**, 3103–3109 (1985).

[32] M. Guenza, "Many chain correlated dynamics in polymer fluids," The Journal of Chemical Physics **110**, 7574–7588 (1999), https://doi.org/10.1063/1.478660.

[33] I. Lyubimov and M. G. Guenza, "First-principle approach to rescale the dynamics of simulated coarse-grained macromolecular liquids," Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **84**, 16–18 (2011), arXiv:1103.3047.

[34] M. Doi and S. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, 1988).

[35] K. Hinsen, A.-J. Petrescu, S. Dellerue, M.-C. Bellissent-Funel, and G. R. Kneller, "Harmonicity in slow protein dynamics," Chemical Physics **261**, 25–37 (2000).

[36]P. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, 1979).
[37]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**, 19–25 (2015).
[38]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: Structure, Function, and Bioinformatics **78**, 1950–1958 (2010).
[39]S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 Å resolution," Journal of Molecular Biology **194**, 531–544 (1987).
[40]N. Tjandra, S. E. Feller, R. W. Pastor, and A. Bax, "Rotational Diffusion Anisotropy of Human Ubiquitin from 1 5 N NMR Relaxation," J. Am. Chem. SOC **117**, 12562–12566 (1995).
[41]B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, "LINCS: A Linear Constraint Solver for molecular simulations," Journal of Computational Chemistry **18**, 1463–1472 (1997).
[42]D. E. Sagnella, J. E. Straub, and D. Thirumalai, "Time scales and pathways for kinetic energy relaxation in solvated proteins: Application to carbonmonoxy myoglobin," Journal of Chemical Physics **113**, 7702–7711 (2000).
[43]E. Wilson, J. Decius, and P. Cross, *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*, Dover Books on Chemistry Series (Dover Publications, 1980).
[44]G. Chevrot, P. Calligari, K. Hinsen, and G. R. Kneller, "Least constraint approach to the extraction of internal motions from molecular dynamics trajectories of flexible macromolecules," Journal of Chemical Physics **135** (2011), 10.1063/1.3626275.
[45]A. Meenakshi and C. Rajian, "On a product of positive semidefinite matrices," Linear Algebra and its Applications **295**, 3 – 6 (1999).
[46]L. Reichl, *A Modern Course in Statistical Physics* (Wiley, 1998).
[47]H. Kramers, "Brownian motion in a field of force and the diffusion model of chemical reactions," Physica **7**, 284 – 304 (1940).
[48]G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, "Hydrophobicity of amino acid residues in globular proteins," Science **229**, 834–838 (1985).
[49]S. Miller, J. Janin, A. M. Lesk, and C. Chothia, "Interior and surface of monomeric proteins," Journal of Molecular Biology **196**, 641 – 656 (1987).
[50]L. Landau and E. Lifshitz, *Fluid Mechanics: Volume 6* (Elsevier Science, 2013).
[51]R. Zwanzig, "Diffusion in a rough potential." Proceedings of the National Academy of Sciences **85**, 2029–2030 (1988), arXiv:1409.4581.
[52]B. Iglewicz and D. Hoaglin, *How to Detect and Handle Outliers*, ASQC basic references in quality control (ASQC Quality Press, 1993).
[53]D. Ruppert, *Statistics and Data Analysis for Finacial Engineering* (Springer, New York, 2010).
[54]G. Bowman, V. Pande, and F. Noé, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Advances in Experimental Medicine and Biology (Springer Netherlands, 2013).
[55]W. C. Swope, J. W. Pitera, and F. Suits, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory," The Journal of Physical Chemistry B **108**, 6571–6581 (2004).
[56]R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, New York, NY, USA, 1986).
[57]B. Peters, "Chapter 14 - discrete stochastic variables," in *Reaction Rate Theory and Rare Events Simulations* (Elsevier, Amsterdam, 2017) pp. 363 – 401.
[58]B. E. Husic, R. T. McGibbon, M. M. Sultan, and V. S. Pande, "Optimized parameter selection reveals trends in markov state models for protein folding," The Journal of Chemical Physics **145**, 194103 (2016), https://doi.org/10.1063/1.4967809.
[59]H. Wu and F. Noé, "Variational approach for learning markov processes from time series data," arXiv preprint arXiv:1707.04659 (2017).
[60]F. Paul, H. Wu, M. Vossel, B. L. de Groot, and F. Noé, "Identification of kinetic order parameters for non-equilibrium dynamics," The Journal of chemical physics **150**, 164120 (2019).
[61]B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, "Estimation and uncertainty of reversible Markov models," Journal of Chemical Physics **143**, 174101 (2015).
[62]C. Gardiner, *Handbook of stochastic methods for physics, chemistry, and the natural sciences*, Springer series in synergetics (Springer, 1994).
[63]C. Baysal and A. R. Atilgan, "Relaxation kinetics and the glassiness of native proteins: Coupling of timescales," Biophysical Journal **88**, 1570 – 1576 (2005).
[64]O. B. Okan, A. R. Atilgan, and C. Atilgan, "Nanosecond motions in proteins impose bounds on the timescale distributions of local dynamics," Biophysical Journal **97**, 2080 – 2088 (2009).
[65]H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, "The energy landscapes and motions of proteins," Science **254**, 1598–1603 (1991).
[66]J. J. Portman, S. Takada, and P. G. Wolynes, "Microscopic theory of protein folding rates. II. Local reaction coordinates and chain dynamics," Journal of Chemical Physics **114**, 5082–5096 (2001), arXiv:0008455 [cond-mat].
[67]Q. Cui and I. Bahar, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, Chapman & Hall/CRC Mathematical and Computational Biology (CRC Press, 2005).
[68]I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, "Global Dynamics of Proteins: Bridging Between Structure and Function," Annual Review of Biophysics **39**, 23–42 (2010).
[69]I. Bahar and R. L. Jernigan, "Cooperative fluctuations and subunit communication in tryptophan synthase," Biochemistry **38**, 3478–3490 (1999).
[70]D. Komander, "The emerging complexity of protein ubiquitination," Biochemical Society Transactions **37**, 937–953 (2009), 164.
[71]D. Komander and M. Rape, "The ubiquitin code," Annual Review of Biochemistry **81**, 203–229 (2012).
[72]Z. Lv, K. M. Williams, L. Yuan, J. H. Atkison, and S. K. Olsen, "Crystal structure of a human ubiquitin e1–ubiquitin complex reveals conserved functional elements essential for activity," Journal of Biological Chemistry **293**, 18337–18352 (2018), http://www.jbc.org/content/293/47/18337.full.pdf+html.
[73]D. D. Boehr, R. Nussinov, and P. E. Wright, "The role of dynamic conformatinal ensembles in biomolecular recognition," Nat Chem Biol. **5**, 789–796 (2009).
[74]J. Monod, J. Wyman, and J.-P. Changeux, "On the nature of allosteric transitions: A plausible model," Journal of Molecular Biology **12**, 88 – 118 (1965).
[75]J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, "Xsede: Accelerating scientific discovery," Computing in Science Engineering **16**, 62–74 (2014).